

Measuring Poverty and Inequality with Reduced Data

A Machine Learning Approach Using Nigerian Household Data

Vanesa Jordá and Miguel Niño Zarazúa

Working paper

No. 275

May 2026

The SOAS Department of Economics Working Paper Series is published electronically by SOAS University of London.

ISSN 1753 – 5816

This and other papers can be downloaded free of charge from:

SOAS Department of Economics Working Paper Series at
<http://www.soas.ac.uk/economics/research/workingpapers/>

Research Papers in Economics (RePEc) electronic library at
<https://ideas.repec.org/s/soa/wpaper.html>

Suggested citation

Jordá, Vanesa and Niño Zarazúa, Miguel (2026), “Measuring Poverty and Inequality with Reduced Data. A Machine Learning Approach Using Nigerian Household Data”, SOAS Department of Economics Working Paper No. 275, London: SOAS University of London.

Department of Economics
SOAS University of London
Thornhaugh Street, Russell Square, London WC1H 0XG, UK
Phone: + 44 (0)20 7898 4730
Fax: 020 7898 4759
E-mail: economics@soas.ac.uk
<http://www.soas.ac.uk/economics/>

© Copyright is held by the author(s) of each working paper.

Measuring Poverty and Inequality with Reduced Data

A Machine Learning Approach Using Nigerian Household Data

Vanesa Jordá¹

Miguel Niño Zarazúa²

Abstract

Reliable measurement of income and consumption is essential for monitoring poverty and inequality in low- and middle-income countries, yet full household surveys are costly and difficult to implement regularly. This paper examines whether reduced survey instruments can preserve key distributional information. We apply Random Forest Recursive Feature Elimination (RF-RFE) to the 2018/19 Nigeria General Household Survey-Panel to identify the income sources, consumption categories and household characteristics that best classify individuals within the welfare distribution. The analysis focuses on three outcomes: poverty status, location in the quintile distribution and position relative to the Gini-based inequality line. The survey's post-planting and post-harvest periods allow us to assess performance under different seasonal contexts. Results show that RF-RFE achieves strong classification accuracy with few predictors. For consumption, poverty status and inequality-line position are accurately predicted using a small set of expenditure categories, while quintile classification reaches about 80 percent accuracy for seasonal consumption and 60--65 percent for annual consumption predicted from a single seasonal visit. For income, poverty status reaches around 90 percent accuracy with five predictors, and inequality-line position is largely captured by labour earnings. The findings suggest that machine-learning methods can help improve survey design and reduce data requirements while retaining much of the distributional information needed to measure and monitor poverty and inequality.

Keywords: consumption measurement; poverty; inequality; Random Forests; household surveys; Nigeria

JEL classification: C38, C55, D31, I32, O55.

¹ Department of Economics, Cantabria University. Email: vanesa.jorda@unican.es

² Department of Economics, SOAS University of London. Russell Square, London WC1H 0XG, UK. Email: mn39@soas.ac.uk

Acknowledgements: The authors are grateful to participants of the 2026 WIDER Development Conference: “Green Industrialisation and Inclusive Growth in a Fractured World Order” in New Delhi, India for their comments and suggestions on early versions of this paper. VJ acknowledges financial Support from the I+D+i project Ref. PID2024-156871NB-I00 financed by MICIU/ AEI/10.13039/501100011033/ FEDER,UE. All errors are solely the responsibility of the authors.

1 Introduction

Accurate measurement of income and consumption is central to poverty and inequality analysis. Income captures the flow of resources available to households through labour, agriculture, enterprises, transfers and other sources, while consumption expenditure reflects how resources are used to meet needs and sustain living standards. Together, these welfare aggregates allow researchers and policymakers to identify deprivation, monitor distributional change, and assess the extent to which economic resources are unequally distributed across individuals, households and communities.

Income and consumption serve distinct analytical purposes. Income data are generally more directly connected to inequality analysis because they capture command over resources at a given point in time and are closely aligned with internationally agreed statistical definitions, such as those established by the Canberra Group (Canberra Group, 2011). Income is also closely linked to the distribution of earnings, capital income and transfers (Atkinson, 1970; Piketty, 2014), and is therefore central to the study of inequality (Cowell, 2011; Bourguignon, 2015). By contrast, consumption expenditure is often preferred for poverty analysis, particularly in low-income settings, because it is typically less volatile than income and may better approximate longer-term living standards (Deaton & Zaidi, 2002a; Ravallion, 1998a). Consumption may smooth temporary income fluctuations and can therefore provide a more reliable indicator of material deprivation, especially at the lower end of the welfare distribution.

In sub-Saharan Africa, household surveys have traditionally prioritised the collection of detailed consumption expenditure data rather than comprehensive income data. Large-scale survey programmes such as the Living Standards Measurement Study (LSMS) have adopted consumption aggregates as the primary welfare indicator (Beegle et al., 2012c; Deaton, 1997a). This reflects both conceptual and practical considerations. In economies characterised by informality, self-employment and subsistence agriculture, income is difficult to measure accurately because it is irregular, seasonal and often derived from multiple activities. Detailed expenditure modules are also costly and time-consuming, but they tend to provide more stable and reliable welfare estimates. As a result, poverty measurement across the region is overwhelmingly consumption-based, while income data, where available, are often used in complementary analyses of earnings, livelihoods and inequality.

However, collecting income and consumption data through nationally representative household surveys requires substantial financial and human resources. Such surveys require trained enumerators, complex questionnaire modules, repeated visits in settings where seasonality matters, data collection technology, field supervision and statistical expertise.¹ These costs limit the frequency with which full welfare surveys can be implemented, particularly in low- and middle-income countries with constrained statistical capacity. They also create a practical need for shorter instruments that can retain key distributional information while reducing respondent burden and fieldwork costs.

¹The cost of implementing LSMS-type surveys in sub-Saharan Africa is significant. While costs vary by sample size, geography and survey design, the World Bank's Independent Evaluation Group estimates the average cost of a standard national household survey at approximately US\$1.7 million (Independent Evaluation Group, 2023). On a unit basis, costs in African survey contexts can range from US\$100 to US\$500 per household, reflecting the high marginal costs of reaching remote rural populations and the increasing use of Computer-Assisted Personal Interviewing technologies (Kilic et al., 2017). The global requirement for monitoring the Sustainable Development Goals through household surveys was projected at nearly US\$945 million between 2016 and 2030 (Kilic et al., 2017; Grosh & Glewwe, 1998).

A growing body of literature has developed alternative strategies to approximate income or consumption using fewer survey questions. These include asset-based wealth indices constructed through principal component analysis (Filmer & Pritchett, 2001; Montgomery et al., 2000; McKenzie, 2005; Vyas & Kumaranayake, 2006), supervised proxy means tests and poverty scorecards calibrated to predict poverty status (Grosh & Baker, 1995; Schreiner, 2010), shortened consumption modules based on selected expenditure categories (Beegle et al., 2012a; Gibson & Kim, 2007), single or bracketed income questions designed to reduce reporting error (Hurd et al., 2004; Kennickell, 1998), and subjective welfare measures based on self-assessed economic status (Pradhan & Ravallion, 2000; Ravallion & Lokshin, 2002). These approaches reduce survey burden, but they are usually designed for specific purposes, such as poverty targeting, broad wealth ranking or rapid monitoring. They are not designed to preserve several distributional classifications simultaneously.

This paper contributes to the literature on welfare measurement by proposing a machine-learning approach to identify a parsimonious set of income sources and consumption categories that can classify individuals within the welfare distribution. Specifically, we implement Random Forest Recursive Feature Elimination (RF-RFE), a supervised feature-selection method that ranks candidate variables according to their predictive contribution and evaluates model performance across progressively smaller sets of predictors. Unlike standard proxy means-testing approaches, which are typically designed to identify households below a poverty threshold, our approach evaluates whether reduced survey instruments can preserve three complementary distributional outcomes: poverty status, quintile classification and position relative to the Gini-based inequality line.

The empirical analysis uses the fourth wave of the Nigeria General Household Survey-Panel, collected in 2018/19. Nigeria provides a relevant setting for this exercise because household welfare is shaped by agricultural seasonality, income volatility, regional heterogeneity and the coexistence of formal, informal and subsistence economic activities. The survey's post-planting and post-harvest visits allow us to examine whether reduced instruments perform differently across seasonal information sets. We construct benchmark income and consumption aggregates from the full survey modules, derive the corresponding poverty, quintile and inequality-line classifications, and then assess how accurately RF-RFE models reproduce these classifications using a much smaller set of predictors.

The results show that reduced models can preserve substantial distributional information, although performance varies across welfare concepts and information sets. For consumption, poverty status and position relative to the Gini-based inequality line can be predicted accurately with a small number of expenditure categories. Quintile classification is more demanding, especially when annual consumption is predicted from a single seasonal visit, but aggregated consumption categories still achieve high accuracy for seasonal welfare rankings. The most informative consumption predictors include food, meals consumed outside the home, roots and tubers, vegetables and rent, with some variation across poverty thresholds and seasons. For income, a small set of variables capturing crop sales, labour earnings, remittances, non-farm enterprise revenue, livestock sales, pensions and rental income accounts for much of the predictive information needed to classify individuals within the income distribution. Income poverty status reaches high accuracy with only a few predictors, while position relative to the Gini-based inequality line is captured particularly well by labour earnings.

Our analysis has two central implications. First, it shows that machine-learning-based feature selection can help identify a small set of welfare variables that retain much of the

information needed for poverty and inequality monitoring. Second, it highlights that the optimal reduced module depends on the welfare concept being measured. Binary poverty and inequality-line classifications are easier to reproduce than quintile classification, and seasonal information matters for consumption-based welfare measurement. Reduced instruments should therefore be designed with explicit attention to the distributional outcome of interest, rather than treated as generic substitutes for full income or consumption surveys.

The remainder of the paper is organised as follows. Section 2 sets out the welfare measurement framework used to construct the benchmark income and consumption aggregates. Section 3 reviews existing approaches to reduced income and consumption measurement. Section 4 presents the RF-RFE framework and defines the target classification outcomes. Section 5 describes the Nigeria GHS-Panel data and the construction of the analytical sample. Section 6 presents the empirical findings for consumption and income classifications. Section 7 reports robustness checks using alternative equivalence scales and machine-learning algorithms. Section 8 concludes with implications for survey design and welfare monitoring.

2 Welfare aggregates and the reduced-data measurement problem

This section sets out the welfare measurement framework used to construct the benchmark income and consumption aggregates against which the reduced-data models are evaluated. We focus on income and consumption expenditure, which capture related but distinct dimensions of household economic well-being. Income measures the flow of resources available to the household for current consumption and saving. Following the Canberra Group framework, household income includes monetary and in-kind receipts received at annual or more frequent intervals that are available for current consumption and do not reduce household net worth (Canberra Group, 2011). This excludes asset sales, borrowing, windfall gains and capital transfers. Consumption expenditure, by contrast, measures the value of goods and services acquired or consumed by the household during the reference period (Deaton & Zaidi, 2002b). In low- and middle-income settings, consumption is often preferred for poverty analysis because it is less volatile than income and may better approximate longer-term living standards, while income remains central for analysing earnings, transfers and inequality (Deaton, 1997b; Ravallion, 1998b).

In the empirical analysis discussed below, household income is defined as:

$$Y_h = Y_h^L + Y_h^A + Y_h^{LV} + Y_h^{NF} + Y_h^T + Y_h^O, \quad (1)$$

where Y_h^L denotes labour income, Y_h^A crop income, Y_h^{LV} livestock income, Y_h^{NF} non-farm enterprise income, Y_h^T remittances and private or public transfers, and Y_h^O other income sources, including pensions, rental income, interest and dividends. Agricultural, livestock and non-farm enterprise income are measured net of production costs where the survey provides the relevant information, in line with standard LSMS practice for welfare measurement in low- and middle-income countries (Grosh & Glewwe, 2000; Carletto et al., 2015).

Second, household consumption expenditure is defined as:

$$C_h = C_h^F + C_h^{NF} + C_h^H, \quad (2)$$

where C_h^F denotes food consumption, C_h^{NF} non-food consumption expenditure, and C_h^H housing-related consumption, including actual rent for tenants and imputed rent for owner-occupiers and non-market tenants. Food consumption includes purchases, own production, gifts and food consumed outside the home. Non-food consumption excludes asset purchases, debt repayments, production inputs, transfers to other households and highly irregular ceremonial expenditures where their inclusion would introduce substantial noise into short-reference-period welfare measurement (Deaton & Zaidi, 2002b). Social transfers in kind and the service flow from non-housing durable goods are not included in the benchmark aggregate because they cannot be consistently valued with the available data.

Both income and consumption components are annualised and aggregated at the household level. Because the distributional analysis is conducted at the individual level, household resources are assigned to household members. The baseline specification uses per capita income and consumption:

$$W_{ih} = \frac{W_h}{n_h}, \quad (3)$$

where W_h denotes either household income or household consumption and n_h is household size. This approach is transparent and widely used in poverty analysis, but it imposes the strong assumption that household needs rise proportionally with household size and that there are no economies of scale in consumption (Deaton, 1997b; Lanjouw & Ravallion, 1995). For this reason, the robustness analysis also considers alternative equivalence scales. First, we use the square-root scale, which divides household resources by the square root of household size and therefore allows needs to increase less than proportionally with household size. Second, we use the modified OECD equivalence scale, which assigns a weight of 1 to the first adult, 0.5 to each additional adult and 0.3 to each child, thereby allowing for both economies of scale and differences in needs between adults and children (Hagenaars & et al., 1994). These alternative scales reflect the long-standing recognition that welfare rankings and inequality estimates can be sensitive to assumptions about household sharing, demographic composition and economies of scale (Deaton & Muellbauer, 1980; Lanjouw & Ravallion, 1995).

