Nowcasting Global Poverty

Daniel Gerszon Mahler
R. Andrés Castañeda Aguilar
David Newhouse
Abstract

This paper evaluates different methods for nowcasting country-level poverty rates, including methods that apply statistical learning to large-scale country-level data obtained from the World Development Indicators and Google Earth Engine. The methods are evaluated by withholding measured poverty rates and determining how accurately the methods predict the held-out data. A simple approach that scales the last observed welfare distribution by a fraction of real GDP per capita growth—a method that departs slightly from current World Bank practice—performs nearly as well as models using statistical learning on 1,000+ variables. This GDP-based approach outperforms all models that predict poverty rates directly, even when the last survey is up to five years old. The results indicate that in this context, the additional complexity introduced by applying statistical learning techniques to a large set of variables yields only marginal improvements in accuracy.

This paper is a product of the Development Data Group, Development Economics and the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dmahler@worldbank.org, acastanedaa@worldbank.org, and dnewhouse@worldbank.org.
Nowcasting Global Poverty*

Daniel Gerszon Mahler
R. Andrés Castañeda Aguilar
David Newhouse

JEL codes: C53, D31, I32, O10.

Keywords: Poverty, Nowcasting, Machine Learning, Measurement.

*All authors are with the World Bank. We are grateful for comments received from Aart Kraay, Andres Fernando Chamorro Elizondo, Benjamin Stewart, Benu Bidani, Christoph Lakner, Dean Jolliffe, Lucas Kitzmueller, Marta Schoch, Minh Cong Nguyen, Nishant Yonzan, Nobuo Yoshida, and Samuel Kofi Tetteh Baah. We are also grateful for feedback received during the special IARIW-World Bank Conference ‘New Approaches to Defining and Measuring Poverty in a Growing World’, the CCS-UN Technical Workshop ‘Nowcasting in International Organizations’, and the 2021 ECINEQ Conference. We gratefully acknowledge financial support from the UK government through the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Programme and from the World Bank through a Research Support Budget grant. The code to reproduce the findings of this paper is available at https://github.com/danielmahler/NowcastingGlobalPoverty.
1 Introduction

Timely and comparable poverty estimates are vital to assess countries’ development progress. International poverty estimates serve as a public good for researchers and inform the development community on efforts to meet the first Sustainable Development Goal, to end extreme poverty by 2030. Within international development organizations, national development agencies, and NGOs, they also inform the allocation of resources and the development of strategic priorities.

Yet timely and comparable estimates of poverty are lacking for many reasons. In some countries, fragility, conflict, and violence make it difficult to conduct household expenditure surveys altogether, while in other countries, lack of monetary resources is the main obstacle. Even when surveys are frequently conducted, the time it takes to field a survey, collect, process, and analyze the data, often implies a two-year lag from the time of data collection to the release of international poverty estimates. With the world changing at an ever more rapid pace, as illustrated by the unexpected onset of COVID-19, this lag risks painting an outdated picture of poverty in a country. As of October 2021, on average across the developing world, the most recent survey with international poverty data was from 2014. In addition, 16 economies with a population greater than 1 million had no international poverty estimate at all. For these reasons, initiatives that reliably and cost-effectively predict what the poverty rate is today are crucial for informed and effective high-level decision-making.

The objective of this paper is to test various methods to estimate extreme poverty in all countries of the world as of the present year and as of the preceding year – at the time of writing, 2021 and 2020. We will refer to estimates of poverty for the present year as nowcasts and estimates of poverty for the preceding year as nearcasts. For nearcasting, one can rely on all data that are produced with less than a one-year time lag. We will use more than 1,000 variables from the World Development Indicators, the World Economic Outlook, and the Google Earth Engine. For nowcasting, only variables that themselves have been nowcasted by others are available, as well as variables that are produced with little or no time lag, such as certain remote sensing indicators.

We combine all of these possible predictors with the PovcalNet database, which contains more than 2,000 international poverty estimates covering 168 countries. We train our model on these past surveys – essentially pretending a subset of them do not exist – and evaluate the models by measuring how well they approximated the true poverty rates of these held-out surveys. The best performing
models are then leveraged to predict global poverty for the nearcasting and now-casting years.

Intuitively, to predict extreme poverty around the world, one would want the models to predict poverty rates directly. Yet this ignores that prior full distributions of consumption or income (henceforth welfare) are available for most countries. We will explore whether greater accuracy can be obtained by predicting poverty indirectly, by, for example, predicting changes in poverty from the last survey, by predicting growth in mean welfare and applying this growth to scale up the past distribution, or by predicting growth in the mean and in the Gini and applying growth incidence curves to scale and stretch the past distributions.

For nearcasting, we find that models that predict poverty rates directly are outperformed by models that predict growth in mean welfare since the last survey and scales the last distribution by this predicted growth. Though this method assumes that inequality remains unchanged since the last survey, explicitly modeling distributional changes by predicting changes in the Gini coefficient does not help. The models that predict changes in inequality next to growth in mean welfare perform slightly worse than models that assume no changes to inequality. The reason for this is that of the 1,000+ candidate variables, none of them contains notable information about changes in inequality. The best performing method overall, which predicts growth in mean welfare using a random forest, gives a mean absolute deviation of 3.65 percentage points. This means that on average over all countries for which we have data, the predicted poverty evaluated at the international poverty line of $1.90 is 3.65 percentage points from the truth.

The discussion above pertained to models using variables that are not always available in the nowcasting year. Yet, we find that a model which simply uses a fraction of growth in real GDP per capita to scale up the last mean – a model which can be used for nowcasting as well – gives a mean absolute deviation of 3.69, about 1 percent worse than the overall best performing model for nearcasting. In other words, conditional on knowing growth in real GDP per capita since the last observed distribution, no other variable contributes significant information about the evolution of poverty rates. Due to its simplicity and ability to be used for both nearcasting and nowcasting, we consider this the preferred method. We show that even when the last survey is as much as five years old, extrapolating forward using a fraction of growth in real GDP per capita is superior to predicting poverty rates directly with 1,000+ variables. For longer extrapolation times, we lack power to determine which method is superior.

On the one hand, the relevance of GDP growth for nowcasting poverty is not
surprising; the impact of growth on poverty reduction has been well-known for decades (Kraay, 2006; Ferreira and Ravallion, 2009). On the other hand, ample evidence has found large inconsistencies between consumption measured in household surveys and national accounts within and across countries (Ravallion, 2003; Deaton, 2005; Ferreira et al., 2010; Pinkovskiy and Sala-i Martin, 2016; Prydz et al., 2021; Deaton and Schreyer, forthcoming) and noted the difficulty of measuring GDP in developing countries (Angrist et al., 2022).

Several factors can explain why we nonetheless find GDP to be such an important predictor of poverty. First, discrepancies between levels of GDP and levels of welfare do not directly affect the accuracy of the method, which is based on growth rates from the two data sources. Second, the average discrepancy between growth rates in the two sources is accounted for by only allowing a fraction of growth in real GDP per capita to ‘pass through’ to welfare as measured in household surveys. The preferred method has a passthrough rate of about 0.7 for consumption-based poverty estimates, and a passthrough rate of 1 for income-based poverty estimates. Third, GDP is probably the statistic in which the global community has invested the most amount of financing and capacity-building to improve quality and ensure cross-country comparability (see Angrist et al. (2022) for a discussion of this). Despite measurement issues and shortcomings, we would expect GDP to have more signaling properties for cross-country measures of average well-being than any other statistic. Fourth, although using growth in real GDP per capita is the best performing method, no method is able to explain more than a quarter of the variation of growth in mean welfare. Hence, a large part of the variation in poverty rates no model is able to predict, and may reflect a substantial amount of random noise in measured welfare, resulting from different definitions of welfare.

