# Inequality and Volatility

Joshua D. Merfeld* Jonathan Morduch†

January 16, 2026

**Abstract**

In most low- and middle-income countries, poverty and inequality are measured through household consumption surveys. We use a mathematical decomposition to show that data collected following expert guidelines leads to measures of inequality that conflate differences in consumption *between* households and the volatility of consumption *within* households over time. We propose a way to calculate both terms, using machine learning to predict annual consumption from cross-sectional data. We validate this method with month-level panel data from rural India. We then apply the method to three waves of the nationally-representative Indian National Sample Survey. We show that changes in observed inequality over time are often due to changes in volatility as much as to differences in resources between households. An application to India's large workfare program suggests that increased volatility also explains why the program appeared to increase measured inequality. More generally, the ML approach shows how existing data collection methodologies need to be—and can be—enhanced to deliver theoretically-consistent inequality measures.

*Keywords*: inequality, machine learning, inequality decomposition, income volatility, consumption smoothing, seasonality, household expenditure
*JEL Codes*: I31, I38, D63, C83

*School of Economics, University of Queensland, and IZA; j.merfeld@uq.edu.au
†Robert F. Wagner Graduate School of Public Service, New York University; jonathan.morduch@nyu.edu (corresponding author)

# 1 Introduction

Artificial intelligence is opening new ways to describe the economy through text analysis and machine learning approaches. We illustrate a second path for using AI to describe the economy. We show how AI can also be used recover an historically-important economic statistic of global importance that, at present, is based only on conventional methods and data. Our focus is on national inequality, a core metric of national economic performance (Pikkety, 2014; Ravallion, 2014; Bardhan et al., 2017; Milanovic, 2024).

We use a machine learning approach to overcome persistent problems with the household data on which inequality measurement often relies. We show that the data problems have led national inequality statistics to diverge from theoretical notions of inequality. Rather than responding by creating an alternative inequality statistic or predicting inequality with "alt-data" (Cukier, 2025), we show how AI can be used to extract more information from basic household data, taking advantage of the structure of sampling methodologies. This allows the recovery of a theoretically-consistent notion of inequality rooted in the experiences of households over the year. We illustrate the approach in three waves of India's National Sample Survey (National Sample Survey Organisation, 2001), and we show the implications for the measured impacts of India's Mahatma Gandhi National Rural Employment Act .

Inequality is generally measured using economic data collected through nationally-representative surveys of households, and most countries base inequality measurement on household consumption levels rather than income. This is true of all low-income countries, 90% of lower-middle income countries, and 62% of upper-middle income countries (Mancini and Vecchi, 2022). We focus on inequality in low- and middle-income countries and thus on household consumption, but a similar set of issues arises with income-based measures.

An immediate challenge for national statistical offices is that households' consumption can vary substantially across the year. (e.g., Khandker 2012, Devereux et al. 2012, and Dercon 2002). Statistical offices then generally take one of three paths. First, some simply decide to base inequality on conditions at a particular point in the year, ignoring variability during the rest of the year. Second, statistical agencies can go in the opposite direction and base inequality on conditions for each

household throughout the year. In practice, this is costly as it requires multiple visits to the same households and is relatively rare (Deaton and Grosh, 2000; Deaton and Zaidi, 2002; Mancini and Vecchi, 2022). The third strategy is to only interview each household once during the year but to collect data on different households at different times of year, thus incorporating the effects of seasons and other sources of instability.

What results is an array of approaches to collecting data which translate into different slices of inequality. The three approaches yield identical results only if households completely smooth consumption during the year, an assumption that finds no support in the data (e.g., Breza et al. 2021, Bryan et al. 2014, Pomeranz and Kast 2024, Fink et al. 2020, Casaburi and Willis 2018, and Augenblick et al. 2024).

We focus on inequality based on data collected following the third strategy. This is the choice adopted by the Indian National Survey Organization and is the basis of expert guidance from global economic agencies (FAO and World Bank, 2018).[1] The expert guidance holds that, in the face of budget constraints, statistical agencies should conduct one-time interviews of households, focusing on short-term consumption (in recognition of respondents' fallible memories). At the same time, statistical agencies should construct samples based on random stratification by period to account for variation across time. Statistical agencies then obtain data rooted in particular periods of the year, and we show how this re-shapes inequality measures.

First, we decompose the Theil-L inequality measure (also known as the mean log deviation or MLD; Theil 1967) to show what is obtained when calculating national inequality with data collected under the guidelines in FAO and World Bank (2018). We show that the resulting national statistics encompass two components. The first is a measure of inequality between households based on their average consumption over the year, a notion that aligns with theoretical concepts of inequality. The second is a measure of consumption volatility within the year, an idea that is of independent interest but which is outside the scope of conventional inequality

[1]The guidance emerges from an international collaborative, created by the Inter-Agency and Expert Group on Food Security, Agricultural and Rural Statistics, convened by the World Bank and UN Food and Agriculture Organization, and endorsed by the forty-ninth session of the United Nations Statistical Commission in 2018 (FAO and World Bank, 2018). See also Mancini and Vecchi (2022). The guidelines were disseminated as a guidebook for the World Bank's Livings Standards Measurement Survey (LSMS) program.

analyses. A consequence is that measured inequality is systematically upward-biased by the inclusion of the volatility component.

Second, we show an approach to retrieving more accurate measures of inequality by coupling existing data with machine-learning algorithms. The approach works even with only one short-term observation from each household in a single period of the year—which is all that is often available in practice. The approach takes advantage of conditions on the predictions that arise due to the randomization process when collecting the data; i.e., means of average consumption are correct in expectation for the population even if incorrect for any particular household.

Our aim is to describe an overlooked problem and show how AI provides a natural, workable solution that builds from existing survey data.[2] With the available data, we can now with confidence create a set of new inequality numbers for India. We thus conclude by reflecting on what kinds of data would allow improvements to the approach and what this means for the work of statistical agencies in a world with AI.

# 2 Measuring inequality

## 2.1 The measurement problem

We start with a single household $i$ which consumes consumption in the amount $c_{it}$ in period $t$. Over a year divided into $T$ periods of equal length, household $i$ thus consumes:

$$c_{i1}, c_{i2}, c_{i3}, ..., c_{iT}, \tag{1}$$

and their total consumption for the year is $c_{i1} + c_{i2} + c_{i3} + ... + c_{iT},$. Dividing by $T$ gives their average consumption for the year, $\bar{c}_i$.

Measures of inequality provide ways to summarize the distribution of the $\bar{c}_i$ in an economy. In the Indian National Sample Survey and survey designs that follow

[2]The problem remains overlooked even though Gibson et al. (2003) described the basic challenge and approached it with corrections based on correlations between the same household's expenditures in different months of the year, in the spirit of Scott (1992). a big advantage of the newer ML approach is that it is more flexible and can draw on a much wider range of data.

