Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

APPLIED ECONOMETRICS AND FORECASTING

MASTER'S FINAL WORK

DISSERTATION

MULTIDIMENSIONAL POVERTY IN BENIN:

EVIDENCE FROM CLASSIC AND MACHINE LEARNING ANALYSIS

LÁGIDA KÓRCIA ALMEIDA COIMBRA MONTEIRO BARBOSA

SEPTEMBER – 2023

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

APPLIED ECONOMETRICS AND FORECASTING

MASTER'S FINAL WORK

DISSERTATION

MULTIDIMENSIONAL POVERTY IN BENIN:

EVIDENCE FROM CLASSIC AND MACHINE LEARNING ANALYSIS

LÁGIDA KÓRCIA ALMEIDA COIMBRA MONTEIRO BARBOSA

SUPERVISION:

ESMERALDA DE JESUS RATINHO LOPES ARRANHADO
JOÃO AFONSO BASTOS

SEPTEMBER – 2023

To my son, Cássio Barbosa,

the driving force behind

all I do.

GLOSSARY

AFDB – African Development Bank

ALE – Accumulated local effects

APE – Average partial effects

DGP – Data generating process

EHCVM – Enquête Harmonisée sur le Conditions de Vie des Ménages

IFAD – International Fund for Agricultural Development

JEL – Journal of Economic Literature

MAPE – Mean absolute percentage error

MLE – Maximum likelihood estimator

MSE – Mean squared error

OLS – Ordinary least square

PPP – Purchasing Power Parity

QMLE – Quasi maximum likelihood estimator

SHAP – SHapley Additive exPlanations

UN MDG – United Nations Millennium Development Goals

ABSTRACT, KEYWORDS AND JEL CODES

This dissertation provides new insights on immediate factors affecting multidimensional poverty in Benin. Ordered probit and fractional probit models are compared to the random forest model, and poverty-targeting indicators are derived for the country, using 2018/2019 individual-level cross-sectional data. In most cases, the effects of regressors on the response variable have the same direction of impact in both *glass box* and *black box* models, whereas accumulated local effects plots on random forest suggest a highly nonlinear relationship between individual's welfare condition and the age of household head and inequality, as well as a nonlinear but non-concave relationship with household size and child dependency ratio. While all models corroborate suggesting that education, agroecological zones, financial access, household size, and employment sector are among most important variables associated with welfare condition, only the *black box* model, through SHAP values, ranked variables with highly nonlinear effects among the most important regressors, as well child dependency ratio. Moreover, the random forest model, by computing more complex interactions between variables, was able to present a broader range of important variables in the top 15. In general, my findings are consistent with most literature on poverty in Africa and Benin, with all models indicating that education is the most important "proximate" determinant of the welfare condition in Benin. The most important poverty-targeting indicators are household size, food diversification, household head without education, households that gather wood for home cooking, and child dependency ratio.

Keywords: Multidimensional Poverty; Ordered Probit; Fractional Probit; Random Forest; Explainable Model Techniques

JEL CODES: I32, C13, C14, C21, C25, C52

ABSTRATO E PALAVRAS-CHAVE

Esta dissertação fornece novos *insights* sobre fatores imediatos que afetam a pobreza multidimensional no Benim, comparando resultados dos modelos probit ordenado e probit fracionado com o modelo de floresta aleatória, e deriva indicadores de pobreza para a política de intervenção, utilizando dados *cross-section* de 2018/2019 ao nível individual. Na maioria dos casos, os efeitos dos regressores têm a mesma direção de impacto esperada nos modelos *glass box* e *black box*. O ALE *plot* aplicado à floresta aleatória sugere uma relação altamente não linear entre a condição de bem-estar do indivíduo e a idade do chefe do agregado familiar e a desigualdade, bem como uma relação não linear, mas não côncava, com a dimensão do agregado familiar e o rácio de dependência infantil. Embora todos os modelos corroborem que a educação, as zonas agroecológicas, o acesso financeiro, a dimensão do agregado familiar e o sector do emprego estão entre as variáveis mais importantes associadas à pobreza no Benin, apenas o modelo *black box*, através de valores SHAP, classificou as variáveis com efeitos altamente não lineares entre os regressores mais importantes, bem assim o rácio de dependência infantil. Além disso, a floresta aleatória, ao calcular interações mais complexas entre variáveis, conseguiu apresentar um leque mais vasto de variáveis importantes no *top* 15. Em geral, os resultados dos modelos são coerentes com a maior parte da literatura sobre a pobreza em África e no Benim, com todos os modelos a indicarem que a educação é o determinante "próximo" mais importante da condição de bem-estar no Benim. Os indicadores do perfil de pobreza mais importantes são a dimensão do agregado familiar, a diversificação alimentar, o chefe de família sem instrução, os agregados familiares que recolhem lenha para cozinhar e o rácio de dependência infantil.

Palavras-Chave: Pobreza Multidimensional; Probit Ordenado; Probit Fracionado; Floresta Aleatória; Interpretação de Modelos *Black Box.*

JEL CODES: I32, C13, C14, C21, C25, C52

TABLE OF CONTENTS

TABLE OF FIGURES

AKNOWLEDGMENTS

I express my gratitude towards my family and friends for their patience and continued encouragement during the duration of my engagement with this endeavor. I am very thankful to my spouse, Danilson Barbosa, in particular.

I would like to thank Professors Esmeralda Arranhado and João Bastos for their guidance and accessibility throughout the entire process.

Furthermore, I am thankful to the Bank of Portugal and the Bank of Cabo Verde for their financial support.

MULTIDIMENSIONAL POVERTY IN BENIN:
EVIDENCE FROM CLASSIC AND MACHINE LEARNING ANALYSIS

By Lágida Barbosa

1. INTRODUCTION

Poverty is a global social problem that affects every nation, albeit more severely in developing nations. Its deleterious vicious cycle makes reducing poverty one of the top priorities of development policy, where eradicating extreme poverty and hunger is the number one goal of the UN MDG. Consequently, it is important to understand which factors can affect the measures of poverty and how they do so. Meanwhile, as a complex multidimensional phenomenon, poverty analysis can be challenging in most conventional statistical methods, as they may struggle to uncover complex pattern in the data. Statistical learning that relies on *black-box* models to analyze poverty have been implemented in literature mostly for policy-targeting approach (see, for instance, Thoplan, 2014; McBride & Nichols, 2016; Sohnesen & Stender, 2017; Engstrom et al., 2017; Fitzpatrick et al., 2018; Alsharkawi et al., 2021; Li et al., 2022). Although causal interpretation of predictive models is often not possible, if a model approximates the DGP well enough, its interpretation should reveal insights into the underlying process (Molnar et al., 2020). Relying on some domain knowledge about the causal structure, on models with good predictive performance, and on suitable visualization tools allow to gain intuition on how regressors affect the response function from a *black-box* model (Zhao & Hastie, 2019).

This dissertation provides new insights on immediate factors affecting multidimensional poverty in Benin, by comparing the outputs of ordered probit and fractional probit models to the random forest model, and derives poverty-targeting indicators for the country. In most cases, the effects of regressors on the response variable have the same expected direction of impact in both *glass box* and *black box* models, whereas the random forest, through accumulated local effects (ALE) plots, provided a deeper intuition on the effects. This non-parametric approach suggests that there is a highly nonlinear relationship between the individual's welfare condition and the age of household head and inequality, as well as a nonlinear but non-concave relationship with household size and child dependency ratio. While all models corroborate suggesting that education (of both household head and individual's mother), agroecological zones,

financial access, household size, and employment sector are among most important variables associated with welfare condition, only the *black box* model, through SHAP values, ranked the variables with highly nonlinear effects among the most important regressors, as well child dependency ratio. Moreover, the random forest, by computing more complex interactions between variables, was able to present a broader range of important variables in the top 15. In general, the findings are consistent with most literature on poverty in Africa and Benin, with all models indicating that education is the most important "proximate" determinant of the welfare condition in Benin. The out-of-sample error rate of the random forest in the classification problem was 2.5%, compared to 37.5% of the ordered probit model, while the MSE of the random forest in the regression problem was 0.002, compared to 0.435 of the fractional probit model. Regarding the poverty profile, the five most important indicators describing a poor Beninese in the 2018/2019 are: household size, food diversification, household head without education, households that gather wood for home cooking, and child dependency ratio.

I performed the analysis using cross-sectional data at individual level, using a set of 21 regressors, selected among determinants of poverty suggested by literature, to gain insights on important variables and how they affect the response function, beyond comparing their out-of-sample predictive performance. These determinants can be seen as “proximate” determinants of poverty, as it is very challenging to uncover the deep roots of poverty (Haughton & Khandker, 2009). I use two dependent variables: 1) a continuous variable representing a multidimensional poverty measure (MPM); and 2) an ordinal variable representing classes of deprivation (MPMc), derived from the previous response variable. The insights from the random forest model are given by two explainable model techniques: 1) SHapley Additive exPlanations (SHAP) values (Lundberg and Lee, 2017); and 2) Accumulated local effects (ALE) plots (Apley and Zhu, 2020). Then, I expanded the dataset to 231 predictors for the poverty-targeting approach, using MPM as target variable, random forest model, and SHAP values. The data are from 2018/2019 EHCVM national survey for Benin, retrieved on World Bank’s microdata database.

This dissertation is divided into the following six sections: The first section is the introduction; the second section provides a brief literature review; the third section describes the methodologies used; the fourth section describes the data and provides

descriptive analysis; the fifth section presents the empirical findings; and the final section provides a conclusion.

2. LITERATURE REVIEW

2.1. Poverty - a complex phenomenon

The concept of poverty, which is associated with the notion of well-being, is very complex. Not surprisingly, literature is vast in terms of the understanding of what is well-being, hence, on how to measure poverty. According to United Nations (2017), 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 (as cited by Haughton & Khandker, 2009) defines poverty as a "pronounced deprivation in well-being".

The most prevalent approach to quantifying poverty defines well-being in monetary terms (a welfarist approach), setting a threshold or poverty line bellow which households/individuals are considered poor. Meanwhile, measuring poverty with a single income or expenditure measure is an imperfect approach to comprehend the deprivations of the poor (United Nations, 2015). Other perspectives give rise to non-monetary dimensions (such as subjective poverty, health poverty, education poverty, etc.) or multidimension approaches to measuring poverty, which may or may not incorporate both the monetary and non-monetary dimension. Indeed, for a set of reasonable axioms on poverty measurement, there may be several poverty indices (Chakravarty, 2009). According to D'Ambrosio (2018), after the work of Sen (1976), several authors suggested postulates for a multidimensional poverty index – for instance, Tsui (2002), Bourguignon and Chakravarty (2003), Chakravarty and Silber (2008), and Alkire and Foster (2011a). In particular, the United Nations and the World Bank approach to measuring the multidimensional poverty takes monetary and non-monetary information into account to provide a more complete picture of poverty (United Nations, 2015; Diaz-Bonilla & Sabatino, 2022).

Another layer of complexity comes from the attempt to understand the deep roots of poverty. According to Haughton and Khandker (2009), we can show that a lack of education causes or increases the risk of poverty, but cannot so easily explain why some people lack education, being also difficult to separate causation from correlation. We can

question whether individuals are poor due to lack of education, whether they lack education because they are poor, or also if there is simultaneity. According to Silva (2008), most economic variables at household level become endogenous, as the time horizon of the analysis increases, with many variables becoming a function of the welfare level to some extent. Therefore, the models can return results for the degree of association or correlation and not for casual relationships.

In consequence, developing a clear understanding of the fundamental causes of poverty is challenging, which is why researchers have concentrated on immediate or "proximate" causes of poverty (Haughton & Khandker, 2009). In this framework, poverty may be caused by (or at least correlates with) national, sector-specific, community, household and/or individual characteristics factors. TABLE I synthetizes the main indicators.

TABLE I

MAIN IMMEDIATE DETERMINANTS OF POVERTY

Regional	Isolation/remoteness, less infrastructure, poorer access to markets and services
	Resource base, land availability/quality
	Weather/environmental conditions
	Regional governance and management
	Inequality
Community	Infrastructure
	Land distribution
	Access to public goods and services
	Social structure and social capital
Household	Household size
	Dependency ratio
	Household head (or adults on average) sex
	Assets
	Employment and income structure
	Household members health/education on average
Individual	Age
	Education
	Employment status
	Health status
	Ethnicity

Source: Haughton and Khandker (2009)

2.2. Determinants of poverty in Africa and Benin from conventional statistical learning

In Africa, empirical research shows the importance of many of the "proximate" factors as determinants (or correlates) of poverty. Although some factors may vary by

region and country, the role of education is systematically evidenced in many studies (for instance, Glewwe, 1991; Grootaert, 1997; Datt et al., 2000; Datt & Jolliffe, 2001; Okurut et al., 2002; Muller, 2002; Mukherjee & Benson, 2003; Geda et al., 2005; Bogale et al., 2005; Sackey, 2005; Adjasi & Osei, 2007; Epo, 2010; Fambon, 2017; Habyarimana et al., 2015; Cho & Kim, 2017). Geda et al. (2005) find education, measured in terms of the highest level attained at the household, to be the most important determinant of poverty in Kenya, while Datt and Jolliffe (2001) stress the role of parent's education to capture intergenerational human capital effects on living standards in Egypt, in addition to the effect of the average years of schooling of the household.

Household size is also found to be one of the main and sometimes the most important determinant of poverty (Datt et al., 2000; Okurut, 2002; Geda et al., 2005; Sekhampu, 2013; Cho & Kim, 2017), while other important determinants are related to the employment sector or status (for instance, Sekhampu, 2013; Fambon, 2017); assets, in particular, land (for instance, Bogale et al., 2005; Sackey, 2005); age (Epo, 2010; Sekhampu, 2013; Habyarimana et al., 2015); gender (for instance, Habyarimana et al., 2015; Cho & Kim, 2017); place of residence (Adjasi & Osei, 2007; Habyarimana et al., 2015); ethnicity (Glewwe, 1991; Muller, 2002; Mededji, 2008); access to credit (Sackey, 2005; Fambon, 2017); and access to infrastructure (Muller, 2002; Mukherjee & Benson, 2003; Epo, 2010).

Benin's empirical research show that the most relevant determinants of poverty go in tandem with empirical findings in Africa. According to Hodonou et al. (2010), the dynamics of poverty status in the country is sensitive to the age and sex of household head, household size, place of residence, possession of durable goods, improved access to housing, electricity, communications, and education. In particular, education and household size, as well for Benin, are found to have a very important role explaining poverty, with empirical research suggesting that the effects of education on poverty do not differ by gender nor by place of residence (Attanasso, 2005); that households with an educated head, with at least a primary education, are at a lower risk of being poor (Alia et al., 2016; Gbinlo, 2020); and that education is one of the main determinants of the time needed to exit poverty (Alia, 2017). On the other hand, the importance of household size in explaining the transition in poverty status is well evidenced in Hodonou et al. (2010), Alia (2017) and Acaha-Acakpo and Yehouenou (2019), with their findings suggesting

that an increase in the household size worsens the household welfare, in line with relevant literature on poverty.