To our knowledge, this is the first paper comparing different ways of nowcasting poverty at a global scale. Other papers, such as Chi et al. (2021), Cuaresma et al. (2018), and Moses et al. (2021) have predicted global or near-global poverty rates but not with the purpose of testing different methods. For other indicators, nowcasting is a more established exercise, such as nowcasts of GDP (Giannone et al., 2008), inflation (Aruoba and Diebold, 2010), and macroeconomic variables more broadly (Giannone et al., 2012). Partially due to the SDG target of ending extreme poverty by 2030, many papers have focused on forecasting global poverty rather than nowcasting it (Edward and Sumner, 2014; Hillebrand, 2008; Ravallion, 2013; Lakner et al., 2020; Sumner and Hoy, 2022).

Our findings contribute to the literature by pointing out that when predicting
poverty at the national level, (1) predicting changes from the past distribution is superior to predicting poverty rates directly, and (2) conditional on knowing real GDP per capita, publicly available remote sensing data carry little additional information.

COVID-19 illustrates both the promises and pitfalls of nowcasts. On the one hand, the pandemic caused an unprecedented temporary stop in the production of household surveys leading to an important gap in timely data and an important role for methods that circumvent the availability of primary data sources. On the other hand, the nature of the shock casts doubts on whether models trained on historical data work in a radically changed environment. Partially due to this uncertainty, the country-level nowcasts we produce should not be understood as a substitute for country-level nowcast exercises. When the objective is to nowcast poverty for a single country, often relying on other methods, such as microsimulation tools, or more granular data will be superior as they can be tailored to country-specific contexts. Yet as those methods are hard to implement consistently across many countries, the present exercise can be attractive when the objective is to compare poverty across a range of countries.

2 Method

We will employ several methods to predict poverty around the world. In this section we outline the methods we use, their advantages and disadvantages, and other important methodological choices. We will focus on three distinct issues: (1) The target variable being predicted and (if not poverty rates directly) how poverty rates are obtained from these predictions, (2) the algorithms used to generate the predictions, and (3) the evaluation criterion used to judge predictive performance. Throughout this paper, when we refer to predicting poverty rates, we are using the international poverty line of $1.90 per day (Ferreira et al., 2016).

2.1 The target variable

When deciding how to predict poverty, it seems intuitive that the target variable to predict should be the poverty rate in each country. Yet, such a method ignores that we have prior poverty estimates from most countries which we could predict changes from. Behind these prior estimates are a full distribution of welfare which might contain relevant information – for example by distinguishing the near poor and the rich. Utilizing some of this past information, we will work with
six different target variable combinations as outlined in Table 1 and explained in more detail in what follows.

<table>
<thead>
<tr>
<th>Target variable</th>
<th>Approach for estimating poverty</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Poverty rates</td>
<td>Predicted directly</td>
</tr>
<tr>
<td>(2) Changes in poverty rates</td>
<td>Apply the predicted change to the most recent poverty rate</td>
</tr>
<tr>
<td>(3) Mean welfare</td>
<td>Scale the past distribution to the predicted mean welfare</td>
</tr>
<tr>
<td>(4) Growth in mean welfare</td>
<td>Scale the past distribution by $1 + \hat{c}_{\text{poverty},t_s,t_n}$</td>
</tr>
<tr>
<td>(5) Mean welfare and Gini coefficient</td>
<td>(i) Assume the distribution is log-normal or log-logistic or (ii) apply a growth incidence curve from the past distribution to match the predicted mean and Gini</td>
</tr>
<tr>
<td>(6) Growth in mean welfare and Gini coefficient</td>
<td>(i) Apply the predicted growth in the mean and Gini to the past mean and Gini and assume that the distribution is log-normal or log-logistic or (ii) apply a growth incidence curve to the past distribution with the predicted growth in the mean and Gini</td>
</tr>
</tbody>
</table>

Notes: The table lists the six different combinations of target variables we will use for the predictions and how poverty rates are backed out based on the target variable(s).

(1) Poverty rates and (2) changes in poverty rates. Predicting the poverty rate directly is the most intuitive and straightforward option; the poverty rates are the ultimate objective of this paper. Predicting poverty rates directly at the nowcasting year, $t_n$, has the advantage that it can yield estimates for countries without any previous poverty estimates at all. This is particularly a concern for countries that do not produce or share household survey data. Predicting changes in poverty from the past survey conducted at time $t_s$, in contrast, needs some assumptions about poverty levels in countries without data to arrive at global poverty rates. Yet, by utilizing the past survey, one can exploit that there is a past estimate to anchor the analysis around. When we predict changes in poverty, $\hat{c}_{\text{poverty},t_s,t_n}$, the nowcasted poverty rates are given by

$$poverty_{t_n}(\hat{c}_{\text{poverty},t_s,t_n}) = poverty_{t_s} + \hat{c}_{\text{poverty},t_s,t_n}$$ (1)

When predicting changes from the past survey, we look at the annualized change in poverty rates in percentage points. This avoids having to predict extreme and undefined values, which often occur when predicting annualized growth in poverty rates, due to countries with poverty rates close to or at zero percent.

(3) Mean welfare and (4) growth in mean welfare. While predicting changes in poverty rates from the past survey exploits some past information available, it
still ignores the fact that a whole distribution of welfare was available in the past. Distributions with high density around the poverty line are likely to experience different magnitudes of changes in poverty than countries with sparse density around the poverty line. By predicting the mean or growth in the mean and scaling the past distribution accordingly, the model takes full advantage of the previous data in the sense that the entire distribution is leveraged. Method (3) works by scaling the welfare of each household, $h$, at the last survey by the ratio of the predicted mean at the nowcasting year, $\hat{\mu}_{tn}$, and the observed mean at the last survey $\mu_t$.

The adjusted distribution is used to estimate poverty at time $t_n$:

$$
poverty_{t_n}(\hat{\mu}_{tn}) = \mathbb{F}(\text{welfare}_{h,t_n} \frac{\hat{\mu}_{tn}}{\mu_t} < 1.9)
$$

Similarly, method (4) works by taking the last observed distribution of welfare and scaling it by the growth in the mean predicted between $t_s$ and $t_n$, $\hat{g}_{\mu,t_n,t_s}$.

$$
poverty_{t_n}(\hat{g}_{\mu,t_s,t_n}) = \mathbb{F}(\text{welfare}_{h,t_n}(1 + \hat{g}_{\mu,t_s,t_n}) < 1.9)
$$

Figure 1 shows a hypothetical example of how this works. In this hypothetical example, we take the latest observed distribution for Botswana from 2016 and show what would happen if we predicted annualized growth in the survey mean of 4% between 2016 and 2021. By shifting the distribution to the right reflecting five years of this growth rate, the poverty rate at $1.90$ declines from about 15% to 9%. As the figure makes clear, predicting growth in the mean has the advantage that the model can be applied to any poverty line. Yet it imposes the assumption that all households have experienced the same growth since the last survey. In other words, it imposes that inequality has not changed since the last survey.