FAO and World Bank (2018), however, only one of the elements in equation (1) are collected, so $\bar{c}_i$ is not available for any household. Nonetheless, inequality can be calculated with the proxy, $c_i^*$, household consumption from the one randomly-selected period:

$$c_i{}^* = \sum_{t=1}^{T} c_{it} \cdot I_{it}, \tag{2}$$

where $I_{it}$ is an indicator which captures the randomized sampling process; $I_{it}$ is equal to 1 in the period that household $i$ was randomly selected to be surveyed and 0 in the other $(T-1)$ periods.[3]

While the sampled $c_i^*$ could be quite far from household $i$'s average consumption for the year ($\bar{c}_i$), taking averages across the $N$ households in the population yields $\mu$ in expectation, the true population average of consumption. Following Scott (1992),

$$\mu^* = \frac{1}{N} \sum_{i=i}^{N} c_i^*, \tag{3}$$

and

$$E[\mu^*] = \frac{1}{N} \sum_{i=i}^{N} E\left[c_i^*\right] = \frac{1}{N} \sum_{i=i}^{N} \sum_{t=i}^{T} \frac{1}{T} c_{it} = \frac{1}{N} \sum_{i=i}^{N} \frac{1}{T} \sum_{t=i}^{T} c_{it} = \frac{1}{N} \sum_{i=i}^{N} \bar{c}_i = \mu, \tag{4}$$

where the expectation relies on randomized sampling and thus that selecting any period is equiprobable with probability $1/T$.

## 2.2 Approximating inequality

Researchers have a choice of inequality indices, including the Gini index, but the Mean Log Deviation (MLD) is often used for policy analysis, thanks to the ability to decompose the MLD by population subgroups (Bourguignon 1979, Ravallion

[3]The quantity $c_i^*$ is based on expenditures near the survey date, but in practice $c_i^*$ may be adjusted to include a share of some items that are larger and less-frequently purchased during the year. In some countries, the value for consumption may be extrapolated to the quarterly or annual level, but the data still reflect the period in which respondents were interviewed.

2014, Milanovic 2024).[4] Relative to the Gini, the MLD puts heavier weight on the poorest parts of populations.

The MLD is a member of the Theil inequality measures (Theil-L; Theil 1967) and can be written as:

$$T_L = \frac{1}{N} \sum_{i=i}^{N} \ln\left(\frac{\mu}{\bar{c}_i}\right) \tag{5}$$

where $\bar{c}_i$ is the yearly mean of monthly consumption for household $i$ and $\mu$ is the yearly mean of monthly consumption of the entire population. With complete equality ($\bar{c}_i = \mu$ for all $i$) and $T_L = 0$.

Statistical agencies with data collected following FAO and World Bank (2018) have to approximate the MLD by subsitituting $c_i^*$ for $\bar{c}_i$ and $\mu^*$ for $\mu$. They thus calculate

$$T_L^* = \frac{1}{N} \sum_{i=i}^{N} \ln\left(\frac{\mu^*}{c_i^*}\right), \tag{6}$$

instead of equation (5). In expectation, the calculation yields a value that is always larger than the true $T_L$. We show that the gap can be interpreted in terms of an MLD version of a household volatility measure, which we label $V_L$:

$$\begin{aligned} E[T_L^*] &= E\left[\frac{1}{N} \sum_{i=i}^{N} \ln\left(\frac{\mu^*}{c_i^*}\right)\right] = \frac{1}{N} \sum_{i=i}^{N} \sum_{t=i}^{T} \frac{1}{T} \ln\left(\frac{\mu}{c_{it}}\right) \\ &= \frac{1}{N} \sum_{i=i}^{N} \frac{1}{T} \sum_{t=i}^{T} \ln\left(\frac{\mu}{\bar{c}_i} \frac{\bar{c}_i}{c_{it}}\right) \\ &= \frac{1}{N} \sum_{i=i}^{N} \ln\left(\frac{\mu}{\bar{c}_i}\right) + \frac{1}{N} \sum_{i=i}^{N} \frac{1}{T} \sum_{t=i}^{T} \ln\left(\frac{\bar{c}_i}{c_{it}}\right) \\ &= T_L + V_L, \end{aligned} \tag{7}$$

where we use scale invariance and equation (4) to simplify. The second expression on the penultimate line is a measure of within-year consumption inequality,

[4]Foster (1983) shows that the Theil index is the only Lorenz-consistent index that is exactly decomposable into inequality between and within groups. The World Bank provides national-level versions of the MLD at https://prosperitydata360.worldbank.org/en/indicator/WB+PIP+mld.

averaged across households, which we label $V_L$. In essence, this is a measure of volatility, but it can also be described as "within-inequality."[5] For farmers, $V_L$ captures variation between lean seasons and harvest seasons. For urban residents, $V_L$ captures times of slow work versus times of peak labor demand. Only when households perfectly smooth consumption within the year (i.e., when $c_{it} = \bar{c}_i$ for each period $t$ for each household $i$) will $V_L$ equal zero. Otherwise, within-year variation in household consumption leads to $V_L > 0$ and $\mathrm{E}[T_L^*] > T_L$.[6]

What the randomized sampling process delivers is thus not a noisy measure of the MLD, centered on the true MLD as in equation (5). Instead, it delivers a measure that is always larger than the true MLD by the quantity $V_L$.

Another way to say this is that what emerges are three distinct inequality quantities. The first is what economists want: *between-inequality* ($T_L$), reflecting differences in average consumption between households, given by the first term on the right hand side of equation (7). The second is essentially a nuisance variable: *within-inequality* ($V_L$), reflecting the shifting conditions of households within the year, given by the second term of equation (7). Neither quantity is seen independently. The third quantity is what statisticians can directly observe: *total-inequality* ($T_L^*$), which encompasses both the between and within components. In the rest of the paper, we show how machine learning can be used to dissaggregate total inequality in order to isolate its components.

# 3 Using ML to calculate between- and within-inequality

The problem in section 2.2 stems from the fact that, following expert guidelines, statisticians generally only observe $c_i^*$ and not $\bar{c}_i$. In the Indian NSS, for example, statisticians observe a single, randomly-chosen month of consumption (expenditures) for each household. The NSS and other household surveys, however, contain a large amount of other data, including modules devoted to demographics, assets,

[5] To our knowledge, the MLD has not been used before as a measure of volatility.

[6] The measure of volatility is particularly sensitive to inequality at the bottom of the distribution. $V_L$ is translation invariant, so if a household at the low end of the distribution has consumption that varies up and down by \$100 during the year, it will increase $V_L$ (and thus increase the measure of inequality $T_L^*$) more than if a better-off household experienced the same \$100 ups and downs. As with the broader MLD, the measure $V_L$ puts more weight on negative deviations from average consumption than positive deviations.

and employment. Our goal is to predict annual consumption for each household using the observed monthly data, for which we use the XGBoost algorithm (Chen and Guestrin, 2016). We do this by predicting consumption in every month for every household and then aggregating to the annual level.

We first validate this approach using a panel of monthly data from rural India that provides a complete set of expenditure data for each household over time. This is the Village Dynamics in South Asia (VDSA) Survey, collected by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). It is a balanced panel of household-level data from 2010-2014 that includes the incomes and expenditures of 945 low-income households collected monthly for at least four years.

The monthly data allows us to estimate annual consumption for each household with data and compare these estimates to the "true" value for each household. ICRISAT is not the ideal dataset for this purpose for two reasons. First, the data is not nationally representative (Merfeld and Morduch, 2025). Second, the sample is relatively small, meaning that our proposed estimation strategy—using a machine learning algorithm—may not perform as well as it would with a larger sample. However, the ICRISAT data is the only dataset we are aware of that contains a long panel of monthly expenditures for households in India, which is valuable for our validation exercise.

