Policy Research Working Paper 10175

Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning

David Newhouse
Joshua D. Merfeld
Anusha Pudugramam Ramakrishnan
Tom Swartz
Partha Lahiri

WORLD BANK GROUP
Development Economics
Development Data Group
September 2022

Policy Research Working Paper 10175

Abstract

Estimates of poverty are an important input into policy formulation in developing countries. The accurate measurement of poverty rates is therefore a first-order problem for development policy. This paper shows that combining satellite imagery with household surveys can improve the precision and accuracy of estimated poverty rates in Mexican municipalities, a level at which the survey is not considered representative. It also shows that a household-level model outperforms other common small area estimation methods. However, poverty estimates in 2015 derived from geospatial data remain less accurate than 2010 estimates derived from household census data. These results indicate that the incorporation of household survey data and widely available satellite imagery can improve on existing poverty estimates in developing countries when census data are old or when patterns of poverty are changing rapidly, even for small subgroups.

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

Small Area Estimation of Monetary Poverty in Mexico Using Satellite Imagery and Machine Learning[*]

David Newhouse[§]
Joshua D. Merfeld[†]
Anusha Pudugramam Ramakrishnan[‡]
Tom Swartz[§]
Partha Lahiri[+]

Keywords: *poverty, small area estimation, poverty mapping, satellite data, machine learning*

JEL codes: C53, C83, I32

[*] We are indebted to Boris Babenko and Jonathon Hersh who made substantial contributions to previous versions. We thank Shiva Makki and the World Bank's Research Support Board for financial support and James Crawford and Orbital Insight, Inc. for support in providing AGEB-level estimates of land classification and poverty. We thank Keith Garrett and seminar participants at the Northeast University Development Conference, the World Bank, the KDI School, and Sungkyunkwan University for helpful comments. All remaining errors are ours.
[§] Development Economics Data Group, World Bank, and IZA
[†] KDI School of Public Policy and Management, IZA
[‡] Consultant, World Bank
[§] Change Research
[+] Joint Program in Survey Methodology and Department of Mathematics, University of Maryland, College Park

1 Introduction

In 2015, all United Nations members adopted the 2030 Agenda for Sustainable Development, which ushered in a new set of sustainable development benchmarks – known as the Sustainable Development Goals (SDGs) – across the world.[1] These SDGs supplanted the previous Millennium Development Goals (MDGs),[2] adding many new goals, including new goals related to the environment. However, one thing did not change with the adoption of the SDGs: Goal 1.1 was still the eradication of extreme monetary poverty.

Key to addressing progress towards the new poverty goals, however, is the accurate measurement of poverty (e.g., Ravallion, 2015). To this end, many household surveys in developing countries now include consumption and/or expenditure modules, designed to measure a specific type of monetary poverty. For example, Malawi's Fifth Integrated Household Survey (IHS5) was a major input into the 2020 Malawi Poverty Report (NSO Malawi, 2021),[3] in large part due to its consumption module.

While household surveys like Malawi's IHS remain key to measuring poverty, difficulties remain. First, large household surveys are expensive and, as such, generally are fielded only every few years, at most. Second, surveys are designed to be representative at certain levels of aggregation. For example, in Mexico, the Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) is designed to give representative estimates for rural and urban areas across states, but not for most municipalities.[4] The inability to consistently estimate more disaggregated statistics of poverty can make targeting difficult, especially when poverty is found in pockets below the level of representativeness. Moreover, many SDGs disaggregate progress by groups (e.g., women and children), which can lead to additional difficulties with reliably estimating these statistics, since many of these subgroups may have relatively few observations. This is also true for the non-poverty benchmarks, like access to health facilities. Therefore, improving the accuracy and precision of survey estimates in these contexts is a first-order question for many development economists and practitioners. In this paper, we employ a common statistical method – small area

[1] https://sdgs.un.org/goals
[2] https://www.un.org/millenniumgoals/
[3] https://microdata.worldbank.org/index.php/catalog/3818/related-materials
[4] https://www.inegi.org.mx/programas/enigh/nc/2018/

estimation (SAE) – and show that it can provide more disaggregated poverty estimates even when using only remotely-sensed satellite data.

This is an important finding, since traditional SAE methods require census data that is often not available in many developing countries or is available only sporadically across time. "Big data" from satellites, on the other hand, are widely available for most countries and is often available in near real-time. Recent advances in the availability of satellite and mobile phone data, as well as increases in computing power, have sparked great interest in combining these data with survey data to generate more precise estimates of socio-economic characteristics (World Bank, 2021; Burke, 2021). Traditionally, small area estimation has been carried out by utilizing survey data to estimate a model and simulating that model in household-level census data (e.g., Elbers, Lanjouw, and Lanjouw, 2003). In the last five years, however, a growing body of innovative research has used geospatial or other big data to predict poverty and generate purely synthetic predictions of poverty or welfare, often generated by convolutional neural networks (Jean et al., 2017; Yeh et al., 2020; Engstrom et al., 2022; McBride et al. 2021). These estimates are typically validated against survey data at the village level, using sample data withheld from the prediction model. The results demonstrate the ability of methods utilizing novel sources of data to extrapolate poverty estimates to countries with no sample data with a reasonable degree of accuracy.

The existing literature is, however, less informative about the performance of methods that seek to combine survey and geospatial data to generate more granular estimates within a country, for two reasons. First, much of the existing literature stops short of combining synthetic predictions with prior estimates based on survey data, which is important to generate more accurate and precise estimates for well-defined administrative levels (Masaki et al., 2022). Second, it is challenging when utilizing neural networks or other machine learning models to properly estimate uncertainty, which is crucial when evaluating the reliability of small area estimates. This paper aims to address both of these issues.

A few studies examine the extent to which combining survey data with geospatial data improves small area estimates of proxies related to household welfare, with generally encouraging results (Steele et al., 2017; Masaki et al., 2022). No studies to date, however, have examined the gain in both accuracy and efficiency from combining survey and remote sensing data for predicting monetary poverty for small areas. Furthermore, we know of no study that has compared predictions

derived from a combination of current survey and geospatial data with older, traditional census-based small area estimates. These are important knowledge gaps given that subnational monetary poverty estimates play a significant role in resource allocation in many countries.

In addition, little is known about how the accuracy of predictions compares for target areas – the area at which we wish to produce estimates – that are covered by the sample and those target areas that lie outside the sample. This distinction is important for two reasons. First, in some contexts, the official household survey used for poverty measurement may not cover all target areas. If predictions for non-sampled areas are inaccurate, this would strengthen the case for both exercising caution when utilizing out-of-sample estimates for policy and for expanding survey data collection to cover all target areas. Second, comparing the accuracy of non-sampled and sampled areas sheds some light on the benefits of using Bayesian or Empirical Bayesian methods, which utilize the sample data as a prior to inform the estimates, rather than relying only on purely synthetic predictions.

Finally, the optimal statistical method to use when combining survey and satellite data remains unsettled. Bayesian methods effectively combine sample and big data in a well-established framework (Steele et al., 2017; Steele et al., 2021; Pokhriyal and Jacques, 2017). There are significant barriers, however, to their widespread use. These include computational complexity, a lack of familiarity among many practitioners, and required distributional assumptions on priors and hyperpriors that introduce additional complications. An appealing alternative is to use Empirical Best Predictor (EBP) models – a name coined by Jiang and Lahiri (2001) and Jiang et al. (2002) – in part because it is a well-established method in the existing literature on small area estimation.[5]

This paper makes four main contributions. First, it assesses the gain in accuracy and precision, relative to direct survey estimates, from combining survey data with geospatial indicators to generate small area estimates of monetary poverty. Second, it compares the performance of different methods for generating estimates for both in-sample and out-of-sample municipalities –

[5] EBPs are also known as EBLUP (Empirical Best Linear Unbiased Predictor) for linear mixed models. See Battese, Harter, and Fuller (1988), Jiang and Lahiri (2008), Molina and Rao (2010), and Van der Weide (2014) for more on their use in small area estimation.

the target areas – in Mexico.[6] Third, it assesses whether the combination of survey and geospatial data generate estimates that improve on small area estimates generated using a census enumerated five years prior as well as a widely available wealth index created by Meta (Chi et al., 2022).[7] Finally, it compares estimates from three different types of small area estimation methods: a household unit-context model that models household per capita income as a function of sub-area and area-level predictors, a sub-area model that utilizes a nested error model at the level of the sub-area, and a Fay-Herriot area-level model (Fay and Herriot, 1979).

We consider these questions in the context of Mexico. Mexico is a useful case study for this analysis because of the richness of publicly available official data. In the Mexican context, sub-areas are AGEBs (Area Geoestadística Básica) and areas are municipalities. The official MCS-ENIGH household survey provides representative statistics at the state level, motivating the use of small area methods to generate municipal estimates. INEGI, the Mexican statistical agency, fielded a large intercensus sample of 5.8 million households in 2015. CONEVAL, the agency that produces poverty statistics in Mexico, used these data to generate official municipal level estimates of monetary poverty, which can be used as a benchmark for evaluation. CONEVAL employed an EBP using a model that included an extensive set of household and municipal characteristics. These type of small area estimates, generated using household-level auxiliary data, are rarely available between census rounds in low- and middle-income countries. They therefore provide a natural benchmark to evaluate whether combining survey and geospatial data produces municipal poverty estimates that are like those produced by a best-practice poverty mapping exercise, based on household data from a large intercensus.

We compare these 2015 benchmark estimates with five sets of municipal poverty estimates. The first is direct sample estimates taken from the 2014 survey. We generate the second, third, and fourth sets of estimates by combining the 2014 sample with geospatial indicators and estimating a household-level "unit-context" model of welfare, a sub-area model of poverty rates, and an area-level model of poverty rates. The fifth and final set of estimates are the official poverty estimates

[6] In other words, municipalities are the level at which we wish to estimate poverty. This is below the level of representativeness of the survey.
[7] https://dataforgood.facebook.com/dfg/tools/relative-wealth-index

for 2010, which were derived by CONEVAL from the 2010 survey and census. Six main results emerge:

1. Combining satellite indicators with household survey data substantially improves the accuracy of municipal poverty estimates relative to estimates using the survey data alone. In the preferred specification, correlation with the benchmark official estimates rises from 0.8 to 0.86 and root mean squared deviation falls 25 percent.

2. Incorporating geospatial predictors greatly improves the precision of the estimates compared with the direct survey estimates. The median coefficient of variation declines from 38.5 to 19.8 in the household level model, which rises to 25 if the mean squared error estimates are adjusted to maintain equal coverage rates across methods. Even after this adjustment, the additional precision achieved by the small area estimation procedure is equivalent to increasing the size of the sample by a factor of roughly 2.4.

3. The household-level model moderately underestimates uncertainty and yields a coverage rate of 77 percent for in-sample municipalities and 83 percent of out-of-sample municipalities. For in-sample municipalities, this is moderately lower than the 86 percent obtained using the direct survey estimates when using the Horvitz-Thompson variance estimator, but much larger than the 39 percent coverage rate when using the standard variance estimator clustered on enumeration areas, which greatly underestimates uncertainty.

4. For all satellite-based estimates, in-sample predictions are substantially more accurate and much more precise than out-of-sample predictions. The correlation between the satellite-based estimates and the official benchmark estimates declines from 0.86 for in-sample municipalities to 0.70 for out-of-sample municipalities, and the median coefficient of variation rises from 19.8 to 33.9.

5. Estimates that combine survey and geospatial data using a household model are more accurate and much more precise than those produced by a sub-area model for in-sample municipalities (correlation of 0.86 vs 0.83, median CV of 19.8 vs. 35.6). The sub-area model estimates, in turn, are more accurate but somewhat less precise than the area-level model estimates for in-sample areas (correlation of 0.83 vs. 0.80, median CV of 35.6 vs

28.1). For out-of-sample municipalities, the household and sub-area models are equally accurate (correlation of 0.7), while the area-level model is moderately less accurate (correlation of 0.66).

6. All geospatial estimates are substantially less accurate than the official 2010 municipal poverty estimates when predicting 2015 municipal poverty rates, particularly for out-of-sample municipalities. The correlation between the 2010 estimates and the 2015 benchmark is 0.91 in-sample and 0.90 out of sample, as compared with 0.86 and 0.70, respectively, for the preferred household model geospatial estimates. This suggests that poverty in Mexico was relatively stable, at least between 2010 and 2015.

Overall, the results confirm that combining survey data and geospatial data can greatly improve the accuracy and precision of monetary poverty estimates. However, accuracy and precision both suffer markedly when making synthetic predictions into out-of-sample municipalities. This underscores the importance of using a statistical method that conditions predictions on the sample data and of including as many target areas as possible in the sample. In this context, household-level models generate more accurate and precise predictions than the basic sub-area and area models considered here.[8] The household-model results, however, are substantially less accurate for predicting 2015 municipal poverty than the official 2010 census-based estimates. This correlation may be overstated because data from the 2010 census are used to generate both the 2010 and 2015 estimates. Nonetheless, this finding underscores the value of regular census data collection in this context and the possible importance of designing surveys to include all target areas if successful small area estimation is one goal of the survey.

While predictions that combine survey and satellite data are not always more accurate than older census-based poverty maps for predicting current poverty rates, they greatly improve upon estimates solely based on survey data and are a viable second-best solution in settings where recent census-based estimates do not exist, which is a common situation in much of the developing world. The relative accuracy of estimates produced from survey and geospatial data, compared with older census-based estimates, depends on several context-specific factors. These include the pace of change in regional patterns of poverty since the previous census, the size of the sample, and the

[8] More complex and sophisticated versions of the area and sub-area levels, such as those that transform the dependent variable, may improve on the accuracy of the simple versions considered here.

predictive power of the auxiliary geospatial indicators. Additional research, drawing on data from different contexts, will help inform the choice between using more recent estimates obtained by combining survey and geospatial data, and older small area estimates derived from a census.

2. Small area estimation

Small area estimation deals with methods to reliably estimate parameters of interest in areas without adequate sample sizes (Molina and Rao, 2015). Here, we discuss the basics of small area estimation, with a particular focus on the Mexican context we study in this paper. Interested readers can find more details in Rao (2003), Molina and Rao (2010, 2015), and Jiang and Rao (2020). Ghosh (2020) details the evolution of small area estimation over the last five decades.

In Mexico, it is worth differentiating four separate levels in the data: households, AGEBs, municipalities, and states. Households are nested within AGEBs, AGEBs are nested within municipalities, which are nested within states. We are interested in estimating poverty rates for municipalities – the target areas – but the only national surveys available are representative at urban-rural areas within states. Therefore, although these estimates are unbiased over repeated samples, a single municipality-level sample is generally too small to provide a reliable estimate of poverty rates.

To estimate these poverty rates, we need to build a model at the level of the municipality or lower. In theory, we could directly estimate poverty rates at the municipality level – that is, taking poverty rates at the municipality level and estimating the model using municipalities as the unit of analysis – which is referred to as an area-level model. We could also estimate poverty in AGEBs and then aggregate AGEB-level poverty rates up to the municipality based on AGEB populations, referred to as a sub-area model. Finally, we could estimate poverty at the household level and then aggregate household-level poverty rates up to the municipality, using a household-level (or unit-level) model.

These methods share some characteristics but also offer different advantages. One similarity across all three methods is that they combine survey data with auxiliary data. This is distinct from much of the recent literature on satellite estimation of poverty (see, for example, Blumenstock et al. (2015)), which tends to estimate poverty using a purely synthetic regression prediction. Small area

estimation methods, on the other hand, "borrow strength" across areas through a regression framework that is quite flexible and can incorporate different types of data, including census data, remote sensing data, and administrative data, to name but a few. In traditional unit-level models like Elbers, Lanjouw, and Lanjouw (2003), the key predictors were household-level values derived from a census. However, this requires that there exists a (recent) census with variables also available in a survey with monetary poverty measures. This is a restrictive requirement, as censuses like this are few and far between in many developing countries. In this paper, we instead use AGEB or municipality aggregate variables when we implement the household-level models. Implementing a household or unit-level model with only aggregate variables is sometimes referred to as a "unit-context" model.