(5) Mean consumption and Gini coefficient and (6) Growth in mean and growth in Gini. The fifth and sixth methods try to deal with the latter issue by also predicting inequality – either directly or by predicting growth in inequality since the last survey. A challenge with this approach is that there are many different measures of inequality and infinitely many ways in which the same level/growth of inequality can materialize. We will use the Gini coefficient as the measure of inequality due its popularity and use some further distributional assumptions to pin down how the Gini shapes the distribution. In particular, we will use two ways of converting Gini predictions into poverty rates.

First, we will use the predicted mean and predicted Gini together with a known two-parameter distributional shape that welfare can follow. We will use the log-normal distribution, which is frequently used for poverty and inequality
Figure 1: Example of recovery of poverty rates from predictions of mean welfare

Notes: The figure shows how we implement models where the target variable is mean welfare or growth in mean welfare. In this particular figure, we use the last observed distribution from Botswana from 2016 and show how that distribution is projected forward by a hypothetical prediction of a growth in mean welfare of 4% per year.

analysis (see for example Bourguignon (2003)) and the log-logistic distribution (also known as the Fisk distribution, after Fisk (1961)). We use either of these distributions to back out the poverty rate given a predicted mean, $\hat{\mu}_{tn}$ and predicted Gini, $\hat{gini}_{tn}$. Formally, we back out poverty with the log-normal distribution as

$$poverty_{tn,lognormal}(\hat{\mu}_{tn}, \hat{gini}_{tn}) = \Phi\left(\frac{\ln(1.9) - \ln(\hat{\mu}_{tn}) + 2[erf^{-1}(\hat{gini}_{tn})]^2}{2erf^{-1}(\hat{gini}_{tn})}\right), \quad (4)$$

and with the log-logistic distribution as:

$$poverty_{tn,loglogistic}(\hat{\mu}_{tn}, \hat{gini}_{tn}) = \left[1 + \left(\frac{\hat{\mu}_{tn} \sin(\pi \hat{gini}_{tn})}{1.9 \pi \hat{gini}_{tn}}\right) \frac{1}{\hat{gini}_{tn}}\right]^{-1} \quad (5)$$

Figure 2 shows the resulting poverty rates we predict as a function of $\hat{\mu}_{tn}$ and $\hat{gini}_{tn}$.

The other method applies a specific growth incidence curve (GIC) from the last observed distribution. GICs plot the growth in welfare as a function of the percentile, $p$, of the initial welfare distribution (Ravallion and Chen, 2003). Downwards-sloping GICs reduce inequality and vice versa. Evidence shows that
GICs often take on approximately linear and convex forms (Lakner et al., 2020; Kakwani, 1993; Ferreira and Leite, 2003). By imposing particular functional forms on the GICs, given a predicted mean welfare and predicted Gini, there is only one possible GIC. When using GICs, the nowcasted poverty rates are backed out as follows:

\[
\text{poverty}_{t_n,GIC}(\hat{\mu}_{t_n}, \hat{\text{gini}}_{t_n}) = F[\text{welfare}_{h,p,t_n} (1 + g_{p,t_n}(\hat{\mu}_{t_n}, \hat{\text{gini}}_{t_n})) < 1.9],
\]

where \(g_{p,t_n}(\hat{\mu}_{t_n}, \hat{\text{gini}}_{t_n})\) are percentile-specific growth rates, which are determined such that the resulting distribution matches the predicted mean and Gini. In addition, for the linear GIC, there is a requirement that \(g_{p,t_n}(\hat{\mu}_{t_n}, \hat{\text{gini}}_{t_n}) = \beta + \delta p\), while for the convex GIC, there is a condition that \(g_{p,t_n}(\hat{\mu}_{t_n}, \hat{\text{gini}}_{t_n}) = (1 - \alpha)(1 + \gamma) - 1 + [\alpha(1 + \gamma)\mu_{t_n}] / \mu_{p,t_n}\). Here \(\beta\) and \(\delta\) or \(\alpha\) and \(\gamma\) are parameters that are estimated to ensure that all the above equations hold (Lakner et al., 2020).

Figure 3 shows two sample GICs – again using the last survey from Botswana in 2016 as the starting point – and how they impact the shape of the nowcasted distribution. Both assume that the nowcasted mean has grown by 4% annually since the last survey and that the nowcasted Gini is 5% lower than the last observed Gini but distribute this growth differently along the distribution.

The above equations pertain to method (5). Analogous equations can be made with method (6) where the target variables are the growth in the mean and Gini.
Notes: The figures show examples of a linear and convex GIC, here applied to the last survey in Botswana from 2016.

Throughout, we predict growth in the Gini rather than changes in the Gini based on preliminary analysis comparing the performance of those two options.

Throughout the analysis we will pay particular attention to a submethod under the fourth category, which is a variant of what the World Bank uses to extrapolate poverty in countries and report on Sustainable Development Goal indicator 1.1.1 – to end extreme poverty by 2030. The method is based on the premise that there is a tight relationship between income or expenditure as measured in national accounts and income or consumption observed in household surveys. This method works by taking the last observed distribution of welfare and scaling the welfare of each household by the growth observed in GDP per capita from national accounts between the survey and the nowcasting year.\(^1\) This model assumes that growth observed in national accounts is fully ‘passed through’ to the welfare observed in household surveys and that the only factor informative for changes in poverty is growth in national accounts. In addition, like all methods under (3) and (4), it assumes that growth accrues to everyone equally, that is, without changing the distribution of welfare. This is problematic if growth was pro-poor or pro-rich in the intervening period.

\(^1\)The World Bank uses Household Final Consumption Expenditure (HFCE) whenever available and GDP otherwise, with the exception of countries in Sub-Saharan Africa where only GDP is used (Prydz et al., 2019). We focus on GDP here since HFCE nowcasts are not available for many countries.
The methods we cover are obviously not comprehensive, and other ways of arriving at predictions of poverty rates exist. We hope that the methods we have chosen cover both a mix of the most intuitive options and the ones that have been applied in prior work.

2.2 Algorithms

In order to predict any of the options presented in Table 1, we rely on a number of frequently used machine learning algorithms. In particular, we will use the lasso (Tibshirani, 1996), the post-lasso (Belloni et al., 2013), CART random forests (Breiman, 2001), conditional inference random forests (Hothorn et al., 2006) and gradient boosting (Friedman, 2001). These methods all have in common that they can predict the outcome variable of interest while being agnostic about which variables are relevant for the predictions.

Since the variables we use suffer from a lot of missing data, it is necessary to find a strategy to deal with this missingness. Simply deleting rows with missing values is not feasible as this would leave no or very few observations left. For the conditional inference random forests and gradient boosting, we rely on imputation methods embedded in the algorithms to deal with missing data. For conditional inference random forests, for example, the algorithm works by sequentially splitting the sample into two based on a variable deemed most predictive of the target variable. The algorithm might judge that countries with a decline in the share of male workers in agriculture is predictive of growth in the mean. For country-years without data on the share of male workers in agriculture, the algorithm will search for the most similar variable in terms of how it relates to the target variable, which in this case could be the share of all workers in agriculture, and split the observations with missing values in the first variable based on the latter.

