



Working Paper Series

**Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments**

Hai-Anh Dang  
Dean Jolliffe  
and Calogero Carletto

**ECINEQ WP 2018 - 456**

# Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments\*

Hai-Anh Dang<sup>†</sup>  
Dean Jolliffe  
Calogero Carletto  
*World Bank, USA*

## Abstract

We offer a review of methods that have been employed to provide poverty estimates of poverty in contexts where household consumption data are unavailable or missing. These contexts range from completely missing and partially missing consumption data in cross sectional household surveys, to missing panel household data. We focus on methods that aim to compare trends and dynamic patterns of poverty outcomes over time. We present the various existing methods under a common framework, with pedagogical discussion on the intuition. Empirical illustrations are provided using several rounds of household survey data from Vietnam. Furthermore, we also offer a practical guide with detailed instructions on computer programs that can be used to implement the reviewed techniques.

**Keywords:** poverty, mobility, imputation, consumption, wealth index, synthetic panels, household survey.

**JEL Classification:** C15, I32, O15.

---

\*We would like to thank Alemayehu Ambel, Andrea Brandolini, Grant Cameron, Shatakshee Dhongde, Sam Freije-Rodriguez, John Gibson, Peter Lanjouw, Haizheng Li, Michael Lokshin, David Newhouse, Franco Peracchi, Kinnon Scott, Jacques Silber, Alessandro Tarozzi, Renos Vakis, and participants at the China meeting of the Econometric Society (Wuhan) and seminars at Georgia Tech and World Bank for helpful discussions on earlier versions. We are grateful to the UK Department of International Development for funding assistance through its Knowledge for Change (KCP) Research Program.

<sup>†</sup>**Contact details:** H. Dang, economist in the Survey Unit, Development Data Group, World Bank, and a non-resident senior research fellow with Vietnam's Academy of Social Sciences; email: [hdang@worldbank.org](mailto:hdang@worldbank.org). D. Jolliffe, lead economist in the Survey Unit, Development Data Group, World Bank; email: [djolliffe@worldbank.org](mailto:djolliffe@worldbank.org). Carletto is the manager of Center for Development Data, Rome, Italy and World Bank; email: [gcarletto@worldbank.org](mailto:gcarletto@worldbank.org).

## I. Introduction

Poverty reduction consistently ranks among the most prioritized tasks of developing countries as well as the international community. For example, the Sustainable Development Goals (SDGs) recently adopted by the United Nations General Assembly call for eliminating poverty by 2030 in its very first goal. But effective and efficient poverty eradication entails the prerequisite of accurate poverty measurement, since it would be hardly possible to claim progress with decreasing poverty—or perhaps good performance with diminishing just about almost all other undesirable phenomena—if we cannot measure it well.

Three questions usually come up in the context of poverty measurement. The first question concerns, unsurprisingly, what represents the best yardstick that reliably measures poverty. For example, should it be household consumption (or income), or household wealth? Second, once we have a good poverty measure, how best to track the trends of poverty over time? Put differently, it is useful to understand the issues that can potentially affect the comparability of poverty estimates over time. Finally, what are the composition of poverty transitions over time? In particular, what is the proportion of the poor in one period that remain poor or escape poverty in the next period? Or what is the proportion of the non-poor that fall into poverty in the next period?

A sizable literature exists on the first question, and some consensus has perhaps been reached that household consumption data can be employed as a good proxy for household living standards—thus such data can provide the types of poverty estimates that are often sought.<sup>1</sup> Yet, the implementation of this idea in practice still remains a challenging task; we thus focus in this

---

<sup>1</sup> Throughout this paper we are focused primarily on consumption as this is the measure of well-being that is most frequently used in the world for assessing poverty status, but we also note that income is used as a measure of well-being in many cases. For example, the 2012 global poverty estimates reported in Ferreira *et al.* (2016) are based on data from 131 countries, of which 99 use consumption as the measure of well-being and 32 use income. In an abuse of terminology, we refer to “consumption” and “income” interchangeably as the measure of household living standards in this paper. See also Ravallion (2016) for a comprehensive discussion on the history of thought on poverty and other measurement approaches.

paper on the remaining other questions. The main reason is simply that household consumption data are often either unavailable, or infrequently collected, particularly for low-income countries.

Using the World Bank's PovCalNet database, which covers the period 1981-2014, we plot in Figure 1 the number of data points of poverty estimates against a country's income level (as measured by its consumption level in the household survey). For better presentation, we also graph the fitted line for the regression of the former outcome on the latter outcome. The estimated slope of this regression line is positive and strongly statistically significant, suggesting that countries with higher incomes more frequently implement household surveys. Indeed, a 10 percent increase in a country's household consumption is associated with almost one-third (i.e., 0.3) more surveys. Figure 1 thus helps highlight the—perhaps paradoxical—fact that poorer countries with a stronger need for poverty reduction also face a more demanding challenge of poverty measurement given their smaller numbers of surveys. This is unsurprisingly consistent with a prevailing perception among some development practitioners that collecting survey data may not be the top priority for many developing countries (see, for example, Devarajan (2013)).

The observed pattern between missing data and national income is also particularly relevant for assessing methods to fill data gaps. As one example, the World Bank monitors the global headcount of people living in extreme poverty and these counts are used by many institutions, including the United Nations, previously for the purposes of monitoring the Millennium Development Goals, and now the Sustainable Development Goals.<sup>2</sup> Jolliffe *et al.* (2015) indicate that the approach used by the World Bank to imputing poverty rates for those countries with missing poverty data is simply to assume that the data are missing at random within each region. So, if poverty data are missing for the Democratic People's Republic of Korea, the global poverty

---

<sup>2</sup> For recent examples of the literature documenting the World Bank's global poverty counts, see Ravallion, Chen and Sangraula (2009) and Ferreira *et al.* (2016).

headcount assumes that the rate of poverty in the Democratic People's Republic of Korea is the average poverty rate in East Asia. The systematic correlation between missing survey data and national income observed in Figure 1 suggests that the assumption of missing at random is untenable, and can result in a downward bias in the global poverty count.

While the availability of household consumption surveys is essential for tracking poverty, the quality of these surveys is no less important. In particular, where household consumption data are available, it should not be taken for granted that these data are comparable over time. For example, Serajuddin *et al.* (2015) find that over the period 2002- 2011, almost one-fifth (i.e., 28) of the 155 countries for which the World Bank monitors poverty data using the WDI database have only one poverty data point.<sup>3</sup> As another example, a recent survey by Beegle *et al.* (2016) points out that just more than half (i.e., 27) of the 48 countries in Sub-Saharan Africa had two or more comparable household surveys for the period between 1990 and 2012. Even countries with a respectable and long-running household survey such as India from time to time run into issues with survey comparability (Deaton and Kozel, 2005; Dang and Lanjouw, forthcoming).<sup>4</sup> Perhaps this lack of survey comparability may be ameliorated if (better) data can be produced more frequently.

A less rosy picture emerges with panel survey data, where far fewer developing countries can afford to collect such data. Tracking the same household (or individual) over time is a costly undertaking for various reasons. For example, households can grow larger with new members (through birth or marriage) or smaller (with previous members dying or migrating) over time, or the whole household can migrate, making it more difficult for interviewers to follow them.

---

<sup>3</sup> In addition, Serajuddin *et al.* (2015) also find that as many as 29 countries do not have any poverty data point in the same period.

<sup>4</sup> Household consumption data may not exactly be comparable across different countries without a common conversion system for prices of goods and services. The International Comparison Program (ICP) is an international effort to address this issue (World Bank, 2015). See also Crossley and Winter (2015) and the 2015 special issue of the *Journal of Development Studies* for further discussion on survey comparability issues.

Consequently, even where panel data exist, the estimates of poverty mobility based on these panel data are often found to be affected by various data quality issues such as national representativeness or cross-country comparability.<sup>5</sup>

Still, financial constraints aside, it is common knowledge that a certain level of technical and logistical capacity is required to implement a high-quality consumption survey. For example, fielding an LSMS-type (Living Standards Measurement Survey) household survey is a demanding task that often involves a number months—if not years—of preparation, which consists of a variety of steps ranging from pre-survey stages (e.g., questionnaire design, sample selection, enumerator training, survey firm selection) to post-survey stages (e.g., data entry, data cleaning and checking, post-survey weighting or stratification).<sup>6</sup> Put differently, available resources may provide the necessary condition, but it is technical/ logistical capacity that forms the sufficient condition for producing high-quality survey data. The latter issue is perhaps a common challenge for which most national statistical agencies are keen on finding a solution.

Furthermore, development practitioners working on smaller-scale projects have been more and more interested in collecting consumption data, say, for project evaluation purposes. Such projects more often than not lack the financial and technical resources, nor the intention, to implement an LSMS-type regular household consumption survey. But properly evaluating such projects would clearly rely on good-quality consumption data. As such, data gaps and the need to come up with innovative measurement methods—particularly those that can be implemented at scale—to fill in these gaps are increasingly receiving more attention.

---

<sup>5</sup> Beegle *et al.* (2016) offer a review of studies that use the existing panel data for African countries and find much variation in the estimates for chronic poverty and transient poverty. Furthermore, it is argued that a considerable proportion of these estimates of poverty mobility are subject to measurement errors in income or consumption (Dercon and Krishnan, 2000; Glewwe, 2012; Lee *et al.*, 2017). See also Dercon and Shapiro (2007) for a brief discussion of various panel data sources for developing countries.

<sup>6</sup> See Grosh and Glewwe (2000) for a systematic treatment of the different steps involved in fielding an LSMS-type survey.

We make several contributions in this paper. On the conceptual front, we offer a review of methods that have been employed to provide estimates of welfare in the absence of consumption data, with a particular focus on comparing trends of poverty outcomes over time. Since the existing studies are presented in a variety of formats and technical approaches, we attempt to consistently present these various methods, some new and some more established, under a common framework. For this purpose, we provide some rather simple, but new theoretical results to help fix ideas for better discussion and also to highlight the nuanced differences between the methods that appears to have received scant, if any, attention in previous studies. For example, most development practitioners would likely implement a full-fledged household consumption survey—if given this choice—to produce poverty estimates, rather than attempt to do so using household assets alone. But to our knowledge, there has been no formal discussion to provide insights into this choice, and possibly other similar choices as well. Our paper attempts to fill in this gap in the literature by offering a review on poverty imputation methods.

Furthermore, we focus on providing the intuition behind each estimation method in an effort to make these techniques accessible to a larger audience. For this objective, we refer to the most important theoretical results in the text, and offer a more detailed technical discussion of these results in Appendix 1. In addition, we mostly review the economics literature in this paper, but will also refer to the statistics literature where it is relevant.

On the empirical front, we provide illustrations using real household survey data from Vietnam. We also offer a practical guide in an appendix offering detailed instructions on computer programs (mostly in Stata) that can be used to implement the reviewed techniques. In an effort to make these methods more accessible, and where there is not yet an established literature, we also aim to provide more specific—and somewhat prescriptive—discussion of best-practices methods.

Given our focus on data-scarce environments, we emphasize methods with operationally feasible implementation, which can be practiced at scale.

We discuss a typology of data scarcity situations, including the overall analytical framework and data source in the next section. This is subsequently followed by more detailed discussion for key data shortage situations, which range from the cross-sectional consumption data being completely missing (Section III) to being partially missing (Section VI). We then discuss the situation of missing panel data in Section V and other related topics in Section VI before concluding in Section VII.

## **II. Typology of Data Scarcity Situations and Imputation Methods**

Given the conceptually broad meaning of household welfare, it is useful to lay out in the beginning the practical strategy that we will employ in this paper. More precisely speaking, we sidestep the theoretical debates about the various approaches to measuring household welfare, and are mostly concerned with the most common practice with measuring household welfare. We will focus accordingly on the money-metric aspect of household welfare that can be measured with household consumption (or income) data that are widely collected from household surveys. In particular, our central premise is that household consumption (or income) data provide the benchmark measure of household welfare.<sup>7</sup> This key assumption helps operationalize and better focus the measurement of household welfare, whereby a household's position in the welfare space can be quantitatively identified by their consumption level. For example, households can be identified as poor if their consumption levels are below a specified threshold (i.e., the poverty line). As another example, since we can rank different households' welfare against each other on a

---

<sup>7</sup> There is perhaps some strong consensus among policy makers regarding this as well. For example, the United Nations most recently decided to monitor the global poverty rate as the proportion of the population living below the international poverty line of US\$1.90 per day. But see also Alkire *et al.* (2015) for a comprehensive discussion of the alternative approach of multi-dimensional poverty.



common scale according to their consumption, we can obtain the relative distribution of household welfare, or measures of inequality.

We next provide an overview of data situations and imputation methods before discussing the analytical framework and data.

## II.1. Overview

### Data Situations

For presentation purposes, we group all situations of missing data into three broad categories as follows (Table 1):

- A. the cross-sectional household consumption data are *completely* missing
- B. the cross-sectional household consumption data are *partially* missing, and
- C. although the cross-sectional consumption data are available, the panel household consumption data are (completely or partially) missing

These categories can also be thought of as being ranked, in a roughly decreasing order, according to their severity of data scarcity. Put differently, we assume that the ideal data scenario is one where we have (high-quality) panel consumption data, the second-best scenario is one where we have cross sectional consumption data (but no panel data), and the least desirable scenario is one where we have no consumption data.

Group A thus represents the most data-scarce situation, where no (cross-sectional) data are collected on household consumption, although some data may be collected on other household characteristics.<sup>8</sup> Clearly, a typical example of the surveys that belong to Group A is non-consumption surveys where no consumption data are collected by design. Group A remarkably covers quite a few surveys that are commonly implemented. For example, these surveys include the popular Demographic and Health Surveys (DHS), (most) Labor Force Surveys (LFS), and other surveys such as school-based surveys. In fact, implementing a household consumption

---

<sup>8</sup> Unless otherwise noted, for brevity we hereafter follow the common practice and do not use the term “cross sectional” when referring to such data, but we explicitly mention the term “panel data” when discussing these data.

module requires considerable resources and logistic arrangement, thus almost all small-scale surveys that are conducted on an *ad hoc* basis would normally fall into this category.

Group B represents data situations that are less data-scarce but, on the other hand, cover a large number of cases. Consequently, it is useful to break this group into the following three smaller sub-cases i) the (cross-sectional) consumption data not comparable across survey rounds, ii) the (cross-sectional) consumption data unavailable in the current survey but available in another related survey, and iii) the (cross-sectional) consumption data unavailable at more disaggregated administrative levels than those in the current survey. Sub-case (i) thus concerns the data situations with most Sub-Saharan African countries and India as discussed earlier. Sub-case (ii) is relevant to all situations where we do not have consumption data in the survey under consideration, but have consumption data in other similar surveys that somehow can be linked to the former. An example of this sub-case is where we have a more recent LFS that has no consumption data but has a similar design to an older household consumption survey.<sup>9</sup>

Given the significantly higher costs of implementing panel surveys, Group C represents situations with least data availability. This is especially true for developing countries. For example, two major household consumption surveys that are commonly employed to provide poverty estimates in China and India—the China Household Income Project (CHIP) survey and the National Sample Survey (NSS)—are both cross-sectional surveys.<sup>10</sup> For countries where panel surveys exist, few such surveys are likely to be representative of the population over a long period of time without a great deal of efforts. One major reason is that the surveyed household unit can

<sup>9</sup> There is thus some overlap between this sub-case and the non-consumption surveys in Group A. But note that we focus on the *absence* of consumption data in Group A, and the *similarity* between consumption surveys and other surveys in Group B. Put differently, we consider non-consumption surveys on their own in Group A but highlight their relationship with other consumption surveys in Group B.

