An Exploration of Targeting Leakage and Under-coverage: Evidence from Mexico's Progresa Cash Transfer Experiment*

Daniel Maggio[†] Jack Cavanagh[‡]

April 24th, 2023

Abstract

In a world of scarce resources, governments must be able to target their social programs efficiently. Over the past 20 years, governments in contexts without high-quality administrative data have increasingly used proxy means testing (PMT) to conduct this targeting due to its ease and relatively small data requirements. In this paper, we explore the potential of neural networks and satellite imagery to reduce the error rate of traditional PMT and test whether those reductions translate to improved effectiveness of programs that rely on the method. We use administrative data to conduct an ex-post targeting exercise of the Progresa cash transfer experiment in Mexico. We find that in our context, the overall gains associated with utilizing neural networks and satellite imagery are relatively small – around an additional three percentage points of variation explained. However, we decompose these gains into different error types and find that our model that overweights households towards the endpoints of the consumption distribution decreases exclusion error with a limited tradeoff in terms of inclusion error. We compare the program effectiveness of Progresa on nutrition and poverty outcomes for the full sample, those targeted by traditional PMT, and those targeted by our preferred model. We find evidence of modest effectiveness gains due to improved program targeting. These gains are centered around outcomes such as poverty intensity and consumption of grains, where we might expect social protection programs to have greater impacts among poorer households.

*This research was made possible by support from Dr. Natalia Volkow and the entire staff at the Instituto Nacional de Estadística, Geografía e Informática (INEGI). The statistics reported within do not reflect official statistics of the Federal Government of Mexico or INEGI.

[†]Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY 14853 USA; Corresponding author

[‡]School of Humanities, Arts and Social Sciences, MIT, Cambridge, MA 02142 USA

1 Introduction

The ability of governments to effectively implement programs depends on their capacity to successfully target recipients, regardless of the program's composition. No government has unlimited resources, meaning there is naturally a tradeoff between spending on beneficiaries and data collection for program targeting. This tradeoff means that cheap and accurate targeting methods are invaluable to resource-constrained governments looking to make cost-effective reductions to poverty.

In high-income countries, program targeting is typically made easier through robust, near-universal administrative data systems. For example, eligibility for income-based welfare programs can be automatically tested using high-quality data on tax returns (Mirrlees, 1971). However, in lower-income countries with limited data systems and/or state capacity, accurate targeting can be both a much more difficult and arguably more important exercise – as the proportion of the country living in extreme poverty increases, so does the necessity of being able to target developmental and humanitarian programs effectively.

To overcome this difficulty, resource-constrained governments often use one or more alternative methods to target social programs, from self- to community-based to statistical/predictive targeting (Hanna and Olken, 2018). In this paper, we hone in on the last of these methods and explore the potential of best-practice machine learning methods and satellite imagery to do two things. First, we test whether using these methods can reduce the error rate of one currently-widespread method of statistical targeting, proxy means testing (PMT). Second, conditional on our method displaying improvement over standard methods, we test whether greater targeting accuracy translates to improved effectiveness of targeted programs and briefly discuss the potential costs and benefits of implementing the improvements at scale.

To do the first, we construct neural network models, train and validate them on survey data with high-quality consumption measures on the left side of the equation, and a combination of satellite images and variables found in the much smaller set of data used to target the program on the right. We test this against a standard PMT exercise on the same data (minus the satellite images). We also vary the loss functions of both models to overweight two groups where misclassification is likely the most costly, those towards the endpoints of the consumption distribution (we refer to these as our "distance-weighted" models). To do the second, we use the trained model and data from the randomized evaluation of the Mexican conditional cash transfer program Progresa to compare the estimated average treatment effects of the program on nutrition and poverty outcomes for three subsamples: i) the full Progresa sample; ii) the set of households that would have been targeted under the method being evaluated; and iii) the set of households that received the intervention but would not have been targeted under the method being evaluated. Additionally, we estimate the treatment effects among households not included in Progresa's original sample but would have been under our methods. To do so, we leverage a secondary targeting exercise called the "densification" process to identify program recipients who were not originally targeted.

We find that all of our models explain more than half of the variation in out-of-sample consumption at the household level. While there are gains to utilizing neural networks and satellite image data, the overall gains are relatively small – around an additional three percentage points of variation explained. We decompose those gains to see how the models perform on different types of targeting errors – inclusion and exclusion – and for two different poverty lines. We find that the distance-weighted neural network results in the best inclusion-to-exclusion-error ratio and select it as our preferred model for testing treatment effects from Progresa.

The results from the Progresa exercise are largely consistent for both the set of nutrition outcomes and the set of poverty outcomes. In both cases, as has been shown before in Hoddinott and Skoufias (2004) and Skoufias et al. (2001), respectively, the program had a statistically and economically significant positive impact on a range of outcomes. This is also true across the three subsamples, though we are more interested in the differences between estimated effects. Tests for coefficient equality find some evidence of modest gains to accurate program targeting, particularly along outcomes like poverty intensity and calories from grains, where we might expect the bigger difference between the properly-targeted poor and the less poor that "leaked" into the Progresa sample.

The paper is structured as follows: section 2 provides a brief review of targeting in practice, as well as recent attempts to improve PMT methods; section 3 provides background information on the program and RCT we use to test our models; section 4 details our methods and data; section 5 provides results; and section 6 concludes.

2 Targeting in Practice

2.1 An overview of available methods

In practice, methods of program targeting can vary on multiple dimensions. The most important of these dimensions are the selection criteria and the targeting actors. On criteria, programs can be targeted toward individuals or "categorically" targeted toward groups. Most often, this targeting is conducted by program administrators, but there are also numerous programs in which communities collectively choose program recipients and programs in which recipients self-identify as eligible. Program designers face various trade-offs in choosing methodologies, and there is no one-size-fits-all approach to recipient selection. As such, programs often combine selection methods. Such is the case of Progresa, whose selection methodology we will discuss more thoroughly below.

In cases where administrators delegate the selection of recipients, targeting often occurs in one of two ways. First, some programs rely on self-targeting, in which potential beneficiaries are required to signal to the program in some way that they fall within its bounds, whether through an application, needs assessment, or direct action as in workfare programs (for example, India's NREGA scheme) (Nichols and Zeckhauser, 1982). While this form of targeting can be effective at reducing inclusion error, there are obvious concerns about optimizing the size of the cost, ensuring the program does not disproportionately exclude a certain subgroup of the intended population (e.g., persons with disabilities unable to participate in workfare programs), and advertising such that the application process is widely known (Alatas et al., 2016).

Second, administrators may delegate recipient selection to the community and ask community members to help select the potential beneficiaries for whom the program is most relevant. Included here are both traditional community-based targeting, in which community members help select the beneficiaries themselves (Karlan and Thuysbaert, 2019), as well as "indirect reporting," in which the members only provide information about indicators of poverty, such as observable assets (Alix-Garcia et al., 2019). These methods can reduce some of the potentially exclusionary trade-offs associated with self-targeting discussed above and provide a source of targeting information that is less likely to be influenced by point-in-time estimates of high-variance indicators like income. However, there are potential drawbacks, including elite capture of the targeting mechanism and differing definitions of the indicator being targeted on (Karlan and Thuysbaert, 2019).

Administrators face a trade-off between targeting groups or individuals when they conduct the targeting. In cases where governments enroll groups into programs, the selection is often geographic, targeting sub-national administrative areas. Baker and Grosh (1994) showed that this method can reduce targeting costs by relying on aggregated data but that potential gains associated with programs increase as the unit of targeting approaches the individual. In this paper, we will examine the usefulness of machine learning and satellite imagery when recipient selection is at the individual level, but there is also a growing literature showing that these methods can be useful when conducting geographic targeting in areas where survey data collection is absent or severely limited (e.g., Engstrom et al. (2022); Newhouse et al. (2022)).

Of course, groupings need not be geographic. In many cases, governments wish to provide benefits to a specific subset of the population. One readily available example of such targeting is old-age pensions such as social security in the US. In the developing country setting, South Africa's Old Age Pension Grant Program offers one well-studied example (Lund, 1993; Duflo, 2000; Case and Deaton, 1998). Categorical targeting of this kind can be useful for governments whose programs have a specific goal in mind, such as supporting an elderly population, can reduce data collection needed to select beneficiaries, and may improve transparency by having well-defined eligibility cutoffs tied to program objectives.

While the above methodologies can be well-suited for many programs, categorical targeting of general poverty reduction programs can be ineffective when poverty is well distributed across groups. Further, geographic targeting cannot account for heterogeneity in income within areas and thus has the potential to include large numbers of non-poor recipients.[1] For these reasons, program administrators often use predictive methods to target programs at a household or individual level based on easily-measured *proxies* of more difficult but informative indicators, like consumption. While there are many different types of proxy-based statistical methods, the one that we will focus on in this paper, and that we will refer to simply as "proxy means testing (PMT)" from here on out is a widespread variation in which, similar to poverty mapping methods developed by Elbers et al. (2003), weights used to calculate predicted consumption are estimated using regressions of survey-based consumption measures on observables available in a near-universal data source, like census data (Brown et al., 2016). These methods are popular because they mitigate some of the exclusionary aspects of self-targeting, remove potential elite-capture concerns of community-based targeting, and, conditional on the existence of periodic representative consumption surveys, can be cheaper to implement than either (Alix-Garcia et al. (2019); Ohlenburg et al. (2022)). As this is the method we focus on in this paper, the following paragraphs will focus on its development, tested accuracy, and recent improvements.

2.2 Testing proxy means testing and recent improvements

Proxy-means testing is widely used by governments across the world, including nations both large (e.g., Indonesia, Pakistan, Nigeria, Mexico, and the Philippines) and small (e.g., Burkina Faso, Ecuador, and Jamaica) (Hanna and Olken, 2018). Because of this wide adoption, there has been a concurrent interest in measuring how well PMT works in practice, in which scenarios it works well, and what can be done in the scenarios where it doesn't (see Brown et al. (2016); Filmer and Pritchett (2001); Grosh and Baker (1995) for a non-comprehensive

[1] As a counterpoint to this intuition, Ravallion and Wodon (1999) use data from Bangladesh to show that geographic targeting of anti-poverty programs can be useful when there are large spatial correlations in well-being unrelated to observable characteristics and structural barriers to migration.

starting list in general, and Coady (2006) and Coady et al. (2004) for previous studies on Progresa targeting).