Such methods of dealing with missing values are not possible or not available in the programs we use for the lasso, post-lasso, and CART random forests, where we will instead multiply impute the entire data set of features and thereby avoid missing values altogether (Rubin, 1976, 2004; Schafer, 1997). For each imputation we calculate a predicted value and average over all of these to obtain a final estimate. Although this multiple imputation could also be used for the other methods, it adds quite a bit of computing time, so we will only use it where necessary.
2.3 Evaluation of performance

In order to tune the algorithms listed above, compare the different approaches for nowcasting poverty presented in Table 1, and report the final out-of-sample errors, we rely on nested 5-fold cross validation (Stone, 1974; Iizuka et al., 2003; Varma and Simon, 2006). Intuitively, nested cross-validation works by iteratively splitting the sample into three different subsamples; a training subsample on which variants of a particular machine learning method are estimated, a testing subsample on which these variants are compared against each other and the best performer is selected, and a validation subsample which for each machine learning method is used to calculate the out-of-sample errors. The data points going into these three subsamples are reassigned multiple times to maximize the power of the data. Nested cross-validation is an extension of regular k-fold cross-validation, which reduces a bias with the latter towards selecting a model with many tuning options and a downward bias in the final out-of-sample errors obtained. Appendix A contains a more thorough and technical discussion of nested cross-validation.

Once we are done with nested 5-fold cross validation, we will have one out-of-sample estimate of the poverty rate for each household survey (except for the earliest one for each country, as the methods that rely on changes since the last data point will not work in those cases), for each machine learning method, and for each approach to estimating poverty from Table 1. To evaluate the performance of the various methods, our primary loss function will be the mean absolute deviation between the predicted and true poverty rate. To assure that we do not select models that only work well for a few countries that happen to have many poverty estimates, we weigh each country by the inverse of number of surveys, such that the total weight for each country equals one.

We use the mean absolute deviation rather than the mean squared error since we are interested in minimizing the deviations between the true and predicted poverty rates while giving equal weight to all deviations. Using the mean squared error tends to give more weight to the prediction of outliers. We think this can be problematic since data incomparabilities can create some strong outliers at times, and we do not want to judge the methods by how well they predict these outliers. We focus on percentage point deviations rather than percentage deviations, as the latter sometimes can be very large for countries with low poverty rates. If a country has a poverty rate of 1% but our model predicts a poverty rate of 2% the error in percentage terms would be 100% which would give this observation a large impact. Our focus on percentage point deviations thus implicitly gives a
larger focus on countries with high poverty rates.

Once we have selected the model that minimizes the mean absolute deviation following the details above, we will apply the selected model to predict poverty in the nearcasting and nowcasting years, which for the present paper are 2020 and 2021.

3 Data

All poverty estimates used in this paper come from PovcalNet, which contains the World Bank’s official country-level, regional, and global estimates of poverty. Most of the data in PovcalNet comes from the Global Monitoring Database, which is the World Bank’s repository of harmonized multitopic income and expenditure household surveys used to monitor global poverty. PovcalNet contains more than 2,000 surveys from 168 countries covering 98% of the world’s population. The data available in PovcalNet are standardized as far as possible but differences exist with regards to the method of data collection, and whether the welfare aggregate is based on income or consumption. By relying on the PovcalNet database, we ensure consistency with the official numbers used by the World Bank and United Nations for monitoring poverty, inequality, and related goals.

To predict poverty, we rely on variables from three databases. First, the World Bank’s World Development Indicators, which contain country-year information on nearly 1,300 variables covering a wide range of topics, such as health, agriculture, education, climate change, infrastructure and more (but as we will explain below, only a small selection of these can be used for the present exercise). Second, we rely on the IMF’s World Economic Outlook (WEO) database, which contains country-year information for about 50 variables related to macroeconomic outcomes, such as inflation, government debt, and unemployment. Third, we rely on remote sensing data from the Google Earth Engine, particularly, data on nighttime lights, rainfall, land surface temperature, impervious surface, crop-land, and normalized difference vegetation, snow, and water indices. In contrast to WEO and WDI, the remote sensing data are both more granular spatially and more frequent temporally. To fit into our framework, they need to be aggregated to the country-year level. We first aggregate them to annual data by calculating the mean, max, min, and standard deviation of each location over a year. Afterwards, we aggregate spatially by taking the mean, max, min, and standard deviation of the annual data for a country. This gives 16 features for each type of variable. Some of these combinations will not be relevant for global poverty
nowcasting but we include all of them here – as we do with WDI and WEO – to remain agnostic about which ones are relevant and which ones are not.

For all of the variables from the three sources above we also calculate annualized growth rates of the variables between two household surveys for a country. For all variables that are expressed as percentages, rates, or indices, we also calculate the annualized change between two household surveys for a country.

The only variables we remove from what is described above are (1) variables with more than 90% missing information in 2020, the nearcasting year, (2) variables with more than 90% missing information for country-years with poverty estimates, and (3) variables that are not comparable between countries, such as variables expressed in local currency units. The first two criteria help speed up the models by removing variables with too little information to be relevant. The latter removes variables where exploiting cross-country variation is not meaningful.

Depending on which month of the year this exercise is carried out, the removal of variables with more than 90% in the nearcasting year removes a large fraction of the WDI. The reason is that early in the year, most variables do not yet have information for the prior year, meaning the information used for nearcasting is not much better than the information that can be used for nowcasting. When conducted in July (as the current exercise is), only about 300 WDI variables meet that criterion. All in all, we are left with a bit more than 1,000 features across the various data sources.

Not all poverty trends within countries are comparable over time due to changes in the survey methodology or the welfare aggregate. This matters for our predictions of changes/growth in a target variable. Even if we knew the exact causes of poverty in a country, if welfare aggregates are not comparable, then we would not be able to predict the changes in poverty between two surveys. Though this also makes it more difficult to predict levels in our target variables, the problem is arguably larger when predicting changes. Though we can restrict our sample to comparable spells only, our main results will cover all data points, even the ones that are not comparable over time. As one of our findings will be that nowcasting poverty by predicting changes from a past distribution gives more accurate results than predicting levels directly, the decision to include non-comparable spells makes this result more conversative, as it precisely penalizes the methods we find superior. As a robustness check we restrict our data to comparable spells only.
4 Results

4.1 Evaluation of the performance of the models

Figure 4 shows the prediction errors across the six combinations of target variables from Table 1 by each of the five machine learning algorithms we use (regular lasso, post-lasso, CARF random forest, conditional inference random forest, and gradient boosting). Note that regardless of which target variable combination we used (i.e. which panel of Figure 4), all the predictions are turned into poverty rates and the errors are evaluated over poverty rates. Hence, they are comparable to each other. As one example, the very first bar shows that when predicting poverty rates directly using a CART random forest (abbreviated as \textit{carf}), then we on average get a predicted poverty rate 4.99 percentage points off the truth.

Note that each bar in panels 5 and 6 of Figure 4 have several possible realizations depending on which growth incidence curve or distributional assumption is used to convert means and Ginis into poverty rates. In Figure B.1 in the Appendix we show the best way of converting predictions of growth/levels in the mean and Gini into poverty rates. The best way of converting predicted levels of the Gini and mean tends to be by using a log-normal distribution, while the best way of converting predictions of growth in the mean and Gini tends to be by applying a linear growth incidence curve. Figure 4 shows the best performing options from Figure B.1 in panels 5 and 6.

Figure 4 reveals that generating poverty rates by predicting from the last observed survey (right column) gives lower errors than predicting poverty rates or the mean (and Gini) without accounting for changes from the last survey (left column). Nearly all methods trying to predict poverty rates directly give a higher error than the worst method trying to predict from the last survey.