The geographical location of households seems to increase the likelihood to emerge from poverty in Benin, when cotton- and rice-producing regions are considered (Acaha-Acakpo & Yehouenou, 2019), although working in the agriculture sector has been associated to a negative impact on the household welfare (Alia, 2017; Acaha-Acakpo & Yehouenou; 2019). Other research (Mededji, 2008; Alia, 2017; Acaha-Acakpo & Yehouenou, 2019) find that the employment sector is the most or among most important determinants affecting the likelihood of escaping from poverty in Benin, while Alinsato and Houedokou (2019) points to unobserved factors related to the participation in labor market segments that influence the poverty status in the country.

Attanasso (2005) focus her research on gender poverty and finds that the likelihood of being poor is lower for female-headed household, even though in some departments they faced more severe poverty. While Alinsato and Houedokou (2019) reach a similar conclusion, Alia et al. (2016), Gbinlo (2020), and Alia (2017) find that there is an increase in the likelihood of being poor for female-headed households. In the literature, it is widely held that female-headed households are among the most vulnerable and are disproportionately represented among those who are poor (Saad et al., 2022), as women may face gender discrimination with respect to education, earnings, rights, and economic opportunities (Barros et al., 1997; as cited in Rajaram, 2009), and female-headed households may be stigmatized in patriarchal societies (Chant, 2007; AbuFarash, 2016; as cited in Saad et al., 2022). Meanwhile, there are some critics regarding this view, as female-headed households present heterogeneous characteristics (Chant, 2004; as cited in Saad et al., 2022); there are practical issues related to identifying the actual head of the household, and female headship is not always correlated with poverty (Buvivnic & Gupta, 1997; as cited in Rajaram, 2009); and the econometric results may present contradictory results depending on the measure of poverty employed (Rajaram, 2009).

Regarding shocks, Minot and Daniels (2005) find statistical evidence of a strong link between cotton prices and rural welfare in Benin, with a price reduction shock leading to an increase in poverty in the short and long run. According to Alia et al. (2016), the vulnerability of the households to various types of shocks may explain the large and rapid

change in poverty status in Benin, between 2006 and 2011, with households moving in and out of poverty, an idea reinforced by Gbinlo (2020) econometric results, which suggest an increase in the likelihood to be poor for households experiencing biophysical shocks.

2.3. Machine learning in poverty analysis

The statistical learning employed in the aforementioned poverty analysis is traditional in sense that it relies on parametric models, regarded as *glass box* models, to allow inference, such as logit/probit, OLS, multinomial or ordered logit/probit, and quantile regression (more recently applied to poverty analysis). However, there has been a developing interest in applying machine learning techniques to empirical research on poverty over the past few years.

As machine learning algorithms, such as random forest, support vector machines, and neural networks, may allow for more effective ways to model complex relationships (Varian, 2014), these non-parametric models, regarded as *black box* models, can accommodate highly non-linear relationships between the target and explanatory variables, making them an alternative to analyze complex multidimensional phenomena such as poverty. However, this comes at a cost. Statistical learning, as a set of approaches for estimating the relationship between a response variable Y as a function of regressors $\boldsymbol{X}$, is implemented for prediction and/or inference purpose (James et al., 2021). In this framework, the confront between the goal of prediction *versus* the goal of inference generates a trade-off between accuracy and interpretability, where non-parametric models often have high accuracy, but inference is more challenging and requires additional techniques to bring out of the *black box* which variables are important for the model's predictions and how they affect the response variable.

The analysis of empirical research conducted so far using *black box* models for poverty analysis suggests that researchers have been more concerned with an accurate targeting of the poor, leading them to employ a poverty-targeting approach, where prediction is the main goal. This approach is a way to overcome the constraint that updated information on income, consumption or expenditure is not readily available, requires human effort and is costly to obtain from survey data collected directly from households (Haughton & Khandker, 2009), contributing to a more effective and efficient

policy intervention towards poverty alleviation (see, for instance, Thoplan, 2014; McBride & Nichols, 2016; Sohnesen & Stender, 2017; Engstrom et al., 2017; Fitzpatrick et al., 2018; Liu et al., 2020; Bakar et al., 2020; Verme, 2020; Alsharkawi et al., 2021; Li et al., 2022; Min et al., 2022; Usmanova et al., 2022).

For the purpose of prediction alone, the only concern is to find the optimal mapping function to achieve some desired level of predictive performance, which is a goal quite different from that of estimating a parameter of a distribution (Vowels, 2021). Moreover, there is no endogeneity concern when generating targeting tools because the goal is not causal inference but rather the out-of-sample performance (McBride and Nichols, 2016). In consequence, the poverty-targeting approach can lead to results, in terms of important variables for the prediction performance of the model, that do not reflect the DGP of the response variable, as the large set of regressors usually employed can include variables that are cause (e.g., a shock to the household) and an effect (e.g., ownership of laptop) of the response variable. Hence, when explainable model techniques for machine learning are used to uncover important features, in this context, only a parsimonious poverty profile can be derived.

Although a causal interpretation of predictive models is often not possible because standard supervised machine learning models are designed to merely exploit associations and most explainable model techniques are designed to interpret the model instead of drawing inferences about the DGP, if a model approximates the DGP well enough, its interpretation should reveal insights into the underlying process (Molnar et al., 2020). Indeed, Zhao and Hastie (2019) propose three prerequisites for making causal interpretations of *black box* models: 1) a model with good predictive performance; 2) some domain knowledge about the causal structure; 3) and suitable visualization tools.

3. METHODOLOGY

3.1. Ordered Probit Model

The ordinal dependent variable MPMc, which is a monotonic transformation of a single continuous outcome that is naturally ordered (Greene & Hensher, 2009), is obtained by collapsing the values of MPM into a set of four categories representing increasing levels of deprivation for individuals. This transformation allows to employ a parametric ordered response model to analyze factors that can impact the well-being of

individuals in different classes of deprivation, in particular, the most deprived ones, such as the ordered probit model (Anderson & Philips, 1981; as cited by Greene & Hensher, 2009).

The ordered probit model is usually derived from a latent variable model. The model's latent variable y_{ih}^{*} is determined by:

(3.1.1) $$y_{ih}^{*} = \boldsymbol{X}_{ih}^{T}\boldsymbol{\beta} + e_{ih}, \qquad e_{ih}|\boldsymbol{X}_{ih} \sim Normal\,(0,1)$$

where i represents individuals, h represents the household cluster, for a sample size $N = 42343$ and a cluster size $H = 8012$; $\boldsymbol{X}_{ih}^{T}$ is a $N \times P$ matrix of exogenous regressors, and does not include the intercept; $\boldsymbol{\beta}$ is a $P \times 1$ vector of regressors coefficients; $P = 21$ explanatory variables; and e_{ih} is the error term following a standard normal distribution.

The four categories of deprivation are determined by y_{ih}^{*} through cut points or threshold parameters α_j:

(3.1.2) $$y_{ih} = j \quad \text{if} \quad \alpha_{j-1} < y_{ih}^{*} \leq \alpha_j\,, \quad j = 1,2,3,4$$

where y_{ih} is the response variable representing MPMc and taking on values $\{1,2,3,4\}$, each one having an ordinal meaning, where 1 represents non deprived individuals and 4 represents most deprived individuals. $\alpha_0 < \alpha_1 < \alpha_2 < \alpha_3 < \alpha_4$, assuming $\alpha_0 = -\infty$ and $\alpha_4 = +\infty$ (Greene & Hensher, 2009; Long & Freese, 2014). Although in practice MPMc is derived from MPM with specified cut points ($\delta_1 = 0$, $\delta_2 = 1/3$, and $\delta_3 = 2/3$), in this model framework the latent variable y_{ih}^{*} is unknown and so the cut points α_1, α_2, and α_3 are estimated by the model. Each cut point is the intercept inside the probit cumulative distribution function Φ, determining the magnitudes of the estimated probabilities and partial effects (Wooldridge, 2010).

The distribution of y conditional on $\boldsymbol{X}$ is derived by computing each response probability, which sum up to unity, based on the standard normal distribution assumption for e_{ih}:

(3.1.3) $$P(y_{ih} = j|\boldsymbol{X}_{ih}) = \Phi\left(\alpha_j - \boldsymbol{X}_{ih}^{T}\boldsymbol{\beta}\right) - \Phi\left(\alpha_{j-1} - \boldsymbol{X}_{ih}^{T}\boldsymbol{\beta}\right)$$

The method of estimation is the maximum likelihood (MLE), based on the maximization of the log-likelihood function:

(3.1.4) $$LL = \sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{j=1}^{J} \mathrm{I}[y_{ih} = j]\ln\{\Phi(\alpha_j - \boldsymbol{X}_{ih}^{T}\boldsymbol{\beta}) - \Phi(\alpha_{j-1} - \boldsymbol{X}_{ih}^{T}\boldsymbol{\beta})\}$$

The MLE is consistent, asymptotically normal and efficient, provided that the conditional mean and conditional distribution of the linear latent variable model are correctly specified. I 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 (Wooldridge, 2010). According to Cameron and Miller (2015), a cluster-robust standard error is also a heteroskedastic-robust standard error.

The partial effects assume a relevant importance in this model, as neither the signs nor the magnitudes of the coefficients are directly interpretable in the ordered choice model (Greene & Hensher, 2009). The partial effects are not constant, as the link function $\Phi(.)$ is non-linear, and hence I rely on the sample average partial effects (APE). The estimated partial effect of a continuous variable x_k on each response probability and for each individual i is:

(3.1.5) $$\widehat{PE_{ik}} = \frac{\partial p_j(\boldsymbol{X})}{\partial x_{kih}} = \hat{\beta}_k[\phi(\alpha_{j-1} - \boldsymbol{X}_{ih}^{T}\widehat{\boldsymbol{\beta}}) - \phi(\alpha_j - \boldsymbol{X}_{ih}^{T}\widehat{\boldsymbol{\beta}})]\,, \quad j = 1,2,3,4$$

where ϕ is the probability density function. The direction of the effect of x_k on the probabilities $P(y_{ih} = 1|\boldsymbol{X}_{ih})$ and $P(y_{ih} = 4|\boldsymbol{X}_{ih})$ is determined by the sign of $\hat{\beta}_k$, whereas the highest outcome has the same sign as $\hat{\beta}_k$ and the lowest outcome has the opposite sign to $\hat{\beta}_k$. Regarding the intermediate probabilities, the direction of the effect may not be inferred from the sign of $\hat{\beta}_k$. For a dummy regressor x_z, the partial effect for individual i is given by:

(3.1.6) $$\widehat{PE_{iz}} = \boldsymbol{P}(y_{ih} = j|\boldsymbol{X}_{ih}, x_z = 1) - \boldsymbol{P}(y_{ih} = j|\boldsymbol{X}_{ih}, x_z = 0)$$

3.2. Fractional Probit Model

As MPM is bounded to the interval [0,1] with no concentration of extremum values in the sample data, the conventional fractional regression model estimated by quasi-maximum likelihood (Papke & Wooldridge, 1996) is an alternative approach to analyze the determinants of the multidimensional poverty in Benin, without the loss of information caused by grouping the individuals into four classes of deprivation.

The model conditional mean is

(3.2.1) $E(y_{ih}|\boldsymbol{X}_{ih}) = G\big(\beta_0 + \beta_1 x_{ih1} + \cdots + \beta_p x_{ihp}\big) = G(\boldsymbol{X}_{ih}^T\boldsymbol{\beta}),$

where y_{ih} is the MPM for individual i and cluster h, and $0 \leq y_{ih} \leq 1$; $\boldsymbol{X}_{ih}^T\boldsymbol{\beta}$ is the index function, where $\boldsymbol{X}_{ih}^T$ is a $N \times (1+P)$ matrix including the intercept and exogenous regressors, and $\boldsymbol{\beta}$ is a $(1+P) \times 1$ vector of coefficients. The link function $G(.)$ is a probit functional form and so a standard normal cumulative distribution function, satisfying the condition $0 < G(.) < 1$, $\forall\, z \in \mathbb{R}$, which ensures that the predicted values of y lie in the interval (0,1):

(3.2.2) $G(\boldsymbol{X}_{ih}^T\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_{ih}^T\boldsymbol{\beta}) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz$

The quasi-maximum likelihood method of estimation (QMLE) is based on the maximization of the Bernoulli log-likelihood function:

(3.2.3) $LL \equiv \sum_{i=1}^{N}\sum_{h=1}^{H}\{y_{ih} \ln[G(\boldsymbol{X}_{ih}^T\boldsymbol{\beta})] + (1-y_{ih}) \ln[1 - G(\boldsymbol{X}_{ih}^T\boldsymbol{\beta})]\}$

The Bernoulli QMLE of $\boldsymbol{\beta}$ is consistent and asymptotically normal, provided that only the conditional mean, which is non-linear in $\boldsymbol{\beta}$, is correctly specified, regardless of the distribution of y conditional on $\boldsymbol{X}$. I use also cluster-robust standard errors.

In this framework the sign of the partial effect is given by the sign of $\hat{\beta}_j$. The estimated partial effects of a continuous variable x_j and a dummy variable x_k for an individual i are, respectively,

(3.2.4) $\widehat{PE_{ij}} = \hat{\beta}_j \frac{\partial \Phi(\boldsymbol{X}_{ih}^T\hat{\boldsymbol{\beta}})}{\partial x_{ihj}} = \hat{\beta}_j \left(\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(\boldsymbol{X}_{ih}^T\hat{\boldsymbol{\beta}})^2}{2}\right]\right)$

(3.2.5) $\widehat{PE}_{ik} = [\Phi\big(\boldsymbol{X}_{ih,x_k=1}^T\widehat{\boldsymbol{\beta}}\big) - \Phi\big(\boldsymbol{X}_{ih,x_k=0}^T\widehat{\boldsymbol{\beta}}\big)]$

3.3. Specification testing in glass box models

The conditional mean of both ordered and fractional probit models may be tested using RESET tests. In this dissertation the test is applied in the versions that adds up to two fitted powers of the linear index. The null and alternative test hypothesis are:

(3.3.1) $H_0: E(y|\boldsymbol{X}) = G(\boldsymbol{X}^T\boldsymbol{\beta})$

$H_1: E(y|\boldsymbol{X}) = G(\boldsymbol{X}^T\boldsymbol{\beta} + \sum_{j=1}^{J}\gamma_j(\boldsymbol{X}^T\boldsymbol{\beta})^{j+1}),$

where $j = 1$ for one fitted power [polynomial $(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})^2$ included in $G(.)$], and $j = 1, 2$ for two fitted powers [polynomials $(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})^2$ and $(\boldsymbol{X}^T\widehat{\boldsymbol{\beta}})^3$ included in $G(.)$]. As H_0 is equivalent to test for $H_0: \gamma_j = 0$, I apply a robust Wald test to make inference about the general model functional form specification. Under H_0, the Wald test follows a χ^2 distribution, with one degree of freedom ($J = 1$) or two degrees of freedom ($J = 2$).

