REM WORKING PAPER SERIES

Multidimensional poverty in Benin

Esmeralda Arranhado, Lágida Barbosa, João A. Bastos

REM Working Paper 0343-2024

September 2024

REM – Research in Economics and Mathematics
Rua Miguel Lúpi 20,
1249-078 Lisboa,
Portugal

ISSN 2184-108X

REM – Research in Economics and Mathematics

Rua Miguel Lupi, 20
1249-078 LISBOA
Portugal

Telephone: +351 - 213 925 912
E-mail: rem@iseg.ulisboa.pt

https://rem.rc.iseg.ulisboa.pt/

https://twitter.com/ResearchRem

https://www.linkedin.com/company/researchrem/

https://www.facebook.com/researchrem/

Multidimensional poverty in Benin

Esmeralda Arranhado[a*] Lágida Barbosa[b] João A. Bastos[a]

[a]ISEG - Lisbon School of Economics and Management and REM/CEMAPRE, Universidade de Lisboa

[b]Banco de Cabo Verde, Cabo Verde

Abstract

We examine an individual-level poverty measure for Benin using cross-sectional data. Since our measure is defined within the interval [0,1], we combine fractional regression models and machine learning models for fractions to examine the factors influencing multidimensional poverty measures and to predict poverty levels. Our approach illustrates the potential of combining parametric models, that inform on the statistical significance and variable interactions, with SHapley Additive exPlanations (SHAP) and Accumulated Local Effects (ALE) plots obtained from a random forest. Results highlight the importance of addressing gender inequalities in education, particularly by increasing access to female education, to effectively reduce poverty. Furthermore, natural conditions arising from agroecological zones are significant determinants of multidimensional poverty, which underscores the need for climate change policies to address poverty in the long term, especially in countries heavily reliant on agriculture. Other significant determinants of welfare include household size, employment sector, and access to financial accounts.

Keywords: Multidimensional Poverty; Benin; Fractional regression model; Machine learning; SHAP values; ALE plots.

1 Introduction

Poverty is a widespread social problem that affects all nations, with a more severe impact on developing economies. The harmful cycle it creates makes poverty reduction a top priority of development policy. In fact, eliminating poverty is presented as Goal 1 among the 17 Sustainable Development Goals of the United Nations 2030 Agenda for Sustainable

*Corresponding author: earranhado@iseg.ulisboa.pt

Development.* Therefore, it is crucial to understand which factors are linked to poverty metrics and how they do impact them. As a complex multidimensional phenomenon, several alternative poverty measures have been introduced, including the Multidimensional Poverty Index (MPI) developed by Oxford University and the United Nations, as well as the Multidimensional Poverty Measure (MPM) by the World Bank. Furthermore, due to its complexity, poverty analysis can be challenging for conventional statistical methods, which may struggle to uncover complex patterns in the data.

The common approach to modeling poverty at an individual (person or household) level generally involves explaining and/or predicting the poverty status, which can be binary (poor/not poor) or have additional categories. As a result, regression models for binary or ordered variables (such as logit or probit, or their ordered versions) are commonly employed.† Moreover, recent studies have used machine learning (ML) models, primarily for policy targeting and prediction purposes [e.g. 19; 28].

In this paper we combine the interpretability of econometric regression models with the predictive capabilities of ML techniques to draw conclusions about the factors influencing multidimensional poverty measures and predicting poverty levels. Our empirical approach to model poverty is based on individual-level cross-sectional data from Benin, gathered from the Benin Harmonized Survey of Household Living Conditions for the period 2018-2019 (EHCVM 2018/19) and retrieved from the World Bank microdata database. According to [49], approximately 83% of Benin's population lived below $6.85 a day (2017 PPP) in 2018, with 19.9% living below $2.15 a day. Additionally, 38.5% of the population were considered monetarily poor based on the national poverty line, with food insecurity affecting 9.6% of the population and chronic malnutrition affecting 32% of young children [50]. Given these statistics, analyzing poverty and its determinants is crucial to offer valuable insights aimed at enhancing living conditions in Benin.

Our strategy involves modeling a MPM-type measure defined on the interval [0,1]. This contrasts with the common approach of modeling a poverty status, circumventing the discretization of a poverty measure that relies on an ad-hoc choice of the threshold to define each status. For a recent analysis of robustness of MPI to changes in the poverty cutoff, see [7]. The bounded and continuous nature of the MPM is described by a fractional regression model (FRM) [37; 38]. To the best of our knowledge, FRM have never been used to model a poverty measurement at the individual level, although Bluhm et al. [11] use them to model country-level poverty headcut ratios and Akire et al. [6] mention FRM in their methodological chapter on regression analysis of multidimensional

*See https://www.un.org/sustainabledevelopment/development-agenda.

†For further guidance on regression analysis of multidimensional poverty measures, refer to the recommendations in Alkire et al. [6, 7].

poverty.

Next, we estimate a random forest (RF), which is a nonparametric ML model based on an ensemble of decision trees. This is a powerful off-the-shelf model that consistently outperforms other machine learning models in a variety of tasks involving tabular data [21; 14]. The nonparametric nature of RF allows for more flexible fits to the data, as no functional form is pre-specified by the analyst. However, this greater flexibility and accuracy come at the cost of lower transparency. Indeed, RF are often considered black boxes because it is difficult to understand which covariates influence the model's output. Therefore, to understand which explanatory variables have the most impact on the multidimensional poverty measure, we need to use recent techniques for explaining black-box models, namely SHapley Additive exPlanations (SHAP) [30; 31] and Accumulated Local Effects (ALE) plots [8]. SHAP evaluates the importance of a regressor by measuring its impact on the model predictions when it is present in or absent from all possible 'coalitions' of covariates. ALE plots provide a visualization of how the target variable changes with the input variables (for instance, they indicate whether this change is positive or negative, linear or non-linear, convex or concave) and they have never been used to explain poverty.

We identify the most relevant determinants of multidimensional poverty in Benin using both the parametric and the nonparametric approach. Our results show a reasonable level of agreement between both approaches regarding the direction of the effects of the explanatory variables on poverty, despite some variables displaying highly nonlinear effects, such as the regional inequality and the age of the household head, as evidenced by ALE on random forests. On the other hand, the analysis of the importance of the regressors highlights the relevance of policies aiming to address gender inequalities in education, climate change issues, population growth, financial inclusion, food deprivation, and structural reforms to diversify the economy as part of the poverty reduction strategy of the country.

This paper is structured as follows. Section 2 provides some background discussion on the measurement of poverty, explanatory factors suggested by regression analysis, and previous ML approaches proposed in this area. Section 3 describes the employed methodology, focusing on the FRM and then addressing the ML techniques used. The dataset and the poverty landscape in Benin are detailed in Section 4. Section 5 discusses the results for both the parametric model and the nonparametric models. Finally, section 6 provides some concluding remarks.

2 Literature review

2.1 Poverty – a complex social phenomenon

According to the United Nations [43], poverty may be understood as a condition in which a person or community is lacking the basic need for minimum standard of well-being, particularly as a result of persistent lack of income. The World Bank defines poverty as a ‘pronounced deprivation in well-being’ [23]. Quantifying poverty to reflect well-being can be challenging. Previous approaches defined well-being in monetary terms by setting an income/expenditure threshold (poverty line) below which households/individuals are considered poor. However, this approach has been acknowledged to be imperfect in understanding the deprivations of the poor [41]. This recognition has led to the consideration of non-monetary dimensions, such as subjective poverty, health poverty, and education poverty, as well as multidimensional approaches.

The most well-known multidimensional poverty measures include the Multidimensional Poverty Index (MPI), developed by Oxford University and the United Nations. The MPI assesses severe multidimensional poverty in developing countries across the dimensions of health, education, and standard of living, with households as the unit of identification [7]. Additionally, the World Bank’s Multidimensional Poverty Measure (MPM), inspired by the MPI, evaluates the percentage of deprived households in a country based on dimensions such as monetary poverty, education, and basic infrastructure services, also using households as the unit of identification [47; 17]. The exclusion of the monetary poverty dimension from the MPI is partly because income data are not typically included in the surveys used, as well to avoid the risk of ‘double-counting’ deprivations [42]. While the MPI is designed to complement monetary poverty measures, the MPM addresses the interplay between monetary and non-monetary deprivations, offering a more comprehensive view of poverty.

An additional challenge in poverty analysis arises from the complexity of poverty’s underlying causes. Developing a clear understanding of poverty’s fundamental roots is difficult, prompting researchers to focus on the more immediate or ‘proximate’ causes of poverty [23]. Within this framework, poverty may be caused by or at least correlated with various factors at different levels: regional factors (such as remoteness, inequality, regional governance and management, weather, and environmental conditions); community factors (including infrastructure availability, land distribution, access to public goods and services); household characteristics (like household size, dependency ratio, employment and income structure, average household health/education); and individual characteristics (such as age, education, employment status, health status, ethnicity).