We proceed as follows:

1. Collect relatively stable variables for households in the ICRISAT data. These include household assets, head demographics, and other variables that are relatively stable over time.

2. For each wave (year), randomly select one observation from each household (i.e. one month). This will be the training sample.

3. Use these observations to estimate the XGBoost algorithm.

    - We use the log of consumption as the dependent variable.
    - We then estimate monthly expenditures for every household.

4. Aggregate these predictions to the household level for the year.

5. Using cross validation, estimate the correlations between actual monthly expenditures and predicted monthly expenditures, and actual mean annual expenditures and predicted mean annual expenditures.

## 3.1 Data

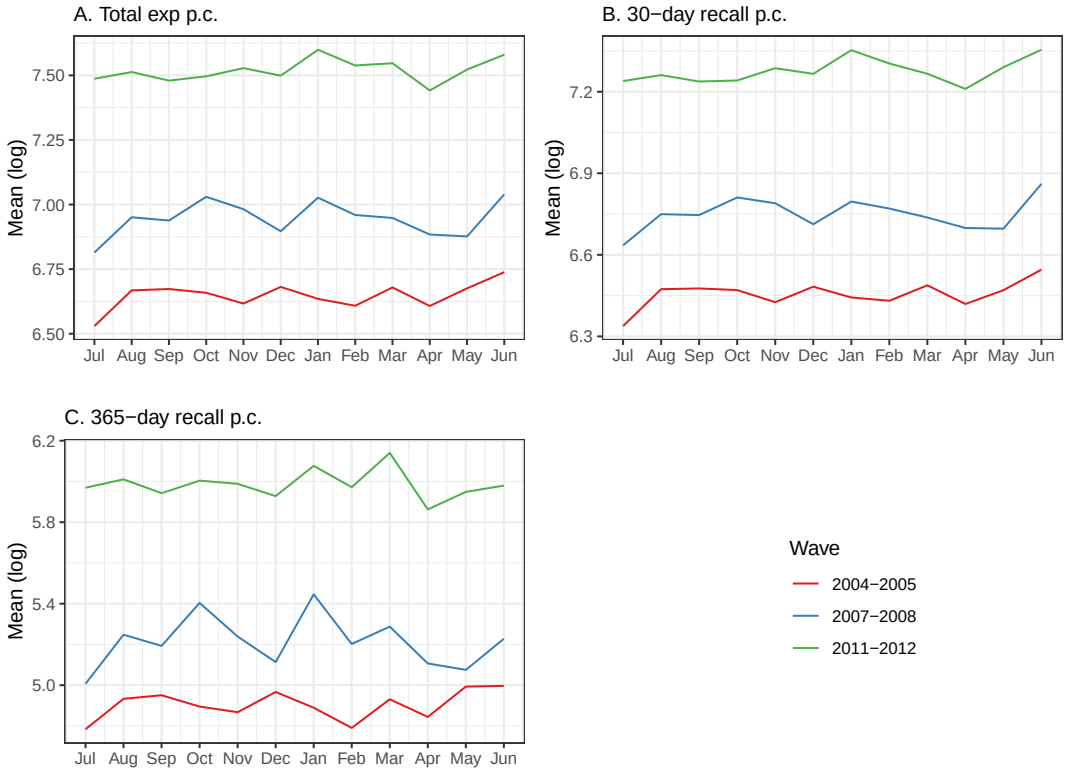
Our main dataset is the Indian National Sample Survey (NSS). The NSS is a large-scale, nationally representative survey that collects data on a wide range of topics, including consumption, employment, and demographics. We use three waves of the NSS data: 61, 64, and 68. The NSS 61 data is from 2004-2005, the NSS 64 data is from 2007-2008, and the NSS 68 data is from 2011-2012. These are the same waves used in previous papers that have studied the impact of the Mahatma Ghandi National Rural Employment Guarantee Scheme (NREGA) program on consumption inequality in India, including Imbert and Papp (2015) and Merfeld (2020).

The NSS was used as the basis for official poverty statistics in India for many years, although not without debate (Deaton and Dreze, 2002; Ghatak, 2022; Sinha Roy and Van Der Weide, 2022). Our assertion in the previous section is that inequality as conventionally measured is actually a mix between between- and within-household inequality. The NSS is an excellent example of why this is the case. Consider, for example, the 61st round of the NSS. The expenditure module collects data on expenditures in two separate ways. First, the survey collects data on consumption for the last 30 days for the most commonly bought/consumed items. This includes food, intoxicants, entertainment, rent, and non-institutional medical expenses. Second, the survey collects data on consumption for the last 365 days for less frequently bought items. This includes items like institutional medical expenses, tuition, clothing, and consumer durables. The 30-day recall constitutes the vast majority of the consumption data in the survey, so most of the data reflects consumption close to the date on which the households were interviewed. In the 61st wave, for example, the share of items asked with 30-day recall makes up slightly more than 77% of total expenditures.

Figure 1 shows log(expenditures) for each month in the 68th wave of the NSS, separately for the 365- and 30-day recall periods. The 365-day recall period shows more variation than the 30-day recall period (standard deviation of 0.045 relative to

a standard deviation of 0.035). However, both tend to move together; the correlation between the deviations is 0.830. Within-household changes in expenditures will contribute to measure inequality, meaning that measured inequality will be higher than if we only observed annual consumption. Our goal is to decompose these two components of inequality.

Figure 1: Monthly expenditures by wave




Notes: Panel A shows the (log of the) mean expenditures per capita for each month and NSS wave. We also split expenditures into the 30-day recall portion (Panel B) and the 365-day recall portion (Panel C).

## 3.2 The XGBoost algorithm

XGBoost (eXtreme Gradient Boosting) is a popular supervised learning algorithm (Chen and Guestrin, 2016). We chose it in part because it is accessible to others and well-documented. Here, we discuss the basics of XGBoost.

To understand XGBoost, it is important to first understand decision trees. Decision trees can be used for both regression and classification. In this paper, we

