Geographic Micro-Targeting of Social Assistance with High-Resolution Poverty Maps

Isabella Smythe[a] **and Joshua E. Blumenstock**[b,1]

[a]School of International and Public Affairs, Columbia University, New York, USA; [b]School of Information, University of California at Berkeley, Berkeley, USA

This manuscript was compiled on January 13, 2022

Hundreds of millions of poor families receive some form of targeted social assistance. Many of these anti-poverty programs involve some degree of *geographic targeting*, where aid is prioritized to the poorest regions of the country. However, in many low-resource settings, policymakers lack the disaggregated poverty data required to make effective geographic targeting decisions. Using several independent datasets from Nigeria, this paper shows that high-resolution poverty maps, constructed by applying machine learning algorithms to satellite imagery and other non-traditional geospatial data, can improve the targeting of government cash transfers to poor families. Specifically, we find that geographic targeting relying on machine learning-based poverty maps can reduce errors of exclusion and inclusion relative to geographic targeting based on recent nationally-representative survey data. This result holds for anti-poverty programs that target both the poor and the extreme poor, and for initiatives of all sizes. We also find no evidence that machine-learning based maps increase targeting disparities by demographic groups, such as gender or religion. Based in part on these findings, the Government of Nigeria used this approach to geographically target emergency cash transfers in response to the COVID-19 pandemic.

Poverty | Targeting | Satellite Imagery | Nigeria

Hundreds of millions of poor and vulnerable families benefit from some form of targeted social assistance (1). Just since the onset of the COVID-19 pandemic, over 3,300 new targeted social protection programs have been launched (1).

A key factor in the success of any anti-poverty program is the degree to which it is accurately *targeted* (2). When truly poor families don't receive benefits (*errors of exclusion*), or when non-poor families do receive benefits (*errors of inclusion*), this undermines the effectiveness of the policy (3).

Unfortunately, many governments in low- and middle-income countries (LMICs) lack recent, reliable data on where poverty is, and where it isn't (4). While most LMICs have access to poverty data that provides comprehensive coverage at the largest administrative subdivision (e.g., the state level in Nigeria, comparable to the state level in the USA), coverage is far less complete at the third administrative subdivision (e.g., the ward level in Nigeria, comparable to municipalities in the USA). In Nigeria, the most recent Demographic and Health Survey (DHS) reaches households in only 13.8% of wards. This problem is present across LMICs: in Peru, for example, 32.0% of the comparable administrative units are covered, and in Indonesia just 16.1%. In practice, this incomplete coverage means that a geographically targeted program must either rely on potentially inaccurate and outdated poverty maps, or accept the efficiency losses of targeting larger administrative units.

In this paper, we ask the question, *Can fine-grained poverty maps, produced by applying deep learning algorithms to high-resolution satellite imagery, improve the accuracy of geographically targeted anti-poverty programs?* Our results are based on analysis done in a high-stakes policy environment, to help the Government of Nigeria determine its emergency COVID-19 response strategy.

Our main results evaluate different geographic targeting mechanisms available to the Nigerian government, which are shared by many policymakers in LMICs. Specifically, we compare the targeting outcomes that would result from using high-resolution machine learning (ML)-based poverty maps to those that would result from using a recent nationally representative household survey (which we refer to as the survey-based "benchmark"). Both approaches are evaluated using a nationally-representative survey of 22,104 Nigerian households; this evaluation data was independently collected and not used to train the ML-based approach or to guide the survey-based benchmark.

We find that the ML-based poverty maps are at least as accurate as the benchmark in targeting benefits to the poor (i.e., those with consumption below the poverty line) and to the extreme poor (consumption below half the poverty line), in regions where benchmark data are available. We also document the main advantage of the ML-based maps, which is that they allow for accurate micro-targeting in all administrative subdivisions of the country—including subdivisions where benchmark data do not exist. This is important, because the survey benchmark does not contain data for 86.2% of Nigerian Wards (the Admin-3 unit) and 18.5% of Local Government Areas (the Admin-2 unit). We document how the accuracy and complete coverage of the ML-based maps make it possible to design a more disaggregated geographic targeting policy than would be possible with survey data alone. This disaggregation directly translates to a higher fraction of benefits being allocated to the poor and extreme poor.

Significance Statement

Many anti-poverty programs use geographic targeting to prioritize benefits to people living in specific locations. This paper shows that high-resolution poverty maps, constructing with machine learning algorithms, can improve the geographic targeting of benefits to the poorest members of society. This approach was used by the Nigerian government to distribute benefits to millions of extreme poor. Since high-resolution poverty maps are now globally available, these results can inform the design and implementation of social assistance programs worldwide.

J.E.B. supervised the project. J.E.B. and I.S. designed the study and wrote the manuscript. I.S. conducted analysis.

[1]To whom correspondence should be addressed. E-mail: jblumenstock@berkeley.edu

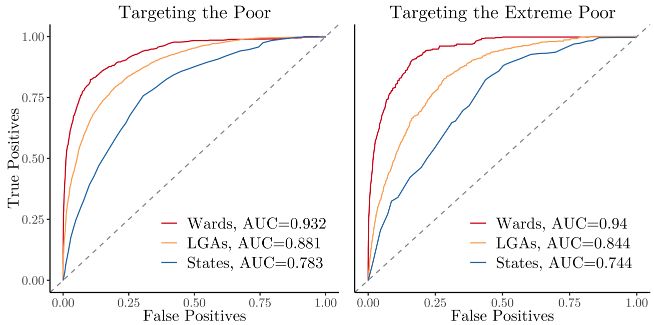

Fig. 1. Targeting performance of policies at different administrative units

Notes: ROC curves show the performance of geographic targeting policies designed at the state (admin-1), LGA (admin-2), and ward (admin-3) level, where all households in a targeted administrative unit receive full benefits and households in untargeted units receive no benefits. The targeting of an administrative unit is determined based on the average wealth of the unit, as calculated from the NLSS survey data. Accuracy is calculated based on the portion of true poor households who are targeted, where poverty status is determined based on the NLSS.

In addition, we assess the fairness of ML-based targeting with respect to several different demographic subgroups. This is to address the concern that targeting approaches that are agnostic to recipients' demographics may over- or under-target certain groups (e.g., female versus male heads of household) (5, 6). Comparing the demographic parity of ML-based and survey-based targeting approaches along several dimensions, and find that ML-based targeting does not decrease fairness overall.

These results build on prior work that develops methods for the construction of high-resolution poverty maps (7–9). However, our focus is different, and more practical. We take the output of prior work (the high-resolution poverty maps) as the input to our analysis, and show how such maps can improve the outcomes of a real-world social assistance program. In January 2021, the Nigerian government chose this approach to guide the expansion of cash transfers to the urban poor (10); our hope is that this analysis can help encourage future efforts to integrate recent innovations in machine learning into humanitarian relief applications.

Results

Benefits of Disaggregation. Our first intuitive result highlights the value of geographic disaggregation in the design of geographic targeting policies. This analysis is shown in Figure 1, where we compare targeting performance at different aggregation levels using hypothetical optimal targeting data. Optimal targeting is simulated by using the same survey data to both perform and evaluate targeting. This allows us to approximate how effectively targeting can be conducted when the true underlying distribution of poverty is known. However, it is important to note that this is a hypothetical exercise: no dataset exists that would allow optimal targeting in practice.

