Policy Research Working Paper 10077

Combining Survey and Geospatial Data Can Significantly Improve Gender-Disaggregated Estimates of Labor Market Outcomes

Joshua D. Merfeld
David Newhouse
Michael Weber
Partha Lahiri

WORLD BANK GROUP
Poverty and Equity Global Practice &
Social Protection and Jobs Global Practice
June 2022

Policy Research Working Paper 10077

Abstract

Better understanding the geography of women's labor market outcomes within countries is important to inform targeted efforts to increase women's economic empowerment. This paper assesses the extent to which a method that combines simulated survey data from urban areas in Mexico with broadly available geospatial indicators from Google Earth Engine and OpenStreetMap can significantly improve estimates of labor force participation and unemployment rates. Incorporating geospatial information substantially increases the accuracy of male and female labor force participation and unemployment rates at the state level, reducing mean absolute deviation by 50 to 62 percent for labor force participation and 25 to 52 percent for unemployment. Small area estimation using a nested error conditional random effect model also greatly improves municipal estimates of labor force participation, as the mean absolute error falls by approximately half, while the mean squared error falls by almost 75 percent when holding coverage rates constant. In contrast, the results for municipal unemployment rate estimates are not reliable because values of unemployment rates are low and therefore poorly suited for linear models. The municipal results hold in repeated simulations of alternative samples. Models utilizing Basic Geo-Statistical Area (AGEB)–level auxiliary information generate more accurate predictions than area-level models specified using the same auxiliary data. Overall, integrating survey data and publicly available geospatial indicators is feasible and can greatly improve state-level estimates of male and female labor force participation and unemployment rates, as well as municipal estimates of male and female labor force participation.

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

Combining Survey and Geospatial Data Can Significantly Improve Gender-Disaggregated Estimates of Labor Market Outcomes

Joshua D. Merfeld[*], David Newhouse[†], Michael Weber[‡], and Partha Lahiri[§]

JEL codes: J21, C13
Keywords: Small Area Estimation, Data integration, Geospatial Data, Labor Force Participation, Unemployment, Mexico

[*] KDI School of Public Policy and Management and IZA
[†] The World Bank and IZA
[‡] The World Bank and IZA
[§] Department of Mathematics and Joint Program in Survey Methodology, University of Maryland

1. Introduction

Low female labor force participation can be an important barrier to economic growth. In Mexico, for example, women's labor force participation is considerably lower than that of men and is also low in comparison to other countries. The female labor force participation rate was just 45 percent in 2019, compared to 77 percent for men, a gap of 32 percentage points. In the OECD, only Turkey and Italy have a lower rate of female labor force participation, while in LAC only Guatemala has a lower value than Mexico. This has important economic impacts. A recent World Bank analysis suggests that if women participated at the same rate as men, per capita income could be as much as 22 percent higher (Inchauste et al., 2021). The same analysis suggests that if Mexico could implement policies to increase the female labor force participation rate by 0.6 percentage points a year – in line with that observed in Spain, Ireland, and Chile – this would help eliminate the gender gap in labor force participation and increase economic growth by 0.4 percent a year, highlighting how increases in female labor force participation are not just an output of the development process, but a driver of it, as well.

Trends in female labor force participation are determined by several key factors, including the relationship between households' economic conditions and female participation, the supply and demand for jobs deemed appropriate for more educated women, national growth strategies, and occupational gender segregation (Klasen, 2018). To better understand from a spatial perspective how policy can address these factors, it is useful to understand where pockets of low rates of female labor participation exist. However, most labor force surveys can only provide adequate estimates for larger areas of aggregation. Mexico's labor force survey is no exception, as it can only provide estimates for urban and rural areas within states, as well as for select larger municipalities. Since Mexico has 32 states and 2,450 municipalities, reliable estimates of labor force participation – for both women and men – at the level of the municipality could greatly inform the geographic targeting of education, labor market programs, and other measures designed to improve female labor force participation. This is also true of other developing countries.

Small area estimation is a branch of statistics that combines survey data with richer auxiliary data to generate more precise and accurate estimates of statistics. It has frequently been applied to measures of poverty and labor market outcomes. An extensive literature has documented that auxiliary data, typically taken from census or other administrative sources, can improve estimates of labor force participation and unemployment (Esteban et al., 2020, Chambers et al., 2016, López-Vizcaino et al., 2015, Ugarte et al., 2009, Molina et al., 2007, Datta et al, 1999). This literature employs a variety of methods, including hierarchical Bayes, empirical Best Predictors, and multinomial logit mixed models, as well as area level models. Intuitively, small area estimation allows surveys to "borrow strength" from auxiliary data that is more geographically comprehensive, and therefore covers unsampled areas. Combining the survey data with model predictions for unsampled areas, derived from the auxiliary data, makes the estimates more accurate and precise. In addition, by modeling the average relationship between the survey and auxiliary data across the entire sample, small area estimation can help correct sampling error even in sampled areas. This method can therefore produce more precise estimates for geographic levels, such as Mexican municipalities, where the labor force survey is too small to generate reliable estimates.

At the same time, the rapid proliferation of publicly available "big data" from satellite and crowd-sourcing applications has made geographically comprehensive data freely available, while survey enumeration area geocoordinates or administrative shapefiles required to link surveys to geospatial data are also becoming increasingly available. These developments have sparked a burgeoning literature – ably reviewed in Burke (2021), McBride et al. (2021), and World Bank (2021) – on the use of satellite imagery for economic measurement. Most early work has focused on the ability of geospatial data to predict agricultural yields and crops (Erciulescu et al., 2019, Lobell et al., 2020), household asset wealth and poverty (Jean et al., 2016, Yeh et al., 2020 Steele et al., 2017, Engstrom et al., 2021), and population (Wardrop et al.., 2020, Engstrom et al., 2020).

This is the first paper to our knowledge that explores whether combining survey data with geospatial indicators can be used for the purposes of improving estimates of labor market outcomes, such as female and male labor force participation and unemployment, through small area estimation. Demonstrating that this methodology works is important, as many developing countries do not have the requisite data traditionally used for small area estimation, such as up-to-date censuses or detailed administrative data. We show that the incorporation of geospatial data – which is available across the globe – is a feasible alternative when preferred sources of auxiliary data are not available.

This paper also contributes to the ongoing discussion over the proper methods for combining survey and geospatial data for the purpose of prediction. Some analysts take an explicitly Bayesian approach (Steele et al., 2017, Pokhriyal and Jacques, 2017, Erciulescu et al., 2019), while others employ an empirical best predictor model (Battese Harter and Fuller, 1988 Masaki et al., 2020).

We undertake this exercise in the context of Mexico, which has publicly available data that makes our approach feasible. We focus on three separate levels of aggregation: the state, the municipality, and the AGEB (Área Geoestadistica Básica, i.e. the Basic Geo-Statistical Area). Importantly, the census data has AGEB-level identifiers for all urban AGEBs in the country. This allows us to match geospatial data to AGEBs. In addition, we simulate a random sample from the census before implementing our preferred small area estimation approach. The use of census data allows us to compare the resulting estimates to the "truth," derived from the census data itself. This type of data is rarely available in developing countries and, as such, Mexico is the perfect context in which to apply our approach.

It is important to note that our results on unemployment and labor force participation differ from official rates in two key ways. First, we include only urban AGEBs due to data limitations. Second, we are using census data, not a labor force survey, so the instruments differ. For example, the age range for which employment questions are asked differs across the instruments. However, despite these differences in the instruments, the underlying concepts of labor force participation and unemployment in the census and labor force survey are similar. There is therefore no reason to believe that the findings on the benefits of incorporating geospatial data would not apply to the official definitions as well.

This paper implements a sub-area model, which is essentially a unit-level model in which the unit is taken to be the AGEB. In particular, we specify a weighted empirical best predictor model, with conditional random effects specified at the municipality level in our main results. Municipal

predictions are then generated by taking the population-weighted mean of the AGEB-level predictions.[1] This approach leverages AGEB-level data, which leads to more accurate and efficient estimates than area-level models specified at the municipal level and requires minor modifications of publicly available software to generate point estimates and confidence intervals.[2]

Six main findings emerge:

1. Combining survey and geospatial data substantially improves the precision and accuracy of state-level estimates of both labor force participation (LFP) and unemployment rates. Estimated mean absolute deviation falls by approximately 62 and 50 percent for female and male participation rates, respectively, and by approximately 52 and 25 percent for female and male unemployment rates.
2. When considering municipal estimates of labor force participation rates, incorporating geospatial data significantly improves accuracy and greatly improves precision. In estimates from repeated simulations, estimated mean absolute deviation falls by about 43 percent for women and 53 percent for men, and rank correlation increases by 0.13 for women and 0.11 for men. After adjusting direct survey estimates to equalize coverage rates, estimated mean squared error for male and female LFP falls by a factor of more than four.
3. Because unemployment rates are very low, municipal estimates of unemployment rates based on a linear model are not reliable. Although estimated mean absolute deviation falls for both man and women, relative bias is high, rank correlation falls significantly for women and the estimates exhibit little variation.
4. Accuracy is substantially lower in out-of-sample municipalities than sampled municipalities.
5. When using a sample that simulates fully enumerating sampled AGEBs, small area estimates offer very minor improvements on direct survey estimates.
6. A model specified at the AGEB level generates more accurate estimates than one specified at the municipal level. The improvement due to using AGEB rather than municipal variation is much larger for male LFP, which is harder to predict, than for female LFP.

