Sarr, Ibrahima; Dang, Hai-Anh; Gutierrez, Carlos Santiago Guzman; Beltramo, Theresa; Verme, Paolo

Working Paper

Using Cross-Survey Imputation to Estimate Poverty for Venezuelan Refugees in Colombia

IZA Discussion Papers, No. 17036

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Sarr, Ibrahima; Dang, Hai-Anh; Gutierrez, Carlos Santiago Guzman; Beltramo, Theresa; Verme, Paolo (2024) : Using Cross-Survey Imputation to Estimate Poverty for Venezuelan Refugees in Colombia, IZA Discussion Papers, No. 17036, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/299964

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17036

Using Cross-Survey Imputation to Estimate Poverty for Venezuelan Refugees in Colombia

Ibrahima Sarr
Hai-Anh H. Dang
Carlos Santiago Guzman Gutierrez
Theresa Beltramo
Paolo Verme

MAY 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17036

Using Cross-Survey Imputation to Estimate Poverty for Venezuelan Refugees in Colombia

Ibrahima Sarr
UNHCR

Hai-Anh H. Dang
World Bank, IZA, Indiana University and London School of Economics and Political Science

Carlos Santiago Guzman Gutierrez
Oxford University

Theresa Beltramo
School of Economics and Management, University of Geneva and UNHCR

Paolo Verme
World Bank

MAY 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Using Cross-Survey Imputation to Estimate Poverty for Venezuelan Refugees in Colombia*

Household consumption or income surveys do not typically cover refugee populations. In the rare cases where refuges are included, inconsistencies between different data sources could interfere with comparable poverty estimates. We test the performance of a recently developed cross-survey imputation method to estimate poverty for a sample of refugees in Colombia, combining household income surveys collected by the Government of Colombia and administrative data collected by the United Nations High Commissioner for Refugees. We find that certain variable transformation methods can help resolve these inconsistencies. Estimation results with our preferred variable standardization method are robust to different imputation methods, including the normal linear regression method, the empirical distribution of the errors method, and the probit and logit methods. We also employ several common machine learning techniques such as Random Forest, Lasso, Ridge, and elastic regressions for robustness checks, but these techniques generally perform worse than the imputation methods that we use. We also find that we can reasonably impute poverty rates using an older household income survey and a more recent ProGres dataset for most of the poverty lines. These results provide relevant inputs into designing better surveys and administrate datasets on refugees in various country settings.

JEL Classification: C15, F22, I32, O15, O20
Keywords: refugees, poverty, imputation, Colombia

Corresponding author:
Hai-Anh H. Dang
World Bank Development Data Group
Washington, DC
USA
E-mail: hdang@worldbank.org

* We would like to thank the UK Foreign Commonwealth and Development Office (FCDO) for funding assistance through the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Program.

1 Introduction

Forcibly displaced populations continue to rise—at the end of 2022 there were 108.2 million Forced Displaced Persons (FDPs) worldwide, of whom 35.3 million are refugees, 62.5 million are Internally Displaced Persons (IDPs), 5.4 million are asylum seekers, and 5.2 million are other people in need of international protection (UNHCR, 2023a). FDPs represent vulnerable population groups that can be even more vulnerable during times of crisis.[1] To design effective assistance programs in response to this urgent humanitarian need, better and more updated poverty estimates for FDPs are indispensable inputs.

Yet, computing poverty figures usually requires high-quality and frequently updated consumption (or income) surveys, which remain scarce in forced displacement settings. The latest UN global report includes only eight countries with comparable, "gold standard" data for refugees. Based on this limited available data, refugees consistently experience higher poverty rates than surrounding nationals (UNHCR, 2023a). This is not surprising as many refugees have specific vulnerabilities that distinguish them from other populations experiencing poverty. They have lost assets and experienced trauma and have limited rights and access to opportunities compared to the host community, and face short-term planning horizons.[2] This data scarcity challenge is due to various reasons including the lack of political will to include forcibly displaced into national statistics exercises, insufficient funding resources, as well as the typically remote and challenging-to-reach locations of FDPs. Furthermore, unique measurement challenges also exist with

[1] Recent phone survey data also show that the COVID-19 pandemic had disproportionately large impacts on refugees compared to hosts along various dimensions of health, education, wages and employment, and food security (JDC, 2021).

[2] In particular, Pape and Verme (2023) observe that due to limitations in humanitarian assistance and government policy, only the refugees who seek assistance may register. This may create a self-selection problem, which likely results in upward biased poverty estimates for the general refugee population if we only use the data from those who registered.

estimating poverty for FDPs. For example, correctly valuing in-kind distributions (e.g., shelters and in-kind food for consumption) and effectively adapting the poverty line to the refugee setting represent challenging tasks.

Against this background, imputation methods have been widely employed in economics to fill data gaps. These survey-to-survey (or cross-survey) methods essentially rely on an existing older consumption survey to build an imputation model using appropriate predictor variables. This imputation model is subsequently employed in combination with the same variables in another (recent or larger) survey that does not collect consumption data to provide poverty estimates for the latter survey. Building on the seminal technique that imputes from a household consumption survey into a census to generate poverty maps (Elbers *et al.,* 2003), cross-survey imputation methods have been used to estimate poverty trends and poverty dynamics in the absence of repeated cross-sectional data or panel data (Dang and Lanjouw, 2023).

In refugee contexts, cross-survey imputation is a useful tool to address the data challenge of missing household consumption surveys. But just a handful of studies current exist that employ imputation. Dang and Verme (2023) find that imputation methods can be employed to estimate poverty and improve targeting efficiency for Syrian refugees living in Jordan, while Beltramo *et al.* (2024) obtain a similar finding for refugees living in Chad. Both these studies impute from a household consumption survey into UNHCR non-consumption, administrative data to obtain poverty estimates for a larger sample of refugees. Employing a machine learning approach for imputation, Altindag *et al.* (2021) also find that administrative data curated by humanitarian organizations can be used to estimate refugee household welfare accurately for targeting purposes with Syrian refugees living in Lebanon.

We make several new contributions to the literature in this paper. We demonstrated that poverty imputation can work for a different refugee population in a different country setting, the Venezuelan refugees in Colombia. Colombia represents an interesting context. Globally, refugees and migrants from Venezuela exceed 6 million, of whom 83% (4.99 million) are hosted in countries in Latin America and the Caribbean. Colombia is by far the largest host country for Venezuelan refugees and migrants, hosting approximately 40% (i.e., 2.5 million persons) of Venezuelans forced to flee (UNHCR, 2023c).[3] Not only do we perform the first poverty imputation for Venezuelan refugees in Colombia, we also extend the nascent literature on measuring poverty for refugees in different respects.

First, we investigate a key assumption behind poverty imputation that has received little attention: the comparability of the variable predictors in the dataset to impute from (the base survey) and in the dataset to impute into (the target survey/ census). While the assumption that these variables should be representative of the same population is the prerequisite for accurate imputation results, it can be violated more often than one might think. For example, inconsistencies between the same surveys over time, or across surveys of different design (e.g., due to different survey questionnaires, or changes to questionnaires to collect better data or updated data on changes in consumption patterns over time) are observed for both poorer and richer countries (Deaton and Kozel, 2005; Moffitt *et al.*, 2022).[4] The violation of this assumption requires harmonization of the two datasets to ensure that their variables have similar distributions. We compare several data adjustment approaches that are employed for this purpose, including raking, matching and variable standardization methods.

[3] The two next largest host countries for Venezuelan refugees and migrants are Peru (976,400), and Ecuador (555,400), which respectively host just about one-half and one-fifth of the refugee populations in Colombia.
[4] Dang and Lanjouw (2023) offer a recent review of studies that investigate survey inconsistencies. We return to more discussion in Section 4.

Second, we impute poverty over time using an older income survey and a newer round of the UNHCR administrative database. While this is often done with imputation studies for the general population, to our knowledge, this has never been done before in a refugee context. Our study thus offers a critical improvement in the literature with practical policy relevance. We estimate the efficiency of imputing poverty with an older, but (far more) expensive base income survey in combination with more recent, but (much less expensive) UNHCR administrative data. Given UNHCR maintains and updates administrative records in most countries, poverty imputation can help justify the large costs of an initial benchmark income survey, which can be subsequently re-utilized with more updated administrative data to produce poverty estimates in a cost-saving manner.

Finally, unlike the few existing studies on refugees discussed above (Altındağ *et al.*, 2021; Dang and Verme, 2023; Beltramo *et al.*, 2024) that work with missing consumption data, we imputes for missing income data. Income imputation offers its own challenges compared with consumption imputation in most contexts (e.g., households can smooth consumption rather than income, so households can consume in time periods when they may have zero income (Deaton, 1997)). These challenges may be amplified especially in refugee settings where refugees are more likely to have zero income.

We assess several variable transformation methods that correct for violation of the comparability assumption and find that some methods, such as the variable standardization method and the raking method, generally perform better than the others. Estimation results with the variable standardization method are robust to different imputation methods, including the normal linear regression method, the empirical distribution of the errors method, and the probit and logit methods. We also employ several common machine learning techniques such as Random Forest,

Lasso, Ridge, and elastic regressions for robustness checks, but these techniques generally perform worse than the imputation methods that we use. We also find that we can reasonably impute poverty rates using an older household income survey and a more recent ProGres dataset for most of the poverty lines.

This paper consists of six sections. We describe the country context in the next section before presenting the data in Section 3. We subsequently discuss the analytical framework in the Section 4, including the imputation method (Section 4.1) and the data adjustment methods (Section 4.2). We discuss the estimation results in Section 5, including imputation for the same point in time using the data in 2019 and 2022 (Sections 5.1 and 5.2) and heterogeneity analysis (Section 5.2). We offer further extension with imputation over time in Section 6. We finally conclude in Section 7.

2 Country context

Colombia hosts 40% (2.5 million) of Venezuelan refugees and migrants forced to flee within Latin America, being the third-largest refugee-hosting country globally (UNHCR, 2023b). Data end 2022 highlights that only 10% of Venezuelan refugees hold a regular status (i.e., having received the (residence) permit); 23% are in the process of getting the permit; 56% have been authorized to receive the permit, and approximately 11% still hold an irregular status (i.e., don't have access to the permit and, therefore, lack access to basic rights). From this universe, close to 15% have sought assistance and support from UNHCR, meaning that they have been registered in ProGres data system. Table A.2 provides the number of forcibly displaced persons (FDPs) in the country.

Despite the implementation of the Temporary Protection Statute (TPS) in 2021, a novel policy tool that granted legal residency for 10 years to Venezuelan refugees and migrants who met certain criteria,[5] refugees and migrants from Venezuela continue to face challenging living conditions and protection risks. A recent report analyzing the socio-economic impact of Venezuelan refugees and migrants in Chile, Colombia, Ecuador, and Peru finds that Venezuelans are better educated and, with the exception of Colombia, more likely to be employed than the host population. For Colombia the study finds some 63 percent of Venezuelans are employed. However, they are often employed in lower-quality jobs than the jobs they had in Venezuela, likely due to a mismatch in skills accreditation- their academic titles are not recognized in Colombia- and they are paid lower wages than nationals for similar jobs (Mejía-Mantilla et al, 2024). The Government of Colombia (GoC) has taken the relatively novel step of including Venezuelan refugees and migrants into the national poverty measures. Based on the estimates from the GEIH 2022 survey, Venezuelan refugees and migrants face higher poverty rates than Colombians: the headcount (income-based) poverty rate is 63% for Venezuelan refugees compared to 39% for Colombian nationals. The relative multidimensional poverty rate is 42% and 13% respectively for the two population groups. Venezuelan refugees also face more barriers when accessing the labor market, resulting in higher participation in informal jobs and unemployment.

Although the TPS is implemented nationwide, the movement patterns of Venezuelan refugees and migrants vary depending on whether they are in border areas. Colombia is a significant passageway as there are seven official entry points along the shared 2,200-kilometer border between the two countries (Chaves-González *et al.*, 2021). Along these borders, three displacement profiles characterize the movements: (i) "In transit", referring to those who aim to

[5] Two conditions were needed to be eligible for the permit: i. to have entered the country before January 2021- either through an irregular or regular border, or ii. To have entered through a regular border before May 2023.

reach another country outside Colombia; (ii) “Pendular”, corresponding to Venezuelans who repeatedly travel to Colombia for less than a month and then return; and (iii) “Intention to stay”, referring to Venezuelans who want to settle in Colombia (R4V, 2023). The primary departments through which Venezuelans enter Colombia are Norte de Santander, La Guajira, and Arauca (R4V, 2023), after which most Venezuelan refugees who intend to stay in Colombia tend to settle in major urban centers like Bogota, Medellin, Cucuta, and Barranquilla where employment opportunities are better (Migration Colombia, 2023). Other border departments like Antioquia and Nariño concentrate the outflow of refugees and migrants towards the United States through the Darién Gap, and South America.

Furthermore, the living conditions of Venezuelan refugees and migrants who settle in border departments are generally more difficult compared to those who settle in the central departments. According to the Unsatisfied Basic Needs (UBN) indicator, Venezuelans in departments such as Norte de Santander, La Guajira, and Arauca have UBN levels above 20%. In contrast, departments like Antioquia and the capital, Bogota, have UBN levels below 10% (DANE, 2021). Moreover, they face greater food insecurity, have limited access to nutritional interventions, and fewer Venezuelan children residing in border areas could enrol in the education system. They also experience more water insufficiency and protection risks on average (R4V, 2023). In general, the inner departments and urban capital hubs like Bogota, Medellin, and Barranquilla provide more opportunities for refugees and migrants.

3 Data

We analyze two data sources, with two datasets (survey/ census rounds) for each in 2019 and 2022: (i) the UNHCR’s Profile Global Registration System (ProGres) database (i.e., proGres19

and proGres22 representing the ProGres database for 2019 and 2022), ii) and Colombia's Great Integrated Household Survey (GEIH) (GEIH19 and GEIH22 representing GEIH dataset for 2019 and 2022). Table A.1 provides an overview of these surveys. We use two UNHCR administrative years (2019 and 2022) to compare the performances of cross-survey methods for two separate years separated by the COVID pandemic health and economic shock. Moreover, using the two different years also allows us to assess whether we can reliably update poverty estimates based on data collected a few years ago.

