Moving targets: When does a poverty prediction model need to be updated?

Emily Aiken[*]
University of California, Berkeley
emilyaiken@berkeley.edu

Tim Ohlenburg[*]
University College London
email@timohlenburg.com

Joshua Blumenstock
University of California, Berkeley
jblumenstock@berkeley.edu

Abstract

A key challenge in the design of effective social protection programs is determining who should be eligible for program benefits. In low and middle-income countries, one of the most common criteria is a Proxy Means Test (PMT) – a rudimentary application of machine learning that uses a short list of household characteristics to *predict* whether each household is poor, and therefore eligible, or non-poor, and therefore ineligible. Using nationwide survey data from six low and middle-income countries, this paper documents an important weakness in this use of machine learning: that the accuracy of the PMT prediction algorithm decreases steadily over time, by roughly 1.5-1.9 percentage points per year. We illustrate the implications of this finding for real-world anti-poverty programs, which typically update the PMT model only every 5-8 years, and then show that the aggregate effect can be decomposed into two forces: "model decay" caused by model drift, and "data decay" caused by changing household characteristics. Our final set of results show how an understanding of these forces can be used to optimize data collection policies to improve the efficiency of social protection programs.

1 Introduction

Each year, over a trillion dollars are spent on social protection programs globally, making up on average 13% of each country's gross domestic product [1]. Many of these programs are *targeted*, providing benefits to only eligible households. In low- and middle-income countries (LMICs), where high-quality poverty data is often unavailable, incomplete, or out-of-date [2], a rudimentary machine learning approach called a *proxy means test* (PMT) is a dominant targeting method. PMTs are currently in use in over fifty LMICs collectively containing over a billion people [3], making them one of the more widespread and consequential use cases of machine learning in government policy.

Proxy-means tests [4] use a machine learning (ML) model to predict per capita household consumption from data on household characteristics, including assets and demographics, collected in a comprehensive *social registry*. The ML model is trained on a representative household sample survey containing consumption expenditure labels; the model is then used to estimate the per capita consumption of every household for which characteristics are available in the registry. Program eligibility is determined by a threshold on estimated consumption.

PMT accuracy, typically quantified through targeting error rates of wrongful inclusion and exclusion, has been evaluated in a large number of papers [5, 6, 7, 4, 8, 9, 10]. However, the vast majority

[*]These authors contributed equally to this work.

Computational Sustainability Workshop at NeurIPS 2023.

of this literature looks at the performance of a PMT at a single point in time — the moment when the PMT data are collected and the PMT decision rule is implemented — which is also when the performance of the PMT is highest. In practice, most PMT-based targeting approaches are updated infrequently [11, 12]. This delay occurs because the updates are costly, involving some combination of (i) updating the household-level data in the registry with a "PMT sweep", which we estimate costs around $57 million purchasing power parity (PPP) in the median country (of the 13 for which we have data); and (ii) recalibrating the ML model with a new sample survey, which has a median cost of $14 million PPP.

In the time gap between PMT data collection and when policy decisions are made based on these data, the living conditions and poverty status of households may change. In this paper, we adapt the *dataset shift* framework [13] to quantify how PMT accuracy is impacted by gaps in time between data collection and PMT deployment. Leveraging 25 rounds of survey data from six countries over twelve years, we find that (i) for each year that a PMT is not updated, it explains 6 percentage points less of the variation in household consumption, resulting in an increase of targeting errors of 1.5-1.9 percentage points, (ii) *data decay* (losses in PMT accuracy due to covariate shift in PMT covariates) is roughly three times as powerful as *model decay* (losses in accuracy due to *model drift* in the learned relationship between the covarites and consumption expenditure), and (iii) based on international information on survey costs, most social protection programs should aim to collect social registry data and recalibrate the PMT model every 1-3 years.