3 Existing approaches to reduced income and consumption measurement

The measurement of income and consumption expenditure has conventionally relied on detailed survey modules covering multiple income sources, expenditure categories and recall periods. These instruments—often containing hundreds of items—remain the conventional benchmark for poverty and inequality analysis (Deaton, 1997b; Deaton & Zaidi, 2002b). However, their implementation is costly, time-consuming, and cognitively demanding for respondents. As a result, a growing literature has emerged proposing alternative approaches that approximate income or consumption expenditure using a reduced number of survey questions. This section discusses this strand of the literature.

3.1 Asset-Based Wealth Indices

One common approach replaces direct measurement of income or consumption with asset ownership and housing characteristics. The seminal work by [Filmer & Pritchett \(2001\)](#) applied principal component analysis (PCA) to asset variables from Demographic and Health Surveys (DHS) to generate a wealth index closely correlated with consumption-based welfare rankings. The first principal component, interpreted as a latent index of long-run economic status, has since been widely used in contexts where income or expenditure data are unavailable ([Montgomery et al., 2000](#); [McKenzie, 2005](#); [Vyas & Kumaranayake, 2006](#)).

Subsequent applications have extended this approach to classify households into national wealth quintiles. Tools such as the *EquityTool* operationalize this approach by selecting a subset of asset variables that reproduces national quintile classifications with relatively high accuracy ([Chakraborty et al., 2016](#)).² These instruments typically reduce 30–40 asset questions to approximately 10–15, thereby lowering survey burden while preserving relative ranking performance.

Asset-based indices offer some advantages. They are relatively stable over time, less affected by short-term income volatility, and easier for respondents to report accurately than detailed expenditure flows. However, empirical evidence shows that asset indices may perform well in ranking households at lowest extremes but exhibit weaker precision in the middle and upper tails of the distribution. Moreover, they also offer limited sensitivity to short-term welfare shocks ([Howe et al., 2008](#); [Harttgen & Vollmer, 2013](#)), while cross-country and inter-temporal comparability can be affected by structural changes in asset ownership patterns, potentially leading to divergence between asset-based inequality measures and those derived from income or consumption aggregates ([Smits & Steendijk, 2011](#); [Sahn & Stifel, 2003](#)). These considerations underscore that asset indices, while operationally attractive, may not fully approximate the distributional properties of consumption or income data used in standard poverty and inequality analysis.

3.2 Proxy Means Tests and Poverty Scorecards

A related but distinct literature develops proxy means tests (PMTs), which predict consumption or income using a small set of observable characteristics selected through supervised statistical models. Early work by [Grosh & Baker \(1995\)](#) formalized regression-based targeting mechanisms that estimate household consumption as a function of demographic, educational, and housing variables. These models are typically calibrated on full consumption surveys and subsequently applied using a reduced questionnaire.

Poverty scorecards represent a simplified implementation of this logic. They use a small number (often 5–15) of observable indicators to estimate the probability that a household falls below a poverty line. Variables are selected to maximize predictive accuracy while minimizing interview time. Compared to PCA-based asset indices, PMTs are explicitly supervised and calibrated to a specific poverty threshold, making them more suitable for estimating poverty headcounts ([Schreiner, 2010](#)). Nevertheless, most PMT applications focus on binary poverty classification rather than approximating the consumption distribution.

²Chakraborty et al. (2016) report Cohen’s Kappa statistics generally falling between 0.60 and 0.75, indicating substantial agreement between the *EquityTool*-based classifications and the full DHS wealth index, according to conventional benchmarks.

3.3 Shortened Consumption Modules

Rather than replacing consumption data with proxy indicators, another strand of the literature seeks to shorten expenditure modules themselves. Experimental evidence suggests that a relatively small subset of high-budget-share items captures a large share of total consumption variance. For example, [Beegle et al. \(2012b\)](#) document how alternative questionnaire designs in Tanzania affect measurement accuracy and respondent burden. Similarly, [Gibson & Kim \(2007\)](#) examine recall periods and item aggregation, highlighting trade-offs between precision and survey length.

These studies suggest that carefully selected consumption categories can reproduce aggregate poverty estimates with limited bias, although inequality measures are typically more sensitive to truncation and aggregation. The emphasis in this literature is on survey design features—recall periods, item aggregation, diary versus recall—rather than systematic selection of items. Nonetheless, the underlying objective is closely aligned with reducing the number of questions required to approximate welfare aggregates.

3.4 Single-Question and Bracketed Income Measures

In contexts where rapid data collection is essential—such as high-frequency phone surveys or labour force surveys—researchers often rely on a single self-reported income question. While appealing in its simplicity, this approach is known to suffer from substantial measurement error. Reviews by [Moore et al. \(2000\)](#) and [Bound et al. \(2001\)](#) document systematic under-reporting, non-response, and rounding (heaping) in income data. Moreover, [Deaton \(1997b\)](#) argues that income is generally more volatile and less reliably measured than consumption in developing-country settings.

To mitigate reporting error, some surveys employ bracketed income questions, asking respondents to select the interval within which their income falls.³ Evidence from unfolding bracket designs shows that this approach substantially reduces item non-response and improves response rates for sensitive financial questions ([Hurd et al., 2004](#); [Kennickell, 1998](#)). However, while bracketing lowers respondent burden, it limits precision for distributional analysis, particularly when estimating inequality indices or poverty headcounts close to the threshold, as grouped income data can bias Gini coefficients downward and distort poverty measures when the poverty line falls within an interval ([Cowell & Mehta, 1982](#); [Ravallion, 1994](#)). Thus, while single-question or interval-based approaches may suffice for coarse ranking or trend monitoring, the literature suggests that they are generally inadequate for reconstructing detailed distributional statistics without strong parametric assumptions ([Kennickell, 1998](#); [Cowell & Victoria-Feser, 1996](#)).

3.5 Subjective Welfare Measures

An alternative strategy elicits perceived economic status through subjective questions, such as self-assessed poverty or placement on an economic ladder. [Pradhan & Ravallion \(2000\)](#) and [Ravallion & Lokshin \(2002\)](#) report that subjective welfare measures correlate with consumption-based poverty status and can be used to derive *implicit* poverty lines.

³For examples of bracketed income questions, see the University of Michigan’s Panel Study of Income Dynamics (<https://psidonline.isr.umich.edu>); the Federal Reserve Board’s Survey of Consumer Finances (<https://www.federalreserve.gov/econres/scfindex.htm>); and the U.S. Census Bureau’s American Community Survey (<https://www.census.gov/programs-surveys/acs/methodology.html>).

Although these questions require minimal survey time, they capture multidimensional perceptions of welfare rather than monetary consumption *per se*. Cultural norms, reference groups, and reporting heterogeneity limit their comparability across contexts. As such, subjective measures are better interpreted as complements to, rather than substitutes for, consumption data in distributional analysis.

Taken together, the literature identifies several strategies for reducing survey length: constructing asset-based wealth indices through PCA; implementing supervised proxy means tests or poverty scorecards; shortening consumption modules by focusing on selected expenditure categories; relying on single or bracketed income questions; and eliciting subjective welfare assessments. Each approach involves trade-offs between survey cost, respondent burden and the precision with which the underlying welfare distribution can be recovered.

Most existing approaches are designed for specific purposes, typically poverty targeting, relative wealth ranking or rapid monitoring. They are less well suited to preserving several distributional properties simultaneously, including poverty status, quintile distribution and position relative to inequality-sensitive thresholds. This gap motivates the RF-RFE approach developed in the following section, which uses supervised feature selection to identify a parsimonious set of income and consumption variables that retains predictive information across multiple welfare classifications.

4 Methodology

Our primary objective is to identify a parsimonious subset of income and consumption components that can reproduce key distributional outcomes derived from the full welfare aggregates. The purpose is not only to reduce questionnaire length, but also to assess whether shorter survey instruments can preserve sufficient information for poverty and inequality analysis.

We organise the empirical exercise around three outcomes. First, we examine whether reduced survey instruments can predict poverty status under alternative poverty thresholds. Second, we evaluate whether they can identify individuals' position relative to the Gini-based inequality line. Third, we examine quintile classification as a broader test of whether reduced models preserve individuals' ranking across the welfare distribution. To implement these tasks, we use Random Forest Recursive Feature Elimination (RF-RFE), a supervised machine learning framework that combines ensemble prediction with iterative feature selection (Breiman, 2001; Guyon et al., 2002; Hastie et al., 2009).⁴

4.1 Random Forest Recursive Feature Elimination

RF-RFE is a wrapper-based feature selection method that uses Random Forests as the base learner (Kohavi & John, 1997; Guyon et al., 2002). The algorithm begins with the full set of candidate predictors, estimates a predictive model, ranks variables according to their contribution to model performance, removes the least informative predictor, and repeats this process until a sequence of increasingly parsimonious models has been evaluated.

⁴For expositional simplicity, all equations are written without explicit sampling-weight notation. In the empirical implementation, however, the benchmark welfare distributions, poverty thresholds, quintile classifications, inequality lines and model-performance statistics are computed using the survey weights described in Section 5.

Let Y_i denote the target welfare outcome for individual i .

Depending on the empirical task, Y_i represents poverty status, position relative to the inequality line, or position in the quintile distribution. Let

$$\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$$

denote the full set of p candidate predictors, including income sources, consumption categories and, where relevant, household characteristics. The objective is to identify a reduced subset of predictors $S \subseteq \{1, \dots, p\}$ such that

$$f(\mathbf{x}_{i,S}) \approx Y_i, \quad |S| \ll p,$$

where $\mathbf{x}_{i,S}$ denotes the vector of predictors for individual i restricted to the variables in S , and $f(\cdot)$ is the prediction function.

We split the data into an 80 percent training sample and a 20 percent held-out test sample. The split is implemented at the household level to avoid allocating members of the same household to both training and test samples. The cross-validation folds are also defined at the household level, ensuring that individuals from the same household are not assigned to both training and validation folds. All feature elimination, model selection and cross-validation procedures are conducted within the training sample. The test sample is reserved exclusively for final out-of-sample evaluation.

At the first step, the algorithm estimates a Random Forest using the full predictor set $S_0 = \{1, \dots, p\}$. Random Forests construct multiple decision trees on bootstrapped samples and aggregate their predictions, which improves predictive accuracy and reduces sensitivity to individual sample realisations (Breiman, 2001). For classification tasks, predictions are obtained by majority voting across trees.

At each iteration, the algorithm computes variable-importance scores. In the baseline classification models, importance is measured using the mean decrease in Gini impurity, a standard variable-importance measure in Random Forest classification models (Breiman, 2001). The predictor with the lowest importance score is removed, the model is re-estimated on the reduced feature set, and the procedure continues until all nested subsets have been evaluated. The result is a learning curve that relates predictive performance to the number of retained variables.

To reduce dependence on a single training split, we embed the elimination procedure within repeated fivefold cross-validation, with $K = 5$ folds and $R = 5$ repetitions (Hastie et al., 2009). For a given subset S , cross-validated risk is defined as:

$$\widehat{R}_{CV}(S) = \frac{1}{R} \sum_{r=1}^R \frac{1}{n_{\text{train}}} \sum_{\kappa=1}^K \sum_{i \in \mathcal{D}_{\kappa}^{(r)}} \mathcal{L}(Y_i, f^{(-\kappa,r)}(\mathbf{x}_{i,S})),$$

where $\mathcal{D}_{\kappa}^{(r)}$ denotes validation fold κ in repetition r , $f^{(-\kappa,r)}$ is the model trained on the remaining folds, n_{train} is the size of the training sample, and $\mathcal{L}(\cdot)$ is the relevant loss function. Cross-validation ensures that feature rankings reflect stable predictive relationships rather than noise in a single training realisation. For classification tasks, we use zero-one loss:

$$\mathcal{L}(Y_i, f(\mathbf{x}_{i,S})) = \mathbf{1}\{Y_i \neq f(\mathbf{x}_{i,S})\}.$$

Under this specification, empirical risk corresponds to the misclassification rate. For binary outcomes, such as poverty status or position relative to the inequality line, accuracy

is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP , TN , FP and FN denote true positives, true negatives, false positives and false negatives.

For quintile classification, the outcome has five categories. In this case, predictive performance is evaluated using multiclass accuracy, defined as the share of individuals correctly assigned to their observed quintile:

$$\text{Accuracy} = \frac{1}{n} \sum_{i=1}^n \mathbf{1} \{ \widehat{Q}_i = Q_i \},$$

where Q_i denotes the observed quintile of individual i and \widehat{Q}_i denotes the predicted quintile.

Rather than selecting a single model solely on the basis of maximum accuracy, we report performance over the full sequence of nested models. This allows us to identify the smallest set of variables that reaches a high and stable level of predictive performance. This is important for survey design, where the relevant trade-off is not only predictive accuracy, but also the number of questions required to obtain that accuracy.