As part of its mandate to protect displaced persons in host countries, UNHCR collects data to monitor the welfare of refugees and other populations of interest and to deliver assistance and services. The ProGres database is the case management tool developed by UNHCR to facilitate the protection of the people that UNHCR serves. It compiles individual and group-level data of refugees, asylum-seekers, IDPs, returnees, and stateless populations and is used for a number of different operational tasks including refugee status determination and targeting of assistance. ProGres also serves as a starting database from which additional information is collected for resettlement or voluntary repatriation where appropriate.

In Colombia, only around 15% of refugees and migrants, IDPs, and returnees are registered in the UNHCR proGres system. While this number appears low, this is a common issue in many refugee settings. Wherever possible UNHCR aims to rely on national systems to limit duplication and prioritize scarce resources appropriately. This is the case in Colombia where the GOC has its own registration system for (i) armed conflict victims (internally displaced persons constitute the main percentage of registered victims) with the Victims Registry (further explored in section 3), and (ii) refugees and migrants from Venezuela with the Single Registry of Venezuelan Migrants

(RUMV by its Spanish acronym).[6] Frequently many upper and middle income countries chose to rely on existing national systems for registration and managing administrative data on refugees and asylum seekers. In the case of Colombia, the proGres database was operationalized primarily to administer assistance to the most vulnerable by UNHCR. As is the case in many welfare programs in developed countries, the identification of those in need for services frequently begins with individuals self-identifying themselves as in need of assistance. The second step generally taken is welfare agencies validate the need of individuals using available public records (e.g. tax records) and/or home visits. In Colombia, a refugee has been entered into the proGres database for two main reasons. First, s/he has requested assistance near the border for temporary shelter in one of UNHCR and GoC sponsored reception centers. Second, the individual has requested assistance from UNHCR via a visit to one of UNHCR's field or branch offices. The accommodation in the reception centers is quite basic with shared facilities and rooms, as such those who could afford not to rely on UNHCR services generally never stop at the Reception Centers requesting temporary accommodation. And those who requested additional assistance from UNHCR field or branch offices, their requests were then validated by UNHCR through personal interviews. As a result of how the UNHCR proGres registry is formed, we believe that the proGres dataset represents the relatively worse off portion of Venezuelans arriving in Colombia.

The UNHCR administrative data collects some limited sociodemographic characteristics including sex, date of birth, date of arrival in the country, legal status, and country of nationality, household size, and education attainment. However, it does not include any welfare indicator, such as income or consumption.

[6] For more information, see Guzmán Gutierrez (2023) for a mapping of various data sources in Colombia.

The second set of data used in this study is the Great Integrated Household Survey (GEIH). In an effort to track the living conditions of Venezuelan refugees and migrants and identify policy priorities, the GoC has included displaced Venezuelans in national household surveys. There are two key household surveys implemented by the Colombia National Department of Statistics (DANE by its Spanish acronym). One of them is the Great Integrated Household Survey (GEIH), which is implemented monthly and provides information regarding labor market outcomes. The sample consists of both Colombians and Venezuelans.

The GEIH is a nationally representative sample covering both nationals and non-nationals (refugees and non-refugees). Such dataset is restricted to Venezuelan refugees for our analysis. The GEIH provides representative information at the national, urban-rural, regional, and departmental levels, as well as for the capitals of each department. The main objective is to provide basic information about the size and structure of the workforce (employment, unemployment, inactivity), as well as a description of sociodemographic characteristics of the population. As such, the survey collects information regarding labor market outcomes (employment, income, social protection) and general characteristics of the population, such as sex, age, marital status, educational level, sources of income, and expenses (what they buy, how often they buy, and where they buy).

The average monthly sample of the GEIH data comprises approximately 21,000 households. The target population consists of individuals living in the national territory. It excludes the most rural areas in which approximately 1% of the population live. The sample is selected following a stratified probabilistic sampling strategy. The country is classified into two sampling frames, with one corresponding to 24 cities with their metropolitan areas and the other corresponding to the rest of the country. Each capital or metropolitan area is self-represented and has a selection probability

of 1. The GEIH collects data on housing, socio-demographic characteristics, and income, among other variables.

Colombia's official poverty numbers are based on income, published once yearly, and have been reported since 2002 except for 2006 and 2007. In line with this practice, we use household income to measure poverty among Venezuelan refugees in Colombia.

It is useful to highlight some key differences between the two datasets. While the GEIH sample covers both economic migrants and FDPs, the UNHCR ProGres database only covers refugees and Other people in need of international protection under UNHCR mandate. Consequently, ProGres is more reliable in terms of statistical representativeness of FDPs (although it is more likely to cover FDPs most in need of assistance as discussed above). Moreover, the ProGres has better representativeness of the refugee population at a lower administrative level.

Table 1 provides the summary statistics for the two datasets in 2019 (Panel A) and 2022 (Panel B). Overall, the means of the variables are generally different. In particular, compared with the GEIH in 2022, the Venezuelan refugees in the ProGres database are more likely to be female, younger, have less tertiary education (but more tertiary education in 2019), and live in households with a smaller size.

4 Analytical framework

4.1 Imputation method

The methodology used in this paper relies on the cross-survey imputation framework that was first introduced by Elbers *et al.* (2003) to generate poverty maps.[7] Most recently, Dang *et al.* (2017)

[7] See also Tarozzi (2007) and Mathiassen (2009) for further improvements and adaptation of this approach (e.g., by estimating the standard errors in a different way). (e.g., Douidich *et al.,* (2016) offer an early study that imputes across types of surveys such as consumption and labor force surveys. Dang *et al.* (2019) and Dang and Lanjouw (2023) offers

built on this literature to propose a model that imposes fewer restrictive assumptions and offers an explicit formula for estimating the poverty rate and its variance. Three new contributions introduced by this study are: (i) it offers a simple variance formula, which is in line with the recent statistical literature; (ii) it can accommodate complex design sampling; and (iii) the framework remains applicable to two surveys with different designs (such as imputing from a household consumption survey into a labor force survey). Finally, the approach allows for different modeling methods, including the standard linear regression model, its variant with a flexible specification of the empirical distribution of error terms, a logit model, and/or a probit model. This method has been validated and applied to data from poor and middle-income countries in different regions ranging from India, Jordan, Tunisia, and Sub-Saharan African countries to Vietnam (Beegle *et al.*, 2016; Cuesta and Ibarra, 2018; Dang and Lanjouw, 2023). Recent applications of this method include providing poverty estimates for refugees (Dang and Verme, 2023; Beltramo *et al.*, 2024).

Let x_j be a vector of characteristics that are commonly observed between two surveys, where j indicates survey type, with 1 and 2 being respectively the base survey (survey with welfare indicator) and the target survey (survey without welfare). We assume that the welfare indicator is a function of household and individual characteristics (x_j):

$$y_j = \beta_j' x_j + \upsilon_{cj} + \varepsilon_j \qquad (1)$$

where y_j is the welfare indicator which is in the framework of this study income per capita per month, β_j is a vector of parameters, υ_{cj} is cluster i random effects, and ε_j is the idiosyncratic error term. We suppress the index for households (and individuals) to make notation less cluttered.

recent reviews of previous imputation studies that discuss the main advantages and different approaches of welfare imputation practices as well as provide useful insights into the imputation process. See also Little and Rubin (2019) for a recent review on related topics in the statistics literature.

This imputation framework is based on two assumptions. Assumption 1, which is critical for poverty imputation, states that measurement of household characteristics in each sample of data is a consistent measure of the characteristics of the whole population. In other words, it stipulates that the surveys considered are representative of the same target population. The second assumption states that changes in x_j between the data collection periods of the two data sets can capture the change in welfare over the period (Assumption 2).

Under these two assumptions, the imputed welfare is

$$y_2^1 = \beta_1' x_2 + v_{c1} + \varepsilon_1 \qquad (2)$$

where y_2^1 represent the imputed welfare when we apply the estimated parameters (β_1') and the estimated distributions of the error terms (υ_{c1} and ε_1) from the base survey to the variables (x_2) in the target survey.

Since Equation (1) is typically estimated with the standard cluster-effects linear regression model, Dang *et al.* (2017) propose different imputation methods for parameter estimation. The first method relies on the assumption of the normal distribution for the two error terms (μ_{cj} and ε_j are uncorrelated and $\upsilon_{cj}|x_j \sim \mathcal{N}(0, \sigma_{\mu_{cj}})$ and $\varepsilon_j|x_j \sim \mathcal{N}(0, \sigma_{\varepsilon_j})$). Hereafter, this method is referred to as the normal linear regression model. An alternative method proposed is the empirical error method, which assumes no functional form for these error terms and uses instead the empirical distribution to estimate the parameters.

Since the estimated parameters are obtained using a different survey from the target survey, we can use simulation to estimate Equation (2) (for a single draw) as follows:

$$\hat{y}_{2,s}^1 = \hat{\beta}_1' x_2 + \tilde{\upsilon}_{c1,s} + \tilde{\tilde{\varepsilon}}_{1,s}. \qquad (3)$$

In Equation (3), $\tilde{\tilde{\beta}}_{1,s}'$, $\tilde{\tilde{\upsilon}}_{c1,s}$, and $\tilde{\tilde{\varepsilon}}_{1,s}$ represent the s^{th} random draw (simulation) from their estimated distributions using the base survey, for s= 1,…, S.

The imputed poverty rate ($\widehat{P}_2$) and its variance ($V(\widehat{P}_2)$) in the target survey are then estimated as:

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

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

where *P(.)* is the probability function that estimates the poverty rate in the population for each simulation. In Equation (5), $\hat{P}_{2,s}$ is similarly defined as follows $\hat{P}_{2,s} = P(\hat{y}_{2,s}^1 \leq z_1)$. To make notation simpler, we do not show the equations with sampling weights. Formulae with weights are shown in Dang *et al.* (2017).

These poverty estimators provide consistent estimates of the parameters. Furthermore, in terms of prediction accuracy, evidence suggests that these estimators outperform the traditional proxy means testing technique, which typically omits the error terms $\upsilon_{c1} + \varepsilon_1$ and results in biased estimates of the welfare indicator (Dang and Lanjouw, 2023). To provide further robustness check, we also employ two alternative modelling methods—the probit model and the logit model. These models place more restrictive assumptions on the error term but estimate poverty figures directly (i.e., Equation (4) and Equation (5)) instead of estimating income first and subsequently obtain poverty estimates using the predicted income.

To check robustness, we also use Machine Learning techniques. More precisely, we use the Random Forest technique, which is a combination of a series of tree structure classifiers and has many good characters, and we also use the Lasso regression.[8] Alternative models which adjust the penalty parameter to estimate ridge and elastic net regressions are also used (Melkumova & Shatskikh, 2017; Tay *et al.*, 2023).

[8] See Mullainathan and Spiess (2017) and Athey and Imbens (2019) for recent reviews of these techniques in economics studies.

Since the ProGres database is limited in terms of socioeconomic and demographic variables, to evaluate the performance of the welfare estimation model, we consider three models that add variables on a cumulative basis. Model 1 includes the household size and the gender of the head of household. Model 2 adds to Model 1 the age of the head of household and its squared. Model 3 adds to Model 2 variables educational attainment of the head. Consequently, Model 2 is richer than Model 1 and Model 3 is richer than Model 2, but they are also more demanding in terms of the control variables. Table A.1 (Appendix A) describes the variables for the two datasets.

4.2 Data adjustment methods

In our context, given the significant difference in the means shown in Table 1 discussed above, Assumption 1 that the ProGres and the GEIH both provide similar estimates for the same population is likely violated. Indeed, formal statistic tests confirm that the variables in the two datasets have different distributions, and these differences are strongly statistically significant (Table 3, column 1 for data without adjustment). In fact, survey design issues that compromise the comparability of poverty estimates are found in various countries including China (Gibson *et al.*, 2003), India (Deaton and Kozel, 2005), Tanzania (Beegle *et al.*, 2012), and Vietnam (World Bank, 2012). Even for richer countries like the U.S., inconsistencies between different surveys are well documented in the literature. For example, Abraham *et al.* (2013) study the differences between employment data between the Current Population Surveys and employer-reported administrative data. Bavier (2014) finds spending and income poverty in the Consumer Expenditure Survey to be an outlier compared with those in other surveys including the Panel Study of Income Dynamics. Moffitt *et al.* (2022) document differences with male earnings volatility between major surveys and administrative data.

We discuss next three techniques that can help remedy the violation of Assumption 1: i) matching method, ii) raking method, iii) variable standardization method. For completeness, we next start with the case without any data adjustment.

No adjustment

With no change, we analyze the datasets as is. By doing so, we accept the fact that most of the variables are different across the two datasets as the means of the variables in ProGres are mostly statistically different from those in the GEIH (Table 3, Column 1). This "naïve" implementation of poverty imputation clearly violates Assumption 1 and typically leads to biased estimates of the poverty rate. It is not straightforward to sign the direction of the bias, which would depend on the complex dynamics of the relationships between household consumption and the control variables as well as the magnitudes of the differences for the same variables across the two datasets.

Raking method

A technique to improve the relation between a sample survey and the population is to adjust the sampling weights of the cases in the sample so that the marginal totals of the adjusted weights on specified characteristics agree with the corresponding totals for the population. This operation is known as raking or sample-balancing, which is a model-based approach using known population totals (usually from a census) that adjusts the sampling weights so the marginal values of a table sum to those known totals (Deville *et al.*, 1993; Anderson & Fricker, 2015). In our case, given the biased sample in the ProGres, we will need to adjust the proGres dataset to make it consistent with the reference GEIH.

Raking assigns a weight value to each survey respondent such that the weighted distribution of the sample is in very close agreement with two or more marginal control variables. For example, in household surveys, the control variables are typically sociodemographic variables. Raking is an

iterative process that uses the sample design weight as the starting weight and terminates when the convergence criterion is achieved. However, the resulting final weight may exhibit considerable variability, with some sampling units having extremely low or high weights relative to most of the other sampling units. This leads to inflated sampling variances of the survey estimates.

We implement the following steps: i) in the target (ProGres) data, we create one matrix that stores the summary statistics of target variables; ii) in the reference (GEIH) survey, we match the target variables with the summary statistics matrix in the reference survey.