2 Methods and Results

2.1 Data

Our analysis relies on publicly available panel surveys from the Living Standards Measurement Study (LSMS) in Ethiopia, Nigeria, Tanzania, and Uganda, the Ghana Panel Survey, and Peru's Encuesta National de Hogares. Each panel contains between three and five rounds, covering between five and eleven years (Table S1). Since our analysis relies on observing changing households conditions, we consider only households that appear in all rounds of the survey in each country. The resulting sample sizes range from 424 households in Tanzania to 3,393 households in Ghana.

2.2 Machine learning approach

Following the standard implementation of a PMT [4, 14, 9], we construct a machine learning experiment in which log-transformed per capita household consumption expenditure[2] is estimated based on "registry covariates" including: (i) housing-related variables, such as the material of the roof, walls, and floor; amenities such as a toilet or electricity; and information on building size and ownership, (ii) asset ownership variables unique to each country's context, and (iii) demographic information such as such as household size, number of children, and characteristics of the household head. Our base specification divides the surveyed households of each country randomly into a training set of 75% of households and an evaluation set of 25%. For each survey round, we train our machine learning models to predict per capita household consumption from the social registry covariates on training set households, and evaluate performance on the test set. Survey weights are used both in training and in calculating the accuracy metrics.

The primary machine learning approach we test is a linear regression paired with stepwise forward selection of input variables. Although a variety of machine learning models are used for PMTs in practice, and recent work has suggested that more complex models may provide marginal accuracy gains [9, 8, 15], linear regression with stepwise forward selection remains a standard approach to calibrating PMTs. We also compare the performance of the stepwise approach with a number of other models, including ordinary least squares with the full set of registry covariates, the LASSO, a random forest, and a gradient boosting machine (Table S2).

To evaluate PMT performance, we calculate the R^2 score and Spearman's rank correlation between predicted and ground truth consumption values on the test set. We also calculate targeting error rates of the PMT for hypothetical cash transfer programs that aim to target the poorest 20% and 40% of

[2]Throughout our consumption expenditure measures are adjusted for inflation across survey years using the consumer price index and then converted to purchasing power parity (PPP) dollars.

Table 1: Quantifying yearly decay for a PMT that is not updated

	Baseline Performance	**Combined Spec.**	**Decomposed Spec.**		
			Data Decay	*Model Decay*	*Interaction Term*
R^2	0.5053	-0.0610***	-0.0544***	-0.0479***	0.0058***
		(0.0064)	(0.0103)	(0.0103)	(0.0021)
Spearman	0.6923	-0.0287***	-0.0347***	-0.0115***	0.0033***
		(0.0040)	(0.0033)	(0.0033)	(0.0007)
TER(20)	0.4667	0.0186***	0.0207***	0.0078***	-0.0020***
		(0.0021)	(0.0018)	(0.0018)	(0.0004)
TER(40)	0.2947	0.0148***	0.0167***	0.0057***	-0.0014***
		(0.0016)	(0.0013)	(0.0013)	(0.0003)

Notes: Left: Average pre-decay performance (across countries). Middle: combined decay parameterization. Right: Decomposing combined decay into data decay, model decay, and their interaction. Stars determine statistical significance: * indicates $p < 0.10$, ** indicates $p < 0.05$, and *** indicates $p < 0.01$.

households.[3] All evaluation metrics are calculated separately for 100 simulations with random data splits, and the average across splits is reported.

2.3 Quantifying decay

To simulate decay caused by temporal gaps between data collection, model calibration, and deployment, we conduct the same machine learning experiment described in the base specification above, but this time introduce lags between training and evaluation sets. We start by estimating combined decay, the setting in which the PMT model and the social registry covariates are equally out-of-date. To simulate combined decay, the ML model is trained on the training set from round w_i, and performance metrics are calculated on the test set using predictions based on covariates from round w_i versus ground truth consumption data from round w_j, for all $i = j$ in the panel survey. To estimate combined decay, we regress the difference between PMT performance with temporal lags and performance using data that matches the time of the evaluation on the temporal lapse.