3.4. Random Forest

Random forest is a combination of tree predictors, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). It is a non-parametric model, a supervised learning algorithm and an ensemble method implemented using bagging and feature randomness. The aim is overcoming the tendency of decision trees to overfit the training data, i.e., their usually feature of suffering from high variance, by enhancing the accuracy of the model through bootstrap aggregation method and creating an uncorrelated forest of decision trees.

A decision tree itself is a non-parametric model that can be applied to both regression and classification problems. It consists of a series of splitting rules (if-then-else rules on regressors), starting at the top of the tree (*root node*) and subsequently generating two *child nodes* from each previous node (*parent node*) until a terminal node (*leaf node*) is reached by some stopping criteria (see, for instance, Hastie et al., 2009; James et al., 2021). For a cut-off value or split-point s on the regressor x_j, the data are divided into two *child nodes* (region spaces R_1 and R_2), whereas observations satisfying the condition $x_j < s$ go to one *child node* $[R_1(j, s) = \{\boldsymbol{X}|x_j < s\}]$ and observations satisfying the condition $x_j \geq s$ go to the other *child node* $[R_2(j, s) = \{\boldsymbol{X}|x_j \geq s\}]$. A *child node* becomes a *parent node* if it is subsequently splited or a *terminal node* if not. The algorithm chooses at each *parent node* the predictor (which j?) and split-point s that minimize the variation on the dependent variable,

(3.4.1) $\sum_{i:\, x_i \in R_1(j,s)}(y_i - \bar{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)}(y_i - \bar{y}_{R_2})^2,$

in the case of regression problem, meaning that we want to maximize the reduction in the variation of y with respect to the *parent node*. In the case of a classification problem, j and s are selected to minimize the node impurity, measured with Entropy or Gini Index,

$$(3.4.2) \quad \begin{cases} Entropy: \frac{n_L}{n_P} E_{C_L} + \frac{n_R}{n_P} E_{C_R}, \;\; E = -\sum_{k=1}^{K} p_k log_2 p_k \\ Gini: \frac{n_L}{n_P} G_{C_L} + \frac{n_R}{n_P} G_{C_R}, \;\; G = 1 - \sum_{k=1}^{K} p_k^2 \end{cases}$$

where E, G, and p_k denote, respectively, the entropy measure, the Gini index measure, and the proportion of observations in a node from the kth class; E_{C_L}, E_{C_R}, n_L, n_R, n_P denote the entropies of the left and right *child nodes*, and the number of observations on the left *child*, right *child* and *parent* nodes, respectively. Minimizing the node impurity leads to the maximization of the information gain with respect to the *parent node*.

Many trees are generated in a random forest. Following Hastie et al. (2009), the mechanism of growing the trees in random forest can be synthetized as follows: for $b = 1$ to B trees, the algorithm draws a bootstrap sample N^* of size N from the training data; then, it grows a random-forest tree T_b to the bootstrapped data, recursively repeating the following steps before each split, until the minimum node size n_{min} is reached: 1) it selects $m \leq P$ variables at random, from the P regressors, as candidates for splitting; 2) picks the best predictor among the m; 3) and splits the node into two *child nodes*.

The output of the ensemble of trees is $\{T_b\}_1^B$ and the prediction at a new point i is the average of trees prediction, in the case of a regression problem

$$(3.4.3) \quad \hat{f}_{RF}^B(i) = \frac{1}{B}\sum_{b=1}^{B} T_b(i),$$

and the majority vote, in the case of a classification problem

$$(3.4.4) \quad \hat{C}_{RF}^B(i) = majority\ vote\ \left\{\hat{C}_b(i)\right\}_1^B,$$

where $T_b(i)$ and $\hat{C}_b(i)$ are, respectively, the predicted value and the class prediction of the bth random-forest tree.

To run the random forest model, the main three hyperparameters must be defined: 1) number of trees B; 2) number of random variables m; and 3) minimum node size n_{min}. I rely on the usual approach of setting $m \approx P/3$ for the regression problem, where the dependent variable is MPM, and $m \approx \sqrt{P}$ for the classification problem, where the dependent variable is MPMc. The number of regressors $P = 21$ as for the *glass box* models, but also $P = 231$ to derive the poverty profile for Benin for the covered survey period. The minimum node size is also the default, defined as $n_{min} = 1$ for the

classification problem and $n_{min} = 5$ for the regression problem. I randomly select 80% of the data for training and 20% for testing, and set the number of trees $B = 500$.

3.5. Model's out-of-sample performance

I analyze the out-of-sample performance of the models using the test dataset, randomly selected with 20% of the total sample observation, and relying on the mean squared error (MSE) as the loss function to compare the prediction accuracy of the regression problem models:

(3.5.1) $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

For classification problem, the comparison of the performance of the models is based on the confusion matrix for each class of deprivation:

(3.5.2) $$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

(3.5.3) $$Error\ rate = 1 - Accuracy$$

(3.5.4) $$F1\ score = \frac{2\times\left[\frac{TP}{TP+FP}\right]\times\left[\frac{TP}{TP+FN}\right]}{\left[\frac{TP}{TP+FP}\right]+\left[\frac{TN}{TN+FN}\right]},$$

where TP, FP, TN, and FN are, respectively, true positives, false positives, true negatives and false negative predictions of deprivation classes. Higher F1-score and accuracy indicate better model performance.

3.6. Explainable model techniques for black box interpretation

3.6.1. SHAP *values*

The Shapley value is a method from the 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. In machine learning context (Lundberg & Lee, 2017), regressors represent the players and prediction represents the payout in the regression analysis. The SHAP value for a regressor value of x_j 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 value of regressor x_j, can be written as

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

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 (Molnar, 2022); $f_{S \cup x_j}$ is the model trained including x_j and $f_{S \cup x_j}(\boldsymbol{X}_S \cup x_j)$ 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:

(3.6.1.2) $I_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j^{(i)}$

Although the random forest does not require scale transformation of regressors, when computing SHAP values, I perform a scale transformation of the numerical variables and one-hot-encoding of categorical variables. Also, to compare its results of the *glass box* models, I adjust the APE by scale transforming the numerical variables and consider absolute APE values.

3.6.2. Accumulated local effects (ALE) plot

The ALE plot algorithm overcomes the lack of interpretability of *black box* models by visually describing the effect of a regressor on the predicted response (Apley & Zhu, 2020), 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, quadratic (U-shaped or inverse U-shaped), etc. Because some *black box* models, such as the random forest, are non-differentiable, partial derivatives are approximated by finite differences in the predicted response variable, within K intervals for regressor x_j, to block the effect of other, usually correlated, features. ALE method reduces the complex prediction function to a function that depends on only one (or two) features and this reduction is performed by averaging the effects of the other features (Molnar, 2022), which means that the effects are not the traditional *ceteris paribus* effects.

The estimated ALE main effect (first-order effect) for regressor x_j is computed by first segmenting the range of values of x_j into K intervals or bins. For $k = 0, 1, \dots, K$, the interval bound values $Z_{k,j}$ are (k/K)-quantiles of the empirical distribution of x_j, whereas $Z_{0,j}$ 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

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

where $k_j(i)$ is the index of the interval into which an observation value i of x_j falls; $n_j(k)$ 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,

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

Plotting $\widehat{ALE}(x_j)_C$ *versus* x_{ji} correctly reveals the true effect of x_j on the predictive function of y. In the analysis, I set $K = 150$.

4. DATA AND DESCRIPTIVE STATISTICS

4.1. Data and poverty measure

The data source for this dissertation is the Benin Harmonized Survey of Household Living Conditions 2018-2019 (EHCVM 2018/19), retrieved from the World Bank microdata database. The survey was conducted with a two-wave approach to account for seasonality of consumption, and used the 2013 Census of Population and Housing as the sampling frame. A two-stage sampling methodology was employed, where in the first stage the enumeration areas were selected with probability proportional to size (measure of size = number of households) and, in the second stage, 12 households were randomly selected in each enumeration area. This methodology does not suggest an endogenous

sampling of the data. The survey design defined the domains as country, urban and rural areas, and each of the 12 regions of the country, containing a sample data with 8,012 households, totalizing 42,343 household members.

I rely on Diaz-Bonilla and Sabatino (2022) to compute the MPM, combining the monetary dimension with two non-monetary dimensions (education and basic infrastructures), while the indicators for parameter cut-off reflect both a different perspective of deprivation and the availability of data in the survey. The sum of the weights from all indicators determines the value of the MPM, where higher values mean more deprived individuals or households. I provide in the annex the explanation of the MPM construction framework.

To create the database, I retrieved data from 25 files available at the survey database, matching the information with a unique generated household code (*hhid*) and household member code (*hhid_i*). The database comprises 231 features (see TABLE VIII in supplementary material), out of which 73 correspond to the exact original data in the survey dataset, 37 to new variables derived from available information in the survey dataset, and 121 to survey responses that I adjusted either to have a categorical variable with a lower number of categories either to impute blank cells with implicit information contained in other response variables of the survey dataset (most of them in the same source file). I also imputed data to households and individuals not reported in some source files. The 231 features cover regional, community, household, and individual characteristics, being a larger set of variables under analysis than those used in previous research for Benin, as no machine learning technique (to the best of our knowledge) has been applied previously to analyze poverty in this country.

TABLE II displays the set of $P = 21$ regressors, from the total 231 features, to estimate the ordered probit and fractional probit models and compare their results to the ones of the machine learning algorithms. In this smaller dataset, I account for agroecological differences between regions and shocks related to nature events to analyze the effect of weather and environmental conditions in the likelihood of an individual to be more or less deprived. In addition to natural disasters, the analysis also considers other forms of shock, such as shock in agriculture, at household level, at macroeconomic level,

TABLE II

SMALL SET OF REGRESSORS (P=21)

Variable	Description	Categories
head_age	HH head age	
hh_size	Household size	
hh_depratio_c	Child dependency ratio	
c_inequality	Gini index of per capita expenditure	
geo_aez	Agro-ecological zone	5
geo_urbrur	Place of residence	2
head_sex	HH head sex	2
head_educ	HH head education	4
ind_edu_mother	HH member mother's education	4
head_emp_sector12m	HH head employment sector/status	5
head_mstat	HH head marital status	5
ind_ethnic	HH member ethnic group	6
ind_migration	HH member previous place of residence	4
hh_trf_receive	Remittances from non-HH members	2
hh_fin_access	HH 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

Note. HH = household; N = 42343 for each variable.
Source: EHCVM 2018/19, World Bank, and author calculations.

to household business, to household income, and security-related shock (see TABLE I in supplementary material).

In determining the likelihood of being more or less deprived, I consider the role of financial access, using as proxy variable the ownership of any type of financial account or prepaid card, as well as of cash remittances from non-household members. Regarding education, I analyze the level of education of the household head, who is male dominant in Benin (84.1% in the sample); and the individual's mother education, in an effort to capture the intergenerational impact of female education on poverty.

By comparing the five major ethnic groups of Benin to a group that includes the remaining 46 ethnic groups, I examine the statistical significance of ethnicity in the likelihood of being more or less deprived. An ethnic group is a social group that shares a common and distinctive history, culture, religion, language, or the like (Olarewaju &

Olarewaju, 2021), which can encompass unobserved factors that may influence the likelihood of being more or less deprived.

I consider in the analysis the role of migration, as a means to improve life conditions, taking into account whether migrants to Benin came from urban/rural areas or abroad, as well as the effects of a proxy of income inequality within subregions; the impact of basic infrastructures (road access, electric distribution and running water networks); and the role of governance and management affecting human capital, using as proxy variables the presence of health and school committees structures at communities. Finally, I investigate the effects of other commonly analyzed regressors, such as age, sex, marital status and employment sector/status of household head, the household size, the child dependency ratio, and the place of residence (urban/rural).

4.2. Descriptive statistics

Benin is one of the poorest countries in the world. The World Bank (2023) estimates that around 83% of its population were living below $6.85 a day (2017 PPP) in 2018, and 19.9% were living below $2.15 a day, whereas at national poverty line the monetary poverty reached 38.5% of its population. The country has an economy reliant on agriculture, which generates around 70% of employment and 30% of GDP, dependent on rainfall and vulnerable to climate change and to unfavorable variations in global cotton and oil prices (IFAD, 2023; AFDB, 2023; World Bank, 2023). The pace of population growth is deemed a challenge because the increasing number of births puts pressure on the economy; young people are forced to leave the rural areas in search of work in urban areas (IFAD, 2023); and extended family members are often forced to live together due to a lack of capital, deteriorating their living conditions (The Borgen Project, 2020).

In the sample, TABLE III shows that the multidimensional poverty measure MPM has a mean value of 0.51 and a standard deviation of 0.26. The minimum value is 0 and the maximum value is 1, which corresponds to 5.16% and 2.44% of the sample observations, respectively, indicating that there is no concentration of extremum values for the fractional variable. The empirical distribution of MPM is slightly skewed and has a low kurtosis, indicating the absence of outliers (see also Figure 1).

TABLE III

MAIN DESCRIPTIVE STATISTICS OF MPM

Mean	Median	SD	Min	Max	Kurtosis	Skewness
0.51	0.50	0.26	0	1	2.15	-0.07

Note. SD = standard deviation.
Source: EHCVM 2018/19, World Bank, and author calculations.

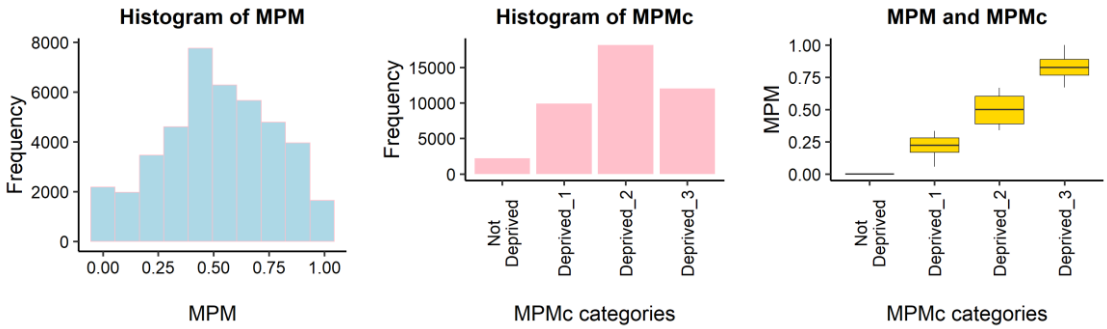


FIGURE 1 - Distribution of MPM and MPMc.

TABLE IV presents the descriptive statistics for the ordinal dependent variable MPMc (i.e., classes of multidimensional poverty). The class of individuals without any form of deprivation ("Not Deprived") has an MPM of 0, whereas the most deprived class of individuals (Deprived_3) comprises about 28.5% of the sample data, with mean MPM of 0.83.

