



Estimating poverty for refugees in data-scarce contexts: an application of cross-survey imputation

Hai-Anh H. Dang^{1,2,3,4,5} · Paolo Verme^{5,6}

Received: 9 July 2020 / Accepted: 13 May 2022 / Published online: 24 June 2022
© The Author(s) 2022

Abstract

The increasing growth of forced displacement worldwide has brought more attention to measuring poverty among refugee populations. However, refugee data remain scarce, particularly regarding income or consumption. We offer a first attempt to measure poverty among refugees using cross-survey imputation and administrative and survey data collected by the United Nations High Commissioner for Refugees (UNHCR). Employing a small number of predictors currently available in the UNHCR registration system, the proposed methodology offers out-of-sample predicted poverty rates that are not statistically different from actual poverty rates. These estimates are robust to different poverty lines, perform well according to targeting indicators, and are more accurate than those based on asset indexes or proxy means tests. They can also be obtained with relatively small samples. We additionally show that it is feasible to provide poverty estimates for one geographical region based on existing data from another similar region.

Keywords Poverty imputation · Syrian refugees · Household survey · Missing data · Jordan

Responsible Editor: Klaus F. Zimmermann

✉ Hai-Anh H. Dang
hdang@worldbank.org

Paolo Verme
pverme@worldbank.org

¹ Data Production and Methods Unit, Development Data Group, World Bank, Washington, USA

² International School, Vietnam National University, Hanoi, Vietnam

³ Indiana University, Bloomington, USA

⁴ IZA, Bonn, Germany

⁵ GLO, Essen, Germany

⁶ Research Program On Forced Displacement, World Bank, Washington, USA

JEL C15 · I32 · J15 · J61 · O15

1 Introduction

The sharp growth in the global count of forcibly displaced people during the past decade has created new challenges for host governments and aid organizations that will require a new approach to the measurement of poverty.¹ Host governments are keen to know the number and status of refugees living in their countries, as they struggle to maintain internal order while assisting the newcomers. Humanitarian organizations charged with managing displacement crises are confronted with increasing financial needs and, when these needs are not met by donors, with budget cuts and a shift from universal assistance to means-tested targeting. The increasingly protracted nature of displacement also challenges development organizations to design sustainable poverty reduction programs for displaced people and host communities. For all these actors, measuring poverty among displaced populations has become a key ingredient of any effective economic policy. It also becomes increasingly clear that achieving the SDGs (Sustainable Development Goals) number 1 goal of poverty reduction will not be possible if the forcibly displaced are excluded from the count.

Measuring poverty among refugees is not an easy task. It is more complex than for regular populations because refugees are more mobile. They also live in areas that are often difficult to reach due to environmental or security barriers. Indeed, the global count of the poor excludes, for the most part, displaced populations because these populations are not usually captured by censuses and, as a consequence, are largely excluded from consumption surveys, the main instruments used to measure poverty. The various challenges related to micro survey data collection, such as survey administration, sampling, and questionnaire design or funding, are exacerbated for displaced populations and will require years of efforts to meet the poverty measurement standards that we are now accustomed to seeing in (most) low-income countries. Not surprisingly, studies on refugee poverty are very rare. Refugee studies tend to either focus on the impact of refugees on host communities (see, e.g., Verme and Schuettler (2021) for a review) or on the impact of various policies including aid on refugees (see, e.g., Alix-Garcia et al. (2019) or Alloush et al. (2017)).

Organizations such as the United Nations High Commissioner for Refugees (UNHCR) and the World Bank are now fully committed to bridging this data gap, but past experiences with measuring poverty in low-income countries suggest that this is going to be a long-term process. For example, the UNHCR has attempted to collect consumption data for the Syrian refugees in Jordan using large-scale surveys that interview as many as 5000 households per month (or 60,000 households

¹ The UNHCR estimates that the number of forcibly displaced people at the end of 2020 was 82.4 million, the largest number since the beginning of records in 1951 (<https://www.unhcr.org/en-us/figures-at-a-glance.html>).

per year). In other refugee contexts, where fewer resources and more logistical challenges exist, such large-scale surveys may not be feasible or sustainable.²

In this paper, we make several new contributions to the poverty measurement literature. First, we address the data challenge in the refugee context by demonstrating that it is feasible to apply cross-survey imputation to obtain poverty estimates for refugees. In particular, we combine census-type administrative data that have no consumption measures with consumption household survey data collected by the UNHCR on Syrian refugees in Jordan in 2014. We subsequently employ a recently developed cross-survey imputation method to estimate poverty among these refugees. To our knowledge, this is the first experiment of its kind.³ Poverty studies that make use of cross-survey imputation methods have now become more frequent (see, e.g., Dang et al. (2019) for a recent review), but none of these studies has shed any light on refugee populations.

Second, we show that it is feasible to provide imputation-based poverty estimates for one geographical location based on the imputation model from another. This question has more practical relevance than one might think. It is well known among survey practitioners that data may often not be collected for a location due to reasons beyond one's control, such as inaccessible roads due to various forms of unexpected natural calamities (i.e., flood, storms or landslides), or conflict and violence. In the context of refugees, aside from these occurrences, even temporarily volatile security situations may also result in data not being collected for specific locations. Or it can simply be that prohibitively expensive survey costs prevent data collection at a specific location. In these cases, if the welfare variable exists for another geographical location that is comparable to the location without these data, we can employ our proposed technique to provide imputation-based poverty estimates for the latter location.

Finally, we provide theoretical and new empirical evidence that relatively small survey samples can be combined with those from the census-type registration system to provide updated estimates of poverty. Moreover, our imputation models are rather parsimonious and use variables that are already available in the UNHCR's administrative database, which is consistent with the findings in recent studies for imputation-based poverty estimates for regular populations.

Our findings show that the imputation-based poverty estimates are not statistically different from the non-predicted consumption-based poverty rates (henceforth, the "true" poverty rate), and even fall within one standard error of the latter in quite a few cases. This result is robust to various validation tests, including alternative poverty lines, disaggregated population groups, and different modelling assumptions. Furthermore, these poverty estimates are found to have smaller standard errors

² Over the past 5 years, these two organizations have sharply increased their cooperation and they recently announced the establishment of a joint data center with the objective of addressing this data challenge.

³ Beltramo et al. (2021) offer another recent application of cross-survey imputation to estimate poverty for refugees in Chad. But that paper mostly focuses on empirical estimation results.

than other poverty measures based on asset indexes or proxy means testing. They also perform better than average for standard targeting indicators such as coverage and leakage rates.

While our estimation results are encouraging, a note of caution is necessary. Our study focuses on Syrian refugees in Jordan because the data available were particularly suitable to test the methodology proposed. It is clear that validating this methodology will require further supportive evidence from other countries, refugee groups, or other sources of data. However, if our proposed imputation method is further validated, it can offer a cost-effective and logistically efficient way to obtain poverty estimates in data scarce environments.

The remainder of the paper consists of four sections. We discuss in the next section the basic theory and analytical framework. We subsequently provide in Sect. 3 the country background, a description of the data, and the empirical results for imputation for the whole population and from one geographic location to another. This section also offers various robustness tests to alternative poverty lines, disaggregated population groups, and a stronger modelling assumption. Section 4 discusses further methodological challenges related to survey sample sizes, and other related poverty measures such as asset indexes, proxy means tests, and targeting ratios. We conclude in Section 5.

2 Analytical framework

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 are available in both survey rounds, we can apply survey-to-survey imputation methods. These methods are mostly built on Elbers, Lanjouw, and Lanjouw's (2003) seminal study that imputes household consumption from a survey into a population census to measure poverty, which is commonly known as "poverty mapping." Various studies subsequently adapt this approach to implement survey-to-survey imputation for poverty estimates, such as Christiaensen et al. (2012) for China, Kenya, the Russian Federation, and Vietnam and Mathiassen (2013) for Uganda.⁴

In this paper, we apply Dang et al. (2017) imputation framework, which builds on the earlier survey-to-census imputation approach (Elbers et al. 2003; Tarozzi 2007) to provide poverty estimates for Jordan. Compared to previous studies, Dang et al.'s

⁴ Variants on this approach exist. For example, Tarozzi (2007) proposes a two-step inverse probability weighting probit estimator, with the relevant weights derived in the first step from the change in the distribution of household characteristics across the two surveys. Mathiassen (2009) also employs a probit estimator, but proposes an exact expression for the standard errors and imposes a stricter parametric functional form on the error term. This imputation method is also related to an established literature on multiple imputation (MI) in statistics (see, e.g., Rubin (1987) and Carpenter & Kenward (2013)). See Dang et al. (2014) for an extension in the context of synthetic panel data, and Dang et al. (2019) for further discussion on these methods. Also see Christiaensen et al. (2021) for a different approach to estimating household consumption and poverty based on demand systems.