Next, to estimate the contributions of model and data decay to combined decay in PMT accuracy, we expand our dataset to include lag combinations where the model and data are unequally out-of-date. In this setting, the ML model is trained on the training set from round w_i, performance metrics are calculated on the evaluation set predictions that use covariates from round w_j, and true consumption data is taken from round w_k, for all $i, j <= k$ in the panel survey. In our decomposed specification, we regress decay on the time since model calibration, the time since data collection, and the interaction of the two.

We find that the PMT's accuracy degrades significantly over time: our combined decay estimates in Table 1 below and Figure S1 indicate that each year that a PMT is not updated results in a reduction of roughly six percentage points of the explained variation in household consumption. Targeting errors increase by between 1.5 and 1.9 percentage points each year, depending on the total coverage of the program. To put these numbers in context, with a program the size of PROGRESA in Mexico (which provides benefits to approximately 2.6 million households annually [16]), a delay of five years between PMT updates would be expected to produce around 200,000 additional exclusion and inclusion errors. We also find that although data decay and model decay contribute approximately equally to decreases in the R^2 score (coefficients of -5.4 and -4.8 percentage points, respectively), data decay is approximately three times more powerful in increasing targeting errors: decay parameters of 1.7-2.1 and 0.6-0.8 percentage points per year, respectively.

[3]Under such a quota approach, inclusion and exclusion error rates are equal and referred to as the *targeting error rate (TER)*, following [14]. We notate the targeting error rate with the relevant quota: for example, the targeting error rate for a program aiming to reach the poorest 20% of households is notated *TER(20)*.

Table 2: Recommended updating policies for selected real-world social protection systems

Country	Budget per Household	Recommendation	Country	Budget per Household	Recommendation
Argentina	$261	2	Ghana	$106	4
Benin	$20	10	Jamaica	$835	1
Colombia	$184	3	Mexico	$1,465	1
Congo	$1,163	1	Peru	$156	3
Costa Rica	$2,651	1	Philippines	$194	3
Ecuador	$1,060	1	Togo	$69	6
El Salvador	$721	1	**Median**	**$261**	2

Notes: Recommended updating policies calculated using program data from the Manchester Social Assistance Database [3], registry coverage data from [17], [18], [19], and [20], population data from the World Bank [21], and average household size data from the UN [22]. We assume 11% of each program's budget goes towards administrative costs [23]. The recommended updating frequencies assume a program coverage level of 20% (although recommended strategies are generally fairly robust across coverage levels). For all real-world programs, the recommended updating policy is symmetrical (that is, the ML model is recommended to be recalibrated at the same frequency as PMT sweeps), though non-symmetrical updating policies may be recommended for alternative benefit levels (see Figure S2). All costs are measured in 2015 USD PPP.

2.4 Assessing model and data refresh policies

Data collection for model recalibration and PMT sweeps is costly, but our results show that accuracy decay leads to a mis-allocation of social assistance towards households that are ineligible. The resulting dilemma for policymakers is how to balance survey cost and accuracy decay economically.

We test a set of 100 model and data refresh policies, representing every combination of social registry data updating and PMT model recalibration frequency from every 1-10 years. We consider the best updating policy as the one that minimizes the total cost of social registry data collection, consumption survey data collection for calibrating the PMT model, and benefit payments lost to mis-targeting.[4] We estimate the costs of consumption sample surveys and social registry data collection based on the median reported in the literature (Tables S3 and S4), coming out to $3.33 PPP per social registry household for the consumption sample suvey to train the PMT, and $13.31 PPP per social registry household for PMT covariate data collection. The best updating policy will also depend on the program's coverage, the size of the social registry, and the benefit amount that the program provides.

Table 2 analyzes thirteen social protection systems for which data on design parameters (including registry coverage and program budget) are available from the Manchester Social Assistance Explorer [3], paired with our estimates of the best updating strategy. While information on status quo updating policies is not generally available a gap of 5-8 years is typical [11]. Our results call for more frequent updating: most programs should collect new registry data and update their ML model every 1-3 years.