2.2 Determinants of poverty in Benin

Empirical studies conducted in Benin indicate that the key factors influencing poverty align with findings observed across Africa. Hodonou et al. [24] highlight the sensitivity of poverty dynamics in the country to various factors, including the age and gender of household heads, household size, place of residence, possession of durable goods, improved access to housing, electricity, communications, and education. Education and household size are important determinants of poverty. Research suggests that the impact of education on poverty is consistent across gender and place of residence [9]. Households led by educated individuals are found to face lower poverty risks [4; 20], and education is crucial in determining the time needed to escape poverty [3]. Conversely, larger household sizes tend to worsen household welfare, influencing the transition in poverty status [24; 3; 1].

Geographical factors also play a significant role, as households residing in cotton or rice-producing regions have a higher likelihood of escaping poverty [1]. However, shocks such as reductions in cotton prices can lead to an increase in poverty [34], and employment in the agricultural sector has been linked to a decline in well-being [3; 1]. Some studies suggest that the employment sector is one of the most important determinants of the likelihood of poverty [3; 1], with Alinsato and Houedokou [5] emphasizing the influence of unobserved factors related to labor market participation in determining poverty. The impact of the gender of household heads produces mixed findings. Some studies suggest a decreased likelihood of poverty for female-headed households [9; 5], while others indicate the opposite [4; 3; 20]. Moreover, biophysical shocks to households also contribute to an increased risk of poverty [20].

Most of these studies examined the determinants of a monetary poverty measure and used parametric models, such as logit/probit, OLS, and multinomial or ordered logit/probit, for statistical inference.

2.3 Machine learning in poverty analysis

Empirical research using machine learning models for poverty analysis has primarily focused on accurately targeting deprived individuals or households, with prediction as the main goal. An example is Thoplan [40], who trained a random forest to predict the poverty status of individuals in Mauritius. The mean decrease in accuracy was used to identify the most important predictive regressors, such as the number of hours worked, age, education, and gender. Mcbride and Nichols [32] examined the out-of-sample performance of quantile regression, probit regression, random forest, and quantile random forest in classifying the poverty status in Bolivia, Timor-Leste, and Malawi. Their findings indicated that stochastic ensemble methods resulted in improved accuracy compared to traditional methods. Sohnesen and Stender [39] used consumption expenditure surveys to compare the out-of-sample error of variable selection methods (stepwise selection and LASSO) and random forests in predicting poverty in Albania, Ethiopia, Malawi, Rwanda, Tanzania, and Uganda. They concluded that random forests often achieve higher accuracy. However, none of the methods consistently delivered accurate predictions of poverty over time. Engstrom et al. [18] examined poverty in urban slums in Accra, Ghana, by creating a slum index with the help of random forest. They estimated poverty at the neighborhood level by first using the LASSO estimator to identify key variables and then selecting the final model through a stepwise procedure. Their findings revealed a strong correlation between living in slums and higher monetary poverty, increased fertility among women, and lower school attendance among children. Additionally, they found that poverty is more common in communities located in lower elevation areas, which in Accra are typically flood-prone regions. Fitzpatrick et al. [19] predicted the poverty status of households in Malawi and Indonesia using linear regression and a diverse set of machine learning algorithms. The out-of-sample performance of the models led them to conclude that there are significant benefits to using machine learning approaches over simple regressions. They also found that ensemble methods consistently produced more accurate models.

In their study, Liu et al. [29] applied the Lindeman, Merenda, and Gold method to multiple linear regression and the permutation-based error reduction on random forest to determine the relative importance of various geographic factors in predicting poverty at the village level in Yunyang, rural China. They found that both models exhibit similar accuracy, and consistently highlight access to industry, access to banks, and access to town as the most important predictors. Bakar et al. [10] predicted poor households in Malaysia using linear regression, decision trees, and random forests. Random forests had the best accuracy, with the top five predictors identified as per capita income, state, ethnicity, strata, and religion, based on permutation variable importance. Li et al. [28] forecast household poverty in Kyrgyzstan using extreme gradient boosting, LASSO, and ridge regressions with varying sets of variables. Their findings indicate that extreme gradient boosting generally outperforms the other methods, and they suggest that including more variables may not necessarily be advantageous for prediction accuracy. Finally, Min et al. [33] predicted individuals poverty for Costa Rica using linear regression, decision trees and random forests. They concluded that random forest performed better, with the most important covariates being the dependency, walls, and average education, as indicated by SHAP values.

3 Methodology

3.1 Fractional regression model

The MPM is bounded to the interval [0,1]. Therefore, a fractional regression model, estimated using quasi-maximum likelihood [37], can be used to analyze the determinants of multidimensional poverty in Benin. The model conditional mean is

$$E(y_{ih} \mid \boldsymbol{x}_{ih}) = G(\boldsymbol{x}_{ih}^T \boldsymbol{\beta}), \tag{1}$$

where $y_{ih} \in [0,1]$ is the MPM for individual i and household cluster h, $\boldsymbol{x}_{ih}$ is a vector of exogenous regressors including the intercept, $\boldsymbol{\beta}$ is a vector of coefficients, and $G(\cdot)$ is a functional form satisfying $0 < G(\cdot) < 1$. The quasi-maximum likelihood method of estimation is based on the maximization of the Bernoulli log-likelihood function:

$$LL \equiv \sum_{i=1}^{N} \sum_{h=1}^{H} y_{ih} \ln \left[G(\boldsymbol{x}_{ih}^T \boldsymbol{\beta}) \right] + (1 - y_{ih}) \ln \left[1 - G(\boldsymbol{x}_{ih}^T \boldsymbol{\beta}) \right], \tag{2}$$

where N is the number of individuals and H the number of households. The Bernoulli quasi-maximum likelihood estimator of $\boldsymbol{\beta}$ is consistent and asymptotically normal, provided that only the conditional mean $G(\cdot)$ is correctly specified, regardless of the distribution of y conditional on $\boldsymbol{x}$. We use a robust variance-covariance matrix to account for any within-cluster correlation due to the household clustering effects on individuals in the sample, while assuming independence between clusters, i.e., no intercluster correlation [46]. A cluster-robust standard error is also a heteroskedastic-robust standard error [13].

The link function $G(\cdot)$ is a probit functional form, representing a standard normal cumulative distribution function $G(\boldsymbol{x}_{ih}^T \boldsymbol{\beta}) = \Phi(\boldsymbol{x}_{ih}^T \boldsymbol{\beta})$. Since $G(\cdot)$ is strictly monotonically increasing, the sign of the partial effect is determined by the sign of $\hat{\boldsymbol{\beta}}_j$. The estimated partial effects of a continuous variable x_j and a dummy variable x_k for an individual i are, respectively, $\widehat{PE_{ij}} = \hat{\beta}_j \frac{\partial \Phi(\boldsymbol{x}_{ih}^T \hat{\boldsymbol{\beta}})}{\partial x_{ihj}}$ and $\widehat{PE_{ik}} = [\Phi(\boldsymbol{x}_{ih,x_{k=1}}^T \hat{\boldsymbol{\beta}}) - \Phi(\boldsymbol{x}_{ih,x_{k=0}}^T \hat{\boldsymbol{\beta}})]$. The conditional mean of the fractional probit model may be tested using RESET tests. In this paper, the test is applied in the versions that adds up to two fitted powers of the linear index, $(\boldsymbol{x}^T \boldsymbol{\beta})^2$ and $(\boldsymbol{x}^T \boldsymbol{\beta})^3$. The null hypothesis is $H_0 : E(y \mid \boldsymbol{x}) = G(\boldsymbol{x}^T \boldsymbol{\beta})$, and a Wald test for the joint significance of the two fitted powers added to the model in test is implemented.

3.2 Random forest model

A random forest [12] is a straightforward yet powerful technique for aggregating multiple individual decision trees. A decision tree is a nonparametric model that partitions the regressor space into distinct and non-overlapping regions $\{R_m\}_{m=1}^M$. The specific region R_m to which an observation belongs is determined by a series of if-then-else tests conducted on the regressor values $\boldsymbol{x}$. Formally, a decision tree model can be expressed by the equation:

$$f(\boldsymbol{x}; \boldsymbol{w}) = \sum_{m=1}^{M} w_m \cdot \mathrm{I}(\boldsymbol{x} \in R_m), \tag{3}$$

where $\mathrm{I}(\boldsymbol{x} \in R_m)$ is an indicator function that yields 1 when its argument is true and 0 otherwise. The weights w_m represent the model's output for observations that belong to region R_m. They correspond to the average y-value of all observations within the estimation data that fall into that particular region:

$$w_m = \frac{\sum_{i=1}^{N} y_i \cdot \mathrm{I}(\boldsymbol{x}_i \in R_m)}{\sum_{i=1}^{N} \mathrm{I}(\boldsymbol{x}_i \in R_m)}. \tag{4}$$

Since MPM is bounded to the interval [0,1], both w_m and $f(\boldsymbol{x}; \boldsymbol{w})$ are as well.