The poor performance of models trying to predict poverty rates directly is probably not due the fact that they have poverty rates on the left-hand side, but more likely that the models we have used with levels of poverty, mean and the Gini on the left-hand side do not use country fixed effects or lagged values. Under one interpretation, this treats these models equally to the models that predict changes from the last distributions, which likewise do not have country-fixed effects or lagged values. Such factors are less important when we have already taken first differences, though. Given that surveys within countries often are incomparable and often have many years between them, lagged information may not provide much additional information. Yet, to the extent that there are country specific effects in levels that are hard to predict, then any model that does not use
fixed effects or lagged values will not perform well.

Among the different ways of predicting from the last survey, predicting changes in poverty rates (panel 2) performs worse than predicting growth in the mean or mean and Gini (panel 4 and 6). Predicting the Gini along with the mean tends to slightly increase the error. In other words, assuming no changes to inequality works better than trying to predict changes in inequality since the last survey.\footnote{This is a slightly counterintuitive result that appears to emerge because the models that also predict the Gini work better on average for countries with low poverty rates but slightly worse for countries with high poverty rates. The latter effect dominates our loss function. Perhaps this is because that most of the variation of changes in the Gini that our models can pick up is from}
The best performing method is a conditional inference random forest which just predicts growth in the mean, giving an error of 3.65 percentage points (panel 4).

### 4.2 Comparison to models just using GDP growth

It is interesting to see how the methods above compare with the method that simply scales the distribution up according to growth in GDP per capita. Such a model is simpler to understand, simpler to implement, and has the additional advantage that it can be used for nowcasting as well, given that GDP growth rates themselves are nowcasted by various institutions. The models presented in the last subsection relied heavily on variables that are not available for nowcasting. Depending on how well the models predict when missing values are present, this will make them perform worse for nowcasting. In Figure 5 we compare the method of only using GDP growth with the best performing machine learning method, the best performing machine learning method that predicts poverty rates directly, as well as three other scenarios. These other three scenarios are as follows. First, a method which only allows a fraction of GDP per capita growth to be passed through to the distribution and where this fraction is estimated through a simple linear regression. The fraction passed through turns out to be 79%. This follows empirical evidence showing that on average, only a fraction of growth in national accounts trickles down to household surveys (Deaton, 2005; Lakner et al., 2020; Pinkovskiy and Sala-i Martin, 2016; Ravallion, 2003). Second, we estimate this passthrough rate separately by consumption and income welfare aggregates. This is motivated by the fact that this interaction was the first variable to enter in the lasso predicting growth in the mean. With this method, 71% of growth is estimated to pass through to consumption aggregates while 97% is estimated to pass through to income aggregates. Third, we compare the predictions with the predictions we would get if we were perfectly able to predict the mean. This is done by shifting a distribution from the beginning of a spell such that it matches the mean of the distribution at the end of the spell. This is what would happen if our predictions of the mean or growth in the mean were error free. As such, it represents the lowest possible prediction error of using method (3) or (4).

We find that the best performing machine learning method only reduces the wealthy countries. Note also that if the distributional assumptions or growth incidence curves we impose are not accurate, then even if we predict the Gini and mean well, it does not mean that we predict the poverty rate well.  

---

17

3A similar method would be to rely on growth in Household Final Consumption Expenditure (HFCE), often the largest component of GDP. HFCE is rarely nowcasted and not available for all countries for nearcasting, making it slightly less attractive from a practical perspective.
Figure 5: Comparing best performing methods to models only using GDP growth

Notes: The figure compares errors of six different models. The best performing method takes the minimum bar from Figure 4 while the best method predicting poverty rates directly takes the minimum bar of panel 1 of Figure 4. Using only GDP growth to shift the mean refers to predicting poverty by adjusting mean consumption by the growth in GDP per capita. The rightmost column reflects a hypothetical scenario of the errors one would get if one was perfectly able to predict growth in the mean.

error by 0.12 percentage points over just using GDP growth, only by 0.05 percentage points if a passthrough rate is used, and only by 0.04 if an income/consumption specific passthrough rate is used. This means that just using GDP per capita growth to nearcast distributions is nearly as accurate as any method using more than 1,000 variables and complex machine learning methods. This is not because the method of first predicting the mean and adjusting the distribution accordingly does not work – if growth in the mean was predicted perfectly, the mean absolute deviation would nearly half to 1.91 percentage points. Rather, it is because nearly all of the variation in growth in mean consumption that can be explained with these 1,000 variables can be explained by growth in GDP per capita. The figure also reveals that just using growth in GDP per capita to shift the mean gives a better performance than any model trying to predict poverty rates directly using 1,000 variables.

Though the findings above apply across countries on average, a point of interest is whether there are cases where this may not apply. We explore this in Figure 6 by plotting the mean absolute deviation for three select models – the best overall, the best at predicting poverty rates directly, and the best model only
using a fraction of growth in GDP per capita – as a function of extrapolation time (the time between two household surveys), poverty rates, year of the data, and annualized growth in real GDP per capita. In all of these figures, the sample of countries changes over the x-axis. For example, the countries with an extrapolation time of one year are mostly in Europe and South America, while the countries with an extrapolation of 10 years tend to be in Sub-Saharan Africa. As a result, looking at the trend for a given line is of little interest. However, the gap between the three lines can be used to explore, for example, for countries with a five-year extrapolation time, which method performs better?

It is plausible that using changes from the last survey is a good strategy when the last survey is only a couple of years old, while it is a less sound strategy when the last survey is a decade old or more. Extrapolation time does not matter when predicting poverty rates directly, but the average prediction error might still be a function of extrapolation time for the reason mentioned above – that the sample of countries changes. Panel (a) of Figure 6 shows that even for extrapolation times of five years, predicting changes from the past survey by just using growth in GDP per capita with a passthrough rate outperforms a model using 1,000+ variables trying to predict the poverty rate directly and performs as well as models predicting growth in the mean using 1,000+ variables. For extrapolation times beyond five years, we have too few observations to say anything with sufficient certainty, but we can rule out that just using GDP growth is significantly worse than the other methods.

In panel (b) we look at whether predicting poverty rates directly could be better for either poor or non-poor countries. Predicting poverty rates directly appears to be worse for all poverty levels except for countries with poverty rates around 30%. Particularly for very poor countries and countries with low extreme poverty, modeling changes through GDP is preferred to predicting poverty rates directly.

Deaton and Schreyer (forthcoming) argue that GDP is increasingly becoming detached from national material well-being. This could imply that using growth rates in GDP per capita has become less attractive, relative to predicting poverty rates directly, in recent years. Panel (c) shows that we do not find this to be the case. A possible reconciliation between these two findings is that, while GDP has become increasingly detached from household surveys, household surveys have become more comparable across rounds within the same country, allowing predictions from one survey to the next not to deteriorate.
Figure 6: Errors as a function of other variables

(a) Extrapolation time

(b) Poverty rates

(c) Year

(d) GDP growth rates

Notes: The figures plot local polynomials of the absolute deviations as a function of other variables. The trend for a particular line is not interpretable, since, for example, the sample of countries with extrapolation time of two years might be very different from the sample of countries with an extrapolation time of four years. Yet the gap between the lines is interpretable. We only show confidence intervals for one model to not clutter the graph. The confidence interval evaluates the uncertainty of the local polynomial fit but does not incorporate the uncertainty regarding the survey-level predicted poverty rates themselves.