The left panel of Figure 1 displays the ROC curves where the objective is to provide benefits to the poor (daily consumption below \$1.05); the right panel provides the ROC curves for the objective of targeting the extreme poor (daily consumption below \$0.57). Substantial increases in area under the curve are produced as the targeting policy shifts from states (the largest administrative unit) to LGAs (the intermediate administrative unit), and from LGAs to wards (the smallest administrative unit). These findings are consistent with work done in other contexts to document the benefits of spatial disaggregation in geographic targeting (11–13): intuitively, programs targeting smaller administrative units are able to more precisely direct benefits to the poorest regions than programs targeting larger ones.

Coverage and Accuracy of ML-Based Poverty Maps. Our second set of results contrasts the coverage of ML-based poverty maps with survey-based alternatives, and compares the accuracy of these two approaches at different spatial scales.

The difference in coverage between survey- and ML-based poverty maps is evident in Panel A of Figure 2, which shows the two versions of Nigerian poverty maps side-by-side, at different levels of geographic aggregation. Grey areas indicate administrative units where no surveys occurred in the benchmark dataset, a nationally-representative DHS household survey of 40,427 households, conducted in 2018. At the state level (Row A), both maps have complete coverage; however, at the LGA level (Row B), the survey-based map loses 18.5% of LGAs, and at the finest level (Row C), surveys cover only 13.8% of all wards in Nigeria. A full tabulation of these results are also shown in the first two columns of Table 1.

The better coverage of ML-based poverty maps does not come at the expense of accuracy. Rather, we find that the ML-based poverty maps measure the spatial distribution of poverty with approximately the same accuracy as the benchmark survey. This can be seen in Panel B of Figure 2, which measures the accuracy of survey-based and ML-based poverty maps using a third, independent source of ground truth data, Nigeria's National Living Standards Survey (NLSS) of 22,110 households, conducted in 2018-2019. At all levels of spatial disaggregation, the correlation with NLSS is similar. Note that we do not expect the ML-based estimates to outperform the DHS-based estimates, since the DHS data were used to train the ML-based model (see *Materials and Methods*, Section C.1). Rather, the main advantage of the ML-based maps is that they allow accurate extrapolation of wealth estimates into the large number of regions not surveyed by the DHS.

We further find that correlations with ground truth for both DHS and ML-based poverty maps increase when we consider only the regions where the evaluation (NLSS) data are most reliable. (This analysis is intended to address one limitation of our empirical setting, noted in *Materials and Methods*, Section E, which is that the ground truth (NLSS) data used to evaluate performance are incomplete). These results are shown in the last two columns of Table 1, which reports the correlation between the two poverty maps with ground truth estimates from the NLSS. While rows 1, 2, and 6 echo the results shown in Figure 2B, the other rows indicate the correlation in specific subsets of the administrative units. In particular, we find that the performance of models evaluated using all of the LGA data (Row 2) is inferior to that of models using only data from LGAs with at least 30 households in the NLSS data (Row 3). This effect is even stronger in Panel C, when we compare the analysis of all wards (Row 6) to that of wards with at least 20 households (Row 7). As expected, both DHS and ML-based poverty maps are more strongly correlated with the NLSS validation data when regions with the fewest households surveyed are excluded.

Perhaps most important, we find that the ML-based esti-

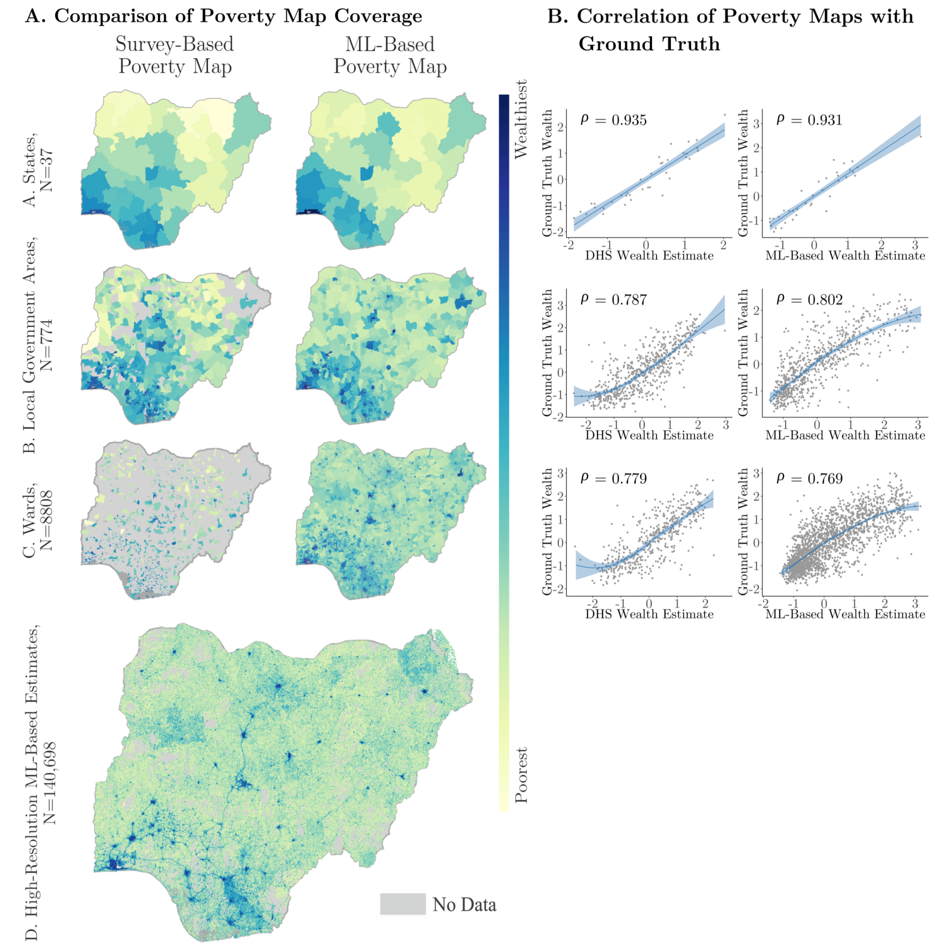

Fig. 2. Coverage and correlations of ML and benchmark poverty maps to NLSS-estimated ground truth poverty maps

Notes: Panel A compares the coverage and estimates of traditional survey-based poverty maps (left column) and ML-based poverty maps (right column) at the three different administrative levels: State (A), Local Government Area (B), and Ward (C). Regions without data shown in gray. Bottom figure (D) shows the high-resolution ML-based estimates, prior to aggregation. For privacy reasons, high-resolution poverty estimates are not generated for grid cells with fewer than 10 inhabitants. Panel B compares the ML and survey benchmark (DHS) wealth estimates of each administrative unit against the NLSS ground truth estimate of that unit's wealth. Pearson's correlation coefficients reported across all relevant units. Fewer observations exist in Panel B because not all LGAs and wards contain households that were surveyed in the DHS.

mates remain accurate even when evaluated in regions where no DHS surveys occurred. The accuracy of ML-based estimates in regions *not* covered by the DHS (but present in the NLSS ground truth and ML-based estimates) can be seen in rows 4-5 and 8-9 of Table 1. For instance, comparing rows 6 and 8, we see that the correlation between the ML-based estimates and ground truth is very similar (0.77 vs. 0.76). There is a slight attenuation in accuracy at the LGA level (row 2 vs. row 4), but this is also likely due to the fact that the NLSS validation data is sparser in regions with no DHS data. Thus, we re-calculate these correlations removing regions with little NLSS data (rows 5 and 9); the gap in accuracy at the LGA level shrinks (row 3 vs. row 5) and disappears at the ward level (row 7 vs. row 9). Overall, there is little evidence that the performance of ML-based maps deteriorates in regions where training data was unavailable.