In short, the results support the use of augmenting survey data with geospatial data when estimating state-level statistics and when estimating municipality-level female and male labor force participation rates. When estimating municipal unemployment rates, however, the estimates are not reliable because the sample contains insufficient information to train a linear model. In particular, the sample contains no unemployed men in 44 percent of municipalities and no employed women in 68 percent of municipalities. Modeling unemployment rates using a linear model is problematic in this setting, highlighting the usefulness of software that can conveniently estimate and apply a wider range of non-linear empirical best predictor models.

[1] While this is in spirit similar to the sub-area model proposed in Torabi and Rao (2018), the approach outlined in that paper is difficult to implement using publicly available software.
[2] Code is available from the authors upon request.

2. Data and Methodology

a. Census and Survey Data

The primary source of data on labor market outcomes is the 2020 Census of Population and Housing, carried out in March 2020 by the Mexican National Institute of Statistics, Geography, and Informatics (INEGI).[3] The census was collected the month before the COVID-19 pandemic led to widespread shutdowns. INEGI publishes census statistics publicly at different geographic levels, with the lowest level being the Área Geoestadistica Básica (AGEB). INEGI publishes aggregate statistics only for urban AGEBs. We use these to generate municipality statistics that are weighted by the population of urban AGEBs. Because only urban AGEBs are included in the analysis, the results in this paper pertain to urban state and municipality-level rates. Urban AGEBs total to around 96.4 million people, or more than 75% of Mexico's estimated 127.6 million people.

We are primarily interested in examining the feasibility and effectiveness of combining geospatial data with survey data to improve municipal estimates of labor force participation rates and unemployment rates, separately for women and men. We also examine the ability of the procedure to improve state-level estimates of these four labor market outcomes. We restrict the analysis to urban AGEBs in order to utilize the census data as a credible benchmark against which to assess the performance of the small area estimates.

From the available urban census data, we construct a data set with six different variables: the total number of women in the labor force, the total number of women employed, the total number of women 12 years or older, plus the same three variables for men.[4] Since this is census data, we take these values to be the true values for each AGEB. We then randomly sample AGEBs from the full set of urban AGEBs to form a pseudo survey. We select AGEBs with probability proportional to size within states, which serve as the strata. We then choose the number of AGEBs per stratum (state) using sample sizes similar to INEGI's labor force surveys.[5] Like INEGI, we assume an average of 2.7 people 12 years of age and older per household when selecting AGEBs.[6] Out of 50,942 AGEBs with non-missing geospatial auxiliary features, 7,642 – or 15.0 percent – are included in the pseudo sample.

After selecting AGEBs, we simulate sampling individuals within each selected AGEB. The census data indicate the number of men and women the number of people 12+ living in each AGEB, as well as the number of people 12+ who are active labor force participants as well as the number employed. Since being employed is conditional on being in the labor force, we can use this information to simulate a sample of men and women within each AGEB based on these variables. We essentially construct the total population and randomly select 40 individuals within each AGEB – or all individuals if there are fewer than 40 men and women 12+ in the AGEB. These are then be used to construct gender-specific labor force participation and unemployment rates in the sample.

---

[3] The data are available online here.
[4] The labor data are collected for those 12 years of age and older, hence why we use 12+ as our population.
[5] More information is available here. One difference between our methodology and theirs – apart from the fact that we do not have access to actual enumeration areas – is that INEGI also includes some larger municipalities as strata. For simplicity, we only use states.
[6] This number comes from INEGI's labor force survey methodology manual in footnote 5.

Table 1 presents summary statistics for the 7,642 AGEBs in the sample, covering 1,034 municipalities. The first column presents statistics for the sample – for both women and men – while the second column presents actual AGEB population values from the census. Two striking facts emerge from the table. First, labor force participation for women is considerably lower than for men, both in the sample and in the population. Second, the number of unemployed (labor force minus employed) is very low across both columns. For example, in the sample the implied unemployment rate is less than 1 percent for women and less than 3 percent for men. The fact that the unemployment rate is close to zero makes it far more challenging to predict unemployment rates than labor force participation using a linear model in this sample. We return to this point below.

Table 1 - Mean number of individuals across AGEBs

	(1) Sample	(2) Population
Female		
Total	20.73	693.07
In labor force	10.62	356.04
Employed	10.48	351.09
Unemployed	0.14	4.94
Male		
Total	19.23	640.56
In labor force	14.54	482.42
Employed	14.21	471.19
Unemployed	0.33	11.23
Number of AGEBs	7,733	45,250
Number of municipalities	1,072	1,619

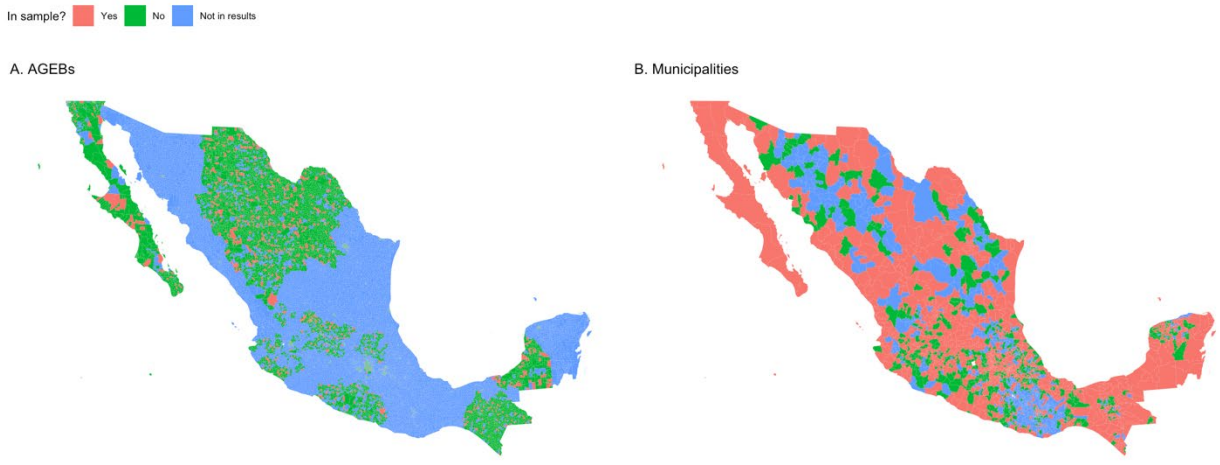

Figure 1 - Appearance in Sample: AGEBs and Municipalities

Figure 1 shows the AGEBs and municipalities that appear in the main sample. Panel A shows AGEBs, with individual AGEBs colored based on whether they are in the estimation sample ("Yes"), are not in our estimation sample but are urban ("No"), or are rural AGEBs that are not considered in the analysis ("Not in results"). Panel B shows the same, but for municipalities. Specifically, a municipality is coded as included in the sample if at least one AGEB within that municipality is contained in the sample. As such, a municipality in our sample could have a relatively small proportion of its total population actually included in our sample. There are 1,641 municipalities that contain at least one urban-AGEB with non-missing data. Of these, 1,034 – or 63.0 percent – have at least one AGEB in the sample.

b. Auxiliary geospatial data

Auxiliary data is drawn from two sources: Google Earth Engine, and Open StreetMap, utilizing the 2020 shapefile of urban AGEBs provided by INEGI. These sources were largely selected because they are publicly available, cover a large portion of the world, and convenient to obtain. However, both contain a large number of candidate predictors that could be plausibly correlated with spatial patterns in labor force participation. From Google Earth engine, we extracted summary statistics by AGEB from six datasets: Nitrogen Dioxide from Sentinel 5P, Normalized Difference Vegetation Index (NDVI) from the Sentinel 2 Multi-spectral Instrument, Nighttime lights from VIIRS, estimated population from WorldPop, land cover classifications from the Copernicus dynamic land cover map, and the year of development, as proxied for by the change in pixels from pervious to impervious surfaces (Gong et al., 2020). All summary statistics were taken over March 2020, except for land cover, which pertains to the period between January and December 2019.