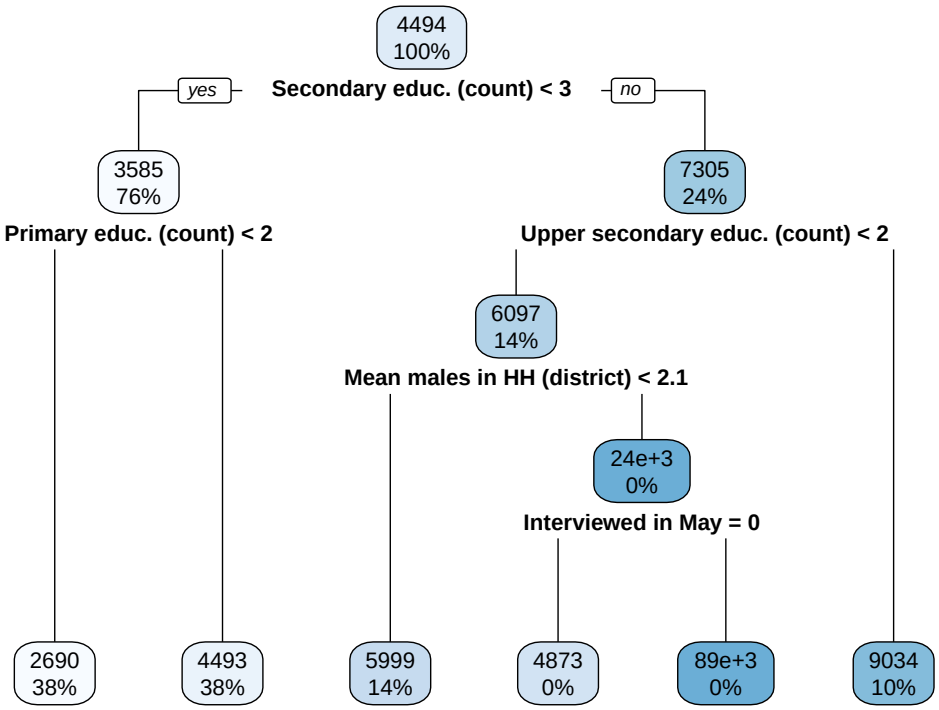
are interested in predicting a continuous outcome, so we ignore the classification aspect. A decision tree works by recursively splitting the data based on the values of independent variables that predict the outcome. The tree is built by selecting the variable that best splits the data into two groups, based on some criterion (e.g. minimizing the sum of squared errors). This process continues until some rule is met (e.g. a minimum number of observations at each split or a minimum decreases in squared errors).

Recall that we are estimating expenditures per capita at the household level, so the tree is built using household-level data. We include as predictors any variables that are arguably time invariant, at least within a single year.[7] For variables that are defined as 1/0 at the individual level, we aggregate the number of individuals in the household that have the characteristic. For example, if a household has three members and two of them have at least primary education, the primary education variable will have a two.[8] We include a large number of predictors, including household assets, demographics, usual occupation, state dummies, and month dummies.

[7]Since we only observe a household once in the NSS data, we do not want to use variables that are likely to change over time. For example, we do not include current employment status, but we do include usual employment status. We discuss this more below.

[8]The number of household members in different demographic categories are also included as predictors.

Figure 2: Example decision tree




Notes: The figure shows an example decision tree that predicts (log) expenditures per capita in the NSS 61 data.

Figure 2 shows an example decision tree that predicts (log) expenditures per capita in the NSS 61 data. Starting at the top, the first split is based on whether the household has more than three individuals with a secondary education. For households that meet this criterion, the split continues to the right, while for households that do not meet the criterion, the split continues to the left. The decision tree then splits again based on the number of individuals in the household with a primary education (left) and the number of people with upper-secondary education (right). The number of splits is referred to as the "depth" of the tree, and this hyperparameter can in theory be arbitrarily large. In practice, it is usually capped to prevent overfitting and, in this example, we only allow for a depth of up to four for exposition.

A single decision tree is often not very accurate, so it is common to use an ensemble of trees to improve prediction accuracy. One way to do this is with

random forests, which build many individual trees and aggregate predictions by e.g. taking the mean prediction (Breiman, 2001; Athey et al., 2019). In practice, this is done by taking random bootstrap samples of the data and building decision trees on the separate samples. This can improve accuracy by reducing the variance of the predictions – i.e. by preventing overfitting.[9]

XGBoost is a more sophisticated example of random forests. XGBoost builds many trees sequentially, with each subsequent tree predicting the *error* from the previous tree, hence the name "gradient boosting." Gradient boosted trees simply proceed in sequence, with each tree predicting the residuals from the previous tree. This process continues until some stopping rule is met, such as a maximum number of trees or a minimum decrease in the loss function. XGBoost adds additional complexity by using regularization to prevent overfitting, similar to lasso, a more commonly used algorithm in economics (Tibshirani, 2018). Since subsequent models predict residuals, final predictions are created by adding together the predictions from each tree, rather than taking the mean.

While the algorithm uses the data to select splits for each tree, there are several parameters, referred to as "hyperparameters," that must be selected by the user. These include things like the maximum depth of the tree, which we discussed above, as well as other values that help prevent overfitting, like whether each tree should use all variables or only a random subset of variables. We discuss how we tune hyperparameters below.

## 3.3 Methods

We predict monthly level expenditures for each household in the NSS data, separately for each wave.[10] We then aggregate these predictions to calculate mean expenditures for the entire year, separately for each household.

A key issue with these predictions is inference. We are not interested in the predictions themselves; instead, we want to make inferences regarding 1) changes in inequality in India across the three waves of the NSS data and 2) the impact of

[9] In machine learning, fitting models to random subsets of the data is referred to as bootstrap aggregation, or "bagging."

[10] We do this separately by wave for two reasons. First, patterns may have changed across years. While XGBoost is flexible enough to capture these changes, we have large enough samples to estimate models separately. Second, predictive features are not the same in each wave of the data.

the NREGA program on inequality. To do this, we use a bootstrap approach. We randomly sample NSS blocks with replacement and estimate the XGBoost model to calculate predicted expenditures. We repeat this process 500 times, saving all the predictions for each household. We tune the hyperparameters once using the actual data to save time – tuning separately for each iteration leads to similar results but is computationally expensive. With all 500 predictions, we then estimate mean inequality by year as well as the effects of NREGA on inequality. We use the bootstrap distribution to calculate confidence intervals. We list the the optimal hyperparameters in appendix Table A1.

The Mahatma Ghandi National Rural Employment Guarantee Scheme (NREGA) was rolled out in India over three years, starting in 2005 (Imbert and Papp, 2015; Merfeld, 2020). The program guarantees 100 days of work per year to every rural household in India. The program was rolled out in phases, with the first phase starting in 2005, the second phase starting in 2006, and the third wave starting in 2007. We use the NSS data to estimate the impact of the program on inequality.

Importantly, we are interested in the effects of inequality at the *district* level, which is the level at which the program was rolled out. To calculate inequality, we calculate each of the three Theil measures, discussed above, collapsing household-level data to the district level. We do this separately for each of the 500 bootstrap predictions. We then estimate a differences-in-differences regression, comparing districts that have received NREGA to districts that have not, with the following regression:

$$Inequality_{dt} = \beta NREGA_{dt} + \sum_{t=2}^{3} I(wave == t) \times \phi_{d,t==1} + \gamma_{ts} + \delta_d + \epsilon_{dt}, \qquad (8)$$

where $Inequality_{dt}$ is the Theil measure for district $d$ in year $t$, $NREGA_{dt}$ is a dummy variable that equals one if the district has received NREGA, $\gamma_t$ are state-by-wave fixed effects, and $\delta_d$ are district fixed effects. Following Imbert and Papp (2015) and Merfeld (2020), we include pre-treatment values of key covariates that helped determine rollout: population, percent rural, percent SC, percent ST, literacy rates, labor force participant rates, casual labor rates, agricultural labor rates, and other work rates. This addition is given by $\sum_{t=2}^{3} I(wave == t) \times \phi_{d,t==1}$, where

$\phi_{d,t==1}$ is the value of the covariate in the first wave of the data for district $d$ and $I()$ is the indicator function. In other words, we allow the effect of pre-treatment characteristics to vary by wave (since the characteristics themselves are collinear with the district fixed effects).