(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). This technique has recently been applied (and validated) using multiple survey rounds from different countries such as various African countries, India, Tunisia, and Vietnam (Beegle et al. 2016; Cuesta and Ibarra 2017; Dang and Lanjouw 2018; Dang et al. 2019, 2021). We briefly describe this imputation method below before discussing its extensions to the refugee context.

Let x_j be a vector of characteristics representing the main observable factors that determine a household’s consumption, 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 $j = 1, 2$.⁵ 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 the household head works, 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, or some variables at the region level. Other community or regional variables can also be added since these can help control for different labor market conditions.

The following linear model is typically employed in empirical studies to project household consumption on household and other characteristics (x_j):

$$y_j = \beta'_j x_j + v_{cj} + \varepsilon_j \tag{1}$$

where v_{cj} is a cluster random effects, ε_j is the idiosyncratic error term, and y_j is household consumption typically modeled in log form. Note that we suppress the subscript that indexes households to make the notation less cluttered.⁶

For convenience, we also refer to the survey that we are interested in imputing poverty estimates for as the *target* survey ($j = 2$), and the survey that we can estimate Eq. (1) on as the *base* survey ($j = 1$). The former survey is usually more recent (or offers more disaggregated information, as in the case of a census) and has no consumption data, while the latter is usually older and has consumption data.

Assume that the explanatory variables x_j are comparable for both surveys (Assumption 1), Dang et al. (2017) define the imputed consumption y_2^1 as

⁵ 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.

⁶ 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.

$$y_2^1 = \beta_1' x_2 + v_1 + \varepsilon_1 \tag{2}$$

and estimate it as

$$\hat{y}_{2,s}^1 = \hat{\beta}_1' x_2 + \hat{v}_{1,s} + \hat{\varepsilon}_{1,s} \tag{3}$$

where the parameters β_1' (and the distributions of v_1 and ε_1) are estimated using Eq. (1), and $\hat{v}_{1,s}$ and $\hat{\varepsilon}_{1,s}$ represent the s .th random draw from these estimated distributions, for $s=1, \dots, S$. Using the same notation as in Eq. (3), the poverty rate P_2 in survey (or period) 2 and its variance can then be estimated as

$$\hat{P}_2 = \frac{1}{S} \sum_{s=1}^S P(\hat{y}_{2,s}^1 \leq z_1) \tag{4}$$

$$V(\hat{P}_2) = \frac{1}{S} \sum_{s=1}^S V(\hat{P}_{2,s}|x_2) + V\left(\frac{1}{S} \sum_{s=1}^S \hat{P}_{2,s}|x_2\right) \tag{5}$$

The intuition behind this poverty imputation method is that we predict the consumption variable in the target survey based on the estimated consumption parameters (and the error term) and their distributions using Eq. (1). Once we obtain the predicted (distribution of the) consumption variable, we use it to estimate the poverty rate as in Eq. (4).

The variance for the estimated poverty rate in Eq. (5) consists of two components, one is the sampling error (i.e., first term on the right hand side), and the other the modelling error (i.e., the second term on the right hand side). If the regression model has a good model fit, the sampling error is likely larger than the modelling error. Notably, the variance $V(\hat{P}_2)$ is related to Rubin’s (1987) variance formula, except for a component due to simulation errors in his formula.⁷ For this reason, Dang et al. (2017) recommend using a large number of simulations to make this component negligible. We follow their recommendation and use 1,000 simulations (i.e., $S=1000$) to obtain our estimates. We also provide robust standard error for the estimated poverty rate \hat{P}_2 by clustering the standard error at the district level.

It is important to check on Assumption 1 before running the models. In our specific case, this assumption is satisfied by the very nature of the data we use, since we restrict our experiment to households that are present in both data sets using personal identifiers so that both data sets contain the same individuals. In other words, since these individuals are identical in both surveys, Assumption 1 that the explanatory variables x_j are comparable in both surveys is satisfied by design.

⁷ Rubin’s variance formula is $V(\widehat{P}_2) = \frac{1}{S} \sum_{s=1}^S V(\widehat{P}_{2,s}|x_2) + V\left(\frac{1}{S} \sum_{s=1}^S \widehat{P}_{2,s}|x_2\right) + \frac{1}{S} \sum_{s=1}^S V(\widehat{P}_{2,s}|x_2)$. When S tends to infinity (or is practically large enough), the third term on the right hand side in this equality due to the simulation (computing) error will vanish, thus yielding Eq. (5) (see Dang et al. (2017) for more discussion).

For imputation on two surveys that are implemented in two different periods, Dang et al. (2017) make an additional assumption that the changes in x_j between the two periods can capture the change in poverty rate in the next period (Assumption 2). Since we use administrative and survey data that were collected by the UNHCR in the same year, this assumption can be modified as changes in x_j between the two data sources can fully capture any difference in the poverty rates estimated from these data sources (Assumption 2'). But as discussed later, since the household survey data is a subset of the administrative data, Assumption 2' is satisfied by design in our case. In summary, Assumptions 1 (and 2') are practically equivalent to, but somewhat more relaxing than, the assumption that the distributions of β_j , v_{cj} , and ε_j are the same for both the administrative and survey data.⁸

As discussed in Dang et al. (2017), while we can specify Eqs. (1) and (2) as a simple OLS model (i.e., with the random effects v_j being subsumed into the error terms), modelling the random effects explicitly helps improve the precision of estimation results. The random effects model offers an advantage over the OLS model by capturing the between-cluster variations thanks to the additional information offered by the cluster random effects. Put differently, v_j is instrumental not only in estimating β_j but also for our estimates of poverty in survey 2 as a component of the predicted household consumption. Also different from the traditional econometric model that estimates the *impacts* of x_j on y_j , our focus is on *predicting* y_j conditional on x_j .⁹ As such, worries about endogeneity of x_j pose far less important, if any, concerns in our context.

It can also be useful to note that in contexts where there are few explanatory variables x_j that are comparable between the two surveys (say, when we impute from a household consumption survey into a labor force survey), the role of the random effects v_j is even more important. In this case, explicitly modelling the random effect term v_j can help better control for the larger variations due to the unobserved cluster characteristics that are not available in both surveys. Indeed, empirical evidence from various countries including Jordan and Vietnam suggests that the estimated variance of v_j tends to be larger where the regression based on Eq. (1) has lower goodness-of-fit (i.e., a lower R^2) (Dang et al. 2017, 2019). We provide a more detailed description of the imputation procedures and the user-written Stata routine in Appendix 1, Part A.

⁸ Assumption 1 is testable, for example, by employing standard t-tests to compare the distributions of the explanatory variables x_j in two different surveys. But Assumption 2' is, by design, not testable since the consumption data are missing in the target survey. However, if consumption data from previous survey rounds are available, Dang et al. (2017) suggest that we can test for both assumptions using these data, which can provide supporting evidence for applying imputation in the present surveys. See also the related discussion on a similar assumption in section 3.2 below when we impute from one region into another.

⁹ Dang et al. (2017) offer empirical evidence for Jordan where omitting such random effects render estimates severely biased. Note that for these two channels of impacts, we can use standard tests such as the Breusch and Pagan test to test for the statistical significance of the random effects v_j on the latter channel (as a component of the predicted household consumption), but not on the former channel (in estimating β_j).

For the purpose of testing this method, we use administrative and survey data that cover the same households which can be matched with unique identifiers. This allows us to split the sample artificially, simulate a cross-survey imputation exercise, and compare predicted poverty with true poverty. This is an ideal data scenario that provides the conditions for a rigorous test of the cross-survey imputation model proposed.