Results of National Targeting Simulations. The third set of results, which are likely the ones most relevant to policymakers, compare targeting outcomes using the ML-based wealth estimates to targeting outcomes using survey benchmark wealth estimates. The analysis is based on simulations of ward-level geographic targeting, where all households in selected wards receive an equal benefit, and no households in unselected wards receive benefits. The data and methods used to construct poverty maps from the ML-based and survey-based data sources are described in *Materials and Methods*, Section C. The details of the targeting simulations used to evaluate both methods are provided in *Materials and Methods*, Section D.

To summarize the main finding: using a variety of different methods for evaluating targeting performance, we find that the ML-based poverty maps would deliver a higher proportion of benefits to the poorest people in Nigeria than the survey-based benchmark. This is true whether the goal of targeting is to provide benefits to the poor (defined as those consuming less than US $1.05 per day) or the extreme poor (consuming less than US $0.57 per day).

The ROC curves in Figure 3 Panel A compare ward-level

	Number of Regions with Estimates		Correlation with Ground Truth Truth Estimates	
	DHS	**ML-Based**	**DHS**	**ML-Based**
Panel A: State-level correlations				
1. All states	37	37	0.935	0.931
Panel B: LGA-level correlations				
2. All LGAs	597	706	0.787	0.802
3. > 30 ground truth households	300	371	0.839	0.863
4. LGAs with no DHS data		109		0.713
5. > 30 ground truth households and no DHS		38		0.812
Panel C: Ward-level correlations				
6. All wards	464	2016	0.779	0.769
7. > 20 ground truth households	95	242	0.894	0.870
8. Wards with no DHS data		1552		0.759
9. > 20 ground truth households and no DHS		147		0.871

Table 1. Coverage and accuracy of different approaches to constructing poverty maps in Nigeria.

Notes: First two columns indicate the number of administrative units for which data exist in the 2019 NLSS ground truth *and* the 2018 DHS survey (column 1) or the ML-based estimates (column 2). Last two columns indicate the Pearson correlation between wealth estimates generated from the ground truth (NLSS) and the DHS survey (column 3) or ML-based estimates (column 4). Correlations are measured across administrative units (i.e., not across households), using NLSS household weights for aggregation at the state level but not at the LGA or Ward level. The three panels indicate different levels of spatial of aggregation of wealth estimates. Rows 3, 5, 7, and 9 restrict analysis to administrative units where the NLSS ground truth contains a minimum of 20-30 households (to remove high variance observations from the ground truth estimate). Rows 4, 5, 8, and 9 evaluate the ML-based estimates on the subset of administrative units where no DHS data exist.

geographic targeting performance using the ML-based maps (AUC=0.87) to ward-level performance using the survey benchmark (AUC=0.81). We also include, for reference, the performance of an "oracle" strategy (AUC=0.93), which indicates the optimal performance that could be achieved with a purely geographic targeting approach. The survey benchmark shown in Figure 3 interpolates wealth estimates into missing wards (see *Materials and Methods*, Section E) to ensure that every ward has a non-zero probability of receiving benefits. In results not shown in the figure, we measure the performance of a survey-based approach that is evaluated only in the 13.8% of wards with DHS data (AUC=0.87). This approach performs similarly to the ML-based approach, but could not be feasibly implemented because it would leave 86.2% of wards ineligible for benefits. Our findings are similar when we evaluate targeting based on the proportion of transfers reaching the extreme poor rather than the poor: ward-level targeting using the ML-based estimates (AUC=0.86) improves on the survey benchmark (AUC=0.80), and performs as well as the DHS when the DHS is only evaluated in the 13.8% of DHS wards (AUC=0.86, not shown).

We find that ML-based maps can improve upon the survey-based benchmark for anti-poverty programs of all sizes. Figure 3 Panel B shows the fraction of transfers going to the poor and extreme poor as the number of beneficiaries increases. We measure program size as a fraction of the total number of poor in Nigeria (currently estimated at 73.5 million). In the left subfigure, the ML-based map performs better than the benchmark irrespective of the size of the program. On the right subfigure, the ML-based map outperforms the benchmark for all extreme-poverty program sizes except those targeting a population of between 11.8 and 25.7 million people.

To more concretely illustrate how the improvements in targeting accuracy from using the ML-based maps translate into better policy outcomes, Table 2 shows targeting precision at

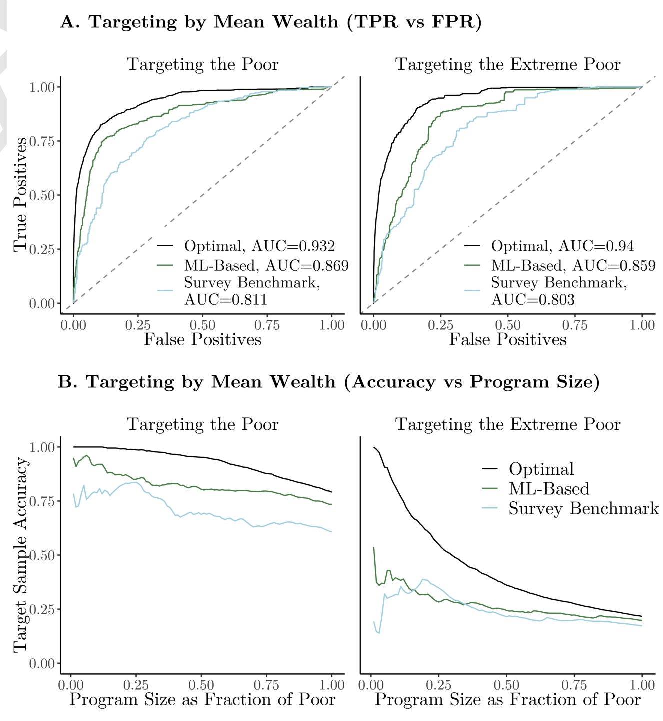

Fig. 3. Ward-level targeting performance

Notes: Three different datasets are used to identify the poorest wards; all residents of the selected wards are then targeted. Panel A shows ROC curves based on whether the NLSS households in targeted wards are poor (left) or extreme poor (right). Panel B shows the fraction of program benefits going to the poor (left panel) and extreme poor (right panel) as the size of the anti-poverty program varies.