TABLE IV

MAIN DESCRIPTIVE STATISTICS OF MPMc CATEGORIES

MPMc	MPM interval	Distribution of **MPMc**			Descriptive statistics of **MPMc** with MPM values					
		Freq.	Percent	Cum.	Mean	SD	Min	Max	Kurtosis	Skewness
Not Deprived	0	2 185	5.2	5.2	-	-	-	-	-	-
Deprived_1	]0 1/3]	9 913	23.4	28.6	0.22	0.08	0.06	0.33	1.82	-0.21
Deprived_2	]1/3 2/3]	18 195	43.0	71.5	0.51	0.10	0.34	0.67	1.68	0.16
Deprived_3	]2/3 1]	12 050	28.5	100	0.83	0.09	0.67	1.00	2.06	0.20
	Total	42 343	100							

Note. SD = standard deviation
Source: EHCVM 2018/19, World Bank, and author calculations.

Approximately 43% of observations consist of Deprived_2 individuals, with mean MPM 0.51 as for the total sample, while the remaining 23.4% of the sample consists of Deprived_1 individuals, the class of less severe multidimensional poor individuals, with

mean MPM of 0.22. The MPM values are nearly symmetrical and highly concentrated within each MPMc category (see also Figure 1).

The descriptive statistics of numerical regressors by MPMc category (see TABLE VIII in annex, and Figure 2) indicate that the mean and median values of the age of the household head, household size, and child dependency ratio are higher for the most deprived class of individuals, while the opposite occurs for inequality. Regarding the total sample, the household head age, household size, child dependency ratio and inequality present, respectively, a mean value of 44.4, 7, 125.3, and 0.33, with a standard deviation of 13.2, 3.6, 95.7, and 0.05.

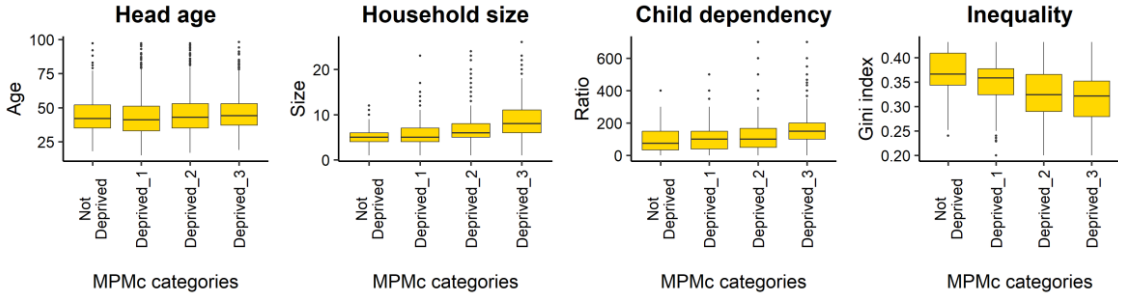
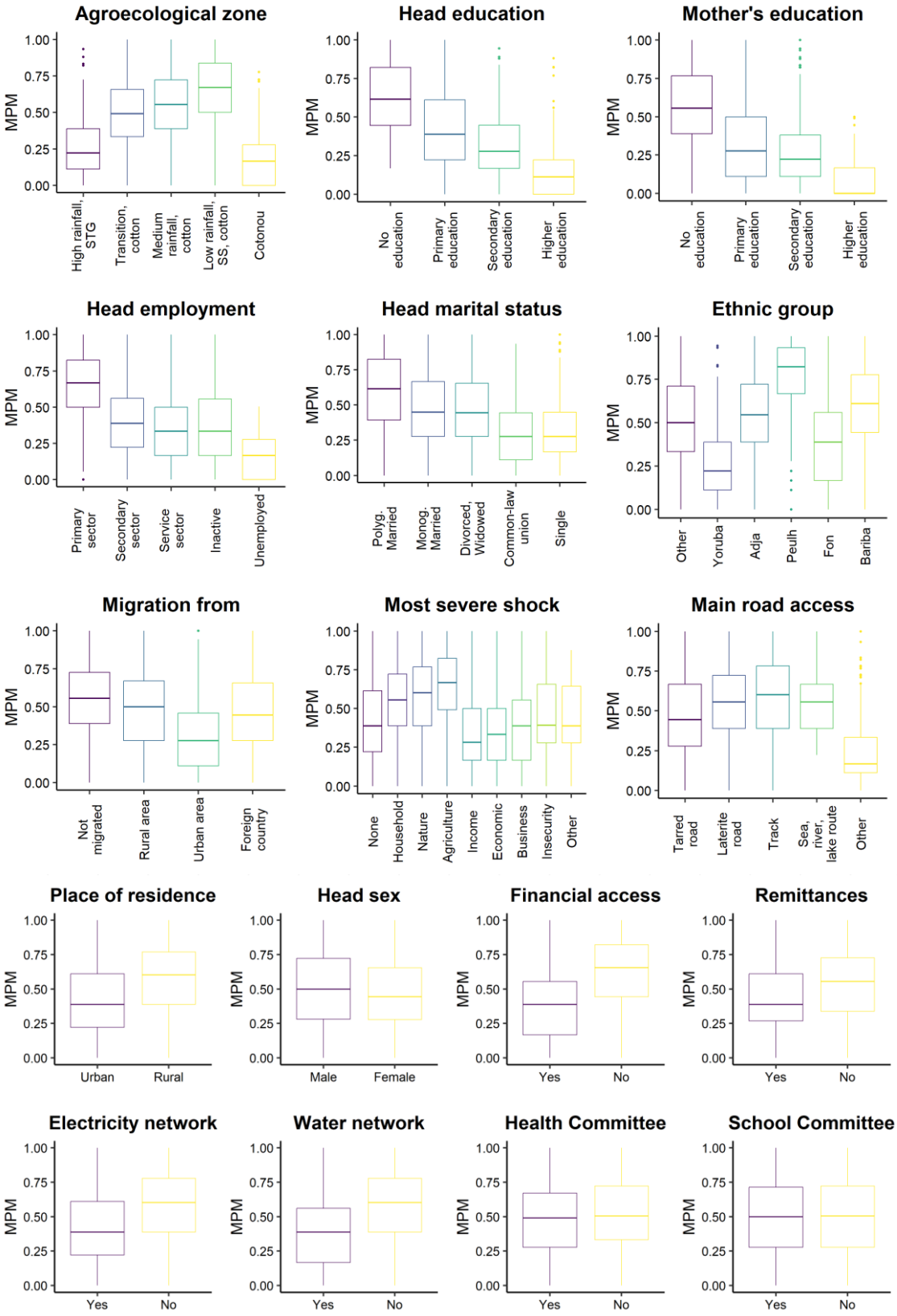


FIGURE 2 - Distribution of numerical variables by MPMc category.

Regarding categorical variables, individuals with higher levels of deprivation reside predominantly in less endowed agroecological zones (see TABLE IX in annex, and Figure 3). In the sample, 48.1% of individuals belonging to the Deprived_3 class reside in Sudan-Sahelian areas, characterized by low rainfall, whereas the class of "Not Deprived" individuals reside predominantly in high rainfall areas (37.4%) and Cotonou (37.0%), the economic hub and largest city of Benin and a littoral area characterized by tropical wet and dry climate.

Similarly, compared to lower levels of deprivation, a relatively higher proportion of individuals in the most deprived MPMc classes reside in rural area (above 60%), did not migrated (above 80%), belong to a household that does not receive cash remittances (85.4% for Deprived_3, compared to 63.7% for "Not Deprived"), has no financial access (around 80% for Deprived_3, compared to only 6% for "Not Deprived") and experience more shocks (with relatively higher proportions of household-related shocks, as well those related to nature and agriculture).



Note. STG = subequatorial-tropical-guinean; SS = Sudan-Sahelian
Source: EHCVM 2018/19, World Bank, and author calculations.

FIGURE 3 - Distribution of categorical variables by MPM.

Concerning education, a relatively higher proportion of individuals in Deprived_3 class live in households headed by a non-educated person (82.4%, compared to 0%, 22.4%, and 65.4% of individuals in "Not Deprived", Deprived_1, and Deprived_2 classes, respectively) and has a non-educated mother (95.8%, compared to 31.1%, 62.9%, and 88.5% of individuals in "Not Deprived", Deprived_1, and Deprived_2 classes, respectively). These individuals have a household head who works mainly in the primary sector (79.4% among Deprived_3, compared to 2.8% of those in "Not Deprived" class), as well reside in communities with lower public infrastructures (such as tarred road and electricity and running water network) and health management structures.

Although most individuals in all MPMc categories live in monogamous households (above 50%), a relatively higher proportion of individuals of the most deprived classes live in polygamous households (39.1% for Deprived_3, compared to 5.9%, 15.7%, and 24.5% for "Not Deprived", Deprived_1, and Deprived_2, respectively).

Against expectations, the descriptive statistics reveals that a relatively lower proportion of "Not Deprived" individuals reside in communities with a school committee (63.6%, compared to 67.3% for Deprived_3 class) and belong to a male-headed household (81.4%, compared to 87.9% for Deprived_3 class). Finally, Figure 3 depicts lower median MPM for the Yoruba ethnic group, with the highest share observed among "Not Deprived" individuals (5.9%, compared to 0.5% for Deprived_3 class).

When conducting econometric analysis, the possibility that regressors are excessively correlated is a natural concern. The pairwise correlation and the variance inflation factor suggest that there is no strong linear dependence between the regressors, thereby suggesting the absence of multicollinearity issues in the models (see TABLE II and III, and Figure 1 in supplementary material).

5. EMPIRICAL RESULTS

5.1. Interpreting the glass box models versus ALE plot

The RESET test on ordered and fractional probit models, using both one and two powers of the fitted index function, suggests a valid model specification for the set of 21 regressors when I consider nonlinear effects on age, household size, and child dependency and some interactions for household head sex, infrastructure variables, and household

head education. TABLE VI provides the APE, their statistical significance, and cluster-robust standard errors, whereas TABLE XI in annex provides the estimation output.

In both ordered and fractional probit models, most variables are statistically significant and exhibit the expected direction of impact, whereas the same intuition can be drawn from the ALE main effects plots in most cases (Figure 4 for random forest regression, and Figure 5 for random forest classification problem).

The estimated APE suggest that individuals are less likely to be in the most deprived classes (Deprived_2 and Deprived_3) if the household head and their mother have any level of education compared to having no education; these effects decrease the MPM values, indicating lower risk of multidimensional poverty and so greater well-being. The same intuition is drawn from Figure 4, while Figure 5 shows that higher education of household head was determinant for the prediction of "Not Deprived" individuals; primary and secondary education of household head were determinant for predicting Deprived_1 individuals; and having no education was determinant for predicting Deprived_3 individuals. On the other hand, primary, secondary and higher education of individual's mother had the highest ALE main effect on classifying Deprived_1 individuals, while no education was determinant to predict the most deprived classes.

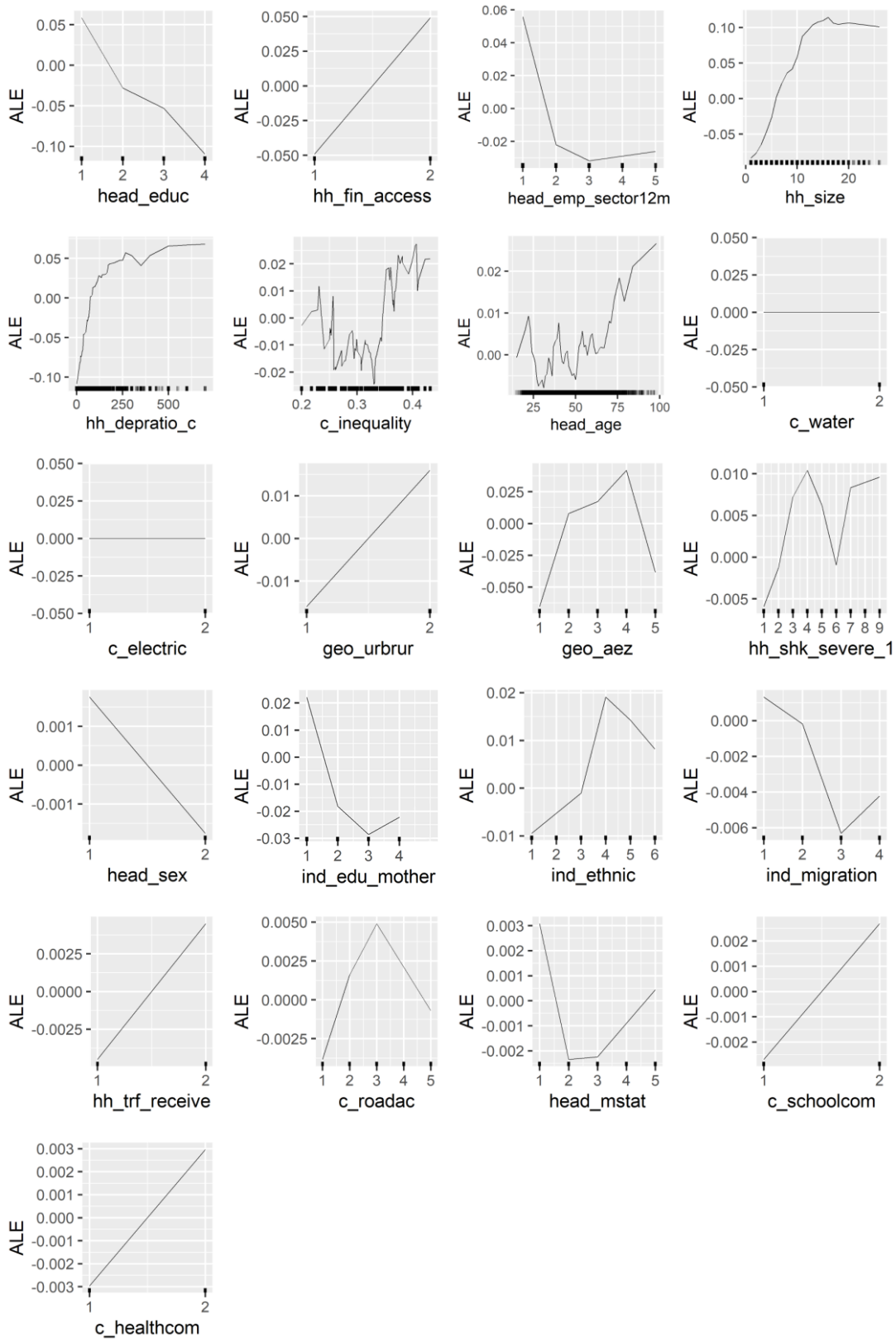
The ordered probit APE suggests that individuals are more likely to be in the most deprived class if the household size is larger, in line with the fractional probit APE. The estimated quadratic relationship, as practice in literature, was statistically significant in both parametric models, which supports the idea of some economy of scale for larger households (Lanjouw & Ravallion, 1995). Figure 4's ALEplot suggests a monotonic deterioration of well-being with increasing household size up to 16 individuals, after which the worsening effect becomes slightly less severe, but it does not suggest an inverted U-shaped relationship. In addition, Figure 5 indicates that larger households contributed to the prediction of Deprived_3 individuals. Similarly, the ALEplot depicts a nonlinear decline in well-being as child dependency increases, and after the ratio of 350 the effects become steadily more severe. This effect was determinant for predicting Deprived_2 individuals. The direction of this effect is the same in the parametric models, which evidence a statistically significant quadratic relationship.