In all three methods discussed here, poverty or welfare is regressed on a selection of predictive features and random effects, which we discuss in more detail in the methods section and the appendices. Thus, each in-sample target area has two separate estimates: the direct survey estimate and the synthetic prediction estimate created through the regression. The key to small area estimation is the combination of these two separate estimates, based on relative variances. Intuitively, when the sample estimates are more precise for an area relative to synthetic predictions, there is "less strength" that can be borrowed from the prediction and the direct sample estimates are given more weight in the estimate. When the sample estimates are less precise relative to the synthetic predictions, on the other hand, areas can "borrow more strength" through the regression (Jiang and Rao, 2020). However, note that not all municipalities in Mexico appear in our sample. As such, this means that some municipalities do not have a direct estimate. For these out-of-sample areas, the sample contains no information and therefore gets zero weight in the estimate, which is a purely synthetic prediction from the regression.

The three methods also differ in sometimes important ways. Sub-area and area-level models can be estimated more quickly and easily than unit-context household models. This comes at a price, however. Area-level models require estimates of the sample variance for each area, which can be unstable when derived from a survey, since the survey is not designed to provide adequate sample size to generate reliable estimates at the area level (Bell, 2008). This instability can be alleviated by variance smoothing, but that adds an additional layer of complexity to the estimation procedure. The sub-area models considered in this paper, meanwhile, require a non-trivial reweighting exercise. This entails both correcting for heteroscedasticity when the number of sample households

varies across AGEBs, and normalizing the weights appropriately within each municipality. Second, area-level models cannot produce estimates at lower levels of geography, such as sub-areas. In addition, area-level and sub-area models of the type implemented here model poverty rates rather than welfare, therefore discarding information on whether non-poor households in a sub-area are near or far from the poverty line. Household-level models predict welfare instead of poverty rates, conditional on the sample, and therefore better preserve this information.

Currently available software for estimating area-level also models cannot easily accommodate different weights for different areas. This can become a constraint when the analyst wishes to give more weight to more populous areas when estimating the model, which may be preferred in some policy settings. Finally, both the standard sub-area and area-level models considered here impose the assumption that poverty rates are a linear function of the predictors, which can also reduce prediction accuracy.

Corral et al. (2021) note that the nested error regression model we use in the household-level model is subject to "omitted variable" model bias, when household-level auxiliary variables are replaced by their respective area-level population averages and when the true data generating process is based on a household-level model. This bias occurs when the population averages used as predictor variables on the right-hand side are correlated with the discrepancy between the area-level means of the population and the particular sample used. As a result, they recommend using area-level models when feasible rather than household-level models. However, as detailed in Appendix F, the exact same omitted variable bias also appears in area-level models, under the same assumptions that the data generating process occurs at the household level and that the predictors are drawn from the population.[9] Thus, because the same omitted variable bias appears in both household and area-level models, this bias is not a sufficient justification for preferring area-level models over unit-context models.

There are two other reasons why the omitted variable bias identified in Corral et al (2021) may not be important in practice. First, if sampling error is exogenous, as one would expect with a random sample, both the household and area level models are consistent, implying that this bias will tend towards zero as the number of AGEBs included in each area becomes large. Furthermore, in both

[9] Auxiliary data in area-level models is typically drawn from administrative data or other sources that represent the population rather than the sample.

household and area models, when one defines bias with respect to both sample design (in our case simple random sampling within AGEB) and the model parameters, the design-model bias vanishes. This occurs because the sample mean of the predictor variables is an unbiased estimator of their population means, under a simple random sample design, and follows from the concept of design-model bias discussed in Valliant et al. (2000, p. 120).

In many applications, it is difficult or even impossible to apply a nested error regression model with household-level auxiliary variables, either because household-level auxiliary variables are not available for the entire finite population or due to the presence of measurement error in the household-level auxiliary variables. In these situations, the unit-context model would be a reasonable option. Like a nested error regression model with household-level auxiliary variables, a unit-context model is very flexible in producing estimates of a wide range of linear and nonlinear parameters (e.g., different FGT indices) using a single model. This is in sharp contrast with an area or sub-area level model, which requires separate models for each estimated parameter. The unit-context model also permits incorporation of uncertainty incurred due to estimation of different model parameters, such as the variance of the sampling error.

3. Data

We evaluate the ability of data derived completely from satellite imagery to predict municipality-level poverty in Mexico. In this section, we discuss the different data sources used to construct the requisite variables as well as the data used to validate the performance of the models.

3.1. Household data

We conduct the main analysis using monetary poverty estimates at the household level from the 2014 MCS-ENIGH Household Survey data. The survey is conducted every two years by the Mexican National Institute of Statistics and Geography (INEGI) to understand the income, spending, socio-demographic, and employment situation of the country. It is a nationally representative, cross-sectional household survey. The 2014 MCS-ENIGH survey covers 58,125 households, of which approximately 75% are urban and 25% are rural.[10] The survey samples 896

[10] Rural areas are defined as localities with fewer than 2,500 inhabitants.

municipalities (out of 2,457 municipalities) and 6,849 AGEBs.[11] Table A1 presents additional details on the geographic structure of Mexico.

Importantly, the publicly available data contains an AGEB-level identifier, which can be linked with the official AGEB shapefile. For rural areas, the data and associated shapefile identify localities, which are points rather than polygons. These were grouped together using the same AGEB shapefile into rural AGEBs. The survey collects income per capita, which is the welfare metric used to calculate the official poverty rate. With the AGEB-level identifiers and merged shapefile, we can link the survey data with predicted poverty rates and land classifications derived from satellite imagery.

To predict poverty for the entire country, we need data on the location of all households. For this, we create a "synthetic census" based on the 2010 census. For each AGEB, we estimate the number of households by dividing the total AGEB population in the 2010 census by the average household size, at the state-urban/rural level, in the 2014 household survey. We round to the nearest integer and use this to construct a synthetic census that links the "households" in an AGEB with the AGEB-level satellite indicators, which serve as the auxiliary data for the household model. This resulting synthetic census consists of just under 28 million records. In addition to the satellite imagery – which serves as the auxiliary data for the prediction exercises – we also include average household size at the state-urban/rural level, estimated with the 2014 MCS-ENIGH survey, which we use as population weights when aggregating across households to generate municipal poverty estimates.

3.2. Benchmark CONEVAL estimates

We compare the estimated poverty rates with a benchmark to evaluate the ability of the satellite imagery to improve poverty predictions. The benchmark used for evaluation is the municipal income poverty estimates, generated by CONEVAL using a combination of the 2014 MCS-ENIGH household survey and the 2015 Intercensus survey. The Intercensal survey is conducted every 5 years between two censuses, to update socio-demographic information at the national and

[11] AGEBs (Áreas Geoestadísticas Básicas) are equivalent to sub areas in this context, while municipalities are target areas, or the areas at which we are interested in predicting poverty. Each municipality is made up of AGEBs, with larger municipalities generally having more AGEBs.

sub-national levels. The 2015 Intercensus sampled 5.8 million households, 56 percent of whom lived in urban areas. The Intercensus contains only household labor income and transfer income, and not total household income, necessitating the use of a small area estimation model to estimate municipal-level poverty rates.

CONEVAL creates these estimates using a unit-level model, with a large set of household and individual-level auxiliary variables, including demographic, labor, housing quality, and municipal characteristics at the individual, household, and municipal level. The municipal characteristics include average per capita income and other poverty estimates. The municipal estimates and detailed documentation on the methodology are available at the CONEVAL website.[12] The 32 Mexican states are divided into 6 groups[13] and a model is estimated for each, with R^2 values ranging from 0.52 to 0.57.

CONEVAL used the 2014 MCS-ENIGH survey in an empirical best household model to construct the small area municipal estimates. This implies that both the estimates that are evaluated below, based on geospatial data, and the CONEVAL benchmark data use the same 2014 survey. However, the R^2 of the unit-level model used to generate the CONEVAL estimates are roughly three times as high as the model used to generate the geospatial-based estimates, since the auxiliary data used by CONEVAL vary at the household level. Therefore, the CONEVAL estimates give far more weight to the predicted values and less weight to the survey, relative to the estimates derived from the satellite data. This reduces the mechanical correlation between the estimates and the benchmark due to the municipal random effects being conditioned on the same 2014 survey data. Furthermore, as a robustness check, we also compare the CONEVAL estimates with geospatial small-area estimates generated using the 2016 MCS-ENIGH survey, which eliminates all mechanical correlation due to the use of the same sample data.

3.3. CNN-Generated Auxiliary Geospatial Data

A key difference between this paper and Masaki et al. (2022) is the nature of the auxiliary data. We train convolutional neural network (CNN) models against high resolution satellite imagery

[12] https://www.coneval.org.mx/Medicion/Paginas/Pobreza-municipal.aspx
[13] The six state groups we use in this paper match those used to develop the official municipal poverty estimates and are listed in Table C1 in appendix C.

from Planet – which features spatial pixel resolution of 3 to 5 meters – and use those predictions as inputs into the small area estimation model. We make use of two different CNNs: one to estimate the poverty rate (defined as extreme or moderate headcount poverty according to the CONEVAL definition) and one to recognize the land classification type of an image. We train the former using the 2014 MCS-ENIGH. Land classification subdivides images into coverage by building, road, water, grassland, forest, and background (all other classes). Because CNNs require square tiles of imagery for prediction, we train the CNN to estimate poverty or land type at the tile level, which are roughly 750 square meters each. The CNN predicts poverty or land type at this tile, and we use an area-weighted sum to aggregate the predictions to the AGEB – or sub-area – level. We then aggregate these AGEB-level predictions to municipalities, weighting by the 2010 census population.[14] The auxiliary information that augments the small area estimation models is therefore AGEB-level and municipal-level CNN predictions of poverty or land type. Summary statistics for the CNN generated auxiliary spatial data are shown in Table 2, and more technical details of the CNN are presented in Appendix A.

Intuitively, the poverty CNN uses all available visual characteristics in an image to predict extreme or moderate poverty as defined by the 2014 MCS-ENIGH. If these visual characteristics reliably predict poverty, then CNN model predictions can be extrapolated to areas not covered by the MCS-ENIGH and used to estimate poverty predictions for these areas. Adding this additional information should allow a small estimation model to have lower variance and higher predictive performance. While Babenko et al. (2017) demonstrate the performance of the CNN poverty predictions against a benchmark, incorporating this information into a small area estimation model improves the predictions in three ways. First, it enables the estimates to also benefit from auxiliary data on land classification. Second, it conditions the predictions on sample information through the empirical best predictor model, making the predictions more accurate. Finally, it enables the appropriate estimation of uncertainty, in the form of mean squared error estimates, from which one can construct reasonably accurate confidence intervals.

[14] An alternative when there is no census available is to use estimated population from WorldPop (https://www.worldpop.org/), which seems to perform well in other applications (e.g., Merfeld et al., 2022).

4. Methods

The key survey data we use is the 2014 MCS-ENIGH. Although we are interested in estimating poverty for all municipalities in Mexico, the survey is not designed to estimate poverty at the municipality level, but rather for urban and rural areas within states. Therefore, the direct estimates from the survey will not consistently estimate poverty accurately for municipalities.

The rest of this section describes how we construct the direct estimates and the household-model estimates, as well as two robustness checks using sub-area and area models.

4.1 Direct estimates

The direct survey estimates for each municipality are the weighted mean, across sample households, of a dummy variable indicating whether the household's per capita income falls below the official poverty line, which is defined separately for urban and rural areas. These are taken for each municipality, using the product of the household sample weight and household size as weights. Estimating uncertainty properly is a bit more complicated. This is because standard methods assume that per capita income is independent across primary sampling units, which can significantly underestimate the uncertainty associated with municipal poverty estimates. We therefore use two methods for estimating uncertainty: the Horvitz-Thompson approximation (Horvitz and Thompson, 1952) and the standard method of clustered standard errors, using AGEBs as the clustering unit (Huber, 1967). As shown below, the Horvitz-Thompson approach gives much more accurate estimates of uncertainty and is therefore preferred.

4.2 Household model estimates

The baseline method for the household model is similar to the method used to predict non-monetary poverty rates in Masaki et al. (2022), in that it relies on a conditional random effect model, which is also often referred to as a mixed model. Many previous iterations of household-level models rely on detailed census data (e.g., Elbers, Lanjouw, and Lanjouw (2003) and Tarozzi and Deaton (2009)). However, such data is usually not available for many developing countries, while satellite imagery is. As such, instead of using household-level covariates to predict poverty,

we use satellite-derived features defined at higher levels of aggregation, such as the AGEB or the municipality.

There are two notable differences between this paper and Masaki et al. (2022). First, the household-level model implements sample weights, both when estimating the mixed effect model and when calculating the empirical best predictors. Introducing sample weights decreases the precision of the sample data relative to the prediction and therefore gives the sample less weight when conditioning the mixed effect on the sample data, as described in Appendix C.[15] Second, because urban and rural poverty lines differ in Mexico, the baseline specification in this paper inflates the sample welfare measure in rural areas by the ratio of the urban to rural poverty lines. This enables us to apply a constant poverty line threshold in the simulation stage.

The estimated model is:

$$G(Y_{smai}) = \beta_1 X_{sma} + \beta_2 X_{sm} + \beta_3 X_s + \eta_{sm} + \varepsilon_{smai}, \quad (1)$$

where $G(Y_{smai})$ is transformed per capita income for a household i in AGEB a, which is located in municipality m and state s. X_{sma} is a vector of AGEB aggregates, X_{sm} is a vector of municipal aggregates, X_s is a vector of state dummies, η_{sm} is a municipality-level random effect, and ε_{smai} is a classical error term.

The estimation process for the household model involves the following eight steps, with additional details in Appendix C:

1. Classify households in the MCS-ENIGH as poor if their per capita income, ictpc, falls below the monetary poverty lines of 1,242.6 pesos per month for urban households, or 868.25 pesos per month for rural households.
2. Multiply the ictpc of rural households by the ratio of the urban to rural poverty lines, which is approximately 1.4314. This step allows us to compare the vector of deflated per capita incomes against a single line, the urban poverty line.[16]

[15] As in Masaki et al. (2020), we normalize weights to sum to the number of sample observations in each municipality. Unlike in standard regressions, the results of mixed models depend on the scale of the weights and normalizing to the sum of sample observations is a common approach.

[16] Alternatives to this procedure, including estimating separate urban and rural models, are considered in section 6 below.

3. Normalize the resulting vector of welfare using Ordered Quantile Normalization, as described in Peterson and Cavanaugh (2020) and implemented in the R bestNormalize package.
4. Set the poverty line as the minimum value of normalized welfare for non-poor households, as defined in step one. Since we are only interested in estimating headcount poverty, we do not need to back transform to recover the original welfare measure. Instead, we can simply compare a household's predicted location in the normalized distribution to the location of the poverty line.
5. Select a model from the pool of candidate variables using the "plug-in LASSO" estimator implemented in Stata 16, allowing for heteroscedasticity, with the dependent variable equal to normalized welfare (Belloni and Chernozhukov, 2013). We use the "plug-in" lasso to be sure to avoid overfitting, since it selects more parsimonious models than those selected using cross-validation (Statacorp, 2021). The weights in the LASSO regression are the sample population weights (household weight times household size), normalized to sum to the number of observations in each municipality. The candidate set of variables include the rates of moderate and extreme poverty predicted by the CNN at both the AGEB and municipality level, predicted land classifications at both the AGEB and municipal level, a dummy for whether the household's AGEB is urban or rural, the share of the population in the municipality that are rural, and state dummies.
6. Using the model selected in step 5 and the poverty line calculated in step 4, estimate headcount poverty rates for each municipality using a modified version of the EMDI software package in R. The modification allows for the use of sample and population weights, as described in Appendix C. The "population" file is based on the synthetic census, constructed as described in the data section.
7. Benchmark the estimates by multiplying the estimated headcount rates and the estimated root mean squared error by a scaling factor calculated for each state. The scaling factor is calculated as a ratio, where the numerator is the official state headcount estimate taken from the survey data. The denominator, meanwhile, is the average, across each state, of the small area municipalities that are contained in the MCS-ENIGH survey, weighted by the sum of the original sample weights for each municipality. These scaling factors are calculated based on the small area estimates from municipalities included in the survey

sample, to preserve legitimate differences in the estimated state poverty rates due to including auxiliary data from non-sampled municipalities.

8. Compare the benchmarked estimator to the official estimates produced by CONEVAL, as described below.

Table 1 displays a set of basic diagnostics for the household-level model. Of the 50 candidate variables, LASSO selects 23 predictors. Eight of these are geospatial predictors of some kind while the remaining 15 variables are state dummies. The first column of Table A3 presents the coefficients from a post-LASSO OLS regression using the sample data at the household level.