3 Discussion

Our work investigates the impact of dataset shift on the widely applied predictive modelling task of proxy means testing. We find that 1.5-1.9% of intended social assistance beneficiaries are excluded if the ML model and registry covariates used in a PMT are allowed to go out-of-date by a single year. By year five, which may be a fair estimate of the lag between data collection rounds [11, 12], that number has risen to around 8.5% of beneficiaries. These figures and the scale of targeted social assistance programs [3, 24] suggest that millions suffer the effects of accuracy decay each year.

There are several extensions of this work that we are currently exploring. First, we are investigating the *welfare impacts* of data and model decay using the framework from [5], and identifying welfare-maximizing PMT updating policies. Second, we are quantifying additional loss in PMT accuracy resulting from the creation of new households over time (see [25] on the importance of complete household lists). Third, and most central to our ongoing analysis, we are exploring methods from the domain adaptation and domain generalization literature [26] to design "decay-robust PMTs" less

[4] In ongoing work, we are also testing an alternative approach to identifying the best recalibration frequency based on optimizing a utility function assuming a fixed budget for data collection and program benefits.

prone to dataset shift than the standard approach. Specifically, we are comparing state-of-the-art approaches from the ML literature (such as invariant risk minimization [27], distributionally robust optimization [27], and CORAL [28]) to simpler feature selection approaches to reduce decay. Our preliminary results indicate substantial decreases in decay from eliminating unstable features, but little improvement from domain adaptation and generalization methods.[5]

[5]This preliminary result is consistent with other recent work that has found little benefit of domain generalization methods "in the wild" [26].

Supplementary Tables and Figures

Table S1: Panel surveys used in our analysis

Country	Survey Waves	Households	Social Registry Covariates
Ethiopia	**Three** waves over **five** years: 2011, 2013, 2015	3,169	52
Ghana	**Three** waves over **eight** years: 2009, 2013, 2017	3,393	54
Nigeria	**Four** waves over **nine** years: 2010, 2012, 2015, 2018	1,237	48
Peru	**Five** waves over **five** years: 2015, 2016, 2017, 2018, 2019	1,575	28
Tanzania	**Five** waves over **eleven** years: 2008, 2010, 2012, 2014, 2019	424	68
Uganda	**Five** waves over **seven** years: 2009, 2010, 2011, 2013, 2015	1,041	32

Notes: Summary statistics on the six panel surveys used throughout our analysis.

Table S2: Estimates of combined decay for different ML models

	Stepwise + LR	LR	LASSO	Random Forest	Gradient Boosting
R^2	-0.0610***	-0.0682***	-0.0580***	-0.0597***	-0.0645***
	(0.0064)	(0.0068)	(0.0066)	(0.0059)	(0.0064)
Spearman	-0.0287***	-0.0307***	-0.0293***	-0.0308***	-0.0324***
	(0.0040)	(0.0039)	(0.0041)	(0.0042)	(0.0043)
TER(10)	0.0173***	0.0180***	0.0166***	0.0162***	0.0198***
	(0.0021)	(0.0023)	(0.0022)	(0.0027)	(0.0027)
TER(20)	0.0186***	0.0200***	0.0184***	0.0204***	0.0229***
	(0.0021)	(0.0023)	(0.0023)	(0.0022)	(0.0023)
TER(30)	0.0170***	0.0185***	0.0181***	0.0187***	0.0193***
	(0.0017)	(0.0019)	(0.0019)	(0.0020)	(0.0020)
TER(40)	0.0148***	0.0165***	0.0155***	0.0151***	0.0163***
	(0.0016)	(0.0016)	(0.0017)	(0.0018)	(0.0018)

Notes: Replication of results on combined decay in Table 1 for five different ML approaches. Stars are determined by statistical significance of the coefficient in the relevant regression specification: * indicates $p < 0.10$, ** indicates $p < 0.05$, and *** indicates $p < 0.01$.