TABLE V

AVERAGE PARTIAL EFFECTS (APE) - MODEL A6

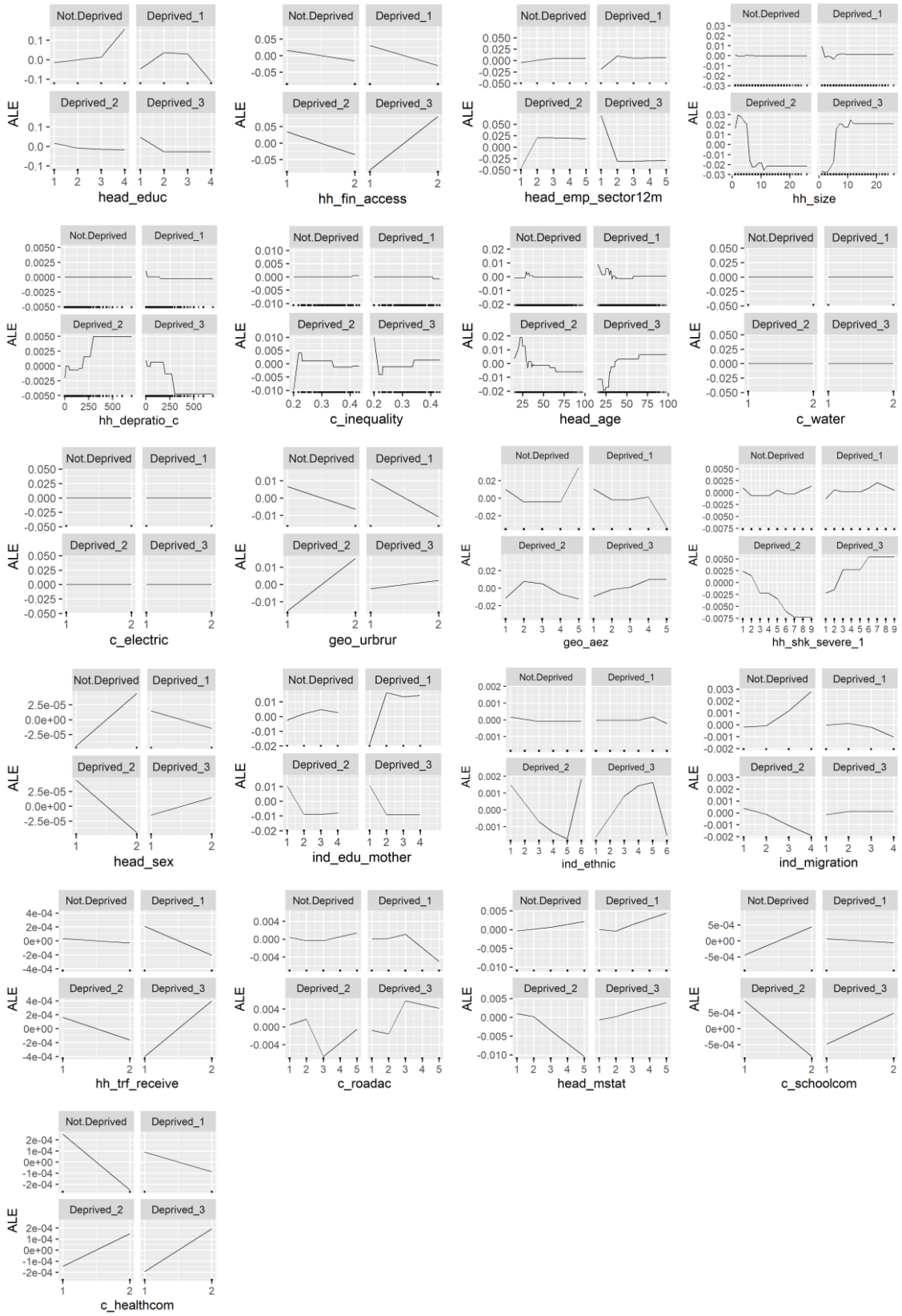
Variables	Ordered Probit								Fractional Probit	
	Not Deprived		Deprived_1		Deprived_2		Deprived_3		MPM	
	dy/dx	SE	dy/dx	SE	dy/dx	SE	dy/dx	SE	dy/dx	SE
head_age	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00	0.000	0.00
hh_size	-0.008 ***	0.00	-0.015 ***	0.00	-0.003 ***	0.00	0.026 ***	0.00	0.019 ***	0.00
hh_depratio_c	0.000 ***	0.00	0.000 ***	0.00	0.000 ***	0.00	0.000 ***	0.00	0.000 ***	0.00
c_inequality	-0.021	0.02	-0.045	0.04	-0.017	0.02	0.082	0.08	0.118 **	0.05
geo_aez (base = High rainfall...)										
Transition, cotton…	-0.039 ***	0.00	-0.099 ***	0.01	0.011 **	0.01	0.126 ***	0.01	0.096 ***	0.01
Medium rainfall, cotton…	-0.042 ***	0.00	-0.110 ***	0.01	0.008 *	0.01	0.144 ***	0.01	0.109 ***	0.01
Low rainfall, cotton…	-0.047 ***	0.00	-0.130 ***	0.01	0.000	0.01	0.177 ***	0.01	0.133 ***	0.01
Cotonou	0.017 **	0.01	0.030 **	0.01	-0.017 *	0.01	-0.030 **	0.01	-0.030 **	0.01
geo_urbrur (base = Urban)	-0.007 ***	0.00	-0.016 ***	0.01	-0.005 ***	0.00	0.028 ***	0.01	0.014 ***	0.01
head_sex (base = Male)	0.015 **	0.01	0.025 ***	0.01	-0.041 ***	0.01	0.001	0.01	-0.019 **	0.01
head_educ (base = No education)										
Primary education	0.023 ***	0.00	0.079 ***	0.01	-0.015 *	0.01	-0.088 ***	0.01	-0.069 ***	0.01
Secondary education	0.046 ***	0.00	0.132 ***	0.01	-0.029 ***	0.01	-0.149 ***	0.01	-0.110 ***	0.01
Higher education	0.084 ***	0.01	0.208 ***	0.05	-0.063 ***	0.02	-0.229 ***	0.04	-0.191 ***	0.03
ind_edu_mother (base = No education)										
Primary education	0.034 ***	0.00	0.079 ***	0.01	0.005 **	0.00	-0.117 ***	0.01	-0.082 ***	0.01
Secondary education	0.042 ***	0.00	0.093 ***	0.01	0.001	0.00	-0.136 ***	0.01	-0.100 ***	0.01
Higher education	0.087 ***	0.01	0.154 ***	0.02	-0.039 ***	0.01	-0.203 ***	0.01	-0.194 ***	0.02
head_emp_sector12m (base = Primary										
Secondary sector	0.014 **	0.01	0.038 ***	0.01	-0.004	0.01	-0.048 ***	0.01	-0.041 ***	0.01
Service sector	0.025 ***	0.01	0.059 ***	0.01	0.001	0.01	-0.085 ***	0.01	-0.065 ***	0.01
Inactive	0.024 ***	0.01	0.046 ***	0.01	-0.012	0.01	-0.059 ***	0.02	-0.048 ***	0.01
Unemployed	0.047 *	0.03	0.070	0.05	-0.021	0.04	-0.096	0.09	-0.121 ***	0.03
head_mstat (base = Polyg. Married)										
Monog. Married	-0.002	0.00	-0.005	0.01	-0.002	0.00	0.009	0.01	0.001	0.01
Common-law union	0.000	0.01	-0.002	0.01	-0.005	0.01	0.007	0.03	-0.008	0.02
Divorced, Widowed	-0.004	0.01	-0.008	0.01	-0.003	0.00	0.014	0.02	0.009	0.01
Single	-0.011 **	0.01	-0.024 **	0.01	-0.012 **	0.01	0.047 **	0.02	0.040 ***	0.01
ind_ethnic (base = Other)										
Yoruba	0.016 **	0.01	0.030 **	0.01	0.006 ***	0.00	-0.052 ***	0.02	-0.032 **	0.01
Adja	-0.012 ***	0.00	-0.028 ***	0.01	-0.013 ***	0.00	0.054 ***	0.01	0.036 ***	0.01
Peulh	-0.019 ***	0.00	-0.047 ***	0.01	-0.025 ***	0.01	0.092 ***	0.02	0.070 ***	0.01
Fon	-0.004	0.00	-0.008	0.01	-0.003	0.00	0.014	0.01	0.006	0.01
Bariba	0.001	0.00	0.003	0.01	0.001	0.00	-0.005	0.01	-0.008	0.01
ind_migration (base = Not migrated)										
Rural area	0.001	0.00	0.001	0.00	0.000	0.00	-0.002	0.01	0.006	0.01
Urban area	0.008 ***	0.00	0.017 ***	0.01	0.005 ***	0.00	-0.031 ***	0.01	-0.019 ***	0.01
Foreign country	0.002	0.00	0.004	0.01	0.001	0.00	-0.007	0.01	-0.007	0.01
hh_trf_receive (base = Yes)	-0.003	0.00	-0.007	0.01	-0.003 *	0.00	0.013	0.01	0.010 **	0.01
hh_fin_access (base = Yes)	-0.031 ***	0.00	-0.089 ***	0.01	-0.022 ***	0.00	0.142 ***	0.01	0.090 ***	0.01
hh_shk_severe_1 (base = None)										
Shk_Household	-0.006 **	0.00	-0.013 **	0.01	-0.004 **	0.00	0.023 **	0.01	0.009	0.01
Shk_Nature	-0.012 ***	0.00	-0.027 ***	0.01	-0.010 ***	0.00	0.049 ***	0.01	0.026 ***	0.01
Shk_Agriculture	-0.013 ***	0.00	-0.029 ***	0.01	-0.011 ***	0.00	0.053 ***	0.02	0.025 ***	0.01
Shk_Income	-0.010	0.01	-0.021	0.01	-0.007	0.01	0.038	0.03	0.013	0.02
Shk_Economic	-0.003	0.00	-0.006	0.01	-0.002	0.00	0.011	0.01	-0.015 *	0.01
Shk_Business	-0.006	0.01	-0.013	0.01	-0.004	0.00	0.024	0.02	0.007	0.01
Shk_Insecurity	0.007	0.01	0.014	0.01	0.002	0.00	-0.023	0.02	-0.021 *	0.01
Shk_Other	-0.009	0.01	-0.019	0.03	-0.006	0.01	0.034	0.06	0.013	0.03
c_roadac (base = Tarred road)										
Laterite road	0.000	0.00	-0.001	0.01	0.000	0.00	0.001	0.01	-0.003	0.01
Track	0.001	0.00	0.003	0.01	0.001	0.00	-0.005	0.01	-0.001	0.01
Sea, river, lake ro..	0.005	0.01	0.011	0.01	0.004	0.00	-0.020	0.02	-0.013	0.01
Other	-0.009 **	0.00	-0.020 **	0.01	-0.009 *	0.01	0.038 **	0.02	0.017	0.01
c_electric (base = Yes)	-0.015 ***	0.00	-0.027 ***	0.01	0.000	0.00	0.042 ***	0.01	0.024 ***	0.01
c_water (base = Yes)	-0.014 ***	0.00	-0.026 ***	0.01	0.000	0.01	0.040 ***	0.01	0.037 ***	0.01
c_healthcom (base = Yes)	0.004 *	0.00	0.004	0.00	-0.005	0.00	-0.003	0.01	-0.006	0.01
c_schoolcom (base = Yes)	-0.002	0.00	-0.004	0.00	-0.002	0.00	0.008	0.01	0.014 ***	0.01
N	42343								42343	
Log pseudolikelihood	-33541								-17978	
Pseudo R-squared	0.35									
Wald Chi-square	4588								8478	
Prob > chi2	0.00								0.00	
RESET TEST J=1 (*p* -value)	0.89								0.82	
RESET TEST J=2 (*p* -value)	0.29								0.06	

Note. ***$p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.1$; SE = cluster-robust standard error



Note. For binary variables: 1 = Yes, Urban or Male; and 2 = No, Rural, or Female.

FIGURE 4 - ALE Main Effects of Regressors on MPM



Note. For binary variables: 1 = Yes, Urban or Male; and 2 = No, Rural, or Female.

FIGURE 5 - ALE Main Effects of Regressors on MPMc classes

Figure 4 shows that inequality evidence highly nonlinear ALE main effects on MPM, with Gini indices greater than 0.35 consistently having a negative effect on individual's well-being; an effect that is evidenced by the estimated linear relationship on the fractional probit regression. Moreover, the ALE main effect on MPM is positive for some lower levels of inequality, which can be explained by subregions with a predominance of poor people at similar levels of poverty. Indeed, Figure 5 supports this idea, as both very low and high values of inequality positively determined the classification of Deprived_3 individuals. Interestingly, low rainfall areas and Cotonou, one of the subregions with the greatest inequality in Benin, were the agroecological zones with the highest ALE main effect for predicting Deprived_3 individuals. Living in a low rainfall, Sudan Sahelian area, compared to high rainfall regions, increase the likelihood of being categorized as Deprived_3, according to ordered probit; and, among the agroecological zones, it has the highest worsening effect on welfare (fractional probit and random forest regression).

Figure 4's ALE main effects evidence that the age of household head, the only variable which was not statistically significant in both econometric models, has also highly nonlinear effect on MPM in the *black box* model. It suggests a negative effect on individual's well-being when household heads are younger than 25 years old and a consistent worsening, although nonmonotonic, of the welfare condition when household heads are older than 62 years old, which is fairly reasonable for a country severely lacking an adequate social protection system. About 85% of the labor force works in the informal economy (World Bank, 2023), which contributes to degrade the living condition of people at retirement age. Figure 5 indicates that older household heads positively determined the classification of Deprived_3 individuals.

Working in an economic sector other than the primary sector appears to reduce the likelihood of experiencing extreme deprivation. Unexpectedly and relative to the primary sector, inactive and unemployed household heads exhibit a decreased APE effect on MPM, unintuitively suggesting better welfare conditions. Meanwhile, Figure 5's ALEplot shows that the main effect of these two factors contributed more for predicting Deprived_2 individuals, who make up 43% of the sample observations and are the second most deprived class. Concerning the financial sector, all models support that financial inaccessibility worsens welfare conditions and increases the likelihood of more severe

poverty. In addition, the fractional probit APE and random forest ALE main effects evidence that individuals whose households do not receive remittances are less well-off.