4.2 Target variables

The RF-RFE framework evaluates whether reduced survey instruments can reproduce three distributional outcomes derived from the full income and consumption aggregates: poverty status, position relative to the Gini-based inequality line, and quintile classification. These outcomes capture complementary dimensions of welfare measurement: identification of deprivation, classification relative to an inequality-sensitive threshold, and ranking across the full welfare distribution.

4.2.1 Poverty status

We begin with poverty status because it is the most common policy application of reduced welfare instruments. A poverty line defines the threshold separating poor from non-poor individuals. For consumption, we evaluate poverty status using both national and international poverty lines. The national poverty line reflects domestic welfare standards and cost-of-living conditions ([National Bureau of Statistics, 2020](#)), while the international poverty line facilitates comparison with global poverty monitoring ([World Bank, 2022](#)). For income, we focus on a relative poverty line, defined as 60 percent of median income, $z^R = 0.6 \times \text{median}(y)$, following common practice in relative poverty measurement ([OECD, 2011](#); [Eurostat, 2021](#)). For a given welfare aggregate y_i , poverty status is defined as:

$$P_i = \mathbf{1} \{ y_i < z \},$$

where z denotes the relevant poverty threshold. The reduced-data model is evaluated by its ability to reproduce this binary classification using a smaller set of income or consumption predictors.

4.2.2 Inequality lines

We also examine whether reduced models can identify whether individuals fall below or above the inequality line. Inequality lines define the level of welfare below which an additional transfer reduces inequality, and above which an additional transfer increases inequality (Lopez-Noval et al., 2024). They provide a distribution-sensitive threshold that differs from poverty lines, which are defined in relation to deprivation.

In this paper, we focus on the inequality line associated with the Gini index. Let $F^{-1}(\cdot)$ denote the quantile function and let G denote the Gini coefficient. The Gini-based inequality line is given by:

$$c_y = F^{-1}\left(\frac{1+G}{2}\right).$$

We then define whether individual i lies above the inequality line as: $R_i = \mathbf{1}\{y_i \geq c_y\}$. The RF-RFE model is evaluated by its ability to reproduce this binary classification using a reduced set of predictors.

4.2.3 Income and consumption quintiles

Finally, we examine whether reduced instruments can reproduce individuals' broader position in the welfare distribution. We develop classification models with five categories corresponding to quintiles of the income or consumption distribution. Each quintile represents 20 percent of the population, ranked from the lowest to the highest level of welfare. Let y_i denote the welfare level of individual i , where y_i may refer to income or consumption. Let

$$F(y) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{y_i \leq y\}$$

denote the empirical distribution function. The k -th quintile threshold Q_k is defined as:

$$Q_k = \inf \left\{ y : F(y) \geq \frac{k}{5} \right\}, \quad k = 1, \dots, 4.$$

These thresholds partition the population into five mutually exclusive groups. The first quintile contains the poorest 20 percent of individuals, while the fifth quintile contains the richest 20 percent. The reduced-data model is evaluated by its ability to assign individuals to the same quintile as the full welfare aggregate.

4.3 From reduced predictors to distributional outcomes

The final step evaluates whether the reduced models preserve the distributional information required for poverty and inequality analysis. We compare predicted and observed classifications for poverty status, position relative to the Gini-based inequality line, and quintile classification. This allows us to assess whether reduced questionnaires can accurately classify individuals into policy-relevant welfare groups and whether a parsimonious set of income and consumption variables can retain enough information to support poverty and inequality monitoring.

Rather than selecting a single model solely on the basis of maximum accuracy, we report performance across the full sequence of nested models. This makes it possible to

identify the smallest set of variables that reaches a high and stable level of predictive performance, which is the relevant trade-off for survey design.

5 Data

The empirical analysis uses the fourth wave of the Nigeria General Household Survey-Panel (GHS-Panel), collected in 2018/19 by the National Bureau of Statistics (NBS) of Nigeria in collaboration with the World Bank Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) programme. The GHS-Panel is a nationally representative household panel survey designed to collect detailed information on agricultural production, household welfare, livelihoods and socio-economic conditions. The 2018/19 round is the fourth wave of the panel, following previous rounds conducted in 2010/11, 2012/13 and 2015/16 ([National Bureau of Statistics, 2019](#)).

The survey is representative at the zonal level, covering Nigeria's six geopolitical zones: three in the North and three in the South. It also distinguishes between rural and urban areas, allowing the analysis to account for important spatial differences in livelihoods, welfare levels and access to markets and services. The analytical sample used in this paper contains 5,049 households with the information required to construct both income and consumption aggregates. All distributional estimates and model-performance statistics are computed using the GHS-Panel sampling weights. These weights account for the survey's stratified multi-stage design, differential probabilities of selection, non-response, and the combination of the refresh and long-panel samples. Because the analysis is conducted at the individual level, household welfare aggregates are assigned to household members and weighted to recover the population-representative individual welfare distribution.

Wave 4 is particularly appropriate for this paper because it combines rich welfare information with detailed agricultural and seasonal data. Households were interviewed twice: first during the post-planting period, between July and September 2018, and then during the post-harvest period, between January and February 2019. This two-visit structure is important in the Nigerian context, where income, agricultural production and consumption can vary substantially across the agricultural calendar. The post-planting visit captures household conditions during a period in which many agricultural households face tighter liquidity constraints and rely on expected harvests, while the post-harvest visit captures realised production, crop sales and consumption decisions after harvest.

The GHS-Panel Wave 4 sample combines a refresh sample with a long-panel component. In Wave 4, 360 new enumeration areas were selected for the refresh sample, with 60 enumeration areas per geopolitical zone and 10 households selected in each enumeration area. This refresh sample was combined with a subsample of households from the original panel to preserve the longitudinal dimension of the survey. The combined Wave 4 sample consisted of 519 enumeration areas ([National Bureau of Statistics, 2019](#)).

The survey includes three main instruments: a household questionnaire, an agriculture questionnaire and a community questionnaire. The household questionnaire provides information on demographics, education, health, labour, food and non-food expenditure, household non-farm enterprises, food security, shocks, safety nets, housing conditions, assets, information and communication technology, remittances and other sources of income. The agriculture questionnaire collects detailed information on land ownership and use, plot characteristics, household and hired farm labour, input use, crop pro-

duction, harvest and disposition, livestock activities, agricultural capital, irrigation and fishing. The community questionnaire records information on local infrastructure, prices, labour conditions, land markets, community organisations, resource management and local events.

For income, the survey contains information on labour earnings, agricultural crop production, livestock activities, non-farm enterprises, remittances, transfers, pensions, rental income, savings and other income sources. For consumption, it includes detailed food expenditure, meals consumed outside the home, non-food expenditure, education, health and housing information. These modules allow us to construct full income and consumption aggregates that serve as benchmark welfare measures against which reduced-data models are evaluated.⁵

In this paper, the post-planting and post-harvest data are used both separately and jointly. This allows us to assess whether reduced survey instruments can recover welfare rankings, poverty status, position relative to the Gini-based inequality line and quintile classifications under different seasonal information sets. The working sample is restricted to households with the information required to construct the income and consumption aggregates and the corresponding distributional outcomes used in the RF-RFE analysis.

6 Results

In this section, we present the primary findings regarding the modelling of the income and consumption distribution using a parsimonious set of variables. Specifically, we identify the key predictors determining individual positions within the distribution, focusing on quintile classification and status relative to both the poverty line and the inequality line.

As detailed in Section 2, baseline results are derived by adjusting household income and consumption by the number of household members to estimate per capita welfare aggregates. The sensitivity of our findings to this specification is addressed in the subsequent robustness section, where we apply alternative equivalence scales to individualise household income. To identify the most relevant predictors, we implemented RF-RFE using a random forest base learner (Breiman, 2001), with model performance validated through repeated five-fold cross-validation.

6.1 Predicting consumption

As detailed in Table B1 in the Appendix, the survey instrument captures household consumption across more than 150 food items, nine categories for meals consumed outside the home, approximately 90 non-food categories, and various expenditures on education and health. This high-dimensional data, comprising over 240 individual items, presents significant computational challenges for the implementation of RF-RFE. To address this complexity, we employ a two-tiered aggregation strategy to streamline the feature space while maintaining economic relevance.

The first level of aggregation reduces the data to 36 categories. This includes 15 food categories and four categories for meals consumed outside the home, as specified in Table B1. Non-food items are consolidated into 13 categories following the structure in Table B2, alongside a dedicated category for paid or imputed rent. Finally, education

⁵For a more detailed discussion of the procedures used to construct the benchmark welfare aggregates, see Appendix 9.

expenditures are grouped into one category and two for health, distinguishing between outpatient costs (consultations, transport, and drugs) and hospital stays.

In instances where this granular level of aggregation fails to yield sufficient predictive accuracy, we implement a more parsimonious model. This second tier further aggregates sub-categories into broader headings, resulting in 18 consumption categories. In this configuration, all food items, all meals consumed outside the home, and all health-related expenditures are consolidated into their respective single variables to optimize model performance and interpretability.

This aggregation strategy is consistent with evidence from survey experiments showing that the design of consumption modules, including item aggregation and recall structure, can substantially affect measured consumption and poverty estimates (Beegle et al., 2012a). It also reflects the practical trade-off emphasised in the welfare-measurement literature: detailed expenditure modules improve coverage but increase respondent burden and may introduce reporting noise, while broader categories can preserve much of the relevant welfare information at lower cost (Deaton & Zaidi, 2002b).

We first focus on predicting the position of individuals within the quintiles of the consumption distribution. Figure 1 illustrates the accuracy of the models as a function of the number of variables, using both PP and PH data. For each dataset, two models are estimated: (1) represents the prediction of the consumption distribution specifically for the period corresponding to the data collection, while (2) is used for models that rely on these seasonal datasets to predict aggregate annual consumption.

The right panel illustrates the results obtained using the 36 consumption categories. These findings suggest that with this level of aggregation, the maximum accuracy for the seasonal models, denoted as (1), is approximately 73 percent. In contrast, when predicting annual consumption (Model 2), precision declines significantly to 54 percent, remaining relatively stagnant even as the number of features increases. When using a less granular set of consumption categories (left panel of Figure 1), predictive accuracy for seasonal data improves markedly, exceeding 80 percent with as few as five variables for both the post-harvest and post-planting periods. However, identifying an individual’s quintile within the annual consumption distribution remains considerably more challenging. Accuracy peaks at approximately 65 percent for post-planting data and 60 percent for post-harvest data.

The stronger performance of the aggregated specification suggests that predictive accuracy does not necessarily increase with more granular expenditure information. In this setting, broader consumption categories appear to reduce idiosyncratic item-level variation while retaining the budget components that most strongly differentiate households across the welfare distribution. This finding is consistent with the survey-design literature, which shows that shorter or more aggregated consumption modules can reproduce broad welfare patterns even though they may be less suitable for precise item-level expenditure analysis (Beegle et al., 2012a; Gibson & Kim, 2007).

We now analyse the primary predictors of consumption quintile classification. Figure 2 displays the five most influential categories for the four models considered across both levels of aggregation. In panel (a), *food*, *meals (outside the home)* and *rent* emerge as the overwhelmingly dominant predictors across all seasonal and annual specifications. The importance of the aggregated food category is nearly double that of the next most significant variable. In the disaggregated model (Panel b), the predictive power is more evenly distributed across a wider set of categories. *roots*, *rent*, and *vegetables* stand out as the primary pillars of the model. Interestingly, *grains* and *fish* appear as critical predictors

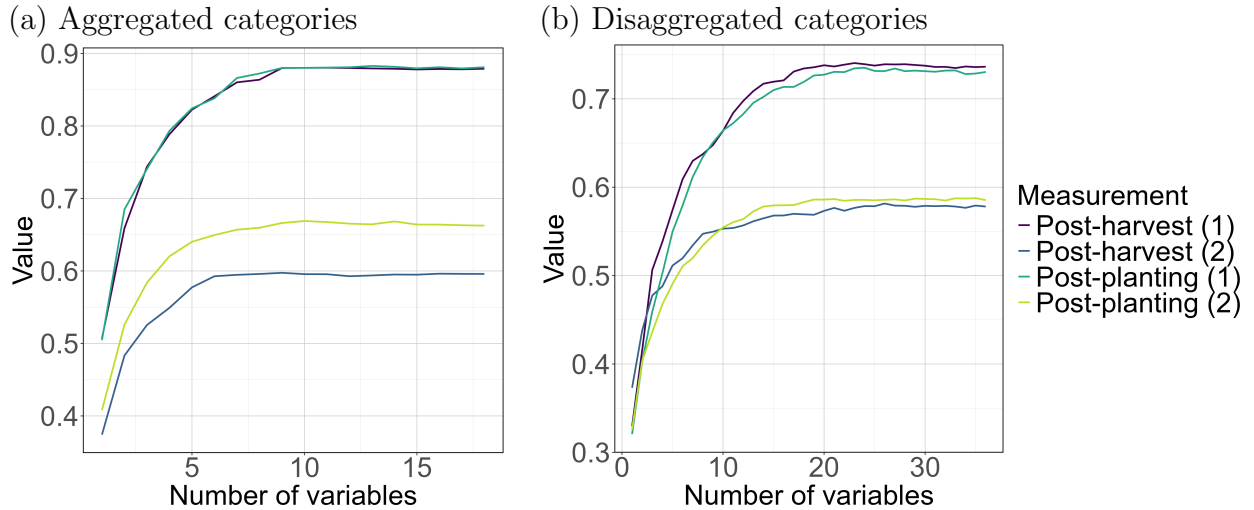


Figure 1: Model accuracy in predicting consumption quintiles by number of variables

Source: Authors' compilation.

specifically for the post-planting (lean season) models. A significant observation across both panels is the role of *rent*. Its importance is consistently high and exhibits less seasonal fluctuation than food-related items, suggesting that housing expenditure provides a relatively stable indicator of longer-term consumption levels.