The predictive power of these variables is weak compared with other studies that use per capita consumption, as marginal R^2 is only about 0.13. This is partly because the welfare variable was spatially deflated in rural areas, meaning there is far less variation between urban and rural areas for the geospatial data to explain. In addition, the welfare measure is income per capita, making the results not comparable to other studies that have used per capita consumption as the welfare measure. The estimated area effect has skewness close to zero but kurtosis close to 3.5, meaning that the tails are slightly heavier than normal. The area effect is estimated very precisely and only accounts for 1.4 percent of the total variance. The complement of the estimated shrinkage factor is about 0.7, suggesting substantial efficiency gains in the estimation of the mixed effect from incorporating geospatial indicators.

Error! Reference source not found. shows the quantile-quantile plots for the estimated residuals of the household and area components of the error term. The estimated distribution of the municipality random effects is quite close to normal. However, there is substantial deviation from normality among the household residuals, both at the bottom and the top. This could potentially cause bias in the estimated poverty rates, since the simulation procedure will incorrectly assume the household residuals follow a normal-tailed distribution. To investigate that possibility further, we compare the predictions against the direct estimates and the CONEVAL benchmark below.

4.3 Criteria for evaluation

We consider six criteria for evaluating the estimators. The first is the simple unweighted mean of the poverty rate estimates across municipalities, prior to benchmarking. This gives an indication of overall bias in the pre-benchmarking estimators. The second and third criteria are measures of

uncertainty: the (unweighted) average estimated mean squared error (MSE) and median coefficient of variation (CV) across municipalities, which is essentially the median relative standard error. In the household model, the estimated MSE is determined by a parametric bootstrap procedure as described in Kreutzmann et al. (2018) and González-Manteiga et al. (2008).[17] Meanwhile, the coefficient of variation for each municipality is defined as the square root of the mean squared error, divided by the estimated poverty rate.

The fourth and the fifth criteria for evaluating the estimators are measures of accuracy. The first of these is the correlation with the official CONEVAL poverty estimates. The second is the root mean squared deviation, defined as:

$$RMSD = \sqrt{\frac{1}{M}\left(\sum_m \left(\widehat{\Theta}_m - \Theta_m^*\right)^2\right)},$$

where M is the number of municipalities, $\widehat{\Theta}_m$ is the estimated poverty rate for municipality m, and Θ_m^* is the official estimate calculated by CONEVAL using the intercensus.

Finally, we examine the coverage rate, defined as the share of target areas for which the official CONEVAL estimate lies within the estimated confidence interval. The upper and lower bounds of the confidence interval are determined by multiplying the square root of the estimated mean squared error by 1.96 and adding and subtracting it from the point estimate.

4.4 Alternative models

As mentioned above, the relative merit of household-level versus area-level models is a controversial topic in the literature. Corral et al (2021) concludes that area-level models should be preferred to household-level models due to the potential of omitted variable bias, even though the same omitted variable bias also appears in area-level models when assuming the same underlying data generating process. Furthermore, that paper shows no empirical evidence comparing the relative performance of the household and area-level models. To the best of our knowledge, the only source of empirical evidence on this issue is Masaki et al (2022), which found that a unit-context model produced moderately more accurate estimates of non-monetary poverty than an

[17] Two different parametric bootstrap methods to estimate uncertainty of EBP were proposed earlier by Arora et al. (1997) and Butar and Lahiri (2003).

area-level model in Sri Lanka and slightly less accurate estimates in Tanzania. Generating additional empirical evidence on the relative performance of different types of statistical models is important because in practice their performance depends on the true data generating process, which in practical applications is unknown. Therefore, we compare in the Mexican context the household model estimates with two alternative small area estimation models: a sub-area model and an area-level model.

The sub-area model uses the same type of empirical best prediction model as the household model. However, the unit of analysis is the AGEB and the dependent variable is the estimated poverty rates for each AGEB from the 2014 MCS-ENIGH survey data. These poverty rates are then modeled as a linear function of covariates, with random effects specified at the area level. The inclusion of the area-level random effect allows the model to "borrow strength" from sample data in sampled municipalities, which contain relevant information for non-sampled AGEBs. Covariates are again selected using the plug-in (rigorous) LASSO procedure. The selected AGEB-level covariates include the predicted shares of extreme and non-poverty from the CNN and the share of pixels classified as buildings at the AGEB level. At the municipal level, selected covariates include extreme poverty; share roads, forest, and grass; share rural; and ten state dummies. These variables explain 28 percent of the variation in AGEB-level poverty rates. The weights adjust both for sample weights and for heteroscedasticity in the dependent variable, due to the differing number of sample observations used to estimate AGEB-level poverty. As in the household model described above, the results are benchmarked to the state-level poverty estimates. The second column of Table A3 presents the coefficients from a post-LASSO OLS regression using the sample data, collapsed to the AGEB (sub-area) level.

In addition, for comparison, we also report results for a simple area-level model as initially described in Fay and Herriot (1979). Candidate predictor variables are aggregated to the municipality level, using 2010 population values from the census as weights. The prediction model is again selected using the "plug-in" rigorous LASSO, allowing for heteroscedasticity. The selected model contains seven predictors: The predicted municipal extreme and moderate poverty rates, the percentage of pixels classified as roads and background, and three state dummies. These variables explain 36 percent of the variation in estimated municipal poverty rates. The variance component of the model is estimated using the adjusted maximum likelihood approach for the linking model proposed by Li and Lahiri (2010), which guarantees positive variance estimates of

the modeled area effect for each small area. The model was estimated using the fayherriot package in Stata (Halbmeier et al., 2019). The third column of Table A3 presents the coefficients from a post-LASSO OLS regression using the sample data, collapsed to the municipal (area) level.

We provide more details on these alternative models in Appendixes C and D.

5. Main Results

5.1 Bias and uncertainty

Table 2 displays evaluation metrics for overall bias and uncertainty for the two direct survey estimates (the Horvitz-Thompson approximation and AGEB-clustered variance estimates), the household model, the sub-area model, the area-level model, and the 2010 official poverty estimates produced by CONEVAL. For reference, the bottom row shows the mean poverty estimates for the official 2015 poverty estimates used as the benchmark.

The first column shows the simple average, across municipalities, of the estimated poverty rate prior to benchmarking, while the second and third column of Table 2 report two measures of the uncertainty of the point estimates – the mean estimated mean squared error and the median coefficient of variation. Remarkably, for sampled municipalities, the mean estimated poverty rate from the household model, prior to benchmarking, is only 0.1 percentage points lower than the direct estimate (28.1% vs 28.2%). This reflects the success of the ordered quantile normalization transformation at virtually eliminating overall bias in the household model, despite the slight deviations in the household error term from normality shown in Figure 1. It also suggests that any bias due to omitting household characteristics from the model in this context does not systematically bias the municipality estimates, at least on average. The mean poverty rate for the sub-area model matches the direct estimates to one decimal place, while the area-level model is substantially more downward biased.

There are four other striking findings in the table. The first is the difference in estimated uncertainty between the Horvitz-Thompson and standard clustered estimates of variance. This reflects the high degree of correlation in per capita income across AGEBs within municipalities. The standard clustered estimates do not take these into account and therefore severely

underestimate the standard clustered estimates of uncertainty. Estimates of uncertainty based on clustered standard errors cannot be trusted in this setting.

A second striking funding is the large reduction of uncertainty for in-sample municipalities when using the household-level model. The mean MSE is substantially lower for the household model, at just 55 percent and 35 percent as large as the MSE for the sub-area and area-level models, respectively. The median coefficient of variation of the household model estimates is 19.8, which is well below that of the sub-area and area-level models (32.1 and 28.1, respectively). As noted below, the mean squared error and coefficient of variation in the household model is moderately underestimated. If the mean squared error is inflated by a constant to achieve an 86 percent coverage rate, however, the median CV of the household model rises to 25, remaining well below the sub-area and area-level models.

A third striking finding from Table 2 is the difference in precision between the estimates for sampled and non-sampled municipalities. For example, in the household-level model, the median CV rises from 19.7 in the sampled municipalities to 33.4 in the non-sampled municipalities, though this is still below the 38.5 median CV for (in-sample) direct survey estimates using the Horvitz-Thompson approximation. Since the sampling strategy for the MCS-ENIGH is partly based on population, one possible explanation is that the smallest municipalities are less represented in the survey but are also systematically different from larger municipalities. The fact that the true poverty rate for out-of-sample municipalities is substantially higher than for in-sample municipalities is consistent with this.

A final notable finding is that the gain in precision from utilizing a household model, rather than sub-area or area-level model, decreases substantially when considering out-of-sample municipalities. The greater precision of the household model for in-sample estimates, compared with the sub-area and area-level model, may reflect the ability of the household model to better distinguish between near-poor and wealthy households, because it is modeling welfare levels. This advantage does not apply when predicting out-of-sample, because the estimates are purely synthetic predictions that are not conditioned on the sample data. This may also explain why the

household model's advantage, in terms of precision, is smaller for out-of-sample municipalities than for in-sample municipalities.

5.2 Accuracy and coverage rates

Error! Reference source not found. presents predicted-true plots for of the poverty predictions for in-sample predictions (left) and out-of-sample predictions (right) for the household model, the sub-area model, and the area-level model. For each model, it is visually clear that the in-sample predictions are more accurate than the out-of-sample predictions. Across models, the household-level model is discernably more accurate in-sample, as the estimates are generally closer to the dotted line than in the sub-area model. Out-of-sample, the differences are more difficult to distinguish.

Table 3 displays summary measures of the accuracy and coverage rates of the estimates from the different models. The second-to-last row shows that the 2010 census-based estimates are the most accurate, especially out of sample, with estimated correlations above 0.9.[18] The household model improves on the accuracy of the direct estimates for in-sample municipalities, with the correlation between the estimates and the true estimates rising from 0.80 to 0.86. Consistent with that, the root mean squared deviation falls 25 percent, from 0.126 to 0.094. The sub-area model fares slightly worse than the estimates from the household-level model with respect to both accuracy measures but are an improvement over the direct estimates. The area-level model estimates are the least accurate and do not improve on the direct estimates in terms of correlation but do improve on the direct estimates in terms of RMSD. This may reflect imprecision in the estimated variance of municipal poverty derived from the sample, which are an input into the area-level model.

Out of sample, the predictive performance of all estimators suffers from the lack of sample data on which to condition the estimated random area effects.[19] The out-of-sample correlation is only 0.70 for the household-level and sub-area level model, and 0.66 for the area-level model. Similarly, root mean squared deviation of the household-level model is nearly twice as high out of sample,

[18] However, derivations of the 2010 poverty estimates are used as predictors in the official 2015 poverty estimates. As such, part of the higher correlation might be mechanical.

[19] The lower predictive performance out of sample is not due to overfitting the model, since the plug-in lasso selects more parsimonious models than cross-validation.

at 0.181, compared with the in-sample value of 0.094. In terms of root mean squared deviation, the household and sub-area models are more accurate than the area-level estimates, which have a root mean squared deviation of 0.198.

Finally, the bottom row shows the correlation of municipality-level poverty with the relative wealth index (RWI) created by Chi et al. (2022) and part of Meta's Data for Good project.[20] This wealth index is available for 56 different low- and middle-income countries across the globe. The correlation between the RWI and poverty rates is worse than any of the other estimates. However, this is not an indication that the RWI is incorrect; instead, it is an important reminder that wealth does not always correlate well with poverty. Interestingly, however, the correlation between the RWI and poverty is more than 12 percent lower in out-of-sample municipalities, even though the RWI was created with completely different data. In other words, the out-of-sample municipalities may simply be systematically different from in-sample municipalities.

Turning to coverage rates, the household model moderately underestimates mean squared error, which is reflected in estimated coverage of 76.9% in sample and 82.5% out of sample. This underestimate of uncertainty stems from the household model failing to account for uncertainty in the estimated variance of the sample. The sub-area model performs substantially better on coverage, due to its higher MSE estimates. Meanwhile, the coverage rate of the area-level model falls between the household and area-level model for in-sample municipalities and is the lowest for out-of-sample municipalities.

While the coverage rate is lowest for the household model, for in-sample municipalities the rate is only 9 percentage points lower than the Horvitz-Thompson direct survey estimates, and nearly twice as large as the very low coverage rate of the direct estimates when using standard clustered variance estimates. Overall, the results indicate that the household-level model is the most accurate and the most precise of the three estimators for in-sample municipalities and is equally accurate as the other models for out-of-sample municipalities. While coverage rates are moderately underestimated, they remain respectable.

Finally, Figure 3 presents boxplots of the relative bias of the estimators, a statistic not reported in the tables. The relative bias is defined as the difference between the predicted poverty rate and the

[20] https://dataforgood.facebook.com/

official CONEVAL benchmark, divided by the latter. Results are reported both for in-sample (left) and out-of-sample (right) municipalities. For out-of-sample municipalities, the relative bias estimates were top-coded at 25.

Both in and out of sample, the official 2010 CONEVAL estimates exhibit markedly less relative bias than the other estimators. For in-sample estimates, the boxplot confirms that the 2010 CONEVAL estimates are the least biased, followed by the household model estimates, the sub-area model estimates, and the area-level estimates. For out-of-sample municipalities, the 2010 estimates do far better than any of the geospatial estimates, for which differences are hard to discern. The out-of-sample estimates contain several large positive outliers among non-sampled municipalities, including cases where the predicted poverty rate is more than ten times above the official CONEVAL estimates. This highlights the value of incorporating survey data in an empirical best framework, rather than simply relying on synthetic predictions, and the value of designing surveys to guarantee inclusion of all target areas in the sample.

6. Robustness checks

6.1 Benchmarking

The results presented in Table 2 and Table 3 used estimates that were benchmarked to match the estimated poverty rates for each state. Benchmarking is appealing to ensure consistency between small area estimates and published survey results, and to remove any systematic bias at the state level. It is appropriate in cases when the small areas of interest define a hierarchical structure (Pfeffermann et al., 2014). Nonetheless, it is useful to test that the main results hold when not applying benchmarking. This is in part because existing software does not allow for benchmarking within the parametric bootstrap procedure when estimating mean squared error in unit-level models.

Table 4 shows the results for the household-level model with and without benchmarking. The mean poverty rate is defined as the unweighted mean estimated poverty rate across municipalities

prior to benchmarking, and therefore is unaffected by benchmarking. Benchmarking has minor effects for in-sample municipalities because of the presence of many state-level dummy variables in the model. Conversely, state dummy variables were not included for many out-of-sample municipalities, and as seen in the mean poverty column in Table 2, the non-benchmarked predictions significantly underestimate poverty in these cases. As in Masaki et al. (2022), the mean squared error was multiplied by the same scaling factor that was used for benchmarking in each municipality, which increases mean squared error and improves coverage rates, particularly in out-of-sample municipalities. For in-sample municipalities, benchmarking improves the correlation with the CONEVAL estimates by about a percentage point, from 0.853 to 0.862, and slightly reduces root mean squared deviation. For out-of-sample prediction municipalities, meanwhile, benchmarking has a more substantial beneficial effect on accuracy, as the correlation with the CONEVAL estimates improves by about 3 percentage points from 0.669 to 0.701, and root mean squared deviation falls from 0.196 to 0.181. Overall, benchmarking improves the accuracy of the estimates, particularly for out-of-sample municipalities, and the benchmarked estimates are therefore preferred.

6.2 Weights

Table 5 considers the impact of using sample and population weights when estimating the household-level model. The population weights in this case are estimated average household size at the state level, obtained by the MCS-ENIGH survey. As noted in section 3, this is assigned to each household in the synthetic census. Sample population weights, defined as the provided household weight times household size, are provided with the MCS-ENIGH data, and are used in the preferred specification when estimating the mixed effect model. Table 8 shows that failing to use sample weights decreases the average poverty rates prior to benchmarking from about 28.1%, which nearly matches the direct estimates, to about 26.4%. This suggests that failing to incorporate weights can moderately increase bias. Interestingly, the use of sample weights makes the estimates slightly less accurate for in-sample municipalities, but slightly improves accuracy for out-of-sample municipalities. This suggests that the method that was used to apply the weights to the sample may not be optimal for generating in-sample predictions in empirical best models. We nonetheless prefer the weighted estimates because they are more accurate for out-of-sample

predictions and because they generate larger and more accurate estimates of uncertainty. Because the cost to accuracy for in-sample municipalities is minor, the larger estimates of uncertainty increase the coverage rate for in-sample municipalities. Failing to weight by state household size when aggregating the simulation results to municipalities has virtually no effect on the estimates because the population weights only vary at the state level and the benchmarking procedure aligns the estimates with the survey-based estimates at that level.