Regarding shocks, both parametric models and the random forest regression model corroborate that a shock in agriculture has the greatest negative impact on well-being, whereas Figure 5 shows that almost all shocks contributed positively to the classification of Deprived_3 individuals, and household-related shocks were positively determinant for predicting Deprived_2 individuals. Individuals residing in rural areas have a greater likelihood of belonging to the most deprived class. In addition, based on the findings, those who migrate from urban areas are less likely to experience extreme poverty, which appears to be associated with a higher standard of living.

The results suggest that unobserved factors related to ethnicity are statistically significant in explaining the likelihood of experiencing more or less deprivation, with the Yoruba ethnic group being associated with better welfare condition, and Peulh and Adja ethnic groups associated with worse welfare condition, relative to other ethnic groups. According to *Wikipedia* (2023), the Yoruba are among the most urbanized people in Africa, and for centuries before the arrival of the British colonial administration most Yoruba already lived in well-structured urban centers organized around powerful city-states. According to Figure 5, Adja, Peulh, and Fon ethnic groups contributed positively for predicting Deprived_3 individuals, and Bariba, Yoruba and other ethnic groups determined positively the classification of Deprived_2 individuals.

Regarding the sex of household head, both parametric APE and Figure 4's ALE effects suggest that female-headed households are associated with greater living standard; meanwhile, it is possible to notice in Figure 5 that although female sex contributed for predicting not deprived individuals, it also determined positively the prediction of the most deprived individuals. This result resembles Attanasso's (2005) findings and supports IFAD's (2023) assertion that poverty is lower among female-headed households in Benin yet women are more vulnerable and lack economic opportunities.

Relative to polygamous household heads, only single household heads were statistically significant for extreme deprivation. Figure 4 shows that polygamous and single household heads are associated with worse welfare conditions, with the former having a positive effect on predicting Deprived_2 individuals and the latter having the

highest ALE positive main effect on predicting Deprived_3 individuals relative to other marital statuses.

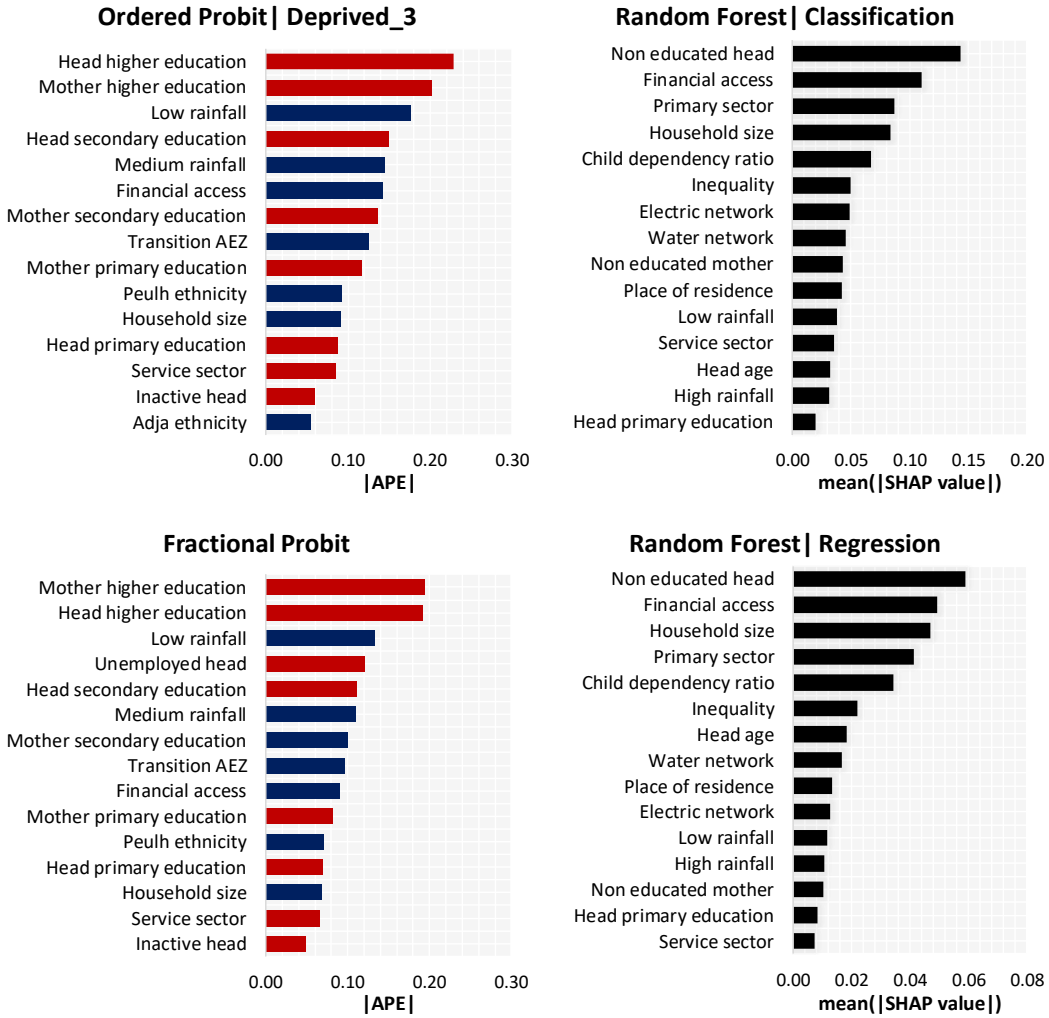
The parametric and non-parametric models diverged mostly on infrastructure effects. Although the public infrastructures of electricity and water are statistically significant in reducing the likelihood of extreme poverty and improving well-being, both ALEplot main effects of the random forest show no effect of these variables in the response functions. Meanwhile, Apley (2021) notes that variables with no main effects may become important when higher order effects are examined because they may interact strongly with other variables to predict the response function.

The econometric results diverge in evidencing statistical significance for road access, and the presence of health and school committees at communities. In line with the expectation, the ALEplot main effects suggest that the absence of public health and education management structures at communities is associated with worsening welfare conditions. It also suggests that track roads were the factor more associated with higher values of MPM, indicating lower welfare condition.

5.2. Important variables according to APE and SHAP values

Figure 6 depicts the most important regressor's effects for each model (see TABLE VI in supplementary material). All models corroborate indicating education (of the household head and individual's mother), financial access, agroecological zone, household size, and employment sector among most important variables.

Variables with highly nonlinear effects (inequality and age) ranked among the top 15 most important effects only in random forest models. In addition, only the random forest models deemed the nonlinear effect of the child dependency ratio to be among the most significant. Moreover, this *black box* model, by computing more complex interactions between variables, were able to present a wider range of important variables. On the other hand, only the parametric models regard ethnicity and employment status among the top 15 factors.



Note. Red bar = negative APE; blue bar = positive APE; black bar = no direction of effect indicated.

FIGURE 6 - Variable Importance | absolute APE and SHAP values.

5.3. Out-of-sample performance

Table VII presents the out-of-sample performance of the models (see also Figure 9 in annex). The ordered probit model classified correctly 62.5% of test observations, compared to 97.5% achieved by the random forest model. Both models had the lowest performance in predicting "Not Deprived" individuals, with an accuracy of 65.7% and F1-score of 42.5% for the ordered probit model, compared to 91.4% and 89% for the random forest, respectively.

The ordered probit had higher accuracy predicting Deprived_1 individuals (74.5%) and better F1-score in predicting Deprived_2 class (64.6%), whereas the random forest

TABLE VI

OUT-OF-SAMPLE PERFORMANCE

Classification

	Ordered Probit			Random Forest		
	F1-score	Accuracy	Error rate	F1-score	Accuracy	Error rate
Not Deprived	42.5	65.7	34.3	89.0	91.4	8.6
Deprived_1	61.1	74.5	25.5	95.3	97.3	2.7
Deprived_2	64.6	67.6	32.4	98.3	98.6	1.4
Deprived_3	63.0	74.0	26.0	99.4	99.6	0.4
Overall		**62.5**	**37.5**		**97.5**	**2.5**

Regression

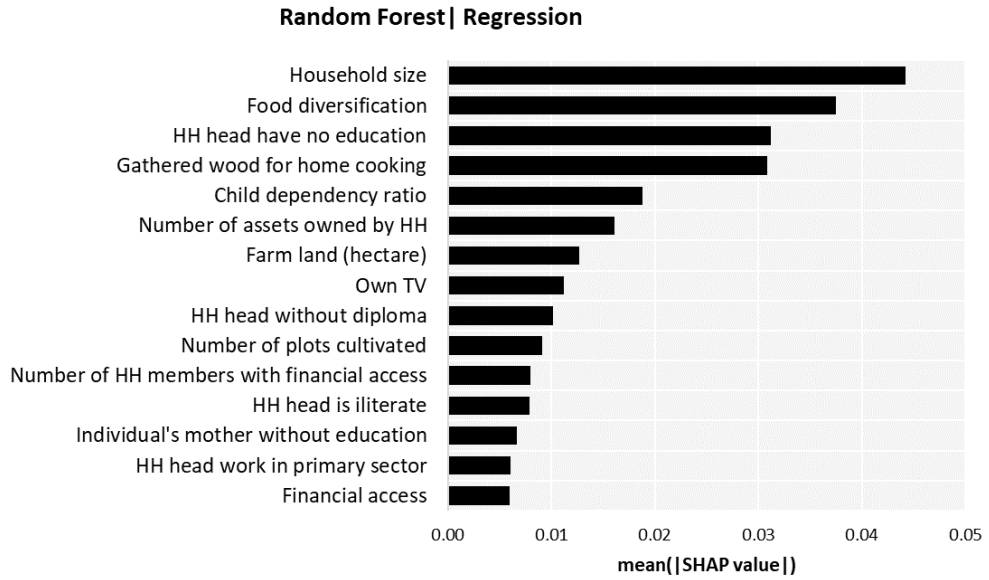
Model	MSE
Fractional Probit	0.435
Random Forest	0.002

had better and consistent performance in predicting Deprived_3 class (accuracy of 99.6% and F1-score of 99.4%). In the regression problem, the fractional probit had a MSE of 0.435, while the random forest predicted MPM with a MSE of 0.002.

5.4. Poverty profile in 2018-2019

Figure 7 depicts the 15 most important indicators of the poverty profile in 2018/2019 for Benin, derived from the large dataset of regressors (see TABLE VII in supplementary material for top 60 indicators). The five most important variables are: household size; food diversification; household head without education; households that gather wood for home cooking; and child dependency ratio. Some of the most important immediate determinants of poverty are still suggested among the top 15 indicators of the poverty profile: household size; education (of both household head and individual's mother); child dependency ratio; financial access; and employment sector.

The pure poverty-targeting classification exercise produced an out-of-sample accuracy of 98.9%, while the MSE stabilizes around 0.001 in the regression problem (see Figure 9 and TABLE XII in annex).



Note. HH = Household

FIGURE 7 - Indicators of poverty profile of Benin | 2018/2019

6. CONCLUSION

In this dissertation, I performed a cross-sectional analysis of multidimensional poverty in Benin, using both traditional and machine learning methods.

In most cases, the effects of regressors on the response function have the same expected direction of impact in both *glass box* and *black box* models, whereas the accumulated local effects plot on random forest provided a deeper intuition on the effects. This non-parametric approach suggests a highly nonlinear relationship between the individual's welfare condition and the age of household head and inequality, as well as a nonlinear but non-concave relationship with household size and child dependency ratio. While all models corroborate suggesting that education (of both household head and individual's mother), agroecological zones, financial access, household size, and employment sector are among most important variables associated with welfare condition, only the *black box* model, through SHAP values, ranked the variables with highly nonlinear effects among the most important regressors, as well child dependency ratio. Moreover, the random forest, by computing more complex interactions between variables, was able to present a wider range of important variables in the top 15 factors. The out-of-sample error rate of the random forest in the classification problem was 2.5%,

compared to 37.5% of the ordered probit model, while the MSE of the random forest in the regression problem was 0.002, compared to 0.435 of the fractional probit model.

From the large dataset of variables, the household size, food diversification, household head without education, households that gather wood for home cooking, and child dependency ratio ranked the five most important indicators describing the profile of a poor Beninese in 2018/19.

I performed the analysis using an adapted measure of multidimensional poverty. In particular, I defined a decreasing exponential function for calculating the monetary parameter weights and I did not evaluate whether a different setting would produce different results for the analysis. Concerning the model for policy-targeting approach, Sohnesen and Stender (2017) find that applying a technical model within a single year may not always be sufficient for accurate predictions of poverty over time. Other machine learning algorithms may be more suited to a different or panel dataset, despite the fact that random forest has demonstrated excellent predictive performance in this cross-sectional analysis.

According to Apley (2021), a venue for deriving a variable importance measure from the ALE function is possible, based on global sensitivity analysis, where the importance of a predictor takes into account not only its main effect but also all interactions it might have. Future research relying on this visualization approach may provide a deeper intuition on immediate determinants of poverty.

Despite these limitations, this study has evidenced results broadly in line with literature in Africa and Benin, with all models indicating that education is the most important "proximate" determinant of the welfare condition in Benin. Moreover, it suggests that is possible to combine simpler statistical learning with machine learning to provide a deeper intuition on the relationship between the regressors and a poverty measure, provided that the outcomes are interpreted with the necessary caution.

REFERENCES

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.

Adjasi, C.K., & Osei, K.A. (2007*). Poverty profile and correlates of poverty in Ghana*. International Journal of Social Economics, 34, 449-471. DOI: 10.1108/03068290710760236

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

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.

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.

Alinsato, A. S., & Houedokou, W. (2019). *Sector of Economic Activity and Poverty in Benin*.

Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., & Alyaman, M. (2021). *Poverty Classification Using Machine Learning: The Case of Jordan*. Sustainability, 13, 1412.

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.

Apley, D.W. (2021). *Interpreting Black box Supervised Learning Models Via Accumulated Local Effects*. INFORMS-QSR and ENBIS Webinar Series. Available at: https://www.youtube.com/watch?v=06knUxoig9Y&t=9s

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.

Bakar, A.A., Hamdan, R., & Sani, N.S. (2020). *Ensemble Learning for Multidimensional Poverty Classification.* Sains Malaysiana, 49, 447-459. http://dx.doi.org/10.17576/jsm-2020-4902-24

Bogale, A., Hagedorn, K., & Korf, B. (2005). *Determinants of poverty in rural Ethiopia*. Quarterly Journal of International Agriculture, 44(2), 101-120.

Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32. DOI: 10.1023/A:1010933404324.