We are interested in inequality which means, by definition, that we cannot estimate equation (8) at the household level. Instead, we must aggregate to a higher level to calculate inequality measures. Given that NREGA is rolled out at the district level and that the district has traditionally been used as a proxy for distinct labor markets (Kaur, 2019), we aggregate data to the district level. We calculate three types of inequality: total-inequality (left hand side of equation 7), between-inequality (first term of equation 7), and within-inequality (second term of equation 7). We calculate total-inequality using the raw NSS expenditures data. Between-inequality is calculated using household-level means, which we estimate using XGBoost and validate in the next sub-section. Within-inequality can be calculated in one of two ways: we can either subtract between-inequality from total-inequality or we can use predicted monthly-level expenditures of households. We show that household averages are much more accurately predicted than monthly-level expenditures, so we opt for the first option.

Due to the timing of the data collection, we code NREGS phase one districts as having received treatment in the second wave of the NSS data, while phases two and three receive treatment in the third wave.

## 3.4 Validation

We hypothesize that household-level mean expenditures will be much more accurate than monthly-level predictions. We validate this hypothesis using the ICRISAT data. There are 23 unique villages in the dataset and we validate the prediction exercise through leave-one-out cross validation. In other words, we take each village, remove it from the dataset, estimate XGBoost with the remaining villages, and predict expenditures for the village that was removed. We do this 23 separate times, calculating expenditures separately for each *wave* of the ICRISAT data for the held-out village. We believe this simulates the use-case with the NSS, where a primary sampling unit (village) is observed in one month of the year but not the others. We are interested in the accuracy of these pure out-of-sample predictions,

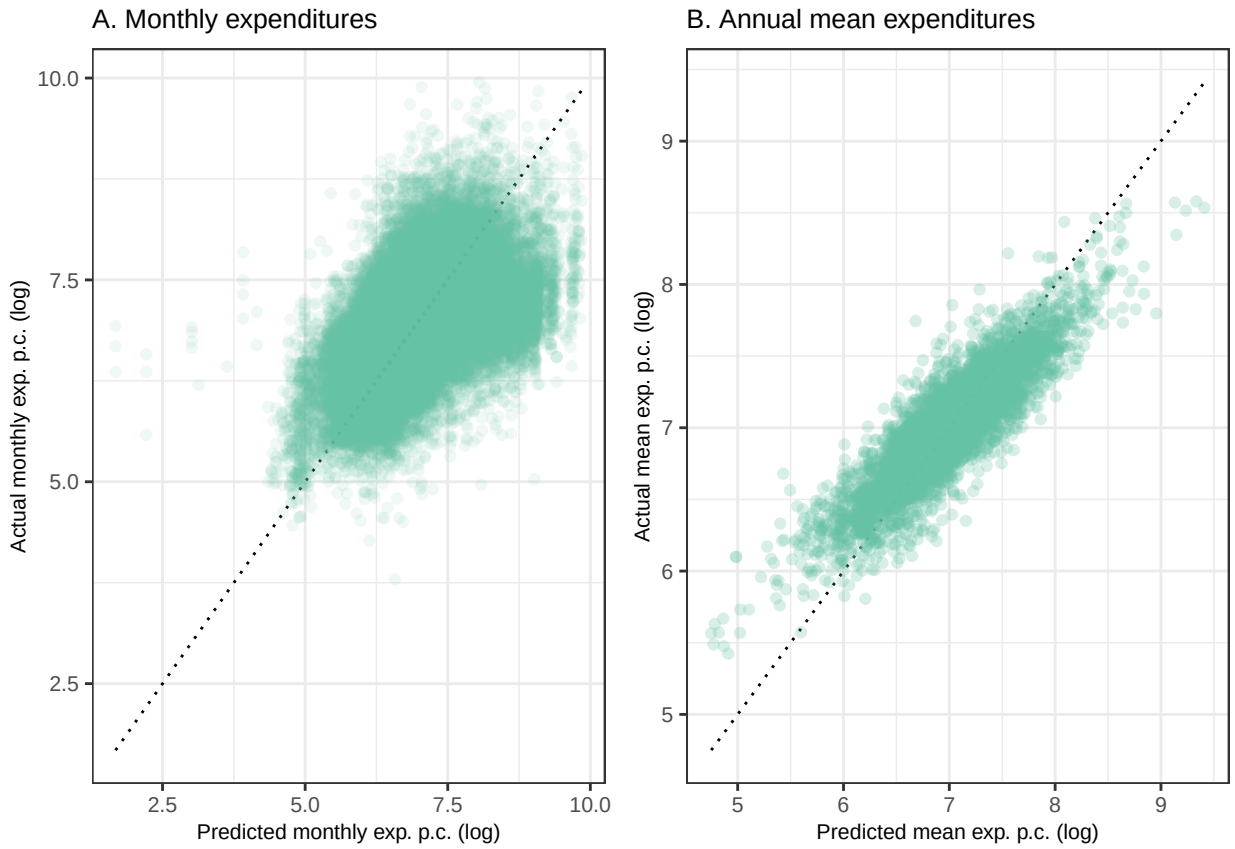



Figure 3: Monthly expenditure predictions

Note: The figure shows out-of-sample predictions of monthly expenditures per capita compared to actual monthly expenditures per capita.

which is what our cross validation exercise is meant to mimic.

Figure 3 shows the validation results from the ICRISAT data. Figure A presents a scatterplot of out-of-sample predictures for monthly expenditures per capita compared to actual expenditures. The overall correlation is 0.568. Figure B, on the other hand, shows the results for annual mean expenditures, calculated using one month of data for each village and imputing the other 11 months with the XGBoost algorithm. The correlation between the predicted and actual values is 0.857. As expected, the correlation is *much* higher for the annual mean than for an individual month.

While it is not possible to validate annual means in the NSS data—since we do not observe households more than once—we can nonetheless validate predictions for the month in which we see actual data on the household. While we want to use information on all households when predicting for the months we do not observe, we do not use data from household $i$ when predicting expenditures for household $i$ in the month in which we see that household; otherwise, this type of information leakage could bias the estimated correlations upward.

To validate the monthly predictions, we randomly hold out primary sampling units (PSUs) and estimate the XGBoost model on the remaining data. We then cross-validate by predicting expenditures for the held-out PSUs and calculating the correlation between the predicted and actual values. We randomly allocate PSUs across 20 folds, estimate the model 20 times—each time leaving one of the folds out of the estimation process—and calculate predicted expenditures for the held-out fold. We save all the results and calculate the overall correlation.

Figure 4 shows the out-of-sample predictions of monthly expenditures per capita compared to actual monthly expenditures per capita, separately by NSS wave and NREGS phase. The out-of-sample correlation, calculated by holding out a random 5 percent of primary sampling units—i.e. with 20 folds—is 0.869 (not shown in the figure). We note that this is for *current* consumption, and not annual consumption, and is much higher than the correlation from the ICRISAT data of 0.560.

We plot wave-specific correlations, separately by NREGA phase, to see if differences in correlation across waves and phases might be biasing the treatment effect, and estimate a difference-in-difference model. The estimated treatment effect on the NREGA dummy is just 0.005, indicating that the correlation between predicted and actual expenditures out of sample is more or less the same across waves and phases.