From Open StreetMap, indicators were obtained representing the total length and number of highways in each AGEB. Two measures of road length were calculated: The first is the total length of the portions of highways contained within the AGEB, and the second is the total length of all highways that intersect the AGEB, including the portions of the road outside the AGEBs. In addition, the total counts of open StreetMap amenities and points of interest, such as hospitals, offices, schools, churches, and so on, were calculated for each AGEB. A final auxiliary variable, besides the Open StreetMap indicators, is the total area of the AGEB of the AGEB in sq km, which was calculated directly from the AGEB level shapefile provided publicly by INEGI. For each auxiliary variable, municipal variables were constructed by taking either population-weighted averages or simple sums (for counts and area) of the AGEB values.

c. Model selection

Variables for prediction were selected using the LASSO (Least Absolute Selection and Shrinkage Operator), with the penalty chosen to minimize the Bayesian Information Criteria (BIC) (Zhang et

al., 2010).[7] LASSO models were fitted separately to four dependent variables. These are the transformed versions of the four labor market outcomes, AGEB male or female labor force participation and unemployment rates, using the arcsine transformation. The candidate predictors included in the LASSO model include all of the AGEB and municipal level geospatial auxiliary data described in the last section, as well as dummy variables for states. The state dummies were included with no penalty, and were included to both improve the model and to prioritize the selection of variables that predict variation in the labor market outcomes within states.

d. Model estimation

We begin by constructing two separate estimates of LFP and unemployment in Mexico:

1. Direct estimates using only the pseudo survey of sampled AGEBs, which can only be constructed for in-sample municipalities.
2. Small area estimates using the pseudo survey of sampled AGEBs and the auxiliary geospatial data, which can be constructed for all municipalities, both in and out of sample.

These estimates are compared with the full census results along several different dimensions to assess their accuracy, as discussed below.

To generate the small area estimates for municipalities, we utilize a standard nested error model at the AGEB level (Battese, Harter, and Fuller, 1988, Jiang and Lahiri, 2006, Molina and Rao, 2010). The transformed version of the labor market outcome at the AGEB level is modeled as a linear function of geospatial auxiliary variables. The dependent variable was transformed using an arcsine transformation, a common transformation for proportions bounded between zero and one.

The model used to generate municipal estimates is specified as follows:

$$(1)\ \mathrm{G}(y_{sma}) = X_{sma}\beta_1 + X_{sm}\beta_2 + X_s\beta_3 + \upsilon_{sm} + u_{sma},\ \text{s=1,…,S,m=1,…,M}_s;\ \text{a=1,…,A}_{sm},$$

where $\mathrm{G}(y_{sma})$ denotes the arcsine of one of the four estimated labor market outcomes (male and female LFP, and male and female unemployment rate) of AGEB a within municipality m and state s, derived from the sample. The model is assumed to hold for the full population of AGEBs although it is estimated using the sample. X_{sma} is a vector of geospatial auxiliary variables specified at the AGEB level, while X_{sm} is a vector of geospatial auxiliary variables aggregated to the municipal level. X_s is a vector of state dummy variables. β_1, β_2, and β_3 are vectors of regression coefficients, υ_{sm} is distributed normally with variance σ_v^2, and is a set of municipality conditional random effects. $u_{sma} \sim N(0, \sigma_u^2)$ is a residual error term. As noted above, the set of predictor variables was selected using the BIC-minimizing lasso procedure from the full set of candidate variables, with the full set of state dummies X_s included with no penalty.

For the model used to generate estimates for states, the specification is similar, except that the conditional random effect is omitted from the model:

[7] The BIC was based on the unpenalized regression of selected variables rather than the penalized regression, and when estimating the lasso, standard errors were clustered at the municipality level.

(2) $G(y_{sma}) = x_{sma}\beta_1 + x_{sm}\beta_2 + X_s\beta_3 + u_{sma}$, s=1,…,S,m=1,…,$M_s$; a=1,…,$A_{sm}$.

The state-level effect υ_s is omitted because of the inclusion of a full set of state dummy variables X_s as independent variables. This mechanically causes the estimated variance of υ_s to be zero, due to the lack of variation across states in the residual after controlling for state dummies, which effectively drops υ_s from the model. In other words, the state-level estimates are generated from what is essentially a fixed effect estimator instead of a conditional random effect estimator. This is possible when generating state-level estimates, but not municipal level estimates, because the survey is sufficiently large to estimate state level results precisely.[8]

To generate the small area point estimates, the procedure simulates 100 draws from a normal distribution for both the random error term u_{sma} and the area effect υ_{sm} (or υ_s for the state level estimates) for each AGEB j. These estimates are then back-transformed, averaged across simulations, and finally aggregated across AGEBs to generate municipal or state-level estimates. When aggregating across AGEBs, we weight using estimated population from WorldPop.[9] Mean squared error estimates are obtained from a parametric bootstrap procedure, as proposed by Gonzalez-Manteiga et al.(2007) and implemented in the R EMDI package (Kreutzmann et al., 2018), based on the theory developed by Butar (1997) and Butar and Lahiri (2003).

When estimating the parameters of models (1) and (2), observations are weighted using normalized inverse probability weights, using the weighting method implemented in the R nlme package (Pinheiro, et al., 2021). Because each municipality is given equal weight when evaluating the estimates, it is necessary to normalize the weights to avoid giving too much weight to more populous municipalities in the sample estimation.[10] We therefore normalize the weights by dividing the inverse probability of selection by the mean inverse probability of selection for the municipality, as follows:

$W_{sma} = \frac{{\pi_{sma}}^{-1}}{\frac{1}{N_m}\sum_{a=1}^{N_m}{\pi_{sma}}^{-1}}$ where N_m is the number of AGEBs contained in the sample for municipality m and π_{sma} is the probability of AGEB a being included in the sample.[11] This normalizes the sum of the weights in each municipality to equal the number of sample observations, which is one recommended method for normalizing weights when estimating conditional random

[8] The same software package was used to estimate both municipal and state level results. We experimented with omitting state-level dummies and the results were worse.

[9] Using estimated population from WorldPop has the added benefit of being applicable to other contexts. While we have census information that would allow us to aggregate using arguably more "correct" data, this would not be possible in other contexts in which census data is missing or out of date. For this reason, we drop any AGEBs with an estimated population of zero. Only 0.6 percent of AGEBs are dropped.

[10] For some policy contexts, it may be desirable to give greater weight to more populous municipalities when evaluating the accuracy of estimates, but in this analysis, we limit consideration to the case where each municipal estimate is given equal weight when evaluating the performance of the estimates as a whole.

[11] We construct the sampling weight as the inverse of the probability of selection, using the formula for sampling with replacement, as an approximation for the actual probability of sampling based on our sampling strategy. In particular, the probability of selection is approximated as $\pi_{ij} = 1 - \left(1 - \frac{Pop_{sma}}{Pop_{total}}\right)^N$, where pop_{sma} = population of AGEB a, pop_{total} = total national population and $N = min(40, Pop_{sma})$, the number of individuals in the sample.

effects models.[12] Below, as a robustness check, we show that giving equal weight to each municipality worsens the predictions.

Finally, we ignore heteroscedasticity in the dependent variable, which is the average AGEB labor market outcome taken from the sample. A sample of 40 individuals was taken from each AGEB, except for the few cases where AGEBs contain fewer than 40 working aged persons, which implies that heteroscedasticity is not a first order concern when estimating labor force participation rates. Different rates of labor force participation across sample AGEBs will affect the number of labor force participants used to calculate unemployment rates from the sample, but we do not adjust for this source of heteroscedasticity.[13]

Population counts for each AGEB taken from the 2020 census were used as population weights when aggregating AGEBs to municipal estimates. We used a modified version 2.0 of the EMDI package in R, modified to incorporate sample and population weights, to estimate the model (Kreutzmann et al., 2018).[14]

e. Evaluating the estimates

To evaluate the estimates, we compare each estimate to the population values from the 2020 census. We calculate the following statistics to compare these four separate estimation methods:

- Estimated mean squared error. For the direct estimates, this is assumed to be equal to the estimated variance of the mean, estimated using the Horwitz-Thompson approximation.[15] For the small area estimates, MSE estimates were generated using the parametric bootstrap procedure.
- The median estimated relative standard error across municipalities, where the relative standard error for each municipality is defined as the square root of the estimated mean squared error divided by the point estimate.
- Coverage rate: This is the share of municipalities for which the actual census value falls within the 95% confidence interval generated by the estimate.

[12] See Rabe-Hesketh and Skondral (2006). In conditional random effect models, unlike in standard random effect models, multiplying the weights by a constant affects the parameter estimates.

[13] This is partly because the high number of municipalities in the sample with zero unemployment is a more serious issue than heteroscedasticity when estimating unemployment rates.

[14] The number of simulations and bootstraps, L and B, were each set to equal 100. The weighting strategy relies on passing through the specified weights to the lme function in the R nlme package that estimates the conditional random effect model (Pinheiro et al., 2020). In addition, we take a weighted mean instead of a simple mean when aggregating point estimates for AGEBs to generates estimates for municipalities. In the latter case, the aggregation weights are the population of the AGEBs.