	Precision at 10% Recall		
Targeting Approach	**Targeting the Poor**	**Targeting the Extreme Poor**	**Coverage**
Panel A: States			
Optimal (NLSS)	0.806	0.277	100%
ML-Based	0.682	0.118	100%
DHS-Based	0.795	0.218	100%
Panel B: LGAs			
Optimal (NLSS)	0.979	0.553	91.2%
ML-Based	0.855	0.390	100%
DHS-Based	0.824	0.301	100%
DHS Upper Bound	0.840	0.359	81.5%
Panel C: Wards			
Optimal (NLSS)	1.000	0.976	22.9%
ML-Based	0.919	0.406	100%
DHS-Based	0.793	0.311	100%
DHS Upper Bound	0.923	0.376	13.8%

Table 2. Precision at 10% recall

Notes: First two columns indicate the fraction of transfers going to poor (extreme poor) individuals when the program budget allows for 10% of the poor (extreme poor) to be targeted. Third column indicates the proportion of each administrative unit for which the relevant dataset provides estimates (e.g., the NLSS conducted surveys in 91.2% of LGAs and 22.9% of wards). Optimal (NLSS) targeting uses the ground truth data to select the poorest administrative units for benefits. ML-based targeting selects units based on the average estimated wealth of those units. DHS-based targeting selects units based on the average wealth of DHS households in that unit, or an interpolated wealth estimate. DHS Upper Bound evaluates targeting performance only in units where DHS surveys occur.

10% recall — i.e., the fraction of transfers that reach the poor when 10% of the poor are targeted. In this Table, we observe a similar pattern as was shown in Figure 1: that targeting performance generally increases as smaller administrative units are targeted (i.e., when results in Panel A are compared to Panel B, and when Panel B is compared to Panel C).*

Most important, we find that the ML-based approach outperforms the benchmark at all levels of geographic targeting except the state level (which would not be a viable approach to geographic targeting in Nigeria, given how large each state is). In targeting the poor, the ML-based approach increases precision relative to the benchmark from 0.82 to 0.86 at the LGA level and from 0.79 to 0.92 at the ward level. In targeting the extreme poor, the increase is from 0.30 to 0.39 at the LGA level and from 0.31 to 0.41 at the ward level.

We also include an upper bound performance estimate for DHS, which evaluates targeting performance only in the 81.5% of LGAs and 13.8% of wards covered by DHS. ML-based poverty maps outperform this upper bound on DHS maps at the LGA level and for targeting the extreme poor at the ward level; the upper bound DHS performs slightly better for targeting the poor at the ward level (precision of 0.923 vs. 0.919). While this accuracy could not be attained with DHS in practice because so many regions lack data, these findings illustrate that the targeting performance of ML-based maps is comparable even to best-case DHS performance.

These increases in precision directly translate into reductions in errors of exclusion and inclusion. For instance, if we compare two geographically-targeted anti-poverty programs that each provide transfers to 7.3 million individuals (i.e., 10% of Nigeria's poor population), the best ML-based approach (ward-level targeting) would correctly target 6,750,920 individuals; 66,735,620 poor individuals would not receive transfers and 597,734 non-poor individuals would be incorrectly included. DHS-based ward-level targeting would correctly target 5,787,802 individuals; 67,698,738 would be incorrectly excluded and 1,560,852 would be incorrectly included. In other words, the ML-based approach would reduce exclusion errors by 1.4% and would reduce inclusion errors by 61.7%, resulting in nearly a million poor individuals receiving aid who otherwise would not have.†

Our finding that ML-based targeting outperforms the survey benchmark is robust to several alternative approaches to targeting. Thus far, performance has been evaluated based on a method's ability to target regions with low average (mean) wealth. When targeting is instead conducted based on median wealth, ML-based maps improve AUC over survey-based maps from 0.808 to 0.863 for targeting the poor, and from 0.803 to 0.854 for targeting the extreme poor. Similar performance is observed for targeting based on the fraction of households in the ward that are (extreme) poor: AUC for ML-based maps is 0.861 for targeting the poor versus 0.802 for survey-based maps, and 0.835 versus 0.774 for targeting the extreme poor. See Appendix for full ROC plots.

Targeting Fairness and Demographic Parity. Our final set of results explore the extent to which different targeting approaches lead to a "fair" distribution of resources, where fairness is assessed based on statistical parity. This is motivated by the fact

*Note that while targeting performance increases with spatial disaggregation in Table 2, we earlier saw in Table 1 that the correlation between the NLSS ground truth and both the ML-based and DHS-based poverty maps *decreased* with spatial disaggregation. This illustrates a bias-variance tradeoff, where the smaller units of analysis imply fewer households are available to calculate the average wealth.

†The analysis in Table 2 is based on wards where both DHS and NLSS surveys contain at least one household (N=464). There are also a large number of wards for which DHS does not contain data but NLSS does (N=1,552). In these regions, it is possible to evaluate ML-based targeting performance but not the survey-based benchmark. We focus on wards where all data are available to facilitate direct comparisons between the DHS and ML model. In analysis not reported in Table 2, we find that the targeting accuracy of the ML-based approach is largely unchanged when evaluatd on the full set of 1,552 wards with NLSS data. For targeting the poor, AUC remains virtually unchanged (0.867, versus 0.869 for DHS wards only). For targeting the extreme poor, we observe a slightly larger decline (to 0.82 from 0.86). Precision at 10% recall is unchanged at 0.92 for targeting the poor, and declines slightly from 0.41 to 0.39 for targeting the extreme poor.

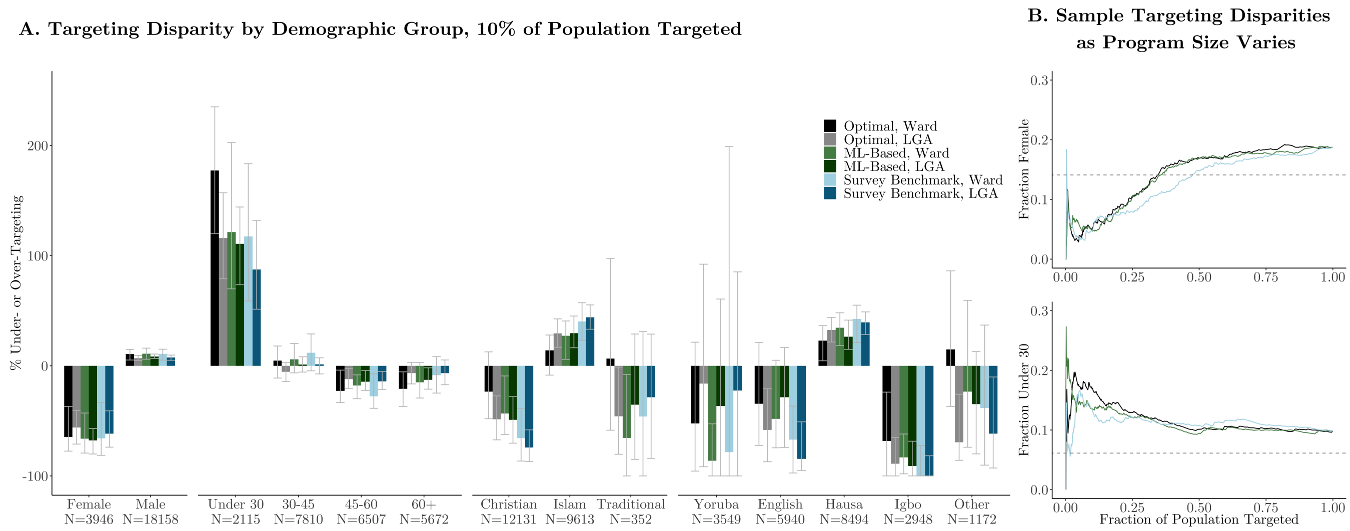

Fig. 4. Comparison of targeting fairness for selected demographic groups (assigned based on head of household)