# 4 Inequality over time in India

In Table 1, we present Theil inequality measures for each of the three waves of the NSS, from 2004 to 2011, distinguishing between total-inequality (*de facto* measured inequality), between-household inequality (what economists generally consider to be inequality), and within-household inequality. In the notation of section 2, the

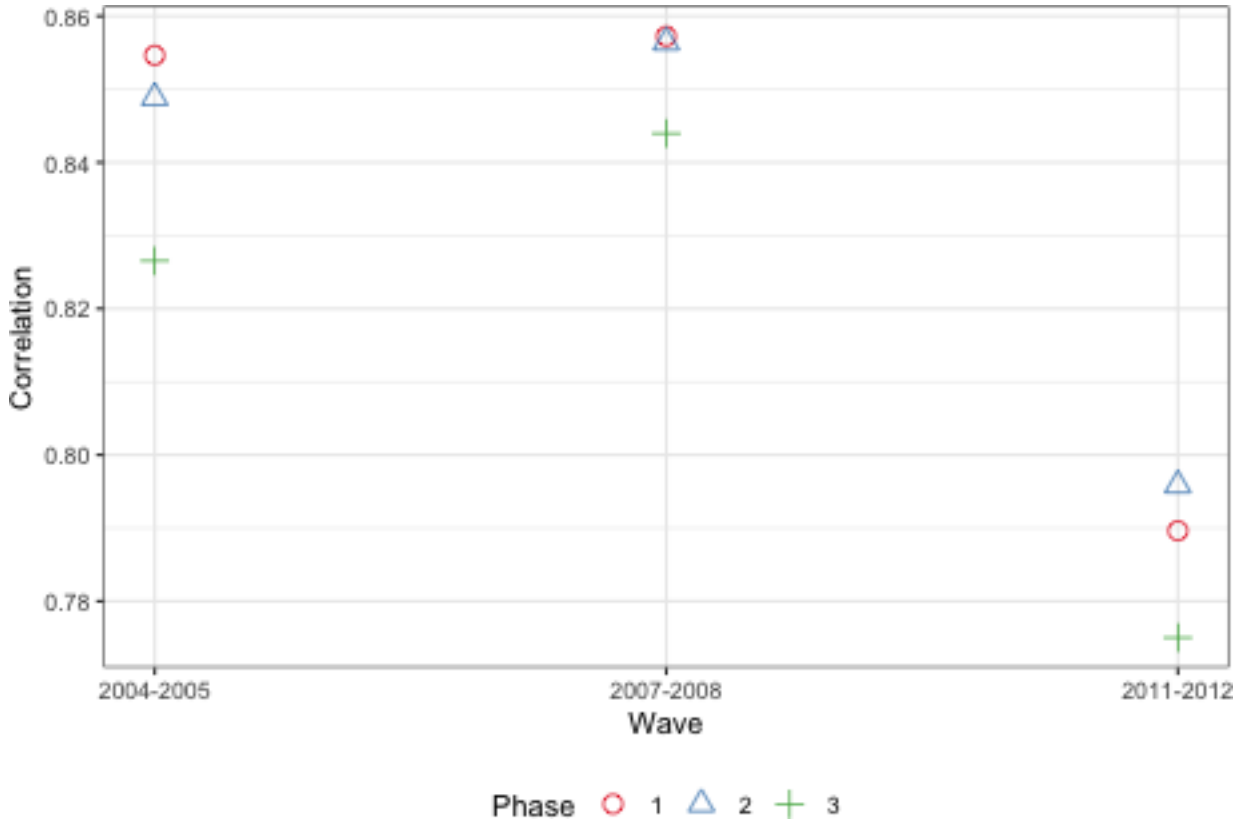



Figure 4: Out-of-sample predictions

Note: The figure shows out-of-sample predictions of monthly expenditures per capita compared to actual monthly expenditures per capita, separately for each NSS wave and NREGA phase.

measures correspond to $T_L^*$, $T_L$, and $V_L$. Confidence intervals are calculated using 500 bootstrap replications.

Table 1: Theil Index Decomposition by Wave

| Wave | Total | Between | Within |
|---|---|---|---|
| 61 (2004-2005) | 0.193 | 0.182 | 0.011 |
| | (0.188, 0.199) | (0.178, 0.187) | (0.010, 0.013) |
| | [0.189, 0.198] | [0.178, 0.187] | [0.010, 0.013] |
| 64 (2007-2008) | 0.186 | 0.161 | 0.024 |
| | (0.175, 0.203) | (0.156, 0.169) | (0.017, 0.038) |
| | [0.176, 0.199] | [0.156, 0.167] | [0.018, 0.036] |
| 68 (2011-2012) | 0.222 | 0.219 | 0.005 |
| | (0.214, 0.231) | (0.211, 0.228) | (0.004, 0.006) |
| | [0.216, 0.229] | [0.212, 0.227] | [0.004, 0.006] |

Note: The table shows overall inequality measures across India from three separate NSS waves. The table shows the inequality estimates, 95-percent CIs in parentheses, and 90-percent CIs in brackets, calculated through 500 bootstrap replications.

Total inequality ($T_L^*$) decreased slightly from 2004-2005 to 2007-2008, then increases in 2011-2012. However, these overall changes in inequality—which is all that statisticians typically observe—are driven differentially by changes in the two constituent indices across time. The AI approach reveals that in 2004-2005 between-inequality (i.e., the conventional notion of inequality) is around 94% of total-inequality. In 2007-2008, however, this drops to just 86.6% of total-inequality as within-inequality increases while between-inequality decreases. In the final wave, between-inequality comprises almost all of total-inequality. This is driven by an increase in between-inequality together with a large decrease in within-inequality, which is only 20% as large in 2011-2012 compared to 2007-2008.

Our main goal is not necessarily to show how much of total-inequality—as usually measured—is in fact between-household inequality. It is not surprising that between-household inequality is a much larger proportion of total-inequality than the within-household component; differences across households are typically much larger than than differences across time for a given household. Instead, our focus is on understanding how the components of inequality move across time and how they affect interpretations of economic change.

Table 2 uses 2001 census characteristics to examine changes in inequality from one wave to the next. We we see very different coefficients across time, even for the same variable. For example, more rural districts see larger increases in total and between inequality from wave 61 to wave 64, but then larger decreases in inequality from wave 64 to 68. However, despite between-household inequality being the majority of total, as-measured, inequality, some coefficients indicate that changes in these two constituent measures are often equally correlated with the variables included in the regressions. For example, rurality shows a relatively larger conditional correlation with changes in within-household inequality than between-household inequality.

We caution that many of these coefficients are imprecisely estimated, so we cannot draw firm conclusions. Nonetheless, the pattern of coefficients across regressions indicates that the breakdown of inequality into its constituent parts is informative and, if we are truly interested in between-household inequality, important.