[15] We use the Horvitz-Thompson approximation of the variance described in Molina and Marhuenda (2015), in which the variance of the sample mean for municipality m is estimated as:

$\hat{V}\left(\hat{\bar{Y}}_m\right) = \sum_{i=1}^{N_m} w_{mi}(1 - w_{mi})y_{mi}^2$

Where m is a particular municipality, i is an individual within municipality m, N_m is the number of relevant individuals in the municipality, w is the inverse probability sample weight, and y_{mi} is the binary outcome variable (participation or unemployment) for individual i.

- Estimated relative bias: This is the deviation of the estimate from its true value, expressed in percentage terms relative to the true value.
- Estimated mean absolute error: This is the absolute value of the difference between the estimated rate and the actual census rate.
- Estimated correlations with census value: This is a simple correlation between a point estimate and the census value and is a measure of the accuracy of the prediction.
- Estimated rank correlations with census value: This is the spearman rank correlation between a point estimate and the census value, a measure of the accuracy of the predicted rank.

Estimated in this context refers to the fact that these statistics are estimated using one sample. In each case, statistics are calculated giving equal weight to each municipality. All comparisons between direct survey estimates and the small area estimates are restricted to in-sample municipalities. This makes for more accurate comparisons, since the direct estimates only pertain to these municipalities. Later, we also show how results for out-of-sample municipalities compare with results for sampled municipalities.

f. Repeated simulations

A concern with drawing inferences about the performance from one sample is that any particular sample may not be representative. Since we construct this sample using a synthetic population, we can simulate the performance of the estimators across multiple samples. We therefore simulate 100 separate samples and calculate point estimates – for both direct estimates and small area estimates – for each of these samples. These are our preferred measure of the accuracy of small area estimates and direct estimates. However, due to computing time considerations, we do not perform a parametric bootstrap to calculate mean squared error in the simulation.[16] Therefore, our estimate of the efficiency gain due to small area estimation comes from only one sample.

g. Second-stage sample size

As noted above, a key issue with respect to modeling unemployment rates is that the indicator is very close to zero in Urban Mexico, especially for women. In our sample, for instance, female unemployment is less than 1.5 percent. However, our sample size for each AGEB is only 40 and around half of those are women, for just 20 women per AGEB. Additionally, labor force participation is about 50 percent, meaning that unemployment is necessarily calculated from a sample of sometimes as small as just 10 women. With an average unemployment rate of less than 1.5 percent, the great majority of these AGEBs have an estimated unemployment rate of zero.[17] The same is true for men.

[16] Drawing the sample, calculating the estimates, and performing the parametric bootstrap takes around two hours for a single sample. Since we simulate 100 different samples, this would take more than eight full days to estimate MSE across all simulations. Instead, we opt to calculate just point estimates – which takes only around 20 minutes for a single sample – and compare the performance of the point estimates.

[17] For male unemployment, approximately 72 percent of AGEBs have a direct estimate of zero. For female unemployment, it is 87 percent.

The low unemployment rates in the population lead to two issues. First, the estimated R^2 of the model is very low, meaning that the linear model predicts little of the variation in unemployment rates in the sample. This is not surprising, given the large amount of sampling error in the AGEB-level unemployment rates in the simple random sample. This does not rule out the possibility that the small area estimates improve on the direct estimates.[18] In this case, however, the large share of sample municipalities with zero unemployment also downwardly biases the variance estimate of the municipal conditional random effect. This in turn leads the predictions to give too much weight to inaccurate predictions from the model, relative to the sample, which in turn generates inaccurate estimates.

The most direct approach to this problem would be to implement a non-linear model such as a two-part model (Belotti et al., 2015), although doing so in an empirical best framework with existing software is not straightforward. However, increasing the size of the second stage of the sample may mitigate this issue by reducing the number of municipalities with zero unemployment in the sample.[19] As a robustness check, we calculate small area estimates using a hypothetical full enumeration of sampled AGEBs. In other words, we use the actual census value of the labor market outcomes in each selected AGEB to estimate the model. In many developing countries, it is not uncommon to perform a full listing of all households in selected enumeration areas before drawing the second stage sample. It may therefore be feasible to collect data on low-probability events like unemployment from all adults in selected enumeration areas, as part of a listing exercise.

3. Results

A. Model diagnostics

Before comparing results against the full census, we check the characteristics of the models themselves. Table 2 shows the number variables selected, the R^2 values of the regression, and a variety of other model diagnostics for the lasso and post-lasso results for all four outcome variables. The first panel presents the results for the state-level estimates, while the second and third panels present results for the municipality-level results. Marginal R^2 represents the variance explained by the auxiliary variables, while conditional R^2 represents the variance explained by both the auxiliary variables and the conditional random effect. The R^2 values for the simple random sample that generates municipal level estimates are generally lower than individual level models, particularly for modeling unemployment.[20] However, the conditional R^2 of the female LFP model is significantly higher, at approximately 0.3, and EBP models can perform well even in cases when the available covariates are not strong predictors (Marhuenda et al, 2017).

[18] This is counterintuitive but follows from Stein's paradox. Stein's paradox implies that a shrinkage estimator that takes a weighted average of direct survey estimates for subpopulations and the grand mean, when there are three or more subpopulations, can substantially improve predictions for the subpopulations (Stein, 1964, Efron and Morris, 1977). This is also consistent with Marhuendra et al.(2017), who note that EB estimates can perform well even in cases where the available covariates are not strong.

[19] The problem can also be fixed by using a non-linear estimator such as a two-part model (Belotti et al., 2015).

[20] R^2s may increase after adjusting for sampling error (Li and Lahiri, 2018).

Appendix Table A1 shows the actual post-lasso model coefficients for the random sample results. In general, areas that are more urban on average are associated with higher female LFP, as mean vegetation intensity is negatively associated with participation, while the correlation is positive for the number of residential roads and street crossings. Consistent with this, areas that have been urbanized more recently, as indicated by the mean year of switching to impervious surfaces, have lower participation rates all else equal. Furthermore, smaller municipalities (by area), many of which are in Southern Mexico, are correlated with lower female labor force participation all else equal.

The relationship between other indicators of urban density and female labor force participation is more complex, however. Areas that are more heterogeneous in terms of vegetation intensity, such as the suburbs, are other things equal associated with higher labor force participation. The relationship between night-time lights and female LFP is another example. The median value of nighttime lights is positively associated with female LFP while the mean value is negatively associated, although the latter is not statistically significant. This nonetheless suggesting that AGEBs with a large number of very bright pixels, perhaps due to the presence of highways, have lower female labor force participation rates.

Some of the infrastructure variables are more easily interpretable, as well, but these are rarely significant. The number of schools is significantly positively correlated with higher male LFP but does not show up for any of the other three indicators.

Appendix Table A2 decomposes model R^2 across different categories of variables. The latter shows that the determinants of male and female LFP differ significantly. For example, the geographic size of the AGEB and municipality, nighttime light luminosity, the mean of the vegetation index, the number of different points of interest recorded in Open StreetMap, and the year that pixels changed to being an impervious surface are important predictors of female LFP. Municipal area is positively associated with female LFP, which makes sense in the Mexican context because many of the smaller municipalities are located in the poorer south of the country where female LFP tends to be lower. Median nighttime luminosity in the municipality is positively associated with female LFP, consistent with nighttime lights indicating economic activity. Mean nighttime lights in the municipality, which conditional on median is a measure of right-skewness or inequality in luminosity, is negatively associated with LFP. This is consistent with women in more developed areas being able to afford not to work. Areas with a greater concentration of vegetation, as proxied by mean NDVI, have lower female LFP, perhaps also reflecting greater wealth in suburban areas with more vegetation. The correlates of male LFP are more difficult to interpret. The share of bare land in the municipality is positively and significantly correlated with male LFP, as is the number of untagged highway points, and the latter may be positively correlated with economic activity.