A random forest is a collection of individual decision trees. First, a specified number of bootstrap samples are generated from the data, each comprising the same number of observations as the original dataset. Then, a decision tree is constructed for each of these bootstrap samples. However, at each step of dividing the data into regions R_m, only a random subset of the regressors is taken into consideration. This increases the diversity among the decision trees. Let's suppose that we have generated B bootstrap samples and let $f_b(\boldsymbol{x}; \boldsymbol{w})$ denote a decision tree estimated on a specific bootstrap sample. The prediction provided by a random forest is the average of the individual predictions given by the B trees:

$$f(\boldsymbol{x}; \boldsymbol{w}, B) = \frac{1}{B} \sum_{b=1}^{B} f_b(\boldsymbol{x}; \boldsymbol{w}). \tag{5}$$

Since $f_b(\boldsymbol{x}; \boldsymbol{w})$ is constrained to the interval [0,1], $f(\boldsymbol{x}; \boldsymbol{w}, B)$ is as well. That is, the predictions for MPM will be constrained within the unit interval as intended. Despite its simplicity, the random forest is a powerful out-of-the-box model, consistently surpassing other machine learning models in many tasks involving tabular data [21; 14].

3.3 Explaining black box models

Random forest models are 'black boxes'. Therefore, we use ALE plots and SHAP values to understand how regressors influence multidimensional poverty based on the random forest model. These methods are preferred over permutation feature importance, partial dependence plots, and marginal plots when the covariates are not independent, as is the situation here [35].

3.3.1 Accumulated local effects (ALE) plots

ALE plots overcome the lack of interpretability of black box models by visually describing the effect of a regressor on the predicted response [8], allowing the researcher to have an intuition on how the regressor impacts the prediction of the dependent variable and so to infer if the relationship may be positive or negative, linear or non-linear, concave or convex, and so on.

The estimated accumulated local effect for regressor x_j is computed by first segmenting the range of values of x_j into K intervals or bins. For $k = 0, 1, \ldots, K$, the interval boundary values $Z_{k,j}$ are the k/K-quantiles of the empirical distribution of x_j, where $Z_{0,j}$ is chosen just below the smallest observation of the regressor and $Z_{K,j}$ is chosen as the largest observation. The formula for the uncentered effect is given by

$$\widehat{ALE}(x_j)_U = \sum_{k=1}^{k_j(i)} \frac{1}{n_j(k)} \sum_{\{i: x_{ji} \in N_j(k)\}} \left\{ f\left(Z_{k,j}, \boldsymbol{X}_{\setminus j}^{(i)}\right) - f\left(Z_{k-1,j}, \boldsymbol{X}_{\setminus j}^{(i)}\right) \right\}, \tag{6}$$

where $k_j(i)$ is the index of the interval into which falls a x_j's value of the ith observation; $n_j(k)$ is the number of training observations falling into the kth interval $N_j(k)$, so that $\sum_{k=1}^{K} n_j(k) = n$; $\boldsymbol{X}_{\setminus j}^{(i)}$ is the set of values of the other features when observation value x_{ji} is considered; $f\left(Z_{k,j}, \boldsymbol{X}_{\setminus j}^{(i)}\right)$ is the model prediction with x_j equal to the upper limit of the interval (bin); and $f\left(Z_{k-1,j}, \boldsymbol{X}_{\setminus j}^{(i)}\right)$ is the model prediction with x_j equal to the lower limit of the bin. All possible differences in the response predictions are averaged and then accumulated over the grid. The $\widehat{ALE}(x_j)_U$ is centered so that the mean effect is zero,

$$\widehat{ALE}(x_j)_C = \widehat{ALE}(x_j)_U - \frac{1}{n} \sum_{i=1}^{n} \widehat{ALE}(x_{ji})_U. \tag{7}$$

Plotting $\widehat{ALE}(x_j)_C$ versus x_{ji} reveals the effect of x_j on the predictive function for y.

3.3.2 SHAP values

The Shapley value is a concept from cooperative game theory to fairly distribute the final payout among players who cooperated in a coalition to obtain that payout, as some players contribute more than others [35]. In machine learning context [30], regressors represent the players and prediction represents the payout in the regression analysis. The SHAP value for a x_j's value is the weighted sum of its marginal contribution to the prediction $\hat{y}$ across all possible coalition of regressors that exclude it, meaning that the algorithm allows to know by how much a regressor's value contributed to the prediction. Given the full set of P regressors ($\boldsymbol{X}$), the set excluding x_j is $\boldsymbol{X}_{\setminus j}$, all possible subsets of $\boldsymbol{X}_{\setminus j}$ are denoted S (i.e., $S \subseteq \boldsymbol{X}_{\setminus j}$), and the formula for the SHAP value ϕ_j, for a x_j's value, can be written as

$$\phi_j = \sum_{S \subseteq \boldsymbol{X}_{\setminus j}} \frac{|S|!(|P| - |S| - 1)}{|P|!} \left[f_{S \cup x_j}(\boldsymbol{X}_S \cup x_j) - f_S(\boldsymbol{X}_S) \right], \tag{8}$$

where f_S is the model trained without x_j and $f_S(\boldsymbol{X}_S)$ is the prediction for feature values in set S that are marginalized over features that are not included in set S [35]; $f_{S \cup x_j}$ is the model trained including x_j and $f_{S \cup x_j}(X_S \cup x_j)$ is the prediction for feature values in set $S \cup x_j$). To obtain ϕ_j all possible differences $[f_{S \cup x_j}(\boldsymbol{X}_S \cup x_j) - f_S(\boldsymbol{X}_S)]$ must be computed. SHAP values are used as feature attribution, where regressors with large absolute SHAP values are important and the global importance (I_j) for a covariate x_j is derived as the average of the absolute SHAP values across the data:

$$I_j = \frac{1}{n} \sum_{i=1}^{n} |\phi_j^{(i)}|. \tag{9}$$

When calculating SHAP values, we standardize numerical variables into Z-scores and perform one-hot encoding for categorical variables. To compare these results with those of the fractional regression model, we adjust the APE by standardizing the numerical variables and focus on absolute APE values.

4 Data

In this study, we used the Benin Harmonized Survey of Household Living Conditions 2018-2019 (EHCVM 2018/19), obtained from the World Bank microdata database [48]. The survey employed a two-wave approach to address the seasonality of consumption and utilized the 2013 Census of Population and Housing as the sampling frame. The dataset comprises 8,012 households, representing 42,343 individuals. Poverty is assessed using an MPM-type measure, calculated for Benin following the methodology outlined in [17]. This MPM integrates the monetary dimension with two non-monetary dimensions, education and basic infrastructures, with the household serving as unit of identification. The MPM value is a weighted sum of the individual indicators, where higher values indicate more deprived individuals or households, consistent with the standard practice in the literature. A detailed explanation of the MPM construction is available in the Appendix.

Figure 1 shows the distribution of MPM for the sample individuals. The MPM has a mean and median of 0.51 and 0.5, respectively, with a minimum value of 0 and a maximum value of 1, representing approximately 5.2% and 2.4% of the sample observations, respectively. Approximately 18.0% of the sample individuals have an MPM below 0.25, while 21.4% have an MPM above 0.75.

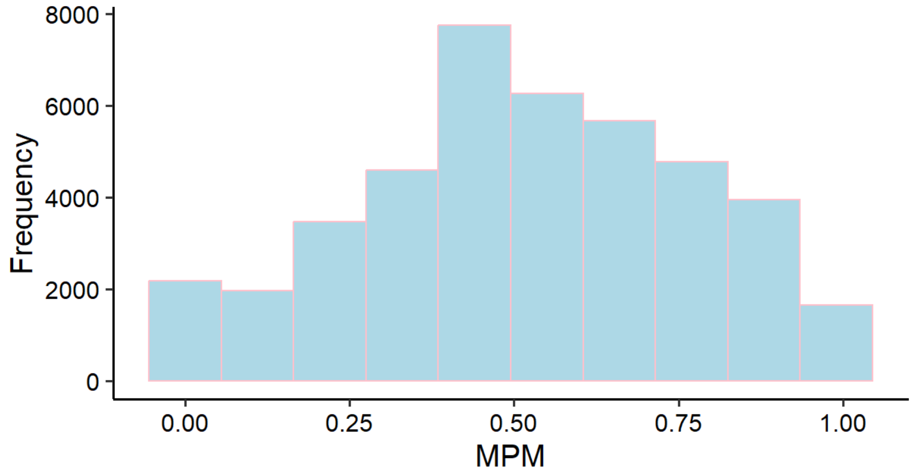

Figure 1: Distribution of Multidimensional Poverty Measure for the sample individuals.

The database contains 231 variables. Of these, 73 correspond to the original data in the survey dataset, 37 are new variables derived from available information in the survey dataset, and 121 are adjusted survey responses. These adjustments include creating categorical variables that aggregate some of the original categories and imputing blank cells with implicit information from other response variables in the survey dataset. The 231 variables cover regional, community, household, and individual characteristics, constituting a more extensive set of covariates than those considered in prior research on Benin (which varies from 5 to 30 on the reviewed literature). This comprehensive set of determinants is not only informed by existing literature in this field but also seeks to capture specific attributes of the Beninese population. The country's economy relies heavily on agriculture, which accounts for approximately 70% of employment and 30% of GDP. This sector is highly dependent on rainfall and is vulnerable to climate change, as well as fluctuations in global cotton and oil prices [2; 49]. The rapid pace of population growth poses a challenge, as the increasing number of births exerts pressure on the economy. This situation leads many young people to migrate from rural to urban areas in search of employment opportunities [25]. Despite a very low official unemployment rate of 2.4% in 2019, underemployment and informal employment rates are significantly high, standing at 72% and 90.1% respectively in the same year [49].