In particular, we take Brown et al. (2016) as a starting point in this paper. They use rigorous consumption measures derived from LSMS surveys in nine countries in sub-Saharan Africa to test how well the canonical version of PMT performs in targeting poor households, as well as testing a few variations, including some (such as quantile regression centered at the poverty line) that explicitly aim to reduce exclusion errors. They test PMT effectiveness using outcomes including the percent of individuals predicted poor at both the 20th and 40th percentiles; the inclusion error rate, or the proportion of those estimated as poor that are not; and the exclusion error rate, or the proportion of those estimated not poor that indeed are. We follow their example and assess our models using similar metrics, more fully defined in section 4.2. They find that even their basic PMT model significantly outperforms their counterfactual of universal coverage in terms of inclusion error but tends to exclude many households erroneously. They conclude that "...econometric targeting typically provides at most modest gains in poverty impacts over other policy-relevant alternatives."

From the above, it is obvious that while PMT methods have made poverty targeting with limited universal data coverage possible, there is still plenty of room to improve processes and decrease error rates. There has been a burgeoning recent literature that we seek to add to that uses machine learning (ML) methods and alternate, often "big," data sources to improve PMT. As with previous evaluations of PMT, the evidence is mixed, with success dependent on context and data sources. Aiken et al. (2022) use survey, satellite, and mobile phone data to target Covid-19 aid in Togo, and find that their method reduced errors of exclusion anywhere from 4-21%. On the other hand, some studies (e.g., Barriga-Cabanillas et al. (2022)) have had more mixed results, particularly in areas where the "big data" is relatively homogeneous across the target population. Here, we take a recent paper by Yeh et al. (2020) as our starting point. In the article, the authors use satellite imagery and neural networks to predict poverty at the village level in countries in sub-Saharan Africa and test the accuracy of those predictions using consumption data from LSMS and DHS surveys. They find that their method can explain approximately 70% of village-level wealth variation. In this paper, we attempt to push beyond this by combining satellite images with survey data in our testing dataset to predict at the household level and test the predictive improvements' value by using data from a targeted RCT.

2.3 A note on costs

Finally, before describing the context of our project, a word should be said on the cost of targeting in general and the costs of increasing the technical requirements of PMT specifically. Surveying is expensive, and surveying to get accurate consumption or income measures can be even more so. Alix-Garcia et al. (2019) estimate that it can cost anywhere from $18 to $300 per household surveyed for data of comparable width and depth to LSMS surveys. In particular relevance to our setting, Caldés et al. (2006) show that the largest line item in Progresa's budget was "beneficiary identification," which made up 34% of the total costs (excluding the cost of the transfers themselves) through the year 2000. One of the benefits of PMT is that it reduces the costs of targeting – only periodic in-depth surveys can be supplemented by predictive models and more universal but shallower data sources like censuses. And recently proposed improvements to the method have emphasized two sources of cost improvements: i) the ability of governments to take advantage of even "cheaper" alternate data sources, such as publicly-available satellite images and available-to-governments mobile phone data (Aiken et al., 2022); and ii) the effectiveness of new methods of reducing the costs of surveys meant specifically for PMT (Ohlenburg et al., 2022).

However, it should be stated that both the standard version of PMT and particularly recent improvements, ours included, are not costless. First, the lack of transparency behind PMT methods has long been a concern of critics and practitioners; as one severe example, Cameron and Shah (2014) argues that mistargeting combined with the method's lack of transparency helped provoke social unrest in Indonesia. It should be noted that perfect transparency may also be undesirable as potential recipients may be able to manipulate their observable characteristics (or survey answers) to increase their likelihood of inclusion. Nevertheless, the increased use of ML methods, which introduce even more opacity towards how weights are calculated, is likely to reduce transparency beyond what would be optimal.

Second, the technical implementation of these targeting methods is not costless. We will discuss in more detail the technical cost of our method in section 6, but two points can be made here. First, as machine learning methods become more computationally advanced and testing and validating datasets become larger, the computing power and storage necessary to train and run the models also increases. OLS regression-based PMT methods that could formerly be run on any laptop with statistical software installed could quickly morph into ML methods that require substantial amounts of expensive and sometimes less secure cloud computing power. Second, increases in the technical sophistication of PMT methods in most cases also necessarily translate to an increase in the base level of technical knowledge required of their implementing team. Learning to export and process satellite images,[2] feed them and tabular data into a mixed-data neural network, and properly train and validate that network involved a non-trivial amount of learning by the authors of this paper. This increased level of technical complexity not only increases the human capital necessary to run targeting methods but also increases the number of potential fault points in the process, points in which human error can have serious consequences on the targeting of much-needed programs.

3 The Progresa Cash Transfer Experiment

In this work, we are interested not only in improving the predictive accuracy of standard program targeting methods but also in understanding the value of these improvements. To do this, we conduct an ex-post targeting exercise of Mexico's Flagship rural development program, *Progresa*, later renamed *Opportunidades*, and eventually *Prospera*. Specifically, we explore the original impact evaluation conducted by the International Food Policy Research Institute (IFPRI) and test program effects within the original targeted sample against those in samples targeted by a standard proxy means test and our improved measure.

3.1 Program Design

The Mexican Government established Progresa (Programa de Educacion, Salud y Alimentacion) in 1997 as the country's flagship rural antipoverty program. The program reached 26 million families by 2000 and included three main components. As the program's name implies, these components are designed to be multi-sectoral and intended to improve households' health, education, and nutrition.

The first component, and potentially the most internationally influential, aimed to increase children's educational attainment through a set of educational grants which effectively acted as a conditional cash transfer. The grants were unrestricted in their use, but their distribution was conditional on a child's enrollment in school and attendance for at least 85 percent of school days for a given month. The cash grants were adjusted every six months to account for inflation and increased as the child advanced in years of schooling. Grants were given for children in the third grade through secondary school graduation. At the time of program implementation, households received 65 pesos for a child enrolled in third grade. For children in secondary schools, the educational grants were slightly larger for females,

[2] Although recent work by Rolf et al. (2021) aims to make this step more accessible, there remain large barriers to entry to utilizing satellite imagery.

such that for children enrolled in the final grade, households received 240 pesos if the child was female and 210 if the child was male.

The second component intended to improve household members' access to basic health services. This component offered household members in-kind health services provided by the Mexican Social Security Institute. This component also included instructional training on health and nutrition called *Platicas*.

The final component aimed to improve household nutrition and included a fixed-value cash transfer. This component began at 95 pesos per household and had increased to 125 pesos by the end of the IFPRI evaluation. These cash transfers were provided conditional on a household's participation in the in-kind health services provided as a part of component two. Finally, this component also included nutritional supplements for all children under the age of 5 and for all women who were either pregnant or lactating.

At the time of its design, Progresa was unique among antipoverty programs globally, and it has since inspired a swath of similar programs concentrated in Latin America. The conditionality of the program was designed so that the program may not only reduce monetary poverty but also create incentives at the household level for human capital development. At the same time, these conditions may have adverse consequences if recipients live in areas in which they do not have access to education or health services. As many of the poorest and most marginalized populations may not have suitable access to these services, a potential downside of these conditionalities is that the poorest populations may be overlooked. As we will discuss in section 3.2, this is part of why the program was targeted using an index of community access to services rather than the standard consumption-based methodology.

3.2 Program Targeting

In this paper, the bulk of our results come from an ex-post targeting exercise of the Progresa program. As such, it is worthwhile to understand the original methodology for targeting the program, which was evaluated using a priority roll-out style randomized experiment. The targeting of Progresa took place in a three-stage process. First, the government selected communities using a principle component analysis-generated index of community marginality. Within selected communities, households were then selected using household survey data. Finally, the selected communities were presented with a list of selected participants and allowed to add households that they felt were excluded (Skoufias et al., 1999).

To ensure that the poorest households were included in the study, the National Institute for Statistics, Geography and Information (INEGI) constructed a marginality index using census-level data from the General Population and Housing Census of 1990, the Population and Housing Count of 1995 and the Geographic Integration Catalog. The marginality index was constructed using seven locality-aggregated variables: 1) illiteracy, 2) access to running water, 3) household drainage capability, 4) household electrical access, 5) the number of household occupants, 6) housing floor material, and 7) labor force participation in the agricultural sector. From these variables, the marginality index was calculated as follows:

$$MI_j = \sum_{i=1}^{7} f_i \frac{x_i^j - \bar{x}_i}{\sigma_i} \tag{1}$$

where x_i^j is the value of variable i for locality j, $\bar{x}_i$ and σ_i are the mean and standard deviation of variable i respectively. f_i is the weight for variable i as determined by the principle component analysis.

Among those communities chosen through the marginality index, they were screened for access to health and education services, as the utilization of these services was a prerequisite for program participation due to the conditionalities discussed in Section 3.1. Communities that did not have a school or health center local to them were still eligible for the program if they had a primary school within 2.5, 3, or 6 kilometers, depending on if the school was in a rural locality, located on a secondary road, or located on a federal route. This same rule applied to secondary schools within 5, 6, or 10 kilometers and health centers within 5, 10, or 15 kilometers.

Among the communities that were selected for the program, individuals were selected using socioeconomic data from the *Encuesta de Características Socioeconómicas de los Hogares* (ENCASEH), which we will describe in detail in Section 4.1. Using this data, INEGI identified eligible households using a discriminant analysis to describe the characteristics of households with a higher tendency to have incomes below the poverty line. Scores from the discriminant analysis ranged from 0 to 100, with lower scores having a higher likelihood of poverty and, thus, being more likely to be included.

Finally, following the initial registration of participants, the administrators of Progresa presented communities with the beneficiary list and allowed those communities to add households they thought had been unfairly excluded.