Table 2 - Model diagnostics by dependent variable and sample type

	Female LFP	Male LFP	Female Unemp	Male Unemp
State - Simple Random Sample within Selected AGEBs				
Number of (non-state) variables selected	41	22	19	10
Marginal R^2	0.127	0.061	0.025	0.028

Of which due to state dummies	0.055	0.018	0.017	0.022
Conditional R^2	0.127	0.061	0.025	0.028
Skewness of state effect	N/A	N/A	N/A	N/A
Kurtosis of state effect	N/A	N/A	N/A	N/A
Estimated variance of state effect	N/A	N/A	N/A	N/A
Estimated household residual	N/A	N/A	N/A	N/A
Ratio of estimated state effect variance to total variance	N/A	N/A	N/A	N/A
Municipality - Simple Random Sample within Selected AGEBs				
Number of (non-state) variables selected	41	22	19	10
Marginal R^2	0.132	0.071	0.038	0.032
Of which due to state dummies	0.055	0.018	0.017	0.022
Conditional R^2	0.293	0.170	0.132	0.077
Skewness of municipal effect	-0.102	0.222	4.047	1.011
Kurtosis of municipal effect	4.840	6.168	36.395	5.679
Estimated variance of municipal effect	0.004	0.002	0.0013	0.0008
Estimated household residual	0.017	0.017	0.012	0.017
Ratio of estimated municipal effect variance to total variance	0.186	0.107	0.098	0.046
Municipality - Full Enumeration of Selected AGEBs				
Number of (non-state) variables selected	33	26	28	39
Marginal R^2	0.277	0.224	0.164	0.158
Of which due to state dummies	0.113	0.060	0.082	0.100
Conditional R^2	0.597	0.451	0.467	0.525
Skewness of municipal effect	-0.610	-0.394	2.460	1.613
Kurtosis of municipal effect	6.128	4.779	24.136	12.472
Estimated variance of municipal effect	0.004	0.001	0.0011	0.0015
Estimated variance of household residual	0.005	0.003	0.0020	0.0020
Ratio of estimated municipal effect variance to total variance	0.442	0.292	0.363	0.436
Number of AGEBs	7,733	7,733	7,733	7,733
Number of municipalities	1,072	1,072	1,072	1,072

Dependent variable was transformed using arcsine transformation.

We focus here on the municipality-level diagnostics. The marginal R^2 values are generally low for the simple random sample, especially for unemployment rates. The low R^2 does not necessarily mean that the small area estimates do not improve on the direct estimates. The smaller the sample and the greater the heterogeneity in outcomes across AGEBs, the more likely a small area

estimation model is to improve upon sample estimates.[21] However, the low R^2 does raise concerns about biased estimates at the tails of the distribution, especially when estimating unemployment rates using the simple random sample. We expect that the combination of low marginal and conditional R^2 may lead to particularly inaccurate estimates for male and female unemployment rates when using the simple random sample. Model predictions for labor force participation using the simple random sample look significantly better for LFP than for unemployment, however, especially for women where the marginal and conditional R^2 values are 0.132 and 0.293 (as opposed to men where the same R^2 values are 0.071 and 0.17). The ratio of the estimated variance of the area effect to total variance is 0.19 for female participation and 0.11 for male participation, while the corresponding ratio for unemployment is about 0.1 for women and about 0.05 for men.

Using the full enumeration sample, not surprisingly, dramatically improves the prediction models. Marginal R^2 rises to about 0.27 for female LFP and 0.22 for men, and to approximately 0.16 for unemployment rates. The ratio of the variance of the estimated area effect to total variance ranges from 30 to 44 percent, depending on the indicator. Compared with the models developed using the simple random sample, more weight is given to the sample, which more accurately reflects municipal outcomes when sample AGEBs are fully enumerated. The higher R^2s of the models reflects the elimination of the considerable noise in the sample data, which drove down the R^2s of the models estimated using the simple random sample.

Finally, Appendix Figure A1 shows quantile-quantile plots for both components of the residuals, which is useful to check whether they appear to be normally distributed as assumed. The residuals are highly non-normal in most cases, except for the random area effect for male and female LFP. The non-normal random effects for unemployment, as well as non-normality in the household error terms may contribute to biased estimates of outcomes and uncertainty, which we will assess by comparing the small area estimates with the census. Future research can explore alternative transformations, such as the rank transformation used in Masaki et al.(2020), that might help achieve a more normal distribution.

B. State-Level Results

We first consider estimates at the state level, the level at which the survey is considered to be representative. We present the seven statistics listed in section 2.e, separately for direct and small area estimates, in Table 3.

The results are striking: Even though the direct estimates are considered to be reliable at the state level, we see large improvements in all seven statistics, across all four indicators, when incorporating geospatial data. The estimated mean squared error is smaller, the estimated correlation with the full-census benchmark increases, and the estimated absolute deviation from the benchmark decreases. Median relative bias declines for all outcomes. Coverage rates improve, sometimes dramatically. The estimates of unemployment rates improve on the direct estimates on most dimensions, even though the R^2 of some models is low. Because the outcomes measured in

[21] For example, suppose as in this case that the sample covers 15 percent of the enumeration areas. Even a model that explains 10 percent of the variation in the remaining 85 percent of the population contributes at least the additional 8.5 percent of the total variation in the population not captured by the sample, in addition to any improvement in the estimates for sampled areas due to averaging samples with model predictions.

the sample contain a large amount of random measurement error, even models with low R^2 can contribute significant amounts of additional information and improve the estimates.

Table 3 - State-level Results

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
Average estimated mean squared error (x1,000)				
Direct	0.3787	0.1983	0.0067	0.0031
SAE	0.0899	0.1263	0.0055	0.0021
Average estimated mean squared error (x1,000) with equalized coverage rates				
Direct	0.4048	1.2653	0.0555	0.0105
SAE	0.0898	0.1264	0.0055	0.0021
Coverage rate				
Direct	0.8750	0.7188	0.5625	0.5000
SAE	0.9375	1.0000	0.9062	0.6875
Estimated mean absolute error				
Direct	0.0179	0.0193	0.0048	0.0032
SAE	0.0088	0.0073	0.0023	0.0024
Median estimated relative bias				
Direct	2.1164	-3.0573	-10.7269	-12.3980
SAE	-0.0757	-0.4322	1.5223	15.1544
Estimated rank correlation				
Direct	0.5652	0.7933	0.7911	0.6635
SAE	0.7185	0.9498	0.8867	0.7529
Estimated Pearson correlation				
Direct	0.6048	0.8141	0.7528	0.6506
SAE	0.7614	0.9583	0.8698	0.5467
States	32	32	32	32

All figures represent unweighted averages or correlation across states. Equalized coverage rates refer to multiplying the estimated MSE of the direct estimates for each state by a constant to achieve the same coverage rate as the small area estimates. Rank or Pearson correlation is the unweighted rank or Pearson correlation between the estimated value and the actual state value. The absolute deviation is the absolute value of the difference between the estimate and the actual state value. Relative bias is 100 times the estimated minus the actual state value, divided by the actual state value.

Table A3 in the appendix presents a robustness check which calculates state-level results by aggregating municipality-level estimates up. The key difference between those results and the ones presented in this section are that the ones in Table 3 are estimated at the state level, without a random effect. The estimates in Table A3 are generally slightly worse than or similar to those in Table 3, with the possible exception of one accuracy indicator for male unemployment. The results in Table A3 give no empirical basis to prefer a method that aggregates the municipal estimates

with a random effect to one that estimates state results directly without a random effect. The latter has the advantage of being much simpler to estimate, particularly when it comes to measures of uncertainty.

C. Municipal-level estimates for sampled municipalities.

We now turn to municipality-level statistics, which the survey is not designed to estimate reliably. All reported estimates in this section pertain to in-sample municipalities only, so that the direct and SAE estimators are compared over the same municipalities. We begin by comparing measures of uncertainty for direct estimates and small area estimates. We present these results for all four labor market statistics in Table 4. The estimated mean squared error is always substantially smaller for the small area estimates than the direct estimates. This difference is largest – in percentage terms – for male labor force participation, where the estimated mean squared error is around a third as large as the direct estimate, roughly equivalent to effectively tripling the size of the sample.

The second row reports coverage rates, which is the share of municipalities for which the true census value is within 1.96 standard errors of the point estimate. Despite the smaller estimated mean squared error for the small area estimates, the coverage rates are also higher, again across all four indicators. For female labor force participation in particular, there is a significant increase in coverage rates from 87 to 98 percent in the small area estimates. This suggests that uncertainty for the direct estimates is underestimated and uncertainty for the small area estimates may be slightly overestimated. Therefore, for reach of the four indicators, we multiply the MSE of the municipal direct estimates by a constant greater than one to replicate the coverage rate of the small area estimates. When coverage rates are equalized in this way, the estimated MSE falls by a factor of four to five for labor force participation.[22]

Table 4 - Measures of Uncertainty and Coverage for Municipal Estimates

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
Average estimated mean squared error (x1,000)				
Direct	9.329	5.494	0.288	0.161
SAE	3.331	4.991	0.179	0.093
Average estimated mean squared error (x1,000) with equalized coverage rates				
Direct	13.895	24.116	N/A	N/A
SAE	3.340	5.042	0.179	0.094
Median estimated relative SE				
Direct	0.122	0.156	N/A	N/A
SAE	0.068	0.143	0.480	0.494

[22] Coverage rates cannot be equalized for unemployment rates because of the large share of direct estimates that are zero.

Coverage rate				
Direct	0.971	0.853	0.409	0.215
SAE	0.988	0.990	0.952	0.976
In-sample municipalities	1,072	1,072	1,072	1,072

Only in-sample municipalities are included. The coverage rate is equal to one in a given municipality if the actual census value is within the calculated confidence interval (point estimates +/- 1.96 times the standard error). MSE with equalized coverage rates refer to MSEs after multiplying the estimated variance of the direct estimates by a constant to match the coverage rate of the small area estimates. Relative SE is missing for unemployment due to the large number of zeros in the direct estimates.