Notes: Under perfect individual targeting, the fraction of transfers going to members of a demographic group would be equal to the fraction of total poor households belonging to that demographic group. Panel A shows the percentage difference between the number of households in each demographic group expected to receive transfers and the number that actually receive transfers, when 10% of the population is targeted. Error bars show bootstrapped 95% confidence intervals. Panel B show how the fraction of transfers going to sample subgroups vary as a function of program size, as a fraction of total population. Results pictured are for ward-level targeting.

that a singular focus on the accuracy of targeting (at reaching the poor) might inadvertently concentrate benefits toward (or away from) specific, potentially marginalized or underserved, subgroups of the population (5, 14–16). We note three results.

First, geographic targeting can create demographic disparities – likely due to the fact that different subgroups of the population concentrate in specific geographic areas. These results can be seen in Figure 4 Panel A, which quantifies the difference in the percentage of households of a certain group that are expected to receive transfers (based on the percent of that group that are truly poor) and the percentage of households of that group that do receive transfers according to a specific targeting method. In the figure, a large number of demographic sub-groups (sets of bars) are statistically over- or under-targeted irrespective of the targeting methodology (indicated by bar color). For instance, even under optimal geographic targeting, Hausa speakers (40.0% of Nigerians per NLSS estimates) are over-targeted and Igbo speakers (11.2% of Nigerians) are under-targeted. We also note significant under-targeting of female-headed households across all targeting strategies.

Second, spatial disaggregation has no clear effect on statistical disparities. With religion, we find that targeting smaller spatial units (i.e., wards) is marginally less disparate than targeting larger spatial units (i.e., LGAs). However, the opposite result appears when considering the age of the head of household. However, across all of these cases, the confidence intervals (indicated by the whiskers) overlap. Thus, the overall impact of disaggregation may depend on the patterns of spatial heterogeneity in the specific regions under consideration.

Third, and perhaps most relevant to the focus of this paper, we find that ML-based targeting leads to disparities that are similar in magnitude and direction as the survey-based benchmark. In Figure 4 Panel A, we see that 95% confidence intervals for ML- and survey-based targeting overlap significantly for all demographic groups. In Panel B, we see similar results when the number of people targeted varies. While parity varies slightly for different program sizes, no systematic differences between targeting approaches are apparent.

Discussion

This paper provides empirical evidence that recent advances in machine learning can improve the geographic targeting of social assistance. Our analysis, done to support the Government of Nigeria's humanitarian response to the ongoing COVID-19 crisis, indicates that programs targeted using ML-based maps direct more transfers to the poorest households than programs targeted using survey-based poverty maps. This improvement in targeting efficiency is due to the fact that the ML-based maps provide accurate estimates of the relative wealth of every administrative subdivision of the country, whereas survey data typically only cover a small fraction of all units; as a result, an ML-based approach can be designed for smaller regions while a survey-based approach can only be designed for larger regions. We also do not find evidence that ML-based poverty maps increase disparities between demographic groups in the Nigerian context.

While promising, we caution that these results should not be misconstrued to suggest that ML-based approaches should replace survey-based methods for measuring poverty. Indeed, the ML-based approach was only feasible because high-quality survey data existed to train the ML model. More broadly, household surveys capture a wide range of information, with much greater nuance, than can be clearly seen in overhead imagery, and which may not be easily modeled with machine learning (17, 18). Rather, these results suggest that the ML-based maps can provide a reliable method for geographic targeting when time and resource constraints prevent bespoke data collection – a frequent consideration in the large number of LMICs without a recent census or comprehensive social registry.

Materials and Methods

A. Related Work. This paper connects a recent strand of the applied machine learning (ML) literature to a rich literature in development economics. The most closely related ML papers explore how ML algorithms can be used to construct estimates of the spatial distribution of wealth and poverty from high-resolution satellite imagery (7–9, 18–22). Also related are papers that construct granular poverty maps from mobile phone (5, 23, 24) and social media data (25), and recent work on evaluating fairness in machine learning-based targeting approaches (5, 6). Broadly speaking, these studies match non-traditional data (such as satellite or phone data) to a survey-based ground truth measure of wealth; train machine learning methods to predict wealth from the non-traditional data; and use the trained model to predict wealth estimates in regions where no ground truth data exists (26).

The second literature, which has a rich history in development economics, studies the targeting of social assistance and government transfers. This body of work provides theory and empirical evaluations of the different *targeting mechanisms* that are commonly used to determine eligibility for benefits (2, 3). The crux of the problem is that central governments often lack recent, reliable, and comprehensive data on the living conditions of each family (4). Thus, a variety of common targeting mechanisms exist to help direct benefits to the neediest households: self-targeting, where benefits are available to anyone, but there is some "ordeal" involved in registering such that only those with the greatest need will choose to participate (27, 28); proxy means tests (PMT), where wealth is estimated based on a small number of easily observed assets and housing characteristics (29, 30); community-based targeting (CBT), where communities are asked to identify their neediest members (31, 32); and geographic targeting, where resources are channeled to the regions with the highest levels of poverty (33–35). A key result from this literature, which we extend in this paper, is that significant efficiency gains can be achieved by targeting small administrative units rather than larger ones (11–13).

This paper connects these two historically disjoint literatures by examining whether novel ML-based poverty maps can improve the targeting of social assistance and humanitarian aid. We build on prior work by Yeh et al. (8), who discuss the potential for ML-based maps in program targeting, but stop short of analyzing a real-world policy decision and do not compare the ML approach to status quo alternatives. We also build on Aiken et al. (36), who show how mobile phone metadata can improve targeting outcomes in Afghanistan. Relative to (36), our approach is likely most relevant in contexts where mobile phone data is not publicly available, or when policy applications require a geographic approach to targeting.

B. Targeting Context: Nigeria. Our analysis was motivated by a specific request for assistance from the Government of Nigeria, who was working with the World Bank to design an emergency social assistance program in response to the COVID-19 crisis. At the onset of COVID-19, there was no single, comprehensive social registry that would allow them to identify the individuals or households with the greatest need for assistance – and in the middle of the pandemic, it was impractical to go door to door to collect this information. Thus, they were interested in evaluating different approaches to geographic targeting.