The prominence of food expenditure is consistent with the central role of food in consumption-based welfare measurement in low- and middle-income settings, where food remains a large budget component and is closely linked to material deprivation (Deaton, 1997b). The importance of rent is also consistent with standard welfare-aggregate practice, which treats housing services as a key component of household consumption and imputes rents for owner-occupiers and non-market tenants when observed rental payments are unavailable (Deaton & Zaidi, 2002b).

We now turn our attention to the prediction of poverty status. Figure 3 illustrates the accuracy of models derived from post-planting and post-harvest data, comparing predictions for seasonal versus annual consumption. The left panel presents the results for the international poverty line. Post-planting and post-harvest datasets demonstrate comparable performance in identifying poverty status, consistently surpassing the 80 percent accuracy threshold for both annual and seasonal consumption models. Specifically, the annual consumption model reaches this benchmark using only three variables, whereas the seasonal consumption model requires five variables to achieve an accuracy exceeding 80 percent. These results suggest that identifying poverty status (a binary classification) is considerably more straightforward than assigning individuals to specific consumption quintiles.

This pattern is consistent with the broader targeting literature. Binary poverty classification is generally less demanding than recovering a household's location across the full welfare distribution, which is why proxy means tests and poverty scorecards are typically designed around poverty status rather than full distributional reconstruction (Grosh & Baker, 1995; Coady et al., 2004). The present results extend this logic by showing that a small set of consumption categories can perform well not only for poverty identification but also, with lower accuracy, for broader welfare ranking.

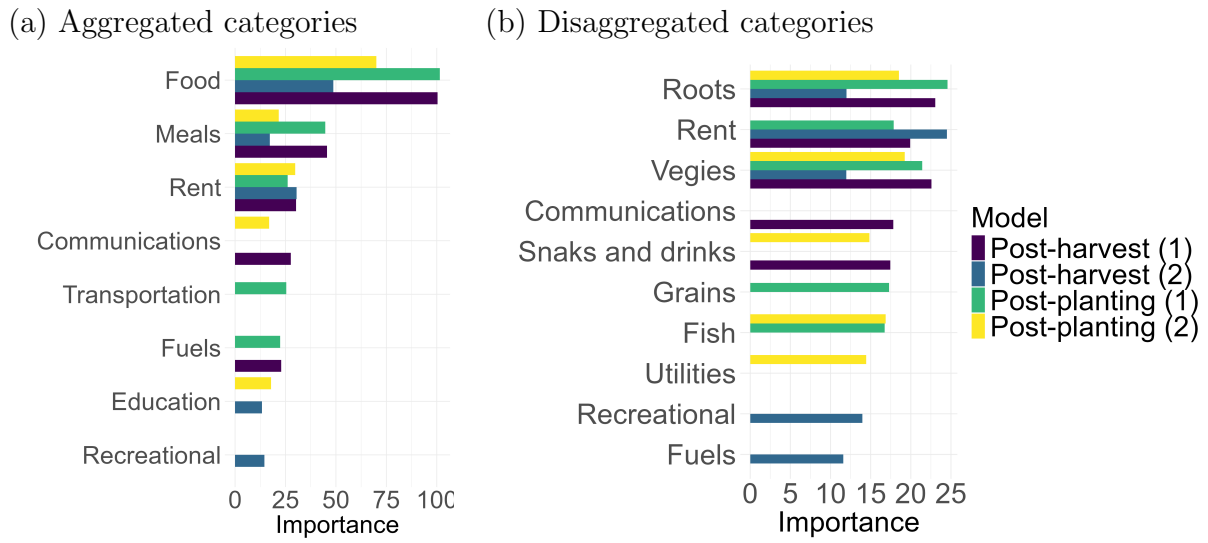


Figure 2: Five most relevant variables to predict consumption quintiles using post-planting and post-harvest data

Source: Authors' compilation.

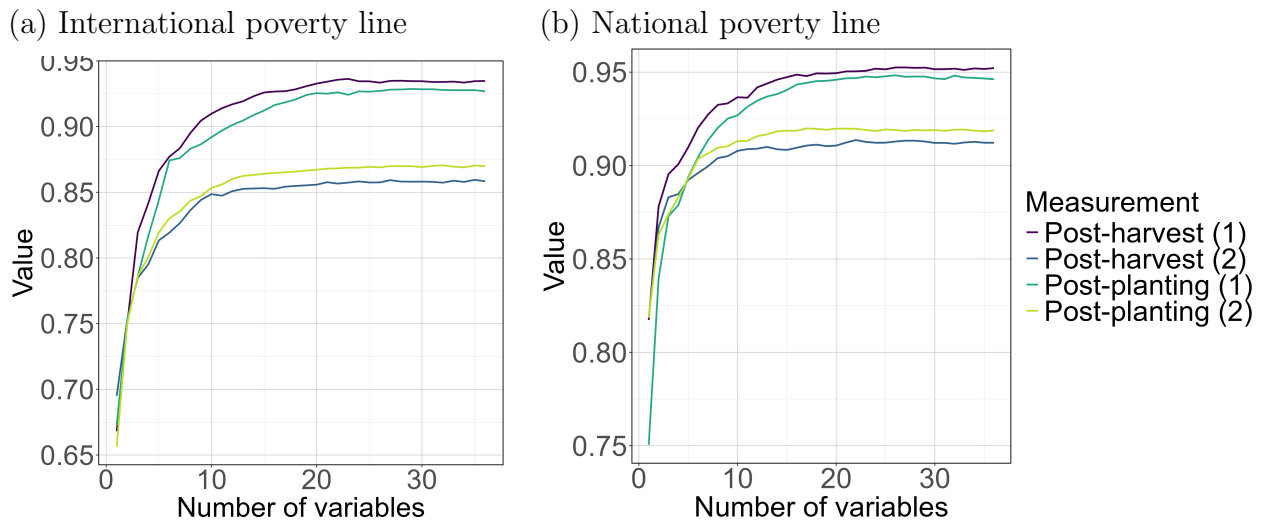


Figure 3: Model accuracy in predicting poverty status by number of variables

Source: Authors' compilation.

Regarding the national poverty line, the results in the right panel suggest that high predictive precision is attainable with minimal features. Specifically, the models achieve over 80 percent accuracy using only two variables across nearly all specifications. The sole exception is the model predicting seasonal poverty status based on post-planting data, which requires three variables to reach this same threshold of accuracy.

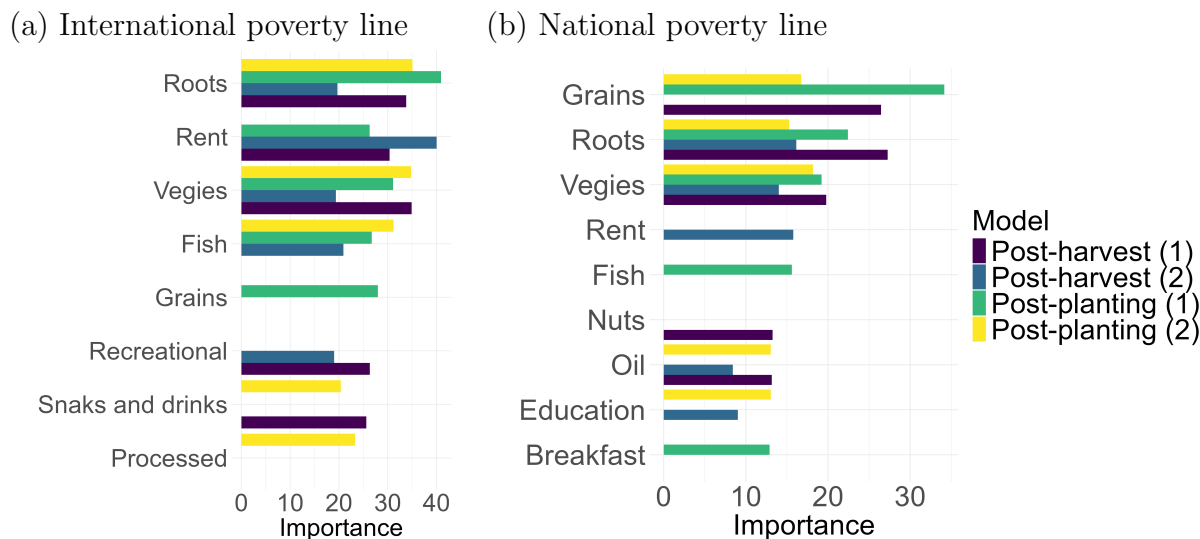


Figure 4: Five most relevant variables to predict poverty using post-planting and post-harvest data

Source: Authors' compilation.

The primary predictors of poverty status, illustrated in Figure 4, reveal that the consumption profiles associated with poverty are highly dependent on both seasonal economic conditions and the specific poverty threshold employed. Several clear patterns emerge from this comparison.

First, *food* staples consistently dominate the identification of poverty under both definitions. *Roots* and *vegetables* rank among the most influential predictors across nearly all specifications, while grains become particularly prominent under the national poverty line. The consistent importance of these categories underscores that staple food consumption remains the core dimension distinguishing poorer households from those with higher welfare levels.

Second, seasonality significantly affects which expenditures gain predictive relevance. Under the international poverty line, post-harvest models assign greater importance to discretionary and higher-value items such as recreation, snacks, drinks, and processed foods. This may be consistent with relaxed liquidity constraints following crop sales. In contrast, post-planting models, which correspond to the “lean season”, place more weight on essential consumption such as fish, reflecting tighter budgets and a shift toward subsistence-oriented spending.

These seasonal shifts are consistent with evidence that agricultural households in developing countries face strong seasonal variation in income, liquidity and consumption opportunities. Existing studies show that the extent to which consumption tracks seasonal income depends on households’ ability to smooth resources over the agricultural cycle and on their exposure to vulnerability and liquidity constraints (Paxson, 1993; Der-

con & Krishnan, 2000). In the Nigerian context, the difference between post-planting and post-harvest predictors therefore reinforces the importance of collecting or modelling seasonal information when using reduced survey instruments.

Finally, the national poverty line places a stronger emphasis on education-related expenditures. Compared with the international line, housing-related expenditure (rent) becomes less dominant, and the predictor set shifts toward a more detailed food basket and basic welfare-related spending.

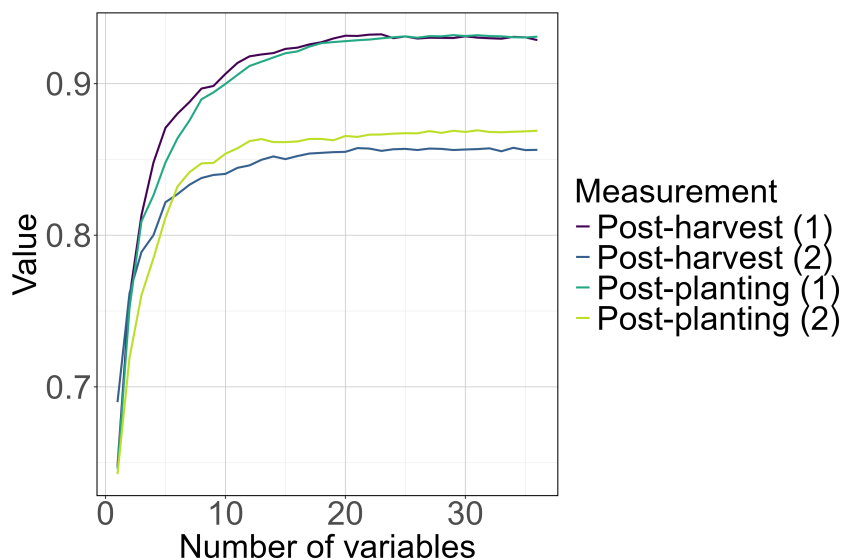


Figure 5: Model accuracy in predicting individuals' position relative to the inequality line

Source: Authors' compilation.

Turning our attention to inequality lines, we explore the predictive accuracy of the models in identifying whether individuals fall above or below this specific threshold. As illustrated in Figure 5, and consistent with our findings regarding the poverty lines, high levels of accuracy are achieved using a limited number of variables. Specifically, when predicting seasonal consumption positions, the models require only three variables to surpass the 80 percent accuracy threshold. In contrast, for annual consumption predictions, five variables are necessary to achieve a comparable level of precision.

The variable importance rankings for the inequality line (Figure 6) illustrate a shift toward housing, infrastructure, and consistent discretionary spending. *Roots* and *rent* stand out as the most dominant features, with rent showing a particularly high importance in the post-harvest annual model (Model 2). This indicates that housing costs are a primary structural differentiator for the upper end of the distribution.