The potentially most important explanatory factors of poverty in Benin are summarized out of 231 features in a more restricted dataset containing the explanatory variables of Table 1. This selection was chosen in order to produce a fractional regression model statistically valid, according to the RESET test, and include typical determinants considered in poverty explanation, such as the education, gender, employment sector, marital status and age of the household head, the household size, child dependency ratio, place

Variable	Description	# categories
head_age	Household head age	
hh_size	Household size	
hh_depratio_c	Household child dependency ratio	
c_inequality	Sub-regional Gini index of per capita expenditure	
geo_aez	Agro_ecological zone	5
geo_urbrur	Place of residence	2
head_sex	Household head gender	2
head_educ	Household head education	4
ind_edu_mother	Household mother's education	4
head_emp_sector12m	Household head employment sector/status	5
head_mstat	Household head marital status	5
ind_ethnic	Household member ethnic group	6
ind_migration	Household member previous place of residence	4
hh_trf_receive	Remittances from non-household members	2
hh_fin_access	Household access to financial account/prepaid card	2
hh_shk_severe_1	Most severe shock	9
c_roadac	Main road access	5
c_electric	Electric distribution network at community	2
c_water	Running water network at community	2
c_healthcom	Health Committee at community	2
c_schoolcom	School Committee at community	2

Table 1: Description of selected regressors. The last columns shows the number of categories for the categorical variables.

of residence, inequality, and those related to access to services and infrastructures, and determinants emerging from Benin specificities, such as ethnicity groups, agroecological zone, and household's most severe shocks. The selection also includes the education of the mother of each individual in the sample to explore the inter-generational effects of female education on poverty.

Variable	Mean	Median	SD	Min	Max
Household head age	44.4	43	13.2	15	98
Household size	7.0	6	3.6	1	26
Household child dependency ratio	125.3	100	95.8	0	700
Sub-regional Gini index	0.33	0.33	0.05	0.20	0.43

Table 2: Descriptive statistics for the numerical explanatory variables.

Descriptive statistics for the numerical explanatory variables are shown in Table 2. Significant variability is observed in the age of the household head, ranging from 15 to 98, and in the household size, ranging from 1 to 26.

Figure 2 presents box plots of MPM for the categorical determinants. Interesting patterns emerge for MPM based on the categorical variables. While certain binary variables like the gender of the household head and the presence of a health or school committee exhibit similar distributions for MPM, higher levels of education, whether of the household

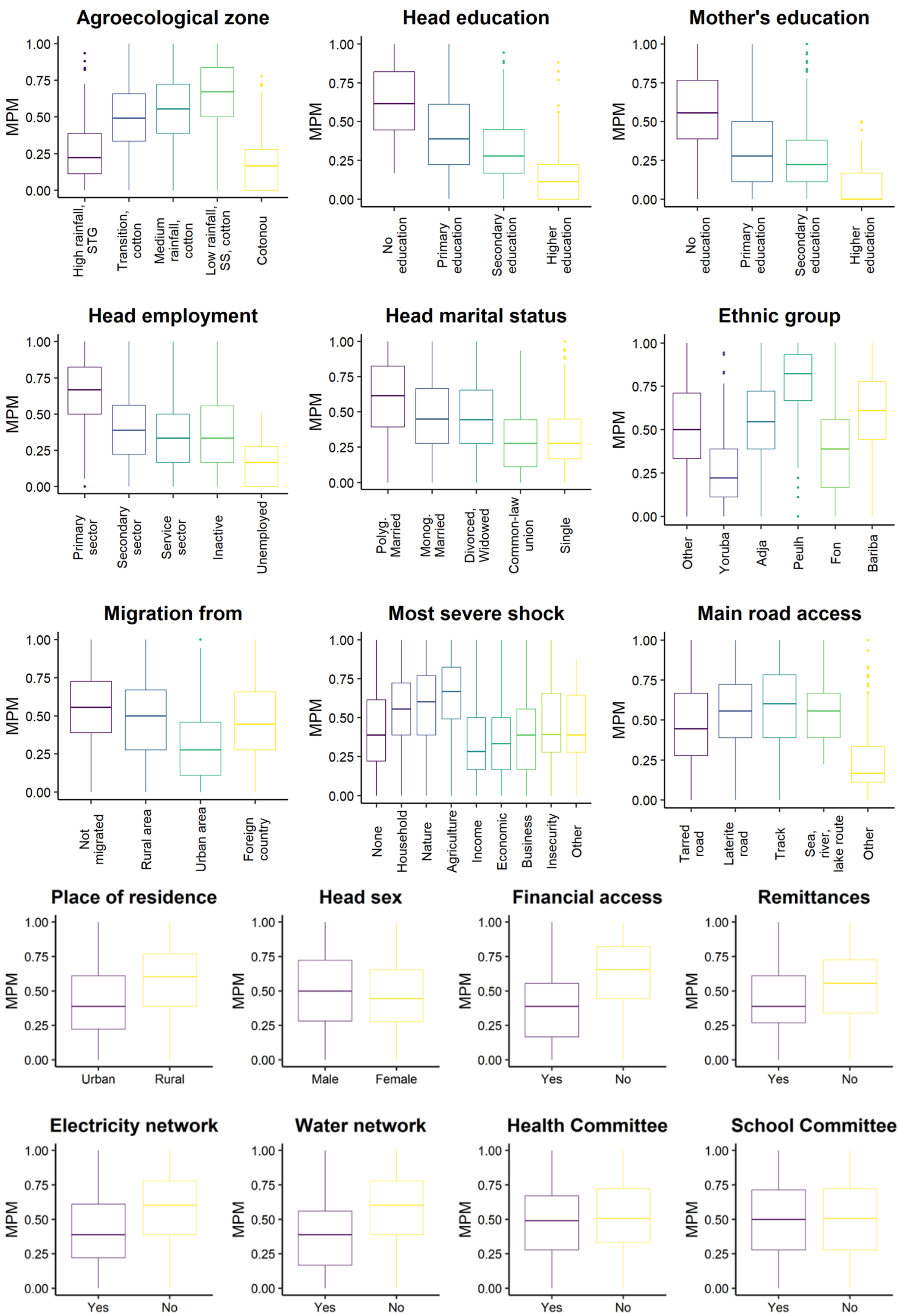

Figure 2: Distribution of categorical variables by Multidimensional Poverty Index (MPM). *Note*: STG = subequatorial-tropical-guinean; SS = Sudan-Sahelian

head or the mother, and residing in an urban area are distinctly linked to lower levels of deprivation. On the other hand, households characterized by the head working in agriculture, having a polygamous marital status, or belonging to the Peulh ethnic group exhibit high levels of deprivation. This is likely influenced by Benin's heavy reliance on agriculture, where any shocks in this sector are associated with increased deprivation.

Natural disasters appear to be closely linked to poverty in Benin. Additionally, the strong correlation between low rainfall areas and increased levels of deprivation underscores the significance of addressing climate change issues in the country's poverty reduction policies. An intriguing finding is the association of unemployment with lower levels of deprivation. This phenomenon may be attributed to the precarious state of social protection in Benin. Poorer families, especially those without strong financial support networks and residing in rural areas with low rainfall, often resort to underemployment for survival [16]. Consequently, they may not be able to afford being officially classified as inactive or unemployed, leading to a biased correlation between MPM and employment status.

5 Empirical results

5.1 Out-of-sample performance

We analyze the predictive accuracy of the models based on the root mean squared error (RMSE), mean absolute error (MAE), and the R^2. Table 3 presents the out-of-sample accuracy metrics obtained through a 10-fold cross-validation.

Model	**RMSE**	**MAE**	R^2
Fractional regression model (selected regressors)	0.162	0.131	0.622
Random Forest (selected regressors)	0.059	0.042	0.957
Random Forest (complete set of regressors)	0.034	0.018	0.985

Table 3: Out-of-sample model accuracy given by a 10-fold cross-validation.

In the dataset with selected regressors, as expected, the random forest model outperforms the FRM in all metrics in terms of predictive accuracy. The random forest model achieves an RMSE of 0.059 and MAE of 0.042, along with a high R^2 of 95.5%, compared to 0.162, 0.131, and 62.2%, respectively, for the FRM. The out-of-sample predictive performance of the nonparametric model improves when we include a larger number of predictors in the dataset, allowing for a more accurate poverty-targeting approach. In this scenario, the random forest RMSE and MAE decrease to 0.034 and 0.018, respectively, and can account for approximately 98.5% of the variability in MPM.

5.2 Interpreting the models

The estimated FRM, presented in Table 4, was validated by the RESET test. The estimated model coefficients are presented in Table A.2 in the Appendix and the APE are displayed on Table 4 and Figure 3. In this model, age and household size, as practice in the literature, are nonlinearly included, as well as the child dependency ratio. In order to allow some partial effects to differ according to the gender of the head of the household, several interaction terms involving this binary variable were considered. Interactions involving the household head's employment sector/status and infrastructure variables were also included. The ALE plots for the nonparametric model are summarized in Figure 4.