<sup>10</sup> See also Gustafsson, Shi, and Sato (2014) for a detailed discussion of the various consumption surveys in China.

change (e.g., household members can die, split off to form a new household, or simply migrate to another place) and it is very costly to track all household members over time. For example, due to attrition, the percentage of households that remain in the panel Russia Longitudinal Monitoring Survey (RLMS) in the first 10 years after it was fielded is around 60 percent; this figure further decreases by half to 29 percent after another 10 years (Kozyreva, Kosolapov, and Popkin, 2016). But as global living standards are rising, this data situation can be improved as more resources may be invested in fielding panel surveys.

### **Imputation Methods**

We briefly show in Table 1 (last column) the imputation methods that can be used to provide poverty estimates in the absence of consumption data. These imputation techniques vary depending on whether the consumption data in need are cross-sectional or panel. For data situations in Group A, the most commonly used method is to generate a wealth index from household assets and the physical characteristics of the house (e.g., the material of the floor or the wall, or which type of toilet is available). For data situations in Group B, techniques have been developed to offer survey-to-survey imputation (i.e., imputation from one survey to another) for sub-cases (i) and (ii), and survey-to-census imputation (i.e., imputation from a survey into a census) for sub-case (iii). Finally, data situations in Group C can be addressed with recently developed methods that construct synthetic panel data from cross-sectional data, which can substitute for actual panel data to some extent. We return to more detailed discussion of each of these imputation methods in the following section.

## **II.2. Analytical Framework**

Following Deaton and Muellbauer (1980), we assume that a household maximizes utility subject to an income budget constraint that includes choice variables such as quantities of goods, durables, and leisure (or labor supply). These in turn are determined by different factors, such as

household tastes. This results in the common practice in most, if not all, household consumption surveys of constructing total household consumption as an aggregate of consumption of different items such as food, non-food (including clothing, education, and/ or health expenses), durable goods, and housing (Deaton and Zaidi, 2002).

For brevity, let  $x_j$  be a vector of characteristics that represent all these factors, where  $j$  indicates the survey type. More generally,  $j$  can indicate either another round of the same household expenditure survey, or a different survey (census), for  $j= 1, 2$ .<sup>11</sup> Subject to data availability,  $x_j$  can include household variables such as the household head's age, sex, education, ethnicity, religion, language (i.e., which can represent household tastes), occupation, and household assets or incomes. Occupation-related characteristics can generally include whether household heads work, the share of household members that work, the type of work that household members participate in, as well as context-specific variables such as the share of female household members that participate in the labor force.<sup>12</sup> Other community or regional variables can also be added since these can help control for different labor market conditions.

It follows that the following linear model is typically employed in empirical studies to project household consumption on household and other characteristics ( $x$ )

$$y_{ij} = \beta' x_{ij} + \mu_{ij} \quad (1)$$

for household  $i$  in survey  $j$ , for  $i= 1, \dots, N$ . Equation (1) thus provides a standard linear model that can be estimated using most available statistical packages.

### II.3. Data Sources

<sup>11</sup> More generally,  $j$  can indicate any type of relevant surveys that collect household data sufficiently relevant for imputation purposes such as labor force surveys or demographic and health surveys. To make notation less cluttered, we suppress the subscript for each household in the following equations.

<sup>12</sup> Regional characteristics related to macroeconomic trends such as (un)employment rates or commodity prices can also be included if such data are available.

We use the latest three rounds of the household survey data from Vietnam—the Vietnam Household Living Standards Surveys (VHLSSs)—in 2010, 2012, and 2014. Being similar to the Living Standards Measurement Study (LSMS) surveys supported by the World Bank, these surveys are implemented biennially by Vietnam’s General Statistical Office (GSO) and collect rich data on household demographics, education, occupation, assets, and consumption. These surveys are regularly employed by the Government of Vietnam and international organizations to provide estimates on household welfare and poverty measures.<sup>13</sup> Since the VHLSSs collect panel data, these surveys offer an ideal setting for us to evaluate the different imputation methods. The key idea is to construct each welfare measure as if we did not have consumption data, and then evaluate the former against the latter.

In particular, we can construct a wealth index and compare how it performs against the actual consumption data in measuring poverty (Category A), impute consumption from one previous round to the next round and compare the imputed consumption against the actual consumption (Category B), and compare the synthetic panels against the actual panel data (Category C). For this category, a particularly useful feature of the VHLSSs is that it has a rotating panel design, whereby approximately half of the households are followed and half are refreshed in each new survey round. We will make use of this feature to provide poverty estimates using the synthetic panels constructed from the cross-sectional component of the VHLSSs, and then validate these estimates against the “true” rates that are based on the panel component of the VHLSSs.

### **III. Completely Missing Consumption Data**

---

<sup>13</sup> An additional advantage of these data is the availability of community targeting, i.e., a household poverty status as classified by the local government. We will make use of this feature and present it together with the analysis based on consumption data.

In the absence of consumption data, wealth indexes can be constructed that can offer a measure of household welfare. We discuss in this section the construction and some main properties of the wealth index (Section III.1) and the application of the wealth index to tracking household welfare over time (Section III.2).

### III.1. Measuring Welfare with Wealth Indexes

#### Definition

Wealth indexes were used in earlier studies to measure household consumption and poverty (see, e.g., Montgomery *et al.* (2000)), but perhaps the study by Filmer and Pritchett (2001) helps significantly popularize its usage. Non-consumption surveys such as the DHS now automatically offer some version of wealth indexes in all their new data releases. Perhaps the main reason behind its popularity is—compared to the typical survey module that consists of hundreds of consumption items required to construct the consumption aggregate—data on a list of assets are both cheaper and easier to collect.

The idea behind a wealth index is, in fact, rather straightforward: it is a single-variable measure for household wealth that can be used to rank household welfare. In practice, it is essentially some combination of the various components of household wealth such as household assets (e.g., whether the household has a television, a car, and a telephone), the living area of the house, and the physical materials out of which the walls (or roof) of the house are constructed (e.g., whether more durable materials such as cement or bricks or less durable materials such as mud or grass are used). The types of a house's facilities such as the toilet are also commonly used since a better type of toilet such as a flush toilet is often observed to proxy for more wealth than other flimsier ones such as a pot toilet or no toilet at all (i.e., open defecation).

To be more precise, consider a variant of Equation (1) where the left-hand side variable, household consumption  $y_{ij}$ , is now missing, but we have data on household assets  $a_{ij}$ —a subset

of  $x_{ij}$ . Still, we want to generate a wealth index  $w_{ij}$  which offers the best combination of (the elements of the different) household assets  $a_{ij}$ . Our problem can then be expressed as follows

$$\gamma' a_{ij} = w_{ij} \quad (2)$$

where we now place the term  $\gamma' a_{ij}$  on the left-hand side to emphasize that  $\gamma$  are the (vector of) weights we place on the  $a_{ij}$  in an effort to generate the best possible measure of wealth. Note that household assets are just a component of the household and community characteristics  $x_{ij}$  in Equation (1).

There are two common approaches to obtaining these weights: the first is to simply let all of them equal 1 (i.e., using simple aggregation that provides a count of the number of assets a household possesses), and the second is to search for a combination of weights that captures the most variation of the  $a_{ij}$  by statistical techniques such as the principal component (PCA) method. Some remarks are, however, in order. The PCA method is data-driven by definition, and offers weights that vary depending on the specific data set that is analyzed. For example, a motorbike would be given a small and perhaps statistically insignificant weight in a setting where it is a commonplace asset (i.e., all households have a motorbike), but would be given a larger weight in another setting where the opposite is true. On the other hand, the simple aggregation method would apply the same equal weights in all the settings. As such, the PCA method appears to be best suited for analysis that is restricted to one setting, while the simple aggregation method may be better for comparison of results across different settings.<sup>14</sup>

Wealth indexes are a useful substitute for household consumption in the absence of the latter, since almost all surveys these days collect the relevant information on household assets and housing characteristics. However, note that wealth indexes are less accurate than household

---

<sup>14</sup> See, for example, Jolliffe (2002) for a comprehensive treatment of PCA methods.

consumption in proxying for household welfare.<sup>15</sup> Since poverty is a function of household consumption, wealth indexes offer biased estimates as well. These results intuitively follow from our assumption that household consumption provides the benchmark measure of household welfare, but can be formalized in the following proposition and corollary.

**Proposition 1:** *Wealth indexes tend to provide biased estimates of poverty rates (as measured by household consumption).*

**Proof:** Appendix 1.

#### *Illustrative Example 1*

We provide an illustrative example where the wealth index is generated using both the simple aggregation (Table 2, Model 1) and the PCA method (Table 2, Models 2 and 3) on the VHLSSs in 2012 and 2014. (Further details on the implementation are provided in Appendix 2.) Each cell in the first five rows shows the proportion of each quintile of the consumption distribution that is correctly captured by each quintile of the wealth index. These quintile divisions may also be considered as different poverty lines. The list of assets for Model 1 includes (whether the household has) a car, a motorbike, a bicycle, a desk phone, a mobility phone, a DVD player, a television set, a computer, a refrigerator, an air conditioner, a washing machine, and an electric fan. Model 2 adds to Model 1 the construction materials for the house's roof and walls, Model 3 adds to Model 2 the type of water and toilet the household has access to.

Estimation results concur with Proposition 1, where each of the quintiles based on the wealth index can only capture around half of the corresponding quintile based on the consumption distribution. For example, the poorest wealth index quintile in Model 3 can correctly capture only 57 and 47 percent of the poorest consumption quintile respectively in 2012 and 2014. Alternatively, we also show the proportion of the cumulative consumption distribution that is

---

<sup>15</sup> In addition, assets take time to depreciate and to be sold/ bought so they are not as smooth as consumption in capturing household welfare.



correctly captured by the wealth quintiles in Appendix 3, Table 3.1.<sup>16</sup> Estimation results are, however, qualitatively similar where taken altogether, the wealth quintiles can correctly capture between 60 percent and 80 percent of the corresponding consumption quintiles. Notably, the correlation between asset indexes and household consumption is higher for the PCA wealth index than for the simple aggregation method (e.g., this correlation is 0.61 for Model 1 in 2012, but increases to 0.67 and 0.70 respectively for Models 2 and 3 in the same year).

These results are broadly consistent with the empirical evidence offered in recent studies. Reviewing 17 studies that analyze 36 different data sets, Howe *et al.* (2009) observe a poor correlation between wealth indexes and household consumption. Filmer and Scott (2012) analyze survey data from 11 countries from four continents and find a similar result. Furthermore, Filmer and Scott (2012) also find that the correlation between wealth indexes and consumption is only stronger under certain, and more demanding, conditions such as urban settings, limited measurement errors, and a small share of individually consumed goods (e.g., food) in total expenditures.<sup>17</sup> The latter caveats can perhaps also be partly explained by the fact that, wealth indexes are a stock measure, rather than a flow measure as with household consumption.

### III.2. Tracking Welfare over Time with Wealth Indexes

#### *Illustrative Example 2*

As earlier discussed, since wealth indexes are biased estimates of household consumption levels, they appear likely to provide biased estimates of consumption trends as well. We provide

<sup>16</sup> In this case, the poverty line can equal the 20<sup>th</sup>, 40<sup>th</sup>, or 60<sup>th</sup> (and so on) percentile of the wealth index distribution.

<sup>17</sup> Some other practical issues with using wealth indexes to measure poverty are that it is unclear how to set the poverty line for wealth indexes (compared with the more established calorie-intake or cost-of-basic-needs with the consumption-based poverty line) or that asset ownership may also be affected by factors related to supply but not household consumption (e.g., inadequate water supply may result in less ownership of flush toilets or washing machines). An alternative approach is to collect data on a reduced set of consumption items that may offer strong correlation with the total consumption aggregate (Morris *et al.*, 2000). Another approach is to produce households' ranks in the population with the number of consumption items they own, if we make the additional assumption that households place an order of importance on their consumption items when having to reduce their consumption expenditure (Deutsch, Silber, and Wang, 2017).

an illustrative example in Table 3, where we compare the growth of consumption (second column) against those of wealth indexes. We use two versions of the wealth indexes, the first is the simple aggregation method (or wealth index 1), and the second the PCA method (or wealth index 2, same as Model 3 in Table 2) after pooling all data for the three survey years. The list of the assets that are used is the same as that used for Model 1 in Table 2.<sup>18</sup> Table 3 shows that the growth rate of mean per capita consumption widely differs from that of the wealth indexes between 2010 and 2012: while the former is around 2 percent, the growth rate of wealth index 1 is higher at 10 percent, and even negative and much higher at more than 100 percent for wealth index 2 (Panel A). The growth rate of mean per capita consumption and the growth rate of the wealth indexes differs as well between 2012 and 2014 (Panel B), but a different result emerges where the former is now larger than that of wealth index 1 but still smaller than that of wealth index 2.

This empirical result qualitatively concurs with that offered by Harttgen, Klasen, and Vollmer (2013), who analyze 160 DHS surveys from 33 African and 34 non-African countries and construct wealth indexes in different ways as well. In particular, Harttgen *et al.* (2013) argue that employing wealth indexes as a proxy for trends in household consumption is subject to several different types of biases. For example, one is due to changing preferences for certain assets (e.g., the increasing ownership of smart phones), another changing relative prices among different assets leading to more demand for one asset at the expense of others (e.g., the dramatically decreasing price of smart phones).<sup>19</sup>

<sup>18</sup> The reason we pool the data is to make the wealth indexes comparable over time, since applying the PCA method to each year separately as with Table 2 would standardize these indexes to a mean 0 for each year. We use the same assets as in Model 1, Table 2 for easier interpretation with the count index; estimation results (not shown) using additional assets in the other models (Models 2 and 3) provide qualitatively similar results.

<sup>19</sup> Although providing biased poverty estimates in a static period, wealth indexes may not necessarily provide biased estimates of trends in household consumption (or poverty) over time. If the degree of bias for both wealth indexes and household consumption is similar over time, wealth indexes can provide unbiased estimates. As an (extreme) example, if the bias of the mean of a wealth index as a measure of mean household consumption remains constant at 10 percent

Still, we end this section on an optimistic note that more methods have recently been developed in an effort to provide more comparable wealth indexes over time. For example, a promising direction of research is offered by Rutstein and Staveteig (2014), who adjust, relative to a reference country, a country-specific wealth index using the country-level relationship between some “unsatisfied basic needs” and ownership of certain basic assets such as a car, a refrigerator, a landline telephone, and a television. This study builds on earlier studies that employ asset indexes to analyze poverty trends over time (e.g., Sahn and Stifel (2000)), but it does not use consumption data to validate its estimates based on the asset indexes.