3 Application to Syrian refugees in Jordan

3.1 Country background and data

The Syrian refugee crisis is one of the largest refugee crises ever recorded in history if we consider the number of displaced people relatively to the population of the country of origin and the countries of destination. The crisis started in the spring of 2011 following clashes between protestors and government forces in several major cities and quickly descended into a complex civil war. By 2014, 6.7 million people had been displaced internally in the country, about 1.5 million people fled the country with their own means, and an additional 3.7 million people were hosted as refugees mostly in neighboring countries. As a result, about half of the Syrian population was considered displaced in 2014. For some countries, Syrian refugees also represented a major population shock. In 2014, Syrian refugees accounted for about 20% of the population of Lebanon and about 10% of the population in Jordan. The incidence of such immigration for these countries is among the highest ever recorded in history (Verme and Schuettler 2021).

The UNHCR has the mandate to protect and assist refugees in host countries and its role in the aftermath of a crisis is to find shelter, provide food and cash assistance and assist with basic services such as health and education. In order to provide these services, the UNHCR employs a system of mandatory registration for all refugees or asylum seekers requiring assistance that implies the collection of personal information. All individuals seeking protection, assistance and refugee status are expected to register with the host government or the UNHCR and, for this purpose, the UNHCR maintains a profile Global Registration System (proGres). This system contains biometric and socio-economic information on asylum seekers and refugees and serves the purpose of identifying the persons most in need and determining the type of protection and assistance they require. ProGres does not offer information on income, consumption or expenditure but contains a rich list of variables that are potentially closely associated with these monetary indicators. This proGres registration system is the most comprehensive database on refugees in any country where the UNHCR manages the registration of refugees.¹⁰ This is the case of Jordan, the country we consider in this paper.

¹⁰ In some countries, such as Turkey, the host government or other agencies manage the registration process.

In addition to the registration system, the UNHCR conducts sample surveys and home visits for a variety of purposes, such as protection of different categories of vulnerable populations or assistance of targeted programs such as the cash or food assistance program. In the case of Jordan and the Syrian crisis, the UNHCR and the World Food Program (WFP) have been conducting a variety of surveys as well as extensive home visits that allowed researchers to analyze refugee conditions as had never been done before.

The paper uses two data sets: the Jordan proGres registration system (PG for short) as of December 2014 and the Jordan Home Visits survey, round II data (HV for short) collected between November 2013 and September 2014. Both data sets were provided by the UNHCR in the context of the joint World Bank-UNHCR study on the welfare of Syrian refugees (Verme et al. 2016). These comprehensive data sets have the distinct advantage that they can be linked by a common identification number. We can therefore trace the same individuals and households across the two sources of data for the same period of time.

The proGres registration system is what we consider the “census” of refugees. This data set has no information on consumption but contains socio-economic characteristics for all registered individuals and households. Variables available in the PG data include, among others, date of birth, place of birth, gender, date and reasons of flight, arrival date in Jordan, registration date, ethnicity, religion, education, professional skills, and occupations in the countries of origin and asylum.

The HV data have been collected in successive rounds since 2013 for the purpose of targeting refugees with cash assistance programs and they contain information on income and expenditure as well as a large set of individual and household socio-economic characteristics. Although this is not a sample survey, for the purpose of this study we will consider this data set as our hypothetical sample survey. The HV data we use cover about one-third of all registered persons in Jordan in 2014 and are a sub-sample of the PG data. Our experiment is restricted to households present in both data sets, a total of approximately 40,000 households. For these households, the socio-economic characteristics of the household and its members are the same by design. This data setup practically implies that Assumption 2' is not needed in our validation context. In fact, when households are interviewed during the home visits, the variables that are common in HV and proGres data sets are expected to be updated in proGres if these variables are outdated.

As unit of observation, we use what the UNHCR refers to as the “case.” A case is a group of individuals who register at the UNHCR together with a Principal Applicant (PA) who takes responsibility for the group. This group may be a family, a household or an extended household. For simplicity and practical purposes, we will consider a case and the PA as a household and its head respectively. The poverty line used is 50 JD/month/person, which is what the UNHCR used in 2014 to select beneficiaries of the cash assistance program. In 2014, this poverty line was higher than the international poverty line and lower than the poverty line used for the Jordanian population. In our case, this poverty line is more relevant than either the national or international poverty line, as it corresponds to what the UNHCR—the UN agency specialized on refugees—considers a sufficient amount to meet basic needs. As

for the welfare aggregate, we use the same aggregate used by Verme et al. (2016), which provides detailed explanation of the consumption aggregate.

3.2 Estimation results

3.2.1 Imputation for the whole population

For the purpose of this paper, the HV data are considered the “survey” data containing information on consumption and the PG registration data are our “census” data containing predictors of consumption but no consumption data. The primary objective of the exercise is, therefore, to test how accurate the estimated poverty figures are using the HV data alone (as both the base and the target survey).

As a first step, we generated two samples by extracting 50% of observations from the HV sample randomly (sample 1) and using the remaining observations as second sample (sample 2). We then impute from sample 1 to sample 2 to obtain the imputation-based poverty rate in sample 2, and we compare this imputed poverty rate with the true poverty rate that can be directly calculated from sample 2 for validation purposes. We also implement this imputation process the other way around by imputing from sample 2 to sample 1 and then compare with the true poverty rate in sample 1. Naturally, given that the two sub-samples are extracted randomly from the same original sample, we should expect these sub-samples to exhibit small differences and provide similar estimates.¹¹ In the next section, we will perform additional tests using samples with higher degrees of heterogeneity.

We consider three model specifications based on different sets of regressors for further comparison. Specification 1 employs the variables that are only available in the PG data set (PG-specific variables), which include case (household) size and the PA’s demographic and employment characteristics (age, gender, different levels of education achievement, occupation group, marital status, and the governorate or city of original residence in the Syrian Arab Republic).¹² Specification 1 also includes variables related to the PA’s immigration status such as the type of border crossing point and the legal status of entry. It is the main model specification. Specification 2 adds to specification 1 several variables that are only available in the HV data and that are related to home ownership, household assets, utilities, and the physical characteristics of the house. These variables include whether the house is rented or owned, the quality status of the kitchen, electricity access, and the ventilation system, the living area of the house (as measured by the number of square meters per person), whether the house is made of concrete, and the availability of tap water and piped sewerage system. Specification 3 further adds to specification 2 HV-specific

¹¹ For valid analysis, it is important that this sub-sample extraction process should be implemented well. Various computing softwares including Stata offer commands such as “sample” that can carry out this task.

¹² We consider the following five levels of education achievement: (1) below 6 years of schooling, (2) 6–8 years of schooling, (3) 9–11 years of schooling, (4) 12–14 years of schooling, and (5) university education or higher.

variables related to the household's shock-coping strategies (i.e., whether receiving humanitarian assistance, help from the host family, or from the host community), whether the household has a valid certificate of asylum, and whether the household receives UNHCR financial assistance.

We are particularly interested in examining whether adding HV-specific variables to the main specification in specification 1 can improve the accuracy of the estimates. If we find that some key predictors of household expenditure—that are not available in the PG data—can improve the accuracy of the poverty predictions significantly, this provides a strong argument for collecting this information upfront when refugees are first registered. Vice versa, if poverty estimates imputed with the PG data are not statistically different from the true rates (i.e., those produced directly from the HV data), this would suggest that existing PG variables are already suitable to produce reliable poverty estimates.

We also use two alternative models to estimate regression errors: one where we assume a standard normal distribution for the error term, and another where we remove this assumption and use the (non-parametric) empirical distribution of the error term instead. If the error term is not distributed normally, our poverty estimates would be biased, and a non-parametric model based on the empirical distribution would likely perform better.