Cho, S., & Kim, T. (2017). *Determinants of Poverty Status in Rwanda*. African Development Review, 29, 337-349. DOI: 10.1111/1467-8268.12260

Cameron, A.C., & Miller, D.L. (2015). *A Practitioner's Guide to Cluster-Robust Inference*. The Journal of Human Resources, 50(2), 317-372. DOI: 10.3368/jhr.50.2.317

D'Ambrosio, C. (2018). Handbook of Research on Economic and Social Well-Being. DOI: 10.4337/9781781953716

Datt, G., Simler, K.R., Mukherjee, S., & Dava, G. (2000). *Determinants of poverty in Mozambique*: 1996-97. DOI: 10.22004/ag.econ.16427

Datt, G., & Jolliffe, D. (2001). *Determinants Of Poverty In Egypt: 1997*. Food Consumption and Nutrition Division Discussion Paper. 75. DOI: 10.22004/ag.econ.94512

Diaz-Bonilla, C., & Sabatino, C. (2022). *April 2022 Update to the Multidimensional Poverty Measure – What's New*. Global Poverty Monitoring Technical Note 22. DOI: 10.1596/37491

Epo, B.N. (2010). *Determinants of Poverty in Cameroon: A Binomial and Polychotomous Logit Analysis*. Urban Economics & Regional Studies & Journal. DOI: 10.2139/ssrn.1424672

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

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

Gbinlo, R.E. (2020). *Drivers of Multidimensional Poverty: New Evidence in Benin*. Journal of Economics and Development, 8.

Geda, A., Jong, N.D., Kimenyi, M.S., & Mwabu, G. (2005). *Determinants of Poverty in Kenya: A Household Level Analysis*. Economics Working Papers. 200544. https://opencommons.uconn.edu/econ_wpapers/200544

Glewwe, P. (1991). *Investigating the determinants of household welfare in Cote d'Ivoire*. Journal of Development Economics, 35, 307-337. DOI: 10.1016/0304-3878(91)90053-X

Greene, W.H., & Hensher, D.A. (2009). *Modeling Ordered Choices: A Primer*. DOI: 10.1017/CBO9780511845062

Grootaert, C. (1997). *The Determinants of Poverty in Cote d'Ivoire in the 1980s*. Journal of African Economies, 6(2), 169–196. DOI: 10.1093/oxfordjournals.jae.a020925

Habyarimana, F., Zewotir, T., & Ramroop, S. (2015). *Determinants of Poverty of Households in Rwanda: An Application of Quantile Regression.* Journal of Human Ecology, 50, 19 - 30. DOI: 10.1080/09709274.2015.11906856

Hastie, T., Tibshirani, R., & Friedman, J.H. (2009). *The Elements of Statistical Learning.* 2nd Ed. Springer New York, NY. DOI: 10.1007/978-0-387-84858-7

Haughton, J.H., & Khandker, S.R. (2009). *Handbook on Poverty and Inequality.* World Bank Publications, 1-446. DOI: 10.1596/978-0-8213-7613-3

Hodonou, A., Damien, M., Gninanfon, A., & Totin, A. (2010). *Poverty Dynamics in Benin: A Markovian Process Approach.* Microeconomics: Welfare Economics & Collective Decision-Making eJournal. DOI: 10.2139/ssrn.1674633

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: with applications in R.* Springer New York, NY. DOI: 10.1007/978-1-0716-1418-1

Lanjouw, P., & Ravallion, M. (1995). *Poverty and household size.* The Economic Journal, 105(433), 1415–1434. DOI: 10.2307/2235108

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. DOI: 10.1016/j.seps.2021.101195

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. DOI: 10.1016/j.spasta.2020.100461.

Long, J.S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata,* 3rd Edition.

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. https://ssrn.com/abstract=2848477

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. DOI: 10.35741/issn.0258-2724.57.1.12

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

Molnar, C. (2022). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable.* ISBN-13 979-8411463330. Available at: https://christophm.github.io/interpretable-ml-book/shapley.html

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.* xxAI@ICML.

Mukherjee, S., & Benson, T. (2003). *The Determinants of Poverty in Malawi, 1998*. World Development, 31(2), 339–358. DOI: 10.1016/s0305-750x(02)00191-2

Muller, C. (2002). *Censored Quantile Regressions of Chronic and Transient Seasonal Poverty in Rwanda*. Journal of African Economies, 11, 503-541. DOI: 10.1093/jae/11.4.503

Okurut, F.N., Odwee, J.J.A.O., & Adebua, A. (2002). *Determinants of Regional Poverty in Uganda*. African Economic Research Consortium, 122.

Olarewaju, T., & Olarewaju, T. (2021). *Ethnic Poverty: Causes, Implications, and Solutions.* In: Leal Filho, W., Azul, A.M., Brandli, L., Lange Salvia, A., Özuyar, P.G., & Wall, T. (eds) No Poverty. Encyclopedia of the UN Sustainable Development Goals, 312–323. Springer, Cham. DOI: 10.1007/978-3-319-95714-2_124

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.

Rajaram, R. (2009). *Female-Headed Households and Poverty: Evidence from the National Family Health Survey.*

Saad, G.E., Ghattas, H., Wendt, A.T., Hellwig, F., DeJong, J., Boerma, T., Victora, C.G., & Barros, A.J. (2022). *Paving the way to understanding female-headed households: Variation in*

household composition across 103 low- and middle-income countries. Journal of Global Health, 12.

Sekhampu, T.J. (2013). *Determinants of poverty in a South African township.* Journal of Social Sciences, 34(2), 145-153. DOI: 10.1080/09718923.2013.11893126

Silva, I.D. (2008). *Micro-level determinants of poverty reduction in Sri Lanka: a multivariate approach.* International Journal of Social Economics, 35, 140-158. DOI: 10.1108/03068290810847833

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. DOI: 10.1002/pop4.169

The Borgen Project (2020). *Homelessness in Benin*. Available at: https://borgenproject.org/homelessness-in-benin-five-things-to-know/

Thoplan, R. (2014). *Random Forests for Poverty Classification*. International Journal of Sciences: Basic and Applied Research (IJSBAR), 17(2), 252–259. Retrieved from https://www.gssrr.org/index.php/JournalOfBasicAndApplied/article/view/2574

United Nations (2015). *Multidimensional Poverty*. Development Issues, 3. Available at: https://www.un.org/en/development/desa/policy/wess/wess_dev_issues/dsp_policy_03.pdf

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. Available at: https://www.un.org/ecosoc/sites/www.un.org.ecosoc/files/files/en/integration/2017/ICPR.pdf

Usmanova, A., Aziz, A., Rakhmonov, D.A., & Osamy, W. (2022). *Utilities of Artificial Intelligence in Poverty Prediction: A Review*. Sustainability 14 (21), 14238. DOI: 10.3390/su142114238

Varian, H.R. (2014). *Big Data: New Tricks for Econometrics*. Journal of Economic Perspectives, 28(2), 3-28. DOI: 10.1257/jep.28.2.3

Verme, P. (2020). *Which Model for Poverty Predictions*. Global Labor Organization (GLO), Essen. Discussion Paper No. 468. Available at: http://hdl.handle.net/10419/213811

Vowels, M.J. (2020). *Misspecification and unreliable interpretations in psychology and social science.* Psychological methods. PMID: 34647760. DOI: 10.1037/met0000429

Wikipedia (2023). *The Yoruba people*. Available at: https://en.wikipedia.org/wiki/Yoruba_people#cite_note-voices-54

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data.* The MIT Press. 2nd Edition. http://www.jstor.org/stable/j.ctt5hhcfr

World Bank (2022). Enquête Harmonisée sur le Conditions de Vie des Ménages 2018-2019. Benin, 2018 - 2019. INSAE. DOI: 10.48529/rn3k-z374. Available at: https://microdata.worldbank.org/index.php/catalog/4291/get-microdata

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

Zhao, Q., & Hastie, T. (2021). *Causal Interpretations of Black Box Models.* Journal of Business & Economic Statistics, 39(1), 272-281. DOI: 10.1080/07350015.2019.1624293

APPENDICES

MPM construction framework

The multidimensional measure of poverty is constructed with three equally weighted dimensions as in Diaz-Bonilla and Sabatino (2022), a World Bank approach, combining the monetary dimension with two non-monetary dimensions (education and basic infrastructures).

The 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, receiving a predefined weight. When selecting the indicators for each dimension, the MPM reflects both a different perspective of deprivation and the availability of data in the survey. I do not use the monetary poor versus non-monetary poor approach, in the monetary dimension, but focus on monetary poor quartiles, setting the indicator weight according to the quartile where the monetary poverty ratio of the poor individual/household falls, as I aim a deeper segregation of the MPM. Under this approach, the maximum weight of the monetary dimension (1/3) is attributed only to poor individuals/households falling in the first quartile, decreasing exponentially until the fourth quartile, where the weight reaches a value close to 1/6, and non-monetary poor individuals/households have 0 weight in all four quartiles. Also in this dimension, I use the national poverty line instead of the international poverty line to determine monetary poor households/individuals, as it is more country tailored.

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, I use the readily available data regarding drinking water and sanitation, which states if the first is potable or not and if the second is a healthy toilet or not, not explicitly relying on the concepts of limited-standard drinking water and limited-standard sanitation. In addition, allowing for a more country tailored indicator, I segregate the access to potable drinking water in terms of dry and rainy seasons.

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. TABLE VII synthetizes the information about the MPM construction framework.

TABLE VII

MPM - DIMENSIONS, INDICATORS, AND WEIGHTS

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

Note. *Percentiles of national monetary poverty ratio per person between [0 1]. This ratio corresponds to per capita total consumption expenditure to national poverty line. Values > 1 indicate non-monetary poor individual.
Source: Adapted from Diaz-Bonilla an Sabatino (2022).

Figure 8 demonstrates that the median value of the monetary poverty ratio decreases as the multidimensional level of deprivation increases, consistent with the methodology for constructing MPM, allowing individuals facing more severe monetary deprivation to be more accurately represented in MPMc classes corresponding to the most severe deprivation.

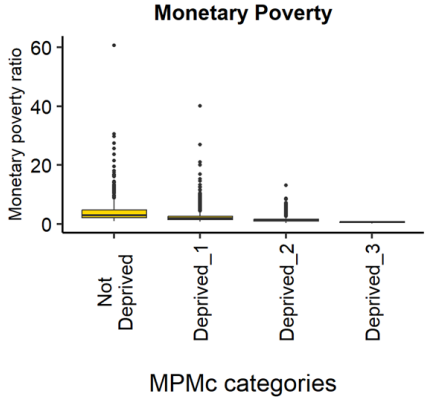


FIGURE 8 - Distribution of monetary poverty measure by MPMc.

TABLE VIII

DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

	MPMc				
Variable	Not Deprived	Deprived_1	Deprived_2	Deprived_3	Total
		Mean value			
head_age	44.0	42.9	44.6	45.5	44.4
hh_size	5.1	5.7	6.7	8.9	7.0
hh_depratio_c	89.0	97.2	121.3	161.1	125.3
c_inequality	0.37	0.35	0.32	0.31	0.33
		Standard deviation			
head_age	12.4	13.3	13.5	12.6	13.2
hh_size	2.0	2.7	3.3	4.0	3.6
hh_depratio_c	75.9	80.9	93.5	102.0	95.8
c_inequality	0.04	0.05	0.05	0.05	0.05
		Kurtosis			
head_age	3.4	3.2	3.3	3.4	3.3
hh_size	3.6	6.9	6.3	4.3	5.7
hh_depratio_c	4.4	5.6	7.0	6.3	6.6
c_inequality	3.2	2.5	2.3	2.3	2.3
		Skewness			
head_age	0.66	0.67	0.67	0.70	0.66
hh_size	0.51	1.16	1.44	1.08	1.35
hh_depratio_c	1.07	1.29	1.45	1.40	1.40
c_inequality	-0.79	-0.49	-0.14	-0.03	-0.23
		Min			
head_age	18	15	17	19	15
hh_size	1	1	1	1	1
hh_depratio_c	0	0	0	0	0
c_inequality	0.24	0.20	0.20	0.20	0.20
		Max			
head_age	97	97	97	98	98
hh_size	12	23	24	26	26
hh_depratio_c	400	500	700	700	700
c_inequality	0.43	0.43	0.43	0.43	0.43

Source: EHCVM 2018/19, World Bank, and author calculations.

TABLE IX

DISTRIBUTION OF SELECTED CATEGORICAL VARIABLES

	MPMc				
Variable	Not Deprived	Deprived_1	Deprived_2	Deprived_3	Total
geo_aez					
High rainfall, 2 rainy seasons, STG	37.4	24.0	6.2	1.0	10.5
Transition, medium rainfall, cotton	14.4	30.9	34.6	22.9	29.4
Medium precipitation, 1 rainy season, cotton	8.3	20.0	29.6	27.6	25.7
Sudan-Sahelian, low rainfall, cotton	2.9	8.4	26.7	48.1	27.3
Cotonou	37.0	16.7	2.9	0.4	7.2
geo_urbrur (Rural)	10.6	29.0	60.2	69.0	52.8
head_sex (Male)	81.4	82.7	82.6	87.9	84.1
head_educ					
No education	0.00	22.4	65.4	82.4	56.8
Primary education	19.2	28.7	19.5	12.7	19.7
Secondary education	43.0	37.2	14.2	4.6	18.3
Higher education	37.8	11.6	0.9	0.4	5.1
ind_edu_mother					
No education	31.1	62.9	88.5	95.8	81.6
Primary education	28.2	20.8	7.8	3.1	10.6
Secondary education	28.6	13.7	3.2	0.8	6.3
Higher education	10.3	1.5	0.0	0.0	0.9
head_emp_sector12m					
Primary sector	2.8	12.4	54.2	79.4	48.9
Secondary sector	13.1	19.0	13.7	5.9	12.7
Service sector	72.8	60.2	28.2	12.6	33.6
Inactive	10.3	8.0	3.9	2.1	4.7
Unemployed	1.1	0.4	0.1	0.0	0.2
head_mstat					
Polyg. Married	5.9	15.7	24.5	39.1	25.6
Monog. Married	72.0	63.2	59.8	50.2	58.5
Common-law union	4.3	2.9	1.0	0.5	1.4
Divorced, Separated, Widowed	13.4	13.2	13.3	9.2	12.1
Single	4.5	5.1	1.4	1.1	2.3
ind_ethnic					
Other	55.6	62.6	65.3	60.0	62.6
Yoruba	5.9	4.2	0.8	0.5	1.8
Adja	3.0	7.3	8.9	8.9	8.2
Peulh	0.1	0.3	2.9	11.9	4.7
Fon	33.1	22.2	13.6	8.6	15.2
Bariba	2.3	3.5	8.5	10.1	7.5
ind_migration					
Not migrated	61.1	69.7	81.9	87.6	79.6
Rural area	7.6	8.5	7.5	6.8	7.6
Urban area	27.3	17.9	6.7	2.9	9.3
Foreign country	3.9	4.0	3.8	2.8	3.6
hh_trf_receive (Yes)	36.3	34.0	25.2	14.6	24.8
hh_fin_access (Yes)	93.9	80.3	45.6	20.4	49.1
hh_shk_severe_1					
None	39.3	24.8	15.9	12.2	18.1
Shk_Household	30.2	33.7	45.1	43.7	41.3
Shk_Nature	3.7	9.1	14.2	17.4	13.4
Shk_Agriculture	0.7	3.5	10.8	18.0	10.6
Shk_Income	6.0	4.6	1.8	1.5	2.6
Shk_Economic	13.6	14.8	6.8	3.0	7.9
Shk_Business	2.7	4.4	2.3	1.3	2.5
Shk_Insecurity	3.7	4.8	2.9	2.8	3.4
Shk_Other	0.2	0.3	0.3	0.2	0.3
c_roadac					
Tarred road	32.5	31.8	25.1	21.8	26.1
Laterite road	23.7	38.0	50.7	49.8	46.1
Track	13.6	13.9	18.0	26.0	19.1
Sea, river, lake route	0.0	0.6	2.9	1.5	1.8
Other	30.2	15.6	3.4	0.9	6.9
c_electric (Yes)	86.9	72.9	42.6	29.6	48.3
c_water (Yes)	80.1	63.0	36.3	20.1	40.2
c_healthcom (Yes)	39.3	40.7	37.1	33.7	37.1
c_schoolcom (Yes)	63.6	69.6	68.9	67.3	68.3

Note. STG = Subequatorial-tropical-guinean; SS = Sudan-Sahelian; Shk = shock.
Source: EHCVM 2018/19, World Bank, and author calculations.