Table S3: Estimates of consumption survey costs

	Total cost (USD Nominal)	**Year**	**Number of HH in Country**	**Cost per HH in Registry (USD Nominal)**	**Cost per per HH in Registry (2015 PPP)**
Afghanistan	$2,289,000	2014	4,125,000	$2.64	$8.97
Bangladesh	$793,600	2010	32,975,809	$0.11	$0.40
Colombia	$1,936,000	2014	13,336,556	$0.69	$1.20
Costa Rica	$1,436,391	2006-2012	1,353,312	$5.05	$8.67
Ethiopia	$1,313,739	2011	20,532,887	$0.30	$1.13
Guatemala	$1,559,790	2014	3,188,816	$2.33	$4.83
Iraq	$3,874,000	2012	4,397,980	$4.19	$9.76
Kyrgyzstan	$245,784	2003	1,200,786	$0.97	$7.59
Malawi	$2,441,929	2010	3,365,799	$3.45	$8.09
Myanmar	$295,200	2015	12,258,083	$0.11	$0.42
Nepal	$1,233,528	2010	6,173,083	$0.95	$3.29
Nicaragua	$773,906	2014	1,193,976	$3.09	$7.97
Niger	$1,188,000	2011	3,757,198	$1.51	$3.36
Nigeria	$1,995,896	2010	35,970,814	$0.26	$0.59
Tanzania	$1,008,885	2014	10,370,317	$0.46	$1.02
Uganda	$1,178,100	2008	6,683,515	$0.84	$2.62
Peru	$2,275,216	2009	7,691,993	$1.41	$3.30
Yemen	$4,291,200	2014	4,142,284	$4.93	$10.87
Median	**$1,375,065**		**5,285,532**	**$1.19**	**$3.33**

Notes: Details of consumption survey cost calculations. Data on survey costs are based on data from the LSMS [29]. Data on the number of households in each country are based on population size are from the World Bank [21], and average household size information from the United Nations [22]. The cost per household in a hypothetical social registry is based on a median global social registry coverage of 21% from [17]. Costs are converted to local currency (with exchange rates from the World Bank [21]), to PPP, and then deflated to 2015 PPP based on the US GDP deflator (PPP exchange rates and GDP deflator are taken from the IMF Economic Outlook Database [30]).

Table S4: Estimates of PMT survey costs

Country	**Year**	**Cost per Survey (USD Nominal)**	**Cost per Survey (2015 PPP)**	**Source**
Burkina Faso	2016	$5.69	$15.52	[31]
Chad	2016	$9.50	$22.58	[31]
Honduras	2008	$2.62	$6.46	[32]
Indonesia	2009	$2.70	$9.93	[6]
Mali	2016	$4.00	$11.11	[31]
Niger	2016	$6.80	$15.52	[31]
Peru	2010	$3.05	$6.32	[32]
Tanzania	2017	$12.00	$29.38	[33]
Median		**$4.85**	**$13.31**	

Notes: Cost per household in the social registry for PMT surveys, from four sources. Survey costs are converted to local currency (with exchange rates from the World Bank Development Indicators [21]), to PPP, and then deflated to 2015 PPP based on the US GDP deflator (PPP exchange rates and GDP deflator are taken from the IMF Economic Outlook Database [30]).

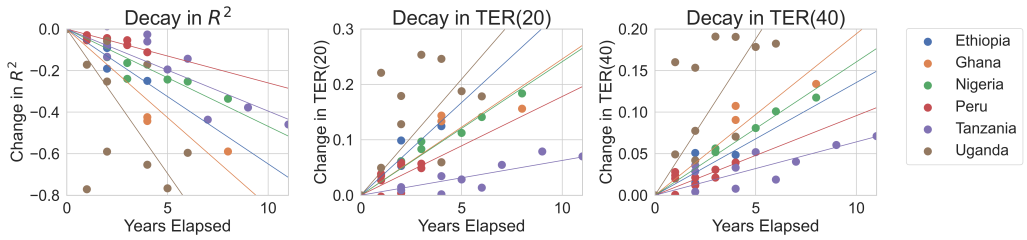

Figure S1: Linear combined decay estimates plotted separately for each country for the coefficient of determination (left) the targeting error rate for a program with 20% coverage (middle), and the targeting error rate fora program with 40% coverage (right).