#### IV. Partially Missing Consumption Data

To further operationalize our estimation framework, we extend Equation (1) to a more general model

$$y_j = \beta_j' x_j + v_{cj} + \varepsilon_j \quad (3)$$

where the error term  $\mu_{ij}$  is now broken down into two components, one ( $v_{cj}$ ) a cluster random effects and the other ( $\varepsilon_j$ ) the idiosyncratic error term. Note that we also suppress the subscript that indexes households to make the notation less cluttered. Conditional on household characteristics, the cluster random effects and the error terms are usually assumed uncorrelated with each other and to follow a normal distribution such that  $v_{cj}|x_j \sim N(0, \sigma_{v_j}^2)$  and  $\varepsilon_j|x_j \sim N(0, \sigma_{\varepsilon_j}^2)$ . While the normal distribution assumption results in the standard linear random effects model that is more convenient for mathematical manipulations and computation, it is not necessary for this type of model. As can be seen later, we can remove this assumption and use the empirical distribution of the error terms instead, albeit at the cost of somewhat more computing time.

---

for all time periods, this bias would cancel out for estimates of the trends in consumption—or the relative growth of the mean. We provide more formal discussion of this result in the proof for Proposition 1.

Without loss of generality, assume for now that consumption data are available for survey 1 but missing for survey 2, for which we are most interested in the poverty estimates. Let  $z_2$  be the poverty line in period 2, and  $y_2^1$  be the imputed consumption for survey 2, the poverty rate  $P_2$  in this period could be estimated with the following quantity

$$P(y_2^1 \leq z_2) \quad (4)$$

where  $P(\cdot)$  is the probability (or poverty) function that gives the percentage of the population that are under the poverty line  $z_2$  in survey 2.

Imputation methods differ in how they estimate  $y_2^1$ . We discuss in this section methods that have been developed to impute consumption, either within the same type of surveys or across different survey types. We start first with discussing a commonly used method, proxy means testing, before discussing how consumption can be imputed to measure poverty trends and “poverty-mapping” (or small area estimation) methods.

#### IV.1. Proxy Means Testing

Most of the estimates based on proxy means testing are usually estimated as

$$y_{ij}^p = \beta_j^{p'} x_{ij,p} \quad (5)$$

where the vector of coefficients  $\beta_j^p$  is often obtained from the regression using another data set (see, e.g., Grosh *et al.* (2008), Coady *et al.* (2014), Brown, Ravallion, and van de Walle (2016)). For example,  $\beta_j^p$  can be obtained from a regression using the household consumption data, and then be imposed on the data from a special—and often smaller—survey that aims at targeting poor households for a social protection program. Alternatively,  $\beta_j^p$  can also be obtained from a regression using data from a neighboring country. But regardless of the specific application, the error terms  $v_{cj} + \varepsilon_j$  (in Equation (3)) are often omitted in Equation (5). As a result, the mean and the variance of the predicted consumption based on proxy means testing would likely provide

biased estimates of those of household consumption. Put differently, the better  $x_{ij,p}$  can capture the variables in  $x_{ij}$ , the less biased the predicted consumption-based proxy means tests is. When  $x_{ij,p}$  is identical to  $x_{ij}$ , (that is, the estimation model for Equation (5) is fully specified except for the error terms  $v_{cj} + \varepsilon_j$ ), there is no bias in the estimated mean consumption, but there is still bias in the estimated variance.

These results have direct implications for poverty estimation, and are formally stated in Proposition 2 below.

**Proposition 2:** *Proxy means testing tends to offer biased poverty estimates.*

**Proof:** (Appendix 1).

As discussed above, if the estimation model is fully specified (i.e.,  $x_{ij,p}$  is a smaller subset of  $x_{ij}$ ), there is no bias in the estimated mean consumption. This result, however, generally does not translate directly into poverty estimates, unless the poverty line is exactly set at the mean consumption level. Nevertheless, proxy means testing offers better estimates of household consumption than those offered by wealth indexes. The intuition is rather straightforward: since we employ an estimation model that is more closely related to household consumption (e.g.,  $\beta_j^p$  is obtained from a regression using household consumption as the dependent variable) with the former, we should expect to have better estimates for household consumption.<sup>20</sup>

Since the empirical example for proxy means testing is closely related to our discussion of imputed consumption, we return to more discussion in the next section.

## IV.2. Monitoring Poverty Trends over Time with Imputed Consumption

The United Nations—through the Sustainable Development Goals (SDGs)—recently set the ambitious targets of eliminating extreme poverty and reducing national poverty levels at least by

---

<sup>20</sup> We offer more formal discussion of these results in the proof for Proposition 2 in Appendix 1.

half by 2030.<sup>21</sup> Such efforts are predicated on an ability to reliably assess and monitor progress since tracking poverty trends can help us understand which policies work and which do not work, and how efficient they are. Producing reliable poverty estimates by conducting household expenditure (consumption) or income surveys, however, requires significant financial and technical resources. As a result, consumption surveys are typically conducted every few years by statistical agencies, and poverty estimates are not available in the intervening years during which surveys have not been implemented. Another challenge to tracking poverty trends is that questionnaire design may change over time, thus making consumption data and poverty estimates not comparable between different rounds.

A notable example of incomparable consumption data over time is perhaps India's NSSs. In the late 1990s, the National Sample Survey Office (NSSO) revised the questionnaire of the NSS in 1999/2000 (55<sup>th</sup> round) to make household survey consumption data more consistent with those from national accounts. Major changes to the questionnaire were implemented such as changing the recall period for household durables and education expenses from a 30-day interval to a 365-day interval, and using both the traditional 30-day recall period as well as a new 7-day recall period for food items. The Government of India subsequently estimated that the poverty rate fell by 10 percentage points between 1993/1994 and 1999/2000. However, researchers provided different estimates ranging from only somewhat lower than the official estimates to a mere three percentage point decline in poverty during the decade of the 1990s (see, e.g., Deaton and Kozel (2005)).<sup>22</sup>

Where consumption data are either incomparable across two survey rounds or missing in one survey round but not the other, but other characteristics ( $x_j$ ) that can help predict consumption data

---

<sup>21</sup> For details see <https://sustainabledevelopment.un.org/sdg1>.

<sup>22</sup> See also Dang and Lanjouw (forthcoming) for a discussion of more recent, albeit less severe, comparability issues with the NSSs in the late 2000s.

are available in both survey rounds, we can apply survey-to-survey imputation methods. In particular, Dang, Lanjouw, and Serajuddin (2017) propose a simple imputation framework that builds on earlier studies (Elbers, Lanjouw, and Lanjouw, 2003; Tarozzi, 2007).<sup>23</sup> Compared to previous studies, the Dang *et al.* (2017) method provides a more explicit theoretical modeling framework, with new features such as model selection and standardization of surveys of different designs (e.g., for imputing from a household survey into a labor force survey).

Assume that the explanatory variables  $x_j$  are comparable for both surveys (Assumption IV.1), and that the changes in  $x_j$  between the two periods can capture the change in poverty rate in the next period (Assumption IV.2). Suppressing the subscript for households in the following equations, Dang *et al.* (2017) define the imputed consumption  $y_2^1$  as

$$y_2^1 = \beta_1' x_2 + v_1 + \varepsilon_1 \quad (6)$$

and estimate it as

$$\hat{y}_{2,s}^1 = \hat{\beta}_1' x_2 + \tilde{v}_{1,s} + \tilde{\varepsilon}_{1,s} \quad (7)$$

where the parameters  $\beta_1'$  are estimated using Equation (3), and  $\tilde{v}_{1,s}$  and  $\tilde{\varepsilon}_{1,s}$  represent the  $s^{\text{th}}$  random draw from their estimated distributions, for  $s = 1, \dots, S$ . The poverty rate  $P_2$  in period 2 and its variance can then be estimated as

$$\text{i) } \hat{P}_2 = \frac{1}{S} \sum_{s=1}^S P(\hat{y}_{2,s}^1 \leq z_1) \quad (8)$$

$$\text{ii) } V(\hat{P}_2) = \frac{1}{S} \sum_{s=1}^S V(\hat{P}_{2,s} | x_2) + V\left(\frac{1}{S} \sum_{s=1}^S \hat{P}_{2,s} | x_2\right) \quad (9)$$

<sup>23</sup> Elbers *et al.* (2003) provide a method that imputes household consumption from a survey into a population census. Adapting this approach for survey-to-survey imputation, Christiaensen *et al.* (2012) impute poverty estimates using data from several countries, including China, Kenya, the Russian Federation, and Vietnam; other studies analyze data from Uganda (Mathiassen, 2013).

Indeed, these estimates of poverty improve on those offered by proxy means testing method. This result is formalized in the following proposition.

**Proposition 3:** *Given assumptions IV.1 and IV.2, the imputed household consumption in equation (7) offers unbiased poverty estimates.*

**Proof:** Appendix 1.

We can test for Assumption IV.2 if consumption data exist in both survey rounds. We can use a decomposition that is similar in spirit to the Oaxaca-Blinder framework (Oaxaca, 1973; Blinder, 1973), where the change in poverty between the survey rounds can be broken down into two components, one due to the changes in the estimated coefficients (the first term in square brackets in Equation (10) below) and the other the changes in the  $x$  characteristics (the second term in square brackets in Equation (10) below). Assumption IV.2 would be satisfied if the poverty change is mostly explained by the latter component. This can be expressed as

$$\begin{aligned} P(y_2) - P(y_1) &= [P(y_2) - P(y_2^1)] + [P(y_2^1) - P(y_1)] \\ &= [P(\beta_2' x_2 + \eta_2) - P(\beta_1' x_2 + \eta_1)] + [P(\beta_1' x_2 + \eta_1) - P(\beta_1' x_1 + \eta_1)] \end{aligned} \quad (10)$$

where  $\eta_j$  is defined as  $\nu_j + \varepsilon_j$ ,  $j = 1, 2$ , for less cluttered notation.

However, we want to predict poverty rate for the current period where consumption data are missing. Thus a practical use of Assumption IV.2 is model selection; in particular, we can test for this assumption in contexts where consumption data exist for earlier survey rounds (i.e., there are two earlier survey rounds with consumption data), and select the imputation model that offers estimation results that are most satisfactory for this assumption.

This imputation method is also related to a larger literature on missing data (or multiple) imputation (MI) in statistics (see, e.g., Little and Rubin, 2002; Carpenter and Kenward, 2013). Official agencies such as the U.S. Census Bureau routinely use imputation to fill in important missing data on various statistics for income (Census Bureau, 2014a) and labor (Census Bureau, 2014b). However, several differences exist between the two literatures. First, MI studies often



employ Bayesian techniques for their estimation, which requires multiple drawing from posterior distributions and consequently, more computing time. Second, MI studies often impute missing data within the same survey where usually less than half of the data are missing, rather than imputing consumption for a whole new survey (census) round. Finally, MI studies, unsurprisingly, do not use economic theory for model selection (as in Dang *et al.*, 2017).<sup>24</sup>

A recent champion of the application of MI methods to poverty is the SWIFT method (Survey of Well-being via Instant and Frequent Tracking), which is developed by Yoshida *et al.* (2015). Besides sharing the features of MI methods discussed above, perhaps SWIFT's most important contribution is its proposal to rapidly collect data on a key and reduced set of variables that can be used for faster and less expensive poverty imputation.<sup>25</sup>

However, it can be useful to note some potential limitations of SWIFT. First, the (distributions of the) parameters  $\beta_1$ ,  $\nu_1$ , and  $\varepsilon_1$  are likely to change more over longer time periods. Consequently, these parameters would need to be updated more frequently for better predictions, which would effectively require the implementation of the full household consumption survey rather than just a reduced set of variables. Second, a full household survey (with or without consumption data) may likely offer more precision with reproducing the whole income distribution

<sup>24</sup> Doudich *et al.* (2016) offer an early application of MI methods to poverty imputation. See also Davey, Shanahan, and Schafer (2001) and Jenkins *et al.* (2011) for studies that apply MI techniques to economic issues. More discussion on the differences between the two literatures and references to earlier imputation studies are provided in Dang *et al.* (2017).

<sup>25</sup> Due to various post-survey data quality control procedures such as data entry, cleaning and checking, household survey data are usually unavailable for use from until half a year to longer with most household surveys. But this situation may improve with the increased use of tablet-based (or hand-held devices-based) data collection. Another notable feature of SWIFT is its statistical approach to model selection (see, e.g., James *et al.* (2013)), and building estimating models (say, by using stepwise regression to screen out variables with low p-values, rather than selecting these variables based on economic theory on household consumption).

and related welfare measures (such as inequality measures). Furthermore, more validation of this method is needed.<sup>26</sup>

We return to more discussion with the illustrative example below.

### *Illustrative Example 3*

We start with checking Assumptions IV.1 and IV.2 before discussing estimation results. Since all the 2010, 2012, and 2014 rounds of the VHLSS share the same sampling frame based on the 2009 Population and Housing Census, and their questionnaire design remains almost identical,<sup>27</sup> Assumption IV.1 for a similar survey design is satisfied. We can test for Assumption IV.2 for the period 2010-2012 using the decomposition provided in Equation (10).

We start first with building the estimation models on a cumulative basis for illustration purposes, with later models sequentially adding more variables to earlier models. Model 1 is the most parsimonious model and consists of household size, household heads' age, gender, highest completed levels of schooling, a dummy variable indicating whether the head belongs to the ethnic majority group, and a dummy variable indicating urban residence. Model 2 adds to Model 1 the household demographics such as the shares of household members in the age ranges 0-14, 15-24, and 25-59 (with the reference group being those 60 years old and older), a dummy variable indicating whether the head worked in the past 12 months. Model 3 adds to Model 2 asset variables, the construction materials for the house's roof and walls, and the type of water and toilet the household has access to, which are the same as those employed in our earlier analysis of wealth indexes. Full model specifications are provided in Appendix 3, Table 3.2.

<sup>26</sup> This practically implies that SWIFT has to rely more heavily on the assumption of constant parameters than most other previous studies. As yet, there appear to be no published validation studies besides those offered in Yoshida *et al.* (2015).

<sup>27</sup> The VHLSSs have a rotating module that collect more detailed data on certain topic in each survey round. For example, the 2014 VHLSS collects more data on land access and ownership. But more importantly for our purposes, the core modules that collect data on household demographics, education, assets, and house materials remain the same over these survey rounds.

We then predict consumption, using the estimated coefficients from the preceding years (i.e., the estimated coefficients are respectively from 2010 and 2012 for our predicted consumption in 2012 and 2014). Decomposition results, shown in Table 3.3, suggest that as the list of control variables becomes richer, the change in poverty that can be explained by the  $x$  characteristics grows proportionately larger. For example, this component increases from around 60 percent in Model 1 to 62 percent in Model 2, and then close to 100 percent in Model 3 for the period 2010-2012. For the period 2012-2014, Models 1 and 2 do not perform as well with less than 20 percent of the change in poverty being explained by the  $x$  characteristics under these models. However, Model 3 offers a good estimation model and can explain almost all the change in poverty.<sup>28</sup> This indicates that Assumption 2 is largely satisfied with Model 3 for both periods, and less likely to be satisfied with the remaining models.