TABLE X

INDEX FUNCTION/FORMULA SPECIFICATION

Variable	Base Model	Model A1	Model A2	Model A3	Model A4	Model A5	Model A6	Model A7	Model A8	Model RF
head_age	•		•	•		•	•	•	•	•
head_age^2							•	•	•	
hh_size	•	•	•	•	•	•	•	•	•	•
hh_size^2	•	•	•	•	•	•	•	•	•	
hh_depratio_c	•		•	•	•	•	•	•	•	•
hh_depratio_c^2							•	•	•	
c_inequality	•		•	•		•	•	•	•	•
geo_aez	•	•	•	•	•	•	•	•	•	•
geo_urbrur	•		•	•		•	•	•	•	•
head_sex	•	•	•	•		•	•	•	•	•
head_educ	•				•	•	•	•	•	•
ind_edu_mother	•	•		•			•	•	•	•
head_emp_sector12m	•	•	•	•	•	•	•	•	•	•
head_mstat	•	•	•	•		•	•	•	•	•
ind_ethnic	•	•	•	•	•	•	•	•	•	•
ind_migration	•	•	•	•	•	•	•	•	•	•
hh_trf_receive	•	•	•	•	•	•	•	•	•	•
hh_fin_access	•	•	•	•	•	•	•	•	•	•
hh_shk_severe_1	•	•	•	•	•	•	•	•	•	•
c_roadac	•	•	•	•		•	•	•	•	•
c_electric						•		•		•
c_water								•		•
c_healthcom			•	•		•		•		•
c_schoolcom		•	•	•		•	•	•		•
c_electric*c_water*c_healthcom*c_schoolcom	•								•	
c_electric*c_water							•			
c_water*c_healthcom							•			
head_sex*head_educ							•			
head_sex*ind_edu_mother							•			
head_sex*head_mstat							•			
head_sex*hh_trf_receive							•			
head_sex*hh_fin_access							•			
head_educ*head_emp_sector12m							•			
RESET TEST J=1 (p-value)										
Ordered Probit	0.05	0.22	0.71	0.79	0.83	0.00	0.89	0.09	0.13	
Fractional Probit	0.01	0.65	0.14	0.08	0.28	0.00	0.82	0.06	0.08	
RESET TEST J=2 (p-value)										
Ordered Probit	0.05	0.06	0.02	0.06	0.55	0.01	0.29	0.05	0.08	
Fractional Probit	0.00	0.38	0.33	0.08	0.05	0.00	0.06	0.01	0.00	

TABLE XI

ESTIMATION – MODEL A6

	Ordered Probit		Fractional Probit	
Variables	Coef.	SE	Coef.	SE
head_age	-0.003	0.01	0.000	0.00
head_age^2	0.000	0.00	0.000	0.00
hh_size	0.179 ***	0.02	0.088 ***	0.01
hh_size^2	-0.004 ***	0.00	-0.002 ***	0.00
hh_depratio_c	0.003 ***	0.00	0.002 ***	0.00
hh_depratio_c^2	0.000 ***	0.00	0.000 ***	0.00
c_inequality	0.373	0.35	0.347 **	0.15
geo_aez (base = High rainfall...)				
Transition, cotton…	0.634 ***	0.06	0.273 ***	0.02
Medium rainfall,	0.712 ***	0.06	0.309 ***	0.03
Low rainfall, cotton…	0.850 ***	0.07	0.377 ***	0.03
Cotonou	-0.193 **	0.09	-0.087 **	0.04
geo_urbrur (base = Urban)	0.128 ***	0.04	0.040 ***	0.02
head_sex (base = Male)	0.275 **	0.13	0.117 **	0.05
head_educ (base = No education)				
Primary education	-0.248 ***	0.07	-0.132 ***	0.03
Secondary education	-0.571 ***	0.08	-0.224 ***	0.03
Higher education	-1.172 ***	0.42	-0.439 **	0.18
ind_edu_mother (base = No education)				
Primary education	-0.639 ***	0.04	-0.264 ***	0.02
Secondary education	-0.728 ***	0.05	-0.314 ***	0.02
Higher education	-1.197 ***	0.12	-0.608 ***	0.06
head_emp_sector12m (base = Primary sector)				
Secondary sector	-0.162 **	0.08	-0.086 ***	0.03
Service sector	-0.347 ***	0.06	-0.164 ***	0.02
Inactive	-0.201 **	0.10	-0.080 **	0.04
Unemployed	-0.391	0.50	-0.302 **	0.14
head_mstat (base = Polyg. Married)				
Monog. Married	0.066	0.05	0.013	0.02
Common-law union	0.112	0.13	0.001	0.05
Divorced, Widowed	0.090	0.10	0.037	0.04
Single	0.263 **	0.10	0.146 ***	0.04
ind_ethnic (base = Other)				
Yoruba	-0.244 **	0.10	-0.093 **	0.04
Adja	0.236 ***	0.06	0.106 ***	0.02
Peulh	0.397 ***	0.09	0.206 ***	0.04
Fon	0.065	0.04	0.018	0.02
Bariba	-0.021	0.06	-0.023	0.03
ind_migration (base = Not migrated)				
Rural area	-0.011	0.03	0.019	0.01
Urban area	-0.141 ***	0.04	-0.055 ***	0.02
Foreign country	-0.033	0.05	-0.021	0.02
hh_trf_receive (base = Yes)	0.084 *	0.04	0.041 **	0.02
hh_fin_access (base = Yes)	0.610 ***	0.04	0.253 ***	0.02
hh_shk_severe_1 (base = None)				
Shk_Household	0.106 **	0.04	0.027	0.02
Shk_Nature	0.220 ***	0.06	0.076 ***	0.02
Shk_Agriculture	0.238 ***	0.07	0.072 ***	0.03
Shk_Income	0.172	0.11	0.039	0.05
Shk_Economic	0.051	0.06	-0.044 *	0.03
Shk_Business	0.108	0.10	0.020	0.04
Shk_Insecurity	-0.111	0.09	-0.062 *	0.04
Shk_Other	0.155	0.25	0.039	0.10

	Ordered Probit		Fractional Probit	
Variables (cont.)	Coef.	SE	Coef.	SE
c_roadac (base = Tarred road)				
Laterite road	0.005	0.04	-0.010	0.02
Track	-0.025	0.05	-0.002	0.02
Sea, river, lake	-0.091	0.11	-0.038	0.04
Other	0.169 **	0.08	0.049	0.03
c_schoolcom (base = Yes)	0.037	0.04	0.040 ***	0.02
c_electric*c_water				
Yes & No	0.249 ***	0.06	0.111 ***	0.02
No & Yes	0.331 ***	0.07	0.107 ***	0.03
No & No	0.378 ***	0.05	0.154 ***	0.02
c_water*c_healthcom				
Yes & No	-0.107 **	0.05	-0.049 **	0.02
No & Yes	-0.018	0.05	-0.004	0.02
head_sex*head_educ				
Female & Prim. educ.	-0.826 ***	0.12	-0.330 ***	0.05
Female & Sec. educ.	-0.532 ***	0.14	-0.254 ***	0.06
Female & Higher educ.	-0.163	0.27	-0.186	0.13
head_sex*ind_edu_mother				
Female & Prim. educ.	0.484 ***	0.08	0.186 ***	0.03
Female & Sec. educ.	0.374 ***	0.10	0.160 ***	0.04
Female & Higher educ.	0.484 *	0.25	0.237 *	0.13
head_sex*head_mstat				
Female & Monog. Married	-0.151	0.13	-0.059	0.06
Female & Common-law	-0.554 *	0.29	-0.145	0.11
Female & Divorced…	-0.158	0.14	-0.067	0.06
Female & Single	-0.349 *	0.21	-0.166 *	0.09
head_sex*hh_trf_receive				
Female & No	-0.156 **	0.08	-0.067 **	0.03
head_sex*hh_fin_access				
Female & No	0.107	0.08	0.025	0.03
head_educ*head_emp_sector12m				
Prim. educ. & Sec. sector	-0.178	0.12	-0.045	0.05
Prim. educ. & Serv. sector	-0.108	0.10	0.000	0.04
Prim. educ. & Inactive	-0.173	0.21	-0.067	0.09
Prim. educ. & Unempl.	-0.025	0.81	-0.018	0.19
Sec. educ. & Sec. sector	-0.255 *	0.13	-0.094 *	0.05
Sec. educ. & Service	-0.220 **	0.10	-0.078 **	0.04
Sec. educ. & Inactive	-0.424 ***	0.16	-0.195 ***	0.07
Sec. educ. & Unempl.	-0.975	0.74	-0.309	0.29
Higher educ. & Sec. sector	0.043	0.48	-0.151	0.21
Higher educ. & Serv.	-0.183	0.43	-0.172	0.18
Higher educ. & Inactive	-0.275	0.49	-0.216	0.21
Higher educ. & Unempl.	0.330	0.82	0.203	0.34
cut1	-0.781 ***	0.22		
cut2	1.286 ***	0.23		
cut3	3.249 ***	0.23		
Constant			-1.095 ***	0.09
Number of obs	42343		42343	
Log pseudolikelihood	-33541		-17978	
Pseudo R-squared	0.347			
Wald Chi-square	4588		8478	
Prob > chi2	0.000		0.000	
RESET TEST J=1 (p-value)	0.891		0.817	
RESET TEST J=2 (p-value)	0.285		0.059	

Note. ***$p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.1$; SE = cluster-robust standard errors

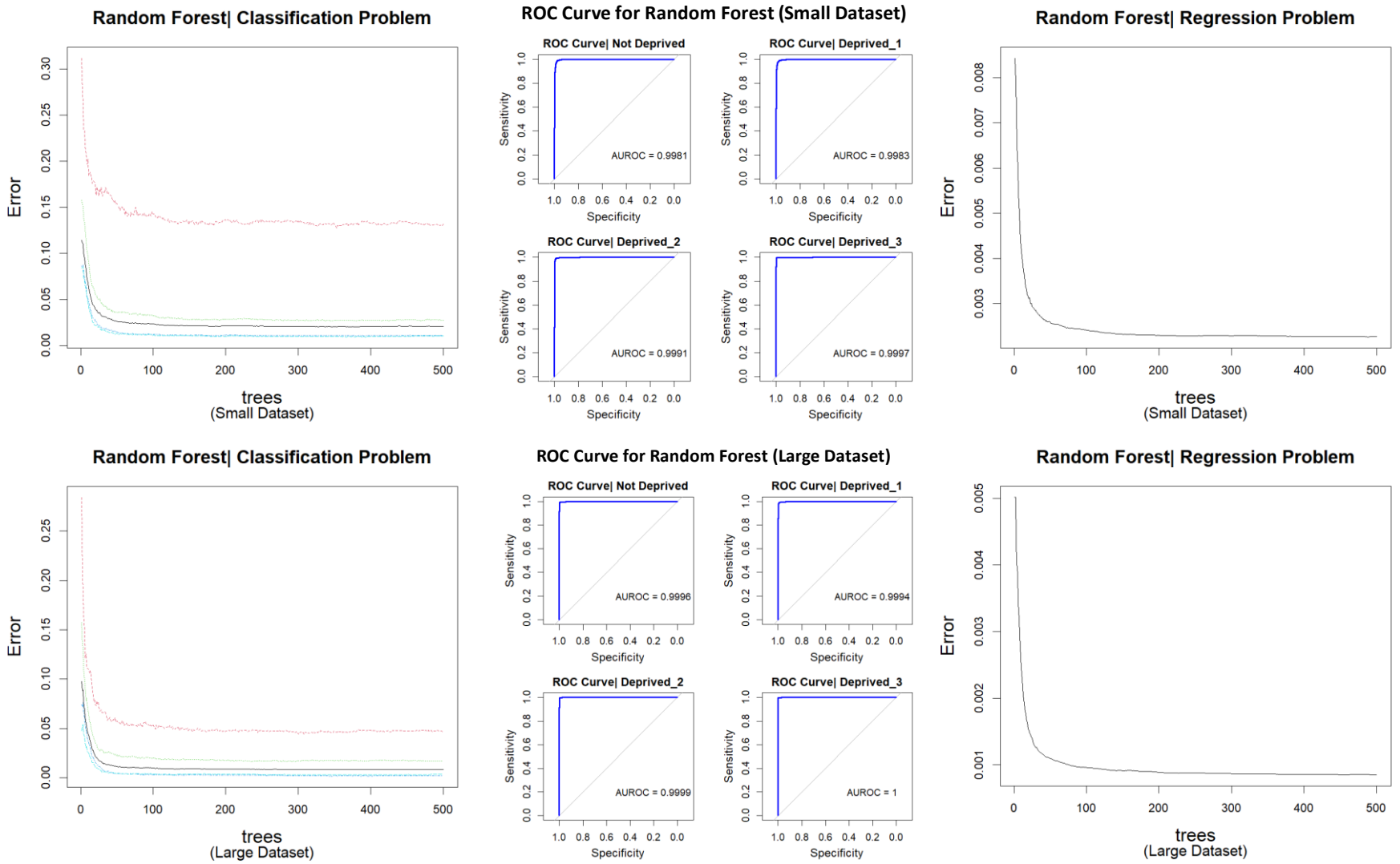


FIGURE 9 - Random forest out-of-sample performance.

TABLE XII

OUT-OF-SAMPLE PERFORMANCE | POVERTY-TARGETING APPROACH

	Random Forest			
	MPMc			
	Not Deprived	Deprived_1	Deprived_2	Deprived_3
Sensitivity	94.7	98.0	99.4	99.5
Specificity	100	99.3	99.0	100
Precision	100	97.6	98.7	100
F1-score	97.3	97.8	99.0	99.8
Accuracy	97.3	98.6	99.2	99.8
Error rate	2.7	1.4	0.8	0.2

	Overall			
	estimate	95% confidence interval		*p*-value
Accuracy	98.9	98.6	99.1	0.00
Error rate	1.1	1.4	0.9	