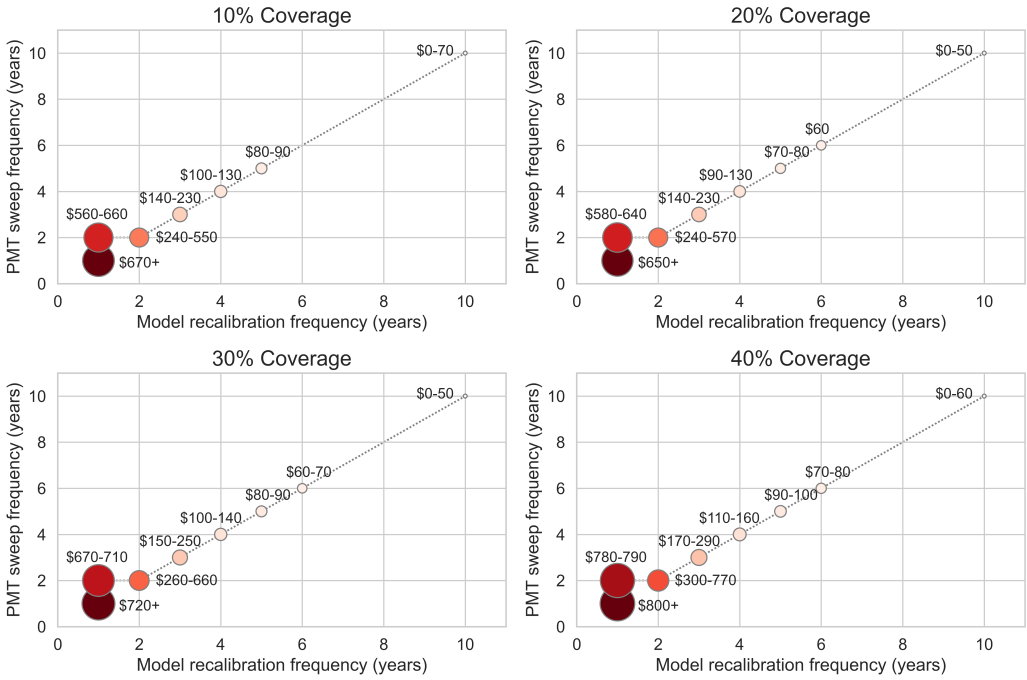

Figure S2: Summary of optimal PMT updating strategies for different program coverage levels (each shown in a different subplot) and different average values of benefits per household in the social registry (shown in colored markers, where markers are colored and sized by the average benefit value). Benefit values are measured in USD 2015 PPP. The X-Y location of the marker on each plot denotes the optimal policy, with the model recalibration frequency (in years) on the x-axis and the frequency of PMT sweeps to collect social registry data (also in years) on the y-axis.

References

[1] ILO. World social protection report 2020–22: Social protection at the crossroads–in pursuit of a better future, 2021.

[2] Morten Jerven. *Poor numbers: how we are misled by African development statistics and what to do about it*. Cornell University Press, 2013.

[3] A Barrientos. Social assistance in low and middle income countries 2000-2015. 2018.

[4] Margaret Grosh and Judy L Baker. Proxy means tests for targeting social programs. *Living standards measurement study working paper*, 118:1–49, 1995.

[5] Rema Hanna and Benjamin A Olken. Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, 32(4):201–26, 2018.

[6] Vivi Alatas, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias. Targeting the poor: evidence from a field experiment in indonesia. *American Economic Review*, 102(4):1206–1240, 2012.

[7] Patrick Premand and Pascale Schnitzer. Efficiency, legitimacy, and impacts of targeting methods: Evidence from an experiment in niger. *The World Bank Economic Review*, 35(4):892–920, 2021.

[8] Alejandro Noriega-Campero, Bernardo Garcia-Bulle, Luis Fernando Cantu, Michiel A Bakker, Luis Tejerina, and Alex Pentland. Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 241–251, 2020.