Nigeria is home to roughly 211 million people, making it the seventh most populous country in the world. Geographically, Nigeria has three different levels of administrative subdivisions (see Figure 2): 37 states (Admin-1), which are subdivided into a total of 774 local government areas (LGAs, Admin-2), which in turn are subdivided into a total of 8808 wards (Admin-3). However, in early 2020, the best poverty data available to the Government of Nigeria could only provide estimates of *state-level* poverty; it did not allow for estimates of rates of poverty at the LGA or ward level.

Based in part on the analysis described in this paper, the Government elected to use our high-resolution poverty maps to target the COVID-19 Rapid Response Registration (RRR) Cash Transfer Project, which began disbursing benefits to the first of an eventual one million recipients in mid-January 2021 (10). The RRR program is designed specifically to help the urban poor; see Appendix for separate evaluation of targeting outcomes for urban areas only.

C. Primary Data Sources and Poverty Map Construction.

C.1. ML-Based Poverty Maps. The high-resolution poverty maps shown in Figure 2 Panel A Row D are constructed using a machine learning approach described in greater detail in Chi et al.(9), which follows an approach similar to that first proposed by Jean et al. (7). To summarize: we start with ground truth survey data from Nigeria's 2018 DHS (see next section for details), which provides information on the wealth of 40,427 households across Nigeria. These "labels" are matched, using geographic markers in the survey dataset, to a rich set of non-traditional geospatial data, including features derived from high-resolution satellite imagery using a convolutional neural net, as well as mobile connectivity data and other topological data. We use a gradient boosted decision tree to predict the labels from the satellite and other geospatial features, using spatially-stratified cross validation. The fitted model is then used to predict the wealth of every 2.4km gridded region in the country of Nigeria.

To produce estimates of the wealth and poverty of the different administrative units of Nigeria (right column of Figure 2 Panel A, A-C), the 2.4km estimates are aggregated using population weights, where the population of each 2.4km grid cell is generated using population estimates from Humanitarian Data Exchange (37). Specifically, the wealth estimate of administrative unit i is calculated as:

$$W_i = \left(\frac{1}{\sum_{t \in T} I(t,i) p_t} \right) \sum_{t \in T} I(t,i) p_t w_t \qquad [1]$$

Where T is the set of all 2.4km satellite tiles, p and w approximate the population and wealth of tile t, and I gives the fraction of tile t that intersects administrative unit i. Because wealth indices are relative and have no meaningful units, they are normalized at each administrative level to have a mean of zero and standard deviation of one.

C.2. "Benchmark" Poverty Maps, from DHS Survey. As a benchmark against which we compare the targeting outcomes of the ML-based maps, we construct a set of poverty maps using data from a recent, nationally-representative household survey. Specifically, we obtain the micro-data from Nigeria's 2018 DHS (38). The DHS is a standardized household survey funded by the U.S. Agency for International Development; the 2018 Nigerian DHS conducted surveys with 40,427 households in 1,360 unique locations across the country. The survey instrument contains detailed questions about the socioeconomic conditions of each household, including a *wealth index*, which provides a scalar measure of the wealth of that household, relative to all other surveyed households.[‡] We also observe the approximate location of each DHS household, where the DHS groups households into clusters (roughly equivalent to villages in rural areas and neighborhoods in urban areas) and provides the geocoordinates of the centroid the cluster of households, after adding up to 5km of jitter to preserve the privacy of individual households.

To construct poverty maps from the household survey data (as shown in the left column of Figure 2, Panel A), we calculate the average wealth index of all surveyed households located in the relevant the administrative unit. For this process, we obtained shapefiles and urban/rural classifications for each administrative unit from the World Bank. Both the NLSS and DHS surveys were designed to provide estimates of population characteristics that are representative at the state level, and each household has an associated survey sampling weight. Thus, for the state-level poverty maps, we use this sampling weight to calculate the weighted average wealth index of all households in the state. When constructing estimates of the wealth of the LGA and ward, we take the simple average of all households in the relevant administrative unit, since

[‡] The wealth index is construct as the first principal component of a vector of assets and household characteristics: air conditioner, animal-drawn cart, bank account, bed, bicycle, boat with a motor, canoe, car or truck, chair, computer, cupboard, electric iron, electricity, fan, generator, landline, motorbike, main floor material, main roof material, main wall material, mobile telephone, motorcycle or scooter, number of members per sleeping room, owns a house, owns land, radio, refrigerator, sofa, source of drinking water, table, television, type of toilet facility, type of cooking fuel, and watch.

the household survey weights were not intended to provide LGA- or Ward-representative inferences.

C.3. Ground Truth Evaluation Data, from NLSS. To evaluate the performance of targeting using the ML-based poverty maps and the survey-based poverty maps, we obtain a separate, independent source of "ground truth" data on living standards in Nigeria. This is the 2019 Nigerian Living Standards Survey (NLSS), an ambitious household survey financed by the World Bank and implemented by Nigeria's National Bureau of Statistics (39). The survey was conducted with 22,110 households in 22,104 unique locations. For each household in this dataset, we observe the exact geocoordinates, as well as a rich set of questions about socioeconomic conditions. We use the responses to these questions to construct a DHS-style wealth index for each NLSS household.[§] The NLSS, which has not yet been publicly released, was never used to train the ML-based poverty maps, and did not influence the collection of the DHS data; it thus provides an objective and out-of-sample means for validating the alternative approaches to geographic targeting.

D. Targeting Simulations. We simulate the geographic targeting of anti-poverty programs in Nigeria using two different approaches – one based on the ML-Based poverty maps (derived from satellite imagery) and the other based on the survey-based benchmark (derived from the 2018 DHS survey). The performance of these two approaches is evaluated using ground truth data derived from the 2018-2019 NLSS, which is thought to be the most comprehensive and up-to-date survey in Nigeria.

Specifically, we assess targeting performance based on the proportion of transfers that would reach *poor* and *extreme poor* households under different approaches to geographic targeting. Using the Nigeria-specific World Bank poverty line of 377 Nigerian Naira per person per day ($1.05 USD in 2018), we estimate from the NLSS data that 40.5% of the population is poor (consumption below the poverty line), and 8.2% is extremely poor (consumption below half the poverty line). Since neither the DHS nor the ML-model provide direct income or consumption data, we use these percentage thresholds to identify the poor and extreme poor using the wealth information provided by the DHS and ML-model.[¶] Thus, households with wealth indices in the bottom 8.2% are classified as extreme poor, and the bottom 40.5% as poor - i.e., poor is inclusive of extreme poor.

Based on these thresholds, we can classify each household in the NLSS evaluation data as extreme poor, poor, and non-poor; we can likewise calculate the fraction of households in each administrative unit that fall into each category of poverty. When calculating state-level poverty rates, we use the survey sample weights; no weights are used to calculate poverty rates at the LGA and Ward level.

We then simulate geographic targeting policies at the state, LGA, and ward level, where the targeting is determined using estimates from the ML-based poverty map ("ML-based method"), the DHS-based poverty map data (the "benchmark method"), and the NLSS-based poverty map (the "oracle method"). Under each approach, we assume that 100% of the households within a given administrative unit will receive the same benefit, which is how the Nigerian government originally envisioned this program would be implemented. Note that this implies that even the oracle method, where geographic targeting is determined by the same dataset used to evaluate targeting, will not be perfectly accurate. This is because there exist non-poor households in even the poorest wards of Nigeria, so providing benefits to everyone in the poorest wards will result in errors of inclusion. Likewise, errors of exclusion will occur whenever poor individuals live in wealthy regions – even if the targeting data can perfectly separate wealthy from poor regions.