More formally, the following optimization problem is solved with a Lagrangian function:

$$\begin{cases} max_W \mathrm{H(W)} = -\sum_{i=1}^{n} w_i * \ln(w_i) \\ such\ \ that\ y_j = \sum_{i=1}^{n} X_{ji} w_i \quad j = 1, \dots, J \\ Weight\_Sum = \sum_{i=1}^{n} w_i \end{cases} \tag{6}$$

where w_i is the new weights to assign as a result of the optimization problem solving; the J constraints given in (6) can be thought of as moment constraints, with y_j being the population mean of the control variables X_j; $Weight\_Sum$ is the sum of the initial weights of the survey data.

Matching method

Suppose two sample datasets, File A (ProGres) and File B (GEIH), are taken from two different surveys. Let File A contain vector-valued variables (X, Y) while File B has vector-valued variables (X, Z). Statistical matching aims to combine these two files to obtain at least one synthetic file containing (X, Y, Z).

Unlike record linkage or exact matching, the two files to be combined are not assumed to have records for the same entities. In statistical matching, the files are considered to have little or no overlap; hence, records for similar, rather than the same, entities are combined (Phua *et al.*, 2006;

Raffo & Lhuillery, 2009). For example, one may match similar individuals on characteristics like gender, age, place of residence, country of origin, and so on.

Regarding matching methods, there are two categories "constrained" and "unconstrained." While constrained statistical matching requires using all records in the two files, unconstrained matching does not. Usually, in an unconstrained match, all the documents from one of the files (say File B) would be used (matched) to "similar" records on the second file (File A). Some records on the second file (File A) may be employed more than once or not at all. More precisely, we need to match observations of the ProGres dataset to similar enough observations of the GEIH dataset.[9]

Variable standardization method

To address the violation of Assumption 1 for surveys of different design, we can standardize the distributions of the ProGres variables by those of the GEIH by following the procedures in Dang *et al.* (2017). Assume further that the overlapping variables between the two surveys follow a normal distribution such that $x_{1t} \sim N(\mu_{1t}, {\sigma_{1t}}^2)$ and $x_{2t} \sim N(\mu_{2t}, {\sigma_{2t}}^2)$, for t= 1,..,T. As such, we standardize the variables as follows:

$$x_{2\to 1,t} = (x_{2t} - \mu_{2t}) * \frac{\sigma_{1t}}{\sigma_{2t}} + \mu_{1t}$$

where x_{jt} respectively represents the observed values for the variables in GEIH (j= 1) and ProGres (j= 2). μ_{jt} and σ_{jt} are respectively the mean and the standard deviation.

5 Estimation results

[9] Matches two columns or two datasets based on similar text patterns. String matching method used is the bigram. For example, 'peter' contains the bigrams 'pe', 'et', 'te' and 'er'. A q-gram similarity measure between two strings is calculated by counting the number of bigrams in common (i.e. bigrams contained in both strings) and divide by either the number of bigrams in the shorter string (called Overlap coefficient2), the number in the longer string (called Jaccard similarity) or the average number of bigrams in both strings (called the Dice coefficient).

5.1 Imputing poverty based on 2019 proGres using the model estimated from 2019 GEIH

Testing model assumption

As a first step, we check whether our data sets are representative of the same underlying population (Assumption 1) by performing means difference tests across critical predictors. As mentioned above, we considered the raw dataset and three alternative methods to make datasets more representative of the same population. We perform a simple t-test to test assumption 1 and show the results in Table 2. They indicate that all the variables, except for the education variables, are significantly different across the two datasets with no adjustment. Unlike the matching method, the raking method and the variable standardization method allow us to correct the violation of Assumption 1 by transforming the distributions of the ProGres variables to make them similar to those of the GEIH variables.

Imputation results

Using the dataset for 2019, Figure 1 provides the estimation results when we simulate the poverty line such that it runs between the 3rd and 99th percentiles of the consumption distribution. Panels A, B, C, and D offer estimation results using the normal linear model with no adjustment for the data (i.e., using the raw data), the matching method, the raking method, and the variable standardization method, respectively. Regardless of the model (Model 1, 2, 3), poverty prediction based on the raw dataset and the matching method, which violates Assumption 1, is only reliable for lower poverty lines. For the raking and the variable standardization methods that result in transformed datasets satisfying Assumption 1, the imputed poverty rates fall within the 95% confidence intervals (CIs) for almost all the different poverty lines and for all the three estimation

models. The imputation estimates based on the variable standardization method, however, are somewhat more accurate.

These results remain similar when we employ the empirical distribution of the errors model (Appendix A, Figure A.1) as well as the probit and logit models (Appendix A, Figure A.2 and Figure A.3). We also employ machine learning (ML) techniques such as Random Forest, Lasso, Ridge, and elastic regressions for robustness checks, but the ML results are generally outside the 95% CIs (Appendix A, Figures A.4 to A.7) for all the variable transformation methods, except for Model 2 and 3 of the variable standardization method.

5.2 Imputing poverty based on proGres22 using the model estimated from GEIH22.

Testing model assumption

We test Assumption 1 using the GEIH and ProGres datasets for the year 2022. Table 3 shows the results of the mean comparison tests. Similar to the testing results with the data for 2019, the raking and variable standardization methods result in transformed variables that are not statistically significantly different.

Imputation results

Figure 2 shows the imputed poverty rates using the normal linear regression model for different poverty lines. Unlike the 2019 data, the predicted poverty rates without data adjustment and using the matching method, which both violate Assumption 1, are largely within the 95% confidence intervals (CIs) for all the poverty lines. On the other hand, the raking and standardization methods predict the poverty rates for different poverty lines more accurately for all three models. Indeed, these two methods' imputed poverty rates are within the 95% CIs for all the poverty lines considered. The results of the empirical distribution of the error term model (Figure A.8), the probit

and models are in line with these results. The results of ML techniques, including Random Forest, Lasso, Ridge, and elastic regressions, do not outperform Dang *et al.* (2017)'s approach (Appendix A, Figures A.9 to A.12).

Further estimates

The previous results suggest that imputation based on the variable standardization method reliably predicts poverty. We further employ this method to provide poverty estimates using two different poverty lines. The first poverty line is the US$1.9 daily poverty line in 2011 PPP, which represents the international poverty line for extreme poverty. The second poverty line is the national poverty line, which corresponds to around US$ 2.86 (World Bank, 2023). We show the estimation results for these two lines respectively in Table 4, Panel A and Panel B.[10] The imputed poverty rates using the two different poverty lines are not statistically different from those obtained directly from the survey consumption data, providing further supportive evidence for the variable standardization method.

5.3 Heterogeneity analysis

We next provides heterogeneity analysis with household heads' gender and geographic locations.

Gender

We consider the 2022 data round and split the data into two samples: a sample of male-headed households and a sample of female-headed households. We implement poverty imputation separately for these two samples and show the results in Figure 3. Poverty rates are well predicted for both female-headed households and male-headed households for all the poverty lines.

[10] See also Table A4 in the annex for the full regression results.

Geography

We expand the heterogeneity analysis by splitting the data into two alternative samples: a sample of households living in border departments and households living in non-border departments. The objective is to gain a better understanding of how robust the method is at predicting poverty for different levels of geographic concentration of refugee and migrant inflows.

Furthermore, we consider three different grouping scenarios based on current displacement dynamics.

i. **Scenario 1:** All borders are considered. This means that we consider as border departments not only the ones through which refugees and migrants enter Colombia, but also those through which they leave the country. In total, there are seven departments under this scenario.

ii. **Scenario 2:** Most important arrival departments are considered. These are the ones with the largest inflow concentration and where refugees and migrants tend to settle at first after arriving to Colombia. In total, there are three departments under this scenario.

iii. **Scenario 3:** Scenario 2 departments plus two additional departments where refugees and migrants transit through to continue their journey to Central and North America. The humanitarian response has shifted to these places, as the transit through the Darien Gap has been increasing since 2022. Figure A.13 in Appendix A presents a visual depiction for these departments.

Figure 4 shows the results for Scenario 1, where all border departments are considered. Among households living in non-border departments, all three models predict poverty well with the imputed poverty rates falling within the 95th CIs. For the households in border departments, the

models also predict poverty well, but with larger CIs. The results are similar for Scenario 2 and Scenario 3 (Appendix A, Figure A.14 and Figure A.15).

6 Poverty imputation over time

Monitoring poverty trends over time is already a challenging undertaking in a developing country setting for various reasons (e.g., household consumption surveys are unavailable, or infrequently collected, or not comparable over time). Yet, this task is even more challenging for refugees, given the typically more mobile nature of their residence. We assess in this section whether the survey-to-survey imputation method reliably tracks changes in poverty rates over time in refugee settings. Specifically, we apply an imputation model that is based on an older income survey (GEIH19) to impute poverty using a more recent dataset without income (ProGres22).

Figure 5 plots the predicted poverty rates for different poverty lines, which are compared to the actual poverty rates as calculated directly from the GEIH 2022 data. Overall, the variable standardization method seems to work best for all the three models, except for lower poverty lines between the 3rd and 25th percentiles of the income distribution. The other methods work mostly when using Model 3 but also for most of the poverty lines. In particular, without any data adjustment or using the matching method, Model 3 works when the poverty line is around the 20th percentile of the income distribution or higher. For the raking method, all three models work when the poverty line is higher, at the 50th percentile of the income distribution or higher. Overall, these results suggest that we can reasonably impute poverty in 2022, using the matching and variable standardization methods, or even without any data adjustments, for most of the poverty lines.

7 Conclusion

Tracking the progress toward SDG Goal 1 of eradicating poverty for all, including forcibly displaced persons, require the availability of high-quality household income/consumption surveys. However, the majority of countries across the world, especially developing countries hosting most refugees, face challenges in collecting poverty data. High-quality consumption surveys that are comparable for forcibly displaced persons (and even the regular populations in many poor countries) are, and will, remain in limited supply, given the monetary costs and survey logistics associated with these types of surveys. In the meantime, cross-survey imputation methods can provide a second-best alternative that can potentially save time and resources.

We combine household income and census-type data on refugees to estimate welfare for refugees in Colombia in two different time periods: 2019 and 2022. Similar to many refugee settings, UNHCR's administrative data proGres may only capture well a portion of the refugee population—those that self-select to register when they are seeking assistance and hence likely represent the poorest of Venezuelan refugees and migrants. This violates a key assumption (Assumption 1) underlying cross survey imputation that the populations in the base survey and the target survey should be similar.

We employ several variable transformation methods and find that the variable standardization method and the raking method are better candidates for correcting the violation of Assumption 1 in the refugee settings in Colombia. Our results also suggest that the predicted poverty figures based on the variable standardization method appears to perform best for all the three imputation models under consideration. These results are robust to different imputation methods, including the normal linear regression method, the empirical distribution of the errors method, and the probit and logit methods. We also employ several common machine learning techniques such as Random

Forest, Lasso, Ridge, and elastic regressions for robustness checks, but these techniques generally perform worse than the imputation methods that we use.

Furthermore, we also find that we can reasonably impute poverty rates using an older household income survey and a more recent ProGres dataset for most of the poverty lines. This result is encouraging and consistent with findings in the poverty imputation literature for the general population (Dang *et al.*, 2019; Dang and Lanjouw, 2023).

Inclusion of refugees into national statistics systems such as the GEIH survey in Colombia represents the ideal data setup to obtain nationally representative poverty estimates for refugee populations. More efforts are under way to expand this data system to other countries. Yet, our imputation exercise still offers valuable opportunities for better data collection and analysis for other similar settings, particularly in upper-middle-income and middle-income countries where governments rely on national systems for refugee registration, and ProGres data only covers a small portion of the total refugee populations. Notably, UNHCR ProGres database can capture operationally relevant information on refugees. These are known as specific needs, which include different household vulnerability profiles such as single parents, female-headed households, or households with a member who has a physical disability. These unique ProGres variables can be exploited in combination with the predicted poverty data to provide more useful insights for specific vulnerable groups within refugee communities.

In fact, in other low-income country settings, UNHCR has used categorical targeting based on demographic criteria and specific needs to design a targeting strategy. For example in Niger UNHCR used categorical targeting based on demographic criteria and specific needs including categorizing those as poor who fall into the following categories: (i) female-headed households, (ii) households with members who suffer from a disability, (iii) households with a lactating mother,

and (iv) households with children under five to target assistance to Malian refugees living in Niger (Beltramo *et al.*, 2023). These examples help illustrate promising uses of poverty imputation methods that can provide better inputs for more effective support for refugees in data-challenging environments.

References

Abraham, K. G., Haltiwanger, J., Sandusky, K., and Spletzer, J. (2013) "Exploring Differences in Employment between Household and Establishment Data". *Journal of Labor Economics*, 31, S129-S172.

Altındağ, O., O'Connell, S. D., Şaşmaz, A., Balcıoğlu, Z., Cadoni, P., Jerneck, M., & Foong, A. K. (2021). Targeting humanitarian aid using administrative data: Model design and validation. *Journal of Development Economics*, *148*, 102564.

Anderson, L., & Fricker, R. D. (2015). Raking: An Important and Often Overlooked Survey Analysis Tool. *Phalanx*, *48*(3), 36–42.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685-725.

Bavier, R. (2014) "Recent Trends in U.S. Income and Expenditure Poverty". *Journal of Policy Analysis and Management*, 33, 700–718.

Beegle, K., de Weerdt, J., Friedman, J., and Gibson, J. (2012) Methods of Consumption Measurement Through Surveys: Experimental Evidence from Tanzania, *Journal of Development Economics*, 98(1), 3-18.

Beegle, K., Christiaensen, L., Dabalen, A., & Gaddis, I. (2016). *Poverty in a rising Africa*. Washington, DC: The World Bank.

Beltramo, T. P., Calvi, R., De Giorgi, G., & Sarr, I. (2023). Child poverty among refugees. *World Development*, *171*, 106340.

Beltramo, T., Dang, H.-A., Sarr, I., & Verme, P. (2024). Estimating poverty among refugee populations: A cross-survey imputation exercise for Chad. *Oxford Development Studies*, 52(1), 94-113.