From this original beneficiary selection, roughly 52 percent of the original IFPRI sample was classified as eligible for the program. However, due to a sense among program administrators that the original classification was unjustly biased against older respondents with fewer children, the program underwent a "densification" process in which they added new households to the list of beneficiaries (Skoufias et al., 2001). The densification increased the proportion of eligible beneficiaries from 52 to 78 percent of the evaluation sample. Below, we will estimate the program's effects on Progresa's originally targeted sample and leverage the densification process to create an, albeit selected, subsample of those households who were not originally targeted but did receive the program.

3.3 Previous Findings on Progresa's Effectiveness

We conduct this ex-post targeting exercise on the Progresa sample for two reasons. The first is the density and quality of data related to the intervention and the setting as a whole. We will discuss this reason more thoroughly in Section 4.1. The second is due to the international familiarity with the structure of the Progresa program. Progresa/Opportunidades/Prospera is potentially the most closely studied social program in decades. Although we do not aim in this work to make direct statements about the effectiveness of the program, our ability to make statements about the link between targeting accuracy and program effectiveness hinges on the strong effects found by others.

Given the size of this literature, it would be outside the scope of this work to describe it holistically,[3] however, it is worthwhile to review some of the works related to the original IFPRI evaluation, as we utilize data from this work.

Potentially the most well-known work related to Progresa, as it is taught in undergraduate development classes, relates to the program's impacts on child education, where Schultz (2004) has shown that the program increased school enrollment and attendance in the short-run. These effects are potentially linked to the program's effects on child labor, where Skoufias et al. (2001) show that children were less likely to participate in labor if their mothers were enrolled in the program. In this work, we are interested primarily in how targeting methods impact the antipoverty potential of social programs. As such, we report the program's effects on standard Foster-Greer-Thorebeck (FGT) poverty measures (Foster et al., 1984) and on food consumption. As such, the work that relates most closely to ours is Hoddinott and Skoufias (2004), in which they show that the program increased dietary quantity, as measured by caloric intake, and dietary quality, as measured by dietary diversity.

It is rare that the literature on a program is as unilaterally supportive of its effectiveness as the literature on Progresa is. For this reason, the program has served as the model for conditional cash transfer programs in over sixty counties. It also makes the program an ideal background for testing targeting methods.

[3] Interested readers should see Parker and Todd (2017) for a comprehensive review of the literature, and Araujo and Macours (2021) for a look at the effects in the long term.

4 Methods

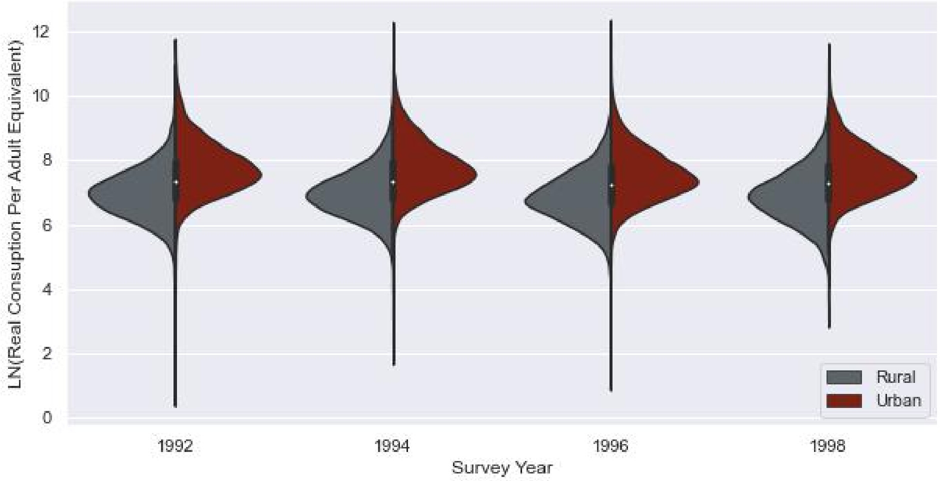
4.1 Data

This work uses two household surveys administered by INEGI throughout the 1990s. The first, the ENIGH instrument, is Mexico's bi-annual nationally representative consumption survey (INEGI, 2023). This survey is a standard household consumption survey conducted on a repeated cross-section, is sub-nationally representative, and includes questions related to household demographics, income, consumption, and assets. We will use this data to train the multi-layer perceptron section of our mixed-data neural network described briefly in section 1 and in-depth in section 4.2 below. We will also use this data to create a predictive model of consumption using a standard proxy means test.

To create our models, we utilize data from the ENIGH surveys conducted in 1992, 1994, 1996, and 1998. Using surveys across multiple years to create our model creates the risk of bias in our predictions if the samples across these survey waves exhibit changes in the distribution of household consumption. We may naturally be "concerned" about consumption growth due to economic development. However, given that this period saw the 1994 Mexican Peso Crisis, we may also be concerned that between 1992 and 1998, consumption in Mexico's rural communities decreased in real terms. To explore this possibility, we present the distribution of household consumption in real terms across survey waves in Figure 1. We see that, in real terms, there was remarkably little change in household consumption across survey waves. This most likely reflects the relative stagnation that Mexico witnessed throughout the 1990s and bodes well for our ability to concatenate data sets to increase the sample size available to train our neural network.

In addition to the ENIGH surveys, we use the INEGI-collected ENCASEH survey, discussed briefly in section 3.2. The ENCASEH served as the baseline for the Progresa evaluation and collected information on household demographics, education, healthcare usage, and income. The survey also included a consumption module. However, this module, unfortunately, did not include a measure of auto-consumption, that is consumption of goods or foods created or grown within the household. It is for this reason that Skoufias et al. (1999) uses the 1996 ENIGH survey to compare Progresa's true targeting to a proxy means test targeting metric. In our context, this is important to note because, later, we will compare the consumption predictions of our neural network and a proxy means test to the consumption reported in the ENCASEH survey. As the ENCASEH consumption did not include auto consumption, and the ENIGH survey did, it is natural to expect our predictions to be higher than the reported consumption. Rather than a bias in our predictions, it is much more likely that this difference represents the missing auto consumption in the ENCASEH sample.

Figure 1: ENIGH Consumption Across Survey Waves



Notes: Figure shows estimated real log consumption per survey year using data from ENIGH surveys. The left-hand (grey) side of each line is estimated distribution of consumption in rural areas; the right-hand (red) side is the estimated distribution of urban areas. Real consumption is calculated using Mexican price indices from the St. Louis Federal Reserve Economic Database (FRED)

Finally, to create the CNN portion of our mixed-data neural network, we utilize data from Google Earth Engine to match images to households at the village level (Gorelick et al., 2017). The images that we take from the database are generated by three separate satellites. The first, NASA/USGS LANDSAT, creates a database of hyper-spectral imagery over an area. The second, the MODIS satellite, creates a vegetation and rangeland conditions database, measured as a composite index called the Normalized Dense Vegetation Index (NDVI). Finally, we utilize images from the DMPS-OLS database to capture information on night light intensity within a given area (Zhao et al., 2022). The images we pull are a 3-year composite over a 255X255 pixel area centered on the relevant survey year. We match these images to localities using geographic coordinates[4,5] provided by INEGI. The images that we use within our network consist of 9 layers. These layers include RBG spectral imagery, infrared imagery, Night light intensity (by pixel), and NDVI (by pixel). An example of the images we use within our network can be found in Appendix A for an anonymized locality

[4] The coordinates used to match images are not publicly available for privacy reasons and cannot be accessed outside of the INEGI microdata laboratory in Mexico City. Thus, a replication package cannot be provided for this work, but the analysis files used are available upon request from any researcher with access to the laboratory.

[5] We begin with a total sample of 48,301 households across the ENIGH waves. For 2,120 (4.3 percent), we cannot identify the municipality code for the household. For another 1,582 (3.2 percent), we do not observe a geographic coordinate for their municipality code. This results in a total sample of 44,599 ENIGH households matched to satellite imagery.

within the Progresa sample.

4.2 Methods

4.2.1 Targeting Methods

There are many versions of the proxy means test, but all share a general procedure. In most cases, a detailed household consumption survey is used to create a predictive model of per-capita or per-adult equivalent consumption. Traditionally, this model is built using OLS linear regression. This model is then projected to a larger administrative dataset, such as a household census or social registry, which has a higher coverage rate but does not contain detailed data on consumption. The predictions from this PMT model are then often used to target social programs or select benefit recipients. In this paper, we follow this methodology and build a predictive model of adult-equivalent consumption at the household level using the repeated cross-section of ENIGH surveys described in section 4.1. We then extend the models to data from the ENCASEH survey to conduct our ex-post targeting exercise of Progresa. We use two methodologies to create our predictive models. In the first, we match the standard OLS version of the PMT. Here, we model consumption using the model:

$$LN(C_i^{AE}) = \alpha + \mathbf{X}_i\beta_1 + \mathbf{Z}_i\beta_2 + \mathbf{W}_i\beta_3 + \varepsilon_i \tag{2}$$

Where C_i^{AE} represents adult-equivalent consumption, X_i is a vector of household demographic characteristics, Z_i is a vector of household structure characteristics, and W_i is a listing of household assets.[6] We include a more detailed description of these characteristics, including summary statistics and a correlation mapping, in Appendix B.

In the second method, we utilize a mixed-data neural network to improve the model's predictive accuracy. By mixed data, we mean that the network takes two data types as inputs, household survey data from the ENIGH surveys and image data from Google Earth Engine, as described in section 4.1. Our network has two branches, the structure of which is described in figure 2. In the first branch, we pass the characteristics described above through a standard artificial neural network (ANN) consisting of 3 layers, with one hidden layer. This simple model architecture was selected through hyperparameter tuning and likely reflects that there are few non-linearities relating our ENIGH characteristics to household consumption.

[6] We acknowledge that by including household assets in our predictive model, we likely eliminate the possibility of extending the model to a household census. However, that information is likely included in many social registries, and in this work, we are interested in understanding how targeting, under ideal conditions, impacts program effectiveness.