6.3 Model heterogeneity

So far, the results shown have come from one estimated national model. However, it is possible that estimating separate models would improve the performance of the model by better modeling the heterogeneity across states or across urban/rural areas. Table 6 shows how the results change due to different types of variants of the estimation and simulation procedure. The baseline model considers estimates from one national model. Mexico has separate poverty lines for urban and rural areas, however, which must be considered when estimating headcount poverty rates during the simulation stage. Because the software package used allows for only one poverty line, the baseline model inflates the per capita income of rural areas in the sample by the ratio of the urban to rural poverty line. After making this adjustment, the urban poverty line threshold is applied to all households in the synthetic census, in both urban and rural areas. The second model considers a similar approach, but instead estimates six models, each corresponding to different groups of states.[21] This may better capture the heterogeneity across these different state groupings. The third model takes a different approach to estimating separate urban and rural poverty lines. This approach simulates urban and rural poverty rates separately, based on a single national model, and then takes the population-weighted average of the urban and rural estimates for each municipality.[22] When taking this population-weighted average, we assume zero covariance between the urban and rural poverty estimates of common municipalities, which further

[21] These are the same six groups used in official government statistics.

[22] It is less straightforward to estimate the mean squared error of the population weighted urban and rural average. Partly this is because the random component of the municipal mixed effect will be drawn independently for rural and urban areas, underestimating the mean squared error. We therefore estimate the covariance of urban and rural poverty rates across municipalities, assuming the poverty estimates are unbiased and that the MSE estimates can be used to estimate variance. We then estimate the covariance between urban and rural municipalities and calculate the variance of the population-weighted average.

underestimates mean squared error. The fourth model is identical to the third, except that it estimates six different models, one for each group of states. The fifth and final model also simulates urban and rural poverty rates separately but differs from the third by estimating two prediction models – one for urban households and one for rural households - instead of a single national model. This allows the estimated relationship between per capita income and the geospatial predictors in the prediction models to vary for urban and rural areas.

In terms of accuracy, as measured by correlation with the official CONEVAL estimates, the results are robust to all alternative specifications for in-sample municipalities. The grouped states model is slightly more accurate than the national model for in-sample municipalities and slightly less accurate for out-of-sample municipalities. The variant that estimates nominal welfare with six group models is the most accurate both in and out of sample but the difference in accuracy with the baseline estimator is modest. In terms of bias (mean poverty), all estimates are virtually unbiased within-sample but all are substantially downwardly biased when predicting out of sample. The average poverty rate of the official estimates for the out-of-sample municipalities reported in Table 2 is .426, while the analogous figure for mean poverty in Table 6 ranges from .355 to .379.

In terms of efficiency, the estimated mean squared error and coefficient of variation is far lower for the nominal welfare model than the deflated welfare model, likely because the former does not account for the positive correlation between the rural and urban areas of each municipality and therefore underestimates uncertainty. The coverage rates for the nominal models are therefore below 70 percent for in-sample municipalities. However, estimating separate urban/rural models makes the estimates substantially less precise, and brings the coverage rate in sample back up to the level of the baseline estimates.

Overall, the results of the household model are generally robust to alternative specifications. In particular, the range of in-sample correlation across the five methods ranges from 0.862 to 0.873. The range of out-of-sample correlations is slightly larger, since the sample does not contribute information, but the range is still only about 1.3 percentage points, from 0.697 to 0.710.

Simulating separate urban and rural models is associated with far greater uncertainty, as reflected in the estimated CV and MSE, but also substantially higher coverage rates. This is because when urban and rural areas are simulated separately, either based on national or separate urban and rural

models, the simulated random component is no longer common to urban and rural areas in the same municipality. This underestimates uncertainty in the model by incorrectly assuming that the income of urban and rural households in the same municipality are independent. To address this, we estimate the covariance across municipalities of the estimated urban and rural poverty rates and account for it when estimating the variance of the municipal poverty estimates. This in turn overestimates the uncertainty associated with the estimates, which counteracts the underestimated uncertainty due to assuming the variance is known rather than estimated. However, the MSE and coverage rate is still slightly below the baseline estimates, which use spatially deflated welfare and a national model, because the latter allow for positive covariance between urban and rural areas of a municipality.

6.4 Using 2016 sample data

The analysis up to this point has all used a single household survey, the 2014 MCS-ENIGH, to estimate municipal-level poverty. While this sample was drawn to generate official measures of poverty, it is also useful to check that the results are robust to the use of an alternative sample. We therefore repeat the analysis using the 2016 MCS-ENIGH instead of the 2014 round, which contains a different set of selected AGEBs. This also eliminates any possible mechanical correlation between the small area estimates and the benchmark CONEVAL estimates, which occurs because both use the 2014 MCS-ENIGH survey to estimate the empirical best prediction model. Table 7 reports the results when using the 2016 sample for the baseline specification. The main difference is that the estimates are moderately less accurate, due to the use of survey data that differs from that used to generate the benchmark. The correlation with the benchmark is now only 0.81, as opposed to 0.86 when using the 2014 survey. The same pattern of results holds, however, when ranking across methods. For in-sample areas, the household model gives moderately more accurate estimates than the sub-area model (correlation of 0.78) while the area-level model and direct estimates give less accurate estimates (correlation of 0.75 each). Furthermore, the household-model is far more precisely estimated for in-sample municipalities (median CV of 21 vs. 31 for the area-level model). Uncertainty is significantly underestimated in

the household model as the coverage rate falls to 59 percent, as opposed to 82 percent for the sub-area model and 67 percent for the area-level model.

Figure 4 shows the boxplot for relative bias. All estimates are slightly biased downward. This is because they are estimated using 2016 data, which is also used for the benchmarking procedure, and then compared against 2015 estimates generated using the 2014 sample data. Therefore, it is not surprising that estimates based on 2016 would be systematically below the benchmark estimates, reflecting the overall trend of poverty decline. Nonetheless, the results confirm that the household-level model generates less biased results than the sub-area and area-level models in sample, which each give less biased results than the direct estimates. Meanwhile, the household-level model does at least as well as – if not slightly better than – the sub-area and area-level models out of sample. When using the 2016 sample data to estimate the model, the estimates produced by the household model are less biased than those produced by sub-area and area level models and represents a considerable improvement in accuracy and efficiency over the direct estimates. However, the uncertainty appears to be substantially underestimated, which is not as true for the sub-area estimates. Thus, while accuracy is best for the household-level model, the sub-area model outperforms in terms of correctly estimating uncertainty.

6.5 Including predictors at various levels.

The baseline specification includes AGEB-level variables, municipal-level variables, and state-level dummies as candidate predictors. To better understand the sensitivity of the household model to including variables at different levels, we estimate five additional models, representing different combinations of including or excluding AGEB, municipal, and state dummies. Two models use only AGEB or only municipal candidate predictors without state dummies, two more only AGEB or only municipal candidate predictors with state dummies, and one uses AGEB and municipal predictors without state dummies. In all cases, we use plug-in LASSO to select a model from the set of candidate predictor variables.

Table 8 reports the results from this exercise. The in-sample results are very robust to different model specifications, as correlations with the benchmark vary only from about half a point. On the other hand, the accuracy of out-of-sample predictions are much more sensitive to the set of candidate predictors. The baseline specification, which includes state dummies, AGEB and

municipal variables, performs the best by between 1.5 and 23 points. The second-most accurate model is the one that include state dummies and municipal variables, suggesting that the AGEB-level variables, conditional on the municipal level variables, is adding only a modest amount of accuracy in this case. In this context, including dummies at the level at which the sample is representative, as well as area and sub-area level variables notably improves out-of-sample predictive performance.

6.6 Design-based simulation

Finally, we also present results from a design-based simulation in Appendix E. Since we use a single sample from 2014, one possible concern is that the results presented here may not hold in other samples. While the results when using the 2016 sample suggest that the main results are robust, we nonetheless present additional results using repeated sampling from the intercensus. For this exercise, because AGEB level identifiers are not available in the intercensus, we use auxiliary data aggregated to the municipal level. The overall results presented in Appendix E are consistent with the findings and conclusions from the 2014 and 2016 surveys. In particular, they indicate that the household level model provides equally accurate estimates in-sample, and significantly more accurate estimates out of sample. In particular, for out-of-sample municipalities, the average correlation with truth in the simulations was 0.80 out-of-sample for the unit-context model, as opposed to 0.75 for the area-level model.

7. Conclusion

This paper examined the ability of methods combining survey and satellite data to generate more accurate estimates of monetary poverty in Mexican municipalities. Mexico is a valuable case study to evaluate small area estimation methodology because official estimates, generated using a best-practice unit-level model with household-level auxiliary data from the 2015 intercensus, provide a solid benchmark for comparison. Unfortunately, these types of comprehensive intercensal surveys are rarely available in developing countries, which often face challenges fielding censuses every ten years. In addition, CONEVAL produces and makes publicly available estimates generated using the 2010 census, which provides a useful point of comparison.

Incorporating even a small number of satellite predictors, with moderate predictive performance, markedly improves the accuracy of the sample estimates. In particular, the correlation with the benchmark official estimates improves from 0.8 to 0.86 when using a household-level "unit-context" model instead of estimates derived from the sample survey. Precision improves dramatically as well, as the coefficient of variation drops by roughly a factor of two, roughly equivalent to quadrupling the size of the survey. The gain in precision is moderately overestimated, but after adjusting for that to maintain the same coverage rate as the direct estimates, incorporating the small area estimates effectively increases the sample size by a factor of about 2.4.

When using satellite data, there is a large difference in accuracy between estimates for in-sample and out-of-sample municipalities. While the preferred household model generated a correlation of 0.86 with the intercensus benchmark in-sample, this fell to 0.7 for out-of-sample. This difference in performance comes for two reasons. First, empirical best predictor models improve accuracy by efficiently combining the survey data with model predictions for each municipality. Second, out-of-sample municipalities are systematically smaller and poorer than sample municipalities, which reduces the accuracy of out-of-sample predictions in this context. These results suggest additional caution when relying on purely synthetic predictions generated from satellite data.[23]

An important benefit of this type of evaluation exercise against a credible benchmark is the opportunity to compare different statistical methods for incorporating geospatial data into small area estimation. In the main set of results, using the 2014 survey data, a household-level model gives more accurate predictions than a sub-area model for sampled areas (correlation of 0.86 vs 0.83) and equally accurate predictions for non-sampled municipalities (correlation of 0.7). When using the 2016 survey, the predictions from the household-level model are more accurate than the sub-area model both in and out of sample. The area-level model is generally less accurate than the sub-area model, with in-sample correlations of 0.8 when using the 2014 data and 0.75 when using the 2016 data. The exception is out-of-sample predictions using the 2016 data, where area-level estimates are more accurate than sub-area estimates but remain less accurate than the household model. Both the household-level and area-level models underestimate uncertainty, relative to the sub-area model, in the 2016 data. In artificial simulations, when compared with predictions from

[23] It is possible that more sophisticated models or methods that allow for non-linearities and interactions may better predict out of sample, especially if the correlations of predictors vary by things like municipality size. This is a topic for further research.

an area-level model, predictions from a household "unit-context" model specified with area-level predictors were equally accurate in sample, but markedly more accurate out of sample. In sum, there is a clear ranking in accuracy across the different methods for incorporating the geospatial auxiliary data: The household model was usually more accurate, and never noticeably less accurate, than the two other methods. This was generally followed by the sub-area model estimates, and finally the area-level model estimates. There is no evidence in any of the settings considered that area-level models give estimates that are more accurate and efficient, or less biased, than household-level unit-context models.

All geospatial small area estimates, however, remain substantially less accurate than those generated using survey and household census data from 2010. This suggests that municipal poverty patterns changed slowly in Mexico during this time, although the methodology used to generate the 2015 benchmark estimates may also not fully capture transient changes in welfare.[24] The correlation between the 2010 estimates and the 2015 estimates is very high: 0.91 in sample and 0.90 out of sample. Thus, if forced to use municipal poverty estimates to inform a pro-poor budget allocation in 2015, the most accurate source would be the 2010 census-based estimates. Analysts should therefore exercise healthy caution before jettisoning moderately older census-based estimates in favor of newer small area estimates based on geospatial data.

The high accuracy of older census-based estimates, however, may not generalize to other contexts, and more research is needed to better understand under what circumstances older census-based estimates give more accurate predictions than recent geospatial poverty estimates. This will depend on the extent to which regional poverty patterns have changed in the country since the last census, as well as the size of the sample survey and the predictive power of the satellite-based auxiliary data, among other factors. It is also likely that the predictive performance of geospatial and other big data will continue to improve as more and better indicators become available.

In addition, surveys can also be fielded in ways that better facilitate integration with geospatial data. For example, increasing the number of households interviewed in each enumeration area would likely further improve the predictive performance of the model and the accuracy of the estimates. Likewise, given the better performance of all small area estimates when predicting in-

[24] This is partly because the model used to generate the official 2015 estimates of municipal poverty rates utilized data from the 2010 census as predictor variables.

sample versus out-of-sample municipalities, expanding sampling to all target areas may yield marked improvements in predictive power. The best survey design for small area estimation with big data thus remains a fruitful area for future research. Extending small area estimation models also remains an important area of research. For example, in the future, sub-area models could better incorporate sampling variability, while unit-context models might benefit from the incorporation of sub-area random effects.

Overall, these results demonstrate, in the Mexican context, that augmenting household survey data with geospatial data substantially improves the accuracy and precision of survey-based estimates of monetary poverty. In countries where census-based estimates are heavily outdated or non-existent, we believe the results of this evaluation exercise are sufficiently encouraging to justify combining survey and satellite data to produce small area estimates of monetary poverty.

References

Arora, V., Lahiri, P. and Mukherjee, K. (1997), Empirical Bayes estimation of finite population means from complex surveys, Journal of the American Statistical Association, 92, 1555-1562.

Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A., & Swartz, T. (2017). Poverty mapping using convolutional neural networks trained on high and medium resolution satellite images, with an application in Mexico. *arXiv preprint arXiv:1711.06323*.

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28-36.

Bell, William R. "Examining sensitivity of small area inferences to uncertainty about sampling error variances." *Proceedings of the American Statistical Association, Survey Research Methods Section*. Vol. 327. 2008.

Belloni, A., and V. Chernozhukov. "High dimensional sparse econometric models: An introduction." Inverse Problems and High-Dimensional Estimation. Springer, Berlin, Heidelberg, 2011. 121-156.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073-1076.

Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. *Science*, *371*(6535).

Butar, F. and Lahiri (2002), On the measures of uncertainty of empirical Bayes small-area estimators, Journal of Statistical Planning and Inference, 112, 63-76.

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, *119*(3), e2113658119.

Corral, P. Himelein, K. McGee and I. Molina (2021). A map of the poor or a poor map? *Mathematics, 9(21),* 2780.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, *71*(1), 355-364.

Engstrom, R., Hersh, J., & Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, *36*(2), 382-412.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269-277.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, *78*(5), 443-462.

Ghosh, Malay. "Small area estimation: its evolution in five decades." *Statistics in Transition. New Series* 21.4 (2020): 1-22.

Halbmeier, C., Kreutzmann, A. K., Schmid, T., & Schröder, C. (2019). The fayherriot command for estimating small-area indicators. *The Stata Journal*, *19*(3), 626-644.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, *47*(260), 663-685.

Jiang, J. and Lahiri, P. (2001), Empirical best prediction for small area inference with binary data, Annals of Institute of Mathematical Statistics, 53(2): 217-243.

Jiang, J., Lahiri, P. and Wan, S. (2002), Jackknifing the mean squared error of empirical best predictor, Annals of Statistics, 30, 1782-1810.

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, *15*(1), 1-96.

Jiang, J., & Rao, J. S. (2020). Robust small area estimation: An overview. *Annual review of statistics and its application*, *7*, 337-360.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

Kreutzmann, A. K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, *91*.

Lahiri, P., 1987, Robust Empirical Bayes Estimation in Finite Population Sampling.

Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of multivariate analysis*, *101*(4), 882-892.

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2022). Small Area Estimation of Non-Monetary Poverty with Geospatial Data, *Statistical Journal of the IAOS*, pre-press.