Table 1 presents the summary results and Appendix Table 1 in Appendix 2 provides the full regression results. Table 1 shows that all the estimates using the normal linear regression model fall within the 95% confidence interval (CI) of the true poverty rate, for both sample 1 and sample 2. In other words, these estimates are not statistically significantly different from the true poverty rates reported at the bottom of the table. Estimates using specifications 2 and 3 with more variables on household assets and house characteristics are somewhat better and closer to the true poverty rate than those using specification 1 for both samples. For example, the poverty estimate using specification 1 (Table 1, first column) is 52.6%, which is 1.1 percentage points larger than the true poverty estimate of 51.5%. The poverty estimate using specification 3 (Table 1, third column) is 52.3%, which is 0.8 percentage points less than the true poverty estimate. This is likely because imputation models that include household assets are usually found to perform better than those that do not (Christiaensen et al. 2012; Dang et al. 2019).¹³

The alternative imputation model based on the empirical distribution of the error terms (Table 1, row 2) performs better than those based on the normal linear regression, although both methods provide estimates within the 95% CI of the true poverty rates. In addition, for both samples, while specification 2 still performs

¹³ The district random effects v_j are estimated to have a variance of 0, which indicates that it has no contribution to the model fit (Appendix 2, Table 2.1). More generally, adding more control variables does not necessarily lead to a better model fit. While this result may appear counter-intuitive, one possible reason is that doing so may overfit the data and thus does not offer more accuracy, which is shown with empirical evidence from India and Jordan (Dang et al. 2019). A recent theoretical study also suggests that for misspecified regressions, adding more variables may result in larger inconsistency (De Luca, Magnus, and Peracchi 2018). Also note that the standard errors around the true poverty estimates are larger than those for the imputation-based estimates, since the latter are model-based.

slightly better than specification 1, specification 3 now performs somewhat worse than specification 1. Yet, since the standard error around the true poverty rate is 2.3 percent for Sample 1 and 2.6 percent for Sample 2, all these differences are in fact still within one standard error of the true poverty estimates. As such, statistically speaking, the differences between the three specifications and the true poverty rates for both samples are negligible. Finally, since the HV data set is originally a non-random subsample of the PG database, we also re-run Table 1 using only variables that are available in the HV data set. The estimation results, shown in Appendix Table 2 in Appendix 2, are very similar to those in Table 1.

In summary, the set of variables available in the PG registration data seems sufficiently powerful to predict the true poverty rate with a 95% accuracy level. This is very encouraging considering that these variables were not selected for this purpose when the registration system was designed.

3.2.2 Imputation from one geographical region to another

We turn next to examining the situation where the consumption data (or the welfare variable) are not available for a particular geographical location, but are available for another similar location. The control variables x_j , on the other hand, are available for both locations. Making similar assumptions, but for two locations instead of two data sources, we can employ the same imputation technique to impute from one location to the other to obtain poverty estimates.¹⁴

We consider two such governorates (regions) in Jordan, the Balqa governorate and the Irbid governorate. The Syrian refugees in these two governorates have very similar consumption levels (i.e., around 150 JD/month/person) and poverty rates (i.e., 51–52%). *t*-Tests suggest that the x_j characteristics are similar mostly for the case sizes and for some, but not all the, other variables (Appendix 2, Appendix Table 3).¹⁵ As such, it can be an empirical question where we can impute from one governorate into another in a similar manner to the imputation exercise with the two samples in Table 1. Notably, in a real-life setting where we do not have consumption data for one region (but say, know from older data that the two regions have comparable income and poverty levels), it is even more important to rely on the assumption of similar x_j characteristics between the two regions.

Table 2 shows that estimates are somewhat less accurate when we impute from the Irbid governorate into the Balqa governorate, but still fall within the 95% CI of the true poverty rate. On the other hand, all estimates for the Irbid governorate are within one standard error of the true poverty rate. Estimates for two other governorates with similar levels of consumption and poverty, Ajloun and Jarash, also perform quite well and fall within one standard error of the true poverty rates (Appendix 2, Appendix Table 4).

¹⁴ See Dang et al. (2019) for reviews of studies that impute from one survey to another over different time periods.

¹⁵ In other words, these test results suggest that the distributions of y_j and x_j are (mostly) similar between the two regions.

Table 1 Predicted poverty rates for Syrian refugees based on imputation, ProGres and HV Data 2014

Method	Sample 1			Sample 2		
	Spec. 1	Spec. 2	Spec. 3	Spec. 1	Spec. 2	Spec. 3
Normal linear regression model	52.6 (2.0)	52.5 (2.0)	52.3 (2.1)	53.1 (2.0)	53.0 (1.9)	53.0 (2.0)
Empirical errors model	51.3 (2.2)	51.3 (2.2)	51.5 (2.3)	51.8 (2.2)	51.8 (2.1)	52.2 (2.1)
Control variables						
Demographics and employment	Y	Y	Y	Y	Y	Y
Household assets and house characteristics	N	Y	Y	N	Y	Y
Shock-coping strategies and receiving UNHCR assistance	N	N	Y	N	N	Y
Overall R^2	0.43	0.48	0.53	0.45	0.50	0.54
N	19,001	19,001	19,001	18,999	18,999	18,999
True poverty rate	51.5 (2.5)			51.6 (2.4)		

The full regression results are provided in Appendix Table 1, Appendix 2. Specification 1 employs variables from the ProGres database only, and specifications 2 and 3 employ variables from both the ProGres and HV databases. The estimation sample is generated by splitting the data into two random samples named sample 1 and sample 2. The imputed poverty rate for sample 1 and sample 2 are shown in the first and second three columns, respectively. The true poverty rate for each sample is shown at the bottom of the table. Robust standard errors in parentheses are clustered at the district level. We use 1000 simulations for each model run

3.3 Robustness checks and extensions

This section provides several robustness tests and extensions for the results presented in Table 1. We offer estimates for different poverty lines, more disaggregated population groups, and alternative estimation methods.

3.3.1 Sensitivity to the poverty line

One important question relates to the performance of the model specifications when the poverty line and the poverty level change. With the poverty rate close to 50%, we have half of the sample below and half above the poverty line. But estimating poverty accurately when the poverty rate is around 5–10% may be more difficult. In Fig. 1, we used variations of the poverty line ranging from 0 to 60% of the population (i.e., 0 to 60th percentile of the consumption distribution) to reproduce poverty estimates using imputations from sample 1 to sample 2 and the two models described. The results show that with a low poverty line and a low poverty rate, the empirical errors model is more accurate in estimating true poverty than the normal linear model, while the normal linear model performs somewhat better when the poverty line and the poverty rate are high. So both methods result in predictions that are within the 95% CI of the true values, but these two methods slightly differ in accuracy as the poverty line and the poverty rate change. Estimation results are

Table 2 Predicted poverty rates for Syrian refugees based on imputation for two different regions, ProGres and Home Visit Data

Method	Balqua			Irbid		
	Spec. 1	Spec. 2	Spec. 3	Spec. 1	Spec. 2	Spec. 3
Normal linear regression model	55.4 (1.6)	55.7 (2.1)	55.5 (2.2)	50.8 (1.0)	51.4 (0.8)	51.8 (1.3)
Empirical errors model	54.4 (1.8)	54.6 (2.3)	54.7 (2.4)	50.4 (1.1)	51.0 (0.9)	51.7 (1.4)
Control variables						
Demographics and employment	Y	Y	Y	Y	Y	Y
Household assets and house characteristics	N	Y	Y	N	Y	Y
Shock-coping strategies and receiving UNHCR assistance	N	N	Y	N	N	Y
R^2	0.45	0.49	0.54	0.49	0.52	0.54
N	1688	1688	1688	10,128	10,128	10,128
True poverty rate	52.0 (2.5)			51.2 (0.9)		

Specification 1 employs variables from the ProGres database only, and specifications 2 and 3 employ variables from both the ProGres and HV databases. The estimation sample is restricted to the Balqua and Irbid regions. The imputed poverty rate for these two regions are shown in the first and second three columns, respectively. The true poverty rate for each sample is shown at the bottom of the table. Robust standard errors in parentheses are clustered at the district level. We use 1000 simulations for each model run

similar if we impute from sample 2 to sample 1 (Appendix Fig. 1). A possible explanation is that, as the number of poor households (sample size) increases, the distribution of the error term approaches a normal distribution. Therefore, as a rule of thumb, we should expect the normal linear model to perform better with larger samples.