In the second branch, we follow Yeh et al. (2020) and utilize a ResNet-18 model architecture for our CNN (Chandola et al., 2021). This model has 72 layers with 18 hidden layers. The residual blocks[7] allow us to create a deep network while minimizing the risk of overfitting and reducing training time. Our model concatenates the weights from these branches and passes them through one fully connected layer to produce a final prediction.

We train our model using an ADAM optimizer and a batch size of 64. We train on 100 Epochs. However, we allow for early stopping.

While increasing the predictive accuracy of the PMT will surely improve the targeting performance of the algorithm, Brown et al. (2016) have shown that PMT algorithms perform particularly poorly towards the endpoints of the consumption distribution. This is likely because households look very similar in characteristics on the lower end of the distribution. At the same time, any program that uses a PMT targeting method must have a cutoff. It is reasonable to believe that there is little heterogeneity in program effectiveness for households in the neighborhood of this cutoff. Therefore, a policymaker looking to maximize program impact may be willing to sacrifice precision near the targeting cutoff for improved accuracy among households towards the endpoint of the consumption distribution. For this reason, we complement our standard OLS and neural network models with what we refer to as a "distance-weighted" model. In these models, we weight observations with a triangular kernel over percentile rank centered at the median such that the objective function can be thought of as:

$$\hat{\beta} = \min_{\beta} \quad \frac{1}{n}\sum^{I}(y_i - X_i\beta)^2|p_{y_i} - 0.5| \tag{3}$$

where p_{y_i} is the percentile of household i's consumption. This model is similar to the poverty-weighted model presented in Brown et al. (2016) except that it attempts to limit program leakage by overweighting both poor and wealthy households rather than just those that are poor.

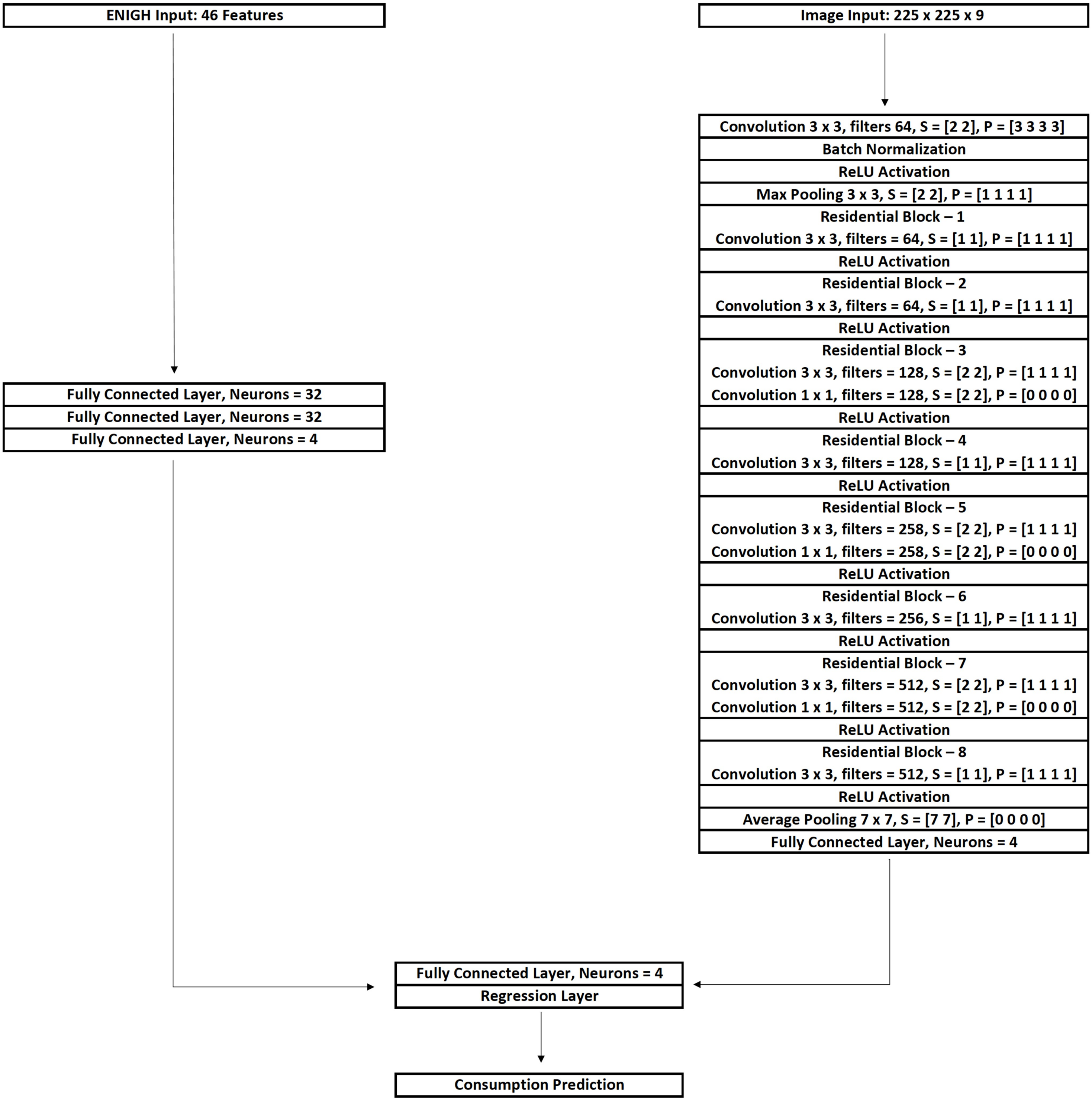
4.2.2 Progresa Effects

In section 5.1 we present an evaluation of the models presented above. We then use these algorithms to predict the adult-equivalent consumption of households within the initial Progresa evaluation sample. In doing so, we are interested in knowing how the program's effectiveness would have been different if the program had been targeted at a household level using one of our PMT algorithms.[8] As such, we match the original Progresa program density

[7] A residual block is a set of layers that can be circumvented if the back-propagation algorithm training the network determines that there is nothing additional to be learned through the layers

[8] We discuss in section 5.1 that we only conduct this exercise using the standard PMT and the predictions created by the "distance-weighted" version of our neural network

Figure 2: Neural Network Model Architecture


ENIGH Input: 46 Features
Image Input: 225 x 225 x 9
Convolution 3 x 3, filters 64, S = [2 2], P = [3 3 3 3]
Batch Normalization
ReLU Activation
Max Pooling 3 x 3, S = [2 2], P = [1 1 1 1]
Residential Block – 1
Convolution 3 x 3, filters = 64, S = [1 1], P = [1 1 1 1]
ReLU Activation
Residential Block – 2
Convolution 3 x 3, filters = 64, S = [1 1], P = [1 1 1 1]
ReLU Activation
Residential Block – 3
Convolution 3 x 3, filters = 128, S = [2 2], P = [1 1 1 1]
Convolution 1 x 1, filters = 128, S = [2 2], P = [0 0 0 0]
ReLU Activation
Residential Block – 4
Convolution 3 x 3, filters = 128, S = [1 1], P = [1 1 1 1]
ReLU Activation
Residential Block – 5
Convolution 3 x 3, filters = 258, S = [2 2], P = [1 1 1 1]
Convolution 1 x 1, filters = 258, S = [2 2], P = [0 0 0 0]
ReLU Activation
Residential Block – 6
Convolution 3 x 3, filters = 256, S = [1 1], P = [1 1 1 1]
ReLU Activation
Residential Block – 7
Convolution 3 x 3, filters = 512, S = [2 2], P = [1 1 1 1]
Convolution 1 x 1, filters = 512, S = [2 2], P = [0 0 0 0]
ReLU Activation
Residential Block – 8
Convolution 3 x 3, filters = 512, S = [1 1], P = [1 1 1 1]
ReLU Activation
Average Pooling 7 x 7, S = [7 7], P = [0 0 0 0]
Fully Connected Layer, Neurons = 4
Fully Connected Layer, Neurons = 32
Fully Connected Layer, Neurons = 32
Fully Connected Layer, Neurons = 4
Fully Connected Layer, Neurons = 4
Regression Layer
Consumption Prediction

Notes: Figure represents the model architecture of the mixed-data neural network used to re-target the Progresa program. S represents the stride length of the convolutions within the CNN branch, P represents the padding added to images during each layer.

and consider a household to be "targeted" by a given method if their predicted consumption falls under the 52^{nd} percentile of the predictions.

We classify households into one of four groups by comparing each targeted sample with Progresa's original targeting. Households are considered "always included" if they were included in Progresa's original targeting and are included in the relevant method's targeted sample. Conversely, the "always excluded" sample is composed of households excluded from Progresa and are excluded by the relevant method. What we refer to as the "leakage" sample comprises households included under Progresa's targeting scheme but should not have been according to the PMT targeting method. Finally, households that did not receive the program but should have according to the method in question makeup what we refer to as the "under-coverage" sample.

In this paper, we are interested in the heterogeneous treatment effects between the groups outlined above, particularly the "always included", "leakage", and "under-coverage" samples. Identifying the effects of the program within the "always included" and "leakage" samples is straight forward, as these households were included in Progrsa's original targeting and received the program outright. To identify the effects in the "leakage sample", we leverage the densification process and consider the treatment effects on households at the intersection of this densification and our targeting assignments. These households were not included in the original recipient selection of Progresa but still received the program, allowing us to identify the effect in the "under-coverage" sample.