In panel (d) we explore whether using GDP growth rates to project forward only works for some levels of growth rates, such as only positive growth rates or only non-extreme growth rates. This could be relevant for years of crises – such as the economic crises following COVID-19 – where one could suspect that models
trained on historical data may be vulnerable. Yet, we do not find it to be the case that using GDP growth to project forward is problematic for some growth rates, suggesting, possibly, that even during recessions such as the one caused due to the pandemic, predicting changes from the last survey might be preferable to predicting poverty rates directly. In sum, we do not find overwhelming evidence that predicting poverty rates directly could be more attractive when compared to simply shifting the mean with adjusted GDP growth rates.

4.3 Predictors of poverty

Though the findings so far suggest that growth in real GDP per capita is the overwhelming predictor of poverty, we can look at this more systematically by analyzing the features that are important for the predictions. For the lasso, this can be analyzed by looking at the order in which variables enter. For forests and gradient boosting, it can be analyzed through feature importance measures. Feature importance measures indicate the importance of a particular feature for the accuracy of the predictions.

In Figure 7, we use the feature importance measures from the conditional inference random forests, as this method generally performed best across the machine learning algorithms (judged by its average rank in Figure 4). For a particular feature, the feature importance is calculated by permuting the feature in the dataset and calculating how much the prediction error increases. The importance measure is standardized such that the variable with the highest value gets a value of 1. The equivalent plots for the other methods look similar.

The only six features that are substantively predictive of growth in the mean are all national accounts variables. They are all identical to, or highly correlated with, growth in GDP. The most predictive variable is growth in final consumption expenditure, which is the sum of two components of GDP, government expenditure and HFCE. The second most predictive variables is growth in HFCE, followed by growth in GDP, GNI, and gross domestic income (GDP measured from the income side). The variable most informative not from national accounting is growth in the employment rate. Yet, it is clear from this figure that just using growth in GDP per capita sums up the information well. Though growth in final consumption expenditure could be used instead to obtain a slightly lower error, final consumption expenditure is often not nowcasted widely, making it only feasible to use for nearcasting.
Figure 7: Variables important for predicting growth in mean welfare

Notes: The variable importance measure comes from the conditional inference random forest. For a particular feature, the value is calculated by permuting the feature in the dataset and calculating how much the prediction error increases. The importance measure is standardized such that the variable with the highest value gets a value of 1.

For completeness, we show what predicts the other five target variables in Figures B.2 and B.3 in the Appendix.

4.4 Implications for global poverty

Though the purpose of this paper is not to analyze current patterns of global poverty, below we show a couple of figures on how the methods we employ can (and, as we will see, cannot) be utilized to get an up-to-date picture of extreme poverty. Figure 8 shows the global and regional trends in extreme poverty from 2014-2021, the nowcasting year as of the time of writing, using GDP growth to shift the mean, the best model overall, and the best model predicting poverty rates directly. For countries without at least one prior estimate of poverty, we use the last method for the former two.

Looking at Figure 8, it stands out that the best model that predicts poverty rates directly deviates for the nowcasting year. This is because the particular model used – gradient boosting – turns out not to work well for 2021 where many of the input variables are missing. Though we could have rerun the model only on variables widely available at the nowcasting year, it is not clear that this would make things better as the model then shifts from one year to the next. For that reason, this model is best only to use for nearcasting. Even for nearcasting, looking closely at the East Asia & Pacific panel reveals that methods trying to predict poverty rates directly might yield unlikely trends. The relatively large increase in the regional poverty rate in 2017 corresponds to the year where survey data
Figure 8: Nowcasted global poverty

Notes: For countries with at least one estimate we use the status quo with separate passthrough rates by income/consumption and for countries without any prior data we use gradient boosting to predict poverty rates directly. Note that the scale of the y-axis differs for each graph.

for China no longer is available. Prior to 2016, the survey estimates for China are used. The jump suggests that the model predicts a higher poverty for China than the official estimates. It is hard to support a case for poverty in East Asia increasing rapidly from 2016 to 2017, though. This, we believe, speaks intuitively in favor of models that are anchored in past estimates.

When comparing the model only using GDP growth and the best model overall, it is evident that for most regions the two are nearly aligned. The exception is South Asia, and as a consequence, the world as a whole, where just using GDP per capita suggests a lower poverty rate for India. In other words, other indicators have progressed less fast in India than growth in real GDP per capita.

Globally, for the best model and the model only using GDP growth, we find an increase in extreme poverty of about 0.6 and 0.8 percentage points in 2020 but that more than half of this increase will reverted in 2021. The increases in 2020 were largest in terms of percentage point in Sub-Saharan Africa and South Asia, but largest relatively speaking in the Middle East & North Africa and Latin America & Caribbean. The nowcasts suggests that both of these regions may see further increases in extreme poverty in 2021.

Figure 9 shows the nowcasted poverty levels around the world using the GDP growth model. The map gives a clear picture of poverty being concentrated in Sub-Saharan Africa, in fact, 23 of the 24 poorest countries are found in Africa (with the Republic of Yemen being the only non-African country). Particularly the
landlocked countries in Africa have high levels of poverty, with South Sudan topping the list with a poverty rate of 82%. High levels of poverty can also be found in other pockets around the world, for example in Yemen (49%), Afghanistan (40%), the República Bolivariana de Venezuela (32%), Papua New Guinea (30%), and Haiti (25%).

Figure 9: Global Poverty Map, 2021

Notes: Nowcasts of extreme poverty around the world using the model that shifts the last distribution by a fraction of growth in real GDP per capita. For countries without any prior data we use gradient boosting to predict poverty rates directly.

4.5 Preferred model and its limitations

Based on the findings of the past subsections, and due to the simplicity and interpretability of just using GDP growth to project the distribution forward, the GDP model with a passthrough rate of around 1 for income aggregates and 0.7 for consumption aggregates is our preferred method to nearcast poverty.

When nowcasting poverty, all methods are likely to perform worse. The machine learning methods will perform worse for two reasons. (1) If missing values are not imputed, the number of features will be much smaller, and if missing values are imputed, the features will be of worse quality because imputations are not a perfect substitute to actual data. (2) Since the features with data in the nowcasting year are based on modeling, extrapolations, or data for only part of the year, they are likely to be less accurate and ultimately less connected to the welfare distributions.
Our preferred model of only using growth in real GDP per capita does not suffer from the first issue but it does suffer from the last issue. The nowcasted growth rates have not yet been realized and are likely to deviate from the growth rates that eventually will be estimated by national authorities. Hence, they will likely be less predictive of welfare changes observed in household surveys. Since our preferred method for nearcasting, in that sense, only suffers from one of the two extra challenges that the other models suffer from when used for nowcasting, we think its advantages stand out even more when nowcasting. In fact, when running all our methods earlier in the calendar year (recall that earlier in the calendar year the informational space available for nearcasting approximates that of nowcasting), just using GDP growth becomes relatively more advantageous.

Using growth in real GDP per capita also comes with risks. If it is the case that nowcasted estimates of GDP are much more off target than nowcasted estimates of other variables, then it is possible that using only GDP growth will be relatively less attractive for nowcasting than nearcasting. Some evidence suggests that nowcasted growth rates by IMF might be too optimistic (Sandefur and Subramanian, 2020).