Merfeld, J. D., Newhouse, D. L., Weber, M., & Lahiri, P. (2022). Combining Survey and Geospatial Data Can Significantly Improve Gender-Disaggregated Estimates of Labor Market Outcomes. *World Bank Policy Research Working Paper 10077*.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369-385.

Rao, J. N., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

McBride, L., Barrett, C. B., Browne, C., Hu, L., Liu, Y., Matteson, D. S., ... & Wen, J. (2021). *Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning* (No. 2388-2021-383).

National Statistical Office (NSO) of Malawi. (2021). Malawi Poverty Report 2020.

Peterson, R. A., & Cavanaugh, J. E. (2020). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics, 47:13-15, 2312-2327*.

Pfeffermann, D., Sikov, A., & Tiller, R. (2014). Single-and two-stage cross-sectional and time series benchmarking procedures for small area estimation. *Test*, *23*(4), 631-666.

Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. Proceedings of the National Academy of Sciences, 114(46), E9783-E9792.

Rao, J. N. K. (2003). "Small Area Estimation". Wiley, London.

Rao, J. N. K. (2014). Small-area estimation. *Wiley StatsRef: Statistics Reference Online*, 1-8.

Ravallion, M. (2015). *The economics of poverty: History, measurement, and policy*. Oxford University Press.

StataCorp. 2021. Stata: Release 17. Statistical Software. College Station, TX: StataCorp LLC.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., ... & Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, *14*(127), 20160690.

Steele, J. E., Pezzulo, C., Albert, M., Brooks, C. J., Erbach-Schoenberg, E., O'Connor, S. B., ... & Tatem, A. J. (2021). Mobility and phone call behavior explain patterns in poverty at high-resolution across multiple settings. *Humanities and Social Sciences Communications*, 8(1), 1-12.

Tarozzi, A., & Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *The Review of Economics and Statistics*, *91*(4), 773-792.

Torabi, M., & Rao, J. N. K. (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, *127*, 36-55.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference, John Wiley & Sons, Inc.

Van der Weide, R. (2014). *GLS estimation and empirical bayes prediction for linear mixed models with heteroskedasticity and sampling weights: a background study for the povmap project*. The World Bank.

World Bank. *World Development Report 2021: Data for Better Lives*. The World Bank, 2021.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., ... & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications*, *11*(1), 1-11.

Tables

Table 1: Household model diagnostics

	Household model
Number of included predictor variables	23
Sample	
Number of households	57,661
Number of AGEBs	6,794
Number of municipalities	892
Synthetic census	
Number of households	27,973,210
Number of municipalities	2,433
Household model diagnostics	
Marginal R^2	0.134
Conditional R^2	0.225
Skewness of area effect	0.041
Kurtosis of area effect	3.447
Wilks-Shapiro P-value	0.043
Skewness of household error	-0.212
Kurtosis of household error	3.749
Wilks-Shapiro P-value	N/A
Estimated variance of area effect	0.009
Estimated variance of household error term	0.659
Ratio of estimated variance of area effect to total variance	0.014
Average complement of shrinkage factor	0.695

Notes: Number of included predictor variables excludes constant. Sample figures reflect the number of households, AGEBs, and municipalities contained in the MCS-ENIGH sample. Synthetic census figures reflect the number of households and municipalities used in the population used to generate municipal poverty estimates, based on the 2010 census. Marginal R^2 reflects the variation explained by model predictors, while conditional R^2 includes the variation explained by conditioning the mixed effect on sample welfare. Skewness and Kurtosis of the area effect and household error term describe the distribution of these components of the error term. The average complement of shrinkage is equal to one minus the average of the estimated shrinkage factors. It is a measure of the average percentage that conditioning the area effect on the sample data shrinks the variance of the mixed effect.

Figure 1 - Quantile-quantile plots of household model residuals

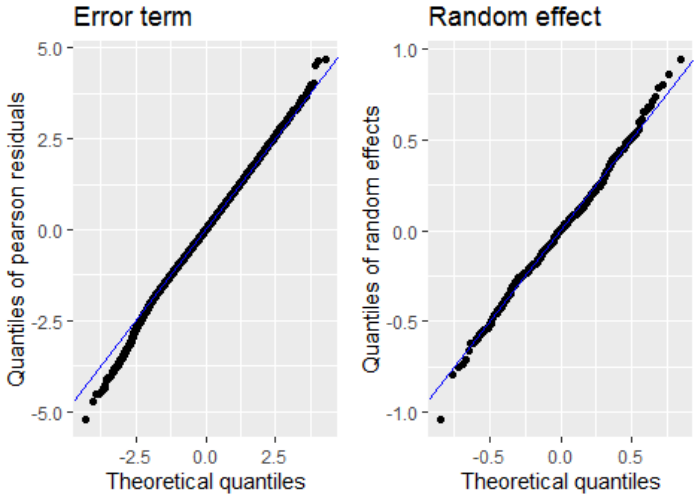


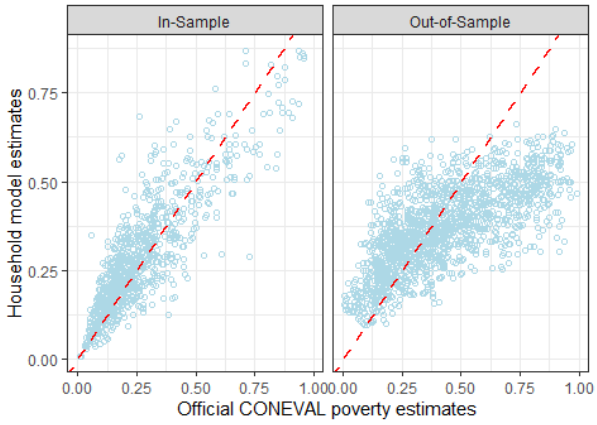
Table 2 - Poverty predictions for municipalities: Mean poverty and precision

	Sampled municipalities			Non-sampled municipalities		
	Mean poverty	Mean MSE	Median CV	Mean poverty	Mean MSE	Median CV
Direct survey estimates						
Horvitz-Thompson approximation	0.282	155.8	38.5	N/A	N/A	N/A
AGEB-Clustered variance estimates		25.2	11.3	N/A	N/A	N/A
CNN-based model estimates						
Household-level EBP model	0.281	35.8	19.8	0.355	150.3	33.9
Bias-corrected Household EBP model						
Sub-area model	0.282	101.5	35.6	0.365	306.2	47.3
Area-level model	0.227	64.7	28.1	0.271	158.1	37.5
Official 2010 estimates	0.266	N/A	N/A	0.459	N/A	N/A
Official 2015 estimates	0.298	N/A	N/A	0.426	N/A	N/A

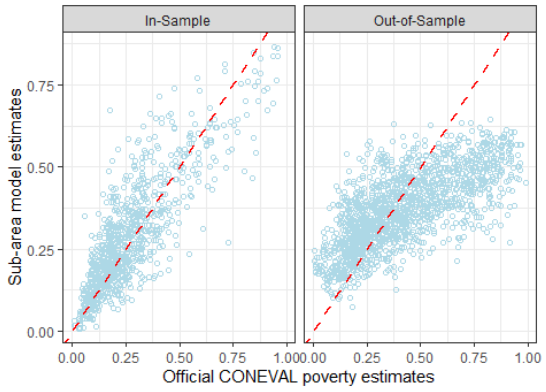
Notes: Table shows key results for direct estimates, household model estimates, sub-area model, and area-level model estimates. In-sample refers to the 893 municipalities contained the 2014 MCS-ENIGH survey, while out of sample refers to the 1,418 municipalities not covered by the survey. Direct estimates are reported separately when variance is estimated using the Horvitz-Thompson approximation and clustering on AGEBs. The mean poverty represents the simple average poverty rate across municipalities prior to benchmarking, which is a measure of the extent of bias in the model-based estimates relative to the direct and official estimates. The mean MSE is the mean estimated Mean Squared Error (times 10,000) across municipalities, Model-based estimates of MSE reflect benchmarking estimated poverty rates to direct estimates at the state level as described in section 4. The median CV is the median across municipalities of the coefficient of variation, defined for each municipality as the square root of the estimated mean squared error divided by the estimated poverty rate. All results are unweighted means or medians across municipalities.

Figure 2 - Predicted-true plots

Panel A: Household model



Panel B: Sub-area model



Panel C: Area -level model

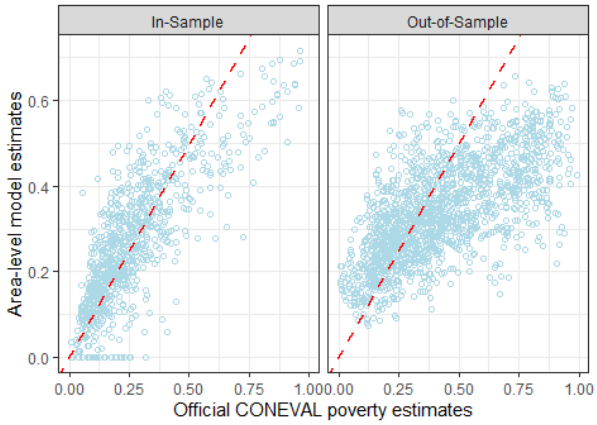
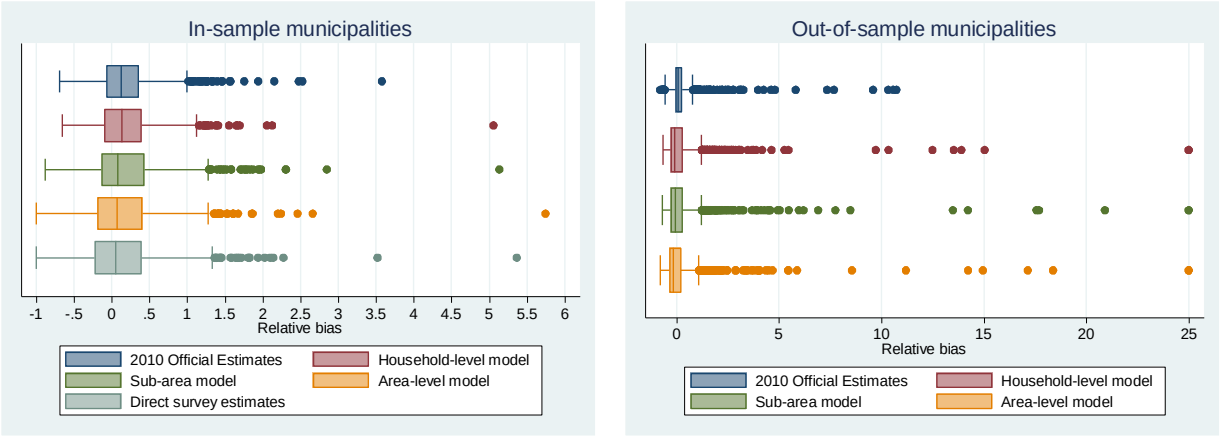


Table 3 - Accuracy and coverage against official 2015 poverty estimates

	Sampled municipalities			Non-Sampled municipalities		
	Corr	RMSD	CR	Corr	RMSD	CR
Direct survey estimates						
H-T	0.800	0.126	0.856	N/A	N/A	N/A
EA-Clustering			0.393	N/A	N/A	N/A
Household-level model	0.862	0.094	0.769	0.701	0.181	0.825
Sub-Area model	0.834	0.103	0.910	0.696	0.183	0.941
Area-level model	0.796	0.110	0.824	0.662	0.198	0.801
Official 2010 estimates	0.912	0.083	N/A	0.904	0.109	N/A
Meta relative wealth index	-0.592		N/A	-0.516		N/A

Notes: Table shows evaluation results for direct estimates, household model estimates, sub-area model, and area-level model estimates against official municipality poverty rates derived from a standard household level small area estimation exercise undertaken by CONEVAL. In-sample refers to the 893 municipalities contained the MCS-ENIGH survey, while out of sample refers to the 1,418 municipalities not covered by the survey. Direct survey estimates are reported separately when variance is estimated using the Horvitz-Thompson approximation and clustering on AGEB. Model-based estimates are benchmarked at the state level as described in section 4. Corr represents the simple correlation between the estimated poverty rates and the official CONEVAL benchmark. RMSD represents the root mean squared deviation of the predictions relative to the official estimates, and is an alternative measure of prediction accuracy. CR represents the coverage rate, which is the share of municipalities for which the official estimate lies within the 95 percent confidence interval of the model-based or survey-based estimates. It is a measure of the accuracy of the estimated mean squared error.

Figure 3 - Box plots of relative bias, 2014 MCS-ENIGH estimates



Note: The figure shows the distribution of relative bias of municipal poverty estimates by estimation method. Relative bias is computed as the ratio of the difference between the estimate and the CONEVAL benchmark to the CONEVAL benchmark. The 2010 official estimates are produced by CONEVAL based on the 2010 census. The household-level, sub-area, and area-level model estimates are produced by combining the 2014 MCS-ENIGH with satellite indicators. The direct estimates are generated using only the 2014 MCS-Enigh survey.

Table 4 - Robustness of household-level model estimates to benchmarking

Household model estimates	Mean poverty	MSE	Median CV	Corr	RMSD	CR
Sampled municipalities						
State-level benchmark	0.281	35.8	19.8	0.862	0.094	0.769
No benchmark	0.281	34.3	19.8	0.853	0.095	0.753
Non-sampled municipalities						
State-level benchmark	0.355	150.3	33.9	0.701	0.181	0.825
No benchmark	0.355	138.5	33.9	0.669	0.196	0.779

Notes: Table shows evaluation statistics for benchmarked and non-benchmarked municipal poverty estimates. In-sample refers to the 893 municipalities contained the MCS-Enigh survey, while out of sample refers to the 1,418 municipalities not covered by the survey. Columns are defined as in Tables 6 and 7. Benchmarking entails multiplying each municipal estimate by the ratio of the weighted average of the state direct estimate to the population-weighted average of the municipal poverty estimates for each state. This ensures that the state average poverty rate for the small area estimates is consistent with the direct survey estimates. The weights in the numerator are the sample weights provided in the sample and the weights in the denominator are taken from the 2010 census. MSE estimates for each municipality are also multiplied by the same ratio that is applied to the poverty estimates.

Table 5 - Results with and without sample and population weights

Household model	Mean Poverty	Mean MSE	Median CV	Correlation	RMSD	CR
Sampled municipalities						
Baseline estimates	0.281	35.8	19.8	0.862	0.094	0.769
Aggregation weights but no sample weights	0.264	32.4	18.2	0.864	0.093	0.748
Neither sample nor aggregation weights	0.264	32.4	18.2	0.865	0.093	0.748
Non-sampled municipalities						
Baseline estimates	0.355	150.3	33.9	0.701	0.181	0.825
Aggregation weights but no sample weights	0.336	145.4	32.9	0.700	0.180	0.824
Neither sample nor aggregation weights	0.336	145.6	32.9	0.700	0.180	0.825

Notes: Table shows evaluation statistics for estimates with different weights applied. The baseline estimates apply both sample weights when estimating the model and estimated household size in the population when aggregating the simulation results across households. Including population weights makes the estimated poverty rates representative of individuals rather than households. Sample weights are taken from the survey and the estimated household size for each AGEB is obtained from the 2010 census. Columns are defined as in Tables 6 and 7.

Table 6 - Robustness of household model results to different specifications

Household model	Mean Poverty	Mean MSE	Median CV	Correlation	RMSD	CR
Sampled municipalities						
Deflated welfare with national model	0.281	35.8	19.8	0.862	0.094	0.769
Deflated welfare with grouped states models	0.280	35.2	19.6	0.870	0.092	0.777
Nominal welfare with national model	0.283	23.4	15.5	0.865	0.093	0.679
Nominal welfare with grouped states models	0.282	23.2	15.3	0.873	0.091	0.671
Nominal welfare with urban/rural models	0.280	33.6	19.1	0.866	0.093	0.767
Non-sampled municipalities						
Deflated welfare with national model	0.355	150.3	33.9	0.701	0.181	0.825
Deflated welfare with grouped states models	0.365	170.6	33.4	0.697	0.173	0.869
Nominal welfare with national model	0.366	143.7	26.3	0.705	0.178	0.751
Nominal welfare with grouped states models	0.379	160.1	25.8	0.710	0.167	0.803
Nominal welfare with urban/rural models	0.356	97.0	26.9	0.707	0.181	0.726

Deflated welfare refers to welfare after multiplying the welfare of rural households by the ratio of the urban to rural poverty lines and using the urban poverty line as the poverty threshold. National model refers to a single model for welfare estimated on the full MCS-ENIGH sample. Grouped states refers to estimating six separate models, one for each group of states as defined in Table B1, for which separate models are estimated. Nominal welfare refers to the use of nominal welfare, with poverty rates simulated separately for urban and rural areas using urban and rural specific poverty lines. In the three nominal welfare models, the final municipal estimates reflect a weighted average of the urban and rural poverty estimates for each model, with weights equal to the population in the 2010 census. The estimated MSE in these cases accounts for the estimated correlation between urban and rural poverty rates in the same municipality, as described in section 6. Urban/rural models refers to separate models for urban and rural areas. Columns are defined as in Tables 5 and 6.