Chaves-González, D., Amaral, J., & Mora, M. J. (2021). Socioeconomic Integration of Venezuelan Migrants and Refugees. *The Cases of Brazil, Chile, Colombia, Ecuador, and Peru. Washington, DC and Panama City: Migration Policy Institute and International Organization for Migration*. https://www.r4v.info/sites/default/files/2021-07/mpi-iom_socioeconomic-integration-venezuelans_2021_final.pdf

Cuesta, J., & Ibarra, G. L. (2017). Comparing cross-survey micro imputation and macro projection techniques: Poverty in post revolution Tunisia. *Journal of Income Distribution*, *25*(1), 1-30.

DANE. (2021). *Caracterización de los migrantes y retornados desde Venezuela a partir del CNPV-2018. National Department of Statistics. Informes de Estadística Sociodemográfica Aplicada. No. 5.* [dataset].

Dang, H.-A. H., Lanjouw, P. F., & Serajuddin, U. (2017). Updating poverty estimates in the absence of regular and comparable consumption data: Methods and illustration with reference to a middle-income country. *Oxford Economic Papers*, *69*(4), 939–962.

Dang, H.-A. H. and Lanjouw, P. F. (2023). "Regression-based Imputation for Poverty Measurement in Data Scarce Settings". In Jacques Silber. (Eds.). *Handbook of Research on Measuring Poverty and Deprivation*. Edward Elgar Press.

Dang, H.-A. H., & Verme, P. (2023). Estimating poverty for refugees in data-scarce contexts: An application of cross-survey imputation. *Journal of Population Economics*, 36(2), 653- 679.

Dang, H. A., Kilic, T., Hlasny, V., Abanokova, K., & Carletto, C. (2024). "Using Survey-to-Survey Imputation to Fill Poverty Data Gaps at a Low Cost". Policy Research Working Paper 10738. World Bank: Washington DC.

Deaton, A. (1997). *The analysis of household surveys: A microeconometric approach to development policy*. World Bank and John Hopkins University Press.

Deaton, A., & Kozel, V. (2005). *The great Indian poverty debate*. New Delhi: Macmillan.

Deville, J.-C., Särndal, C.-E., & Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, *88*(423), 1013–1020. h

Douidich, M., Ezzrari, A., Van der Weide, R., & Verme, P. (2016). Estimating quarterly poverty rates using labor force surveys: A primer. *World Bank Economic Review*, *30*(3), 475–500.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, *71*(1), 355–364.

Gibson, J., Huang, J., & Rozelle, S. (2003). Improving estimates of inequality and poverty from urban China's household income and expenditure survey. *Review of Income and Wealth*, *49*(1), 53-68.

Guzman Gutierrez, Carlos Santiago. (2023). *Assessment of Key Datasets & Web Portals for Immediate Research Opportunities in Colombia. UNHCR*

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Mathiassen, A. (2009). A model-based approach for predicting annual poverty rates without expenditure data. *Journal of Economic Inequality*, *7*(2), 117–135.

Mathiassen, A., & Wold, B. K. G. (2021). Predicting poverty trends by survey-to-survey imputation: The challenge of comparability. *Oxford Economic Papers*, *73*(3), 1153–1174.

Mejia-Mantilla, C., Gonzalez Rubio, S. D. C., Lendorfer, J., & Rodriguez Guio, D. F. (2023). *Venezuelans in Chile, Colombia, Ecuador and Peru: A Development Opportunity*. World Bank and UNHCR report.

Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, *201*, 746–755.

Migration Colombia. (2023). *Distribución de Migrantes Venezolanas(os).* https://www.migracioncolombia.gov.co/infografias-migracion-colombia/distribucion-de-migrantes-agosto--2023

Moffitt, R., Abowd, J., Bollinger, C., Carr, M., Hokayem, C., McKinney, K., ... & Ziliak, J. (2022). Reconciling trends in us male earnings volatility: Results from survey and administrative data. *Journal of Business & Economic Statistics*, *41*(1), 1-11.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

OECD. (2023). *Development Finance for Refugee Situations: Volumes and trends, 2020-21*. https://www.oecd.org/dac/development-finance-refugee-situations-2020-2021.pdf

Phua, C., Lee, V., & Smith-Miles, K. (2006). The personal name problem and a recommended data mining solution. *Encyclopedia of Data Warehousing and Mining*, 24.

R4V. (2023). *Caracterización de Movimientos Mixtos hacia Centro y Norte América—Darién, Febrero 2023.* https://www.r4v.info/es/document/gifmm-colombia-caracterizacion-de-movimientos-mixtos-hacia-centro-y-norte-america-darien

Raffo, J., & Lhuillery, S. (2009). How to play the "Names Game": Patent retrieval comparing different heuristics. *Research Policy*, *38*(10), 1617–1627.

Tarozzi, A. (2007). Calculating Comparable Statistics From Incomparable Surveys, With an Application to Poverty in India. *Journal of Business & Economic Statistics*, *25*(3), 314–336.

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, *106*.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

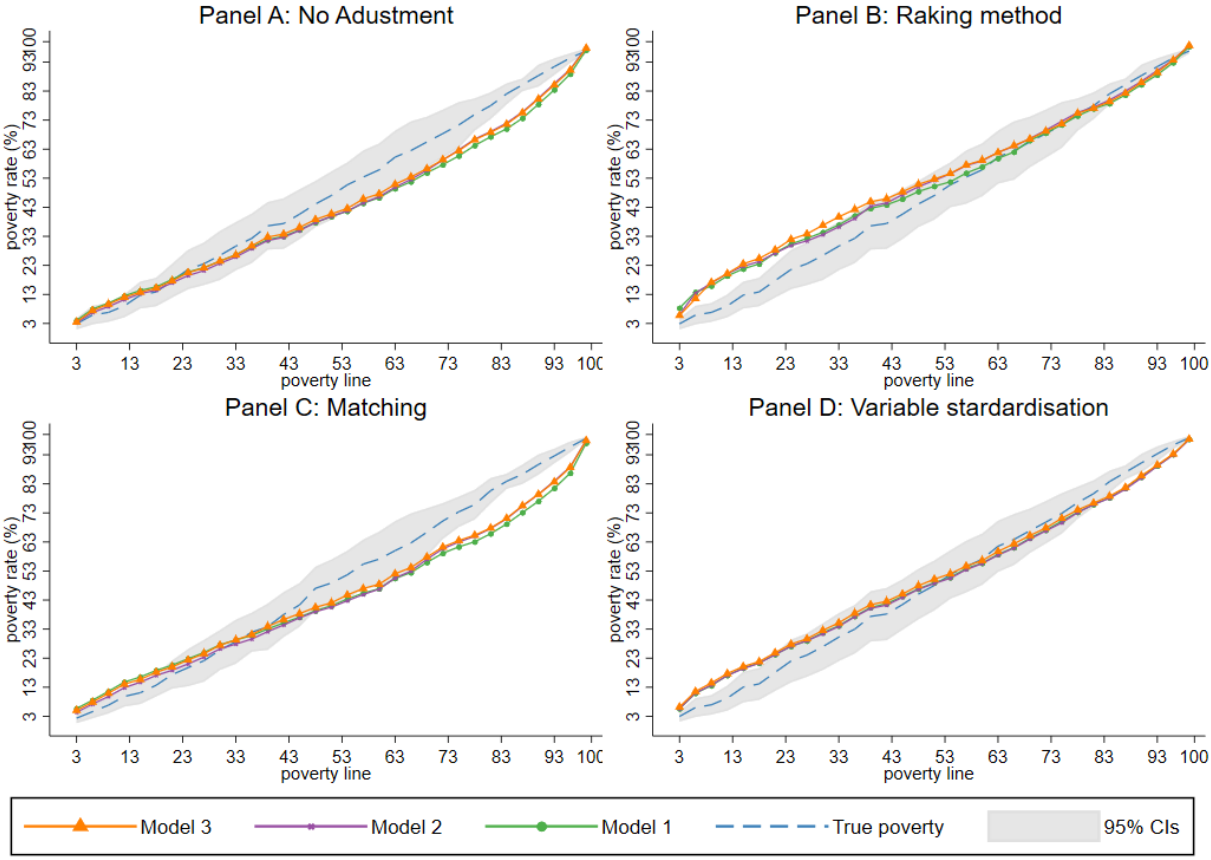
UNHCR. (2023a). *2023 Global Compact on Refugees Indicator Report*. https://www.unhcr.org/what-we-do/reports-and-publications/data-and-statistics/indicator-report-2023

UNHCR. (2023b). *Global trends—Forced Displacement in 2022*. https://www.unhcr.org/sites/default/files/2023-06/global-trends-report-2022.pdf

World Bank. (2023c). *Poverty & Equity Brief-Colombia*.

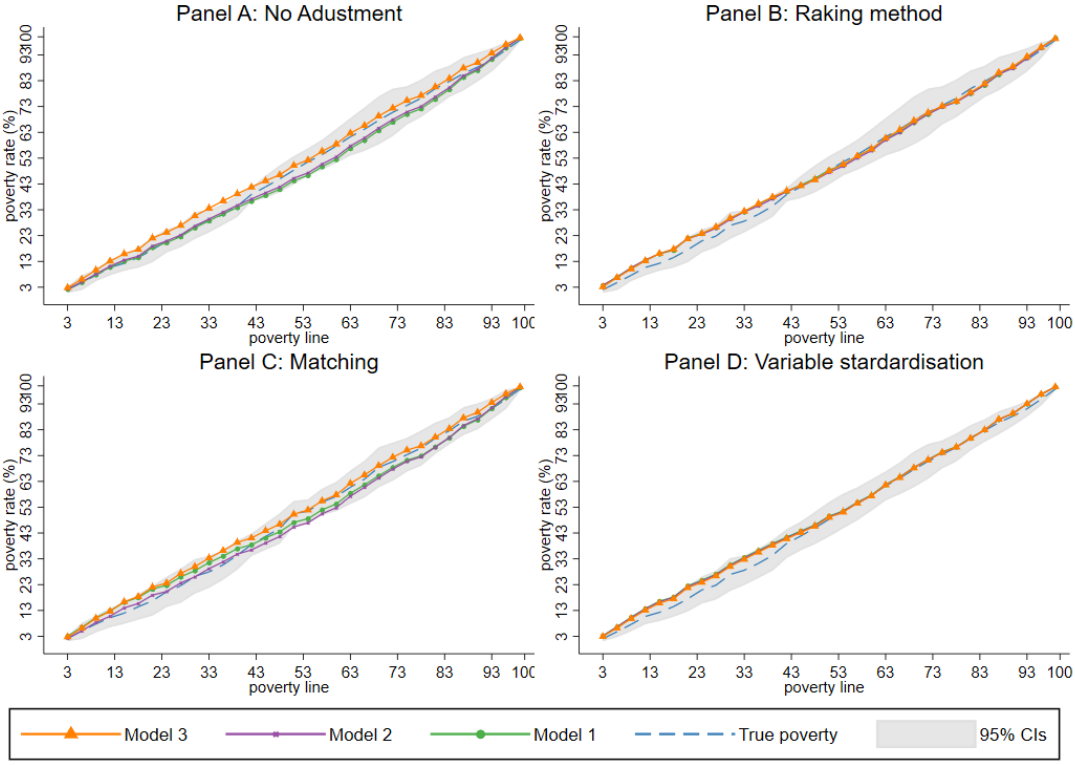
World Bank. (2012). "Well Begun, Not Yet Done: Vietnam's Remarkable Progress on Poverty Reduction and the Emerging Challenges". *Vietnam Poverty Assessment Report 2012*. Hanoi: World Bank.

Figure 1: Imputed income for different poverty lines based on proGres19 using the model estimated from GEIH19, by variable transformation method



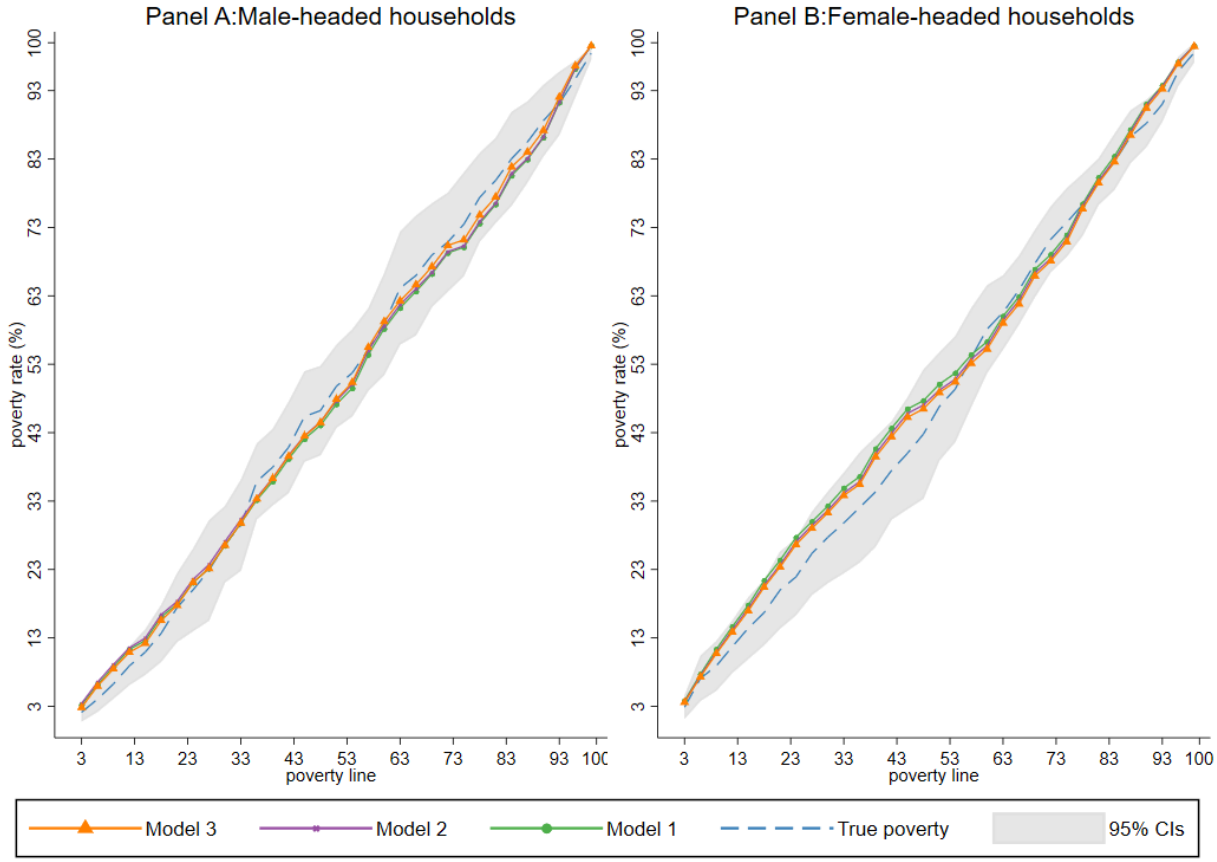
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure 2: Imputed income for different poverty lines based on proGres22 using the model estimated from GEIH22, by variable transformation method