We note that both male and female unemployment rates are not normally distributed in our sample, even after transforming them, due to the mass at zero. The empirical best predictor method used here assumes a normal distribution. This may contribute to uncertainty if anything being slightly overestimated, as indicated by coverage rates that lie between 95 and 99 percent.

We next present statistics intended to measure the accuracy of the estimates, relative to results in the full census. We start with estimated relative bias and absolute deviation in Table 5. These two measures differ along two key dimensions: The former is both relative, in the sense that it gives deviations from low true values greater weight than deviations from high true values, and directional, in the sense that negative and positive bias will cancel out. Estimated absolute deviation on the other hand is both measured in absolute terms and uses an absolute value and is therefore non-directional. It is also important in both cases to distinguish means from medians. Small (census) values of unemployment can lead to very large estimates of relative bias, since the bias formula contains the true value in the denominator.[23] For most policy applications relating to targeting, relative bias may be a misleading measure of accuracy for unemployment rates. This is because relative bias gives absolute discrepancies greater weight in municipalities with low true rates, which does not necessarily correspond to the objective of allocating resources to minimize unemployment.

For labor force participation, both estimated mean and median relative bias is modestly higher in the SAE estimates participation than in the direct estimates, though median relative bias is under 1 percent in both cases. For unemployment rates, estimated mean relative bias is much higher, reflecting the lack of explanatory power of the unemployment models, which in turn leads to large relative bias in areas where true unemployment rates are low. However, when it comes to absolute deviation the SAE estimates outperform the direct estimates by a substantial margin, for both LFP and unemployment. For female LFP, for example, the SAE procedure reduces estimated mean absolute error by about 45 percent, from 0.069 to 0.039, and for female unemployment rates the estimated absolute deviation falls by over half.

To get a further sense of the accuracy of the estimates, the bottom rows of the table report estimated rank and Pearson correlations. For the labor force participation estimates, the small area estimation increases rank correlation by 0.11 points for men, and by 0.13 points for women, from 0.59 to 0.73. For Pearson correlations the improvement is slightly smaller, 0.08 points for men and 0.11 points

[23] For example, the smallest non-zero value for female unemployment is 0.0006349. An estimated unemployment rate of 0.005 would lead to a bias value of 688.

for women, but still sizeable. When it comes to the unemployment rate estimates, the estimated rank correlations also show improvements when comparing the SAE estimates to the direct estimates, increasing by 0.9 for men and 0.6 for women (from an admittedly low level). However, the estimated Pearson correlations show a marked decline, falling by 0.22 for men and 0.065 for women. This again reflects errors at the left and right tails of the true unemployment rate distribution, which the model does a very poor job of predicting.

Table 5 - Measures of Accuracy for Municipal Estimates

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
	Mean (median)	Mean (median)	Mean (median)	Mean (median)
Estimated Relative Bias				
Direct	0.039	-1.017	-4.363	26.429
	(0.306)	(-1.288)	(-59.069)	(-100.000)
SAE	-0.507	1.423	30.732	99.948
	(-0.622)	(-0.726)	(9.651)	(40.522)
Estimated mean absolute error				
Direct	0.061	0.071	0.022	0.017
	(0.046)	(0.054)	(0.015)	(0.009)
SAE	0.028	0.040	0.009	0.006
	(0.022)	(0.028)	(0.006)	(0.005)
Estimated Rank correlation				
Direct	0.471	0.623	0.363	0.324
SAE	0.541	0.750	0.497	0.264
Estimated Pearson correlation				
Direct	0.493	0.659	0.418	0.385
SAE	0.553	0.770	0.415	0.614
In-sample municipalities	1,072	1,072	1,072	1,072

Only in-sample municipalities are included. Reported statistics are unweighted means, medians (in parentheses), or correlations across municipalities. Relative Bias is defined as 100 times the deviation of the estimated value and the true value divided by the true value. Rank or Pearson correlation is the unweighted rank or Pearson correlation between the estimated value and the actual municipal value.

The direct estimates of unemployment actually show negative estimated relative bias. This is likely driven by the fact that most AGEBs with a true unemployment rate of 1 percent are likely to have a sample with an unemployment rate of zero, leading to a consistent underestimate of the true rate. This is clear in Figure 2, which plots bias for male and female unemployment. Values of -100 indicate estimates of zero for unemployment. The direct estimates cluster disproportionately at this lower bound, indicating a large number of zeros. Although this averages out somewhat over municipalities (and different draws, which we return to below), the estimates for a single

municipality are often well off the mark. The figure shows that the small area estimates clearly reduce the estimated mean absolute deviation over the direct estimates for all four indicators, especially for small municipalities, despite high mean bias values for unemployment rates.

Overall, comparing results in sampled municipalities shows that the small area estimation procedure substantially improves the precision and accuracy of the estimates for labor force participation.

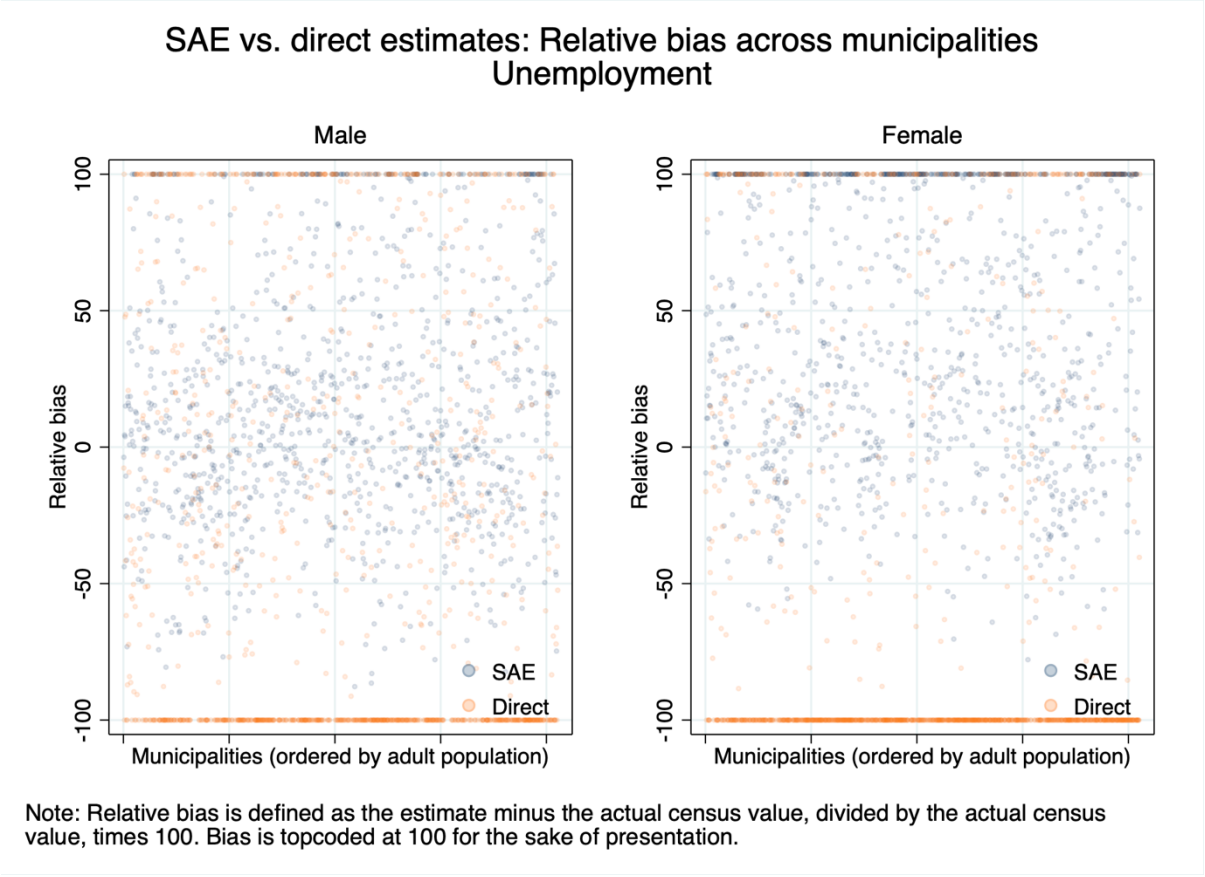

Figure 2 – Estimated bias for unemployment

D. Out-of-sample municipalities

The comparisons in the previous section examined direct estimates and small area estimates for sampled municipalities, but did not address whether out-of-sample predictions are accurate. We first compare selected statistics of these out-of-sample municipalities with the in-sample direct estimates. We present unweighted summary statistics of the estimated MSE, coverage rate, estimated absolute deviation, and estimated rank and Pearson correlation in Table 6.