Table 7 - Robustness of model results to use of 2016 sample data

	Mean Poverty	Mean MSE	Median CV	Correlation	RMSD	Coverage rate
Sampled municipalities						
Direct survey estimates (Horvitz-Thompson)	0.209	84.2	38.2	0.751	0.136	0.699
Household model	0.216	21.1	21.1	0.808	0.109	0.588
Sub-area model	0.210	62.2	40.3	0.777	0.120	0.818
Area-level model	0.166	38.0	30.7	0.749	0.126	0.672
Non-sampled municipalities						
Household model	0.315	116.9	35.8	0.669	0.217	0.667
Sub-area model	0.307	227.5	50.7	0.643	0.226	0.792
Area-level model	0.228	100.2	36.5	0.655	0.233	0.617

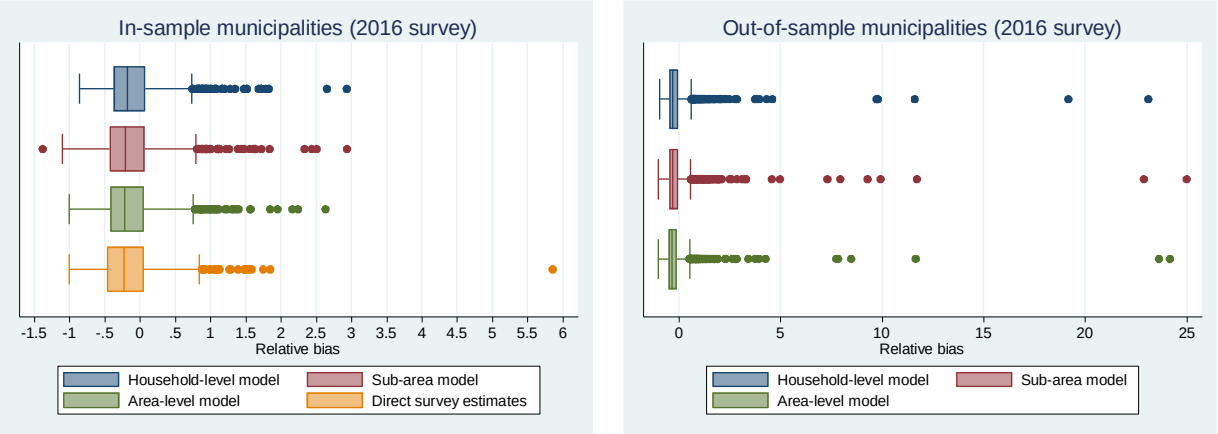
Notes: Table shows key results using the 2016 MCS-ENIGH household survey to estimate the model instead of the 2014 survey. In-sample refers to the 944 municipalities contained in the 2016 MCS-ENIGH survey, while out of sample refers to the 1,512 municipalities not covered by the survey. Direct estimates are reported using the Horvitz-Thompson approximation to generate variance estimates. Model-based estimates of MSE reflect benchmarking estimated poverty rates to direct estimates from the MCS 2016 survey at the state level. Columns are defined as in Tables 6 and 7.

Table 8 - Robustness of household model results to different sets of predictors

	Mean Poverty	Mean MSE	Median CV	Correlation	RMSD	Coverage rate
Sampled municipalities						
AGEB and municipal predictors with state dummies (baseline)	0.281	35.6	19.6	0.863	0.094	0.767
Municipal predictors with state dummies	0.281	36.0	19.6	0.859	0.095	0.768
AGEB predictors with state dummies	0.281	36.1	19.6	0.856	0.096	0.756
AGEB and municipal predictors, no state dummies	0.280	37.5	20.3	0.862	0.094	0.779
Municipal predictors only, no state dummies	0.280	37.3	20.3	0.860	0.095	0.778
AGEB predictors only, no state dummies	0.280	37.8	20.3	0.857	0.096	0.771
Non-sampled municipalities						
AGEB and municipal predictors with state dummies (baseline)	0.355	150.1	33.8	0.701	0.181	0.824
Municipal predictors with state dummies	0.352	153.6	34.2	0.687	0.184	0.831
AGEB predictors with state dummies	0.341	154.1	35.4	0.631	0.198	0.801
AGEB and municipal predictors, no state dummies	0.341	173.3	37.0	0.602	0.204	0.817
Municipal predictors only, no state dummies	0.343	175.2	37.4	0.607	0.201	0.824
AGEB predictors only, no state dummies	0.322	175.3	39.6	0.469	0.225	0.788

Notes: Table shows key results when using different sets of candidate variables to select the model. In-sample refers to the 944 municipalities contained in the 2016 MCS-ENIGH survey, while out of sample refers to the 1,512 municipalities not covered by the survey. Estimates of MSE reflect scaling by benchmarking factors. Columns are defined as in Tables 6 and 7.

Figure 4 - Box plots of relative bias, 2016 MCS-ENIGH estimates



Note: See notes to Figure 3. The household-level, sub-area, and area-level model estimates are produced by combining the 2016 MCS-ENIGH with satellite indicators. The direct estimates are generated using the 2016 survey alone.

Appendix A

Table A1 - Geographic structure of Mexico

Name	Sub-areas AGEBS		Areas Municipality	
	Sample	Population	Sample	Population
All states				
National	6,794	57,661	892	2,433
Urban	5,949	51,961	617	2,311
Rural	851	14,545	494	2,380
State groupings				
Group 1	938	7,586	57	105
Group 2	1,521	13,952	100	244
Group 3	1,845	16,810	188	363
Group 4	777	5,716	97	211
Group 5	1,221	14,358	283	757
Group 6	492	8,084	167	753

Notes: Figures give the number of AGEBs and municipalities in the sample and population of Mexico, broken out by urban and rural areas for the top panel. In the bottom panel, the six state groups match those used to develop the official municipal poverty estimates. More information is in Table C1 in appendix C. Population figures are taken from the 2010 census.

Table A2 - Descriptive Statistics for geospatial auxiliary data

	Household variables		**AGEB mean**	**Municipal mean**
Variable	Mean	SD	SD	SD
Sample				
Per capita income	3513.9	8100.3	4250.0	1680.3
Normalized per capita income	0.0	1.0	0.6	0.4
Poverty status	0.195	0.396	0.205	0.155
Predicted percent extremely poor	0.185	0.082	0.082	0.074
Predicted percent moderately poor	0.319	0.043	0.043	0.034
Predicted percent not poor	0.495	0.112	0.112	0.098
Percentage of pixels classified as:				
Building	0.110	0.103	0.103	0.079
Forest	0.068	0.134	0.134	0.116
Grassland	0.129	0.109	0.109	0.093
Road	0.126	0.086	0.086	0.069
Rural	0.244	0.430	0.429	0.334
Number	57,661		6,794	892
Population				
Predicted percent extremely poor	0.181	0.082	0.082	0.071
Predicted percent moderately poor	0.319	0.043	0.043	0.030
Predicted percent not poor	0.497	0.111	0.111	0.095
Percentage of pixels classified as:				
Building	0.124	0.113	0.113	0.088
Forest	0.063	0.130	0.130	0.106
Grassland	0.128	0.110	0.110	0.088
Road	0.135	0.090	0.090	0.072
Rural	0.221	0.415	0.415	0.266
Number	27,966,071		66,496	2,433

Notes: Mean and Standard Deviation reflect population-weighted mean and standard deviation across households. Third and fourth columns report standard deviations of AGEB and municipal means, across AGEBs and municipalities respectively. Means of AGEB and municipal means are identical to means of household variables reported in second column and therefore not separately shown. Population statistics are weighted by estimated average household size per AGEB, taken from the 2010 census. Sample statistics are weighted using sample weights.

Table A3 - Household, sub-area, and area-level models, post-LASSO OLS coefficient

Dependent Variable:	**Household model** Normalized household welfare	**Sub-area model** AGEB poverty rate	**Area-level model** Municipal poverty rate
Auxiliary variables - AGEB average			
CNN Predicted percent extremely poor	-0.34	0.06	
CNN Predicted percent not poor	0.79***	-0.26**	
Percent building	0.66***	-0.06	
Percent forest	-0.25***		
Auxiliary variables - Municipal average			
CNN Predicted extreme poverty rate	-0.97***	0.24	1.41***
CNN Predicted moderate poverty rate			-1.21***
Percent building	0.03		
Percent road		0.02	-0.37*
Percent forest		0.15**	
Percent grass	-0.55***	0.20**	
Percent background			-0.19**
Percent rural	-0.24***	0.08*	
State dummy variables			
Baja California Sur	0.18***	-0.06***	
Chihuahua	-0.39***	0.18***	0.16***
Mexico City	-0.12***		
Coahuila	0.15***	-0.03*	
Colima	-0.09***		
Durango	-0.18***	0.09***	
Guerrero	0.10***	-0.04***	-0.08**
Jalisco	-0.20***		
México	0.17***		
Nayarit	0.12***		
Nuevo León	-0.21***	0.12***	0.10***
Oaxaca	-0.23***	0.07*	
Puebla	0.17***	-0.06***	
Querétaro	0.09***		
Sonora	0.20***	-0.07***	
Tabasco		-0.02*	
Number of observations	57,660	6,792	893

Notes: Results are post-LASSO OLS regression results. Stars indicate statistical significance at 10, 5, and 1 percent. In the left panel, the dependent variable is normalized household per capita consumption as described in section 4. In the middle and right panels, the dependent variable is estimated AGEB and municipal poverty rates from the 2014 MCS-ENIGH survey, respectively.

Appendix B: Construction of CNN-Based Land Type Estimation and Poverty Predictions

One goal of this project was to predict poverty rates directly from imagery. Previous work in Sri Lanka predicted poverty rates using satellite-derived features such as number of cars and building density (Engstrom et al., 2022). Similarly, Jean et al. (2016) trained a CNN to predict the density of nighttime lights in several African countries. They then used features produced by that network to predict household income using a linear regression. This project, originally written up in Babenko et al. (2017), was an early attempt to train a CNN end-to-end to directly predict poverty rates. End-to-end training refers to methods where the entire training procedure is performed directly by the CNN's optimizer. Alternatives include applying post processing to the results of the CNN or using results from intermediate layers of the CNN as features in another model.

We used the administrative unit codes present in both the survey data and the shapefiles of the administrative units to match poverty rates to imagery. The images that we received from the satellite providers can cover hundreds of square kilometers and be irregularly shaped. Traditional computer vision algorithms require rectangular images that can fit onto computational hardware. We split the world up into rectangular tiles of a reasonable size that can be fed to the computer vision algorithms. These tiles are designed to have a similar number of pixels per tile. As such, their geographic extent varies with pixel size. A single image captured by a satellite may encompass thousands of these tiles. Figure B1 below shows a Planet tile in blue, a Digital Globe tile in green, and a sample urban AGEB in purple for comparison.

Figure B1 – Digital Globe and Planet Tile Sizes, Michoacán

For every tile that intersected with a surveyed administrative unit, we assigned that administrative unit's poverty distribution to the tile as the ground truth. In cases where more than one administrative unit intersected a tile, we assigned an average of those distributions weighed by geographic overlap and the total sampling weights of the administrative units. Tiles must have contained a certain amount of spatial intersection with the surveyed administrative units to be included in training or evaluation. We allowed this threshold to vary based upon the imagery provider to ensure sufficient tiles for training.

B.1 CNN Training

The CNNs were then trained using largely standard deep learning techniques. For each image, a loss that quantifies the difference between the predicted distribution and the true distribution was calculated. We used stochastic gradient descent, a popular optimization algorithm for neural networks, to minimize loss over the training data set. This effectively means iteratively adjusting the weights (coefficients) of the model to reduce loss on small subsets of the training data.

The computer vision algorithm was trained at the tile level. Upon the completion of training, we aggregated the tile level predictions to the administrative unit level. As administrative units commonly intersected with more than one tile, we took an average of those tiles to be the predicted distribution for that administrative unit. We weighed those tile predictions by how much of the administrative unit spatially intersects with the tile. In Figure B2, below, tile 5 would receive a high weight in the averaging process while tiles 7 and 9 would have little effect on the AGEB's final predicted distribution.

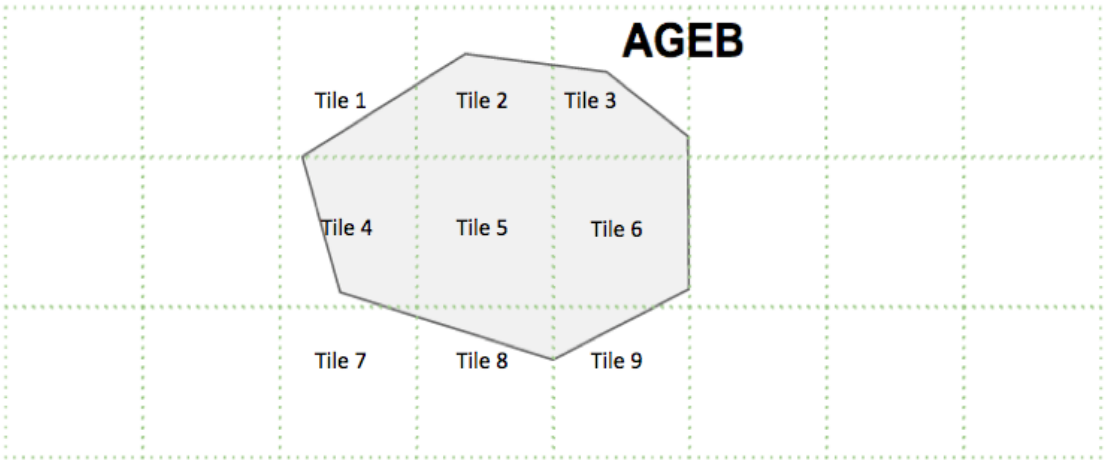


Figure B2 - Example Tile Aggregation to AGEB

B.2 Land Use Prediction

In addition to the end-to-end poverty prediction described above, we provided land-use predictions for all areas in Mexico that had cloud free imagery. We trained a land-use classifier that, for each pixel in an image, produces a likelihood that it belongs to one of several classes. This classifier was trained with Planet imagery from North America and Europe. For each pixel, the classifier assigns a probability distribution across 6 classes: building, road, water, grassland, forest, and background (all other classes). An example visualization of those results can be seen below in Figure B3. The first panel shows the RGB bands of a Planet image (this classifier also uses the Near-Infrared band when making classifications). The second panel shows the assigned likelihood that each pixel is a road. The third panel does the same analysis for buildings.

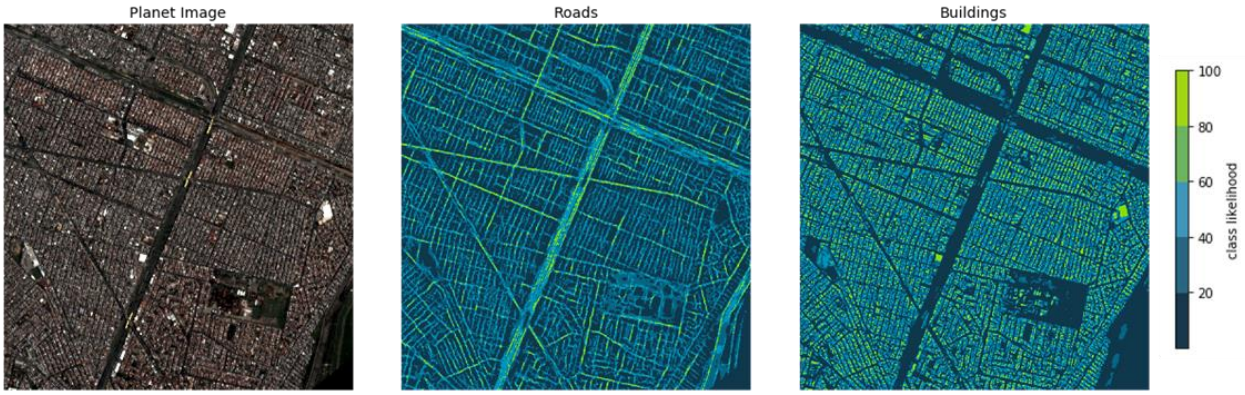


Figure B3 - Example Pixel-level Land-use Results, Mexico DF (Satellite image (c) 2017, Planet)

For each 3-5 m2 pixel in Mexico, the classifier produces 6 scores that sum to one. As such, the total volume of data in this process is substantial and large-scale analysis at this level is computationally expensive. We took the following approach to make large-scale analysis for policy and visualization purposes more feasible. We split each Planet tile into a 5x5 grid. For each grid cell, we took the distribution of pixels that were assigned to those six classes. Each grid cell contains six values, which sum to 1, which represent what percentage of the spatial area of that cell was predicted to belong to that class. Each grid cell was approximately 750 m2 in size. An example of these aggregated results is demonstrated below in Figure B4 overlaid on high-resolution imagery. The colors of the grid cells correspond to the proportion of the pixels in the cell classified as likely to be a road.