Another concern arises if GDP numbers are subject to quality concerns and are not deemed credible. If this is the case, then they may likewise be less relevant for both nearcasting and nowcasting. The fact that using growth rates in real GDP per capita worked well when trained on historical data suggests that historically this has not been frequently the case (or at least not more so than quality concerns with other indicators). Nonetheless, Figure 8 did show a large discrepancy for India when using growth in real GDP per capita and using more explanatory variables. Though this may simply be because true growth in real GDP per capita is less connected with household consumption in India than in other countries, evidence suggests that GDP growth rates in India in recent years are non-credibly high (Subramanian, 2019). If this is indeed the case, then this serves as another argument against only relying on growth in real GDP per capita.

Despite of these arguments against just using national accounts we think the arguments in favor – particularly the relatively high accuracy, simplicity, ability to work for both nearcasting and nowcasting, ability to work with any poverty line, and ability to avoid large shifts when moving from survey estimates to extrapolations – create a compelling case.
5 Robustness checks

In this section we test some of our main results under different methodological choices: first by only using comparable survey spells and second by using the mean squared error as the lost function.

5.1 Using only comparable survey spells

First, we rerun all results using only comparable survey spells. As we mentioned earlier in the text, non-comparable spells tend to make it more difficult to predict changes from the past survey, meaning that we would expect the inferiority of predicting poverty rates directly to hold even stronger when restricting the samples to comparable spells. We show the two main figures from the main text in Figure B.4 and Figure B.5 in the Appendix. Note that the mean absolute deviations are not comparable to the equivalent figures in the main text given that the countries with comparable spells tend to be less poor (implying that the mean absolute deviations will be lower). However, within Figure B.4 it is still possible to compare across and within panels.

Our main results all hold. Predicting poverty rates directly is far from optimal, predicting changes in poverty is worse than predicting growth in the mean or growth in the mean and trying to predict the Gini as well does not help above and beyond the distribution neutral assumption. Once more, a model which just applies a passthrough rate to growth in GDP per capita performs nearly as well as the best of all machine learning methods.

5.2 Using the mean squared error

A more common loss function than the one we applied is the root mean squared error (RMSE). We did not use the RMSE since it tends to give larger weight to the prediction of outliers, which in the context of poverty measurement probably are estimates suffering from measurement error. We do not want to tailor our models towards these data points. Figure B.6 and Figure B.7 in the Appendix test our main results using the RMSE. Some of our main results still hold, but now predicting poverty rates directly with gradient boosting is relatively more attractive (with an error of 6.86). In fact, it is now better performing that shifting the mean with a fraction of growth in GDP per capita (6.92). Once more, however, the potential gain in accuracy of any machine learning model hardly merits the extra complexity they add. It remains the case that predicting changes in poverty
is worse than predicting growth in the mean, and that trying to predict inequality as well does not help.

6 Conclusion

In this paper, we have analyzed how best to nowcast poverty around the world. We applied statistical learning techniques to the World Bank’s collection of international poverty estimates utilizing more than 1,000 development indicators as features. We looked at whether predicting poverty rates directly had a higher accuracy than predicting poverty indirectly, by, for example, predicting changes in poverty from the last survey, by predicting growth in mean welfare and applying this growth to scale up the past distribution, or by predicting growth in the mean and in the Gini and applying growth incidence curves to scale and stretch the past distributions.

Our findings revealed that a model which simply uses a fraction of growth in real GDP per capita since the last observed household survey to shift the entire distribution performs better than any model that predicts poverty rates directly using all the variables mentioned above, and nearly as well as all models trying to predict growth in the mean and in the Gini using all the variables mentioned above. This suggests that conditional on knowing growth in real GDP per capita, no other variable can substantially increase the predictive accuracy.

On the one hand, given the decades long literature documenting the importance of growth for poverty reduction (Kraay, 2006; Ferreira and Ravallion, 2009), this is not surprising. Partially as a result of this literature, variants of the model we prefer have been used both in the literature and for the World Bank’s official global poverty monitoring. On the other hand, several papers have documented important gaps between national accounts data and household survey data, suggesting that GDP might not be that connected to welfare as measured in household surveys (Ravallion, 2003; Deaton, 2005; Ferreira et al., 2010; Pinkovskiy and Sala-i Martin, 2016; Prydz et al., 2021; Deaton and Schreyer, forthcoming). What can explain these seemingly contradictory statements?

First, although the model using growth rates in GDP per capita performs better than nearly all other models, it does not imply that the model’s error is low.4 Note that this pertains to the error of country-level nowcasts. The results do not directly assess the accuracy of the World Bank’s regional and global nowcasts, since a large portion of the country-specific errors will average out when aggregating across many countries. However, the results do suggest that there are substantial discrepancies between nowcasts and survey-based measures for individual countries, and that the country nowcasts should be interpreted with ap-
In fact, our preferred model explains only about 25% of the out-of-sample variation in growth in mean welfare. The models using 1,000+ variables perform at about the same level or slightly better. In other words, most of the variation we are trying to predict cannot be predicted by any information we are utilizing.

Why is poverty so hard to predict by any model? We hypothesize five possible answers. First, the available features may not be well-suited to predict growth in welfare. Though we rely on predictors from different databases including remote sensing data, it is possible that the most important features are contained in other data sources, such as mobile phone data or proprietary remote sensing data. Second, it may be that other algorithms or ensemble learning might perform better. Third, this analysis has only considered data at the country level, and it is possible that models specified at the subnational level may perform better.

Fourth, the reason for the seemingly poor performance of all methods may be due to measurement error and measurement differences in welfare aggregates across countries. Issues like whether a recall or diary is used, number of consumption items asked about, the treatment of rent and durables, and differential non-response to surveys vary between and within countries and can have significant bearings on poverty rates (Beegle et al., 2012; Jolliffe, 2001). This affects both poverty rates in the cross-section as well as changes. Accounting for these issues would require a detailed database on the methodology used to construct welfare aggregates, which is not currently available. Finally, the fact that the distributions of many countries have a lot of mass around the international poverty line means that even small changes to the distribution of welfare can yield large changes to the poverty rate. This makes it difficult to predict the poverty rate with great precision.

Regardless of which, if any, of these hypotheses can explain why no model is able to explain a significant fraction of the variation in the growth in welfare, a very simple model utilizing growth in real GDP per capita and a prior welfare distribution remains an appealing method.
References


A Evaluation of model performance

In order to tune the machine learning algorithms we use, select among the different modeling methods presented in Table 1, and report the final out-of-sample errors, we rely on nested 5-fold cross validation (Stone, 1974; Iizuka et al., 2003; Varma and Simon, 2006). To this end, we split our data into five folds (the outer folds) and split each of these five folds into five folds themselves (the inner folds). Hence, each datapoint is part of two nested folds. For each machine learning algorithm, we tune the models using the inner folds and use the outer folds to select among the different methods and report the out-of-sample errors.