Table 2: Correlates of Changes in Inequality

| | Total | | Between | | Within | |
|---|---|---|---|---|---|---|
| Waves: | 61 to 64 | 64 to 68 | 61 to 64 | 64 to 68 | 61 to 64 | 64 to 68 |
| Rural (prop.) | 0.306 | -0.200 | 0.125 | -0.080 | 0.181 | -0.119 |
| | (0.065, 0.671) | (-0.489, -0.015) | (0.029, 0.281) | (-0.238, 0.027) | (0.011, 0.492) | (-0.353, -0.000) |
| | [0.085, 0.611] | [-0.429, -0.031] | [0.043, 0.251] | [-0.201, 0.012] | [0.018, 0.440] | [-0.307, -0.007] |
| Literacy (prop.) | 0.115 | 0.182 | 0.087 | 0.182 | 0.027 | -0.006 |
| | (-0.010, 0.258) | (0.026, 0.320) | (0.003, 0.181) | (0.050, 0.311) | (-0.061, 0.114) | (-0.055, 0.032) |
| | [0.012, 0.227] | [0.058, 0.297] | [0.014, 0.163] | [0.075, 0.292] | [-0.031, 0.088] | [-0.046, 0.026] |
| LFP (prop.) | -0.113 | 0.094 | 0.006 | -0.024 | -0.120 | 0.114 |
| | (-0.481, 0.148) | (-0.149, 0.413) | (-0.206, 0.178) | (-0.223, 0.173) | (-0.385, 0.033) | (0.011, 0.304) |
| | [-0.406, 0.124] | [-0.120, 0.336] | [-0.157, 0.153] | [-0.195, 0.137] | [-0.343, 0.023] | [0.018, 0.259] |
| Ag Laborers | 0.033 | -0.296 | 0.000 | -0.269 | 0.033 | -0.028 |
| (prop.) | (-0.188, 0.252) | (-0.507, -0.087) | (-0.192, 0.189) | (-0.473, -0.066) | (-0.042, 0.117) | (-0.097, 0.025) |
| | [-0.155, 0.218] | [-0.474, -0.111] | [-0.171, 0.149] | [-0.430, -0.112] | [-0.030, 0.111] | [-0.083, 0.018] |
| Own-Account | 0.612 | -0.598 | 0.210 | -0.276 | 0.403 | -0.310 |
| Workers (prop.) | (-0.012, 1.493) | (-1.289, -0.056) | (-0.069, 0.626) | (-0.669, 0.099) | (0.004, 1.110) | (-0.849, -0.023) |
| | [0.059, 1.372] | [-1.153, -0.116] | [-0.038, 0.516] | [-0.619, 0.038] | [0.019, 1.046] | [-0.732, -0.047] |
| Pop. (log) | 0.052 | -0.025 | 0.020 | -0.009 | 0.032 | -0.017 |
| | (0.009, 0.118) | (-0.067, 0.007) | (0.002, 0.050) | (-0.033, 0.008) | (0.004, 0.084) | (-0.052, 0.002) |
| | [0.013, 0.104] | [-0.062, 0.003] | [0.003, 0.042] | [-0.029, 0.006] | [0.006, 0.077] | [-0.045, 0.000] |

Note: The table shows the results of a regression of the Theil index on various household characteristics. The table shows the coefficient estimates, 95-percent CIs in parentheses, and 90-percent CIs in brackets, calculated through 500 bootstrap replications.

# 5 An application to NREGA

We now turn to an application to India's large workfare program, the Mahatma Gandhi National Rural Employment Guarantee Act, NREGA (Imbert and Papp, 2015). The program aimed to provide a guarantee of at least 100 days of unskilled wage labor per year to at least one member of every Indian rural household. The program was thus both an employment scheme and provided resources to smooth consumption.

Table 3: NREGA Effects on Theil Index Components

| | Total | Between | Within |
|---|---|---|---|
| Point Estimate | 0.024 | 0.005 | 0.019 |
| 95% CI | (-0.021, 0.083) | (-0.023, 0.040) | (-0.007, 0.064) |
| 90% CI | [-0.014, 0.071] | [-0.020, 0.034] | [-0.005, 0.058] |

Note: The table shows the results of a regression of the Theil index on the NREGA treatment variable, controlling for several pre-treatment district-level characteristics. All regressions are at the district level. The table shows the coefficient estimates, 95-percent CIs in parentheses, and 90-percent CIs in brackets, calculated through 500 bootstrap replications.

As before, we estimate 500 separate regressions using the 500 bootstrapped predictions, using a simple two-way fixed effects framework. We present these results in Table 3. We note, again, that between-household inequality is the majority of as-measured total inequality. However, despite this, the point estimates indicate that the implementation of NREGA had a larger effect on within-household inequality than between-household inequality. Taking the point estimates at face value, as-measured total inequality increased by 0.024 after the implementation of NREGA. However, this is not driven by true, between-household, inequality; the coefficient on within-household inequality indicates that 80% of the increase is driven by within-household inequality, not between-household inequality. While these coefficients are imprecisely estimated, this nonetheless indicates that we can come to very different conclusions about effects of NREGA on inequality if we use the naive, total-inequality measure derived from the National Sample Survey instead of the between-household inequality measure isolated with the aid of the machine learning approach.

Why did NREGA lead to an increase in inequality? There are several possible explanations. First, the program faced implementation problems; for example, Narayanan et al. (2017) note that rationing and delayed payments were common. Insofar as the program was designed to supplement incomes and expenditures during lean times of the year, delayed wages could shift the actual receipt of income to higher-income times of the year, leading to an increase in within-inequality. Delayed wages were a key problem with NREGA; even in 2016-2017, after the implementation of electronic wage payments—rather than in-person wage payments that prevailed at the beginning of the program—Narayanan et al. (2019) show that only 21% of wages were paid on time, with the average delay being 51 days for the central government alone. This could again shift the actual receipt of payments to higher-income times of the year, increasing within-household inequality.

Second, turning to between-inequality, there are several suggestions of elite capture at the beginning of the program. For example, Rajasekhar et al. (2012) argue that, despite checks and balances in NREGA implementation, misuse of NREGA funds was nonetheless prevalent at the outset. Similarly, Mukherji (2019) shows that households more connected to local leaders are more likely to obtain a jobs card and work in the program and on average receive their wages more quickly.

# 6 Concluding thoughts

Statistical bureaus, like India's Ministry of Statistics and Program Implementation, serve fundamental roles in modern societies. They regularly collect data on a country's residents, doing the hard work of knocking on doors and collecting data by voice and computer. The nationally-representative data is part of the accountability of governments to their citizens, and the reliability of the data is basic to the mission.

By mathematically decomposing a popular inequality measure, we showed that, even when the work of statistical bureaus is completed faithfully following expert guidelines, it can still fail to yield measures of national inequality as understood by economists. One response made possible by AI is to replace face-to-face data collection and the construction of nationally-representative surveys based on interviews that can stretch for hours. In their place, AI promises other kinds of inequality

measures based on alternative kinds of data that may be naturally collected by mobile phone providers or via satellite imagery, for example (Blumenstock, 2016). These ideas hold great promise.

Yet, the conventional work of interviewing households across a country continues to hold significance. Our AI-based approach does not replace that work but aims to make it more useful. We have shown how an accessible algorithm, XGBoost, can be employed to extend work that many statistical bureaus lack the budget and capacity to complete. We have applied the approach with India's National Sample Survey and validated it with data from a panel of rural households. It performs relatively well, even if the set-up is not ideal. In particular, the rural data are not nationally-representative and were not collected for this purpose. In the future, statistical bureaus that embrace the AI approach could enhance their roles by collecting the same kind of data that they collect now, following expert guidelines like those in FAO and World Bank (2018), but adding to the core data by also collecting separate datasets with a longitudinal dimension to train and validate algorithms. This approach would have the added benefit of permitting estimation of inequality alongside within-year volatility, a concern of rising importance with climate change and weather-related instability (Intergovernmental Panel on Climate Change, 2021).