Additionally, these results can be used to visualize the land-use of Mexico and smaller areas of interest. This visualization is simplest and most effective for each class individually. We can make a heatmap that, for the class of interest, shows the percent of each grid cell that is made up by that class. In this case, areas that have a high proportion of that class will show up as bright and areas where that class is absent will show up as dark. An example is shown below of the roads class for all areas of Mexico in 2016 which had a cloud-free image. The second panel shows a zoomed in area around Mexico City along with a base map of the area.

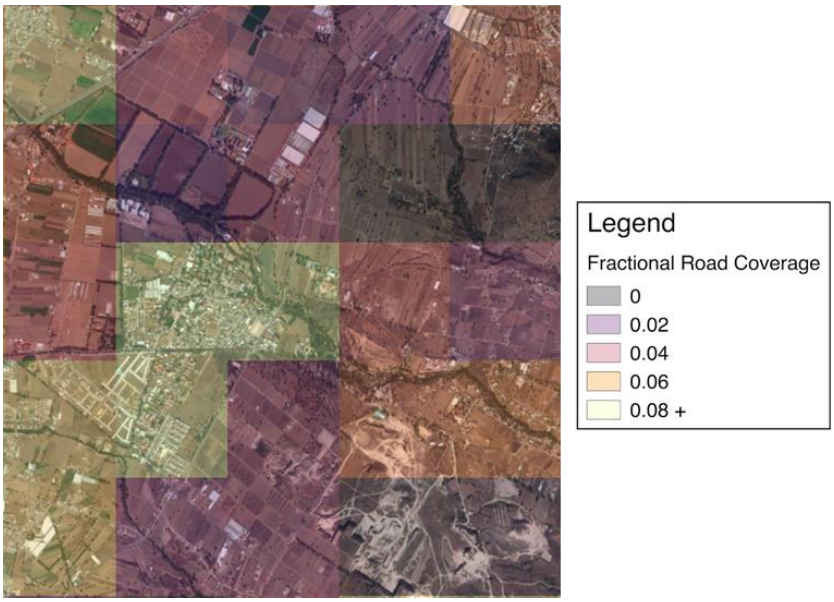


Figure B4 - Example of aggregated results, Mexico DF (Satellite image (c) 2017, Google)

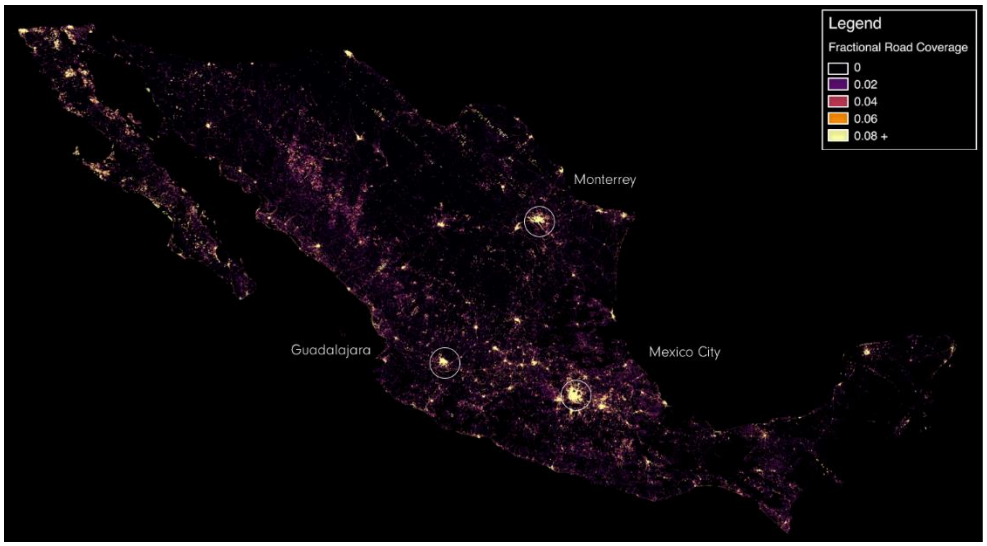


Figure B5 - Fractional Road Coverage, Mexico

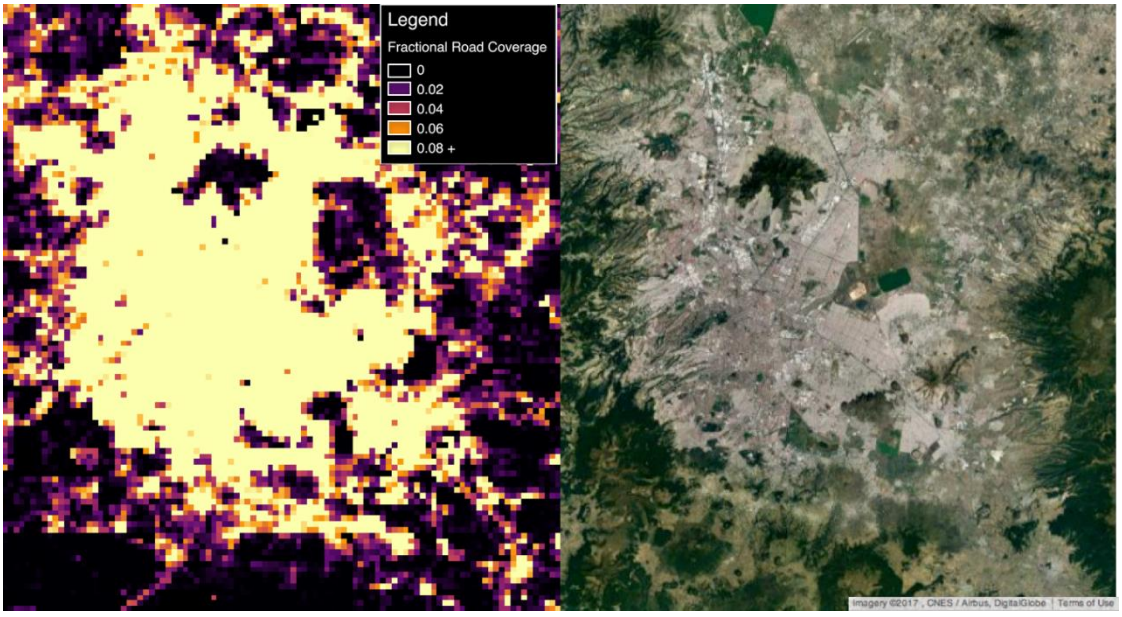


Figure B6 - Fractional Road Coverage, Close-Up of Mexico City Area and Base map

Appendix C: Implementation of the household model

As noted in the text, the unit-level model estimates presented in the paper are estimated using a modified version of the EBP function contained in the R EMDI package. There are two main modifications to the standard version of the package available at CRAN. The first incorporates census weights when aggregating up simulation results across households to the municipality level. The weights used in this application are the estimated average number of household members in each of the AGEBS, obtained by dividing the total population by the number of households in the 2010 census. In rural areas, the census data only contains total population and number of households by location instead of by AGEBS. Therefore, the population and number of households were aggregated up to calculate the total rural population and number of households per AGEB using publicly available AGEB-level shapefiles, under the assumption that no rural households lived outside of the localities provided in the census.

The second main modification to the standard EMDI package allows for household weights in the model estimation process. This was implemented in the following way:

1. Population weights were created as the product of household weights (factor_hog) and the number of household members (tot_integ), and normalized to sum to the number of sample observations in each municipality.

2. These weights (denoted "smp_weight") were incorporated into the estimation of the mixed effects model by adding the following option to the lme function called by the point_estim function in the EMDI package.

weights=varComb(varFixed(as.formula(paste0("~1/",smp_weight))),varIdent(~1)))

3. The calculation of the shrinkage factor g_m for each municipality in the point_estim function was adjusted according to the following formula, following You and Rao (2002) and Van der Weide (2014)

$$\gamma_m = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 * \frac{\sum_{i=1}^{n_m} w_i^2}{\left(\sum_{i=1}^{n_m} w_i\right)^2}} \qquad \text{(C.1)}$$

Where w_i is equal to smp_weight for household i in municipality m, and n_m is the number of sample households in municipality m. When w_i is constant, this is equal to the formula given in Molina and Rao (2010) and implemented by default in the EMDI package, which is:

$$\gamma_m = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_\varepsilon^2}{n_m}} \qquad \text{(C.2)}$$

As described in the text, we created a synthetic census based on the estimated number of households in each AGEB. The synthetic census consists of one observation per household, and contains the AGEB level satellite indicators as well as average AGEB household size, to use as a weight when aggregating simulation results across AGEBs. The average AGEB household size was estimated by dividing AGEB population from 2010 by the average household size for the associated state estimated in the MCS ENIGH survey. This was done separately for urban and rural areas. The number of households in each AGEB was set equal to the total population divided by the average state household size, rounded to the nearest integer.

Because of the large size of the synthetic census, it was much more efficient computationally to divide it into smaller groups when performing the estimation. Unfortunately, unlike the R SAE package, the EMDI package does not allow for estimation on a subset of domains. As a workaround, we divided households into six groups based on their state code, as listed in Table B1 below. The six groups are the same that were used by CONEVAL when generating the official municipal poverty estimates. For each group, the population was constructed by taking all households from the group's states in the synthetic census and appending a single randomly selected household from each municipality outside those states. This was done because the EBP function of the EMDI package requires the population data to contain at least one household from all areas contained in the sample. This procedure therefore produced six sets of estimates, one for each group. These were then appended together to create the final estimates, after discarding all municipal estimates based on a single household outside each group. Each set of estimates used the same national model based on the full MCS-ENIGH sample. As a robustness check, we also selected and estimated six separate models, one for each group, as shown in Table 8. Estimating a separate model for each group yielded qualitatively similar results, although mean squared error is slightly higher when estimating six separate models.

The baseline models "inflated" welfare in rural areas by multiplying them by the ratio of the urban to rural poverty lines. This was necessary because the EBP function in the EDMI package only allows for a single poverty line threshold, and after inflating rural welfare, the urban poverty could be used for both urban and rural areas. As a robustness check, however, we divided the urban and rural population for each group into separate subpopulations and estimated poverty rates separately for urban and rural areas. The results are shown in Table 8. The estimated poverty rates for each municipality were calculated as the population-weighted average of the urban and rural poverty rates. However, since estimated poverty rates in urban and rural areas of the same municipality are positively correlated, this procedure introduces an additional complication when estimating the uncertainty of the municipal poverty estimate. Urban and rural poverty rates within a municipality are correlated, so assuming that they are independent would substantially underestimate uncertainty. We therefore estimated the covariance of the urban and rural areas using the sample of municipalities that contain both rural and urban areas, by calculating the estimated covariance of the urban and rural headcount poverty rates across municipalities. To estimate the mean squared error of the municipal poverty rates, we took the weighted average of the estimated MSEs and added two times the estimated covariance. Therefore, for each municipality the poverty estimate is constructed as:

$$\hat{p} = w_u \hat{p}_u + (1 - w_u)\hat{p}_r \qquad \text{(C.3)}$$

$$MSE_{\hat{p}} = w_u^2 MSE_{\hat{p},u} + (1 - w_u)^2 MSE_{\hat{p},r} + 2w_u(1 - w_u)Cov(\hat{p}_u, \hat{p}_r), \qquad \text{(C.4)}$$

where w_u is the share of the population in urban areas, and $\hat{p}_u$ and $\hat{p}_r$ are poverty rates in urban and rural areas respectively.

As seen in Table 8, <note to check table number> this procedure gives similar point estimates but significantly higher estimates of uncertainty and higher coverage rates than the baseline procedure. This is partly because the baseline procedure underestimates uncertainty by failing to fully account for correlation across households in the same AGEBs. But it is also because this procedure applies the same covariance, estimated across all municipalities, to combine the estimated urban and rural poverty rates. Furthermore, this procedure assumes that bias is zero and that the MSE is therefore a variance estimate. These additional complications and assumptions led us to prefer the "inflated" welfare models, based on spatially deflated welfare applied to a single poverty line, as the baseline estimates.

Table C1 - State groupings

Group	
1	Coahuila, Distrito Federal, Nuevo León
2	Baja California, Baja California Sur, Colima, Jalisco, Quintana Roo, Sinaloa, Sonora
3	Aguascalientes, Chihuahua, Guanajuato, México, Morelos, Nayarit, Querétaro, Tamaulipas
4	Campeche, Durango, San Luis Potosí, Yucatán
5	Hidalgo, Michoacán, Puebla, Tabasco, Tlaxcala, Veracruz, Zacatecas
6	Chiapas, Guerrero, Oaxaca

Appendix D: Implementation of the sub-area model

The sub-area model is implemented as a standard nested error regression model, where the poverty rate estimated at the sub-area is modeled as a function of geospatial indicators specified at the sub-area level. The model is specified as follows:

$$\hat{p}_{ij} = x_{ij}\beta + v_i + u_{ij},\ i=1,\ldots,m;\ j=1,\ldots,J_i, \qquad (D.1)$$

where $\hat{p}_{ij}$ denotes the estimated poverty rate of sub-area j within area i, derived from the sample, X_{ij} is a vector of auxiliary variables specified at either the sub-area or area level, b is a vector of regression coefficients, $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ is an area random effect, and $u_{ij} \overset{iid}{\sim} N\left(0, c_{ij}\sigma_u^2\right)$, with c_{ij} assumed known, is a sub-area residual error term. In this paper, we choose $c_{ij} = n_{ij}^{-1}$, where n_{ij} is the number of sample households in AGEB j. Note that the subarea model (1) is different from the one proposed by Torabi and Rao (2014), which incorporates sampling variability of the estimated proportions, but is more complex to implement than subarea model (1). Auxiliary variables are selected using the rigorous LASSO procedure from the full set of candidate variables. The candidate variables include predicted extreme and moderate poverty and land classification estimates at both the AGEB and municipal variables, and nine state dummy variables.[25] The variables selected were extreme poverty and the building share at the AGEB level, the share building, forest, grass, and rural at the municipal level, and ten state dummies. A modified version of the EMDI package in R, modified to account for heteroscedasticity, was used to estimate the model.

[25] More details on the predicted poverty estimates and land classification variables can be found in Appendix A. The state dummies selected in the model were Colima, Chiapas, Federal District of Mexico City, Guerrero, Jalisco, Nuevo León, Oaxaca, Quintana Roo, and Tamaulipas.

Appendix E: Design-based simulations

In this appendix, we report the results of a design-based simulation using the intercensus survey. Because AGEB-level identifiers are not publicly available in the intercensus data, we cannot simulate either a household model with sub-area auxiliary data or a sub-area model. We therefore simulate a household-level model with municipal-level auxiliary data and compare the performance against a standard Fay-Harriot area-level model. The results are useful to verify whether the household-level model, even when using area-level covariates, generates predictions that are as accurate as an area-level model in repeated simulations. Second, the results give an indication of whether the area-level results reported in Table 2, Table 3, and Table 7 hold in repeated simulations.

The welfare variable used is household per capita labor income, which differs from the total household income measure used in the previous sections. We use the value of per capita labor income at the 28.2 percentile, when weighting by household weights, as a poverty line to match the 2014 monetary poverty rate. Households with per capita income below this threshold are classified as poor, while those above it are not. These classifications can then be used to compute the "true" poverty rates for each municipality in the intercensus, weighting by the product of the household weights and the number of household members for each household.