In more detail, we take four of the five outer folds, and on this subset, run a grid of models with various tuning parameters on four of the five inner folds. We estimate the error on the fifth inner fold, and repeat, sequentially leaving out a fifth of the already subset data. This will allow us to find optimal tuning parameters without having touched the fifth outer fold. We repeat this five times holding out a different outer fold, thus getting five different optimal tuning parameters, each of which has not used one outer fold. Thus far we have run $5 \times 5 \times g$ models, where $g$ is the size of the tuning grid. Subsequently, we take the average of the five optimal tuning parameters, run the optimal model on four of the five outer folds, estimate the error of that model of the held-out outer fold, and repeat over the five outer-folds. This gives us the out-of-sample error of the tuned model.$^5$

Nested k-fold cross validation avoids a problem with using regular k-fold cross-validation for tuning and model selection at the same time. If using regular k-fold cross validation for both of these tasks, then there will be a bias towards selecting a model with a large tuning grid and a downward bias in the final out-of-sample errors obtained. To see this, suppose we want to compare two machine learning methods that on expectation perform equally well. Suppose further that one of the methods can be run with 50 different tuning parameter options ($g = 50$) that all perform equally well on expectation while the other method has no tuning parameters ($g = 1$). When using cross-validation for model selection and model evaluation at the same time, the method with tuning parameters is likely to be chosen 50 out of 51 times instead of half of the time. In addition, the error of the final model is likely to be unlikely low, since it by chance was the best performing of all iterations. If we instead had evaluated the methods on data not

$^5$For random forests, rather than creating inner folds, we rely on the out-of-bag error. The out-of-bag error leverages the fact that each tree in a forest is only based on a subset of observations. To get an out-of-sample error for an observation, the trees in which an observation did not figure are utilized. This avoids having to run $5 \times 5 \times g + 5$ models, and rather only $5g + 1$ models.
used for the model selection, the out-of-sample error would not be downwards biased.

An alternative to using nested k-fold cross-validation could be to reserve the last spell for each county, applying cross-validation on the remaining data to tune models, and use the last spell for choosing methods and reporting predictive performance. Reserving the last spell of each country mimics what Bergmeir and Benítez (2012) call last-block validation. This would leave us with only about 150 data points to evaluate the error, of which more than half have a poverty rate less than 3%. We think this would result in a high uncertainty around the final error, which might lead us to choose the wrong method (Bergmeir and Benítez, 2012).

An assumption necessary for regular and nested cross-validation to work is that the residuals are i.i.d (Arlot and Celisse, 2010). This assumption is violated if there are spatial or temporal patterns in the dataset unaccounted for by the models. This assumption is likely violated in our methods, particularly so when we predict the levels of an outcome variable. The reason for this is simple: given that a country is represented repeatedly in our data, a country’s residual in year $t$ will likely not be independent of that same country’s residual in $t - 1$. This could occur, for example, if part of a country’s welfare aggregate, such as durable goods, is systematically excluded. This country’s poverty rates would all be higher than what we would expect had the welfare aggregate been constructed with durable goods. Supposing none of our features are able to fully pick up this country-specific idiosyncrasy, the residuals from this country’s poverty rate predictions will not be i.i.d. When predicting changes or growth in an outcome variable, it is likely that some of the spatial and temporal patterns disappear, meaning that the i.i.d assumption is less of a concern.

A method to partially overcome the issue of the residuals not being i.i.d would be to use a temporal block cross-validation – assigning observations to folds in blocks of time – or a spatial block cross-validation – assigning all estimates for a country to the same fold. Yet, to the extent that our models are able to explain at least part of the country- or time-specific peculiarities, such models would likely result in less relevant nowcasts. Continuing the example from before, suppose that our models actually were able to find a variable that predicts the relatively high poverty in our hypothetical country where durable goods are excluded from the welfare aggregate. Arguably, our nowcasts for this country should include this country-specific bias, since – had an actual poverty rate been available for the country in the nowcasting year – it would likely have continued the country-specific idiosyncrasy. If using spatial block cross-validation, this country would
not have data in the estimation folds and held-out folds at the same time, meaning that the models that utilize the variable particular to the country’s idiosyncrasies will perform poorly (supposing it does not work for well other countries), and a model not utilizing this variable will likely be chosen.

It is not clear whether this is attractive or not. On one hand, one could argue that if we want to predict some latent unobserved poverty rate of which actual estimates are only a noisy signal, then a nowcast which ignores potential biases from this signal is preferred. On the other hand, the prior estimates of poverty often carry some authority and are treated as the ground-truth by governments. Hence continuing the country-specific particularities into the nowcasting year may be preferable. Since we are of the latter belief, we will not use blocked cross-validation. Bergmeir et al. (2018) have shown that as long as the data are fitted well by the model, cross-validation without any modification works well in practice. This gives us at least some reassurance that i.i.d violations may not be too problematic.

That said, when predicting poverty for a country without any prior poverty estimates, we are likely to get more accurate estimates if using a spatial block cross-validation. For those countries, using models which fit to country-specific peculiarities are likely to work less well than using models that are evaluated on data fully based on new countries, akin to what a spatial block cross-validation does. Since we are more concerned with nowcasting poverty for countries with prior estimates (where, as we have argued, we think blocked spatial cross-validation is not preferable) than predicting poverty for countries without prior poverty estimates (which concerns less than 2% of the world’s population), we think this speaks in favor of not using spatial block cross-validation.
B  Further results and robustness checks

Figure B.1: Performance of methods predicting level/growth in mean and Gini

Notes: carf = CART random forest, cirf = conditional inference random forest, grbo = gradient boosting, plas = post-lasso, rlas = regular lasso.
Figure B.2: Variables important for predictions (1/2)

(a) Poverty rates

(b) Changes in poverty rates

(c) Mean consumption

Notes: The variable importance measure comes from the conditional inference random forest and measures the total decrease in residual sum of squares from splitting on the variable, averaged over all trees. The importance measure is standardized such that the variable with the highest value gets a value of 1.
Figure B.3: Variables important for predictions (2/2)

(a) Gini coefficient

(b) Growth in gini coefficient

Notes: The variable importance measure comes from the conditional inference random forest and measures the total decrease in residual sum of squares from splitting on the variable, averaged over all trees. The importance measure is standardized such that the variable with the highest value gets a value of 1.
Figure B.4: Performance of various methods using only comparable spells

1. Levels of poverty rates
2. Changes in poverty rates
3. Levels of mean
4. Growth in the mean
5. Levels of mean & Gini
6. Growth in the mean & Gini

Notes: carf = CART random forest, cirf = conditional inference random forest, grbo = gradient boosting, plas = post-lasso, rlas = regular lasso.
Figure B.5: Performance of subset of methods using only comparable spells

Notes: The figure compares errors of six different models. The best performing method takes the minimum bar from Figure B.4 while the best method predicting poverty rates directly takes the minimum bar of panel A of Figure B.4. Using only GDP growth to shift the mean refers to predicting poverty by adjusting mean consumption by the growth in GDP per capita. The rightmost column reflects a hypothetical scenario of the errors one would get if one was perfectly able to predict growth in the mean.
Figure B.6: Performance of various methods using the RMSE

Notes: carf = CART random forest, cirf = conditional inference random forest, grbo = gradient boosting, plas = post-lasso, rlas = regular lasso.
Figure B.7: Performance of subset of methods using the RMSE

Notes: The figure compares errors of six different models. The best performing method takes the minimum bar from Figure B.6 while the best method predicting poverty rates directly takes the minimum bar of panel A of Figure B.6. Using only GDP growth to shift the mean refers to predicting poverty by adjusting mean consumption by the growth in GDP per capita. The rightmost column reflects a hypothetical scenario of the errors one would get if one was perfectly able to predict growth in the mean.