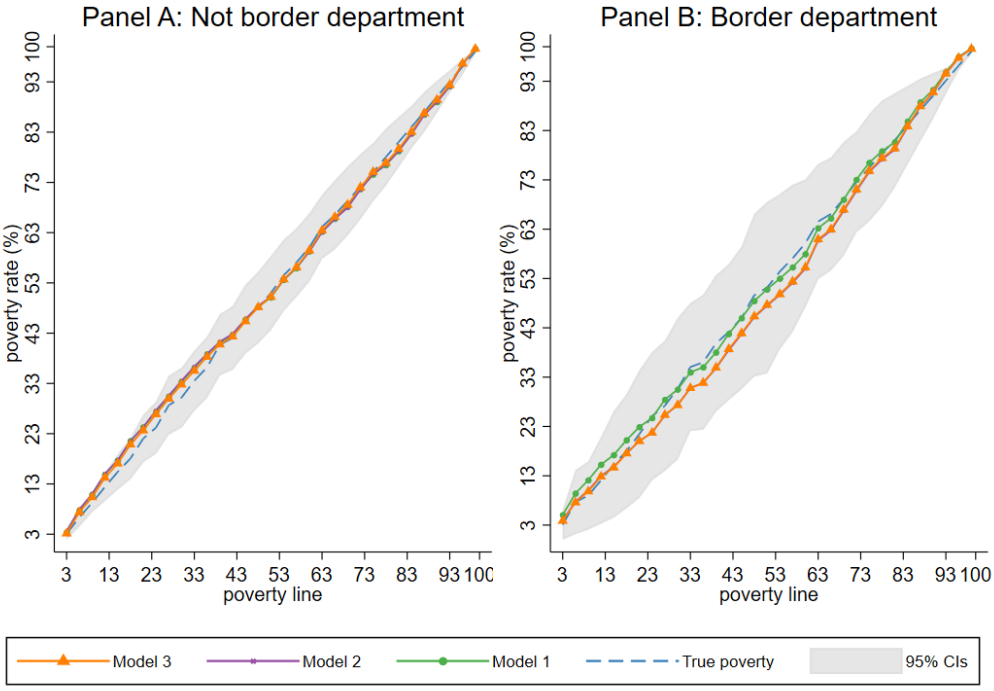
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure 3: Female headed household vs. male headed households



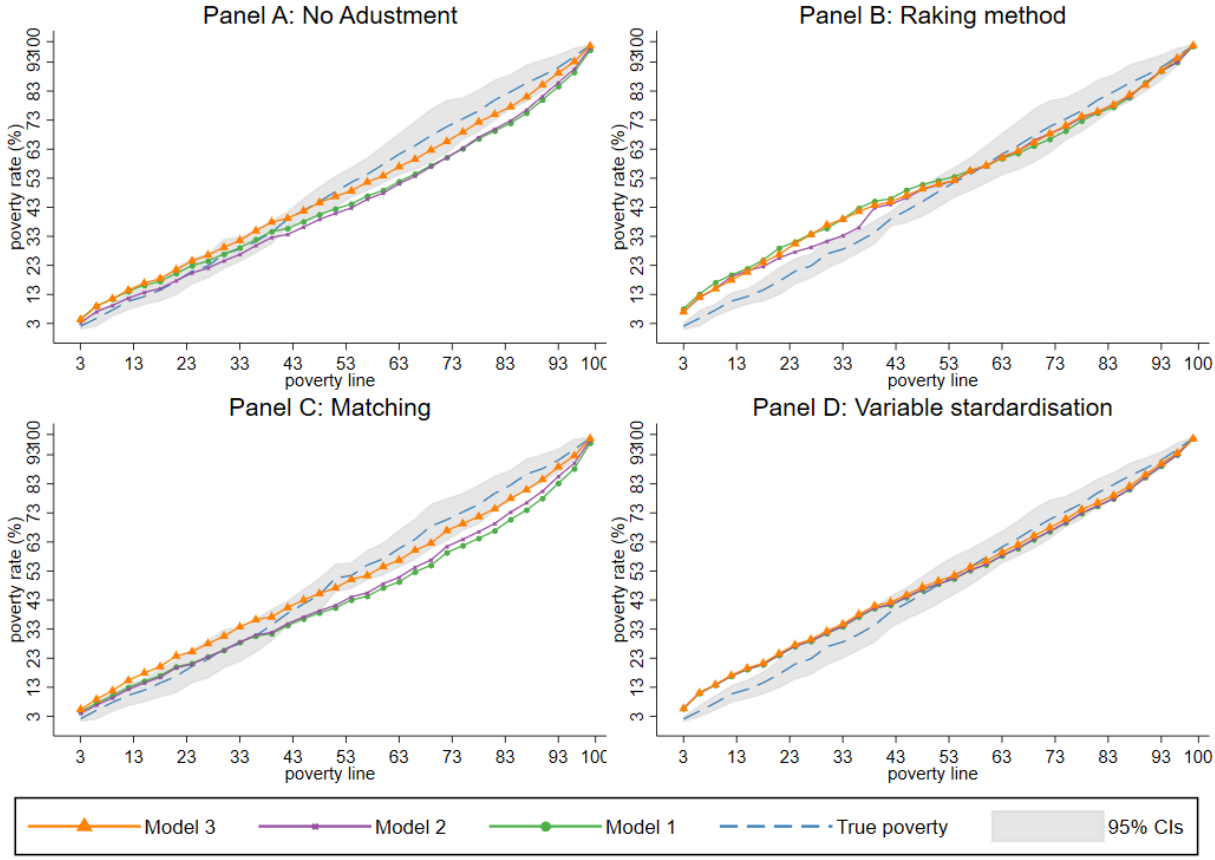
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure 4: Scenario 1 – Non-border departments vs. border departments



Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure 5: Updating poverty figures for proGres22 data, using imputation model based on GEIH19



Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Table 1: Summary statistics

Panel A: 2019 data

	Survey dataset			proGres dataset		
	Mean	Std. dev.	# Obs	Mean	Std. dev.	# Obs
Female	0.30	0.46	1859	0.64	0.48	6491
Age	35.20	11.68	1859	36.66	11.28	6491
No Education	0.02	0.13	1859	0.01	0.11	6491
Primary	0.14	0.35	1859	0.12	0.32	6491
Secondary/High School	0.21	0.41	1859	0.19	0.39	6491
Tertiary/University	0.25	0.43	1859	0.29	0.46	6491
HH sizes	4.04	2.25	1859	2.00	1.49	6491
Log (Income)	12.88	1.11	1859			

Panel B: 2022 data

	Survey dataset			proGres dataset		
	Mean	Std. dev.	# Obs	Mean	Std. dev.	# Obs
Female	0.44	0.50	1861	0.66	0.47	58602
Age	36.04	11.37	1861	32.44	12.71	58602
No Education	0.02	0.14	1861	0.03	0.17	58602
Primary	0.12	0.33	1861	0.22	0.41	58602
Secondary/High School	0.20	0.40	1861	0.36	0.48	58602
Tertiary/University	0.14	0.34	1861	0.00	0.06	58602
HH sizes	2.91	1.58	1861	2.17	1.41	58602
Log (Income)	13.21	1.00	1859			

Table 2: Mean comparison for 2019- proGres19 data and GEIH19 data

	Method			
	No adjustment	Matching	Raking Method	Standardization
Female	-0.343***	-0.321***	0.000	0.001
	(-18.224)	(-13.183)	(0.000)	(0.043)
Age	-1.454***	-1.806***	0.000	0.000
	(-2.840)	(-3.187)	(0.000)	(-0.091)
None	0.005	0.487***	-0.005	0.004
	(1.239)	(67.134)	(-0.526)	(0.970)
Primary	0.023	0.012	0.000	0.000
	(1.509)	(0.864)	(0.000)	(0.013)
Secondary/High School	0.018	0.005	0.000	0.002
	(0.323)	(0.084)	(0.000)	(0.045)
Tertiary or University	-0.046	-0.002	0.000)	0.001
	(-1.106)	(-0.036)	(0.000)	(0.033)
HH size	2.032***	1.337***	0.000	0.010
	(6.832)	(4.018)	(0.000)	(0.028)

*Note: The table compares the mean difference between the GEIH and proGres. Each column represent a given data adjustment method and provides the mean difference between GEIH and proGres. The stars indicate whether the difference is significant or not. Standard errors in parentheses, * p < 0.10, ** p < 0.05, *** p < 0.01.*

Table 3: Mean comparison for 2022- proGres22 data and GEIH22 data

	Method			
	No adjustment	Matching	Raking Method	Standardization
Female	-0.215***	-0.208***	-0.000	0.001
	(-2.968)	(-3.011)	(-0.000)	(0.17)
Age	3.586***	13.550***	0.000	0.000
	(4.978)	(24.124)	(0.000)	(0.068)
None	-0.012***	0.469***	0.000	-0.011
	(-2.333)	(81.871)	(0.000)	(-1.714)
Primary	-0.096***	-0.220***	0.029***	0.011
	(-6.287)	(-13.748)	(-3.507)	(0.052)
Secondary/High School	-0.154***	-0.088**	0.000	-0.002
	(-5.849)	(-2.336)	(0.000)	(-0.097)
Tertiary or University	0.132***	0.407***	0.000	-0.003
	(17.294)	(47.976)	(0.000)	(-0.195)
HH size	0.737***	0.661***	0.000	0.015
	(3.097)	(3.343)	(0.000)	(0.061)

Note: The table compares the mean difference between the GEIH and proGres. Each column represent a given data adjustment method and provides the mean difference between GEIH and proGres. The stars indicate whether the difference is significant or not. Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Imputed Poverty Rates Using the International and National Poverty Lines*

		2022		
		Model 1	Model 2	Model 3
		1	2	3
Panel A: Poverty rates using international standard				
Normal linear regression model		12.3	12.2	12.5
		(1.1)	(1.1)	(1.1)
Empirical error model		7.7	7.8	7.9
		(0.8)	(0.8)	(0.8)
Survey-based poverty estimate				
	7.9			
	(2.0)			
Panel B: Poverty rates using national standard				
Normal linear regression model		45.82	45.7	46.9
		(2.2)	(2.1)	(2.1)
Empirical error model		47.05	46.6	46.4
		(2.4)	(2.3)	(2.3)
Survey-based poverty estimate				
	47.81			
	(3.9)			
Control Variables				
Size and Gender		Y	Y	Y
Age		N	Y	Y
Education		N	N	Y
adjusted		0.114	0.12	0.14
N		43194	43194	43194

Note: The international total poverty line is $2.152017 PPP or 3517.5 Colombian Pesos per person per day, while the national poverty line is 11421 Colombian Pesos per person per day, equivalent to 354037 per person per month. Robust standard errors in parentheses are clustered at the department level. We use 1,000 simulations for each model run.

Source: Authors' calculations.

Appendix A: Additional tables and figures

Table A.1: Summary of data

Name and Year	**proGres19**	**GEIH19**	**proGres22**	**GEIH22**
Year	2019	2019	2022	2022
Type of survey	Census	survey	Census	survey
Producer of dataset	UNHCR	DANE(NSO)	UNHCR	NSO
Number observations				
Existence of income information	No	Yes	No	Yes
Relevant Variables to poverty imputation available	1. HH size 2. Age of HHH 3. Gender of HHH 4. Education of HHH	1. HH size 2. Age of HHH 3. Gender of HHH 4. Education of HHH 5. HH Income	1. HH size 2. Age of HHH 3. Gender of HHH 4. Education of HHH	1. HH size 2. Age of HHH 3. Gender of HHH 4. Education of HHH 5. HH Income

Source: Authors calculations

Table A.2: Distribution of forcibly displaced persons by category in Colombia

Type	Number	Proportion
Refugee and Asylum seeker	1,562	1.13%
Returnees	279	0.20%
Returned IDP	144	0.10%
IDPs	3,653	2.63%
Other of concern	132,692	95.94%
Total	138,312	100%

Source: Authors' calculations, ProGres. The table excludes the category “Not of concern”

Table A.3: Regression coefficients of Model 3- GEIH19

VARIABLES	(1) No adjustment	(2) Matching	(3) Raking Method	(4) Standardization
HH size	-0.14***	-0.20***	-0.14***	-0.23***
	(0.01)	(0.02)	(0.01)	(0.02)
Female	-0.05	-0.10	-0.05	-0.01
	(0.05)	(0.06)	(0.05)	(0.05)
Age	-0.02*	-0.01	-0.02*	0.07***
	(0.01)	(0.02)	(0.01)	(0.02)
Age squared	0.00***	0.00	0.00***	
	(0.00)	(0.00)	(0.00)	
Primary	-0.19**	-0.28***	-0.19**	-0.13
	(0.08)	(0.10)	(0.08)	(0.09)
Secondary/High School	-0.13**	-0.10	-0.13**	-0.12*
	(0.07)	(0.08)	(0.07)	(0.07)
Tertiary/University	0.36***	0.36***	0.36***	0.29***
	(0.06)	(0.07)	(0.06)	(0.06)
Constant	13.59***	13.64***	13.59***	12.63***
	(0.22)	(0.34)	(0.22)	(0.15)
Observations	1,809	1,360	1,809	1,809