The data also reveals the importance of lifestyle and mobility indicators. Transport and fuels appear as significant predictors, categories that typically require higher levels of disposable income. Furthermore, discretionary items like snacks and drinks, recreational expenses, and processed foods maintain high importance across both the post-planting and post-harvest periods. This suggests that, unlike the seasonal patterns observed for poverty, these expenditures remain a persistent feature of household budgets for those above this threshold.

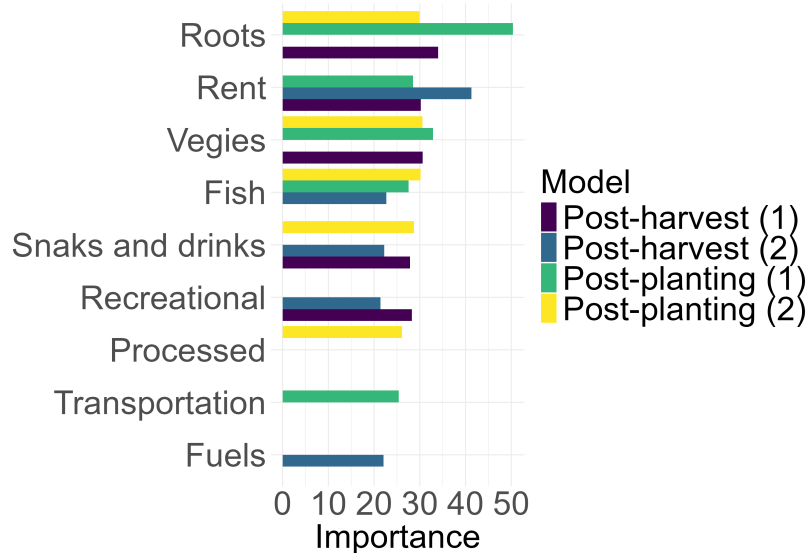


Figure 6: Five most relevant variables for predicting individuals' position relative to the inequality line based on post-planting and post-harvest data

Source: Authors' compilation.

This result is consistent with the interpretation of inequality lines as distribution-sensitive thresholds rather than deprivation thresholds. Unlike poverty lines, which identify minimum welfare shortfalls, inequality lines identify the point in the distribution at which additional income or consumption changes the level of inequality (Roope, 2021; Lopez-Noval et al., 2024). It is therefore unsurprising that variables associated with more stable housing conditions, mobility and discretionary expenditure become more informative than the basic consumption items that dominate poverty classification.

6.2 Predicting income

We now turn to predicting income groups using a parsimonious set of variables. Following the procedure outlined in Appendix 9 to estimate income aggregates, we identify 36 potential predictors of an individual's position along the income distribution. Half of these variables represent positive income streams, while the remaining half comprises agricultural and production costs that enter negatively into the household income calculation.

Unlike consumption, which is conventionally evaluated across both the post-harvest and post-planting periods to capture seasonal variation, earnings from the primary occupation represent the only income category available for both periods. Consequently, we include these two seasonal earning components as separate potential predictors.⁶

We first focus on predicting an individual's classification within the quintiles of the income distribution. The left panel of Figure 7 illustrates model accuracy as a function of the number of variables included. The results suggest that the maximum predictive accuracy for income quintiles converges around 80 percent. Notably, a model with just six variables surpasses the 75 percent accuracy threshold, whereas reaching the 80 percent

⁶We also estimated the model using a single aggregated income variable constructed as the average of post-planting and post-harvest earnings. The relationship between model accuracy and the number of predictors remains qualitatively similar to the results presented in Figures 7-9.

plateau requires a 12-variable specification.

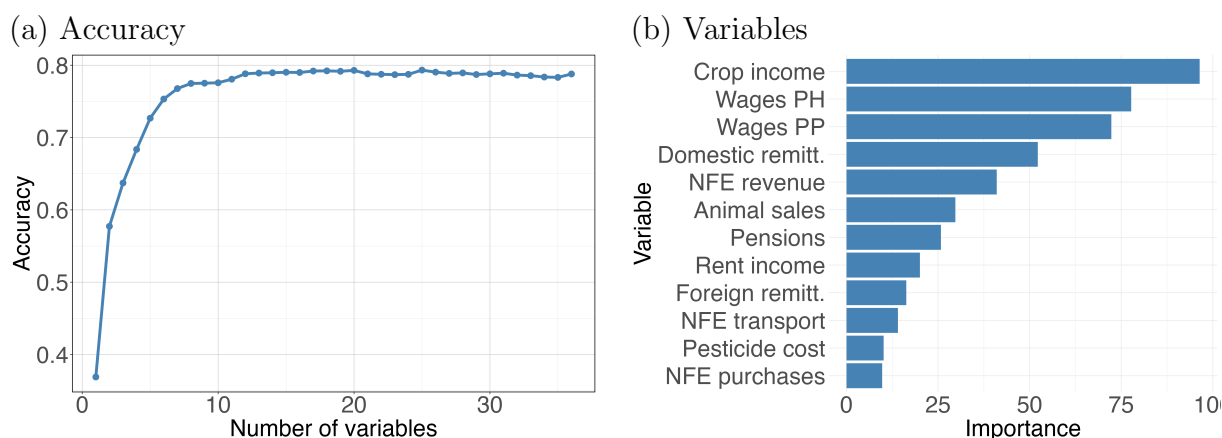


Figure 7: Model accuracy and most relevant variables in predicting income quintiles

Source: Authors' compilation.

The right panel of Figure 7 displays the most influential predictors of income quintiles. The nine most important predictors are positive income streams derived from crop sales, wages, remittances, non-farm enterprise (NFE) revenue, livestock sales, pensions, and rental income. The subsequent categories consist of various production costs, specifically three related to NFEs and two associated with pesticide and fertilizer acquisitions.

The prominence of crop sales, wage earnings, remittances and non-farm enterprise revenues is consistent with the literature on rural livelihoods in Sub-Saharan Africa, which emphasises that household welfare is shaped by diversified income portfolios rather than by agricultural production alone (Barrett et al., 2001). The results therefore suggest that a reduced income module should preserve the main revenue-generating activities across farm, labour-market, enterprise and transfer channels.

We now turn our attention to the prediction of poverty status. Because the international poverty line is conventionally applied to welfare aggregates expressed in comparable purchasing-power-parity terms, we restrict the income analysis to a relative poverty threshold defined as 60 percent of the median income aggregate. As we focus here exclusively on income, we present only the results for the relative income poverty line, defined as 60 percent of the median income aggregate. The left panel of Figure 8 illustrates the predictive accuracy of our models as a function of the number of included variables. The results demonstrate that a remarkably parsimonious model can capture poverty status with high precision. Relying on only three predictors, the model correctly classifies more than 80 percent of individuals' poverty status. When expanding the specification to just five variables, the predictive precision reaches 90 percent.

This high accuracy is consistent with the logic of supervised poverty-targeting tools, which exploit a small number of highly informative welfare correlates to classify households around a threshold (Coady et al., 2004). However, the present ML approach differs from standard proxy means testing because the same reduced income variables are also evaluated for their ability to preserve broader distributional rankings.

An analysis of the feature importance rankings (right panel of Figure 8) reveals striking structural consistency: the nine most influential variables for predicting poverty status coincide exactly in order of importance with those selected by the quintile models. This

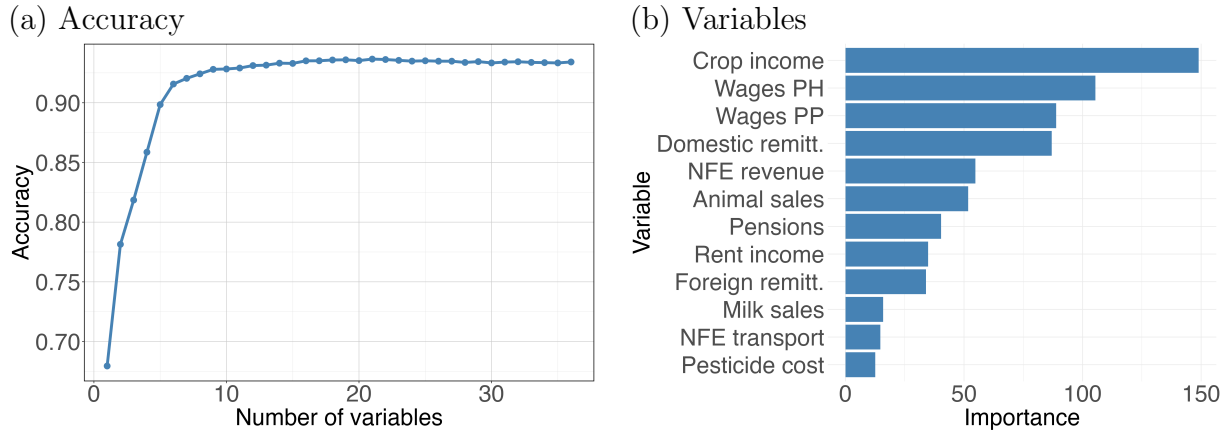


Figure 8: Model accuracy and most relevant variables in predicting individuals' income position relative to the national poverty line

Source: Authors' compilation.

overlap carries important policy and methodological implications. First, from a data-collection perspective, it suggests that the core drivers of structural economic positioning remain identical whether modelling broad distribution brackets (quintiles) or a binary welfare threshold (poverty status). Consequently, the same reduced questionnaire could achieve dual analytical goals: mapping overall inequality and targeting the poor.⁷ Second, because these top nine predictors are predominantly positive income streams (such as crop sales, wages, and non-farm enterprise revenues) rather than agricultural input or production costs, it implies that poverty status in this context is fundamentally driven by severe revenue deficits rather than variations in operational expenses. This finding suggests that social protection programmes and targeted interventions should prioritise income-generating capacity over input-subsidisation schemes alone.

We conclude this section by analysing the predictive accuracy of our models in identifying whether individuals fall above or below the relative inequality line (RIL) corresponding to the Gini index. As illustrated in Figure 9, and consistent with our previous findings regarding absolute poverty lines, high levels of accuracy are achieved using a very limited set of variables. Notably, post-harvest earnings alone successfully predict an individual's position relative to this threshold in 94 percent of cases. Incorporating post-planting earnings into the specification further increases the predictive precision to over 96 percent.

While several of the most influential predictors coincide with those observed in the quintile and national poverty line models, their relative rankings vary substantially. For this threshold, labour earnings eclipse crop sales in predictive power. Furthermore, pensions and rental income assume a more prominent position, displacing remittances and livestock sales relative to their rankings in previous models.

⁷Crucially, this structural alignment may reflect the specific socioeconomic profile of Nigeria, and further cross-country validation is required before general scaling conclusions can be drawn.

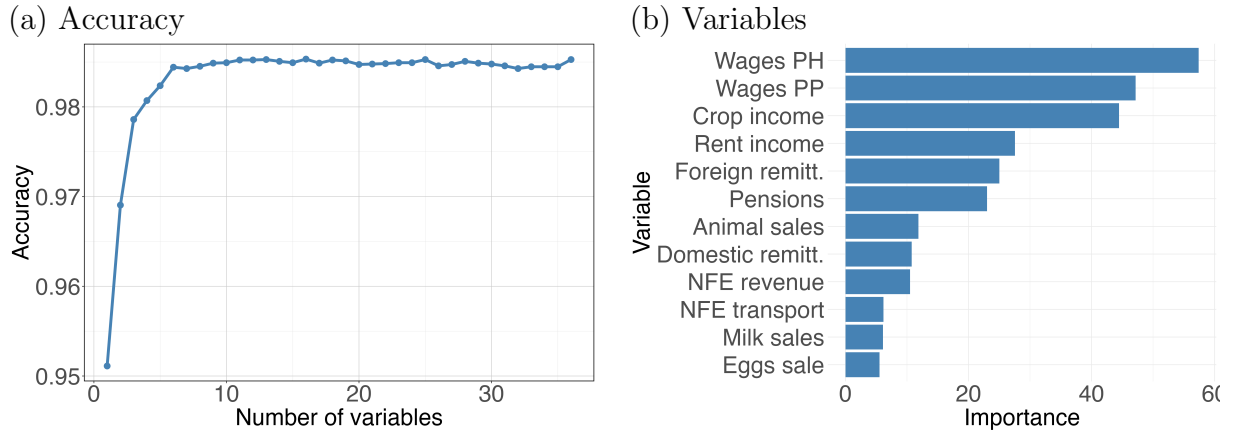


Figure 9: Model accuracy and most relevant variables in predicting individuals' income position relative to the relative inequality line

Source: Authors' compilation.