[9] Linden McBride and Austin Nichols. Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, 32(3):531–550, 2018.

[10] Michael Hillebrecht, Stefan Klonner, and Noraogo A Pacere. The dynamics of poverty targeting. *Journal of Development Economics*, 161:103033, 2023.

[11] V Barca and M Hebbar. On-demand and up to date? dynamic inclusion and data updating for social assistance. *GIZ (https://socialprotection.org/sites/default/files/publications_files/GIZ_DataUpdatingForSocialAssistance_3.pdf)*, 2020.

[12] Ignacio Irarrázaval et al. Sole information systems on beneficiaries in latin america. Technical report, Inter-American Development Bank, 2011.

[13] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

[14] Caitlin Brown, Martin Ravallion, and Dominique Van de Walle. A poor means test? econometric targeting in africa. *Journal of Development Economics*, 134:109–124, 2018.

[15] Ana Areias and Matthew Wai-Poi. Machine learning and prediction of beneficiary eligibility for social protection programs. *Revisiting Targeting in Social Assistance*, page 507, 2022.

[16] Emmanuel Skoufias. *PROGRESA and its impacts on the welfare of rural households in Mexico*, volume 139. Intl Food Policy Res Inst, 2005.

[17] Margaret Grosh, Phillippe Leite, Matthew Wai-Poi, and Emil Tesliuc. *Revisiting targeting in social assistance: A new look at old dilemmas*. World Bank Publications, 2022.

[18] Heidi Berner and Tamara Van Hemelryck. Social information systems and registries of recipients of non-contributory social protection in latin america in response to covid-19. 2021.

[19] Kathleen Beegle, Aline Coudouel, and Emma Monsalve. *Realizing the full potential of social safety nets in Africa*. World Bank Publications, 2018.

[20] Phillippe Leite, Tina George, Changqing Sun, Theresa Jones, and Kathy Lindert. *Social registries for social assistance and beyond: a guidance note and assessment tool*. World Bank, 2017.

[21] World Bank. World bank open data. `https://data.worldbank.org/`, 2023.

[22] United Nations. Household size and composition around the world. *Economic and Social Affairs*, 2017.

[23] Isabel Ortiz, Fabio Duran, Karuna Pal, Christina Behrendt, and Andrés Acuña-Ulate. Universal social protection floors: Costing estimates and affordability in 57 lower income countries. *ILO Extension of Social Security Working Paper*, (58), 2017.

[24] Ugo Gentilini, Mohamed Bubaker Alsafi Almenfi, TMM Iyengar, Yuko Okamura, John Austin Downes, Pamela Dale, Michael Weber, David Locke Newhouse, Claudia P Rodriguez Alas, Mareeha Kamran, et al. Social protection and jobs responses to covid-19. 2022.

[25] Adama Bah, Samuel Bazzi, Sudarno Sumarto, and Julia Tobias. Finding the poor vs. measuring their poverty: Exploring the drivers of targeting effectiveness in indonesia. *The World Bank Economic Review*, 33(3):573–597, 2019.

[26] Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324, 2022.

[27] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[28] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017.

[29] Talip Kilic, Umar Serajuddin, Hiroki Uematsu, and Nobuo Yoshida. Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity. *World Bank Policy Research Working Paper*, (7951), 2017.

[30] IMF. World economic outlook database. `https://www.imf.org/en/Publications/WEO/weo-database/2023/April/download-entire-database`, 2023.

[31] Pascale Schnitzer and Quentin Stoeffler. Targeting for social safety nets: Evidence from nine programs in the sahel. *Available at SSRN 4017172*, 2022.

[32] Dean Karlan and Bram Thuysbaert. Targeting ultra-poor households in honduras and peru. *The World Bank Economic Review*, 33(1):63–94, 2019.

[33] Nina Rosas, Mariana Pinzón-Caicedo, and Samantha Zaldivar. Evaluating tanzania's productive social safety net: targeting performance, beneficiary profile, and other baseline findings. *Washington (DC): World Bank*, 2016.