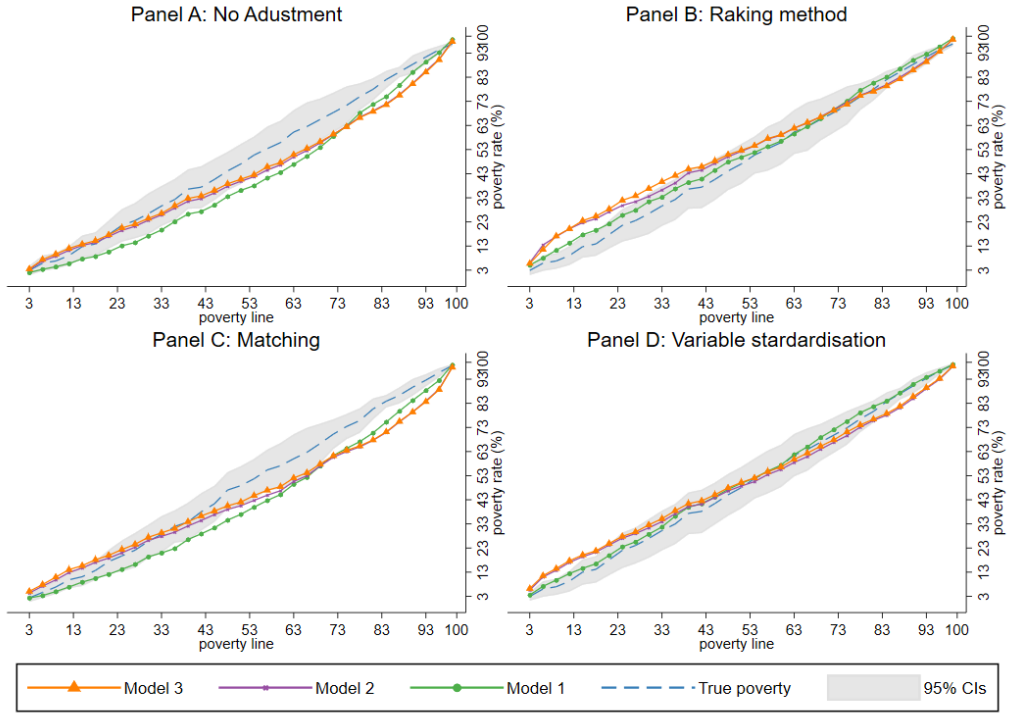
Note: The table compares the mean difference between the GEIH and proGres. Each column represent a given data adjustment method and provides the mean difference between GEIH and proGres. The stars indicate whether the difference is significant or not. Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4: Regression coefficients-Model 3-2022, GEIH22

VARIABLES	(1) No adjustment	(2) Matching	(3) Raking	(4) Standardization
HH sizes	-0.18***	-0.18***	-0.18***	-1.22***
	(0.01)	(0.02)	(0.01)	(0.08)
Female	-0.13***	-0.06	-0.13***	-0.01
	(0.04)	(0.06)	(0.04)	(0.05)
Age	-0.03***	-0.02	-0.03***	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)
Age squared	0.00***	0.00**	0.00***	
	(0.00)	(0.00)	(0.00)	
Primary	-0.16**	-0.27***	-0.16**	-0.10
	(0.07)	(0.09)	(0.07)	(0.07)
Secondary/High School	-0.25***	-0.12*	-0.25***	-0.10*
	(0.05)	(0.07)	(0.05)	(0.06)
Tertiary/University	0.35***	0.34***	0.35***	2.10***
	(0.06)	(0.07)	(0.06)	(0.44)
Constant	14.15***	13.69***	14.15***	12.85***
	(0.18)	(0.26)	(0.18)	(0.11)
Observations	1,861	1,600	1,861	1,809

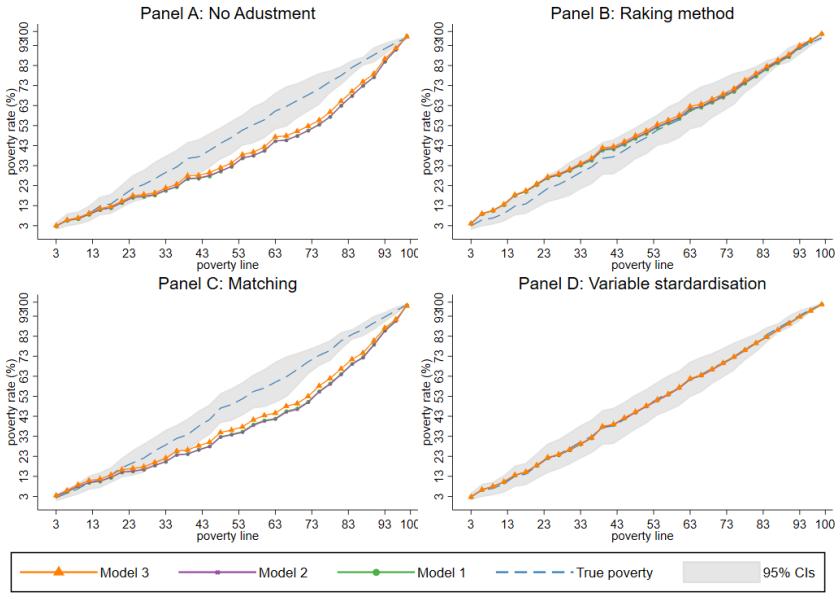
*Note: The table compares the mean difference between the GEIH and proGres. Each column represent a given data adjustment method and provides the mean difference between GEIH and proGres. The stars indicate whether the difference is significant or not. Standard errors in parentheses, * p < 0.10, ** p < 0.05, *** p < 0.01.*

Figure A.1: Imputed poverty for different poverty lines for 2019, Empirical method



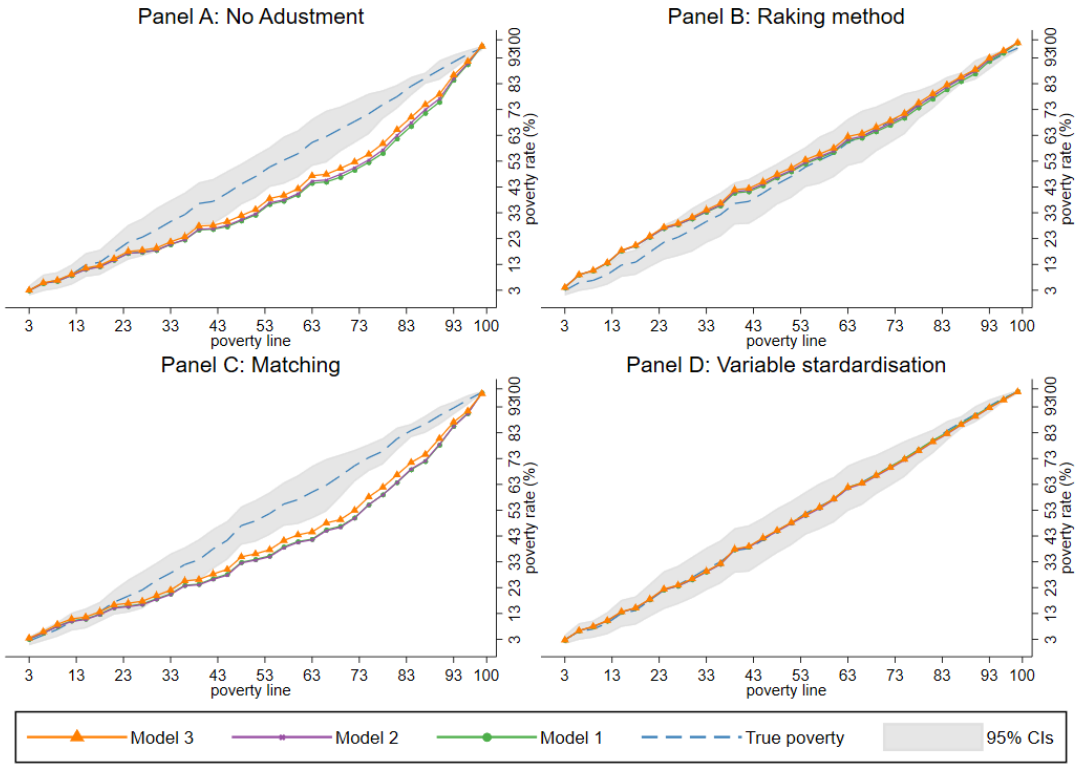
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure A.2: Imputed poverty for different poverty lines for 2019, method probit



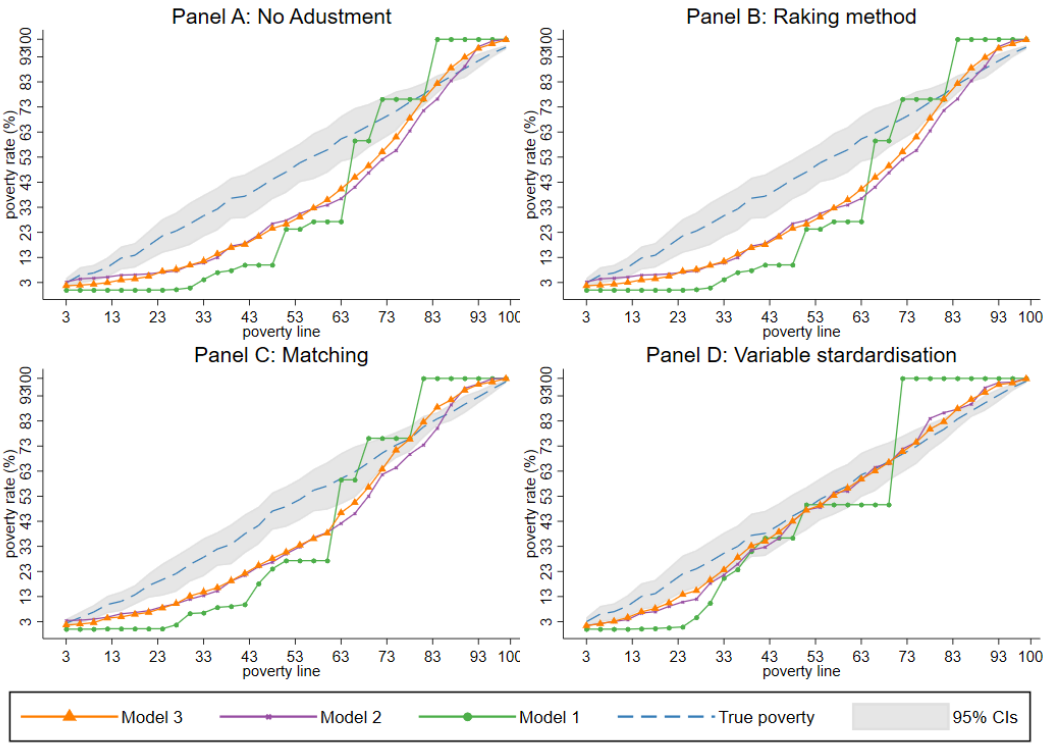
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure A.3: Imputed poverty for different poverty lines for 2019, Logit method



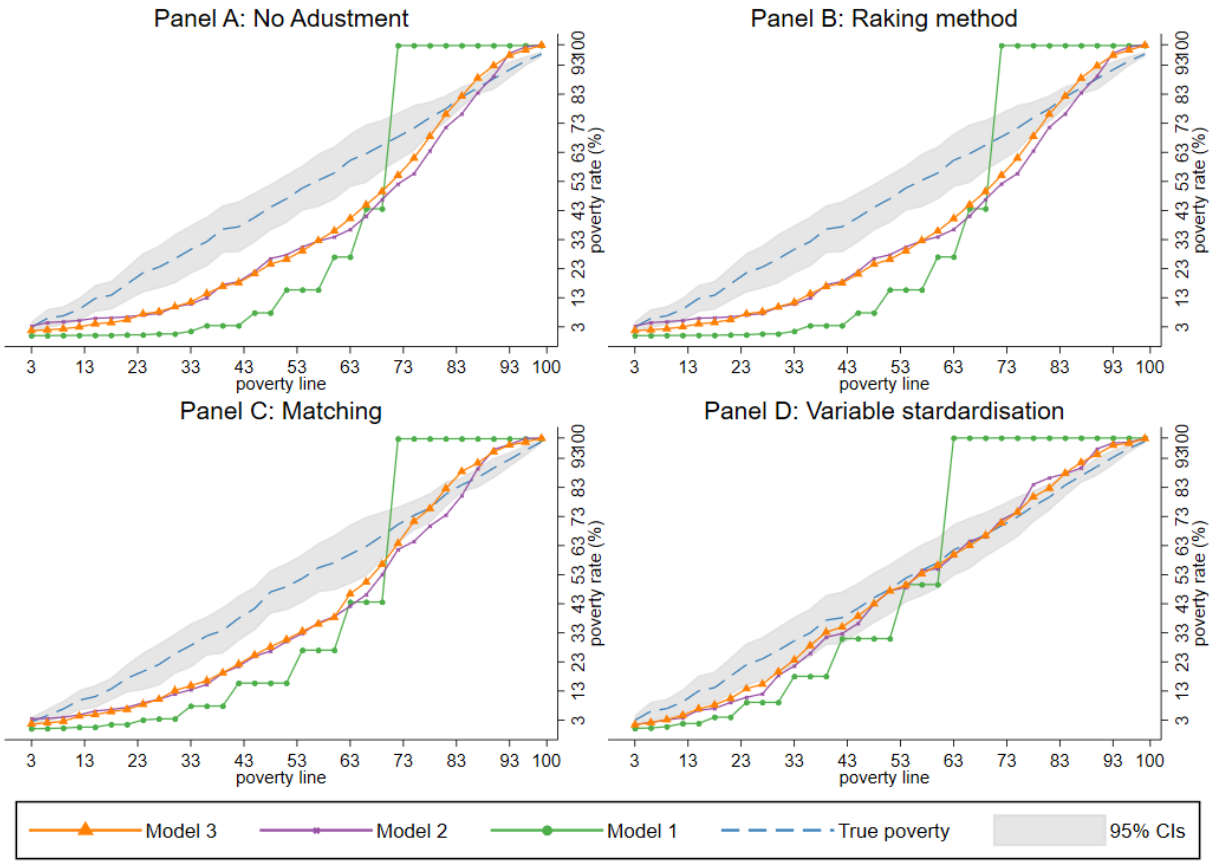
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure A.4: Imputed poverty for different poverty lines for 2019, Machine Learning (Random Forest)