3.3.2 Disaggregated population groups

The next question is whether the results are sensitive to changes in the specified population groups. We know from our regressions that the most important predictor of poverty is case size (see also Verme et al. 2016). If the prediction capacity of the model specification is sensitive to changes in household characteristics, changing case size would likely have the most impact. We impute from sample 1 to sample 2 and re-estimate poverty for each of the case sizes. To ensure that the estimation sample size is reasonable, we combine all the cases with eight or more individuals into a single group (which makes up roughly 6% of the estimation sample). We employ the two error estimation models and plot the estimated poverty rates against case size in Fig. 2.

Both methods provide similar results and both sets of results are within the 95% CI of the true values. In this case, we do not observe any sharp difference between the two error estimation models. As before, we repeat the exercise imputing from sample 2 to sample 1 (Appendix Fig. 2) and find that the results are virtually

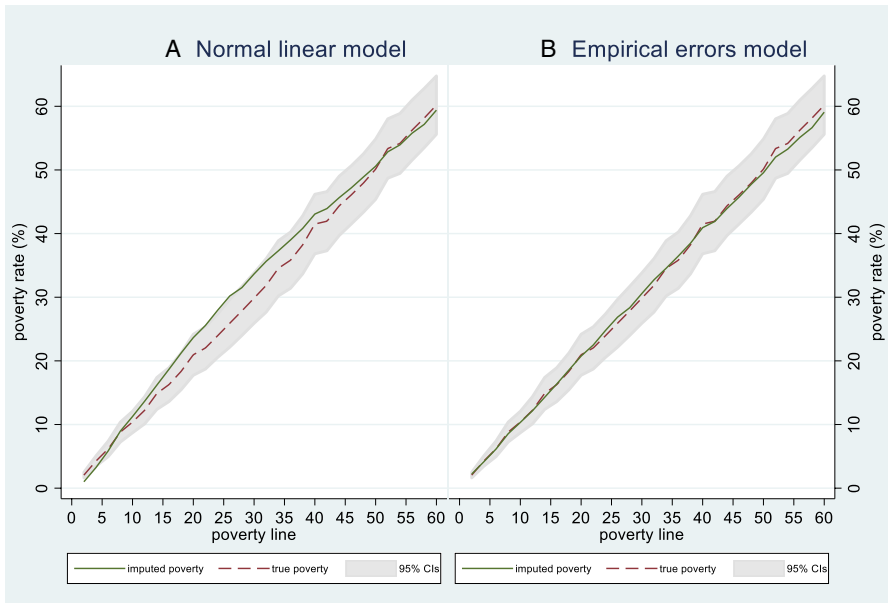


Fig. 1 Predicted poverty rates for different poverty lines

unchanged. Given the association between case size and poverty, both estimation models seem to perform reasonably well.

3.3.3 Models with a stronger parametric assumption

One alternative approach to the present poverty estimation models is to directly run a probit or logit model on poverty status rather than a linear model on expenditure (and subsequently convert the predicted expenditure into poverty estimates). In this case, the population is first divided into poor and non-poor groups using the poverty line and this variable is then used as the dependent variable in a logit or probit model to predict poverty. The difference with a probit (or logit) model is that we need to make a stronger parametric modeling assumption on the dependent variable, which can result in more accurate estimation results if this assumption is correct. But the disadvantage with such models is that estimation results may be worse if the modeling assumption is violated. Furthermore, the conversion of the continuous expenditure variable into a binary variable indicating poverty status can result in loss of information and generally less efficient estimation (Ravallion 1996). Indeed, Appendix Table 5 in Appendix 2 shows that while the estimates using the probit and logit models are still within the 95% CI of the true rates, they are somewhat less accurate than those obtained using the empirical errors model in Table 1. For example, the estimated poverty rate using specification 1 and sample 2 for the logit model is 53.1%, which is 1 percentage point larger than the corresponding figure of 51.8% for the empirical errors model (compared with the true poverty rate of 51.6%).

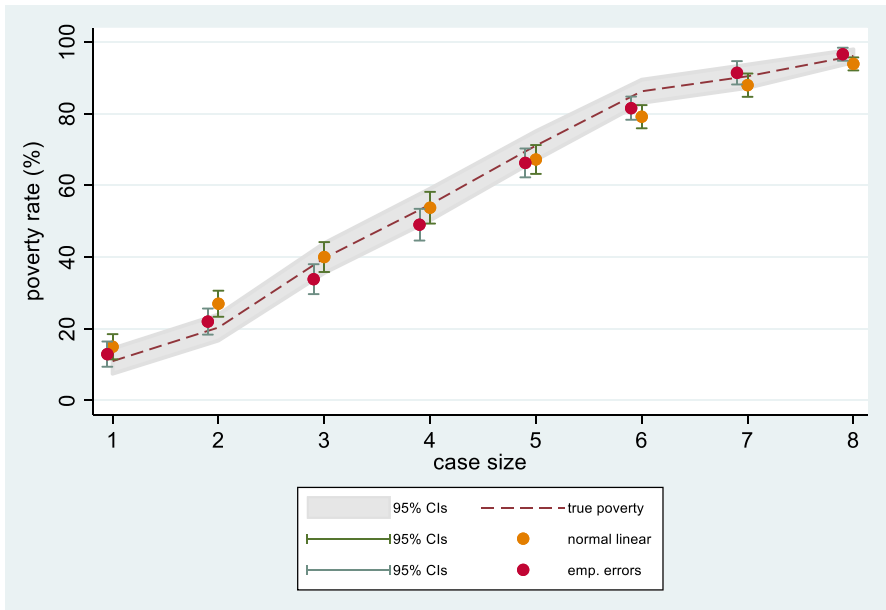


Fig. 2 Predicted poverty rates for different population sub-groups

4 Challenges for applications in other contexts

The data on Syrian refugees in Jordan that we analyze are of relatively high quality in the context of refugee populations. In this section, we discuss methodological challenges in other contexts where data quality may not be as good and some potential for applying our method to other contexts with similar data.

4.1 Small survey sample sizes

One practically relevant question is how large the imputation sample should be to obtain accurate poverty estimates.¹⁶ On the one hand, a large sample size can provide estimates with more accuracy and generally better statistical properties than a small sample size; but on the other hand, it is also more expensive and demands more logistical and technical resources to implement. A balance should be reached between these trade-offs. In most conflict situations, however, the logistical and technical constraints may pose especially severe challenges for data collection efforts.

¹⁶ Note that this challenge of finding an appropriate sample size is in the context of predicted values based on regression models, which is different from calculating the sample sizes for other purposes, such as hypothesis testing. For the latter, see, e.g., Cohen (1988) for a textbook treatment.

Park and Dudycha (1974) offer some theoretical guidance on selecting the appropriate sample size for obtaining regression-based prediction estimates. In particular, we want to find the sample size n such that

$$\Pr[(\rho^2 - \rho_c^2) \leq \varepsilon] = \gamma \tag{7}$$

where ρ^2 is the maximum (or true) multiple correlation coefficient (R^2) possible for Eq. (1) in the population, and ρ_c^2 is the correlation between the predicted value using Eq. (1) and the original y variable. ρ_c^2 is usually referred to as the squared cross-validity correlation coefficient.¹⁷ A good sample size would ensure that the probability of obtaining an estimate within an acceptable error interval (ε) around ρ^2 has reasonably good power (γ). In other words, after we specify some (acceptable) values for ε and γ , the sample size n that satisfies Eq. (7) can be derived as follows:

$$n = \left\lceil \delta^2 \frac{1 - \rho^2}{\rho^2} \right\rceil + p + 2 \tag{8}$$

where δ^2 is the noncentrality parameter for the noncentral Student's t distribution with $p-1$ degrees of freedom associated with Eq. (7), and p is the number of predictors (i.e., explanatory variables) in the estimation model. We provide a more detailed description of Park and Dudycha's (1974) derivations in Appendix 1, Part B.