Table 4 provides the estimation results for the predicted poverty rates based on the imputed data. Estimation results show that our estimates using Model 3 indeed fall within one standard error of the true poverty rates for both 2012 and 2014. For example, our predicted poverty rate for 2014 is 13.1 percent using the normal linear regression model, which is inside the one-standard-error interval of the true poverty rate of [13.0, 14.0] for the same year. As a further robustness check, we also show the estimation results where we predict consumption for 2014, but use the estimated coefficients from 2010 (instead of those from 2012). The longer time interval may reduce the precision of the estimates if the changes in the estimated coefficients dominate the

---

<sup>28</sup> The switch from positive to negative of these models can suggest that Assumption IV.2 is more flexible compared to the commonly made assumption of constant parameters in previous studies. This is further supported by the Wald test that rejects the latter assumption for all models (Table 3.3). But also note that model specifications where the changes in the explanatory variables  $x$  can explain much more than 100 percent of the changes in poverty may also indicate model overfitting. Furthermore, Dang *et al.* (2017) observe that it is generally ill-advisable to include certain assets whose correlations with consumption change dramatically over the two periods due to other factors such as technology. A notable example is that in certain developing countries cell phones could get mass produced quickly and their prices were lowered to the extent that they could no longer be considered a luxury good in the second period.

change in the  $x$  characteristics. However, the estimates shown under Model 3 in Table 3.4 show that we have similar, or even somewhat better, results.

For further comparison, we also provide estimates using MI methods in Table 3.5. Two imputation models are used, one is the normal linear regression and the other the predictive mean matching, which offer the closest corresponding modeling options to the normal linear regression model and the empirical distribution of the error terms model provided in Table 4. Estimation results suggest that both methods, in particular the predictive mean matching method, work reasonably well (using our preferred Model 3). In fact, the predictive mean matching method may offer even slightly better estimates than those in Table 4. However, there are two limitations with MI methods: the variance for MI estimates are larger than those offered in Table 4, and it takes more computer time to obtain these estimates.<sup>29</sup>

It can be useful to briefly discuss estimation results based on proxy means testing. As discussed earlier, proxy means testing would likely provide biased poverty estimates (Proposition 2). The intuition behind this result is that predicted poverty does not appropriately adjust for the error term, and is based on the deterministic part of Equation 1 only (see Equation 5). Indeed, estimation results (Table 3.6) show that the estimated poverty rates are much lower than the true poverty rates, with the difference ranging around 5 percentage points for Model 3 in both years.

### IV.3. Spatial Targeting with Poverty Map

While tracking poverty trends over time requires the temporal imputation of poverty, monitoring poverty on a larger and more disaggregated scale would require the spatial imputation of poverty. Government agencies in richer countries such as the U.S. Census Bureau regularly rely

---

<sup>29</sup> Furthermore, as earlier discussed, these estimates are based on Model 3, which is selected using the method proposed in Dang *et al.* (2017). These results are also consistent with MI estimation results using household survey data from Jordan offered in this study.

on statistical techniques known as “small-area” estimation methods (see, e.g., Bell *et al.* (2007)) to impute poverty numbers at different administrative levels. These methods essentially impute from a household (income or) consumption survey into a population census to provide more spatially disaggregated measures of consumption and poverty for better targeting purposes. Put differently, consumption data are imputed using a data source that is representative at a higher administrative level to another data source that is representative at a lower administrative level. The higher administrative level can be the state or province level, for example, while the lower administrative level can be the district or community level.

This approach has also been widely employed by international organizations such as the World Bank to impute poverty in a developing country context. Since poverty rates are often visually represented on the country map, it is also commonly known as the “poverty mapping” approach. The most commonly used framework developed by Elbers, Lanjouw, and Lanjouw (2003) proposes that we estimate Equation (3) using the household consumption survey. Then these estimated (distributions of the) parameters can be imposed on the same variables in the population census in a similar spirit as with the temporal imputation of poverty. But importantly, techniques differ on how this is implemented. In particular, Elbers *et al.* (2003) suggest a Bayesian estimation method where the distribution of  $\hat{\beta}'_1$ , rather than  $\beta'_1$ , is employed for the imputation. In fact, various methods have also been employed for the imputation model.<sup>30</sup>

However, Tarozzi and Deaton (2007) note that two assumptions are relevant for the implementation of the poverty map method. The first assumption is similar to Assumption IV.1 discussed earlier, and states that the explanatory variables  $x_j$  are comparable for both the survey

---

<sup>30</sup> The Elbers *et al.* method has also been employed to provide other development outcomes such as child malnutrition (Fujii, 2010). See also Rao and Molina (2015) for a text-book treatment of the statistical literature on small area estimation.

and the census. The second assumption is more specific, and states that the conditional distribution of household consumption given  $x_j$  is the same for both the survey and the census. This latter assumption is in fact indispensable for the inference from the estimates for the larger areas (as representative in the survey) to the smaller areas (as representative in the survey).

#### *Illustrative Example 4*

Figure 2—produced by the World Bank Poverty Team in Vietnam—provides a map of poverty estimates by imputing from the VHLSS in 2008 into the population census in 2009.<sup>31</sup> The differing poverty rates at the province level are depicted by the color intensity, where a darker color indicating a higher poverty rate. Figure 2 shows that provinces with the highest concentration of poverty are in the Northern region, the North West region, and the Central Highland. On the other hand, provinces in the coastal regions including the Red River Delta and the Mekong Delta have the lowest poverty rates.

### **V. Missing Panel Consumption Data**

We now turn the last category of missing panel data (Table 1, Category C). New methods have recently been proposed to deal with situations with missing panel household consumption by constructing synthetic panels from cross sections that can help provide insights into the dynamics of poverty (Dang *et al.*, 2014; Dang and Lanjouw, 2013) and other related welfare outcomes such as vulnerability and shared prosperity (Dang and Lanjouw, 2016 and 2017). These synthetic panels have also been increasingly employed in a variety of contexts.<sup>32</sup>

<sup>31</sup> See Demombynes (2015) for an introduction on this map.

<sup>32</sup> Recent applications and further validations include Ferreira *et al.* (2013), Cruces *et al.* (2015) and Vakis *et al.* (2015) for Latin American countries, Martinez *et al.* (2013) for the Philippines, Garbero (2014) for Vietnam, Bourguignon and Moreno (2015) and Foster and Rothbaum (2015) for Mexico, Cancho *et al.* (2015) for countries in Europe and Central Asia, Dang and Ianchovichina (2016) for countries in the Middle East and North Africa region, Dang and

Let  $x_{ij}$  be a vector of household characteristics observed in survey round  $j$  ( $j= 1$  or  $2$ ) that are also observed in the other survey round for household  $i$ ,  $i= 1, \dots, N$ . These household characteristics can include such time-invariant variables as ethnicity, religion, language, place of birth, parental education, and other time-varying household characteristics if retrospective questions about the round-1 values of such characteristics are asked in the second-round survey. To reduce spurious changes due to changes in household composition over time, we usually restrict the estimation samples to household heads age, say 25 to 55 in the first cross section and adjust this age range accordingly in the second cross section. Note that this age range is usually used in traditional pseudo-panel analysis but can vary depending on the cultural and economic factors in each specific setting.

Let  $y_{ij}$  represent household consumption or income in survey round  $j$ ,  $j= 1$  or  $2$ . The linear projection of household consumption (or income) on household characteristics for each survey round is given by

$$y_{ij} = \beta_j' x_{ij} + \varepsilon_{ij} \quad (11)$$

Let  $z_j$  be the poverty line in period  $j$ . We are interested in knowing the unconditional measures of poverty mobility such as

$$P(y_{i1} < z_1 \text{ and } y_{i2} > z_2) \quad (12)$$

which represents the percentage of households that are poor in the first survey round (year) but nonpoor in the second survey round, or the conditional measures such as

$$P(y_{i2} > z_2 \mid y_{i1} < z_1) \quad (13)$$

---

Dabalen (forthcoming) and Dang, Lanjouw and Swinkels (2017) for countries in Sub-Saharan Africa, and Dang and Lanjouw (2017 and forthcoming) for India, Vietnam, and the United States.

which represents the percentage of poor households in the first round that escape poverty in the second round.

If panel data are available, we can estimate the quantities in (12) and (13); but in the absence of such data, we can use synthetic panels to study mobility. To operationalize the framework, we make two standard assumptions. First, we assume that the underlying populations being sampled in survey rounds 1 and 2 are identical such that their time-invariant characteristics remain the same over time (Assumption V.1). More specifically, coupled with Equation (11), this implies the conditional distribution of expenditure in a given period is identical whether it is conditional on the given household characteristics in period 1 or period 2 (i.e.,  $x_{i1} = x_{i2}$  implies  $y_{i1}|x_{i1}$  and  $y_{i1}|x_{i2}$  have identical distributions). Second, we assume that  $\varepsilon_{i1}$  and  $\varepsilon_{i2}$  have a bivariate normal distribution with positive correlation coefficient  $\rho$  and standard deviations  $\sigma_{\varepsilon_1}$  and  $\sigma_{\varepsilon_2}$  respectively (Assumption V.2). Quantity (12) can be estimated by

$$P(y_{i1} < z_1 \text{ and } y_{i2} > z_2) = \Phi_2 \left( \frac{z_1 - \beta_1' x_{i2}}{\sigma_{\varepsilon_1}}, -\frac{z_2 - \beta_2' x_{i2}}{\sigma_{\varepsilon_2}}, -\rho \right) \quad (14)$$

where  $\Phi_2(\cdot)$  stands for the bivariate normal cumulative distribution function (cdf) (and  $\phi_2(\cdot)$  stands for the bivariate normal probability density function (pdf)). Note that in Equation (14), the estimated parameters obtained from data in both survey rounds are applied to data from the second survey round ( $x_2$ ) (or the base year) for prediction, but we can use data from the first survey round as the base year as well. It is then straightforward to estimate quantity (13) by dividing quantity

(12) by  $\Phi \left( \frac{z_1 - \beta_1' x_{i2}}{\sigma_{\varepsilon_1}} \right)$ , where  $\Phi(\cdot)$  stands for the univariate normal cumulative distribution function (cdf).



In Equation (14), the parameters  $\beta_j$  and  $\sigma_{\varepsilon_j}$  are estimated from Equation (1), and  $\rho$  can be estimated using an approximation of the correlation of the cohort-aggregated household consumption between the two surveys ( $\rho_{y_{c1}y_{c2}}$ ). In particular, given an approximation of  $\rho_{y_{c1}y_{c2}}$ , where  $c$  indexes the cohorts constructed from the household survey data, the partial correlation coefficient  $\rho$  can be estimated by

$$\rho = \frac{\rho_{y_{i1}y_{i2}} \sqrt{\text{var}(y_{i1}) \text{var}(y_{i2})} - \beta_1' \text{var}(x_i) \beta_2}{\sigma_{\varepsilon_1} \sigma_{\varepsilon_2}} \quad (15)$$

Note that the standard errors of estimates based on the synthetic panels can in fact be even smaller than that of the true (or design-based) rate if there is a good model fit (or the sample size in the target survey is significantly larger than that in the base survey; see Dang and Lanjouw (2013) for more discussion).

Equation (14) can be extended to the more general case of vulnerability. For example, we can estimate the percentage of poor households in the first period that escape poverty but still remain vulnerable in the second period (joint probability) as

$$P(y_{i1} < z_1 \text{ and } z_2 < y_{i2} < v_2) = \Phi_2\left(\frac{z_1 - \beta_1' x_{i2}}{\sigma_{\varepsilon_1}}, \frac{v_2 - \beta_2' x_{i2}}{\sigma_{\varepsilon_2}}, \rho\right) - \Phi_2\left(\frac{z_1 - \beta_1' x_{i2}}{\sigma_{\varepsilon_1}}, \frac{z_2 - \beta_2' x_{i2}}{\sigma_{\varepsilon_2}}, \rho\right) \quad (16)$$

Other formulae and more detailed derivations for other measures of vulnerability dynamics are provided in Dang and Lanjouw (2017).

#### *Illustrative Example 5*

Table 5 provides the estimates for both unconditional and conditional poverty mobility using the synthetic panels, and then compares these estimates with those based on the true panels. Several remarks are in order for this table. First, the synthetic panel estimates approximate those of the

true panel estimates reasonably well. In particular, all the point estimates based on the synthetic panels fall within the 95 percent confidence intervals (CIs) of those based on the true panels, and even fall within one standard error of the latter in one case. That is, the unconditional probability estimate for those who remain poor in both years is 10.8 percent, which is almost identical to the corresponding estimate based on the true panel of 10.7 percent. Second, all the 95 percent CIs of the synthetic panel estimates overlap with those of the true panels for at least half or more. This is due to the fact that the synthetic panel estimates are model-based estimates, thus these estimates generally should have smaller standard errors than those using the actual panels (which are usually referred to in the survey statistics literature as the design-based estimates). Finally, estimates for the conditional probabilities are somewhat less accurate than those for the unconditional probabilities since obtaining the latter entails prediction for both the numerator and the denominator of (12)-type quantities.<sup>33</sup>

## VI. Other Topics

We briefly review some other issues that are related to poverty measurements that are increasingly receiving attention.

### *Measurement Errors*

Measurement errors are known to be a common issue with micro survey data, and can bias poverty estimates (see, e.g., Deaton, 1997; Chesher and Schluter, 2003). Particularly with low-quality panel data sets, these errors may result in spurious movements that can be incorrectly attributed to poverty mobility. While some researchers argue that up to 50 percent of the transitory poverty may be accounted for by measurement error in income or consumption (Dercon and Krishnan, 2000; Glewwe, 2012; Lee *et al.*, 2017), others suggest the magnitudes of bias rely on

---

<sup>33</sup> This is also consistent with the theoretical and empirical evidence provided in Dang and Lanjouw (2013).

the particular form of errors (Gottschalk and Huynh, 2010).<sup>34</sup> Thus much research still appears to be needed on this topic. One promising direction can be the application of synthetic panels to correct for “bad” panel data.

### *Big Data*

Various Big Data sources have been proposed as viable substitutes for missing household consumption data. For example, Jean *et al.* (2016) combine nighttime maps with high-resolution daytime satellite images to provide poverty maps for five African countries, while Blumenstock *et al.* (2015) use mobile phone records to estimate poverty. Most recently, Steele *et al.* (2017) employ cellphone data and satellite data to evaluate poverty mapping for Bangladesh. These recent advances thus offer a promising avenue where intensive machine learning techniques can be introduced to substitute for traditional econometric methods.

### *Subjective Poverty*

Subjective well-being data can be employed to provide richer analysis of poverty. In fact, it can be argued that if poverty is multi-dimensional, subjective assessment of poverty is an integral part of its measurement. It is even proposed that a “social subjective poverty line”, below which people tend to think they are poor, but above which they do not, can be a conceptual alternative to defining poverty (Ravallion, 2014). But on the other hand, some empirical evidence suggests that poverty may not necessarily overlap with unhappiness in a number of developing countries including India, Mexico, Peru, and the Russian Federation (Banerjee and Duflo (2007), Rojas

---

<sup>34</sup> See also Jantti and Jenkins (2015) for a recent overview of other studies that investigate the impacts of measurement errors on intra-generational mobility, and Nybom and Stuhler (2016) for an analysis in the context of inter-generational mobility.

(2008), Graham, 2010)).<sup>35</sup> Still, one potentially promising direction for further research is to employ subjective well-being data to provide better imputed household consumption.