While ward-level targeting theoretically has a higher upper bound on performance, estimates at the LGA and state level can draw on more data and thus may be more accurate. It is useful to analyze targeting performance for these administrative units as well to quantify the trade-off between greater targeting precision (at the ward level) and potentially more accurate wealth estimates (at the LGA/state level).

D.1. Alternative targeting criteria. In addition to the poverty maps used in our main specification, which estimate mean poverty of each administrative unit, we create two additional ward-level poverty maps from each data source as a robustness check. The first estimates median poverty. Because the NLSS and DHS sample weights are not representative at the ward level, we use the poverty level of the unweighted median household in each ward. Median wealth is calculated from the ML-based map using the median of the wealth estimates of each one kilometer satellite tile, weighted by estimated population in that tile. The second additional map estimates the fraction of households in each ward that are (extreme) poor. NLSS and DHS households are classified as (extreme) poor based on the percentile of their wealth index (see Section D). The unweighted fraction of households in each ward that are (extreme) poor is used as the targeting criteria for NLSS and DHS. For the ML-based map, each one kilometer satellite tile is classified as (extreme) poor based on the percentile of its estimated wealth index. The fraction of people in each ward that are (extreme) poor is calculated as the fraction of people who live in satellite tiles classified as (extreme) poor. Note that unlike for mean and median wealth, separate maps are generated for the fraction of poor households and the fraction of extreme poor households. This means that the rank order of wards to target may vary depending on which of the two criteria (poor or extreme poor) are optimized for.

E. Issues of Incomplete Survey Coverage. One limitation of the surveys — both the DHS data used to construct benchmark estimates of poverty, and the NLSS data used to evaluate targeting performance — is that the data are sparse. As we discuss in greater detail below, only 13.8% of Nigerian wards have one or more surveyed DHS households, and only 22.9% of wards have one or more household in the NLSS.

E.1. Incomplete evaluation data. While great care was taken in the design of the NLSS survey to ensure that the survey population was representative of the full population of Nigeria (and also representative of each state), we are unable to evaluate the performance of the ML-based and survey-based poverty maps in the 8.8% of LGAs and 77.1% of wards where no ground truth NLSS data exists. To ensure that results are comparable for all targeting approaches, we further limit our results on targeting accuracy to the 77% of LGAs and 5% of wards where both the NLSS and DHS surveys include at least one household. When evaluating targeting for LGAs and wards, we also report results when performance is measured only on the subset of wards where the the NLSS contains at least 20 households, and the subset of LGAs where the NLSS contains at least 30 households. This effectively removes the wards and LGAs where our ground truth estimates of poverty have the highest variance.

E.2. Incomplete benchmark data. In a practical policy setting, the incomplete coverage of the survey-based poverty maps implies that those data could not be used in isolation to determine a national ward- or LGA-level geographic targeting policy. Instead, the policy would either need to be designed at the state level (where DHS coverage is complete); or some form of interpolation would be required to make decisions about LGAs and wards where data does not exist.

In the targeting simulations, we will evaluate the performance of the benchmark approach using two different methods. The first, which we consider to be a theoretical upper bound on the accuracy of the benchmark approach, considers only those administrative units where both DHS and NLSS surveys were conducted (5.3% of all wards and 76.9% of all LGAs). This should be interpreted as the performance that would be achieved in the benchmark case, if a much larger-scale survey were conducted that could reach households in 100% of wards, and was of comparable quality to the DHS survey data.

In the second method, we estimate the real-word performance of the benchmark approach by accounting for the fact that it does not measure households in all wards or LGAs. To do so, we interpolate estimates for a fraction of the wards and LGAs, and measure the resulting change in targeting accuracy. Specifically, for each LGA i,

[§] The NLSS is more detailed than the DHS, and contains a superset of the DHS asset questions. We therefore use the PCA weights from the DHS wealth index to calculate the wealth index of each NLSS household (rather than calculating a new set of eigenvectors from the NLSS data).

[¶] The best approach to measuring wealth and well-being in low- and middle-income countries is a hotly contested topic (40). To evaluate consumption poverty using information on wealth requires that the two are monotonically related.

we impute a wealth estimate as:

$$\hat{W}_{\mathrm{L}}(i) = \left(\frac{1}{A_{i,\mathrm{L}}}\right) \sum_{h \in H} (a_h w_h) \mathbb{1}\{\mathrm{S}(h) = \mathrm{S}(i),\ \mathrm{L}(h) \neq i\} \qquad [2]$$

$$A_{i,\mathrm{L}} = \sum_{h \in H} a_h \mathbb{1}\{\mathrm{S}(h) = \mathrm{S}(i),\ \mathrm{L}(h) \neq i\}$$

Here, h is a survey household in state $\mathrm{S}(h)$ and LGA $\mathrm{L}(h)$, with survey weight a_h and wealth index w_h. Intuitively, this gives the survey-weighted mean wealth of households in the same state as a given LGA, but not within the LGA itself. Analogously, ward i is interpolated as:

$$\hat{W}_{\mathrm{W}}(i) = \begin{cases} \left(\frac{1}{A_{i,\mathrm{W}}}\right) \sum_{h \in H} w_h \mathbb{1}\{\mathrm{L}(h) = \mathrm{L}(i), \mathrm{W}(h) \neq i\} & \exists h \in H_{i,\mathrm{W}} \\ \hat{W}_{\mathrm{L}}(L(i)) & \nexists h \in H_{i,\mathrm{W}} \end{cases}$$

Where $A_{i,\mathrm{W}}$ and $H_{i,\mathrm{W}}$ are defined analogously to Eq. (2). Above, $\mathrm{W}(h)$ gives the ward in which survey household h is located. Thus, if at least one household exists in the survey data that is in the same LGA as a given ward, the ward is interpolated as the simple average of all such households. Otherwise, the ward is interpolated as the survey-weighted average of all households in the same state as the ward but not the ward itself.

These interpolated estimates are generated for each ward in our sample – i.e., wards in which both the benchmark (DHS) and evaluation (NLSS) surveys contain at least one household. To approximate the real-world performance of the benchmark approach, we choose a subset of its wards at random to replace with their interpolated values. The number of wards replaced is selected so that the fraction of the sample that is interpolated is equal to the fraction (86.2%) of wards in the full country for which the benchmark survey does not contain data (and would thus require interpolated estimates in practice). We repeat this randomization process 1,000 times, and report results for the iteration with the median performance, as determined by the area under the ROC curve.

F. Estimating Demographic Parity. We estimate the "fairness" of targeting based on statistical *statistical parity* (14), which defines a "fair" allocation as one in which the fraction of households in a specific group receiving transfers is equal to the fraction of households in that group which are truly poor. We acknowledge that other notions of fairness exist and may conflict with this focus on statistical parity (41).