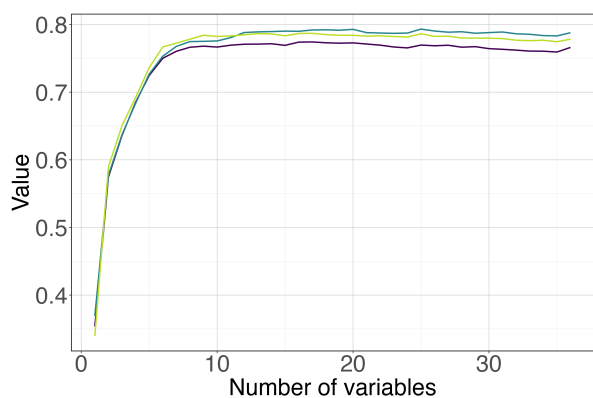
7 Robustness Checks

To ensure that the predictive performance of our parsimonious models is not an artefact of specific demographic assumptions, it is crucial to evaluate their robustness across alternative definitions of individual welfare. The baseline assumption of per capita income implies a completely linear relationship between household size and resource requirements. However, in settings with significant shared resources (such as housing, fuel, or durable goods) or variations in demographic composition (such as the ratio of children to adults), a per capita metric may underestimate individual welfare in larger households. Conversely, non-linear equivalence scales acknowledge economies of scale within the household unit. Because our machine learning models rely heavily on variables deeply intertwined with household size (such as total crop sales, labour earnings, and specific production costs) shifting the target variable's distribution via different scales could theoretically alter total predictive accuracy.

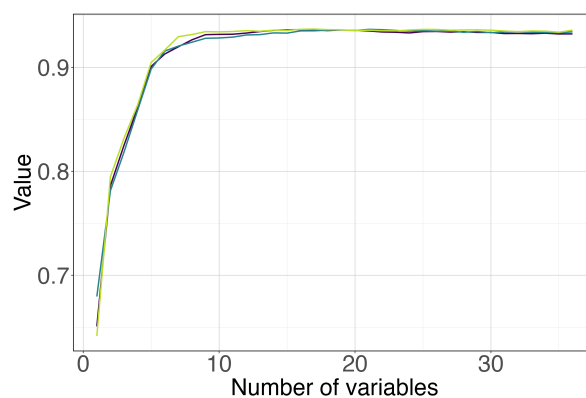
Figure 10 illustrates the predictive accuracy of the models as a function of the number of included variables for the three adjustment methods: per capita, square root, and the OECD equivalence scale. For the three alternative classifications by income groups (income quintiles in Panel a, the national poverty line in Panel b, and the relative inequality line in Panel c) the models exhibit remarkable structural stability. The learning curves closely mirror one another, demonstrating that the speed of model convergence and the ultimate accuracy peak are virtually invariant to the chosen equivalence scale.

The maximum accuracy for predicting income quintiles consistently stabilises at approximately 80 percent across all scales, with a model of just six variables comfortably passing the 75 percent accuracy threshold in every scenario. This consistency is even more pronounced when analysing the national poverty line. The models achieve a matching 90 percent accuracy with only five variables.

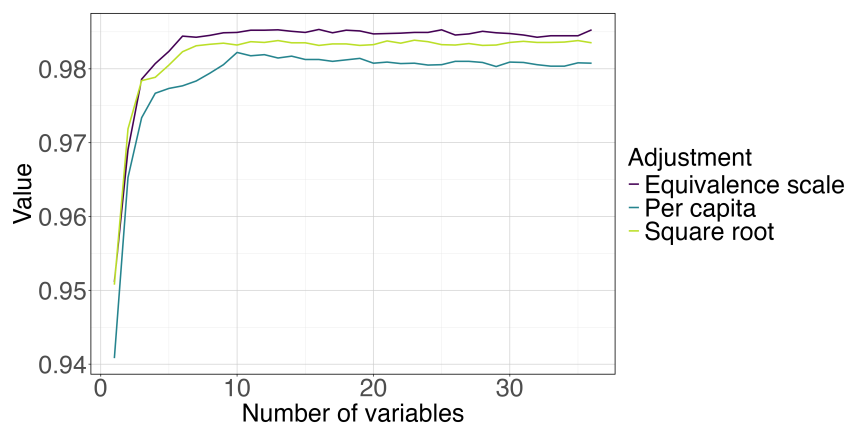
Relative inequality line models displayed in Panel (c) show a marginal divergence, though the overall precision remains exceptionally high. Here, the equivalence-scale specifications yield slightly superior predictive performance compared to the baseline per capita metric, maximising accuracy at roughly 98.5 percent. The per capita line trails



(a) Quintiles



(b) National poverty line



(c) Relative inequality line

Figure 10: Model accuracy in predicting individuals' income position for different equivalence scales

Source: Authors' compilation.

marginally behind, stabilizing just above 98 percent after the inclusion of ten variables.

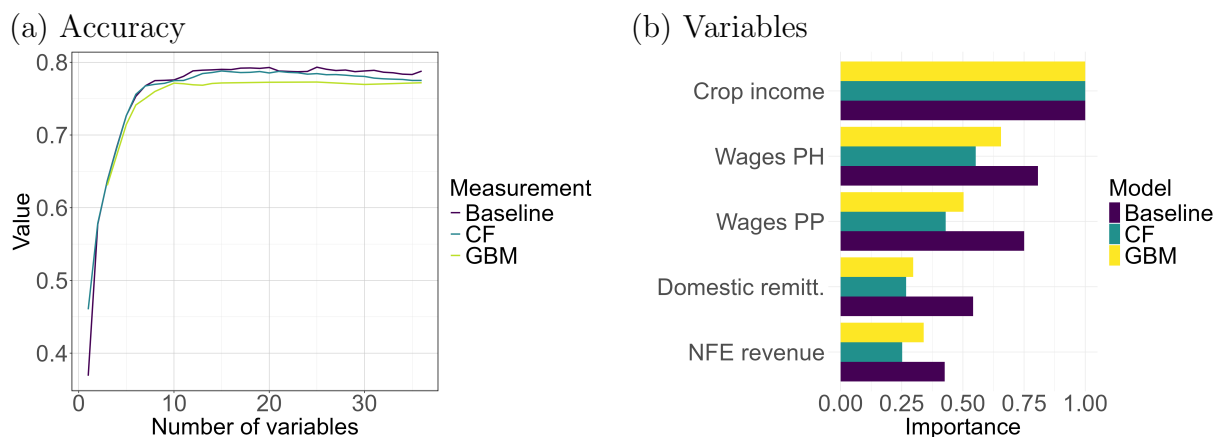


Figure 11: Model accuracy and most relevant variables in predicting income quintile classification across alternative machine-learning methods

Source: Authors' compilation.

Note: Variable importance measures were normalised so that 1 represents the maximum importance, ensuring comparability across methods.

For robustness purposes, alternative machine-learning algorithms are considered in addition to the baseline random forest specification. The objective is to assess whether the predictive performance and the ranking of the most relevant variables are sensitive to the choice of classification algorithm. The baseline specification uses a conventional random forest classifier. Random forests construct a large number of decision trees using bootstrap samples and random subsets of predictors and aggregate predictions across trees. This approach has several advantages, including high predictive accuracy, robustness to overfitting, the ability to capture complex nonlinear relationships and interactions, and relatively limited tuning requirements. However, conventional random forests may exhibit some limitations. Variable importance measures can be biased toward variables with many possible splitting values or variables displaying high variance. Furthermore, standard random forests treat predictors symmetrically and do not explicitly account for potential biases arising from correlated variables.

To address these limitations, two alternative models are considered. Conditional inference forests replace the split-selection procedure used in conventional random forests with a statistical testing framework (Hothorn et al., 2006). Instead of selecting variables according to impurity reduction criteria, splits are chosen based on permutation tests. This procedure reduces the tendency of standard random forests to favour predictors with many possible split points or larger variability (Strobl et al., 2007). Consequently, conditional inference forests may produce more reliable variable importance rankings when predictors differ substantially in scale or distribution. However, these models are computationally more demanding and may occasionally yield slightly lower predictive performance than conventional random forests.

Gradient boosting (GB) constructs trees sequentially rather than independently. Each new tree is estimated to improve the prediction errors generated by previous trees (Friedman et al., 2001). This iterative process often leads to stronger predictive performance and may capture subtle non-linear patterns that standard random forests fail to iden-

tify. Nevertheless, GB models require more extensive hyperparameter tuning and may be more sensitive to overfitting if the number of trees, learning rate, or tree depth are not appropriately selected.

Figure 11 illustrates the predictive accuracy of the alternative estimation frameworks alongside our baseline methodology in predicting individuals' positions within the income distribution. The results indicate that the ensemble methods based on random forests exhibit comparable classification performance, while gradient boosting yields slightly lower performance.

The right panel of Figure 11 presents the relative importance scores of the top five predictors driving these income quintile classifications. Because raw feature importance figures are not directly comparable across different methodologies, we normalised these, so that the maximum value equals one within each method. Interestingly, all three approaches identify the same key predictors of income quintile classification. Crop income emerges as the most important variable, followed by earnings in the post-harvest period and, closely thereafter, earnings in the post-planting period. For the forest-based methodologies, the next most critical predictors are domestic remittances within Nigeria, followed by the revenue generated from non-farm enterprises. In the case of gradient boosting, however, the relative importance of these last two variables is reversed.

8 Conclusion

Reliable poverty and inequality measurement requires detailed information on household income and consumption, yet collecting such data through conventional household income and expenditure surveys is costly, time-consuming and demanding for both statistical agencies and respondents. This paper has examined whether reduced survey instruments can retain sufficient distributional information to support poverty and inequality measurement and monitoring. Using data from the Nigeria General Household Survey, we applied Random Forest Recursive Feature Elimination to identify a parsimonious set of income sources and consumption categories capable of reproducing key classifications derived from full welfare aggregates.

The analysis focused on three distributional outcomes: poverty status, quintile classification and position relative to the Gini-based inequality line. The results show that reduced models can achieve strong predictive performance, although accuracy varies across welfare concepts and information sets. For consumption, poverty status and inequality-line position are classified accurately with a small number of expenditure categories. Quintile classification is more demanding, especially when annual consumption is predicted from a single seasonal visit, but the aggregated consumption specification still achieves high accuracy for seasonal welfare rankings. The results also show that broader consumption categories can outperform more granular item-level information, suggesting that aggregation may reduce idiosyncratic reporting noise while preserving the budget components most relevant for welfare classification.

The consumption results highlight the central role of food expenditure and housing-related costs in reduced-data welfare measurement. Food, meals consumed outside the home and rent are consistently among the most informative predictors of consumption quintiles, while roots, vegetables and grains are especially important for poverty classification. The analysis also shows that seasonal information matters. Post-planting and post-harvest models differ in the relative importance assigned to essential food items, dis-

cretionary spending and housing-related expenditure, reinforcing the need to account for agricultural seasonality when designing shorter welfare modules in rural and semi-rural contexts.

The income results point to a similarly concentrated set of informative variables. Crop sales, labour earnings, remittances, non-farm enterprise revenues, livestock sales, pensions and rental income account for much of the predictive information needed to classify individuals within the income distribution. Income poverty status can be predicted with high accuracy using only a few variables, while position relative to the Gini-based inequality line is captured particularly well by labour earnings. These findings suggest that reduced income modules should preserve information on the main revenue-generating activities across agriculture, labour markets, household enterprises and transfers, rather than focusing narrowly on a single source of income.

The paper contributes to the literature on welfare measurement in two ways. First, it moves beyond approaches designed primarily for poverty targeting or relative wealth ranking by evaluating whether a reduced set of variables can preserve multiple distributional classifications simultaneously. Second, it provides evidence that machine-learning methods can support survey design by identifying the smallest set of questions that achieves a high and stable level of predictive performance. This is particularly relevant for countries where the cost of full household income and expenditure surveys limits the frequency of poverty and inequality monitoring.

Several caveats are important. The analysis is based on a single country, and the selected predictors may partly reflect Nigeria's specific economic structure, agricultural calendar and survey design. The results should therefore be interpreted as evidence of feasibility rather than as a universal reduced questionnaire. Cross-country validation is needed to assess whether similar sets of variables perform well in other low- and middle-income contexts. In addition, the current exercise evaluates classification accuracy rather than the direct recovery of continuous welfare levels or inequality indices. Reduced instruments may classify households accurately around relevant thresholds while still compressing the tails of the distribution, an issue that deserves further investigation.

Overall, the findings suggest that carefully designed reduced survey instruments, supported by supervised feature selection methods, can retain a substantial amount of the information required for poverty and inequality monitoring. Such tools should not replace full household surveys, which remain essential for constructing benchmark welfare aggregates and recalibrating reduced instruments. However, they can complement comprehensive surveys by enabling more frequent, lower-cost and operationally feasible monitoring of welfare classifications between full survey rounds. In this sense, RF-RFE and related machine-learning methods offer a promising route for strengthening statistical capacity and improving the timeliness of distributional evidence in resource-constrained settings.

Acknowledgements

The authors are grateful to participants of the WIDER Development Conference 2026 "Green Industrialisation and Inclusive Growth in a Fractured World Order" in New Delhi, India for their comments and suggestions on early versions of this paper. VJ acknowledges financial support from the I+D+i project Ref. PID2024-156871NB-I00 financed by MICIU/AEI/10.13039/501100011033/FEDER,UE. All errors are solely the responsibility of the authors.