## VII. Conclusion

We offer in this paper a review of poverty measurement methods in contexts where consumption data are missing or have inadequate quality. Some of the methods we reviewed are more established, but some are rather recent. While micro survey data are becoming available and more frequently collected in developing countries, we expect these methods to be useful at least in the immediate term, and particularly when a need arises for backcasting consumption from a more recent survey for better comparison with older surveys.

In addition, these imputation methods may also be appropriate in contexts where survey costs and/or survey implementation pose a challenge. For example, perhaps most national statistical agencies are keen on producing annual poverty statistics. But few, if any, developing countries can afford the associated expenses and demanding logistics of fielding a household consumption survey every year; rather, they are likely to implement the household consumption survey every few years. In such contexts, poverty rates can be imputed for the intervening years between the surveys at just a fraction of the cost of fielding a full-fledged consumption survey by, say using other non-consumption data or implementing a lighter (non-consumption) version of the survey in the spirit of the SWIFT approach.<sup>36</sup>

Seen in this light, imputation techniques can offer a low-cost and arguably wieldy approach to poverty estimation. While we should be mindful of the various assumptions underlying imputation

---

<sup>35</sup> There appears no consensus yet even on the use of subjective well-being data. While some researchers (see, for example, Bond and Lang (2014) and Gibson (2016)) remain skeptical, others are more optimistic (see, for example, Allin and Hand (2017) and Helliwell, Layard, and Sachs (2017)).

<sup>36</sup> Kilic *et al.* (2017) estimate the average cost of implementing a recent household consumption survey (in 2014 or later) to range from approximately US\$800,000 to US\$5 million, depending on the context and sample sizes. On the other hand, applying poverty imputation methods would most likely require only researchers' time costs.

methods as discussed in this paper, we would cautiously call for more attention to further developing these methods, and particularly validation studies to provide richer evidence on contexts where these methods may or may not work, or how well these methods work. Besides providing the typical (regional or national) capacity building exercises for staff at different national statistical agencies, a possibly useful way to proceed is perhaps some subsequent selective pairing of international experts with these staff, who can form small teams that provide further analysis with low costs. Another approach is to incorporate these imputation methods in existing popular software platforms with an accompanying guidebook (see, for example, Foster *et al.* (2013)) that can provide user-friendly and self-contained access to development practitioners.

## References

- Alkire, Sabina, James Foster, Suman Seth, Jose Manuel Roche, and Maria Emma Santos. (2015). *Multidimensional Poverty Measurement and Analysis*. USA: Oxford University Press.
- Allin, Paul, and David J. Hand. (2017). "New Statistics for Old?—Measuring the Wellbeing of the UK." *Journal of the Royal Statistical Society: Series A*, 180(1): 3-43.
- Banerjee, Abhijit and Esther Duflo. (2007). "The Economic Lives of the Poor." *Journal of Economic Perspectives*, 21(1): 141–167.
- Beegle, Kathleen, Luc Christiaensen, Andrew Dabalen, and Isis Gaddis. (2016). *Poverty in a Rising Africa*. Washington, DC: The World Bank.
- Bell, William, Wesley Basel, Craig Cruse, Lucinda Dalzell, Jerry Maples, Brett O'Hara, and David Powers. (2007). "Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties". US Census Bureau Report. Retrieved from the Internet on March 2, 2017 at <https://www.census.gov/did/www/saipe/publications/methods.html>.
- Bierbaum, Mira and Franziska Gassmann. (2012). "Chronic and transitory poverty in the Kyrgyz Republic: What can synthetic panels tell us?" *UNU-MERIT Working Paper #2012-064*.
- Blinder, A. S. (1973) "Wage Discrimination: Reduced Form and Structural Estimates". *Journal of Human Resources*, 8, 436–455.
- Blumenstock, Joshua E., Gabriel Cadamuro, and Robert On. (2015). "Predicting Poverty and Wealth from Mobile Phone Metadata". *Science*, 350(6264): 1073-1076.
- Bond, Timothy N., and Kevin Lang. (2014). "The Sad Truth about Happiness Scales". National Bureau of Economic Research, *Working Paper No. 19950*.
- Bourguignon, Francois, Chor-Ching Goh, and Dae Il Kim. (2004). "Estimating Individual Vulnerability to Poverty with Pseudo-Panel Data", *World Bank Policy Research Working Paper No. 3375*. Washington DC: The World Bank.
- Brown, Caitlin, Martin Ravallion, and Dominique van de Walle. (2016). "A Poor Means Test? Econometric Targeting in Africa". *World Bank Policy Research Working Paper No. 7915*. Washington DC: The World Bank.
- Cancho, Author César, María E. Dávalos, Giorgia Demarchi, Moritz Meyer, and Carolina Sánchez Páramo. (2015). "Economic Mobility in Europe and Central Asia: Exploring Patterns and Uncovering Puzzles". *World Bank Policy Research Paper No. 7173*.
- Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and its Application*. Chichester: John Wiley & Sons.

- Census Bureau. (2017a). Survey of Income and Program Participation, Data Editing and Imputation. Accessed on the Internet on March 2, 2017 at <http://www.census.gov/programs-surveys/sipp/methodology/data-editing-and-imputation.html>
- . (2017b). Current Population Survey, Imputation of Unreported Data Items. Accessed on the Internet on March 2, 2017 at <https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/imputation-of-unreported-data-items.html>
- Chesher, Andrew and Christian Schluter. (2002). “Welfare Measurement and Measurement Error”. *Review of Economic Studies*, 69(2): 357-378.
- Christiaensen, Luc, Peter Lanjouw, Jill Luoto, and David Stifel. (2012). "Small Area Estimation-based Prediction Models to Track Poverty: Validation and Applications." *Journal of Economic Inequality* 10, no. 2:267-297.
- Coady, David, Margaret Grosh, and John Hoddinott. (2014). “Targeting Outcomes Redux”. *World Bank Research Observer*, 19:61–85.
- Crossley, Thomas F. and Joachim K. Winter. (2015). “Asking Households About Expenditures: What Have We Learned?” in Carroll, C., T. F. Crossley and J. Sabelhaus. (Eds.). *Improving the Measurement of Consumer Expenditures. Studies in Income and Wealth*, Volume 74. Chicago: University of Chicago Press.
- Cruces, Guillermo, Peter Lanjouw, Leonardo Lucchetti, Elizaveta Perova, Renos Vakis, and Mariana Viollaz. (2015). “Estimating Poverty Transitions Repeated Cross-Sections: A Three-country Validation Exercise”. *Journal of Economic Inequality*, 13:161–179.
- Dang, Hai-Anh and Andrew L. Dabalen. (forthcoming). “Is Poverty in Africa Mostly Chronic or Transient? Evidence from Synthetic Panel Data”. *Journal of Development Studies*.
- Dang, Hai-Anh and Elena Ianchovichina. (2016). “Welfare Dynamics with Synthetic Panels: The Case of the Arab World in Transition”. *World Bank Policy Research Paper no. 7595*, World Bank, Washington, DC.
- Dang, Hai-Anh and Peter Lanjouw. (2013). “Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections”. *World Bank Policy Research Working Paper No. 6504*, World Bank, Washington, DC.
- . (2016). “Toward a New Definition of Shared Prosperity: A Dynamic Perspective from Three Countries”. In Kaushik Basu and Joseph Stiglitz. (Eds.). *Inequality and Growth: Patterns and Policy*. Palgrave MacMillan Press.
- . (2017). “Welfare Dynamics Measurement: Two Definitions of a Vulnerability Line and Their Empirical Application”. *Review of Income and Wealth*, 63(4): 633-660.

- . (forthcoming). "Poverty and Vulnerability Dynamics for India during 2004-2012: Insights from Longitudinal Analysis Using Synthetic Panel Data". *Economic Development and Cultural Change*.
- Dang, Hai-Anh and Minh Cong Nguyen. (2014) "POVIMP: Stata Module to Provide Poverty Estimates in the Absence of Actual Consumption Data." *Statistical Software Components S457934*. Boston College, Department of Economics.
- Dang, Hai-Anh, Peter Lanjouw, Umar Serajuddin. (2017). "Updating Poverty Estimates at Frequent Intervals in the Absence of Consumption Data: Methods and Illustration with Reference to a Middle-Income Country." *Oxford Economic Papers*, 69(4): 939-962.
- Dang, Hai-Anh, Peter Lanjouw, and Rob Swinkels. (2017). "Who Remained in Poverty, Who Moved Up, and Who Fell Down? An Investigation of Poverty Dynamics in Senegal in the 2000's." In Machiko Nissanke and Muna Ndulo. (Eds). *Poverty Reduction in the Course of African Development, Festschrift for Erick Thorbecke*. Oxford University Press.
- Dang, Hai-Anh, Peter Lanjouw, Jill Luoto, and David McKenzie. (2014). "Using Repeated Cross-Sections to Explore Movements in and out of Poverty". *Journal of Development Economics*, 107: 112-128.
- Davey, Adam, Michael J. Shanahan, and Joseph L. Schafer. (2001). "Correcting for Selective Nonresponse in the National Longitudinal Survey of Youth Using Multiple Imputation." *Journal of Human Resources*, 36: 500-519.
- Deaton, Angus. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. MD: The Johns Hopkins University Press.
- Deaton, Angus and Valerie Kozel. (2005). *The Great Indian Poverty Debate*. New Delhi: Macmillan.
- Deaton, Angus and John Muellbauer. (1980). *Economics and Consumer Behavior*. Cambridge University Press.
- Deaton, Angus and Salman Zaidi. (2002). *Guidelines for Constructing Consumption Aggregates for Welfare Analysis* (Vol. 135). Washington, DC: World Bank Publications.
- Demombynes, Gabriel. (2015). "Mapping Vietnam's Poverty Indicators". World Bank's East Asia & Pacific on the Rise Blog. <http://blogs.worldbank.org/eastasiapacific/mapping-vietnam-poverty-indicators>
- Dercon, Stefan and Pramila Krishnan. (2000). "Vulnerability, Seasonality and Poverty in Ethiopia." *Journal of Development Studies* 36 (6): 25-53.
- Dercon, Stefan and Joseph S. Shapiro. (2007). "Moving On, Staying Behind, Getting Lost: Lessons on Poverty Mobility from Longitudinal Data." In *Moving Out of Poverty: Cross-*



- Disciplinary Perspectives*. D. Narayan and P. Petesch. (Eds). Washington, DC: World Bank.
- Deutsch, Joseph, Jacques Silber, and Guanghua Wan. (2017). “Curbing One’s Consumption and the Impoverishment Process: The Case of Western Asia”. *Research on Economic Inequality*, 25: 1-24.
- Devarajan, Shantayanan. (2013). “Africa's Statistical Tragedy”. *Review of Income and Wealth*, 59(1): S9-S15.
- Doudich, Mohamed, Abdeljaouad Ezzrari, Roy van der Weide, and Paolo Verme. (2016). “Estimating Quarterly Poverty Rates Using Labor Force Surveys: A Primer.” *World Bank Economic Review*, 30(3): 475-500.
- Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw. (2003). “Micro-Level Estimation of Poverty and Inequality.” *Econometrica*, 71(1): 355-364.
- Ferreira, Francisco H. G., Julian Messina, Jamele Rigolini, Luis-Felipe López-Calva, Luis Felipe López-Calva, and Renos Vakis. (2012). *Economic Mobility and the Rise of the Latin American Middle Class*. Washington DC: World Bank.
- Ferreira, Francisco H. G., Shaohua Chen, Andrew Dabalen, Yuri Dikhanov, Nada Hamadeh, Dean Jolliffe, Ambar Narayan, Espen Beer Prydz, Ana Revenga, Prem Sangraula, Umar Serajuddin, and Nobuo Yoshida. (2016). “A Global Count of the Extreme Poor in 2012: Data Issues, Methodology and Initial Results.” *Journal of Economic Inequality*, 14(2): 141–172.
- Filmer, Deon and Lant Pritchett. (2001). “Estimating Wealth Effects without Expenditure Data— or Tears: An Application to Educational Enrollments in States of India”. *Demography*, 38(1): 115–132.
- Filmer, Deon and Kinnon Scott. (2012). “Assessing Asset Indices.” *Demography*, 49 (1): 359–92.
- Foster, James E. and Jonathan Rothbaum. (2015). “Using Synthetic Panels to Estimate Intergenerational Mobility”. *Working paper No. 013/2015*. Espinosa Yglesias Research Centre.
- Foster, James, Suman Seth, Michael Lokshin, and Zurab Sajaia. (2013). *A Unified Approach to Measuring Poverty and Inequality--Theory and Practice: Streamlined Analysis with ADePT Software*. Washington, DC: World Bank.
- Fujii, Tomoki. (2010). “Micro-Level Estimation of Child Undernutrition Indicators in Cambodia”. *World Bank Economic Review*, 24(3): 520-553.
- Garbero, Alessandra. (2014). “Estimating Poverty Dynamics Using Synthetic Panels for IFAD-supported Projects: A Case Study from Vietnam”. *Journal of Development Effectiveness*, 6(4): 490-510.

- Gibson, John. (2016). "Poverty Measurement: We Know Less than Policy Makers Realize". *Asia & the Pacific Policy Studies*, 3: 430–442.
- Glewwe, Paul. (2012). "How Much of Observed Economic Mobility Is Measurement Error? IV Methods to Reduce Measurement Error Bias, with an Application to Vietnam." *World Bank Economic Review* 26 (2): 236–64.
- Gottschalk, Peter and Minh Huynh. (2010). "Are Earnings Inequality and Mobility Overstated? The Impact of Nonclassical Measurement Error". *Review of Economic and Statistics*, 92: 302–315.
- Graham, Carol. (2010). *Happiness around the World: The Paradox of Happy Peasants and Miserable Millionaires*. New York: Oxford University Press.
- Grosh, Margaret and Paul Glewwe. (2000). *Designing Household Survey Questionnaires for Developing Countries*. Washington, DC: World Bank.
- Grosh, M., C. Del Ninno, E. Tesliuc, and A. Ouerghi. (2008). *For Protection and Promotion: The Design and Implementation of Effective Safety Nets*. Washington, DC: World Bank.
- Gustafsson, Bjorn, Li Shi, and Hiroshi Sato. (2014). "Data for studying earnings, the distribution of household income and poverty in China". *China Economic Review*, 30, 419–431.
- Harttgen, Kenneth, Stephan Klasen, and Sebastian Vollmer. (2013). "An African Growth Miracle? Or: What Do Asset Indices Tell Us about Trends in Economic Performance?" *Review of Income and Wealth*, 59(S1): S37–S61.
- Helliwell, John, Richard Layard, and Jeffrey Sachs. (2017). *World Happiness Report 2017*. New York: Sustainable Development Solutions Network.
- Howe, Laura, James R. Jargreaves, Sabine Gabrysch, and Sharon Huttly. (2009). "Is the Wealth Index a Proxy for Consumption Expenditure? A Systematic Review." *Journal of Epidemiology and Community Health*, 63(11): 871–80.
- Independent Evaluation Group. *The Poverty Focus of Country Programs- Lessons from World Bank Experience*. Washington DC: World Bank.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. (2013). *An Introduction to Statistical Learning*. New York: Springer.
- Jäntti, Markus, and Stephen P. Jenkins. (2015). "Income Mobility." In *Handbook of Income Distribution*, ed. Anthony B. Atkinson, and Francois Bourguignon. Vol. 2. Elsevier: 807-935.
- Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. (2016). "Combining Satellite Imagery and Machine Learning to Predict Poverty". *Science*, 19(353): 790-794.