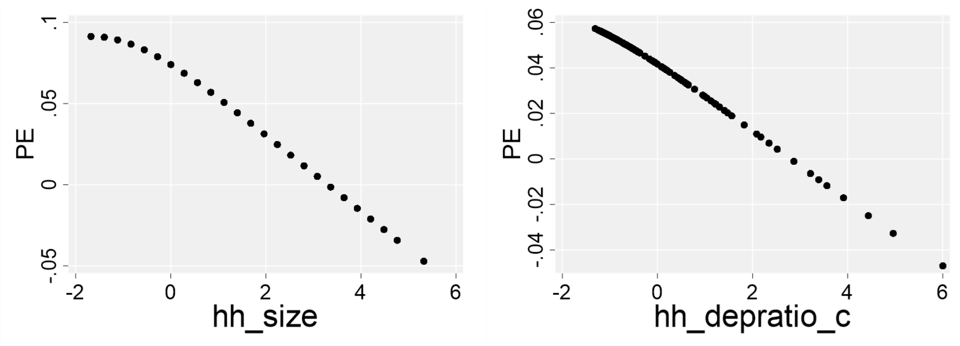

Figure 3: Nonlinear Partial Effects (PE) for FRM. *Note*: The x-axis displays standardized values.

Individual statistical significance tests for the FRM suggest that most of the potential determinants of the MPM are significant. These significant variables coincide with ALE plots displaying a variety of shapes (Figure 4) that are far from being flat. In the FRM, age (head_age) and road access (c_roadac) are not significant. However, they display a nontrivial relationship with MPM in the ALE plots. This mixed evidence in terms of relevance is corroborated by the analysis of variable importance, where none of these two variables appear in the top 5 most important variables, as expected.

The numerical variables display a nonlinear behavior, as evidenced by the ALE shapes in Figure 4. As for the age of the household head and inequality (c_inequality), the effects are highly nonlinear, with a negative impact on well-being when household heads are younger than 25 and a nonmonotonic negative impact for those older than 62, while Gini indices greater than 0.35 consistently display a negative effect on deprivation. The ALE plot has a positive slope for lower levels of inequality, which can be explained by subregions with a predominance of poor people at similar levels of poverty. The APE for inequality also indicates a negative effect on well-being of higher Gini index values. The ALE plots further shows a monotonic decline in well-being for household size up to 16,

Variables	Global APE	Interaction APE head_sex Male	Interaction APE head_sex Female
head_age	0.000		
hh_size	0.066***		
hh_depratio_c	0.037***		
c_inequality	0.006**		
geo_aez (base = High rainfall...)			
Transition, cotton...	0.097***		
Medium rainfall, cotton...	0.109***		
Low rainfall, cotton...	0.133***		
Cotonou	-0.031**		
geo_urbrur (base = Urban)	0.014***		
head_sex (base = Male)	-0.019**		
head_educ (base = No education)			
Primary education	–	-0.05***	-0.172***
Secondary education	–	-0.095***	-0.195***
Higher education	–	-0.207***	-0.274***
ind_edu_mother (base = No education)			
Primary education	–	-0.092***	-0.027***
Secondary education	–	-0.11***	-0.051***
Higher education	–	-0.209***	-0.12***
head_emp_sector12m (base = Primary sector)			
Secondary sector	-0.041***		
Service sector	-0.063***		
Inactive	-0.046***		
Unemployed	-0.124**		
hh_trf_receive (base = Yes)	–	0.014**	-0.008
hh_fin_access (base = Yes)	0.09***		
c_electric (base = Yes)	0.015**		
c_water (base = Yes)	0.037***		
c_healthcom (base = Yes)	0.002		
c_schoolcom (base = Yes)	0.014***		
N	42343		
Log pseudolikelihood	-17984		
Wald Chi^2 for global significance (p-value)	0.000		
RESET Test $J = 2$ (p-value)	0.053		

Table 4: Average partial effects (APE) for a selected group of covariates.

Note: ***$p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.1$. Numerical variables are scaled into Z-scores. APE for relevant nonlinearly included variables are detailed in Figure 3. APE of determinants with multiple categories where most of the categories are not significant at the 5% level are not displayed. This is the case of c_roadac (no relevant categories), hh_shk_sevare_1 (only nature, agriculture and economic shocks are relevant), ind_migration (only from urban area is relevant), head_mstat (only single is relevant), and ind_ethnic (only Yoruba, Adja and Peulh are relevant). For c_healthcom and c_electric, included with interactions for c_water, APE are displayed for absence of water network. Complete results are available from the authors upon request.

beyond which the deteriorating effect lessens slightly, while the FRM suggests an inversion of the partial effects at size 19 (Figure 3), very close to that of ALE. These results are in line with the literature that suggests some scale economy for higher household size [e.g. 27]. Additionally, the ALE plots suggest a nonlinear decline in well-being as child dependency increases, with some beneficial effects for ratios higher than 250 and up to 350, after which the effects become more severe. Due to its specified quadratic functional form, the FRM only suggests beneficial partial effects after the ratio of 400 (Figure 3).

Both parametric and nonparametric models indicate that individuals in female-headed households (head_sex) experience greater well-being, as shown by the APE and the ALE plots. This variable is interacted with the variables measuring education (head_educ) and the variable indicating the reception of financial support (hh_fin_access).

Increasing education of either the head of the household or the mother leads to an increasing reduction of deprivation, with more pronounced effects for female (male) heads of household in the former (latter) measurement of education. Interestingly, APE suggest that differences between the effect of education according to the head of household gender are attenuated as the level of education increases. Notice that for education of the head of household (the mother) at the primary level, being a female (male) head of household promotes a reduction of MPM approximately 3 times higher than in male (female) headed household. For secondary education, one partial effect doubles the other, and for higher education, the effect of both genders becomes similar, concerning the head of household education, while the intergenerational effect of female education, for male and female household heads, is almost the double for that of males. This suggests that the intergenerational benefits of having an educated mother offer greater impact for male headed households, with the difference for female headed households remaining high as education rises. Conversely, education of the head of household offers more benefits for female headed households, but the differences relative to males tend to vanish at a higher rate as the level of education is improved.

APE and ALE plots suggest that working in an economic sector other than the primary sector improves the welfare condition in Benin. While this result is in line with literature, unintuitively, households with inactive and unemployed head of household exhibit negative effects on MPM. The association of unemployment with higher welfare was identified in the descriptive statistics of section 4: heads in poorer households might resort to underemployment activities in subsistence agriculture for survival. On the other hand, the absence of access to the financial sector increases deprivation, as expected. Another positive, and thus adverse effect on MPM emerges from the absence of remittances (hh_trf_receive), but only in households led by men. Households led by women remain unaffected.

Natural conditions of the region have significant effects on MPM. The results suggest that living in a low rainfall, Sudan-Sahelian area, as opposed to high rainfall regions, increases deprivation, as it has the highest worsening effect on welfare among the agroecological zones, despite the production of cotton in the area. This adverse effect is even more pronounced than the negative impact of merely residing in rural areas. Concerning migration, it appears that individuals who migrate from urban areas are less likely to experience poverty.