# 7 Acknowledgements

Ingrid Leiria and Harry Keehoon Jung provided excellent research assistance in cleaning the data from rural India. We thank the Mastercard Impact Fund, in collaboration with the Mastercard Center for Inclusive Growth, and the KDI School for initial research funding. We alone are responsible for all views and any errors.

# References


Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized Random Forests," *The Annals of Statistics*, 47, 1148–1178.

Augenblick, N., B. Jack, S. Kaur, F. Masiye, and N. Swanson (2024): "Retrieval Failures and Consumption Smoothing: A Field Experiment on Seasonal Hunger," Working paper.

Bardhan, P., T. Srinivasan, and A. S. Bali (2017): "Poverty and Inequality in India: An Overview," in *Poverty and Income Distribution in India*, ed. by A. Banerjee, P. Bardhan, R. Somanathan, and T. Srinivasan, New Delhi: Juggernaut, 567–603.

Blumenstock, J. (2016): "Fighting poverty with data," *Science*, 353, 753–754.

Bourguignon, F. (1979): "Decomposable Income Inequality Measures," *Econometrica*, 47, 901–920.

Breiman, L. (2001): "Random forests," *Machine Learning*, 45, 5–32.

Breza, E., S. Kaur, and Y. Shamdasani (2021): "Labor Rationing," *American Economic Review*, 111, 3184–3224.

Bryan, G., S. Chowdhury, and A. M. Mobarak (2014): "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh," *Econometrica*, 82, 1671–1758.

Casaburi, L. and J. Willis (2018): "Time versus State in Insurance: Experimental Evidence from Contract Farming in Kenya," *American Economic Review*, 108, 3778–3813.

Chen, T. and C. Guestrin (2016): "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cukier, K. (2025): "The Pulse of the Planet," *Finance  Development*, 19–23.

Deaton, A. and J. Dreze (2002): "Poverty and Inequality in India, A Re-Examination," *Economic and Political Weekly*, 3729–48.

Deaton, A. and M. Grosh (2000): "Consumption," in *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*, ed. by M. Grosh and P. Glewwe, Washington, DC: World Bank.

Deaton, A. and S. Zaidi (2002): "Guidelines for Constructing Consumption Aggregates For Welfare Analysis," Living standards measurement study working paper no. 135, World Bank.

Dercon, S. (2002): "Income Risk, Coping Strategies, and Safety Nets," *World Bank Research Observer*, 17, 141–166.

Devereux, S., R. Sabates-Wheeler, and R. Longhurst, eds. (2012): *Seasonality, Rural Livelihoods and Development*, London and New York: Earthscan/Routledge.

FAO and World Bank (2018): "Food Data Collection in Household Consumption and Expenditure Surveys: Guidelines for Low- and Middle-Income Countries," LSMS guidebook, World Bank and FAO, Rome.

Fink, G., B. K. Jack, and F. Masiye (2020): "Seasonal Liquidity, Rural Labor Markets, and Agricultural Production," *American Economic Review*, 110, 3351–92.

Foster, J. E. (1983): "An axiomatic characterization of the Theil measure of income inequality," *Journal of Economic Theory*, 31, 105–121.

Ghatak, M. (2022): "Introduction to e-Symposium: Estimation of poverty in India," *Ideas for India*, 10 October.

Gibson, J., J. Huang, and S. Rozelle (2003): "Improving Estimates of Inequality and Poverty From Urban China's Household Income and Expenditure Survey," *Review of Income and Wealth*, 49, 53–68.

Imbert, C. and J. Papp (2015): "Labor market effects of social programs: Evidence from India's employment guarantee," *American Economic Journal: Applied Economics*, 7, 233–263.

Intergovernmental Panel on Climate Change (2021): "Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (Summary for Policymakers)." Tech. rep., Cambridge, UK and New York.

Kaur, S. (2019): "Nominal Wage Rigidity in Village Labor Markets," *American Economic Review*, 109, 3585–3616.

Khandker, S. (2012): "Seasonality of income and poverty in Bangladesh," *Journal of Development Economics*, 97, 244–256.

Mancini, G. and G. Vecchi (2022): "On the Construction of a Consumption Aggregate for Inequality and Poverty Analysis," Report, World Bank.

Merfeld, J. and J. Morduch (2025): "Poverty at Higher Frequency," working paper, University of Queensland and New York University.

Merfeld, J. D. (2020): "Moving Up or Just Surviving? Nonfarm Self-Employment in India," *American Journal of Agricultural Economics*, 102, 32–53.

Milanovic, B. (2024): "The three eras of global inequality, 1820–2020 with the focus on the past thirty years," *World Development*, 177, 106516.

Mukherji, R. (2019): "Local leadership and public good: evidence from the National Rural Employment Guarantee Scheme in India," *Journal of Quantitative Economics*, 17, 311–329.

Narayanan, R., S. Dhorajiwala, and R. Golani (2019): "Analysis of payment delays and delay compensation in MGNREGA: Findings across ten states for financial year 2016–2017," *The Indian Journal of Labour Economics*, 62, 113–133.

Narayanan, S., U. Das, Y. Liu, and C. B. Barrett (2017): "The "discouraged worker effect" in public works programs: Evidence from the MGNREGA in India," *World Development*, 100, 31–44.

National Sample Survey Organisation (2001): "Concepts and Definitions Used in NSS. (Golden Jubilee Publication)," Tech. rep., Ministry of Statistics and Programme Implementation, Government of India, Delhi.

Pikkety, T. (2014): *Capital in the 21st Century*, Cambridge, MA: Belknap Press of Harvard University Press.

Pomeranz, D. and F. Kast (2024): "Savings Accounts to Borrow Less: Experimental Evidence from Chile," *Journal of Human Resources*, 59, 70–81.

Rajasekhar, D., M. D. Babu, and R. Manjula (2012): "Are Checks and Balances in MGNREGS Effective? EDCBA," *Social Sciences*, 73.

Ravallion, M. (2014): "Income inequality in the developing world," *Science*, 344, 851–855.

Scott, C. (1992): "Estimation of Annual Expenditure from One-month Cross-sectional Data in a Household Survey," *Inter-Stat*, 8, 57–65.

Sinha Roy, S. and R. Van Der Weide (2022): "Poverty in India Has Declined Over the Last Decade But Not as Much as Previously Thought1," *Available at SSRN 4427524*.

Theil, H. (1967): *Economics and Information Theory*, Chicago: Rand McNally and Company.

Tibshirani, R. (2018): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

# Appendix A

Table A1: Optimal XGBoost hyperparameters

| Hyperparameter | Value |
|---|---|
| Number of trees | 150 |
| Maximum depth | 4 |
| Learning rate | 0.3 |
| Minimum child weight | 1 |
| Subsample | 0.7 |
| Column subsample | 0.7 |
| Gamma | 0 |

Note: The table shows the optimal hyperparameters, selected through cross-validation.