Leveraging the densification process poses two challenges to comparing effects across groups. First, we know that the process was designed to include older households and those households with fewer children. This means the sample who received transfers is a biased subset of the full "under-coverage" sample. For this reason, we control for household size and the age of the head of household in the regressions below. More pressing is the issue of treatment fidelity. According to Skoufias (2005), of the 3,350 households that were eligible for the program but never received transfers, 85.7 percent (2,872 households) were part of the densification process. To cope with this issue of treatment fidelity, we estimate the local average treatment effect within each group by instrumenting for program participation with treatment assignment at the village level[9]

To conduct our heterogeneous analysis, we utilize household data collected as a part of

[9] There is also the issue of program exposure as households enrolled as part of the densification process may have begun receiving the transfers later than those in the originally targeted sample. Indeed, payment data associated with the program indicated that payments in densification households began four months after the program's roll-out. However, payment records indicate that all households who would eventually receive a transfer were enrolled by December 1998, six months before the first evaluation survey used in our pooled analysis.

the original program evaluation and estimate a pooled regression of the form:

$$Y_{i,v,s} = \alpha + \beta \hat{T}_{i,v} + \mathbf{X}_i \gamma + \mathbf{P}_v \delta + \varepsilon_{i,v,s} \tag{4}$$

where Y_i represents the outcome of interest for household i in village v during survey wave s.[10] $\hat{T}_{i,v}$ represents the program participation of household i, instrumented by the treatment assignment of village v. $\mathbf{X}_i$ is a vector of pre-program characteristics at the household level.[11] $\mathbf{P}_v$ is a vector of locality-level food prices.[12] Finally, $\varepsilon_{i,v,s}$ is an idiosyncratic error term clustered at the locality level, the same level as treatment assignment.

We test equation 4 on the full Progresa sample, the always-included, leakage, and undercoverage samples for the standard PMT and distance-weighted network versions of our targeting metrics. We also report p-values associated with the hypothesis that the treatment effects are equal across these groups. We conduct these tests on two sets of outcomes. First, we replicate the results from Hoddinott and Skoufias (2004) on mean caloric availability in household diets, and more specifically, calories from grains, vegetables, and animal-sourced foods.[13] In the second, we report the treatment effects on poverty measures in the Foster et al. (1984) tradition[14]

5 Results

5.1 Model Performance

We begin our exploration of targeting effectiveness by reporting the performance of the standard PMT and our adjusted methods on out-of-sample ENIGH households. Put more simply, Table 1 reports the performance of our targeting methods on ENIGH observations that were not input into the model for training and were kept aside for evaluation. We present the R^2 associated with each model and the targeting performance on two separate poverty

[10] In this analysis, we utilize data from the Encuesta Evaluation de los Hogares (ENCEL) household surveys conducted in June 1999 and November 1999

[11] These characteristics include log household size; proportions of children 0–2, 3–5; boys 6–7, 8–12, 13–18; girls 6–7, 8–12, 13–18; women 19–54; men 55 and older; women 55 and older); characteristics of the head (education, age, occupation, ethnicity, marital status, gender)

[12] locality-level prices include tomatoes, onions, leafy vegetables, orange, tortillas, corn, milk, white bread, local bread, rice, beans, chicken, and eggs.

[13] In their analysis, Hoddinott and Skoufias (2004) report Progresa treatment effects across survey waves. Here, we pool our analysis across the June 1999 wave and November 1999 wave to ensure adequate statistical power to detect differences across groups. We also restrict our analysis to households classified as poor before Progresa's densification process. For this reason, our results differ from the results in Hoddinott and Skoufias (2004) in both sample size and value.

[14] For these measures, we use the minimum cost of a *standard food basket* from Skoufias et al. (1999) of 320 Pesos per adult equivalent per month in nominal terms as the poverty line.

lines. In the more restrictive case, we set the poverty line at the consumption level of the 20^{th} percentile. We also report performance on a more inclusive 40^{th} percentile cutoff. The results presented here are accompanied by scatter plots of true versus predicted consumption levels for each model in Appendix C.

Table 1: ENIGH Out-of-Sample Model Performance

	Standard OLS	Distance-Weighted OLS	Standard Neural Network	Distance-Weighted Neural Network
R-Squared	0.599	0.579	0.630	0.617
Percent Predicted Poor				
20% Threshold	14.7%	19.8%	15.0%	18.7%
40% Threshold	35.6%	38.1%	34.7%	41.6%
Correctly Included (As a % of Actual Poor)				
20% Threshold	53.9%	64.9%	55.9%	64.7%
40% Threshold	70.6%	73.4%	70.0%	78.1%
Inclusion Error (As a % of Inclusion)				
20% Threshold	27.2%	34.9%	26.2%	31.3%
40% Threshold	22.2%	24.5%	21.1%	26.5%

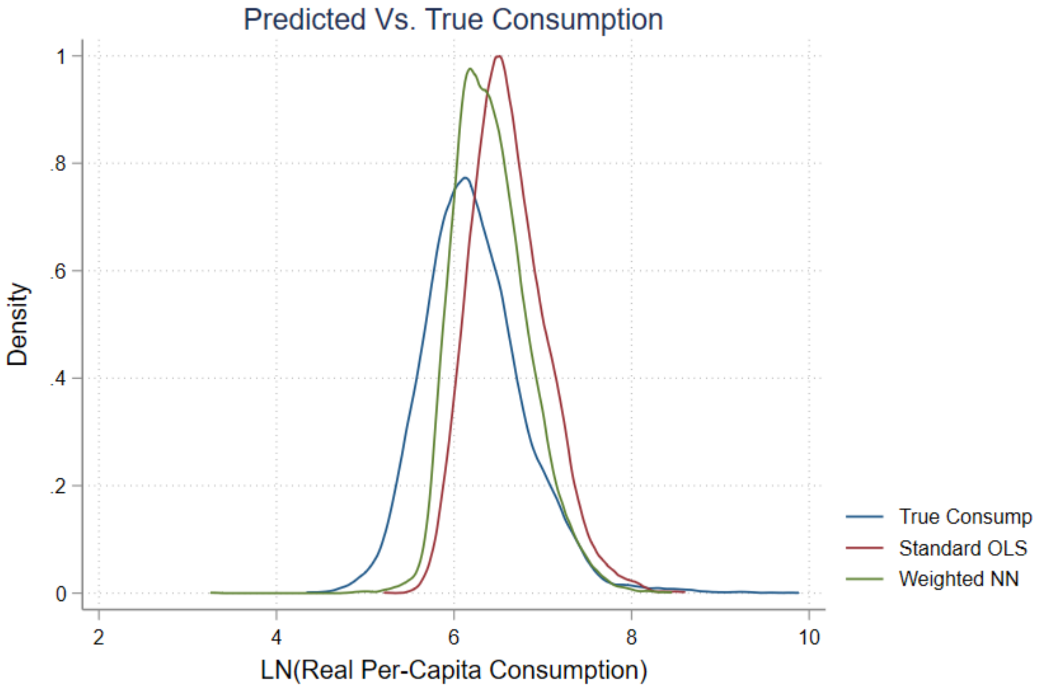
"Percent predicted poor" represents the number of households predicted to be under the relevant threshold as a proportion of the total sample. "Correctly included" represents the proportion of households predicted to be under the relevant threshold as a proportion of the households under that threshold. "Inclusion error" represents the proportion of households incorrectly predicted to be under the threshold as a proportion of all households predicted to be under the threshold.

We see that across our models, we can explain more than half of the variation in out-of-sample consumption. Our models each display an R^2 of between 0.58 and 0.63. Compared to the model performance from Yeh et al. (2020), which reports an R^2 of 0.67 at the village level using a CNN, we consider 0.63 at the household level to be a strong performance. We observe that there are gains to utilizing our neural network and including image data, but they are small, an increase of just 3.1 percent of variation explained in the unweighted case.

We observe potentially large gains from the inclusion of the percentile weights. We see that the distance weights make the targeting more inclusive for both the OLS and neural network models. This naturally means that more people are correctly classified as poor, but it comes with tradeoffs in terms of inclusion error. However, we see that in the case of the neural network, the distance weighting increases correct inclusions by 10.8 percent over the standard PMT while only increasing inclusion errors by 4.1 percent. For this reason, we consider the distance-weighted neural network our model of choice and limit our analysis of Progresa outcomes to this model and the standard version of the PMT.

Armed with our preferred model, we extend our predictions to the ENCASEH survey. Figure 3 reports the distributions of our model's predictions against the truly measured consumption in the ENCASEH survey. Two things are worthy of note here. First, we see that our predicted distributions are biased upward for both models compared to the true consumption. This likely reflects the lack of auto consumption included in the calculation of consumption from the ENCASEH survey. It may also, in part, reflect that the Progresa communities were specifically selected for their low socio-economic status, whereas the ENIGH surveys are nationally representative. Second, we see that this bias is smaller in the case of the distance-weighted neural network. This is likely a reflection of the fact that this model places greater weight on poorer households that bear greater resemblance to the ENCASEH households.

Figure 3: Predicted vs. True ENCASEH Consumption



Notes: Figure displays the distribution of log per-capita consumption for the sample i) calculated from the ENCASEH survey (blue line); ii) predicted using our preferred method, a distance-weighted neural network (green line); and iii) standard OLS PMT (red line)

While the comparison of distributions is interesting, as the program is targeted based on a poverty line, households are targeted subject to their position relative to others. As such, we are interested in how well our algorithms predict household ranking. In Table 2, we present the Spearman rank correlation coefficients between our measures and the ENCASEH measure of consumption. To understand how our methods compare to Progresa's original targeting methodology, we also include the marginality index discussed in section 3.2.

Table 2 suggests that our targeting methods significantly outperform the marginality index used to originally target the program and are themselves highly correlated. In terms of correlation with true consumption, we see that the models perform similarly, and the standard PMT slightly outperforms our preferred model. However, as the ENCASEH survey does not include auto consumption, it is unclear how closely the estimates in Table 2 relate to the rank correlations with a more full measure of household consumption and how much weight we should place on small differences.

Table 2: Progresa Consumption Rank Correlations

	True Consumption	Marginality Index	Standard OLS	Distance-Weighted Neural Network
True Consumption	1			
Marginality Index	-0.2516	1		
Standard OLS	0.4991	-0.3817	1	
Distance-Weighted Neural Network	0.4557	-0.4032	0.9175	1

Notes: Coefficients represent the Spearman rank correlation coefficients between the level of consumption reflected in the ENCASEH survey and the levels predicted by various targeting methods.