- Jenkins, Stephen P., Richard V. Burkhauser, Shuaizhang Feng, and and Jeff Larrimore. (2011). "Measuring Inequality Using Censored Data: A Multiple-imputation Approach to Estimation and Inference." *Journal of the Royal Statistical Society: Series A*, 174(1): 63–81.
- Jolliffe, Ian T. (2002). *Principal Component Analysis*. 2nd ed. New York: Springer.
- Jolliffe, Dean, Peter Lanjouw, Shaohua Chen, Aart Kraay, Christian Meyer, Mario Negre, Espen Prydz, Renos Vakis, and Kyla Wethli. (2015). *A Measured Approach to Ending Poverty and Boosting Shared Prosperity: Concepts, Data, and the Twin Goals*. Washington DC: The World Bank.
- Journal of Development Studies. (2015). Special issue of "*Statistical Tragedy in Africa? Evaluating the Data Base for African Economic Development*". Volume 51, Issue 2.
- Kijima, Yoko and Peter Lanjouw. 2003. "Poverty in India during the 1990s: A Regional Perspective." World Bank Policy Research Working Paper no. 3141, World Bank, Washington, DC.
- Kilic, Talip, Umar Serajuddin, Hiroki Uematsu, and Nobuo Yoshida. (2017). "Costing Household Surveys for Monitoring Progress toward Ending Extreme Poverty and Boosting Shared Prosperity." *World Bank Policy Research Paper no. 7951*, World Bank, Washington, DC.
- Kozyreva, Polina, Mikhail Kosolapov, and Barry M. Popkin. (2016). "Data Resource Profile: The Russia Longitudinal Monitoring Survey—Higher School of Economics (RLMS-HSE) Phase II: Monitoring the Economic and Health Situation in Russia, 1994–2013". *International Journal of Epidemiology*, 395–401.
- Lee, Nayoung, Geert Ridder, and John Strauss. (2017). "Estimation of Poverty Transition Matrices with Noisy Data". *Journal of Applied Econometrics*, 32: 37–55.
- Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*. 2nd Edition. New Jersey: Wiley.
- Mark R. Montgomery, Michele Gragnolati, Kathleen A. Burke, and Edmundo Paredes. (2000). "Measuring Living Standards with Proxy Variables". *Demography*, 37(2): 155-174.
- Martinez, Arturo Jr., Mark Western, Michele Haynes, Wojtek Tomaszewski. (2013). "Measuring Income Mobility Using Pseudo-Panel Data". *Philippine Statistician*, 62(2): 71-99.
- Mathiassen, Astrid. (2013). "Testing Prediction Performance of Poverty Models: Empirical Evidence from Uganda". *Review of Income and Wealth* 59, no. 1:91–112.
- Morris, Saul S., Calogero Carletto, John Hoddinott, and Luc JM Christiaensen. (2000). "Validity of Rapid Estimates of Household Wealth and Income for Health Surveys in Rural Africa." *Journal of Epidemiology & Community Health*, 54(5): 381-387.

- Nayar, Reema, Pablo Gottret, Pradeep Mitra, Gordon Betcherman, Yue Man Lee, Indhira Santos, Mahesh Dahal, and Maheshwor Shrestha. (2012). *More and Better Jobs in South Asia*. Washington, DC: World Bank.
- Nguyen, Minh Cong, Paul Corral, Quinghua Zhao, and Joao Pedro de Azevedo. (2017). *The PovMap Stata Command*. Working paper.
- Nybom, Martin and Jan Stuhler. (2016). “Biases in Standard Measures of Intergenerational Income Dependence”. *Journal of Human Resources*, doi: 10.3368/jhr.52.3.0715-7290R.
- Oaxaca, R. (1973) “Male-female Wage Differentials in Urban Labor Markets”. *International Economic Review*, 14, 693–709.
- Rama, Martin, Tara Bêteille, Yue Li, Pradeep K. Mitra, and John Lincoln Newman. (2015). *Addressing Inequality in South Asia*. Washington, DC: World Bank.
- Rao, J. N. K. and Isabel Molina. (2015). *Small Area Estimation*, 2nd edition, New York: Wiley.
- Ravallion, Martin. (2014). “Poor or Just Feeling Poor? On Using Subjective Data in Measuring Poverty”. In Andrew E. Clark, Claudia Senik. (Eds). *Happiness and Economic Growth: Lessons from Developing Countries*. UK: Oxford University Press.
- . (2016). *The Economics of Poverty: History, Measurement, and Policy*. New York: Oxford University Press.
- Ravallion, Martin, Shaohua Chen, and Prem Sangraula. (2009). “Dollar a Day Revisited.” *World Bank Economic Review*, 23(2): 163–184.
- Rencher, Alvin C. (2002). *Methods of Multivariate Analysis*. USA: John Wiley & Sons.
- Rojas, Mariano. (2008). “Experienced Poverty and Income Poverty in Mexico: A Subjective Well-Being Approach”. *World Development*, 36(6): 1078-1093.
- Rutstein, Shea, and Sarah Staveteig. (2014). “Making the Demographic and Health Surveys Wealth Index Comparable”. *DHS Methodological Report 9*, ICF International, Rockville, MD.
- Sahn, David E. and David C. Stifel. (2000). “Poverty Comparison over Time and across Countries in Africa”. *World Development*, 28(12): 2123-2155.
- Serajuddin, Umar, Hiroki Uematsu, Christina Wieser, Nobuo Yoshida, and Andrew Dabalen. (2015). “Data deprivation: another deprivation to end.” *World Bank Policy Research Paper no. 7252*, World Bank, Washington, DC.
- Steele, Jessica E., Pål Roe Sundsøy, Carla Pezzulo, Victor A. Alegana, Tomas J. Bird, Joshua Blumenstock, Johannes Bjelland, Kenth Engø-Monsen, Yves-Alexandre de Montjoye, Asif M. Iqbal, Khandakar N. Hadiuzzaman, Xin Lu, Erik Wetter, Andrew J. Tatem, and Linus

Bengtsson. (2017). “Mapping Poverty Using Mobile Phone and Satellite Data”. *Journal of the Royal Society Interface*. DOI: 10.1098/rsif.2016.0690.

Tarozzi, Alessandro. 2007. “Calculating Comparable Statistics from Incomparable Surveys, With an Application to Poverty in India”. *Journal of Business and Economic Statistics* 25, no. 3:314-336.

Vakis, Renos, James Rigolini, and Leonardo Lucchetti. (2015). *Left Behind: Chronic Poverty in Latin America and the Caribbean*. Washington, DC: World Bank.

World Bank (2015). *Purchasing Power Parities and the Real Size of World Economies – A Comprehensive Report of the 2011 International Comparison Program*. Washington, DC: World Bank.

**Table 1: Categories of Missing Household Consumption Data and Commonly Employed Imputation Methods**

Type	Extent of Missing Consumption Data	Typical Situation	Example	Imputation Method
A	Completely missing	i) Non-consumption surveys	Demographic and Health Surveys	Wealth index
		ii) Most small-scale surveys		
B	Partially missing	i) Consumption data not comparable across survey rounds	Some rounds of India's National Sample Surveys	Survey-to-survey imputation
		ii) Consumption data unavailable in current survey but available in another related survey	The annual LFS does not have consumption data, but the household consumption survey is implemented every few years	
		iii) Consumption data unavailable at more disaggregated administrative levels than those in current survey	Population census data are representative at lower administrative level than a household consumption survey, but does not collect consumption data.	
C	Available cross sections, but missing panel data	Most surveys in developing countries do not offer panel data		Synthetic panels

**Note:** LFS stands for Labor Force Surveys.

**Table 2: Population Distribution by Asset Indexes vs. Consumption, Vietnam 2012-2014 (percentage)**

Per capita consumption	2012			2014		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Poorest quintile	64.3	52.9	57.1	43.8	43.2	46.6
Quintile 2	24.7	29.7	30.5	42.4	26.3	29.2
Quintile 3	24.6	27.4	28.7	23.5	32.0	28.6
Quintile 4	36.5	30.7	31.0	19.5	32.9	36.7
Richest quintile	39.6	55.4	57.2	47.2	60.8	62.9
Correlation with household consumption	0.61	0.67	0.70	0.59	0.64	0.67
N	9,396	9,324	9,324	9,399	9,348	9,348

**Note:** Each cell in the first five rows shows the percentage of the population that would be correctly captured for each consumption quintile if asset index was used. Model 1 provides a simple count of the number of assets a household possesses, while Models 2 and 3 construct the asset index using principal component method. The list of assets for Model 1 include car, motorbike, bicycle, desk phone, mobility phone, DVD player, television set, computer, refrigerator, air conditioner, washing machine, and electric fan. Model 2 adds to Model 1 the construction materials for the house's roof and wall, Model 3 adds to Model 2 the type of water and toilet the household has access to. All estimates are weighted by population weight.

**Table 3: Growth in Asset Indexes vs. Growth in Consumption, Vietnam 2010- 2014 (percentage)**

	Year/ Growth	Consumption (D'000)	Wealth index 1	Wealth index 2
<b>Panel A</b>	2010	18,683	5.24	0.16
	2012	19,026	5.76	0.14
	Growth rate	1.8	9.9	-110.5
<b>Panel B</b>	2010	19,026	5.76	0.14
	2012	20,941	6.08	0.18
	Growth rate	10.1	5.6	28.0

**Note:** Wealth index 1 provides a simple count of the number of assets a household possesses, while Wealth index 2 constructs the asset index using principal component method after pooling data for all three years. The list of assets for both wealth indexes include car, motorbike, bicycle, desk phone, mobility phone, DVD player, television set, computer, refrigerator, air conditioner, washing machine, and electric fan. All estimates are weighted by population weight.



**Table 4: Predicted Poverty Rates Based on Imputation, Vietnam 2012-2014 (percentage)**

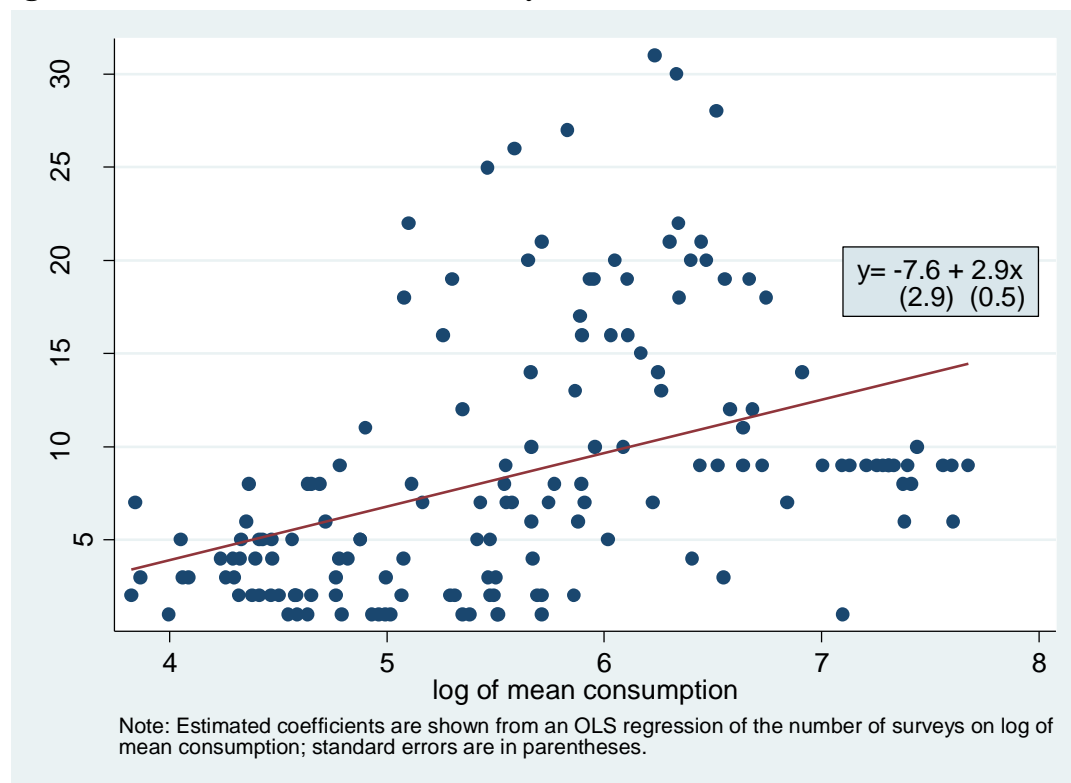
Method	2012			2014		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
1) Normal linear regression model	21.3 (0.5)	21.1 (0.5)	17.5 (0.5)	16.7 (0.5)	16.6 (0.5)	13.1 (0.4)
2) Empirical distribution of the error terms	21.1 (0.5)	20.9 (0.5)	17.5 (0.5)	16.4 (0.5)	16.4 (0.5)	13.0 (0.4)
<i>Control variables</i>						
Parsimonious	Y	Y	Y	Y	Y	Y
Demographics & employment	N	Y	Y	N	Y	Y
Household assets & house characteristics	N	N	Y	N	N	Y
<b>True poverty rate</b>	17.2 (0.5)			13.5 (0.5)		

**Note:** Standard errors in parentheses are adjusted for complex survey design. All estimates are obtained with population weights. Method 1 uses the normal linear regression model with the theoretical distribution of the error terms and Method 2 uses the empirical distribution of the error terms. Both methods employ commune random effects. Imputed poverty rates for 2012 use the estimated parameters based on the 2010 data, and imputed poverty rates for 2014 use the estimated parameters based on the 2012 data. 1,000 simulations are implemented. The underlying regression results are provided in Appendix 3, Table 3.2. True poverty rate is the estimate directly obtained from the survey data.