We apply Eqs. (7) and (8) above and calculate the sample sizes where ε ranges from 0.01 to 0.05, and γ ranges from 0.90 to 0.99.¹⁸ These ranges should cover most of the cases of interest, with a smaller value for ε and a larger value for γ requiring a larger sample size. In particular, the smallest sample size given these values would be where ε and γ are respectively 0.05 and 0.90, or the probability that ρ_c^2 falls within a bandwidth of 0.05 around the true value of ρ^2 is 0.90. Increasing this probability to, say, 0.95 and tightening ε to 0.02 would require a larger sample size. We also assume that ρ^2 is 0.45 and the number of predictors p is 27, which are the parameters obtained under specification 1 for sample 2 in Table 1. The estimates provided in Table 3 suggest that the minimum sample size is 389 observations (where ε and γ are respectively 0.05 and 0.90), and a reasonably good sample size is 1,068 observations (where ε and γ are respectively 0.02 and 0.95). Table 3 also indicates that the largest sample size required to increase γ to its maximal value of 0.99 and reduce ε to its minimal value of 0.01 is 2,509 observations.

While Park and Dudycha's formulae provide useful theoretical guidance on the appropriate sample size, these formulae were originally developed for the simple OLS model. As such, their model does not explicitly take into account the cluster

¹⁷ The intuition is that, since the best job that we can do with prediction is to reproduce the original y variable, the correlation between the original y variable and its predicted value should always be less than or equal to the true correlation in the population.

¹⁸ Pituch and Stevens (2016) consider 0.05 (or smaller) and 0.90 (or larger) are respectively good values for ε and γ .

random effects model. Thus, it remains an empirical question whether these formulae can apply to our context.

We address this question and show estimation results in Fig. 3. The estimates in this figure are restricted to sample 2 from which 10 sub-samples of different sizes—including 200, 400, 600, 800, 1000, 1500, 2000, 3000, 4000, and 5000 observations—have been extracted randomly. The first five samples represent situations ranging from less than the theoretical minimum sample size (200) to less than the theoretically ideal sample (1000), and the last five samples represent situations ranging from the theoretically ideal sample (1,500) to a common and reasonably good sample size in practice (5000). Specification 1 is then re-run on each sub-sample, the underlying regression results are provided in Appendix 2, Appendix Table 6.

The results show that almost all the poverty estimates fall within one standard error of the true poverty rate, and that there appears no strong relationship between the number of observations and the accuracy of the results.¹⁹ Yet, plotting all the estimation results with the normal linear and empirical errors models in Fig. 3 yields two additional observations. The first is that estimates fluctuate less around a sample of 1000 observations with both estimation methods, and the second is that the normal linear model tends to overestimate the true value more than the empirical errors model.²⁰ We can also observe from Appendix Table 6 that the estimated R^2 of the model specifications tends to decline and also stabilize as the number of observations increases, which is consistent with the well-known statistical result that estimates for R^2 in smaller samples may be larger than their population counterparts (see, e.g., Pituch and Stevens (2016)). In essence, good estimates can also be obtained with very small samples but samples of medium size, around 1,000 observations in our case, seem to offer reasonably stable estimates while containing survey costs. This sample size is also consistent with the theoretical results offered in Park and Dudycha (1974).

These results have practical relevance. The HV data used in this study were collected with field visits that covered about 5000 households per month, or 60,000 households per year. We have shown that covering about one-sixtieth of this number, or 1000 households per year, may be sufficient to provide reliable poverty statistics.²¹

¹⁹ All estimates fall within the 95 percent CI of the true poverty rate but are not shown for lack of space.

²⁰ Note that we are only considering a single summary statistics for the whole population (the poverty rate). If we were to estimate disaggregated statistics by geographical areas or population groups for example, sample sizes would have to be reconsidered.

²¹ This result should not be interpreted as suggesting that 1,000 observations are sufficient for a multi-purpose survey. In our case, we estimate this number to be sufficient to estimate one statistic (the poverty rate) whereas most surveys have typically multiple objectives and require the correct estimation of multiple statistics. The latter are the reasons behind common tasks associated with designing a survey such as power calculations, stratification, and clustering of the sample.

Table 3 Theoretical sample size as a function of the population parameters

Epsilon	Gamma		
	0.99	0.95	0.90
0.01	2509	2137	1954
0.02	1253	1068	976
0.03	835	711	650
0.04	625	533	487
0.05	500	426	389

Estimates are based on the formulas provided in Park and Dudycha (1974). We use the given parameters, the R^2 value of 0.45, and the number of predictors of 27 under specification 1 from Table 1



Fig. 3 Predicted poverty rates for different sample sizes

4.2 Related measures of poverty

How does our proposed poverty imputation method compare with alternative estimation methods such as asset (wealth) indexes and proxy-means tests? We examine in this section each of these two alternatives, together with the related exercise of targeting. This is a particularly important question for the UNHCR, which uses asset indexes to measure well-being in place of consumption in many places where consumption is not available. Other development organizations such as the WFP also often employ asset indexes to target food assistance programs for

refugees; one such recent application was for the Malian refugees in Niger (Belt-ramo et al. 2019).

4.2.1 Asset index

We consider a variant of Eq. (1) where the left-hand side variable, household consumption y_j is now missing but we have data on household assets a_j , which is a subset of x_j . Still, we want to generate a wealth index w_j which offers the best combination of (the elements of the different) household assets a_j . Suppressing the household index to make the notation less cluttered, this can be expressed as follows

$$\alpha' a_j = w_j \quad (9)$$

where α is the (vector of) weights we place on the a_j to generate the wealth index w_j . A common way to derive α is through Principal Component Analysis (PCA), another way is just to sum up all the assets available in a_j .

We briefly describe here a couple of reasons that make asset indexes more likely to result in biased estimates of poverty. First, the wealth index w_j does not include the non-asset components, which is equivalent to the well-known issue of omitted variable bias. Second, β_1 and α are generally different from each other, since the estimator for α maximizes the variance in a_j , while the estimator for β maximizes the variance in y_j .²² Finally, in a refugee context, the temporary nature of displacement likely affects refugees' behaviors in terms of accumulation and use of assets. For example, refugees may choose not to invest as much in high-quality durables as regular households do. This practical aspect may further make assets (alone) an even less reliable data source for poverty estimation in a refugee context.

Table 4 provides an illustrative example where we generate the wealth (assets) index using both the simple counting method (Table 4, model 1) and the PCA method (Table 4, models 2 and 3) on the two samples. 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. In other words, the five quintiles provide five different slices of the consumption distribution. The list of assets for model 1 and model 2 include the status of the kitchen, electricity, ventilation system, whether the house is made of concrete, and the availability of tap water and piped sewerage system. Model 3 adds to model 1 the house size and the condition of household furniture.

Consistent with our earlier discussion, the quintiles based on the wealth index can only capture between 12 and 35% of the corresponding quintile based on the consumption distribution. For example, the poorest wealth index quintile in model 3 can correctly capture only 32% (34%) of the poorest consumption quintile in sample 1 (sample 2). The correlation between asset indexes and household consumption is not

²² See Rencher (2002, pp. 389) for a graphical illustration of the general difference between principal component analysis and OLS methods, and Dang et al. (2019) for further discussion on asset indexes.

very strong, ranging between 0.21 and 0.23.²³ These are half as strong as a correlation of roughly 0.44 and 0.48 (respectively for specification 1 and specification 3 in Table 1) between the original household consumption and the predicted consumption obtained from our method. This provides supportive evidence for our earlier discussion that asset indexes may not be good predictors of household welfare and poverty, particularly in a refugee context.