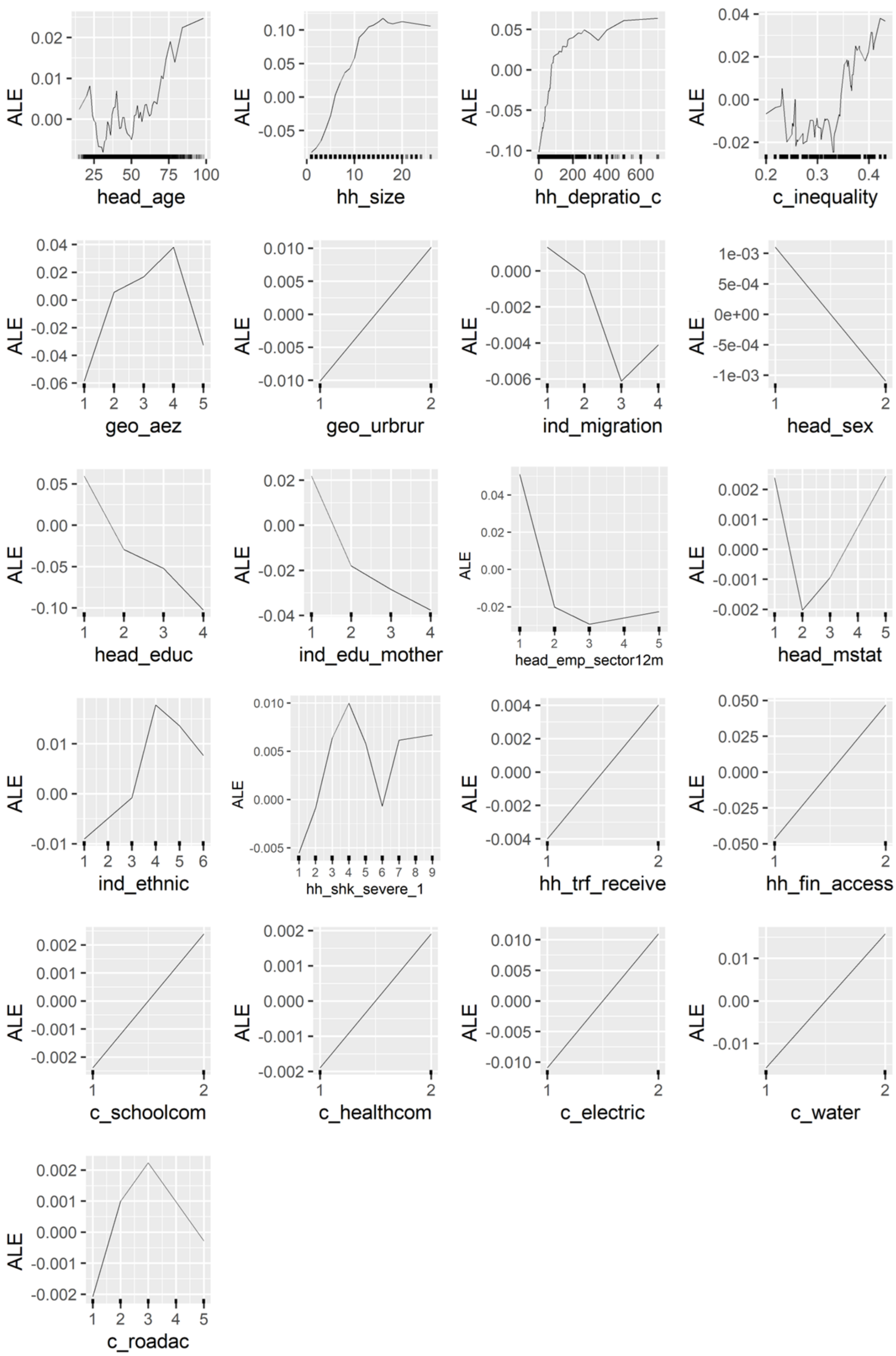

Figure 4: ALE Main Effects of Regressors on MPM. *Note*: For binary variables: 1 = Yes, Urban or Male; and 2 = No, Rural, or Female.

Finaly, the ALE plots indicate that polygamous and single household heads are associated with worse welfare conditions, whereas FRM exhibit a negative APE for households led by a single head, relative to polygamous household heads. Also, both models corroborate that being classified in the Peulh ethnic group has the highest negative effect on the individuals' well-being, while the Yoruba ethnic group is associated with better outcomes. Additionally, an agricultural shock has the highest negative impact on well-being among various shocks, followed by nature shocks, and individuals in communities without a school committee are more likely to face lower well-being. Both models also corroborate that the absence of basic public infrastructures, such as water and electricity networks in the community leads to higher level of deprivation.

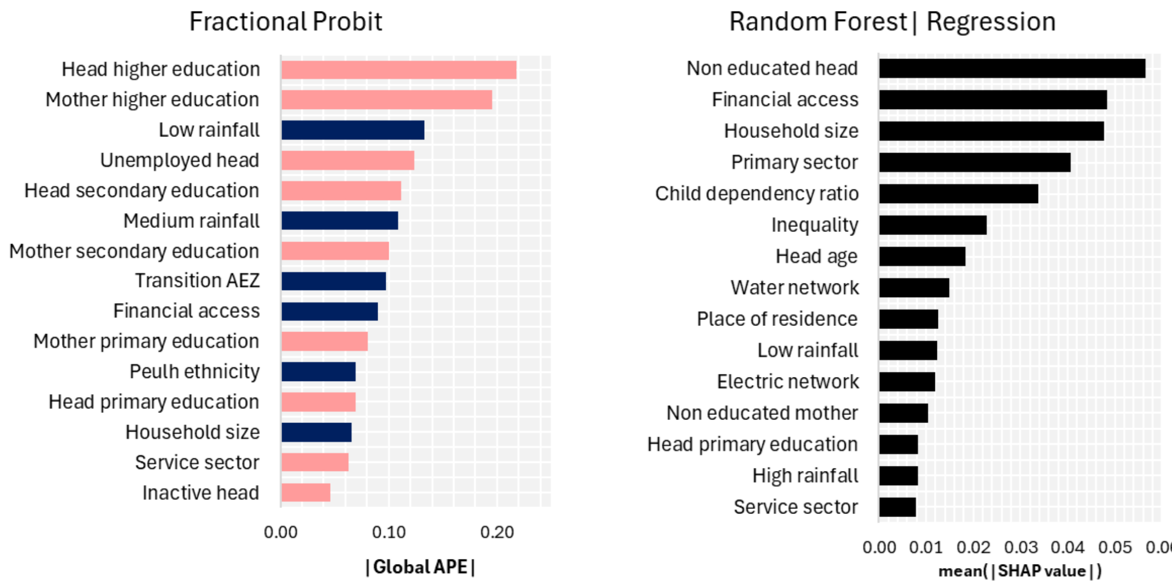

Figure 5: Variable Importance given by absolute APE and SHAP values. *Note*: Red bar = negative APE; blue bar = positive APE; black bar = no direction of effect.

5.3 Variable importance

Concerning the variables importance, Figure 5 depicts the most important regressor effects for each model. Both models indicate that education (of the household head and individual's mother), financial access, agroecological zone, household size, and employment sector are among the most important variables. Variables with highly nonlinear effects (inequality and age) are ranked among the top 15 most important variables only in the random forest model. In addition, only the random forest models deemed the nonlinear effect of the child dependency ratio to be among the most important. Moreover, this nonparametric model, by computing more complex interactions between variables, was able to present a wider range of important variables. On the other hand, only the parametric model considers ethnicity and employment status among the top 15 factors.

Finally, changing the focus to a poverty-targeting approach, Figure 6 depicts the 15 most important indicators of the poverty profile in 2018/2019 for Benin. By considering the complete set of 231 potential determinants, one covers a variety of possible factors, which the parametric regression model, due to the need to use a restricted group of regressors, does not incorporate. Some of the most important regressors examined in the context of the restricted set of regressors are still suggested in the outcomes of the poverty profile, such as household size, education (of both household head and individual's mother), child dependency ratio, financial access, and employment sector.

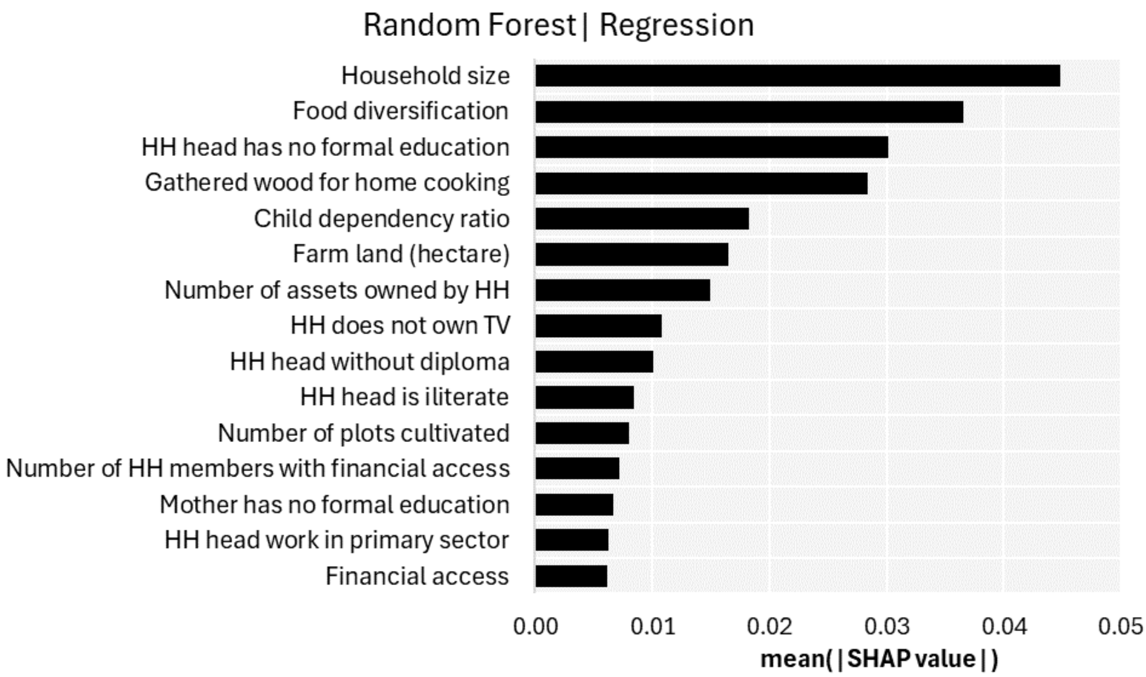

Figure 6: Most important indicators of poverty profile of Benin according to SHAP values for 2018/2019. *Note*: HH = Household.