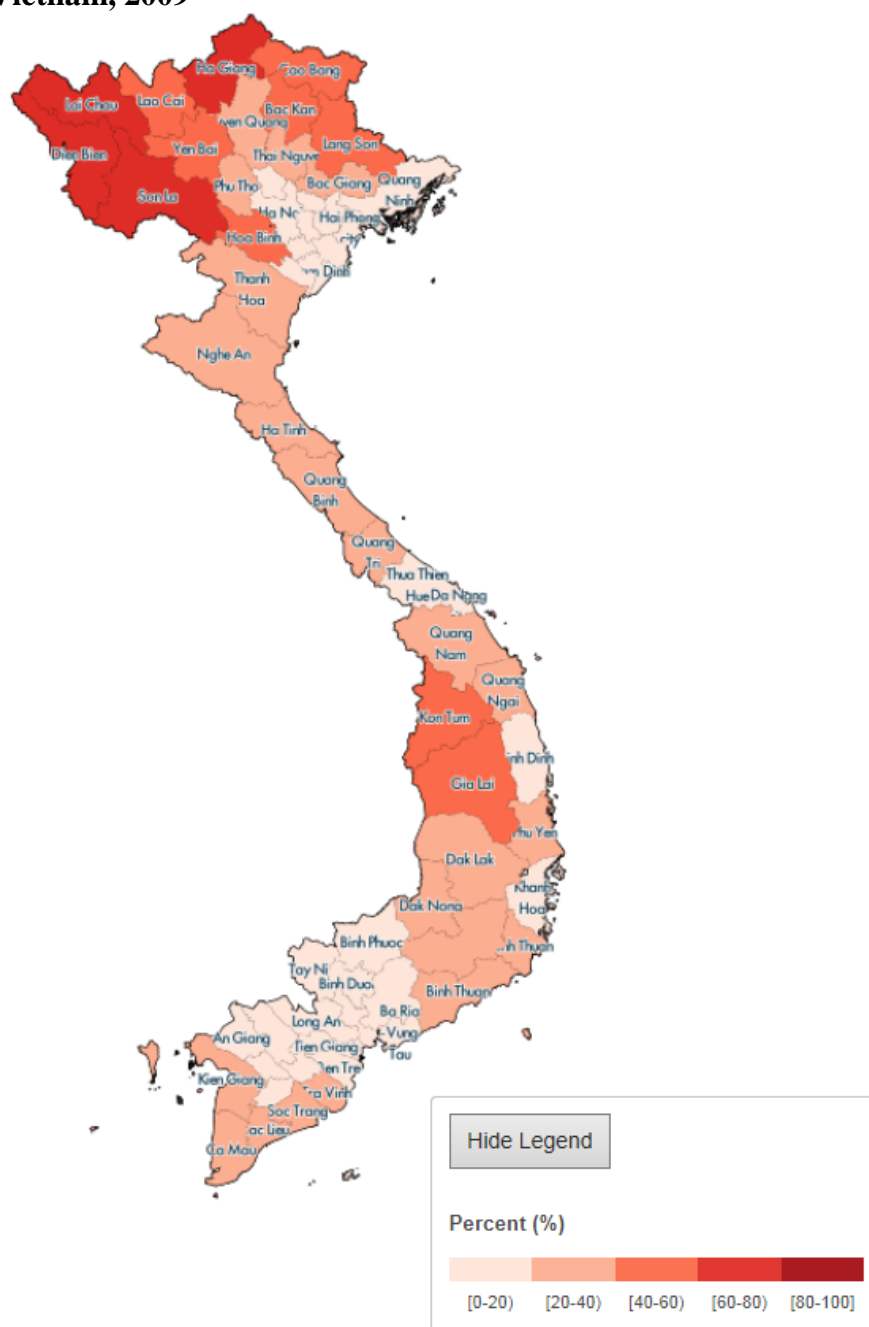
Table 5: Poverty Dynamics Based on Synthetic Data, Vietnam 2012-2014 (percentage)

Panel A: Unconditional Probabilities			Panel B: Conditional Probabilities		
Poverty Status	Actual panel	Synthetic panel	Poverty Status	Actual panel	Synthetic panel
Poor, Poor	10.7 (0.8)	10.8 (0.3)	Poor--> Poor	61.1 (2.2)	64.9 (1.3)
Poor, Nonpoor	6.8 (0.5)	5.9 (0.1)	Poor--> Nonpoor	38.9 (2.2)	35.1 (0.6)
Nonpoor, Poor	4.6 (0.5)	4.0 (0.1)	Nonpoor--> Poor	5.6 (0.6)	4.8 (0.1)
Nonpoor, Nonpoor	77.9 (1.0)	79.3 (0.4)	Nonpoor--> Nonpoor	94.4 (0.6)	95.2 (0.3)
<i>Goodness-of-fit Tests</i>					
Within 95% CI		4/4	Within 95% CI		4/4
Within 1 standard error		1/4	Within 1 standard error		0/4
Coverage of 50% or more		4/4	Coverage of 50% or more		4/4
Coverage of 100%		2/4	Coverage of 100%		1/4
N	2,639	3,519	N	2,639	3,519

**Note:** Synthetic panels are constructed from the cross sectional component. The first survey round is used as the base year for imputation. Standard errors are obtained adjusting for complex survey design. All estimates are obtained with population weights. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the "Within 1 standard error" row shows a similar figure but using one standard error around the estimates based on the actual panels. The "Coverage of 50% or more" row shows the number of times that half or more of the 95% CI around the synthetic panel estimates overlap with those based on the actual panels overlap with those based on the actual panels; the "Coverage of 100%" row shows a similar figure for the number of times that the former fall completely inside the latter.

**Figure 1: Number of Household Surveys vs. Countries' Income Level, 1981- 2014**

**Data source:** PovCalNet, 2017.

**Figure 2: Poverty Map of Vietnam, 2009**

**Note:** Produced by Poverty Team, World Bank, Vietnam (<http://www5.worldbank.org/mapvietnam/>)

**Disclaimer:** Country borders or names do not necessarily reflect the World Bank Group's official position. This map is for illustrative purposes and does not imply the expression of any opinion on the part of the World Bank, concerning the legal status of any country or territory or concerning the delimitation of frontiers or boundaries.

## Appendix 1: Proofs

### Proposition 1

The linear projection of household consumption onto household (and community) characteristics  $x_{ij}$  is given as follows (Equation (1))

$$y_{ij} = \beta' x_{ij} + \mu_{ij} \quad (1.1)$$

We provide three following sets of results in this proof

- 1a. the wealth index  $w_{ij}$  provides a biased estimator of household consumption  $y_{ij}$*
- 1b. poverty estimates based on  $w_{ij}$  tend to provide biased estimates of the poverty rates based on  $y_{ij}$ , and*
- 1c. wealth indexes may not necessarily provide biased estimates of trends in household consumption (or poverty) over time (as discussed in footnote 20).*

*Part 1a. The wealth index  $w_{ij}$  provides a biased estimator of household consumption  $y_{ij}$*

Since  $x_{ij}$  is composed of different characteristics such as household tastes, labor supply, and assets (Deaton and Muellbauer, 1980; Deaton and Zaidi, 2002), we can also write out  $x_{ij}$  more clearly as consisting of two components, assets ( $a_{ij}$ ) and non-assets variables ( $z_{ij}$ )

$$y_{ij} = \beta'_1 a_{ij} + \beta'_2 z_{ij} + \mu_{ij} \quad (1.2)$$

Thus by the Gauss-Markov theorem (see, e.g., Greene (pp. 60, 2012)), the least square estimator  $\hat{\beta}$ , which is  $(\hat{\beta}'_1, \hat{\beta}'_2)'$ , provides the (minimum variance) linear unbiased estimator of  $\beta$ .

On the other hand, reversing the two sides of Equation (2) for presentation purposes, we obtain wealth indexes using the following equation

$$w_{ij} = \gamma' a_{ij} \quad (1.3)$$

where  $\hat{\gamma}$  provides the estimator for  $\gamma$  using PCA methods. Taking the expectations of (1.2) and (1.3) and comparing results, we have

$$E(y_{ij}) = E(\beta'_1 a_{ij} + \beta'_2 z_{ij} + \mu_{ij}) = E(\beta'_1 a_{ij} + \beta'_2 z_{ij}) \neq E(w_{ij}) = E(\gamma' a_{ij}) \quad (1.4)$$

Thus the wealth index  $w_{ij}$  provides a biased estimator of household consumption  $y_{ij}$ . Intuitively, this result is due to two reasons. First, the wealth index  $w_{ij}$  does not include the non-asset component  $z_{ij}$ , which is equivalent to the well-known issue of omitted variable bias. Second,  $\beta_1$  and  $\gamma$  are generally different from each other, since the estimator for  $\gamma$  maximizes the variance in  $x_{ij}$ , while the estimator for  $\beta$  maximizes the variance in  $y_{ij}$ .<sup>37</sup>

*Part 1b. Poverty estimates based on  $w_{ij}$  tend to provide biased estimates of the poverty rates based on  $y_{ij}$*

Since the wealth index  $w_{ij}$  provides a biased estimator for household consumption  $y_{ij}$ , poverty estimates based on  $w_{ij}$  tend to provide biased estimates of the poverty rates based on  $y_{ij}$ . Put differently, let  $P(\cdot)$  denote the poverty function, and  $z_w$  and  $z_y$  respectively the poverty lines for the wealth index and household consumption, the following equation would generally hold true

$$P(w_{ij} \leq z_w) \neq P(y_{ij} \leq z_y) \quad (1.5)$$

<sup>37</sup> See also Rencher (2002) for a graphical illustration of the general difference between principal component analysis and OLS methods.

Furthermore, while Equation (1.5) may be violated under some rare theoretical circumstance (e.g., by choosing a cutoff point below which the proportions of the population are equal for both  $w_{ij}$  and  $y_{ij}$ ), the identification of  $z_w$  is no easy task. While there is an established theory underlying the construction of  $z_y$  (say, using a minimum basic needs or calorie-intake approach (see, e.g., Ravallion, 2016)), no such reliable theory currently exists for the construction of  $z_w$ . In practice, identifying  $z_w$  usually involves assuming a given level of poverty rate and set  $z_w$  at the corresponding percentile in the distribution (see, for example, Sahn and Stifel (2000)).

*Part 1c. Wealth indexes may not necessarily provide biased estimates of trends in household consumption (or poverty) over time*

It is in fact rather straightforward to see that wealth indexes may not necessarily provide biased estimates of trends in household consumption (or poverty) over time (as discussed in footnote 20). Let us specify the bias between the wealth index  $w_{ij,t}$  and household consumption  $y_{ij,t}$  at time  $t$  as follows

$$E(y_{ij,t}) - E(w_{ij,t}) = B_{ij,t} \quad (1.6)$$

Given the assumption that the degree of bias is similar for each cross-sectional survey round, or equivalently, we have the following hold regardless of time  $t$

$$B_{ij,t} = B_{ij} \quad (1.7)$$

Thus, taking the difference between time  $t$  and time  $t-1$  for Equation (1.4), and plugging in Equation (1.5), we can estimate the trends in household consumption as follows

$$E(y_{ij,t}) - E(y_{ij,t-1}) = [E(w_{ij,t}) - B_{ij,t}] - [E(w_{ij,t-1}) - B_{ij,t-1}] = E(w_{ij,t}) - E(w_{ij,t-1}) \quad (1.8)$$

A similar result follows for poverty where we replace the wealth index  $w_{ij,t}$  and household consumption  $y_{ij,t}$  with the poverty function  $P(.)$  in Equation (1.6).

## Proposition 2

Household consumption is predicted using proxy means tests as follows (i.e., Equation (5))

$$y_{ij}^p = \beta^{p'} x_{ij,p} \quad (1.9)$$

We provide two following sets of results in this proof

2a. Proxy means testing tends to offer biased poverty estimate.

2b. The extent of bias using proxy means testing tends to be less than those using wealth indexes

*Part 2a. Proxy means testing tends to offer biased poverty estimate.*

Taking the expectation of (1.1), and (1.9) we have

$$E(y_{ij}) = E(\beta' x_{ij} + \mu_{ij}) = \beta' x_{ij} \quad (1.10)$$

$$E(y_{ij}^p) = \beta^{p'} x_{ij,p} \quad (1.11)$$

When the model is fully specified, the estimator for  $\beta^p$  is identical for that for  $\beta$  (i.e. using the same nationally representative household survey). When the model is not fully specified,  $x_{ij,p}$  is a smaller subset of  $x_{ij}$ . This would result in  $\beta^p$  being a biased estimate for  $\beta$ , the mean of household consumption based on proxy means testing would provide a biased estimate for mean household consumption.

For the variance, taking the variance of (1.1), and (1.9) we have

$$V(y_{ij}) = V(\beta' x_{ij} + \mu_{ij}) = V(\beta' x_{ij}) + V(\mu_{ij}) \quad (1.12)$$

$$V(y_{ij}^p) = V(\beta^p x_{ij,p}) \quad (1.13)$$

When the model is not fully specified, we have

$$V(\beta_{x_{ij,p}}^p) \leq V(\beta' x_{ij}) \leq V(\beta' x_{ij}) + V(\mu_{ij}) = V(y_{ij}) \quad (1.14)$$

since  $x_{ij,p}$  is a smaller subset of  $x_{ij}$ .

Following a similar argument as with Equation (1.5), the results above imply that proxy means testing tends to offer biased estimates of poverty.

*Part 2b. The extent of bias using proxy means testing tends to be less than those using wealth indexes*

The bias between household consumption and predicted consumption obtained by proxy means testing methods is

$$E(y_{ij}) - E(y_{ij}^p) = E(\beta' x_{ij}) - E(\beta^p x_{ij,p}) \quad (1.15)$$

On the other hand, the bias between household consumption and the wealth index is

$$E(y_{ij}) - E(w_{ij}) = E(\beta' x_{ij}) - E(\gamma' a_{ij}) \quad (1.16)$$

If the estimation model is fully specified for the proxy means test, the bias in Equation (1.15) would be zero. However, if the estimation model is not fully specified, assume that  $x_{ij,p}$  includes all the household assets  $a_{ij}$  and some other relevant household characteristics (e.g., education). By the Gauss-Markov theorem, the least square estimator for  $\beta^p$  provides the linear unbiased estimator, while  $\gamma$  is not. Consequently, the bias in Equation (1.15) is smaller than that in Equation (1.16). Replacing the expectation function in Equations (1.15) and (1.16), and using a similar argument as with Equation (1.5), the results above imply that proxy means testing tends to offer less biased estimates of poverty.

### Proposition 3

The poverty rate  $P_2$  in period 2 and its variance can then be estimated as

$$\hat{P}_2 = \frac{1}{S} \sum_{s=1}^S P(\hat{y}_{2,s}^1 \leq z_1) \quad (1.17)$$

where the imputed consumption  $y_2^1$  is estimated as

$$y_2^1 = \beta_1' x_2 + v_1 + \varepsilon_1 \quad (1.18)$$

We provide two following sets of results in this proof

- 3a. *Given assumptions IV.1 and IV.2, the imputed household consumption in equation (1.17) offers unbiased poverty estimates*
- 3b. *The imputed consumption provides less bias and a better variance than the predicted consumption obtained by proxy means testing methods.*

*Part 3a. Given assumptions IV.1 and IV.2, the imputed household consumption in equation (1.17) offers unbiased poverty estimates*

The proof for this part is given in Dang *et al.* (2017).

*Part 3b. The imputed consumption provides less bias and a better variance than the predicted consumption obtained by proxy means testing methods.*

If the estimation model is fully specified for the proxy means test, there would be no bias between household consumption and imputed household consumption. Indeed, writing out the notation for each period, the bias between household consumption and imputed consumption is

$$E(y_2) - E(y_2^1) = E(\beta_2' x_{i2} + \mu_{i2}) - E(\beta_1' x_{i2} + \mu_{i1}) = E(\beta_2' x_{i2}) - E(\beta_1' x_{i2}) \quad (1.19)$$

The corresponding bias between household consumption and predicted consumption obtained by proxy means testing methods is

$$E(y_2) - E(y_2^p) = E(\beta_2' x_{i2}) - E(\beta_1^{p'} x_{i2,p}) \quad (1.20)$$

If the estimation model for the proxy means test is identical to that used for imputing consumption, the biases in Equations (1.19) and (1.20) are the same. But if the estimation model for the proxy means test is not fully specified (i.e.,  $x_{ij,p}$  is a subset of  $x_{ij}$ ), we know from Proposition 2 that  $E(\beta_1' x_{i2})$  would provide a better estimator for  $E(\beta_2' x_{i2})$  than  $E(\beta_1^{p'} x_{i2,p})$ . Consequently, the bias is smaller for imputed household consumption.