Our analysis assesses statistical parity for four demographic characteristics that are recorded in the NLSS survey: gender, age, religion, and language (a proxy for ethnicity, which is not recorded). We observe these characteristics just for the head of household, so our evaluation focuses on the extent to which households with a household head of a certain type are under- or over-targeted.

The fraction of households in each ward in each demographic category are estimated using NLSS data. NLSS data are also used to estimate the total fraction of poor households that belong to each demographic group. This reference statistic is calculated for the subset of wards in which targeting simulations occur (i.e., those with coverage in both the NLSS and DHS surveys); thus, it may not accurately reflect the country-level demographics of poor households.

For each targeting approach (optimal (NLSS), survey benchmark (DHS), and ML-based), we calculate the fraction of targeted households in each demographic group as the percentage of the population targeted varies, using the NLSS estimates for ward-level demographics. We also calculate a snapshot of parity for a program targeting 10% of the population. For demographic group d, we calculate the extent of over or under-targeting using:

$$\frac{\text{Fraction of targeted in } d - \text{Fraction of poor in } d}{\text{Fraction of poor in } d} \cdot 100 \qquad [3]$$

We generate confidence intervals by bootstrap sampling equation Eq. (3) 1,000 times.

Acknowledgments

We thank Tara Vishwanath and Jonathan Lain for helpful suggestions on this project. Emily Aiken, Guanghua Chi, and Raesetje Sefala provided valuable input on the empirical analysis. We are grateful for financial support from data.org and the National Science Foundation under award IIS – 1942702.

Bibliography

1. U Gentilini, M Almenfi, I Orton, P Dale, Social protection and jobs responses to covid-19. (2021).
2. D Coady, M Grosh, J Hoddinott, Targeting outcomes redux. *The World Bank Res. Obs.* **19**, 61–85 (2004).
3. R Hanna, BA Olken, Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *J. Econ. Perspectives* **32**, 201–26 (2018).
4. M Jerven, *Poor numbers: how we are misled by African development statistics and what to do about it.* (Cornell University Press), (2013).
5. E Aiken, S Bellue, D Karlan, CR Udry, J Blumenstock, Machine learning and mobile phone data can improve the targeting of humanitarian assistance, (National Bureau of Economic Research), Technical report (2021).
6. L Kondmann, XX Zhu, Under the radar–auditing fairness in ml for humanitarian mapping. *arXiv preprint arXiv:2108.02137* (2021).
7. N Jean, et al., Combining satellite imagery and machine learning to predict poverty. *Science* **353**, 790–794 (2016).
8. C Yeh, et al., Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nat. communications* **11**, 1–11 (2020).
9. G Chi, H Fang, S Chatterjee, J Blumenstock, Micro-estimates of wealth for all low- and middle-income countries. (2021).
10. NAO Nigeria, 1 million nigerians to benefit from covid-19 cash transfer programme. *Pulse Niger.* (2021).
11. M Ravallion, Poverty alleviation through regional targeting: a case study for indonesia. *The Econ. Rural. Organ. Theory, Pract. Policy* (1993).
12. DP Coady, The welfare returns to finer targeting: The case of the progresa program in mexico. *Int. tax public finance* **13**, 217–239 (2006).
13. C Elbers, T Fujii, P Lanjouw, B Özler, W Yin, Poverty alleviation through geographic targeting: How much does disaggregation help? *J. Dev. Econ.* **83**, 198–213 (2007).
14. C Dwork, M Hardt, T Pitassi, O Reingold, R Zemel, Fairness through awareness in *Proceedings of the 3rd innovations in theoretical computer science conference.* pp. 214–226 (2012).
15. A Noriega-Campero, et al., Algorithmic targeting of social policies: fairness, accuracy, and distributed governance in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* pp. 241–251 (2020).
16. S Barocas, M Hardt, A Narayanan, *Fairness and Machine Learning.* (fairmlbook.org), (2019) http://www.fairmlbook.org.
17. J Blumenstock, Don't forget people in the use of big data for development (2018).
18. A Head, M Manguin, N Tran, JE Blumenstock, Can human development be measured with satellite imagery? , 8–1 (2017).
19. B Babenko, J Hersh, D Newhouse, A Ramakrishnan, T Swartz, Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in mexico. *arXiv preprint arXiv:1711.06323* (2017).
20. R Engstrom, J Hersh, D Newhouse, Poverty from space: using high-resolution satellite imagery for estimating economic well-being (2017).
21. JL Abitbol, M Karsai, Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nat. Mach. Intell.* **2**, 684–692 (2020).
22. W Nordhaus, X Chen, A sharper image? estimates of the precision of nighttime lights as a proxy for economic statistics. *J. Econ. Geogr.* **15**, 217–246 (2015).
23. J Blumenstock, G Cadamuro, R On, Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
24. JE Steele, et al., Mapping poverty using mobile phone and satellite data. *J. The Royal Soc. Interface* **14**, 20160690 (2017).
25. M Fatehkia, et al., Mapping socioeconomic indicators using social media advertising data. *EPJ Data Sci.* **9**, 22 (2020).
26. JE Blumenstock, Fighting poverty with data. *Science* **353**, 753–754 (2016).
27. D Nichols, E Smolensky, TN Tideman, Discrimination by waiting time in merit goods. *The Am. Econ. Rev.* **61**, 312–323 (1971).
28. V Alatas, et al., Self-targeting: Evidence from a field experiment in indonesia. *J. Polit. Econ.* **124**, 371–427 (2016).
29. M Grosh, JL Baker, Proxy means tests for targeting social programs. *Living standards measurement study working paper* **118**, 1–49 (1995).
30. D Filmer, LH Pritchett, Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of india. *Demography* **38**, 115–132 (2001).
31. H Alderman, Do local officials know something we don't? decentralization of targeted transfers in albania. *J. public Econ.* **83**, 375–404 (2002).
32. V Alatas, A Banerjee, R Hanna, BA Olken, J Tobias, Targeting the poor: evidence from a field experiment in indonesia. *Am. Econ. Rev.* **102**, 1206–40 (2012).
33. JL Baker, ME Grosh, Poverty reduction through geographic targeting: How well does it work? *World development* **22**, 983–995 (1994).
34. NR Schady, Picking the poor: indicators for geographic targeting in peru. *Rev. income wealth* **48**, 417–433 (2002).
35. T Bedi, A Coudouel, K Simler, *More than a pretty picture: using poverty maps to design better policies and interventions.* (World Bank Publications), (2007).
36. EL Aiken, G Bedoya, A Coville, JE Blumenstock, Targeting development aid with machine learning and mobile phone data. (2020).
37. Continent of africa: High resolution population density maps (2020).
38. National Population Commission - NPC, ICF, Nigeria Demographic and Health Survey 2018 - Final Report, (NPC and ICF, Abuja, Nigeria), Technical report (2019).

39. NB of Statistics (NBS) Federal Government of Nigeria, Nigeria living standards survey 2018-2019, (NBS - Federal Government of Nigeria, Nigeria), Technical report (2018-2019).
40. A Deaton, Measuring poverty. *Underst. poverty*, 3–15 (2006).
41. J Kleinberg, S Mullainathan, M Raghavan, Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).