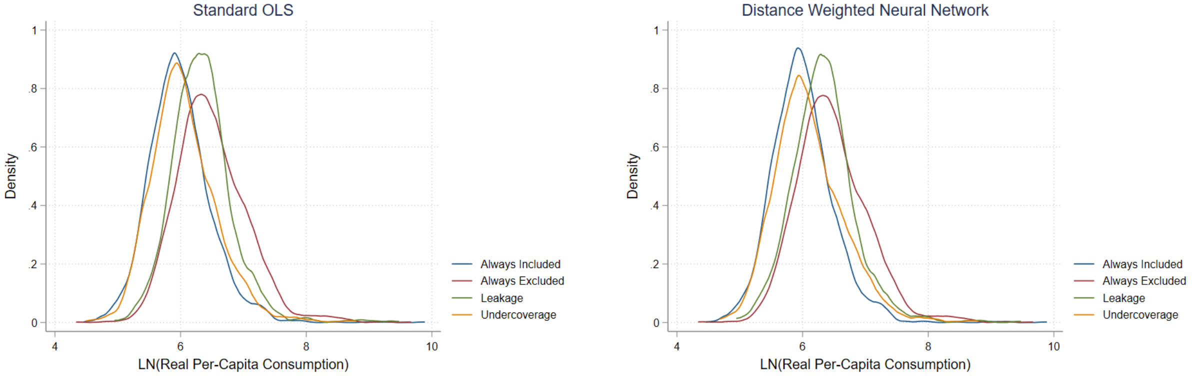
Finally, after classifying households using our predicted consumption levels, we can compare measured consumption levels across the four types of samples that we described in section 4.2.2. In Figure 4, we display the distribution of consumption in these groups for both of our reported measures. Two observations are important from this figure. First, each set is non-empty, implying that our methods differ from Progresa's original targeting. Second, for both panels, we see that the distribution in the leakage sample dominates the distribution of consumption in the under-coverage sample. This implies that by applying our PMT-style targeting methods, we include poorer households that were originally excluded and exclude wealthier households that were originally included. If we believe that anti-poverty programs are more effective in poorer households, this may suggest a program targeted with our methods may have larger effects than one targeted using Progresa's marginality index. For further clarity, in sections 5.2 and 5.3, we compare the effects of Progresa among households that make up the blue distribution to those that make up the green and yellow distributions.

5.2 Progresa Effects: Results on Diets

We begin our analysis of the Progresa effects across groups by looking at the effects on diets. Table 3 reports these effects for the full Progresa sample, the correctly included, leakage, and under-coverage samples for the standard OLS and distance-weighted neural network models.

First, we observe that in line with Hoddinott and Skoufias (2004), there are large dietary improvements associated with Progresa participation. Looking at the differences between the correctly included and leakage samples for each of our models, we observe small differences in point estimates across outcomes. First, we see that the effect sizes for both targeting methods are larger in the always-included sample than in the leakage sample. Although we cannot reject the hypothesis of equality of coefficients for any of these outcomes, We see

Figure 4: Leakage and Under Coverage across Targeting Methods



Notes: A comparison of predicted log real per-capita income for both standard OLS PMT (left) and our distance-weighted neural network model (right) for four sub-groups: those included in both the full Progresa sample and the targeting method (blue); those included in the Progresa sample but deemed ineligible in the targeting method (green), those excluded from Progresa but deemed eligible by the targeting method (yellow), and those excluded from Progresa and deemed ineligible by the method (red)

that the differences are largest for animal-sourced foods. This result is intuitive because we see from Figure 4 that the always included sample makes up the poorest household in the program, and demand for animal-sourced foods is likely to be the most elastic of all the groups tested as these foods act as a luxury good in households' utility functions, while other types of foods act as normal goods. While these differences are modest and only comprise one component of a healthy diet, they nonetheless suggest that gains are associated with improved targeting. This may be particularly true for households with young children, as animal sources foods contain large amounts of Choline, a vital nutrient for bone formation and cellular growth.

The results for the under-coverage sample are less straightforward. Here, we observe the largest increases in caloric intake and see that those changes are driven largely by increases in calories from grains. Conversely, we observe mixed results for animal-sourced foods and the smallest improvements in the consumption of vegetables. While the p-values for the test of equality with the leakage sample are borderline for grains and calories, these results suggest that there may be modest gains associated with improving program targeting on the margins. It may also suggest that these gains are centered around dietary quantity rather than dietary quality.

5.3 Progresa Effects: Results on Poverty

Although the quality of diets and nutrition generally makes up one of Progresa's stated goals (the others being education and health), the program was designed primarily as an

Table 3: Progresa ITT Effects on Calorie Intake Across Targeting Samples

		Standard OLS						Distance Weighted Nerual Network					
	(1)	(2)	(3)	(4)	Differences			(5)	(6)	(7)	Differences		
	Full Sample	Always Included	Leakage	Undercoverage	(2)-(3)	(3)-(4)	(2)-(4)	Always Included	Leakage	Undercoverage	(5)-(6)	(6)-(7)	(5)-(7)
LN(Total Calories)	0.0514***	0.0557***	0.0401***	0.0819***	0.28	0.06	0.16	0.0522***	0.0504***	0.0721***	0.90	0.33	0.32
	(0.00828)	(0.00899)	(0.0138)	(0.0197)				(0.00896)	(0.0134)	(0.0211)			
LN(Calories from Grains)	0.0322***	0.0369***	0.0164	0.0812***	0.32	0.02	0.06	0.0304**	0.0341	0.0825***	0.85	0.11	0.04
	(0.0122)	(0.0122)	(0.0215)	(0.0250)				(0.0124)	(0.0212)	(0.0283)			
LN(Calories from Animal Sourced Foods)	0.174***	0.200***	0.122***	0.209**	0.11	0.35	0.92	0.207***	0.115**	0.0807	0.06	0.71	0.14
	(0.0462)	(0.0518)	(0.0463)	(0.0970)				(0.0504)	(0.0483)	(0.0994)			
LN(Calories from Vegetables)	0.344***	0.353***	0.335***	0.265***	0.67	0.32	0.17	0.362***	0.319***	0.249***	0.28	0.29	0.08
	(0.0371)	(0.0420)	(0.0390)	(0.0679)				(0.0413)	(0.0374)	(0.0679)			
Observations	15142	11395	3747	2460				11594	3496	2323			

Notes: Columns (1) - (7) represented the estimated Progresa ITT effect within the stated sample. P-values of the test of equality between the coefficients accompany the main ITT estimates. Household controls include log household size; proportions of children 0–2, 3–5; boys 6 7, 8–12, 13–18; girls 6–7, 8–12, 13–18; women 19–54; men 55 and older; women 55 and older); characteristics of the head (education, age, occupation, ethnicity, marital status, gender). Price controls include the village-level prices for tomatoes, onions, leafy vegetables, orange, tortillas, corn, milk, white bread, local bread, rice, beans, chicken, and eggs.

anti-poverty intervention. As such, we may be most keenly interested in the program's effects on poverty measures and how this effectiveness is impacted by the method used to target the program. As such, we report the effects of Progresa on FGT-style poverty measures in Table 4. This table follows the same structure as Table 3. Again, we find large impacts of Progresa on the full sample. We see that households assigned to a treatment village were 2.6 percentage points less likely to have consumption below the poverty line, accompanied by a decrease of 35.41 pesos in the poverty gap and 18461.4 pesos^2 in poverty intensity. We see that the reduction in the poverty headcount is largest in the leakage sample, although these differences are not statistically significant at the 5% level. This likely reflects that households in the leakage sample were closer to the poverty line in the pre-program period and, thus, more likely to be pushed over the threshold by the program. This reflects an interesting paradox in which poorly targeted programs may look more effective if only measured through the lens of poverty rates.

This relationship flips when looking at more detailed measures of poverty. While the decrease in the poverty gap is nearly identical across the always-included and leakage samples, we see that the always-included sample displays a larger decrease in poverty intensity than the leakage sample.

Curiously, we see that the under-coverage sample displays the smallest point estimates across outcomes. While it is unclear why we might observe the smallest effects among these households, it may relate to program exposure and the timing of the cash transfers. As the households in the under-coverage sample began receiving their transfers at least four months after the other samples, it is possible the effects had not fully passed through to consumption by the time of the survey.

Table 4: Progresa ITT Effects on Poverty Measures Across Targeting Samples

		Standard OLS						Distance Weighted Nerual Network					
	(1)	(2)	(3)	(4)	Differences			(5)	(6)	(7)	Differences		
	Full Sample	Always Included	Leakage	Undercoverage	(2)-(3)	(3)-(4)	(2)-(4)	Always Included	Leakage	Undercoverage	(5)-(6)	(6)-(7)	(5)-(7)
Poverty Headcount	-0.0275*** (0.00798)	-0.0232*** (0.00673)	-0.0460*** (0.0173)	-0.0148 (0.0220)	0.18	0.21	0.70	-0.0252*** (0.00727)	-0.0477*** (0.0158)	-0.00194 (0.0242)	0.15	0.08	0.33
Poverty Gap	-44.12*** (5.281)	-45.43*** (5.498)	-44.39*** (6.493)	-26.06** (12.23)	0.87	0.12	0.09	-45.93*** (5.462)	-47.18*** (6.040)	-24.33** (12.18)	0.84	0.05	0.06
Poverty Intensity	-23735.4*** (2776.5)	-25858.4*** (3087.3)	-19526.5*** (2650.4)	-14325.1** (6367.5)	0.04	0.39	0.06	-25999.7*** (3014.7)	-20968.7*** (2526.7)	-14443.6** (5918.9)	0.09	0.26	0.05
Observations	15142	11395	3747	2460				11594	3496	2323			

Notes: Columns (1) - (7) represented the estimated Progresa ITT effect within the stated sample. P-values of the test of equality between the coefficients accompany the main ITT estimates. Household controls include log household size; proportions of children 0–2, 3–5; boys 6 7, 8–12, 13–18; girls 6–7, 8–12, 13–18; women 19–54; men 55 and older; women 55 and older); characteristics of the head (education, age, occupation, ethnicity, marital status, gender). Price controls include the village-level prices for tomatoes, onions, leafy vegetables, oranges, tortillas, corn, milk, white bread, local bread, rice, beans, chicken, and eggs.