From Equation (1.14), we have

$$V(y_{ij}^p) = V(\beta_1^{p'} x_{i2,p}) \leq V(\beta_1' x_{i2} + \mu_{i1}) = V(y_2^1) \quad (1.21)$$

Regardless of whether the estimation model for the proxy means test is fully specified or not, it is smaller than, or at most equal to, the variance of the imputed household consumption. In particular, the variance  $V(y_{ij}^p)$  does not take into account the variance of the unobserved household effects, which can be considerable in practice. Thus the variance of the imputed household consumption is likely to provide a better approximation for the variance of household consumption.



## Appendix 2: Practical Note for Implementation

We provide some quick examples for the different poverty imputation methods that are reviewed in this paper.

### 2.1. Principal component analysis (PCA) method for generating wealth indexes

Let  $a_{ij}$  represent the list of household assets for household  $i$  in survey  $j$ . The list of assets includes (whether the household has) a car, a motorbike, a bicycle, a desk phone, a mobility phone, a DVD player, a television set, a computer, a refrigerator, an air conditioner, a washing machine, an electric fan, the construction materials for the house's roof and wall, and the type of water and toilet the household has access to. We use the Stata command "*pca*" to estimate the parameters in Equation (2) as follows

```
pca varlist, components (1) vce(norm)
```

The first component is specified using the option "*components(1)*", and the VCE of the eigenvalues and vectors is computed assuming multivariate normality.

After the model is estimated, the wealth index can be generated using the following command

```
predict wind3
```

More detailed discussion on the options available with the "*pca*" command is provided in the Stata (2016) manual.

### 2.2. Imputing consumption when consumption is partially missing

To implement the Dang *et al.* (2017) imputation method, we can install the "*povimp*" (Dang and Nguyen, 2014) from within Stata by typing

```
ssc install povimp
```

We can stack the different years of data such that the same variables have the same names, and use a year variable to indicate the different years. For example, let this year variable have two values 2010 and 2012, and assume that household consumption is unavailable for 2012, but are available for 2010.<sup>38</sup>

We have different modeling options for imputing household consumption in 2012 based on Equations (8) and (9). We can estimate the normal linear regression model as

```
povimp depvar varlist, by(year) from(2010) to(2012) pline(pline) cluster(com) strata(strata) wt(hhszwt) method(normal) rep(1000)
```

<sup>38</sup> Incidentally, it is rather straightforward to implement poverty estimation using proxy means testing in Stata. After stacking the data from two different years in the same data set, we can obtain the estimated parameters in the household consumption model from one year, and impose these parameters on the  $x$  variables in the other year. For example, we can type "*xtreg depvar varlist if year== 2010, i(household id) re*", and then "*predict depvarhat if year== 2012*". The imputed poverty rate can then be obtained using the predicted consumption variable and the appropriate poverty line.

where “*pline*” is a variable indicating the poverty line. Other survey design variables are specified as cluster (“*com*”), strata (“*strata*”), and population weight (“*hhszwt*”). We use 1,000 simulations.

We can estimate the model with the empirical distribution of the error terms as

```
povimp depvar varlist, by(year) from(2010) to(2012) pline(pline) cluster(com) strata(strata)
wt(hhszwt) method(empirical) rep(1000)
```

More detailed discussion on the options available with the “*povimp*” command is provided in its associated help file (i.e., type “*help povimp*” after installing it in Stata).

### 2.3. Imputing consumption when panel consumption data are missing

Do files to implement an earlier version of this method (i.e., the bounds approach in Dang *et al.* (2014)) can be downloaded on the following website

<https://sites.google.com/site/decrghdmckenzie/datasets>

These do files can be modified to provide point estimates of poverty mobility as proposed in Dang and Lanjouw (2013). Given a good approximation for  $\rho$ , we can use the following Stata command to estimate Equation (14)

```
gen double p1p2= binormal((z1- b1x2)/sigep1, -(z2- b2x2)/sigep2, -rho)
```

where the various parameters are defined as earlier (e.g., *z1* stands for  $z_1$ , *sigep1* for  $\sigma_{\varepsilon_1}$ , and so on).

Equation (16) can be estimated similar as

```
gen double p1v2= binormal((z1- b1x2)/sigep1, (v2- b2x2)/sigep2, rho)- binormal((z1-
b1x2)/sigep1, (z2- b2x2)/sigep2, rho)
```

A Stata program to automate the estimation process will be made available in due course.

### 2.4. Spatial Targeting with Poverty Map, or Small Area Estimation

A commonly used software—PovMap—that can implement poverty mapping is described in detail in Zhao and Lanjouw (2008). Efforts are under way to offer a Stata routine to automate this procedure in Stata (Nguyen *et al.*, 2017).

An R package “*sae*” that can estimate various small area estimation models is also offered by Rao and Molina (2015).

**Appendix 3: Additional Tables****Table 3.1: Cumulative Population Distribution by Asset Indexes vs. Consumption, Vietnam 2010- 2014 (percentage)**

Per capita consumption	2012			2014		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Poorest quintile	12.9	10.6	11.4	8.8	8.6	9.3
Bottom 40 percent	24.4	21.3	22.4	22.8	19.2	20.3
Bottom 60 percent	37.6	33.7	34.9	40.3	33.9	34.4
Bottom 80 percent	54.0	47.8	48.9	59.5	52.1	52.3
All distribution	73.1	67.1	68.2	82.8	75.4	75.6
Correlation with household consumption	0.61	0.67	0.70	0.59	0.64	0.67
N	9,396	9,324	9,324	9,399	9,348	9,348

**Note:** Each cell in the first five rows shows the percentage of the population that would be correctly captured by the cumulative quintiles of the consumption distribution if asset index was used. Model 1 provides a simple count of the number of assets a household possesses, while Models 2 and 3 construct the asset index using principal component method. The list of assets for Model 1 include car, motorbike, bicycle, desk phone, mobility phone, DVD player, television set, computer, refrigerator, air conditioner, washing machine, and electric fan. Model 2 adds to Model 1 the construction materials for the house's roof and wall, Model 3 adds to Model 2 the type of water and toilet the household has access to. All estimates are weighted by population weight.

**Table 3.2: Estimation of Consumption Model Using the VHLSSs, Vietnam 2010-2012**

	2010			2012		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Household size	-0.090*** (0.00)	-0.064*** (0.00)	-0.122*** (0.00)	-0.081*** (0.00)	-0.061*** (0.00)	-0.124*** (0.00)
Head's age	0.003*** (0.00)	0.001 (0.00)	-0.001** (0.00)	0.003*** (0.00)	0.001*** (0.00)	-0.001** (0.00)
Head is female	0.022* (0.01)	0.032*** (0.01)	0.033*** (0.01)	-0.007 (0.01)	0.012 (0.01)	0.039*** (0.01)
Head belongs to ethnic minority group	-0.436*** (0.02)	-0.429*** (0.02)	-0.232*** (0.01)	-0.431*** (0.02)	-0.428*** (0.02)	-0.183*** (0.01)
Head completed primary school	0.187*** (0.01)	0.168*** (0.01)	0.056*** (0.01)	0.196*** (0.01)	0.172*** (0.01)	0.053*** (0.01)
Head completed lower secondary school	0.320*** (0.01)	0.274*** (0.01)	0.087*** (0.01)	0.322*** (0.01)	0.273*** (0.01)	0.073*** (0.01)
Head completed upper secondary school	0.489*** (0.02)	0.450*** (0.02)	0.139*** (0.01)	0.481*** (0.02)	0.441*** (0.02)	0.124*** (0.01)
Head has (some) college education	0.787*** (0.02)	0.755*** (0.02)	0.263*** (0.02)	0.782*** (0.02)	0.756*** (0.02)	0.226*** (0.02)
Share of household members age 0-14		-0.438*** (0.04)	-0.445*** (0.03)		-0.371*** (0.04)	-0.360*** (0.03)
Share of household members age 15-24		0.044 (0.03)	-0.039 (0.03)		0.123*** (0.03)	0.046* (0.03)
Share of household members age 15-24		0.278*** (0.03)	0.079*** (0.02)		0.314*** (0.02)	0.087*** (0.02)
Head worked in the last 12 months		-0.055*** (0.02)	-0.020 (0.01)		-0.023 (0.02)	-0.019 (0.01)
Household owns a car			0.760*** (0.04)			0.594*** (0.03)
Household owns a motorbike			0.151*** (0.01)			0.149*** (0.01)
Household owns a bicycle			-0.026*** (0.01)			-0.031*** (0.01)
Household owns a desk phone			0.034*** (0.01)			0.066*** (0.01)
Household owns a cell phone			0.117*** (0.01)			0.123*** (0.01)
Household owns a DVD player			0.036*** (0.01)			0.050*** (0.01)
Household owns a television			0.016 (0.01)			0.093*** (0.01)
Household owns a computer			0.137*** (0.01)			0.132*** (0.01)
Household owns a refrigerator			0.151*** (0.01)			0.154*** (0.01)
Household owns an airconditioner			0.273*** (0.02)			0.225*** (0.02)
Household owns a washing machine			0.113*** (0.01)			0.117*** (0.01)
Household owns an electric fan			0.044*** (0.01)			0.035*** (0.01)
Log of residential area			0.197*** (0.01)			0.184*** (0.01)
House wall materials			0.021*** (0.00)			0.026*** (0.00)
Access to drinking water			0.013*** (0.00)			0.011*** (0.00)
Type of toilet			0.045*** (0.00)			0.047*** (0.00)
Urban	0.354*** (0.02)	0.339*** (0.02)	0.087*** (0.01)	0.339*** (0.02)	0.322*** (0.01)	0.079*** (0.01)
Constant	9.395*** (0.03)	9.452*** (0.05)	8.485*** (0.05)	9.470*** (0.03)	9.403*** (0.05)	8.396*** (0.05)
$\sigma_e$	0.40	0.39	0.30	0.40	0.38	0.30
$\sigma_u$	0.29	0.27	0.20	0.28	0.27	0.19
$\rho$	0.34	0.33	0.31	0.33	0.33	0.28
$R^2$	0.43	0.46	0.69	0.41	0.45	0.69
N	9,211	9,211	9,211	9,324	9,324	9,324

**Note:** \* p<0.10, \*\* p<0.05 \*\*\* p<0.01. Standard errors are in parentheses. All estimation employs commune random effects models. House wall material is assigned numerical values using the following categories: 6 "cement", 5 "brick", 4 "iron/wood", 3 "earth/straw", 2 "bamboo/board", and 1 "others". The types of toilet are assigned numerical values using the following categories: 6 "septic", 5 "suiabh", 4 "double septic", 3 "fish bridge", 2 "others", and 1 "none".

**Table 3.3: Decomposition of Changes in Poverty, Vietnam 2012-2014 (percentage)**

	2012			2014		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
<b>A. Normal linear regression model</b>						
Due to characteristics	59.9	62.2	96.6	14.2	17.2	110.4
Due to coefficients	40.1	37.8	3.4	85.8	82.8	-10.4
Total	100	100	100	100	100	100
<b>B. Wald test for constant parameters</b>						
F value	9.5	13.4	7.2	10.2	2.9	2.0
p value	0.00	0.00	0.00	0.00	0.00	0.00
<i>Control variables</i>						
Parsimonious	Y	Y	Y	Y	Y	Y
Demographics & employment	N	Y	Y	N	Y	Y
Household assets & house characteristics	N	N	Y	N	N	Y
Adjusted R2	0.43	0.46	0.69	0.41	0.45	0.69
N (base survey)	9,211	9,211	9,211	9,324	9,324	9,324
N (target survey)	9,324	9,324	9,324	9,348	9,348	9,348

**Note:** The decomposition of the changes in poverty for Panel A is implemented using respectively Equation (10) and the Wald test as discussed in the text. All estimates adjust for complex survey design with cluster sampling and stratification. Full model specification is provided in Appendix 3, Table 3.2.

**Table 3.4: Predicted Poverty Rates Based on Imputation, Vietnam 2012-2014 (percentage)**

Method	2014		
	Model 1	Model 2	Model 3
1) Normal linear regression model	19.8 (0.5)	19.8 (0.5)	13.2 (0.4)
2) Empirical distribution of the error terms	19.5 (0.5)	19.5 (0.5)	13.1 (0.4)
<i>Control variables</i>			
Parsimonious	Y	Y	Y
Demographics & employment	N	Y	Y
Household assets & house characteristics	N	N	Y
<b>True poverty rate</b>		13.5 (0.5)	

**Note:** Standard errors in parentheses are adjusted for complex survey design. All estimates are obtained with population weights. Method 1 uses the normal linear regression model with the theoretical distribution of the error terms and Method 2 uses the empirical distribution of the error terms. Both methods employ commune random effects. The imputed poverty rates for 2014 use the estimated parameters based on the 2010 data. 1,000 simulations are implemented. The underlying regression results are provided in Appendix 3, Table 2.2. True poverty rate is the estimate directly obtained from the survey data.

**Table 3.5: Predicted Poverty Rates Based on MI Methods, Vietnam 2012-2014 (percentage)**

Method	2012			2014		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
1) Normal linear regression model	21.5 (0.7)	21.3 (0.7)	17.8 (0.6)	16.9 (0.7)	16.9 (0.6)	13.3 (0.6)
2) Predictive mean matching model	20.4 (0.6)	20.3 (0.7)	17.1 (0.6)	16.4 (0.7)	16.4 (0.6)	13.2 (0.6)
<i>Control variables</i>						
Parsimonious	Y	Y	Y	Y	Y	Y
Demographics & employment	N	Y	Y	N	Y	Y
Household assets & house characteristics	N	N	Y	N	N	Y
<b>True poverty rate</b>		17.2 (0.5)			13.5 (0.5)	

**Note:** Standard errors in parentheses are adjusted for complex survey design. All estimates are obtained with population weights. Method 1 uses the normal linear regression model and Method 2 uses the predictive mean matching model, both with 50 simulations. Method 2 sets the number of closest observations (i.e., nearest neighbors) equal to 5. The imputed poverty rates for 2012 use the estimated parameters based on the 2010 data, and the imputed poverty rates for 2014 use the estimated parameters based on the 2012 data. True poverty rate is the estimate directly obtained from the survey data.

**Table 3.6: Predicted Poverty Rates Based on Proxy Means Testing, Vietnam 2012-2014 (percentage)**

Method	2012			2014		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Proxy mean tests	11.0 (0.4)	11.1 (0.4)	12.5 (0.5)	7.6 (0.4)	8.0 (0.4)	8.8 (0.4)
<i>Control variables</i>						
Parsimonious	Y	Y	Y	Y	Y	Y
Demographics & employment	N	Y	Y	N	Y	Y
Household assets & house characteristics	N	N	Y	N	N	Y
<b>True poverty rate</b>	17.2 (0.5)			13.5 (0.5)		

**Note:** Standard errors in parentheses are adjusted for complex survey design. All estimates are obtained with population weights. The imputed poverty rates for 2012 use the estimated parameters based on the 2010 data, and the imputed poverty rates for 2014 use the estimated parameters based on the 2012 data. True poverty rate is the estimate directly obtained from the survey data.