References

- Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3), 244–263.
- Barrett, C. B., Reardon, T., & Webb, P. (2001). Nonfarm income diversification and household livelihood strategies in rural africa: Concepts, dynamics, and policy implications. *Food Policy*, 26(4), 315–331. doi: 10.1016/S0306-9192(01)00014-8
- Beegle, K., De Weerd, J., Friedman, J., & Gibson, J. (2012a). Methods of household consumption measurement through surveys: Experimental results from tanzania. *Journal of Development Economics*, 98(1), 3–18. doi: 10.1016/j.jdeveco.2011.11.001
- Beegle, K., De Weerd, J., Friedman, J., & Gibson, J. (2012b). Methods of household consumption measurement through surveys: Experimental results from tanzania. *Journal of Development Economics*, 98(1), 3–18.
- Beegle, K., De Weerd, J., Friedman, J., & Gibson, J. (2012c). *Methods of household consumption measurement through surveys in developing countries*. The World Bank.
- Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics, volume 5* (pp. 3705–3843). Elsevier.
- Bourguignon, F. (2015). *The globalization of inequality*. Princeton University Press.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. doi: 10.1023/A:1010933404324
- Canberra Group. (2011). *Canberra group handbook on household income statistics*. United Nations Economic Commission for Europe.
- Carletto, C., Jolliffe, D., & Banerjee, R. (2015). From tragedy to renaissance: Improving agricultural data for better policies. *The Journal of Development Studies*, 51(2), 133–148. doi: 10.1080/00220388.2014.968140
- Chakraborty, N. M., Fry, K., Behl, R., & Longfield, K. (2016). A new approach to measure socioeconomic status in low- and middle-income countries using the equitytool. *Global Health: Science and Practice*, 4(4), 503–515.
- Coady, D., Grosh, M., & Hoddinott, J. (2004). *Targeting of transfers in developing countries: Review of lessons and experience*. Washington, DC: World Bank.
- Cowell, F. A. (2011). *Measuring inequality* (3rd ed.). Oxford University Press.
- Cowell, F. A., & Mehta, F. (1982). The estimation and interpolation of inequality measures. *Review of Economic Studies*, 49(2), 273–290.
- Cowell, F. A., & Victoria-Feser, M.-P. (1996). Robustness properties of inequality measures. *Econometrica*, 64(1), 77–101.
- Deaton, A. (1997a). *The analysis of household surveys: a microeconomic approach to development policy*. World Bank Publications.

- Deaton, A. (1997b). *The analysis of household surveys: A microeconomic approach to development policy*. Baltimore, MD: World Bank and Johns Hopkins University Press.
- Deaton, A., & Muellbauer, J. (1980). *Economics and consumer behavior*. Cambridge university press.
- Deaton, A., & Zaidi, S. (2002a). *Guidelines for constructing consumption aggregates for welfare analysis*. World Bank.
- Deaton, A., & Zaidi, S. (2002b). *Guidelines for constructing consumption aggregates for welfare analysis*. Washington, DC: World Bank.
- Dercon, S., & Krishnan, P. (2000). Vulnerability, seasonality and poverty in ethiopia. *Journal of Development Studies*, 36(6), 25–53. doi: 10.1080/00220380008422653
- Eurostat. (2021). *Glossary: At-risk-of-poverty rate*. Statistics Explained. (At-risk-of-poverty threshold defined as 60 percent of national median equivalised disposable income after social transfers)
- Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of india. *World Bank Economic Review*, 15(1), 115–132.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer.
- Gibson, J., & Kim, B. (2007). Measurement error in recall surveys and the relationship between household size and food demand. *American Journal of Agricultural Economics*, 89(2), 473–489.
- Grosh, M., & Baker, J. (1995). *Proxy means tests for targeting social programs: Simulations and speculation* (Tech. Rep. No. 118). World Bank, Living Standards Measurement Study (LSMS).
- Grosh, M., & Glewwe, P. (1998). Data watch: The world bank’s living standards measurement study household surveys. *Journal of Economic Perspectives*, 12(1), 187–196.
- Grosh, M., & Glewwe, P. (Eds.). (2000). *Designing household survey questionnaires for developing countries: Lessons from 15 years of the living standards measurement study*. Washington, DC: World Bank.
- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. In *Machine learning* (Vol. 46, pp. 389–422).
- Hagenaars, A., & et al. (1994). Statistical matching of microdata files: Theory and practice. *Journal of Official Statistics*, 10(1), 53–64.
- Harttgen, K., & Vollmer, S. (2013). A comparison of asset indices and consumption expenditure as welfare measures in developing countries. *World Development*, 41, 1–13.

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. doi: 10.1198/106186006X133933
- Howe, L. D., Hargreaves, J. R., & Huttly, S. R. A. (2008). Measuring socio-economic position for epidemiological studies in low- and middle-income countries: A methods of measurement in epidemiology paper. *International Journal of Epidemiology*, *37*(4), 871–881.
- Hurd, M. D., Kreuter, F., & Winter, J. (2004). Unfolding brackets in surveys: Efficiency gains and reduced nonresponse. *Journal of Official Statistics*, *20*(3), 595–612.
- Independent Evaluation Group. (2023). *Poverty mapping: Innovative approaches to creating poverty maps from new data sources* (Tech. Rep.). Washington, DC: World Bank.
- Kennickell, A. B. (1998). *Multiple imputation in the survey of consumer finances* (Tech. Rep.). Board of Governors of the Federal Reserve System. (Available at: <https://www.federalreserve.gov/econres/scfindex.htm>)
- Kilic, T., Serajuddin, U., Uematsu, H., & Yoshida, N. (2017). *Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity* (Policy Research Working Paper No. 7951). World Bank.
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324. doi: 10.1016/S0004-3702(97)00043-X
- Lanjouw, P., & Ravallion, M. (1995). Poverty and household size. *The Economic Journal*, *105*(433), 1415–1434.
- Lopez-Noval, B., Niño-Zarazúa, M., Roope, L., & Tarp, F. (2024). *From the bottom 40 to inequality lines: Sharing prosperity globally and domestically* (WIDER Working Paper No. 2024/77). Helsinki: UNU-WIDER. Retrieved from <https://doi.org/10.35188/UNU-WIDER/2024/540-0> doi: 10.35188/UNU-WIDER/2024/540-0
- McKenzie, D. J. (2005). Measuring inequality with asset indicators. *Journal of Population Economics*, *18*(2), 229–260.
- Montgomery, M. R., Gagnolati, M., Burke, K. A., & Paredes, E. (2000). Measuring living standards with proxy variables. *Demography*, *37*(2), 155–174.
- Moore, J. C., Stinson, L. L., & Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, *16*(4), 331–361.
- National Bureau of Statistics. (2019). *Nigeria General Household Survey-Panel 2018/19, Wave 4*. World Bank Microdata Library. (Reference ID: NGA_2018_GHSP-W4_v03_M)
- National Bureau of Statistics. (2020). *2019 poverty and inequality in nigeria: Executive summary* (Tech. Rep.). Abuja.

- OECD. (2011). *Society at a glance 2011: Oecd social indicators*. Paris: OECD Publishing. doi: 10.1787/soc_glance-2011-en
- Paxson, C. H. (1993). Consumption and income seasonality in thailand. *Journal of Political Economy*, 101(1), 39–72. doi: 10.1086/261865
- Piketty, T. (2014). *Capital in the twenty-first century*. Harvard University Press.
- Pradhan, M., & Ravallion, M. (2000). Measuring poverty using qualitative perceptions of consumption adequacy. *Review of Economics and Statistics*, 82(3), 462–471.
- Ravallion, M. (1994). *Poverty comparisons*. Chur, Switzerland: Harwood Academic Publishers.
- Ravallion, M. (1998a). *Poverty lines in theory and practice*. World Bank.
- Ravallion, M. (1998b). *Poverty lines in theory and practice* (Living Standards Measurement Study Working Paper No. 133). Washington, DC: World Bank.
- Ravallion, M., & Lokshin, M. (2002). Self-rated economic welfare in russia. *European Economic Review*, 46(8), 1453–1473.
- Roope, L. S. J. (2021). First estimates of inequality benchmark incomes for a range of countries. *PLOS ONE*, 16(3), e0248178. doi: 10.1371/journal.pone.0248178
- Sahn, D. E., & Stifel, D. (2003). Exploring alternative measures of welfare in the absence of expenditure data. *Review of Income and Wealth*, 49(4), 463–489.
- Schreiner, M. (2010). *A simple poverty scorecard for Country: Design and use*. St. Louis, MO: Microfinance Risk Management, L.L.C.
- Smits, J., & Steendijk, R. (2011). Towards a better understanding of socioeconomic inequality in child health: Asset-based wealth indices versus consumption-based measures. *Social Science & Medicine*, 72(3), 353–361.
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, 25. doi: 10.1186/1471-2105-8-25
- Vyas, S., & Kumaranayake, L. (2006). Constructing socio-economic status indices: How to use principal components analysis. *Health Policy and Planning*, 21(6), 459–468.
- World Bank. (2022). *Poverty and shared prosperity 2022: Correcting course*. Washington, DC: World Bank.

9 Appendix A: Construction of income and consumption aggregates

This section provides additional detail on the construction of the benchmark income and consumption aggregates used in the empirical analysis.

A1. Income data

Household income is constructed from six main components: labour income, agricultural crop income, livestock income, non-farm enterprise income, remittances and transfers, and other income sources. All components are converted to annual values and aggregated at the household level.

Labour income. Labour income is constructed using information from the labour modules of the GHS-Panel, which collect information on wage and salaried employment activities, occupation, sector of employment, payment frequency, hours worked, and remuneration received in cash and in kind. Let Y_h^L denote labour income of household h . Labour income is computed as:

$$Y_h^L = \sum_{i \in h} (Y_i^C + Y_i^{IK}), \quad (4)$$

where Y_i^C denotes annual cash earnings received by individual i , and Y_i^{IK} denotes the annualised value of in-kind payments received through wage employment.

The survey records labour earnings using different payment frequencies, including hourly, daily, weekly, biweekly, monthly, quarterly and annual remuneration. To ensure comparability, all earnings are converted to annual values. Hourly earnings are annualised using information on usual weekly working hours, while daily earnings are annualised assuming six working days per week. In-kind remuneration is valued using the monetary amounts reported by households. Labour supplied to household farms, household enterprises, apprenticeships and unpaid family activities is not directly monetised within this component.

Non-farm enterprise income. Non-farm enterprise income is constructed from the post-harvest household questionnaire, since the non-farm enterprise module is collected only during the post-harvest visit in Wave 4. The module covers household businesses and income-generating activities outside agriculture, including petty trade, transport services, handicrafts, manufacturing, food processing, retail commerce and other informal self-employment activities.

Household non-farm enterprise income is defined as net enterprise income during the reference period. Let Y_h^{NF} denote non-farm enterprise income of household h . This component is computed as:

$$Y_h^{NF} = \sum_{e \in h} (R_{eh} - C_{eh}), \quad (5)$$

where R_{eh} denotes gross revenue generated by enterprise e , and C_{eh} denotes operating costs, including expenditures on raw materials, transport, utilities, hired labour, maintenance and other business-related expenses. Enterprise income is measured at the enterprise level and reflects joint returns to labour, capital and entrepreneurial activity.

Agricultural income. Agricultural income is constructed using the post-harvest agricultural modules, which collect information on crop production, expected harvests and agricultural input expenditures. Let Y_h^A denote agricultural income of household h . Agricultural income is defined as the net value of crop production:

$$Y_h^A = V_h^A - C_h^A, \quad (6)$$

where V_h^A denotes the gross value of agricultural production and C_h^A denotes total agricultural input expenditures.

The value of agricultural production is constructed from the crop harvest modules. For each crop cultivated by the household, the survey records the estimated value of the quantity already harvested from each plot. Let V_{hpc}^H denote the reported value of harvested crop c from plot p , and Q_{hpc}^H the corresponding quantity harvested, converted into standardised units. The crop-specific unit value is computed as:

$$P_{hpc} = \frac{V_{hpc}^H}{Q_{hpc}^H}. \quad (7)$$

Where part of the crop remains unharvested at the time of interview, the expected remaining harvest is valued as:

$$V_{hpc}^E = P_{hpc} \times Q_{hpc}^E, \quad (8)$$

where Q_{hpc}^E denotes the expected quantity still to be harvested, also expressed in standardised units.⁸

Total crop production value is therefore defined as:

$$V_h^A = \sum_{p,c} (V_{hpc}^H + V_{hpc}^E). \quad (9)$$

Agricultural production costs include purchased inputs, such as fertilisers, pesticides and seeds; rental costs of agricultural land; rental costs of equipment and machinery; and transport costs associated with input acquisition.⁹ Since the survey does not provide sufficient information to fully value unpaid family labour, land services or capital depreciation, this measure should be interpreted as a net cash-based approximation of agricultural income rather than full economic profit.

Livestock income. Livestock income is constructed using the livestock production modules, which collect information on animal sales, milk production, egg production and livestock-related expenditures. Let Y_h^{LV} denote livestock income of household h . This variable is defined as:

$$Y_h^{LV} = R_h^{LV} - C_h^{LV}, \quad (10)$$

⁸Crop quantities may be reported under different conditions, such as shelled versus unshelled maize or paddy versus processed rice. To ensure consistency in valuation, expected production is imputed only when the condition of the harvested and expected crop quantities coincides.

⁹Cash rental payments reported for a six-month period are annualised by multiplying the observed payment by two. Payments associated with rainy-season and dry-season contracts are treated as covering a full agricultural cycle and are therefore not further annualised.

where R_h^{LV} represents gross livestock revenues and C_h^{LV} denotes livestock production expenditures.