4.2.2 Proxy means test

Most of the estimates based on proxy means testing start from a general equation that can be described as follows:

$$y_j^p = \beta_j^{p'} x_{j,p} \quad (10)$$

where the vector of coefficients $\beta_j^{p'}$ is obtained from the regression using another survey (see, e.g., Coady et al. 2014; Ravallion 2016; Brown, Ravallion, and van de Walle 2018). As such, proxy means tests are rather similar to the poverty imputation model expressed in Eq. (1) in terms of the deterministic part $\beta_j^{p'} x_{j,p}$. Yet, one key difference between the two methods is that the error terms $v_{cj} + \varepsilon_j$ in Eq. (1) are often omitted in Eq. (10). Consequently, the mean and the variance of the predicted consumption based on proxy means testing would likely provide biased estimates of household consumption. Even when $x_{j,p}$ is identical to x_j —or when the error terms ($v_{cj} + \varepsilon_j$) are negligible—there is no bias in the estimated mean consumption, but there is still bias in the estimated variance.²⁴

Table 5 provides poverty estimates using the proxy means test method as in Eq. (10). A couple of remarks are in order to illustrate the results. First, the estimates fall outside the 95% CI of the true poverty rate for both samples, which suggests that the error terms $v_{cj} + \varepsilon_j$ in Eq. (1) are not negligible. On the other hand, consistent with our theoretical discussion above, the standard errors for the poverty estimates in Table 5 range from 2.5 to 2.9%, which are roughly 10 to 25% larger than those based on the poverty imputation methods shown in Table 1.

4.2.3 Targeting ratios

The importance of modeling the error terms can be further appreciated when we estimate such targeting ratios as the percentage of the poor population that are correctly identified (i.e., coverage rate) and the percentage of the population identified as poor who are not poor (i.e., leakage rate). Note that just as with the poverty rate,

²³ These correlation coefficients between the wealth indexes and consumption are weaker than those observed in Filmer and Scott (2012) for 11 other countries around the world (which range from 0.39 to 0.72 for these countries). Indeed, assets may capture different aspects of household welfare other than consumption, which could result in the weak correlation between the wealth indexes and consumption.

²⁴ Dang et al. (2019) offer more detailed discussion and more formal proofs of these results.

Table 4 Population distribution by asset indexes vs. consumption

Per capita consumption	Sample 1			Sample 2		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Poorest quintile	32.9	32.8	32.4	34.7	33.7	34.0
Quintile 2	27.2	27.0	22.6	26.1	26.9	21.7
Quintile 3	26.5	22.9	19.0	28.1	23.8	21.6
Quintile 4	12.6	12.4	22.2	13.5	12.7	22.1
Richest quintile	19.9	23.6	25.8	19.4	24.2	26.1
Correlation with household consumption	0.21	0.22	0.22	0.21	0.22	0.23
<i>N</i>	19,001	19,001	18,558	18,999	18,999	18,610

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 models 1 and 2 include the status of the kitchen, electricity, ventilation system, whether the house is made of concrete, and the availability of tap water and piped sewerage system. Model 3 adds to model 1 the house size and the condition of household furniture

we need to do multiple simulations to estimate these targeting rates. In particular, the formulae for the coverage rate and the leakage rate are as follows:

$$coverage = \frac{1}{S} \sum_{s=1}^S \frac{1}{N} \sum_{i=1}^N I(\hat{y}_{2i,s}^1 \leq z_1 | y_{2i,s}^1 \leq z_1) \quad (11)$$

$$leakage = \frac{1}{S} \sum_{s=1}^S \frac{1}{N} \sum_{i=1}^N I(\hat{y}_{2i,s}^1 \leq z_1 | y_{2i,s}^1 > z_1) \quad (12)$$

where $I(\cdot)$ is the indicator function, “|” inside the parentheses is the conditional operator, and the subscript i indicates households.

Estimates based on the empirical errors model, shown in Table 6, suggest that Specification 1 can provide a reasonable coverage rate of 70%, and a leakage rate of roughly 32%. As we add more control variables to this specification, these rates unsurprisingly improve. In particular, the coverage rate increases by almost 4 percentage points, while the leakage rate decreases by 3 percentage points when we switch from Specification 1 to the richer Specification 3. These rates compare favorably with recent estimates of the coverage rate and leakage rate of 64% and 31%, using the proxy-means test for a similar poverty rate of 40% for nine African countries (Brown et al., 2018).

4.3 Potential application to other settings

The methodology proposed by this paper can be replicated in most countries hosting refugees. As an example, Table 7 reports proGres data on nine refugee-hosting countries in Sub-Saharan Africa including Cameroon, Chad, the Republic of Congo, the Democratic Republic of Congo, Ethiopia, Kenya, Niger, Rwanda, and the United Republic of Tanzania. Some of these countries such as the Republic of Congo, the

Table 5 Predicted poverty rates for Syrian refugees based on proxy means test, Home Visit Data 2014

Method	Sample 1			Sample 2		
	Spec. 1	Spec. 2	Spec. 3	Spec. 1	Spec. 2	Spec. 3
Proxy means test	59.5 (2.7)	59.0 (2.8)	57.1 (2.9)	60.5 (2.7)	59.6 (2.5)	58.1 (2.5)
Control variables						
Demographics and employment	Y	Y	Y	Y	Y	Y
Household assets and house characteristics	N	Y	Y	N	Y	Y
Shock-coping strategies and receiving UNHCR assistance	N	N	Y	N	N	Y
R^2	0.43	0.48	0.53	0.45	0.50	0.54
N	19,001	19,001	19,001	18,999	18,999	18,999
True poverty rate	51.5 (2.5)			51.6 (2.4)		

The full regression results are provided in Appendix Table 1, Appendix 2. The estimation sample is generated by splitting the data into two random samples named sample 1 and sample 2. We then impute from sample 1 to sample 2 and vice versa to obtain the imputed poverty rate for each sample. The true poverty rate for each sample is shown at the bottom of the table. Robust standard errors in parentheses are clustered at the district level. We use 1000 simulations for each model run

Democratic Republic of Congo, and Chad are countries that typically suffer from lack of quality data. This table reports the numbers of observations and the percentages of total frequencies for eight key variables that can generally be used in estimating the household consumption models (i.e., Eq. (1)). Almost all these variables in the nine countries considered have sufficient observations to be used in modeling except for a few countries where these variables are understandably under-covered or non-existent (e.g., occupation in DRC or ethnicity in Rwanda). Table 7 also shows the latest available refugee survey for each country, which collects information on case size and socio-economic characteristics of the PA, in addition to other characteristics. For all these countries, the latest surveys covering refugees are quite recent, ranging from 2017 to 2020. Since the proGres data are administrative data and are updated for all these countries on a continuous basis, our proposed imputation method may be applied in all the listed countries to fairly recent data. In fact, a first experiment in that direction has been implemented for Chad with rather encouraging results (Beltramo et al. 2021).

5 Conclusion

We provide a first application of survey imputation methods to obtain poverty estimates for the Syrian refugees living in Jordan. Our results show that imputation-based poverty estimates are statistically not different from the non-predicted consumption-based poverty rates, and this result is robust to various validation tests. These estimates are found to perform better or have smaller standard errors than other poverty measures based on asset indexes or proxy means testing, and our

Table 6 Coverage and leakage rates based on imputation, ProGres and Home Visit Data

	Spec. 1	Spec. 2	Spec. 3
Coverage rate	70.0	71.3	73.5
Leakage rate	32.4	30.9	29.4
Control variables			
Demographics and employment	Y	Y	Y
Household assets and house characteristics	N	Y	Y
Shock-coping strategies and receiving UNHCR assistance	N	N	Y
R^2	0.44	0.48	0.54
N	18,992	18,992	18,992

The full regression results are provided in Appendix Table 1, Appendix 2. Specification 1 employs variables from the ProGres database only, and specifications 2 and 3 employ variables from both the ProGres and HV databases, using the empirical errors model. The estimation sample is generated by splitting the data into two random samples named sample 1 and sample 2. The imputed targeting rates are obtained using the empirical errors on sample 2. Robust standard errors in parentheses are clustered at the district level. We use 1000 simulations for each model run

imputation models are rather parsimonious and use variables that are already available in the UNHCR's global registration system. These encouraging results are consistent with the findings in recent studies for imputation-based poverty estimates for regular populations.

The estimation results also point to the need for further research on an alternative and promising method of obtaining poverty estimates for refugees where it is expensive or logistically challenging to implement a large-scale survey. We provide both theoretical and empirical evidence for Jordan that relatively small surveys may be fielded for refugees, and data from this survey can be combined with those from the census-type registration system to provide cost-effective and updated estimates of poverty. While these results are encouraging, they are not definitive and should be replicated in other contexts, possibly using surveys that have a more detailed consumption module. If further validated in other contexts, including some sub-Saharan countries with available and similar ProGes data on refugees, these findings can potentially lead to significant reductions in data collection costs in the context of refugee operations.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00148-022-00909-x>.