The five most important indicators of the poverty profile, based on the large dataset, are household size, food diversification, head of the household without formal education, households that gather wood for home cooking and the child dependency ratio. Household size and dependency ratio reflect the structural characteristics of the household. Large families with many children are naturally at greater risk of deprivation. On the other hand, education is one of the main determinants of poverty, strongly present in previous literature and in the FRM estimated for Benin. Food diversification and the fact that the household collects firewood for cooking are indicators linked to food deprivation, which can be considered one of the most visible and dramatic faces of poverty. This is why the end of hunger is the second Goal in the list of 17 Sustainable Development Goals of the United Nations 2030, following Goal 1 of ending poverty. Furthermore, since the country is heavily reliant on agriculture and poorer families are dependent on subsistence agriculture, it is not surprising that farm land possession and the number of plots cultivated appear in the top 10 indicators.

6 Discussion and concluding remarks

This paper uses a fractional regression model as a new econometric tool to analyze multidimensional poverty phenomena and complements its results with the outcomes from a nonparametric machine learning model. For the first time in poverty analysis, ALE plots are used to understand the potential determinants of poverty according to a machine learning model. The machine learning model further addresses poverty prediction, the traditional issue of interest of the literature in this area, using a large set of covariates.

Building on individual cross-sectional data, our analysis of multidimensional poverty in Benin suggests that, in most cases and with a limited set of regressors, both parametric and nonparametric models generally convey the same insights in explaining poverty. However, as expected, the predictive performance of the random forest models surpasses that of the fractional regression model. The direction of the effects validates the use of machine learning techniques to provide insights into the determinants of poverty, beyond the advantage of improved predictive performance. Simultaneously, it confirms the robustness of fractional regression models in providing a comprehensive view of the relationship between potential explanatory factors and the poverty measure.

Complex, nonlinear effects are evident in the ALE shapes based on random forest analysis, particularly for inequality and the age of the head of household. These effects are also observed, albeit on a smaller scale, for household size and the child dependency ratio. The fractional regression model indicates that the age of the head of household is not statistically significant. This may be attributed to its parametric functional form, which limits the flexibility of the specified quadratic relationship for these variables, or it may reflect the fact that, despite the shape displayed by the ALE plots, the effects of age are not statistically relevant. Moreover, SHAP values based on random forest rank variables with highly nonlinear effects, such as inequality and the age of the household head, along with the child dependency ratio, among the top 15 determinants of poverty. However, despite the flexibility of random forest techniques to incorporate a wide range of interactions between explanatory variables, their results do not explicitly reveal those interactions. In contrast, the fractional regression model yields intriguing conclusions, particularly regarding the effects of education. While previous literature identifies education as a primary determinant of poverty, this paper demonstrates that, in Benin, the beneficial effects of education vary based on the gender of the head of household. Specified interactions indicate that the benefits of educating mothers extend intergenerationally, particularly in male-headed households. However, the education of the household head has a greater impact in female-headed households, with this gender difference converging at higher levels of education. These results underscore the importance of addressing gender inequalities in education, particularly by increasing access to female education, to effectively reduce poverty in the country.

Other significant determinants of welfare in Benin include household size, employment sector, financial access, and agroecological zones. While the importance of the first three factors is well established in the literature, the relevance of natural conditions arising from agroecological zones in explaining multidimensional poverty highlights the urgent need for climate change policies to address poverty in the long term. This is particularly crucial in a country heavily reliant on agriculture, where natural disasters and agricultural shocks are more significant determinants of deprivation than idiosyncratic shocks to households.

Finally, machine learning analysis using a large set of regressors identifies structural characteristics of the household (such as size and the proportion of children), education, and indicators of food deprivation and/or malnutrition (including food diversification and the collection of natural resources for cooking) as the most important predictors of multidimensional deprivation in Benin. This large-scale analysis of potential determinants reveals new indicators beyond those identified in the analysis based on a limited set of regressors, emphasizing the importance of addressing food insecurity and malnutrition as part of policies aimed at alleviating and reducing poverty in the country.

In general, the combination of parametric and nonparametric models employed in this paper demonstrates the potential of this integrated methodology for analyzing multidimensional poverty measures at the individual level. The parametric models provide statistical significance and interpretable interactions, while the flexible shapes produced by ALE and the ability of SHAP values to capture a diverse range of determinants in datasets contribute to a rich and adaptable framework for explaining and predicting poverty.

Acknowledgements

E. Arranhado and J.A. Bastos were partially supported by the Project CEMAPRE/REM - UIDB/05069/2020 - financed by FCT/MCTES through national funds.

References

[1] Acaha-Acakpo, H., Yehouenou, J. (2019). Determinants of household transition into and out of poverty in Benin. Journal of Development and Agricultural Economics, 11(5), 122-139.

[2] African Development Bank (2023). African Economic Outlook - Country Notes. Available at: https://www.afdb.org/en/countries-west-africa-benin/benin-economic-outlook

[3] Alia, D.Y. (2017). Progress toward the sustainable development goal on poverty: assessing the effect of income growth on the exit time from poverty in Benin. Sustainable Development, 25(6), 495-503.

[4] Alia, D.Y., Alia, K.A., Fiamohe, E.R. (2016). On poverty and the persistence of poverty in Benin. Journal of Economic Studies, 43, 661-676.

[5] Alinsato, A. S., Houedokou, W. (2019). Sector of Economic Activity and Poverty in Benin.

[6] Alkire, S., Foster, J., Seth, S., Santos, M., Roche, J., Baloon, P. (2015). Multidimensional Poverty Measurement and Analysis. Oxford, 2015; online edn, Oxford Academic, 20 Aug. 2015.

[7] Alkire, S., Kanagaratnam, U., Nogales, R., Suppa, N. (2022). Revising the Global Multidimensional Poverty Index: Empirical Insights and Robustness. Review of Income and Wealth Series 68, Number S2, December 2022.

[8] Apley, D.W., Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4), 1059-1086.

[9] Attanasso, M.O. (2005). Analysis of the Determinants of Monetary Poverty Among Female-Headed Households in Benin. Poverty and Economic Policy Research Network Working Paper No. PMMA-2005-06.

[10] Bakar, A.A., Hamdan, R., Sani, N.S. (2020). Ensemble Learning for Multidimensional Poverty Classification. Sains Malaysiana, 49, 447-459.

[11] Bluhm, R., de Crombrugghe, D., Szirmai, A. (2018). Poverty accounting. European Economic Review, 104, 237–255. https://doi.org/10.1016/j.euroecorev.2018.03.003

[12] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[13] Cameron, A.C., Miller, D.L. (2015). A Practitioner's Guide to Cluster-Robust Inference. The Journal of Human Resources, 50(2), 317-372.

[14] Curth, A., Jeffares, A., van der Schaar, M. (2024). Why do random forests work? Understanding tree ensembles as self-regularizing adaptive smoothers. arXiv:2402.01502v1.

[15] D'Ambrosio, C. (2018). Handbook of Research on Economic and Social Well-Being.

[16] Danish Trade Union Development Agency - DTDA (2021). Labour Market Profile Benin - 2021/2022. Available at: https://www.ulandssekretariatet.dk>wp-content>uploads>2021>04>LMP-Benin-2021-Final.pdf

[17] Diaz-Bonilla, C., Sabatino, C. (2022). April 2022 Update to the Multidimensional Poverty Measure – What's New. Global Poverty Monitoring Technical Note 22.

[18] Engstrom, R., Pavelesku, D., Tanaka, T., Wambile, A. (2017). Monetary and non-monetary poverty in urban slums in Accra : Combining geospatial data and machine learning to study urban poverty. World Bank

[19] Fitzpatrick, C. A., Bull, P., Dupriez, O. (2018). Machine learning for poverty prediction: A comparative assessment of classification algorithms. Available at: https://github.com/worldbank/ML-classification-algorithms-poverty

[20] Gbinlo, R.E. (2020). Drivers of Multidimensional Poverty: New Evidence in Benin. Journal of Economics and Development, 8(1), pp. 81-88.

[21] Grinsztajn, L., Oyallon, E., Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In Proceedings of NeurIPS 2022 – Neural Information Processing Systems. New Orleans, USA.

[22] Hastie, T., Tibshirani, R., Friedman, J.H. (2009). The Elements of Statistical Learning. 2nd Ed. Springer New York, NY.

[23] Haughton, J.H., Khandker, S.R. (2009). Handbook on Poverty and Inequality. World Bank Publications, 1-446.

[24] Hodonou, A., Damien, M., Gninanfon, A., Totin, A. (2010). Poverty Dynamics in Benin: A Markovian Process Approach (Dynamique De La Pauvrete AU Benin: Approche Par Le Processus Markovien). PEP PMMA Working Paper No. 2010-01. Available at: https://ssrn.com/abstract=1674633

[25] International Fund for Agricultural Development (2023). Benin. Available at: https://www.ifad.org/en/web/operations/w/country/benin

[26] James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An introduction to statistical learning: with applications in R. Springer New York, NY.

[27] Lanjouw, P., Ravallion, M. (1995). Poverty and household size. The Economic Journal, 105(433), 1415–1434.

[28] Li, Q., Yu, S., Échevin, D., Fan, M. (2022). Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan. Socio-Economic Planning Sciences (81), 101195.

[29] Liu, M., Hu, S., Ge, Y., Heuvelink, G.B., Ren, Z., Huang, X. (2020). Using multiple linear regression and random forests to identify spatial poverty determinants in rural China. Spatial Statistics, 100461.