Table 6 - In-sample direct estimates vs. out-of-sample small area estimates

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
Average estimated mean squared error (x1,000)				
Direct (in sample only)	9.329	5.494	0.288	0.161
SAE (in sample only)	3.331	4.991	0.179	0.093
SAE (out of sample only)	6.736	10.458	0.379	0.202
Coverage rate				
Direct (in sample only)	0.971	0.853	0.409	0.215
SAE (in sample only)	0.988	0.990	0.952	0.976
SAE (out of sample only)	0.974	0.956	0.940	0.979
Estimated mean absolute error				
Direct (in sample only)	0.061	0.071	0.022	0.017
SAE (in sample only)	0.028	0.040	0.009	0.006
SAE (out of sample only)	0.046	0.073	0.013	0.009
Estimated rank correlation				
Direct (in sample only)	0.471	0.623	0.363	0.324
SAE (in sample only)	0.541	0.750	0.497	0.264
SAE (out of sample only)	0.225	0.312	0.309	0.207
Estimated Pearson correlation				
Direct (in sample only)	0.493	0.659	0.418	0.385
SAE (in sample only)	0.553	0.770	0.415	0.614
SAE (out of sample only)	0.164	0.298	0.163	0.131
In-sample municipalities	1,072	1,072	1,072	1,072
Out-of-sample municipalities	569	569	569	569

Coverage rate is equal to one if the true census value is within 1.96 standard errors of the point estimate.

There are several patterns to note. First, looking at estimated mean squared error, estimates for non-sampled municipalities are approximately half as precise as those from sampled municipalities, and - except for male LFPs - are less precise than the direct estimates for sampled municipalities. Second, for unemployment rates, the coverage rate is markedly better for the out-of-sample small area estimates relative to the direct estimates. Third, small area estimates are far less accurate out of sample than in-sample. For example, for female LFP estimated absolute deviation is approximately twice as large and correlations are nearly twice as low for out-of-sample municipalities compared with in-sample municipalities. Comparisons between survey-based in-sample municipalities and out-of-sample small area estimates are more mixed. With the exception of female LFP, out-of-sample estimates have slightly lower estimated mean absolute deviation than in-sample survey estimates, but far lower correlations.

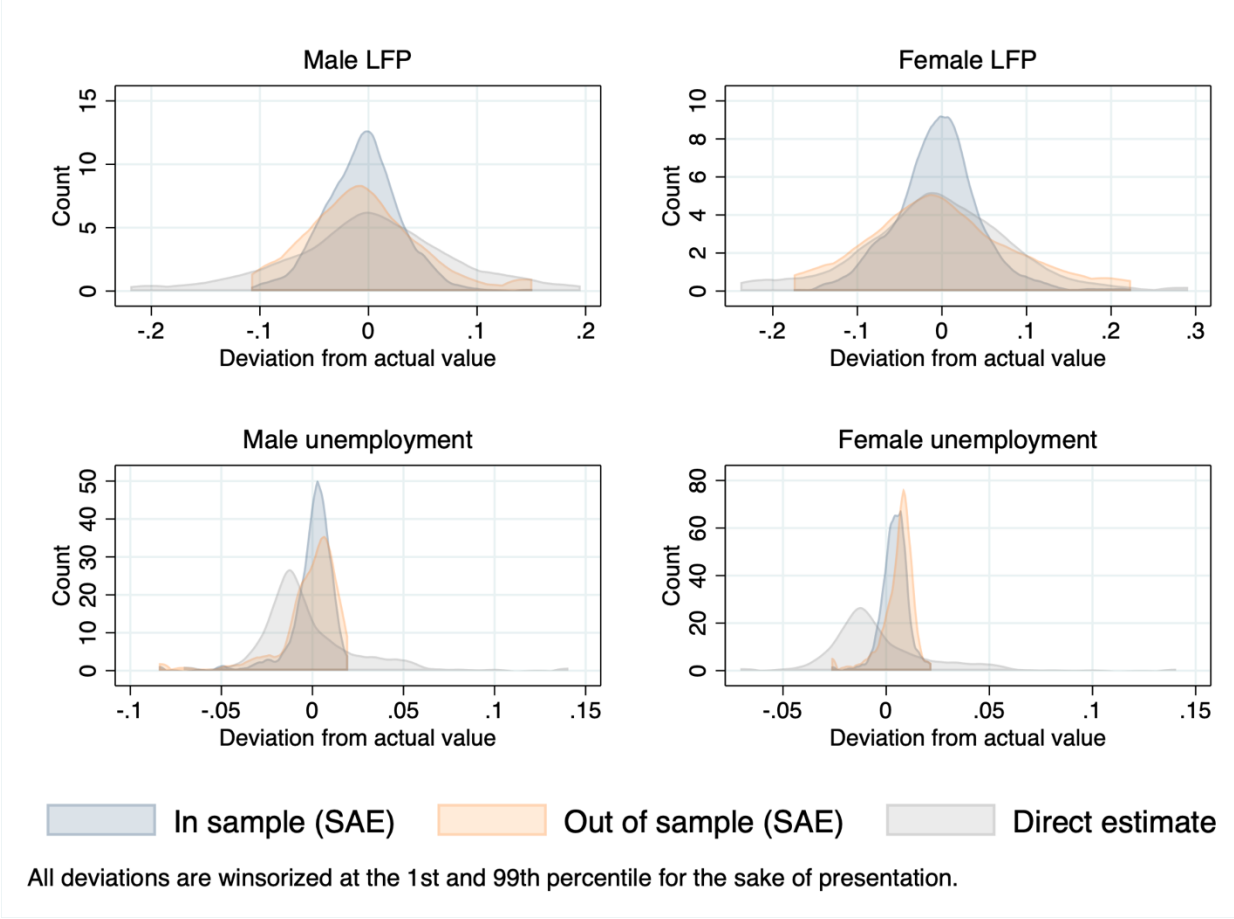

Figure 3 - In sample SAE, out of sample SAE, and direct estimates

Figure 3 presents kernel density graphs of deviations from the true value for three separate estimates: in-sample SAE, out-of-sample SAE, and direct estimates (which are only available in-sample). We see the tendency of direct estimates to underestimate the true outcome for unemployment, due to the large number of municipalities with zero unemployment in the sample. The out-of-sample estimates appear to be reasonable when compared to the direct estimates, which are in sample. The in-sample small area estimates clearly perform better than either of the other estimates for LFP, but the difference is less stark between in-sample and out-of-sample SAE for unemployment, especially for female unemployment. Overall, small area estimates for sampled municipalities are far more precise and accurate than both out-of-sample predictions and direct survey estimates.

E. Full enumeration of sampled AGEBs

As discussed above, a key issue complicating the prediction of unemployment is that unemployment rates in urban Mexico are very low. As a result, the sample contains a large number of municipalities with no unemployed persons, which adversely affects the accuracy of both the direct estimates and small area estimates. The large number of zeroes makes it difficult to predict

variation in unemployment rates, and also leads the small area estimation procedure to underestimate the variance of the municipal random effect, which in turn gives these inaccurate predictions more weight relative to the survey data in generating the estimates. Therefore, the small area estimates for unemployment suffer from high levels of relative bias and are more weakly correlated with the census values than the direct estimates.

While a non-linear estimation model might improve these estimates, a natural question is whether larger samples might improve the small area estimates for unemployment. This ties into a larger question about the potential benefits of expanding the second stage of samples when combining survey data with alternative data sources. To shed some light on this question, we repeat the small area estimation exercise after simulating a sample in which every household in selected AGEBs is included in the sample. This will reduce measurement error in the dependent variable, which substantially improves the quality of the predictions (Lobell et al., 2021, Engstrom et al., 2021). The exercise in intended to provide an upper bound estimate of the impact of expanding the second stage of the sample. But since listing exercises are common parts of household surveys in many developing countries, it may be possible to include selected questions as part of the listing exercise. The full enumeration results also shed light on the potential benefits of expanding the second stage of samples when combining survey data with alternative data sources.