Gross livestock revenues include live animal sales, slaughtered animal sales, and revenues derived from livestock products such as milk and eggs. Where livestock products are consumed within the household, values are imputed using household-specific unit values derived from observed sales where available, or median unit values among households reporting positive sales. Egg income is measured conservatively using only the reported value of egg sales, since the survey does not identify the final destination of eggs not sold.

Livestock production costs include vaccination, veterinary services, water, purchased feed, hired labour, compensation payments and other livestock-related operating costs. Purchases of animals are treated as asset acquisitions and are not deducted as current production costs. The current value of owned livestock is also excluded to avoid conflating income flows with wealth holdings.

Remittances, transfers and other income. Remittances and private transfers include domestic and international transfers received by the household in cash or in kind. Domestic transfers are reported in Nigerian Naira. Foreign remittances reported in U.S. dollars, euros or pounds sterling are converted into Nigerian Naira using average annual exchange rates for 2018.¹⁰ Transfers received in kind are incorporated using the monetary values reported by respondents.

Other income sources include rental income, pensions, returns on savings and investments, social assistance programmes and miscellaneous receipts. Rental income includes both agricultural and non-agricultural property income. Agricultural rental income includes cash and in-kind payments received for renting out agricultural plots, while non-agricultural rental income includes earnings from residential or commercial property. Social assistance includes government support programmes, food assistance and other transfers received by household members, valued using reported monetary amounts where available.

A2. Consumption data

Household consumption is constructed as the total monetary value of food consumption, non-food expenditure and housing services:

$$C_h = C_h^F + C_h^{NF} + C_h^H, \quad (11)$$

where C_h^F denotes food consumption, C_h^{NF} denotes non-food consumption expenditure, and C_h^H denotes housing-related consumption.

Food consumption. The food consumption modules, collected in both post-planting and post-harvest visits, include 16 food categories and more than 160 individual food items. Food consumption includes purchases, own-produced food, gifts and meals consumed outside the home. Since households report food quantities in multiple standard and local units, quantities are first converted into standardised units using the conversion factors available in the survey.

¹⁰Foreign remittances are converted using average exchange rates of 360 NGN/USD, 424 NGN/EUR and 482 NGN/GBP, respectively.

Food values are estimated using the quantity purchased in the most recent purchase and the corresponding reported value. Where household-specific purchase information is unavailable, national average prices are assigned, taking into account the size of the product where this information is available. The final food aggregate includes the value of purchased food, own-produced food and food received as gifts, valued at market prices.

Expenditure on meals consumed outside the home is reported for the previous seven days and includes breakfast, lunch, dinner, side dishes, dairy-based beverages, vegetables, and alcoholic and non-alcoholic drinks. These expenditures are annualised using a factor of 365/7.

Non-food consumption. Non-food consumption is constructed from a broad set of expenditure items grouped into 13 categories. The survey records these expenditures over different recall periods, including the last seven days, last 30 days, last six months and last 12 months. Expenditures are annualised using the corresponding recall-period conversion factors.

Items that do not correspond to current consumption are excluded. These include mortgages, donations, production inputs and highly irregular ceremonial expenditures. Education expenditure includes school fees, uniforms, textbooks, meals and lodging, transport, gifts to teachers, private tutoring and other school-related costs. Where detailed education components are unavailable, the aggregate education expenditure variable is used. Health expenditure includes consultations, medicines, laboratory exams, hospitalisation charges and transport to health facilities. Health expenses reported for the previous four weeks are annualised using a factor of 52/4.

Housing. Housing consumption is measured as the value of housing services received by the household. For tenants, actual rent paid is used as the value of housing services. For owner-occupiers and non-market tenants, rent is imputed using a log-linear hedonic model estimated among market tenants:

$$\log(r_i) = \alpha + \beta_1 R_i + \beta_2 X_i + \epsilon_i, \quad (12)$$

where r_i denotes reported rent for dwelling i , R_i is a vector of regional indicators, X_i is a vector of dwelling characteristics, including the number of rooms, wall, floor and roof materials, and sanitation type, and ϵ_i is the error term. Parameter estimates are presented in Table A1.

Because the model is estimated in logarithms, predicted values are retransformed using Duan's smearing estimator:

$$\hat{r}_i = \exp\left(\hat{\alpha} + \hat{\beta}_1 R_i + \hat{\beta}_2 X_i\right) \cdot \hat{E}[\exp(\epsilon_i)], \quad (13)$$

where the second term is the sample average of the exponentiated residuals from the log-linear model.

A3. Adjustments to household income and consumption

All income and consumption components are harmonised to annual values. Nominal values are then expressed in real terms using Consumer Price Index data from the Nigerian Bureau of Statistics. These adjustments account for both temporal price differences across

Table A1: Ordinary Least Squares Estimates

	Estimate (Std. error)
Constant	7.2031 (0.2319)***
Number of rooms	0.0595 (0.0262)*
<i>Zone (Ref.: North Central)</i>	
North East	0.5329 (0.0263)***
North West	0.1770 (0.1200)
South East	0.2972 (0.0556)***
South South	0.3032 (0.0563)***
South West	0.1660 (0.0519)**
<i>Type of dwelling (Ref.: Separate house)</i>	
Semi-detached house	0.0635 (0.1636)
Apartment/flat	0.6825 (0.2545)**
Compound house	-0.1737 (0.0629)**
Other	-0.2046 (0.1015)*
<i>Walls material (Ref.: Mud)</i>	
Unburnt bricks	0.2995 (0.0894)***
Burnt bricks	0.3409 (0.2785)
Cement or concrete	0.2384 (0.0647)***
Other	0.4675 (0.3153)
<i>Roof material (Ref.: Corrugated iron sheets)</i>	
Cement or concrete	0.1920 (0.2038)
Asbestos sheet	0.4578 (0.0792)***
Long/short span sheets	0.2471 (0.1297).
Other	0.2025 (0.3545)
<i>Floor material (Ref.: Sand, dirt, straw)</i>	
Smoothed mud	-0.0417 (0.2834)
Smooth cement/concrete	0.5936 (0.2642)*
Tile	1.1249 (0.3485)**
Terrazo	0.9163 (0.4877).
Other	1.4570 (0.3763)***
<i>Sanitary type (Ref.: Flush to piped sewage system)</i>	
Flush to septic tank	-0.1443 (0.1488)
Flush to pit latrine	-0.4586 (0.1920)*
Pit latrine with slab	-0.7509 (0.0908)***
Pit latrine without slab/open pit	-0.5269 (0.1857)**
Hanging toilet/hanging latrine	-0.6757 (0.0830)***
No facilities, bush or field	-1.1707 (0.2352)***
Other	-0.3087 (0.0511)***
Adjusted R-squared	0.5721
Number of observations	1006
F-statistic	44.2***

Note: Standard errors clustered at the zone level in parentheses. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$.

interview months and spatial price differences between rural and urban areas, ensuring that welfare aggregates are comparable across households interviewed at different times and in different locations.

Appendix B

Table B1: Items Included in Food Categories

Item	Code	Item	Code
Grains (1)			
Guinea corn/sorghum	10	Maize (Unshelled)	20
Millet	11	Maize (Shelled)	22
Rice - local	13	Other grains and flour	23
Rice - Imported	14	Yam flour	17
Maize flour	16	Cassava flour	18
Wheat flour	19		
Processed Foods & Snacks (2)			
Bread	25	Biscuits	28
Cake	26	Meat Pie/Sausage Roll	29
Buns/Pofpof/Donuts	27		
Roots and Tubers (3)			
Cassava - roots	30	Sweet potatoes	36
Yam - roots	31	Potatoes	37
Gari - white	32	Other roots and tuber	38
Gari - yellow	33	Plantains	35
Cocoyam	34		
Nuts and Pulses (4)			
Soya beans	40	Groundnuts (Shelled)	44
Brown beans	41	Other nuts/seeds/pulses	45
White beans	42	Coconut	46
Groundnuts (Unshelled)	43	Cashew nut	48
Kola nut	47		
Oils and Fats (5)			
Palm oil	50	Other oil and Fat	53
Butter/ Margarine	51	Animal fat	56
Groundnuts Oil	52		
Fruits (6)			
Bananas	60	Mangoes	62
Orange/tangerine	61	Avocado pear	63
Pineapples	64	Other fruits	66
Pawpaw	67	Watermelon	68
Apples	69		
Vegetables (7)			
Tomatoes	70	Garden eggs/egg plant	73
Tomato puree (canned)	71	Okra - fresh/dried	74/75
Onions	72	Fresh/Dry Pepper	76/77
Leaves (Cocoyam, Spinach)	78	Other vegetables	79

Continued on next page

Table B1: Items Included in Food Categories (continued)

Item	Code	Item	Code
Poultry and Eggs (8)			
Chicken	80	Agricultural eggs	83
Duck	81	Local eggs	84
Other domestic poultry	82	Other eggs (not chicken)	85
Meat (9)			
Beef	90	Goat	93
Mutton	91	Wild game/bush meat	94
Pork	92	Canned beef/corned beef	95
Other meat	96		
Fish and Seafood (10)			
Fish - fresh	100	Fish - dried	103
Fish - frozen	101	Snails	104
Fish - smoked	102	Seafood (lobster, crab, etc)	105
Canned fish/seafood	106	Other fish or seafood	107
Milk and Dairy (11)			
Fresh milk	110	Baby milk powder	112
Milk powder	111	Milk tinned (unsweetened)	113
Cheese (wara)	114	Other milk products	115
Coffee, Tea and Breakfast (12)			
Coffee	120	Tea	122
Chocolate drinks (Milo)	121	Sugar	130
Honey	132	Other sweets/confectionary	133
Miscellaneous and Spices (13)			
Salt	141	Melon (shelled/unshelled)	145/146
Unground/Ground Ogbono	142/143	Other Spices	148
Ground Pepper	144		
Non-Alcoholic Beverages (14)			
Bottled water	150	Soft drinks (Coca Cola, etc)	153
Sachet water	151	Fruit juice canned/Pack	154
Malt drinks	152	Other non-alcoholic drinks	155
Alcoholic Beverages (15)			
Beer (local and imported)	160	Pito	162
Palm wine	161	Gin	163
Other alcoholic beverages	164		

Table B2: Items Included in Non-Food Categories

Item	Code	Item	Code
<i>Smoking</i>			

Continued on next page

Table B2: Items Included in Non-Food Categories (continued)

Item	Code	Item	Code
Cigarettes or tobacco	101	Wages paid to staff/maid/lawnsboy	325
Matches	102	Repairs & maintenance to dwelling	327
		Repairs to household and personal items	328
		Building items – cement, bricks, etc.	508
<i>Utilities and Fuels</i>			
Electricity	305	Infant clothing	401
Light bulbs/globes	311	Baby nappies/diapers	402
Water	312	Boys tailored clothes	403
Kerosene	301	Boys dress (ready made)	404
Palm kernel oil	302	Girls tailored clothes	405
Gas (lighting/cooking)	303	Girls dress (ready made)	406
Other liquid cooking fuel	304	Men tailored clothes	407
Candle	306	Men dress (ready made)	408
Firewood	307	Women tailored clothes	409
Charcoal	308	Women dress (ready made)	410
Petrol	309	Ankara/George materials	411
Diesel	310	Other clothing materials	412
Lubricants (oil, grease, etc.)	330	Tailoring charges	417
		Laundry and dry cleaning	418
		Hand loomed: ASO-OKE	431
<i>Personal Care</i>			
Soap and washing powder	313	Boy's shoes	413
Toilet paper	314	Men's shoes	414
Personal care goods	315	Girl's shoes	415
Vitamin supplements	316	Lady's shoes	416
Insecticides, disinfectant, cleaners	317	Repairs of footwear	432
<i>Recreation and Communication</i>			
Books (not for school)	426	Bowls, glassware, plates, silverware	419
Night lodging (house/hotel)	428	Cooking utensils	420
Newspapers and magazines	103	Cleaning utensils (brooms, brushes)	421
Gambling, lotto, raffles	105	Torch/flashlight	422
Cinema, DVD rental	322	Umbrella and raincoat	423
Sports equipment, musical instruments, toys	506	Paraffin lamp	424
Film, processing, camera	507	Stationery (not for school)	425
Postal (stamps, courier)	318	House decorations	427

Continued on next page

Table B2: Items Included in Non-Food Categories (continued)

Item	Code	Item	Code
Recharge cards	319	Electric kettle	433
Landline charges	320	Coal pot/non-electric appliance	434
Internet services	321	Repairs of appliances	435
Cell phone handset	440	Bed sheets, blankets	436
Personal computer	441	Pillow	437
		Curtains and other linen	438
		Carpet and floor covering	439
		Carpet, rugs, drapes	501
		Linen (towels, sheets, blankets)	502
		Sleeping mat	503
		Mosquito net	504
		Mattress	505
<i>Transport and Insurance</i>			
Public transport (commuting)	104	Health insurance	510
Motor vehicle service/repair/parts	323	Auto insurance	511
Bicycle service/repair/parts	324	Home insurance	512
		Life insurance	513
<i>Taxes and Fines</i>			
Council rates	509	Fines or legal fees	514