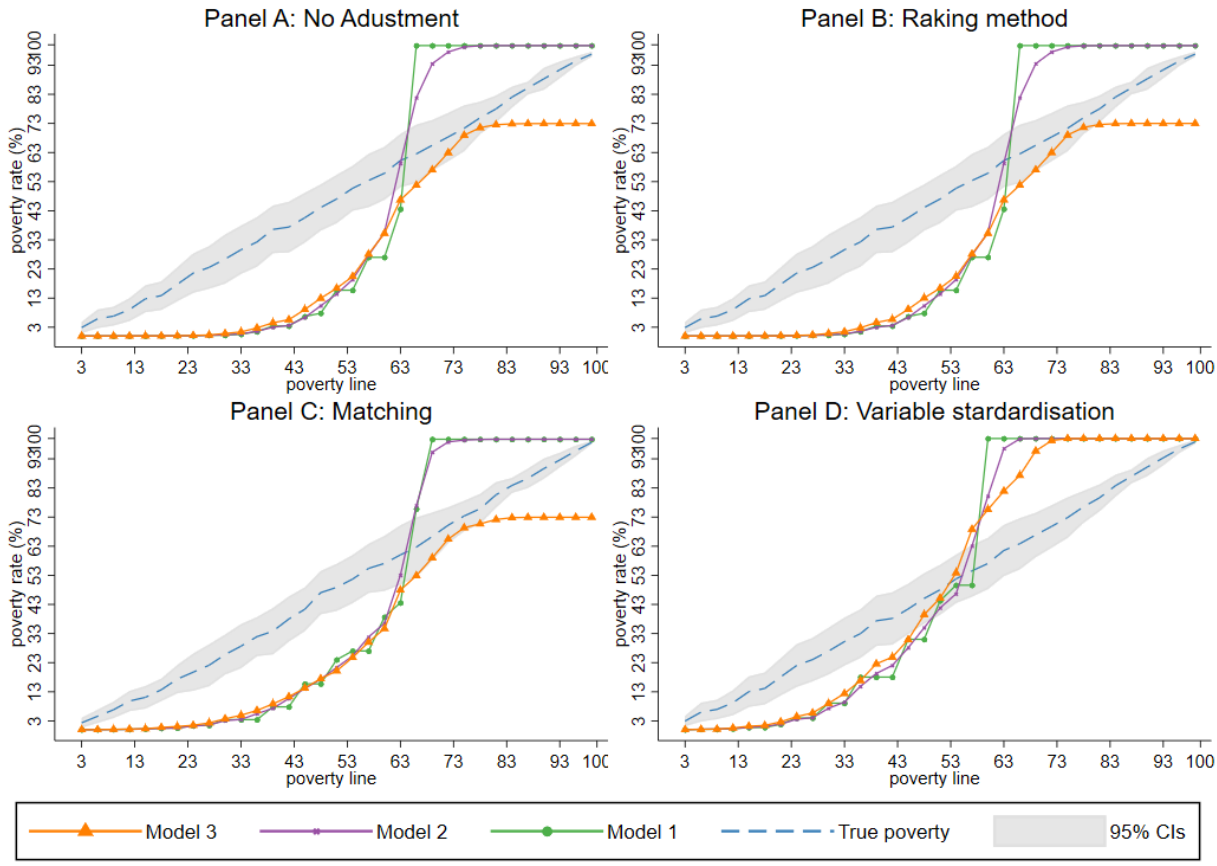
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure A.5: Imputed poverty for different poverty lines for 2019, Lasso



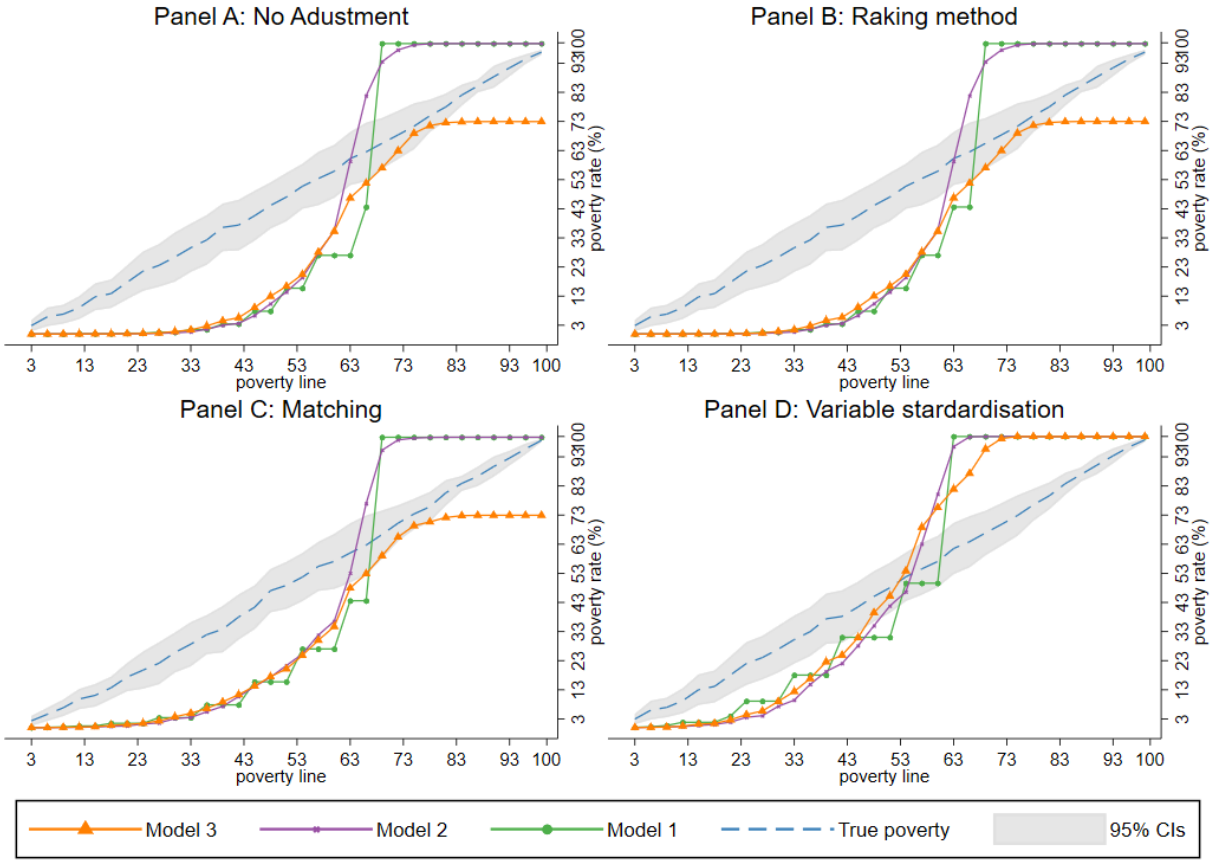
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the progress

Figure A.6: Imputed poverty for different poverty lines for 2019, Ridge regressions



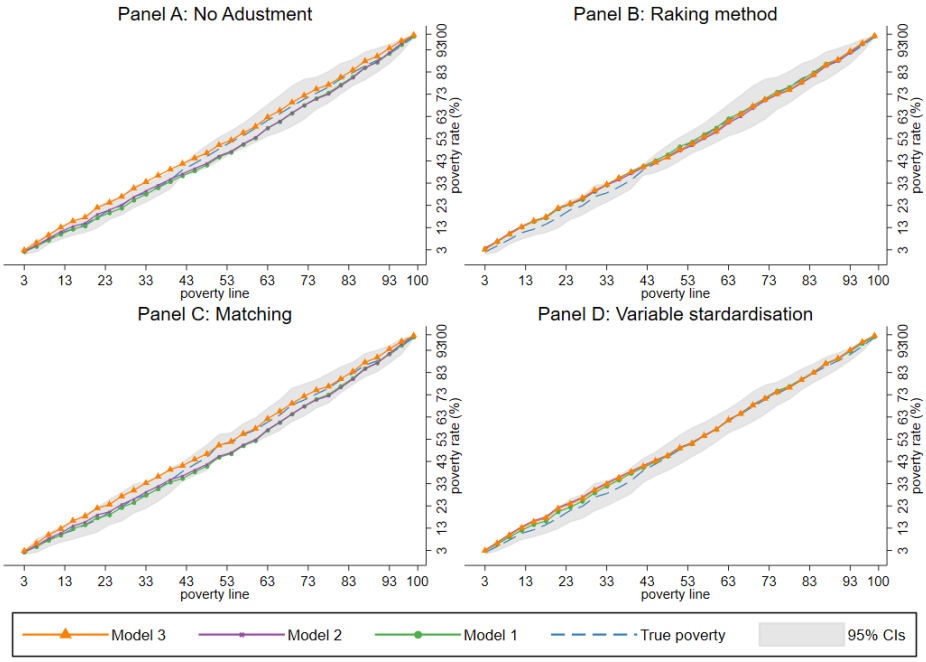
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the progress

Figure A.7: Imputed poverty for different poverty lines for 2019, Elastic regression



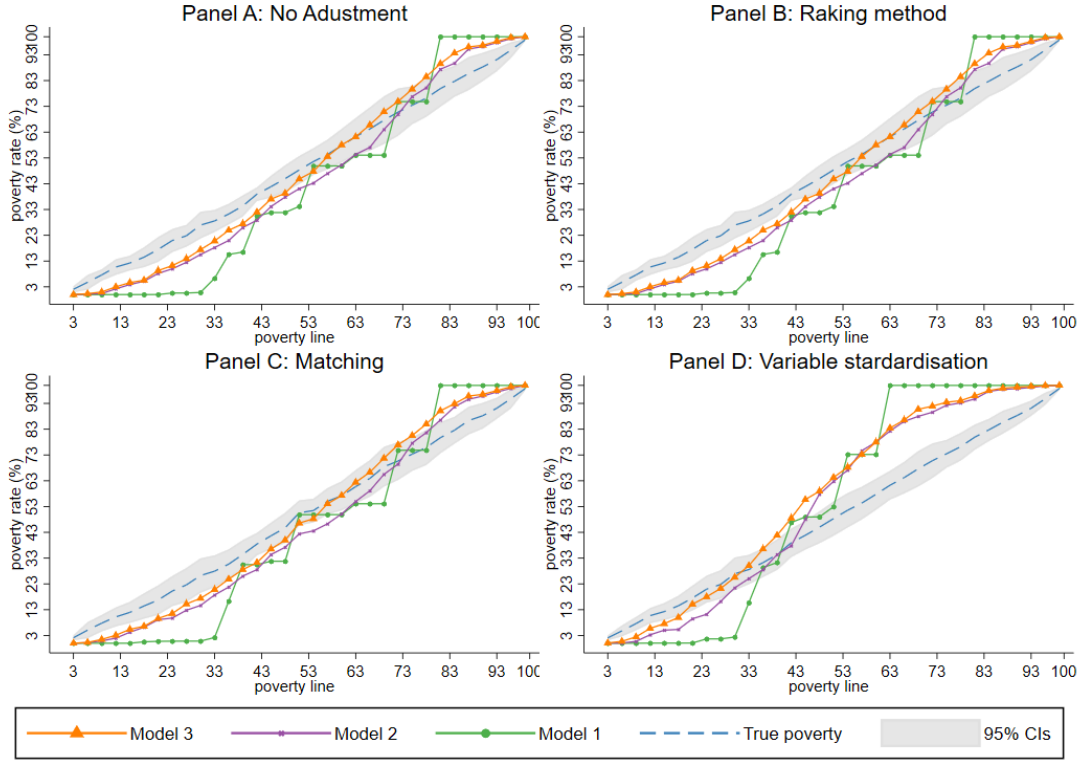
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the progress

Figure A.8: Imputed poverty for different poverty lines for 2022, Empirical method



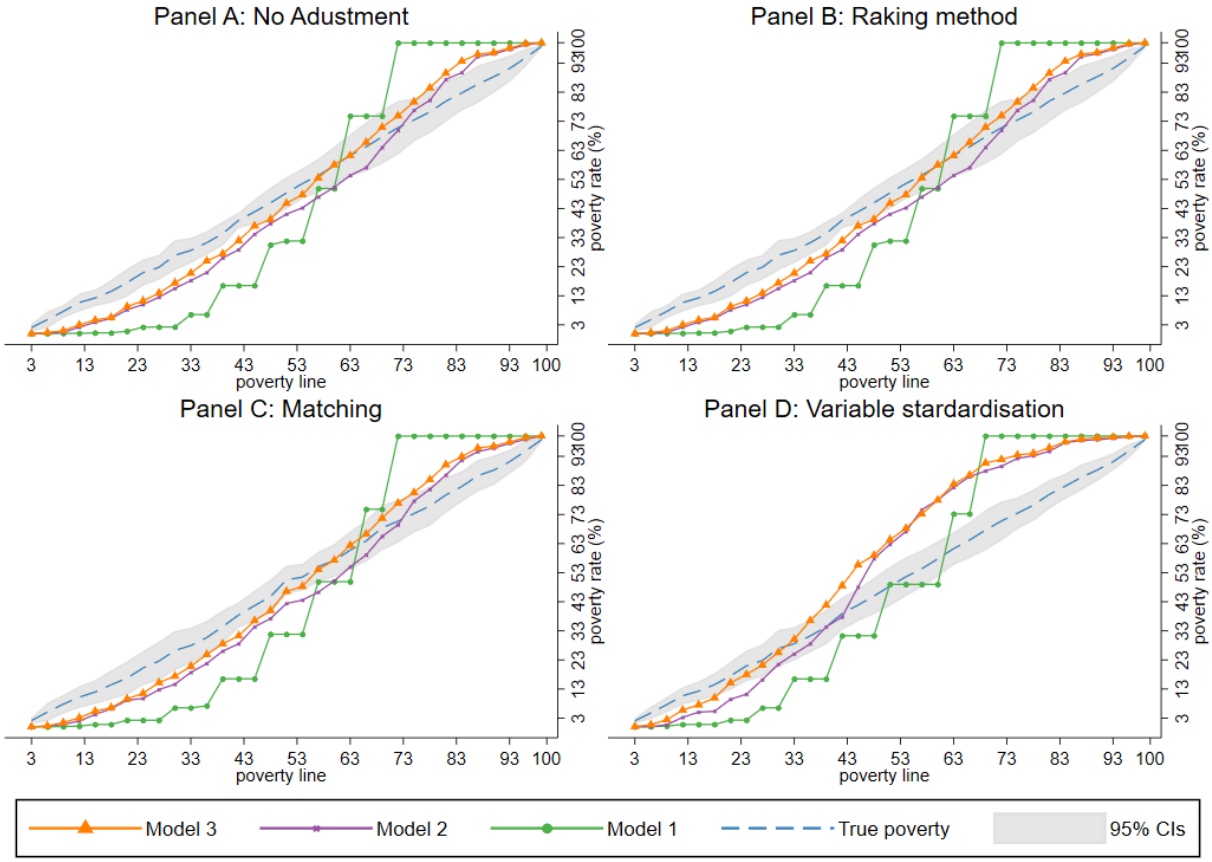
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure A.9: Imputed poverty for different poverty lines for 2022, Machine Learning (Random Forest)