[30] Lundberg, S.M., Erion, G.G., Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems.

[31] Lundberg, S.M., Erion, G., Chen, H. et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2, 56–67.

[32] Mcbride, L., Nichols, A.L. (2016). Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning. World Bank Policy Research Working Paper Series No. 7849.

[33] Min, P.P., Gan, Y.W., Hamzah, S.F., Ong, T.S., Sayeed, M.S. (2022). Poverty prediction using machine learning approach. Journal of Southwest Jiaotong University, 57(1).

[34] Minot, N., Daniels, L. (2005). Impact of global cotton markets on rural poverty in Benin. Agricultural Economics, 33, 453-466.

[35] Molnar, C. (2022). Interpretable Machine Learning - A Guide for Making Black Box Models Explainable. Available at: https://christophm.github.io/interpretable-ml-book/

[36] Molnar, C., Konig, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B. (2020). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models.

[37] Papke, L.E., & Wooldridge, J.M. (1996). Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. Journal of Applied Econometrics, 11(6), 619-632.

[38] Ramalho, E., Ramalho, J.J.S., Murteira, J.M.R. (2011). Alternative Estimating and Testing Empirical Strategies for Fractional Regression Models. Journal of Economic Surveys, 25(1), 19–68.

[39] Sohnesen, T.P., Stender, N. (2017). Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment. Poverty & Public Policy 9 (1), 118-133.

[40] Thoplan, R. (2014). Random Forests for Poverty Classification. International Journal of Sciences: Basic and Applied Research (IJSBAR), 17(2), 252-259.

[41] United Nations (2015). Multidimensional Poverty. Development Issues, 3.

[42] United Nations (2016). Chapter 4: Multidimensional Poverty and its measurement. Guide on Poverty Measurement. Conference of European Statisticians. Seminar on poverty measurement.

[43] United Nations (2017). Eradicating Poverty - Leaving no one behind. International Committee for Peace and Reconciliation (ICPR). NGO in Special Consultative Status with ECOSOC of the United Nations.

[44] Varian, H.R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2), 3-28.

[45] Vowels, M.J. (2020). Misspecification and unreliable interpretations in psychology and social science. Psychological methods. PMID: 34647760.

[46] Wooldridge, J.M. (2010). Econometric Analysis of Cross Section and Panel Data. The MIT Press. 2nd Edition.

[47] World Bank (2018). Poverty and shared prosperity 2018: Piecing together the poverty puzzle. World Bank. https://www.worldbank.org/en/publication/poverty-and-shared-prosperity-2018

[48] WAEMU Commission, Harmonized Survey on Households Living Standards, Benin 2018/19. Ref. BEN_2018_EHCVM_v02_M. Dataset downloaded from www.microdata.worldbank.org on February 2023.

[49] World Bank (2023). The World Bank in Benin. Available at: https://www.worldbank.org/en/country/benin/overview

[50] World Food Program (2024). Benin. Available at: https://www.wfp.org/countries/benin

[51] Zhao, Q., Hastie, T. (2021). Causal Interpretations of Black Box Models. Journal of Business & Economic Statistics, 39(1), 272-281.

Appendix

MPM construction framework

The MPM is based on three equally weighted dimensions as in the World Bank approach [47; 17]. A monetary dimension is combined with two non-monetary dimensions, education and basic infrastructures. Indicators in each dimension are parameters for deprivation cut-off, defined as 1-0 variables, where "1" means the individual or household is deprived in that indicator. Each of these parameters receives a weight, defined in Table A.1. All indicators are addressed at household level, meaning that all individuals in the same household would be classified, for instance, deprived in education if at least one individual in that household is deprived in one of the indicators of that dimension.

Dimension	**Dimension weight**	**Deprivation cut-off parameter**		**Parameter weight**
Monetary	1/3	Percentiles*	[0 0.25]	1/3
			]0.25 0.50]	$0.8 \times 1/3$
			]0.50 0.75]	$(0.8)^2 \times 1/3$
			]0.75 1.00]	$(0.8)^3 \times 1/3$
Education	1/3	At least one	School-age child up to the age of grade 8 is not enrolled in school	1/6
			Adult in the household (age of grade 9 or above) did not complete primary education	1/6
Basic infrastructure	1/3	Household lacks access to	Potable water in dry season	1/18
			Potable water in rainy season	1/18
			Healthy toilets	1/9
			Electricity	1/9

Table A.1: MPM - Dimensions, Indicators, and Weights.

Note: Adapted from [17]. *Percentiles of national monetary poverty ratio per person in the interval [0,1]. This ratio corresponds to per capita total consumption expenditure to national poverty line. Values >1 indicate non-monetary poor individual.

Monetary poverty results from weighting based on the quartile where the monetary poverty ratio of the poor individual/household is classified. Thus, the maximum weight of this dimension (1/3) is attributed to poor individuals/households falling in the first quartile, decreasing until the fourth quartile, where the weight reaches a value close to 1/6. Non-monetary poor individuals/households have 0 weight in all four quartiles. The national poverty line is used to define monetary poor households/individuals. In the dimension of education, the parameter cut-off for primary education considers at least one adult living in the household (instead of all adults) with age of grade 9 or above, allowing to have a MPM where not deprived individuals or households do not face any type of deprivation at the three dimensions. In the basic infrastructure dimension, the readily available data regarding drinking water and sanitation are used, which states if the first is potable or not and if the second is a healthy toilet or not.

Model Estimation

Variables	Coef.	SE	Variables (cont.)	Coef.	SE
head_age	0.000	0.01	hh_trf_receive (base = Yes)	0.040**	0.02
head_age2	0.000	0.00	hh_fin_access (base = Yes)	0.257***	0.02
hh_size	0.194***	0.01	hh_shk_severe_1 (base = None)		
hh_size2	-0.029***	0.01	Shk_Household	0.028	0.02
hh_depratio_c	0.109***	0.01	Shk_Nature	0.077***	0.02
hh_depratio_c2	-0.019***	0.00	Shk_Agriculture	0.072***	0.03
c_inequality	0.019**	0.01	Shk_Income	0.036	0.05
geo_aez (base = High rainfall...)			Shk_Economic	-0.044*	0.03
Transition, cotton...	0.274***	0.02	Shk_Business	0.020	0.04
Medium rainfall, cotton...	0.310***	0.03	Shk_Insecurity	-0.061	0.04
Low rainfall, cotton...	0.376***	0.03	Shk_Other	0.039	0.10
Cotonou	-0.090**	0.04	c_roadac (base = Tarred road)		
geo_urbrur (base = Urban)	0.041***	0.02	Laterite road	-0.011	0.02
head_sex (base = Male)	0.089***	0.03	Track	-0.002	0.02
head_educ (base = No education)			Sea, river, lake	-0.038	0.04
Primary education	-0.139***	0.02	Other	0.048	0.03
Secondary education	-0.229***	0.03	c_schoolcom (base = Yes)	0.040***	0.02
Higher education	-0.593***	0.04	c_electric (base = Yes)	0.106***	0.03
ind_edu_mother (base = No education)			c_water (base = Yes)	0.105***	0.03
Primary education	-0.262***	0.02	c_healthcom (base = Yes)	-0.050**	0.02
Secondary education	-0.315***	0.02	head_sex*head_educ		
Higher education	-0.614***	0.06	Female & Prim. educ.	-0.338***	0.05
head_emp_sector12m (base = Primary sector)			Female & Sec. educ.	-0.279***	0.06
Secondary sector	-0.104***	0.02	Female & Higher educ.	-0.191	0.13
Service sector	-0.167***	0.02	head_sex*ind_edu_mother		
Inactive	-0.102***	0.04	Female & Prim. educ.	0.183***	0.03
Unemployed	-0.355**	0.17	Female & Sec. educ.	0.163***	0.04
head_mstat (base = Polyg. Married)			Female & Higher educ.	0.251*	0.13
Monog. Married	0.006	0.02	head_sex*hh_trf_receive (Female & No)	-0.065**	0.03
Common-law union	-0.020	0.05	head_educ*head_emp_sector12m		
Divorced, Widowed	0.013	0.03	Sec. educ. & Sec. sector	-0.070	0.05
Single	0.104***	0.04	Sec. educ. & Service sector	-0.071*	0.04
ind_ethnic (base = Other)			Sec. educ. & Inactive	-0.167***	0.07
Yoruba	-0.093**	0.04	c_electric*c_water (No & No)	-0.063*	0.03
Adja	0.107***	0.02	c_water*c_healthcom (No & No)	0.055**	0.03
Peulh	0.203***	0.04	Constant	-0.290***	0.04
Fon	0.018	0.02			
Bariba	-0.023	0.03			
ind_migration (base = Not migrated)					
Rural area	0.018	0.01			
Urban area	-0.057***	0.02			
Foreign country	-0.020	0.02			
Number of obs	42343				
Log pseudolikelihood	-17984				
Wald Chi^2 for global significance (p-value)	0.000				
RESET Test $J = 2$ (p-value)	0.053				

Table A.2: Fractional Probit Model. *Note*: $^{***}p \leq 0.01$, $^{**}p \leq 0.05$, $^{*}p \leq 0.1$; SE = cluster-robust standard errors.