For each simulation, we draw a three-stage sample of households from the 5.8 million households in the intercensus as follows:

1. Draw 896 municipalities sequentially, without replacement, with probability proportional to size. The measure of size is municipal population taken from the 2010 census.
2. Draw up to nine PSUs (UPM) using simple random sampling. Sample all available PSUs if there are fewer than nine in the municipality included in the 2015 intercensus. We select nine to match the average number of PSUs per municipality in the MCS-ENIGH 2014 survey.
3. Draw up to seven households in each PSU using simple random sampling. Sample all available households if there are fewer than seven in the PSU included in the 2015

intercensus. We select seven to match the average number of households per PSU in the MCS-ENIGH 2014 survey.

4. Calculate weights as the reciprocal of the approximate inverse probability that a household was selected. We approximate the probability that a household was selected as:

$$w_h = \left[1 - \left(1 - \frac{Pop_m}{\sum_m Pop_m}\right)^{896}\right] * \frac{min(9, N_{psu})}{N_{psu}} * \frac{min(7, N_{hh})}{N_{hh}},$$

where Pop_m is the population of the municipality in which household h resides, N_{psu} is the number of PSUs in household h's municipality, and N_{hh} is the number of households in household h's PSU, according to the 2010 census. The first term approximates the probability that a municipality is selected.[26] The second and third term are the probability that the PSU and household are selected, respectively. To estimate the household model, these weights were normalized by dividing by their mean so that they sum to the total number of observations, as is common when estimating mixed models.

5. Normalize household per capita labor income using the ordered quantile normalization.
6. Estimate the poverty line as the 28.2nd percentile of the transformed household per capita labor income in the sample.
7. Estimate municipal poverty rates using a household unit-context model using the sample data and the intercensus population as the population data. The sample weight is the product of the original household weight from the intercensus and the additional sampling weight calculated in step 4. The population aggregation weight is the household weight times the number of household members provided in the 2015 intercensus data.
8. Calculate direct estimates using the sample, with the same weights as the previous step, and variances calculated using the Horvitz-Thompson approximation.
9. Merge the direct estimates with a version of the intercensus collapsed to the municipality level.
10. Estimate a Fay-Herriot model to generate predicted poverty rates using Stata's fayherriot command, utilizing the Li and Lahiri (2010) adjusted maximum likelihood method.

[26] This formula assumes sampling with replacement, which makes it an approximation.

11. Repeat steps 1 through 10 ninety-nine times, storing the predictions each time, for a total of 100 simulations.

The final step is to merge the actual municipal "poverty rates" from the intercensus data (based on low per capita labor income) and generate evaluation statistics.

Table E1 - Mean poverty and accuracy in simulation study

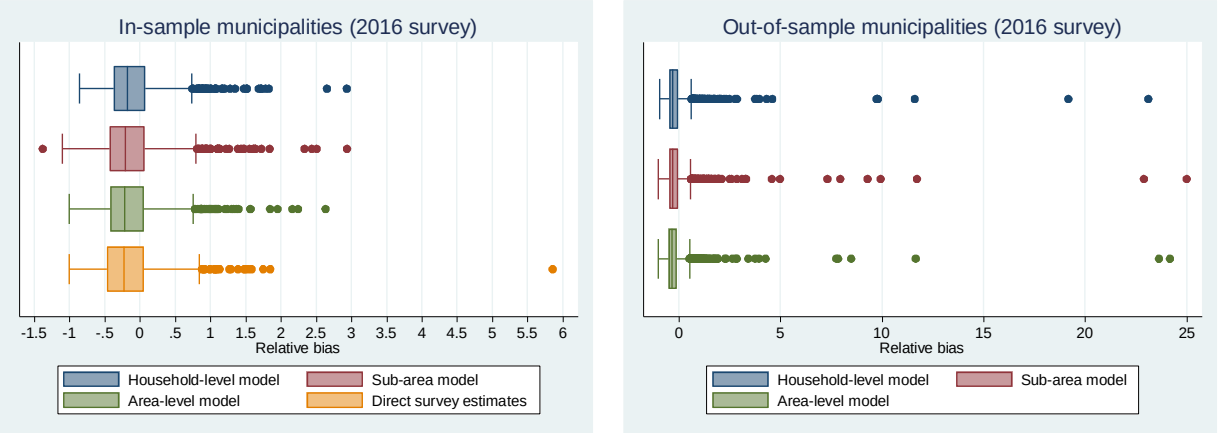
Average over 100 Simulations	Mean Poverty	RMSD	Correlation
Sampled municipalities			
Direct survey estimates (H-T)	0.368	0.294	0.926
Household model	0.358	0.272	0.941
Area-level model	0.336	0.274	0.937
Intercensus benchmark	0.379	0.000	1.000
Non-sampled municipalities			
Household model	0.456	0.380	0.803
Area-level model	0.434	0.405	0.749
Intercensus benchmark	0.522	0.000	1.000

Notes: Table shows results from one hundred simulations based on repeated samples from the intercensus survey, as described in section 7. Mean poverty refers to the simple average of poverty rates prior to benchmarking, RMSD refers to mean squared deviation, and Correlation refers to correlation, relative to the intercensus benchmark.

The evaluation results are displayed in Table E1. For sampled municipalities, the household model, direct estimates, and Fay-Herriot estimates all perform very well, as the correlations with the actual full intercensus are 0.941 for the household model, 0.937 for the area-level model, and 0.926 for the direct estimates. These results are substantially better than the correlations for the two MCS-ENIGH samples shown in Table 2, Table 3, and Table 7, particularly for the direct estimates. This is partly because the sample is drawn directly from the intercensus survey, the same data source used for evaluation. Out of sample, however, the area-level model performs worse than the household-level model, with a correlation of 0.75 compared to 0.8 for the household-level model. The household-level model also produces somewhat less biased estimates of poverty rates, for both

in-sample and out-of-sample municipalities. In sample, the average poverty rate for the household-level model, prior to benchmarking, is 35.8%, which is closer to the true value of 37.9% than that given by the area-level model (33.6%). The out-of-sample estimates are more biased than the in-sample estimates for both sets of estimates, but once again the household-level model estimates are less biased than the area-level estimates.

Finally, Figure E1 shows box plots for relative bias using the household and area-level simulations. These boxplots confirm that the relative bias is very similar for the household and area-level models for in-sample municipalities, each of which noticeably improves on the direct estimates. For out-of-sample municipalities, the household-level model exhibits markedly less relative bias than the area-level results, perhaps because it better incorporates sample weights when estimating model parameters.



Note: See notes to Figure 3. The household-level, sub-area, and area-level model estimates are produced by combining the 2016 MCS-ENIGH with satellite indicators. The direct estimates are generated using the 2016 survey alone.

Figure E1 - Box plots of relative bias, simulation results

Appendix F: Potential Omitted Variable Bias under Area-Level Models

Assume that data are generated by the following household-level data generating process:

$$y_{ah} = \beta_0 + \beta_1 x_{ah}^1 + \cdots + \beta_p x_{ah}^p + \eta_a + e_{ah}; \text{ h=1,…,}N_a\text{; a=1,…,A} \quad \text{(F1)}$$

Let $\overline{x}_a^k = \frac{1}{n_a}\sum_{h\in S_a} x_{ah}^k$ and $\overline{y}_a^k = \frac{1}{n_a}\sum_{h\in S_a} y_{ah}$ be the sample means of x^k and y_{ah} for area a, where S_a is the set of survey sample households and n_a is the number of sample households in area a. Let $\bar{X}_a^k = \sum_{h=1}^{N_a} x_h^k$ and $\overline{Y}_a = \sum_{h=1}^{N_a} y_{ah}$ equal the population means of x^k and y for area a, where N_a is the number of population households in area a . η_a and e_{ah} are area effects and household error terms assumed to be normally distributed.

Taking the average of (F1) across sample households in each area yields the data generating process for sample means:

$$\overline{y}_a = \beta_0 + \beta_1 \overline{x}_a^1 + \cdots + \beta_p \overline{x}_a^p + \eta_a + e_a \quad \text{(F2)}$$

Where $\eta_a \sim N(0, \sigma_\eta^2)$ and $e_a \sim N(0, \sigma_{e_a}^2)$. σ_η^2 and $\sigma_{e_a}^2$ are assumed to be fixed, and η_a and e_a are assumed to be independent of all predictor variables $\overline{x}_a^k$.

An area-level model fits:

$$\overline{y}_a = \beta_0 + \beta_1 \overline{X}_a^1 + \cdots + \beta_p \overline{X}_a^p + \eta_a + e_a \quad \text{(F3)}$$

Where, as noted above, $\overline{y}_a$ is the direct estimate for area a from the sample and $\overline{X}_a^k$ is the population average taken for example from administrative or geospatial data.

This omits variables

$\tilde{x}_a^k = \overline{x}_a^1 - \overline{X}_a^k$ from the model.

Model (F2), the assumed true area-level model, can be written in matrix notation as follows:

$$\boldsymbol{y} = \boldsymbol{x}\beta + \eta + \boldsymbol{e}, \quad \text{(F4)}$$

where

$$\boldsymbol{y} = \begin{bmatrix} \bar{y}_1 \\ \dots \\ \bar{y}_A \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} 1 & \bar{x}_1^1 & \dots & \bar{x}_1^p \\ \dots & \dots & \dots & \dots \\ 1 & \bar{x}_A^1 & \dots & \bar{x}_A^p \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_P \end{bmatrix}, \eta = \begin{bmatrix} \eta_1 \\ \dots \\ \eta_A \end{bmatrix}, \text{and } \boldsymbol{e} = \begin{bmatrix} e_1 \\ \dots \\ e_A \end{bmatrix}.$$

Adding and subtracting $\boldsymbol{X}\beta$ from the right hand side of (F4) yields the equivalent expression of:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \tilde{\boldsymbol{X}}\beta + \eta + \boldsymbol{e}, \quad \text{(F5)}$$

where

$$\widetilde{\boldsymbol{X}} = \begin{bmatrix} \bar{x}_1^1 - \bar{X}_1^1 & \dots & \bar{x}_1^P - \bar{X}_1^P \\ \dots & \dots & \dots \\ \bar{x}_A^1 - \bar{X}_A^1 & \dots & \bar{x}_A^P - \bar{X}_A^P \end{bmatrix}$$

On the other hand, model (F3), the estimated model, can be rewritten as follows:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \eta + \boldsymbol{e}, \tag{F6}$$

where:

$$\boldsymbol{X} = \begin{bmatrix} 1 & \bar{X}_1^1 & \dots & \bar{X}_1^P \\ \dots & \dots & \dots & \dots \\ 1 & \bar{X}_A^1 & \dots & X_A^P \end{bmatrix} \text{ and } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{bmatrix}$$

The Fay-Herriot model estimates the model coefficients $\hat{\beta}$ using weighted least squares:

$$\hat{\beta} = (\boldsymbol{X}'V^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'V^{-1}\boldsymbol{y}, \tag{F7}$$

where

$$V = \begin{bmatrix} V_1 & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & V_A \end{bmatrix}$$

and $V_a = \sigma_\eta^2 + \sigma_{e\,a}^2$ and where σ_η^2 is assumed to be constant for all areas and $\sigma_{e\,a}^2$ varies across areas.

Substituting (F5) into (F7) yields

$$\hat{\beta} = \beta + (\boldsymbol{X}'V^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'V^{-1}\widetilde{\boldsymbol{X}}\beta + (\boldsymbol{X}'V^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\eta + \boldsymbol{e}) \tag{F8}$$

And taking expectations gives:

$$E[\hat{\beta}] = \beta + (\boldsymbol{X}'V^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'V^{-1}\widetilde{\boldsymbol{X}}\beta \tag{F9}$$

Since $E[\boldsymbol{X}'\eta] = E[\boldsymbol{X}'\boldsymbol{e}] = 0$.

This means that the bias is equal to

$$Bias(\hat{\beta}) = E[\hat{\beta}] - \beta = (\boldsymbol{X}'V^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'V^{-1}\widetilde{\boldsymbol{X}}\beta \tag{F10}$$

In matrix format, $\boldsymbol{X}'V^{-1}\boldsymbol{X}$ can be rewritten as $\frac{1}{\sigma_\eta^2}\begin{bmatrix} \sum_a \gamma_a & \sum_a \gamma_a \bar{X}_1^1 & \dots & \sum_a \gamma_a \bar{X}_a^P \\ \sum_a \gamma_a \bar{X}_a^1 & \sum_a \gamma_a \bar{X}_a^{1\,2} & \dots & \sum_a \gamma_a \bar{X}_a^1 \bar{X}_a^P \\ \dots & \dots & \dots & \dots \\ \sum_a \gamma_a \bar{X}_P & \sum_a \gamma_a \bar{X}_a^1 \bar{X}_a^P & \dots & \sum_a \gamma_a \bar{X}_a^{P\,2} \end{bmatrix}$ (F11)

Where γ_a is the shrinkage factor for area a: $\frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_{\bar{e}}^2{}_a}$.

The second term of the bias, $X'V^{-1}\tilde{X}\beta$, can be expressed as a P X 1 matrix:

$$X'\,V^{-1}\tilde{X}\beta = \frac{1}{\sigma_\eta^2}\begin{bmatrix} \sum_{a=1}^{A} \gamma_a \sum_{k=1}^{P} \beta_k \\ \sum_{a=1}^{A} \gamma_a \bar{X}_a^1 \sum_{k=1}^{P} \tilde{x}_a^k \beta_k \\ \dots \\ \sum_{a=1}^{A} \gamma_a \bar{X}_a^P \sum_{k=1}^{P} \tilde{x}_a^k \beta_k \end{bmatrix} \quad \text{(F12)}$$

Meaning that the bias in total is:

$$Bias(\hat{\beta}) = E[\hat{\beta} - \beta] = \begin{bmatrix} \sum_{a=1}^{A} \gamma_a & \sum_{a=1}^{A} \gamma_a \bar{X}_1^1 & \dots & \sum_{a=1}^{A} \gamma_a \bar{X}_a^P \\ \sum_{a=1}^{A} \gamma_a \bar{X}_a^1 & \sum_{a=1}^{A} \gamma_a \bar{X}_a^{1\,2} & \dots & \sum_{a=1}^{A} \gamma_a \bar{X}_a^1 \bar{X}_a^P \\ \dots & \dots & \dots & \dots \\ \sum_{a=1}^{A} \gamma_a \bar{X}_P & \sum_{a=1}^{A} \gamma_a \bar{X}_a^1 \bar{X}_a^P & \dots & \sum_{a=1}^{A} \gamma_a \bar{X}_a^{P\,2} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{a=1}^{A} \gamma_a \sum_{k=1}^{P} \beta_k \\ \sum_{a=1}^{A} \gamma_a \bar{X}_a^1 \sum_{k=1}^{P} \tilde{x}_a^k \beta_k \\ \dots \\ \sum_{a=1}^{A} \gamma_a \bar{X}_a^P \sum_{k=1}^{P} \tilde{x}_a^k \beta_k \end{bmatrix} \quad \text{(F13)}$$

The resulting area-level predictions $\hat{\theta}_a$ are a weighted average of the direct estimate and the synthetic prediction:

$$\hat{\theta}_a = \bar{X}_a\hat{\beta} + \hat{\eta}_a = \gamma_a\bar{y}_a + (1-\gamma_a)\bar{X}_a\hat{\beta} \quad \text{(F14)}$$

Because $\sigma_{\bar{e}}^2{}_a > 0$ for all areas a, $\gamma_a < 1$, and any bias in the estimated coefficients $\hat{\beta}$ will also be present in the small area estimate $\hat{\theta}_a$. It is straightforward to show that the bias is equal to:

$$Bias(\hat{\theta}_a) = E[\hat{\theta}_a - \theta_a] = (1-\gamma_a)\bar{X}_a Bias(\hat{\beta}) \quad \text{(F15)}$$

Which is exactly equal to the bias identified in Corral et al (2021, p.39) for unit-context models. Thus area-level models that use population predictors, when the data generating process is derived from a household-level model, are susceptible to the same omitted variable bias. This model-based bias results from the discrepancy between sample and population means of the predictors, coupled with the assumption of a household-level data generation process, rather than the choice of a household versus area-level model. For both household and area-models, design-model bias, which considers bias prior to drawing the random sample, is zero.