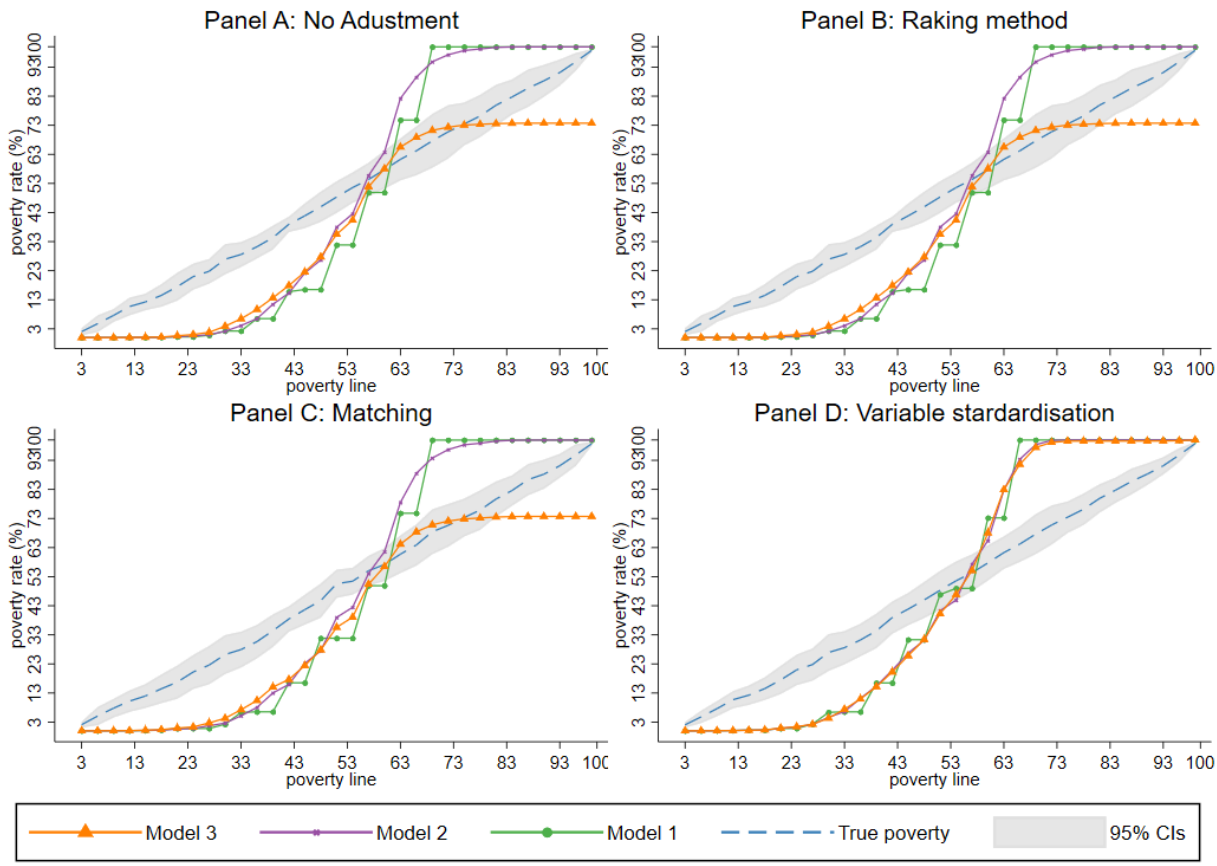
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the progress

Figure A.8: Imputed poverty for different poverty lines for 2022, Lasso



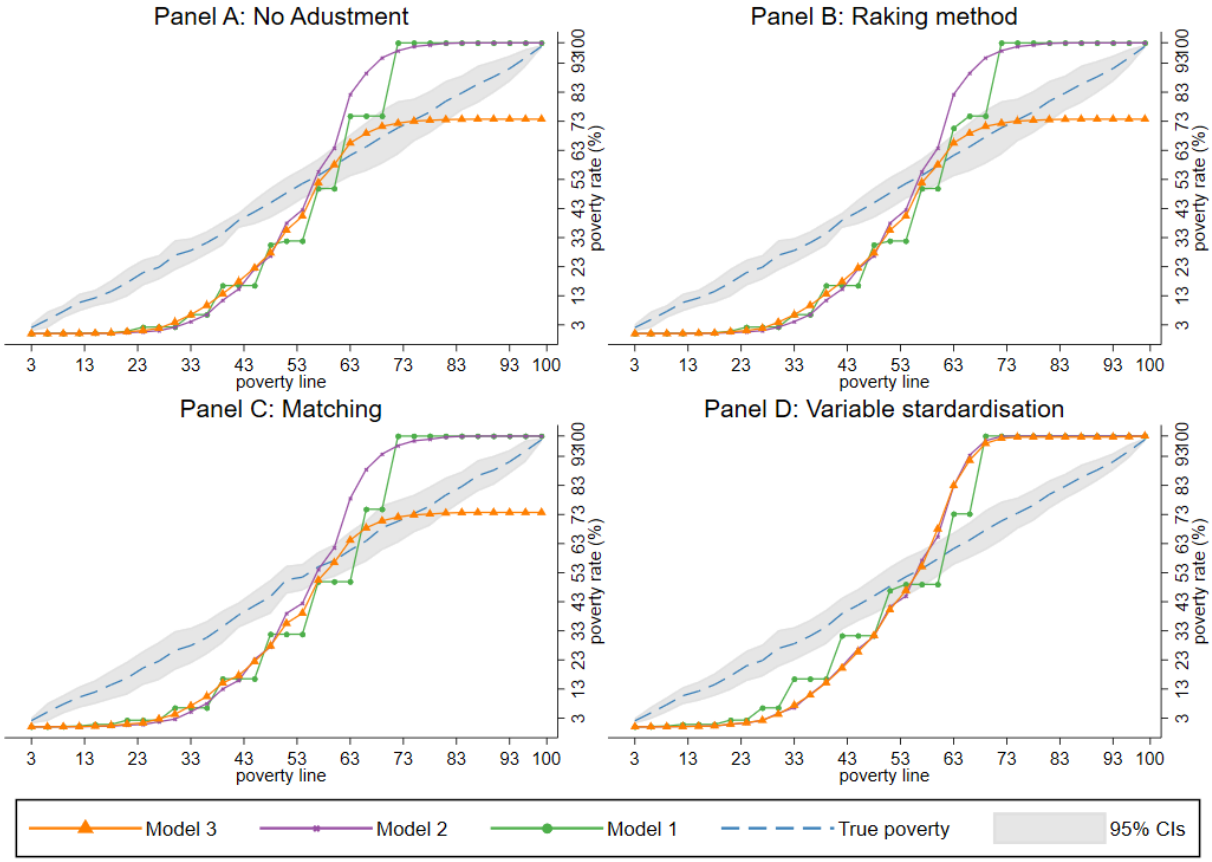
Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the progress

Figure A.9: Imputed poverty for different poverty lines for 2022, Ridge regression



Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the progress

Figure A.10: Imputed poverty for different poverty lines for 2022, Elastic regression

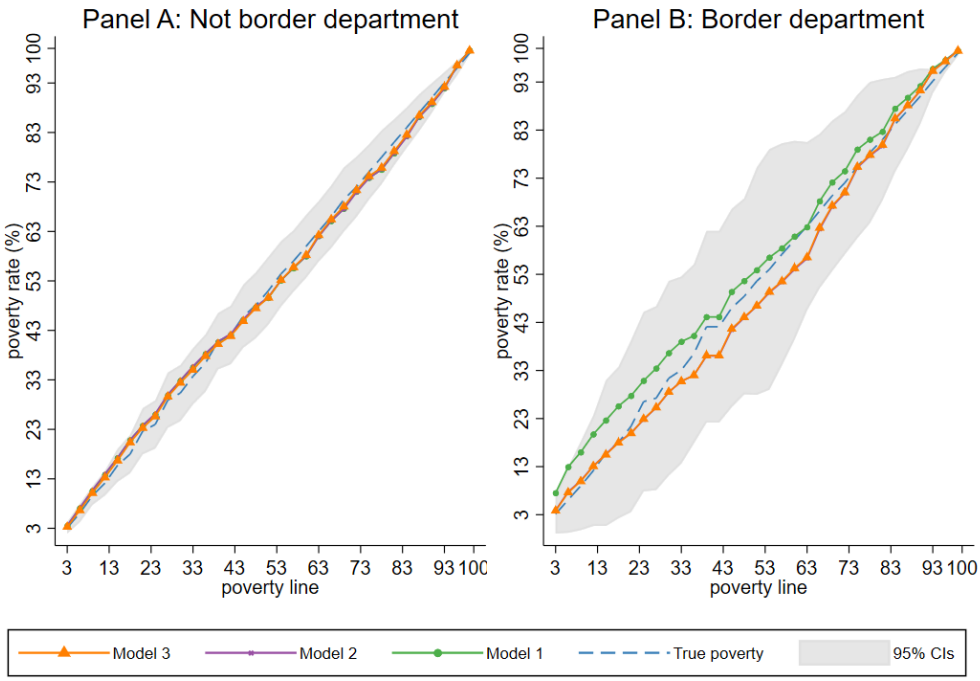


Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure A.13: Departments

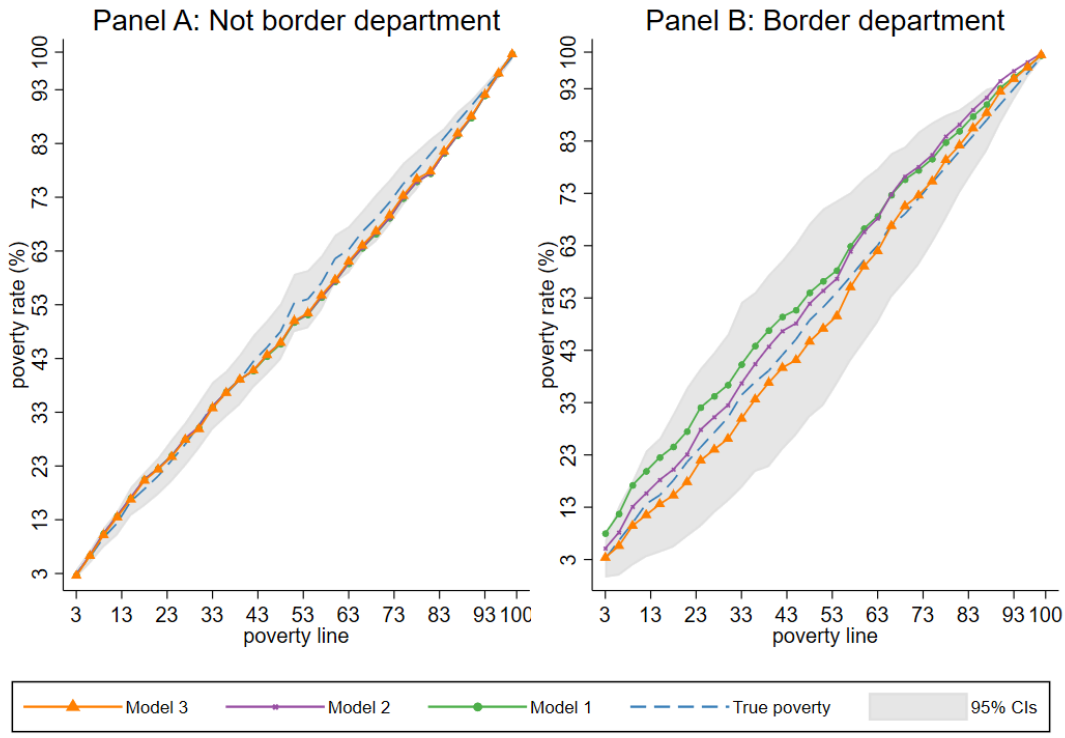
Note: This map depicts the percentage of Venezuelans residing in each department of Colombia. The yellow circle represents the proportion of Venezuelans in a department, calculated by dividing the number of Venezuelans in the department by the total number of Venezuelans in the country. The departments have been divided into four categories: (i) Not border departments, which are those that do not share a border with Venezuela and have a low inflow of Venezuelan migrants and refugees; (ii) Most important arrival departments, which are the ones that have the highest concentration of Venezuelan arrivals; (iii) Response departments, which are not border departments but are important for people continuing their journey to Central and North America through the Darien Gap, and (iv) Other borders.

Figure A.14: Scenario 2 - Non-border departments vs border departments



Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres

Figure A.15: Scenario 3 – Non-border departments vs border departments



Note: The blue dashed curve presents the actual poverty rates derived from the GEIH survey, meaning that the blue dashed curve presents poverty rates derived from observed income of the GEIH. The green solid curve with circle symbol represents the imputed poverty rates from Model 1 with observations from ProGres. The indigo solid curve with symbol "x" represents the imputed poverty rates from Model 2 with the proGres while the orange solid curve with the triangle symbol represents the imputed poverty rates from Model 3 with the proGres