6 Discussion

In this work, we have aimed to leverage recent advances in machine learning to improve traditional methods of program targeting. Motivated by Yeh et al. (2020), we have combined the traditional proxy means test with a convolutional neural network that allows us to incorporate satellite imagery, thus allowing us to incorporate information that is not readily available in traditional surveys. We have also adjusted our algorithms to place greater weight on households at the endpoints of the consumption distribution. The result is a preferred algorithm that is slightly more precise than the standard PMT but is substantially more inclusive of poor households, with limited trade-offs in terms of program leakage.

We test the value of this new algorithm and the standard proxy means test by conducting an ex-post targeting exercise of the Progresa cash transfer experiment. In this exercise, we compare the sample of households that would have been targeted under consumption-based targeting methods to the true targeted sample created using an index of marginality at the community level. We show that consumption-based methods target a sample of households that is poorer than those targeted by Progresa's original method. At the intersection of these targeted groups, we leverage a randomized experiment to test the effects of Progresa among households who are always included and those who received the program but should not have. We also leverage a second targeting exercise to test the program's effects in a sample not included under Progresa's original method but still received the program as they would have under a consumption-based method. We show that the heterogeneous effect sizes across the groups are similar for most dietary and poverty outcomes. However, there is evidence that households included in both consumption-based and marginality-based methods benefited more from the program than those that should have been excluded in terms of consumption. These improved benefits relate to the consumption of expensive calories and the intensity of the poverty they face. These differences are particularly pronounced when studying the group targeted by our preferred distance-weighted neural network algorithm. The results we observe across the leakage and under-coverage samples are more mixed. On the one hand, we find larger gains in food consumption in the under-coverage sample. However, these results are reversed for FGT poverty measures.

Our results suggest that in terms of pure targeting metrics, there are modest gains to be made to traditional targeting methods by utilizing machine learning methods and incorporating data, such as satellite imagery, that have been previously underutilized. They also suggest that improved program targeting improves social protection programs' measured effects. However, the differences are modest and only exist on margins where programs have the potential to make the largest improvements.

This work has several limitations. First, what we refer to as our "under-coverage" sample has several differences in treatment implementation from our other samples. While we attempt to address these differences by controlling for explicit differences and forgoing the intent-to-treat estimation for the local average treatment effect, we cannot completely rule out the possibility that differences in implementation drive the differences we observe with this sample. Second, as we do not have precise and disaggregated data on the costs of surveying or the server costs associated with implementing our algorithms, we cannot make statements about the cost-effectiveness of the modest improvements in program effect sizes that we observe as a result of improved program targeting. Ideally, this work would provide a justification for a randomized control trial testing the impact of targeting methodology on program effectiveness.

Our results contribute to a large literature on program targeting, specifically how to improve existing methods and how these improvements translate to program effects. There are a large number of non-measurable costs associated with the methods we have used here. Not the least, the lack of transparency these methods provide program recipients and the difficulty of their implementation. There are equity concerns associated with accurate program targeting that are beyond the scope of this paper. However, for a policymaker seeking to maximize program impacts and cost efficacy, our results highlight the need for a more careful analysis of the trade-off between targeting accuracy and the resources dedicated to selecting program recipients.

References

Aiken, E., S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock (2022, March). Machine learning and phone data can improve targeting of humanitarian aid. *Nature 603*(7903), 864–870. Number: 7903 Publisher: Nature Publishing Group. 5, 6

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, R. Purnamasari, and M. Wai-Poi (2016, April). Self-Targeting: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy 124*(2), 371–427. Publisher: The University of Chicago Press. 3

Alix-Garcia, J., K. Sims, and L. Costica (2019). Better to be indirect? Testing the accuracy and cost-savings of community leader surveys for poverty targeting. 3, 4, 6

Araujo, M. C. and K. Macours (2021, December). Education, Income and Mobility: Experimental Impacts of Childhood Exposure to Progresa after 20 Years. Technical report, Inter-American Development Bank. 10

Baker, J. L. and M. E. Grosh (1994, July). Poverty reduction through geographic targeting: How well does it work? *World Development 22*(7), 983–995. 3

Barriga-Cabanillas, O., J. E. Blumenstock, T. J. Lybbert, and D. Putman (2022). Digital Breadcrumbs and Dietary Diversity: Testing the Limits of Cell Phone Metadata in Poverty and Impact Assessment. 5

Brown, C., M. Ravallion, and D. v. d. Walle (2016, December). A Poor Means Test? Econometric Targeting in Africa. Technical Report w22919, National Bureau of Economic Research. 4, 5, 14

Caldés, N., D. Coady, and J. A. Maluccio (2006, May). The cost of poverty alleviation transfer programs: A comparative analysis of three programs in Latin America. *World Development 34*(5), 818–837. 6

Cameron, L. and M. Shah (2014, January). Can Mistargeting Destroy Social Capital and Stimulate Crime? Evidence from a Cash Transfer Program in Indonesia. *Economic Development and Cultural Change 62*(2), 381–415. Publisher: The University of Chicago Press. 6

Case, A. and A. Deaton (1998, September). Large Cash Transfers to the Elderly in South Africa. *The Economic Journal 108*(450), 1330–1361. 4

Chandola, Y., J. Virmani, H. S. Bhadauria, and P. Kumar (2021, January). Chapter 4 - End-to-end pre-trained CNN-based computer-aided classification system design for chest radiographs. In Y. Chandola, J. Virmani, H. S. Bhadauria, and P. Kumar (Eds.), *Deep Learning for Chest Radiographs*, Primers in Biomedical Imaging Devices and Systems, pp. 117–140. Academic Press. 14

Coady, D., M. Grosh, and J. Hoddinott (2004). *Targeting of Transfers in Developing Countries : Review of Lessons and Experience*. Washington, DC: World Bank. 5

Coady, D. P. (2006, May). The Welfare Returns to Finer Targeting: The Case of The Progresa Program in Mexico. *International Tax and Public Finance 13*(2), 217–239. 5

Duflo, E. (2000, May). Child Health and Household Resources in South Africa: Evidence from the Old Age Pension Program. *American Economic Review 90*(2), 393–398. 4

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Micro–Level Estimation of Poverty and Inequality. *Econometrica 71*(1), 355–364. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00399. 4

Engstrom, R., J. Hersh, and D. Newhouse (2022, May). Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being. *The World Bank Economic Review 36*(2), 382–412. 3

Filmer, D. and L. H. Pritchett (2001, February). Estimating wealth effects without expenditure data–or tears: an application to educational enrollments in states of India. *Demography 38*(1), 115–132. 4

Foster, J., J. Greer, and E. Thorbecke (1984). A Class of Decomposable Poverty Measures. *Econometrica 52*(3), 761–766. Publisher: [Wiley, Econometric Society]. 10, 17

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017, December). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment 202*, 18–27. 12

Grosh, M. E. and J. L. Baker (1995, July). *Proxy means tests for targeting social programs.* The World Bank. 4

Hanna, R. and B. A. Olken (2018, November). Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries. *Journal of Economic Perspectives 32*(4), 201–226. 1, 4

Hoddinott, J. and E. Skoufias (2004, October). The Impact of PROGRESA on Food Consumption. *Economic Development and Cultural Change 53*(1), 37–61. Publisher: The University of Chicago Press. 2, 10, 17, 20

INEGI (2023). Encuesta Nacional de Ingresos y Gastos de los Hogares. type:dataset. 11

Karlan, D. and B. Thuysbaert (2019, February). Targeting Ultra-Poor Households in Honduras and Peru. *The World Bank Economic Review 33*(1), 63–94. 3

Lund, F. (1993). State social benefits in South Africa. *International Social Security Review 46*(1), 5–25. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-246X.1993.tb00358.x. 4

Mirrlees, J. A. (1971). An Exploration in the Theory of Optimum Income Taxation. *The Review of Economic Studies 38*(2), 175–208. Publisher: [Oxford University Press, Review of Economic Studies, Ltd.]. 1

Newhouse, D. L., J. D. Merfeld, A. Ramakrishnan, T. Swartz, and P. Lahiri (2022, October). Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning. 3

Nichols, A. L. and R. J. Zeckhauser (1982). Targeting Transfers through Restrictions on Recipients. *The American Economic Review 72*(2), 372–377. Publisher: American Economic Association. 3

Ohlenburg, T., J. Pinxten, D. Fricke, and F. Caccioli (2022, December). Fewer Questions, More Answers: Truncated Early Stopping for Proxy Means Testing. 4, 6

Parker, S. W. and P. E. Todd (2017, September). Conditional Cash Transfers: The Case of *Progresa/Oportunidades. Journal of Economic Literature 55*(3), 866–915. 10

Ravallion, M. and Q. Wodon (1999). Poor Areas, or Only Poor People? *Journal of Regional Science 39*(4), 689–711. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-4146.00156. 4

Rolf, E., J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang (2021, July). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications 12*(1), 4392. Number: 1 Publisher: Nature Publishing Group. 7

Schultz, T. P. (2004, June). School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of Development Economics 74*(1), 199–250. 10

Skoufias, E. (2005). PROGRESA and its impacts on the welfare of rural households in Mexico. Research Report 139, International Food Policy Research Institute, Washington D.C. 16

Skoufias, E., B. Davis, and Jere Behrman (1999). An evaluation of the selection of beneficiary households in the Education, Health, and Nutrition Program (PROGRESA) of Mexico. *IFPRI Discussion Paper*. 8, 11, 17