Acknowledgements We would like to thank the editor Klaus F. Zimmermann, three anonymous reviewers, Theresa Beltramo, Jose Cuesta, Talip Kilic, Peter Lanjouw, Christoph Lakner, David Newhouse, Franco Peracchi, Shinya Takamatsu, Matthew Wai-Poi, Tara Wishvanath, and participants at the IARIW-WB conference on poverty measurement and the WB-UNHCR training course on poverty imputation for helpful comments and discussions on earlier versions. This work is part of the program "Building the Evidence on Protracted Forced Displacement: A Multi-Stakeholder Partnership."

Funding The program is funded by UK aid from the United Kingdom's Department for International Development (DFID); it is managed by the World Bank Group (WBG) and was established in partnership

Table 7 Availability of ProGres and Home Visit Data in other countries

ProGres variables	Cameroon	Chad	Congo, Republic	Democratic Republic of the Congo	Ethiopia	Kenya	Niger	Rwanda	United Republic of Tanzania
Total frequencies	319,704	389,690	60,382	319,313	64,079	488,274	143,412	163,350	411,583
Case size	319,704	389,690	60,382	319,313	64,079	488,274	143,412	163,350	411,583
PA age	319,428	389,487	60,158	319,103	63,979	487,928	143,309	163,234	411,547
PA sex	319,703	389,689	60,381	319,311	64,079	488,271	143,412	163,350	411,583
PA highest education (years)	258,757	316,042	55,066	31,233	64,033	407,052	8,416	71,170	69,041
PA marital status	318,662	388,325	59,417	318,209	63,632	486,701	143,068	162,858	411,140
PA occupation/qualification	234,528	311,343	47,809	6,034	63,412	98,089	13,057	62,437	26,096
PA ethnicity	312,662	382,097	35,568	244,524	63,523	470,434	139,715	0	355,217
PA (religion)	308,180	383,290	46,497	280,132	62,295	480,961	143,073	145,722	351,817
% of total frequencies									
Case size	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
PA age	99.9	99.9	99.6	99.9	99.8	99.9	99.9	99.9	100.0
PA sex	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
PA highest education (years)	80.9	81.1	91.2	9.8	99.9	83.4	5.9	43.6	16.8
PA marital status	99.7	99.6	98.4	99.7	99.3	99.7	99.8	99.7	99.9
PA occupation/qualification	73.4	79.9	79.2	1.9	99.0	20.1	9.1	38.2	6.3
PA ethnicity	97.8	98.1	58.9	76.6	99.1	96.3	97.4	0.0	86.3
PA (religion)	96.4	98.4	77.0	87.7	97.2	98.5	99.8	89.2	85.5
Latest available survey year	2020	2018	2017	2019	2017	2020	2020	2019	2019
Data repository	UNHCR	WB-UNHCR	UNHCR	UNHCR	WB	UNHCR	UNHCR	UNHCR	UNHCR

PA principal applicant

with the United Nations High Commissioner for Refugees (UNHCR). The scope of the program is to expand the global knowledge on forced displacement by funding quality research and disseminating results for the use of practitioners and policy makers. Additional funding support is provided by DFID through its Knowledge for Change (KCP) program. This work does not necessarily reflect the views of DFID, the WBG, or UNHCR.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alix-Garcia J, Walker S, Bartlett A (2019) Assessing the direct and spillover effects of shocks to refugee remittances. *World Dev* 121:63–74
- Alloush M, Taylor JE, Gupta A, Valdes RIR, Gonzalez-Estrada E (2017) economic life in refugee camps. *World Dev* 95:334–347
- Beegle K, Christiaensen L, Dabalen A, Gaddis I (2016) Poverty in a rising Africa. The World Bank, Washington, DC
- Beltramo T, Wieser C, Gliigliarano C, Heyn R (2019) Identifying poor refugees for targeting of food and multi-sectoral cash transfers in Niger. *mimeo*.
- Beltramo T, Dang H-A, Sarr I, Verme P (2021) Estimating poverty among refugee populations: a cross-survey imputation exercise for Chad. *IZA Discussion Paper No.* 14606
- Brown C, Ravallion M, van de Walle D (2018) A poor means test? Econometric targeting in Africa. *J Dev Econ* 134:109–124
- Carpenter J, Kenward M (2013) Multiple imputation and its application. John Wiley & Sons, Chichester
- Christiaensen L, Ligon E, Sohnesen TP (2021) Consumption subaggregates should not be used to measure poverty. *World Bank Econ Rev*. <https://doi.org/10.1093/wber/lhab021>
- Christiaensen L, Lanjouw P, Luoto J, Stifel D (2012) Small area estimation-based prediction models to track poverty: validation and applications. *J Econ Inequal* 10(2):267–297
- Coady D, Grosh M, Hoddinott J (2014) Targeting outcomes redux. *World Bank Res Obs* 19:61–85
- Cohen J (1988) *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale, NJ, Erlbaum
- Cuesta J, Ibarra GL (2017) Comparing cross-survey micro imputation and macro projection techniques: poverty in post revolution Tunisia. *J Income Distrib* 25(1):1–30
- Dang H-A, Lanjouw P (2018) Poverty and vulnerability dynamics for India during 2004–2012: insights from longitudinal analysis using synthetic panel data. *Econ Dev Cult Change* 67(1):131–170
- Dang H-A, Lanjouw P, Serajuddin U (2017) Updating poverty estimates at frequent intervals in the absence of consumption data: methods and illustration with reference to a middle-income country. *Oxf Econ Pap* 69(4):939–962
- Dang H-A, Jolliffe D, Carletto C (2019) Data gaps, data incomparability, and data imputation: a review of poverty measurement methods for data-scarce environments. *J Econ Surveys* 33(3):757–797
- Dang H-A, Lanjouw P, Luoto J, McKenzie D (2014) Using repeated cross-sections to explore movements in and out of poverty. *J Dev Econ* 107:112–128

- Dang H-A, Kilic T, Carletto C, Abanokova K (2021) Poverty imputation in contexts without consumption data: a revisit with further refinements. *World Bank Policy Research Paper No. 9838*. Washington, DC: The World Bank
- De Luca G, Magnus JR, Peracchi F (2018) Balanced variable addition in linear models. *J Econ Survey* 32(4):1183–1200
- Elbers C, Lanjouw JO, Lanjouw P (2003) Micro-level estimation of poverty and inequality. *Econometrica* 71(1):355–364
- Filmer D, Scott K (2012) Assessing asset indices. *Demography* 49(1):359–392
- Mathiassen A (2009) A model based approach for predicting annual poverty rates without expenditure data. *J Econ Inequal* 7(2):117–135
- Mathiassen A (2013) Testing prediction performance of poverty models: empirical evidence from Uganda. *Rev Income Wealth* 59(1):91–112
- Park CN, Dudycha AL (1974) A cross-validation approach to sample size determination for regression models. *J Am Stat Assoc* 69(345):214–218
- Pituch KA, Stevens JP (2016) *Applied multivariate statistics for the social sciences: analyses with SAS and IBM's SPSS*. Routledge, New York
- Ravallion M (1996) Issues in measuring and modelling poverty. *Econ J* 106(438):1328–1343
- Ravallion M (2016) *The economics of poverty: history, measurement, and policy*. Oxford University Press, New York
- Rencher AC (2002) *Methods of multivariate analysis*. John Wiley & Sons, USA
- Rubin DB (1987) *Multiple imputation for nonresponse in surveys*. Wiley, New York
- Tarozzi A (2007) Calculating comparable statistics from incomparable surveys, with an application to poverty in India. *J Business Econ Stat* 25(3):314–336
- Verme P, Schuettler K (2021) The impact of forced displacement on host communities: A review of the empirical literature in economics. *J Dev Econ* 150:102606
- Verme P, Gigliarano C, Wieser C, Hedlund K, Petzoldt M, Santacroce M (2016) *The welfare of Syrian refugees: evidence from Jordan and Lebanon*. World Bank, Washington, DC

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.