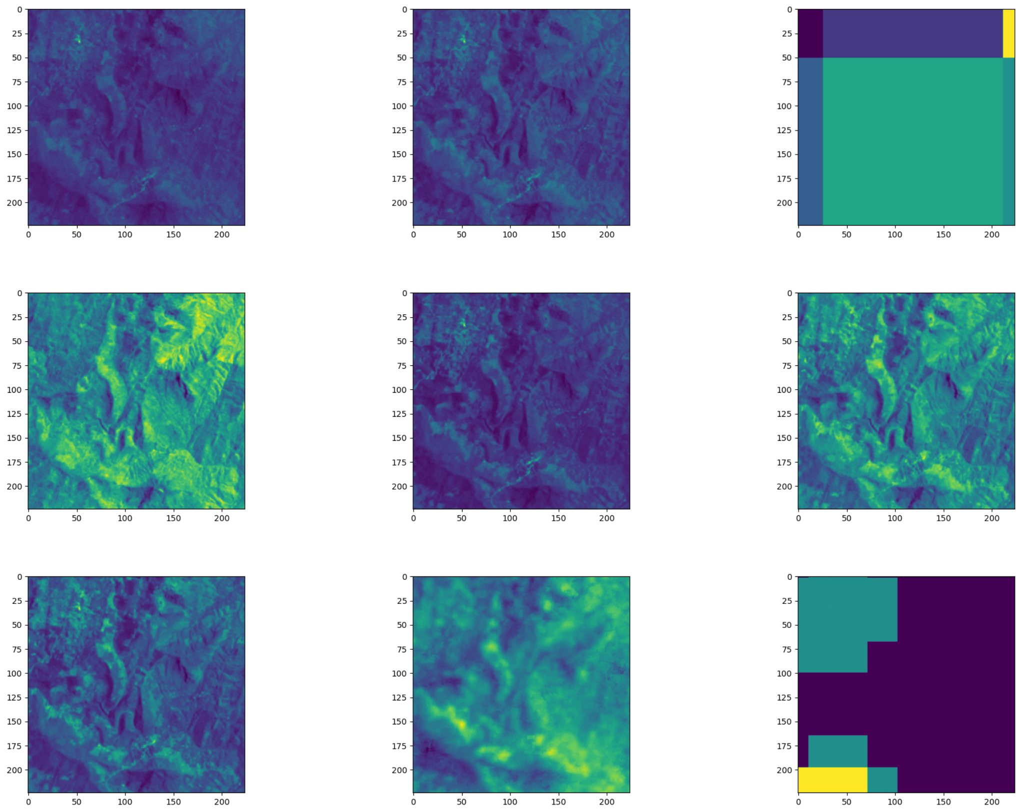
Skoufias, E., S. W. Parker, J. R. Behrman, and C. Pessino (2001). Conditional Cash Transfers and Their Impact on Child Work and Schooling: Evidence from the PROGRESA Program in Mexico [with Comments]. *Economía 2*(1), 45–96. Publisher: Brookings Institution Press. 2, 9, 10

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020, May). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications 11*(1), 2583. Number: 1 Publisher: Nature Publishing Group. 5, 14, 18, 23

Zhao, C., X. Cao, X. Chen, and X. Cui (2022, July). A consistent and corrected nighttime light dataset (CCNL 1992–2013) from DMSP-OLS data. *Scientific Data 9*(1), 424. Number: 1 Publisher: Nature Publishing Group. 12

A Appendix A: Example Images

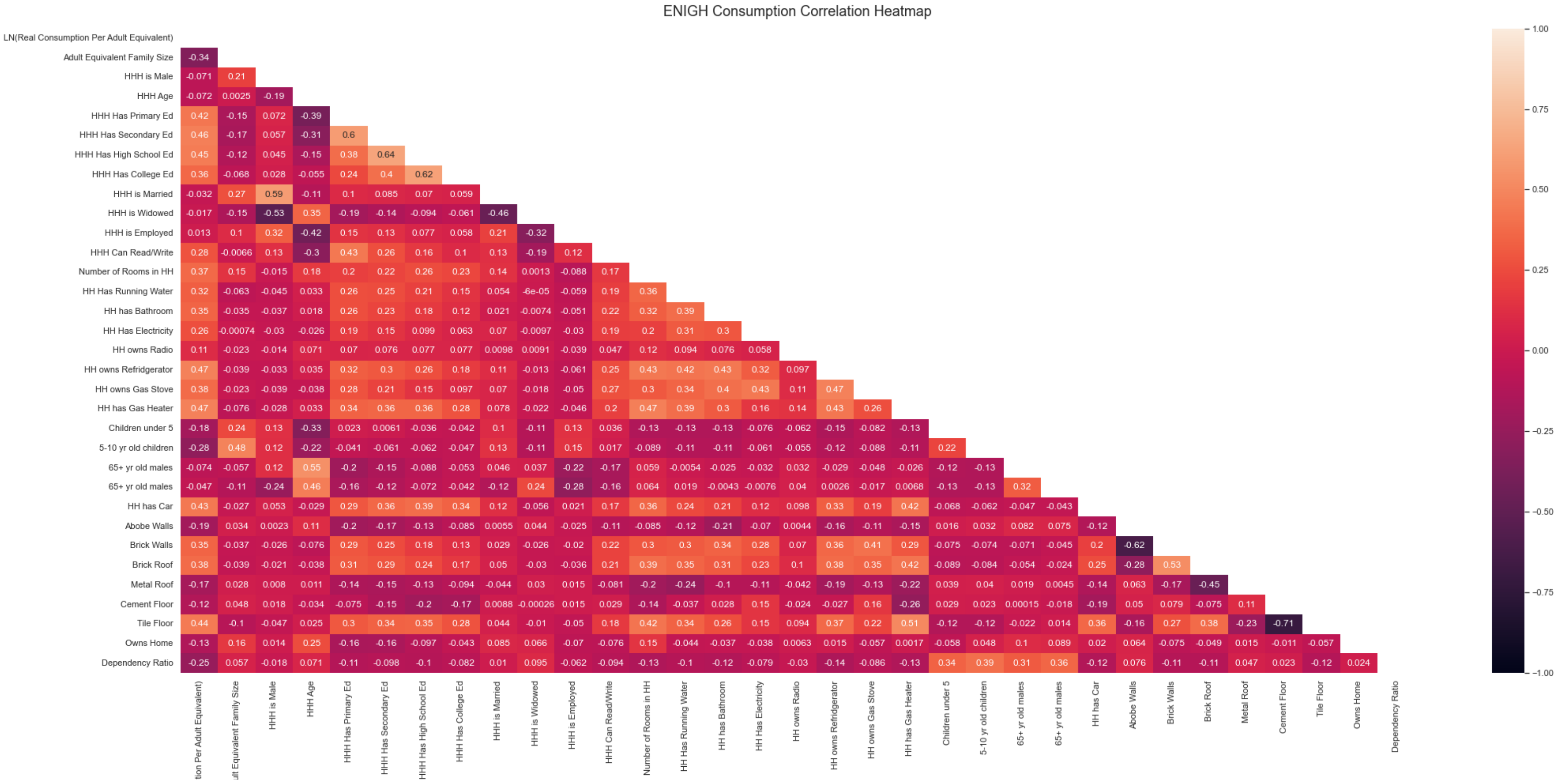
Figure A1: Example satellite images



Notes: This figure displays satellite images from one randomly selected village in the sample. Going from left to right and from top to bottom, the bands are Blue surface reflectance, Green surface reflectance, NDVI, Near infrared surface reflectance, Red surface reflectance, Shortwave infrared surface reflectance 1 (1.55-1.75 um), Shortwave infrared surface reflectance 2 (2.08-2.35 um), Brightness temperature, and Nightlights

B Appendix B: ENIGH Description

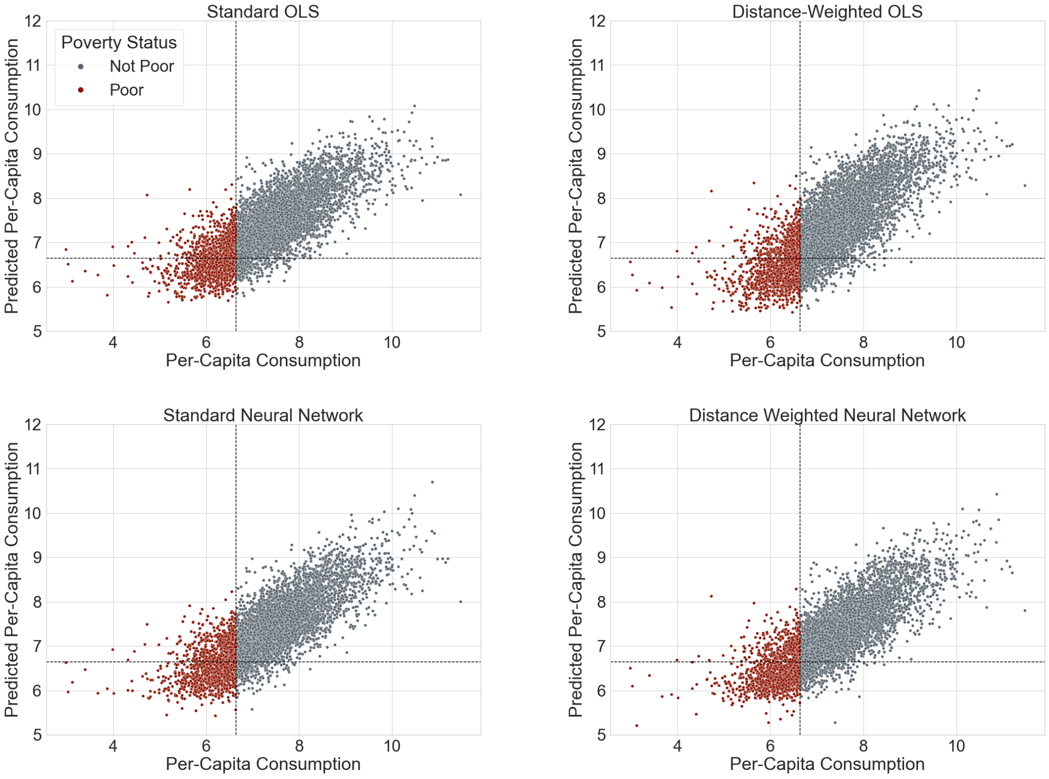
Figure B1: ENIGH Feature Correlations



Notes: Heatmap of correlations between variables found in ENIGH data used to train the neural network.

C Appendix C: Model Performance

Figure C1: ENIGH Model Prediction Scatter Plots



Notes: Scatter plots of true vs. predicted per-capita consumption. Axis lines denote the poverty line – bottom left quadrant are the predicted and truly poor; top left are truly poor but predicted non-poor; top right are truly non-poor and predicted non-poor; bottom right are truly non-poor but predicted poor.