Table 7 - Sample of sampled AGEBs vs. full enumeration of sampled AGEBs

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
Panel A: All municipalities				
Average estimated mean squared error (x1,000)				
Random sample SAE	4.512	6.887	0.248	0.131
Full enumeration SAE	1.306	3.450	0.147	0.049
Coverage rate				
Random sample SAE	0.983	0.978	0.948	0.977
Full enumeration SAE	0.951	0.943	0.941	0.960
Estimated relative bias (median)				
Random sample SAE	-0.759	-0.816	9.858	56.367
Full enumeration SAE	-0.510	-0.275	8.363	12.187
Estimated mean absolute error				
Random sample SAE	0.034	0.051	0.010	0.007
Full enumeration SAE	0.026	0.042	0.008	0.004
Estimated rank Correlation				
Random sample SAE	0.436	0.659	0.426	0.241
Full enumeration SAE	0.673	0.748	0.633	0.599
Panel B: In-sample municipalities				

Average estimated mean squared error (x1,000)				
Random sample SAE	3.331	4.991	0.179	0.093
Full enumeration SAE	0.869	2.014	0.084	0.033
Full enumeration direct	0.240	0.510	0.043	0.045
Coverage rate				
Random sample SAE	0.988	0.990	0.952	0.976
Full enumeration SAE	0.970	0.966	0.953	0.965
Full enumeration direct	0.543	0.538	0.537	0.547
Estimated median relative bias				
Random sample SAE	-0.622	-0.726	9.651	40.522
Full enumeration SAE	-0.491	-0.325	7.281	10.281
Full enumeration direct	-0.041	0.000	0.000	-0.186
Estimated mean absolute error				
Random sample SAE	0.028	0.040	0.009	0.006
Full enumeration SAE	0.017	0.026	0.006	0.003
Full enumeration direct	0.019	0.030	0.006	0.004
Estimated rank correlation				
Random sample SAE	0.541	0.750	0.497	0.264
Full enumeration SAE	0.819	0.866	0.805	0.741
Full enumeration direct	0.830	0.852	0.808	0.722

Table 7 presents summary statistics comparing the sample SAE and full enumeration SAE, as well as the direct estimates using the full enumeration of sampled AGEBs. Panel A compares across all municipalities, while Panel B compares only across in-sample municipalities. Not surprisingly, both the small area estimates and the direct estimates using the full enumeration sample perform much better than the small area estimates using the smaller, more typical, sampling strategy. The main finding from this exercise, however, is that the small area estimation only slightly improves on the direct estimates in this case, except for the coverage rate because uncertainty is estimated more accurately. For example, for female LFP, the small area estimates only reduces absolute deviation by 0.3 pp, from 2.9 to 2.6 pp, and only improves rank correlation from 0.86 to 0.875. For other indicators, the benefit of SAE when the second stage of the sample is fully enumerated is even smaller.

Figure 4 compares the accuracy of the random sample and full enumeration small area estimates, relative to the full census. The blue dots clearly demonstrate that the random sample SAE predictions for unemployment are close to the overall average, due to the poor predictive power of the unemployment models. In contrast, the orange dots representing the full enumeration unemployment results exhibit far more variation and better track the true rate, as do the LFP results for the random sample. This reflects the absence of variation available in the smaller sample to train an accurate model when predicting unemployment rates.

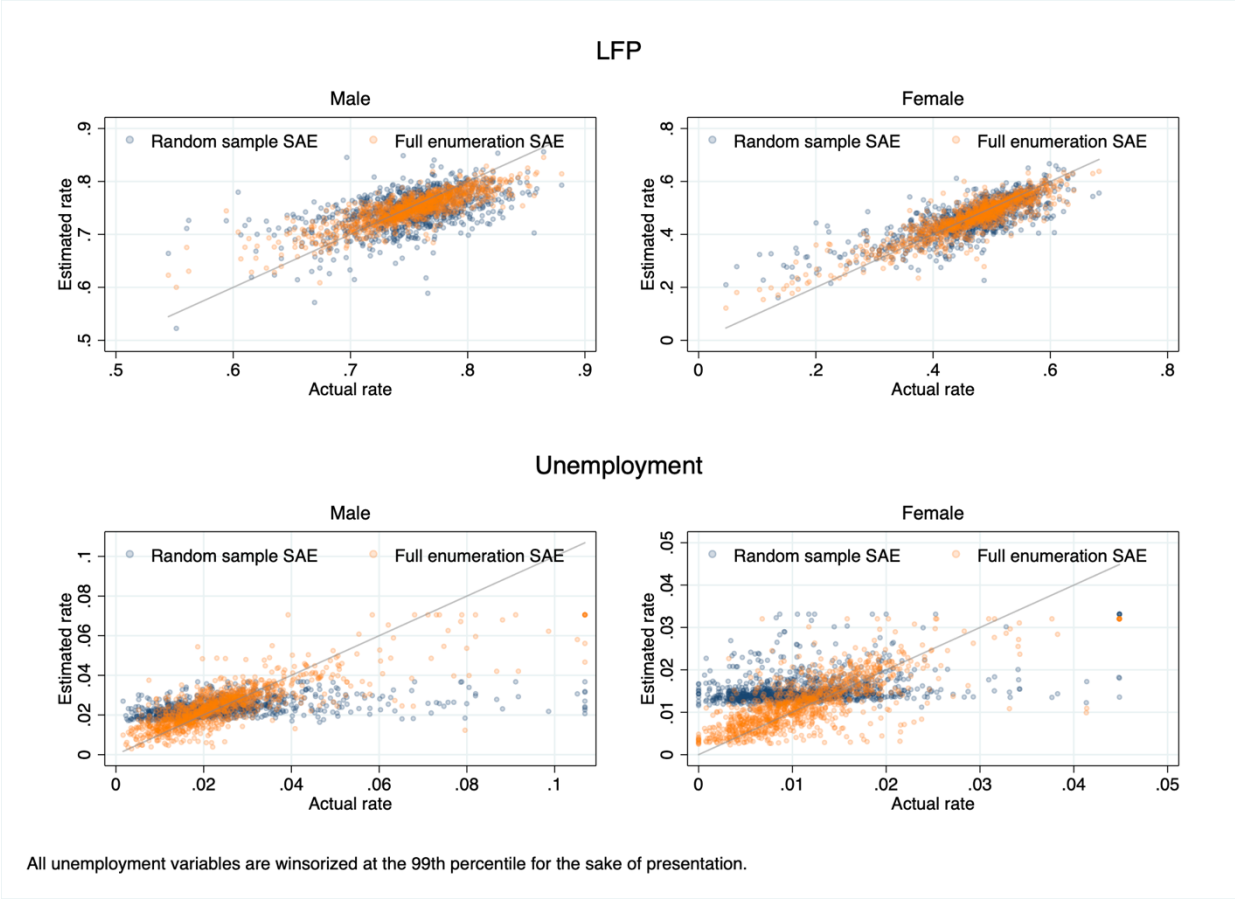

Figure 4 - Sample vs. Full Enumeration Small Area Estimates

To sum up, the results show that in most cases the use of this publicly available set of geospatial data improves estimates of labor force statistics in the context of urban Mexico. This improvement is seen at the state level – the level at which the pseudo survey is reliable, as well as at the municipality level, a level at which the pseudo survey is not considered to be reliable. In addition, small area estimation also generates synthetic predictions for non-sampled municipalities, although these are about half as accurate as the estimates for in-sample municipalities, when looking at estimated mean absolute deviation and correlation. When considering in-sample estimates for municipal unemployment obtained from the simple random sample, small area estimates do not always perform well. In this case, the estimated simple correlations between the small area estimates and the full census are markedly lower than those for direct estimates. Because true unemployment rates are close to zero and the sample is small, the simple random sample does not contain sufficient information to generate accurately municipal unemployment estimates with a linear mixed model. This is in part because the method gives too much weight to the inaccurate model predictions vis-à-vis the sample.[24]

[24] This is not an issue for the state level estimates because even though the model is not predictive, it includes state-level fixed effects which effectively incorporates the sample data into the estimates.

4. Robustness checks

A. Design-based simulations

The results in the previous section use just a single sample to compare results from two estimators: a direct estimator, using just survey results, and a small area estimator, which uses auxiliary information to improve the accuracy and precision of the survey results. However, it is difficult to come to a firm conclusion when using just a single sample. Therefore, we simulate the small area estimates 100 times, using 100 different samples and calculating 100 different point estimates for direct and small area estimates. We do not calculate MSE on each simulation due to computing limitations.

Table 8 - Simulation results for municipal estimates

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
	mean (median)	mean (median)	mean (median)	mean (median)
Simulated relative bias				
Direct	0.186	-0.222	0.128	1.451
	(0.534)	(-0.302)	(-58.406)	(-100.000)
SAE	-0.274	1.755	35.627	71.286
	(-0.656)	(-0.561)	(11.734)	(24.950)
Simulated mean absolute error				
Direct	0.060	0.070	0.023	0.015
SAE	0.028	0.040	0.009	0.005
Simulated rank correlation				
Direct	0.443	0.614	0.366	0.361
SAE	0.555	0.742	0.449	0.389

Rank correlations are the average rank correlations across the 100 simulations.

Table 8 presents three separate statistics – bias, absolute deviation from the census value, and the correlation with the true value – for direct estimates and small area estimates. Simulated mean and median bias is quite small for both direct estimates and small area estimates for LFP, with the highest rate being 2 percent mean bias for female LFP. The means are always higher than the medians, especially for the small area estimates for unemployment, as very low base rates can lead to very high values in some simulations.

In terms of mean absolute error, small area estimates perform markedly better. For LFP, simulated mean absolute error for the small area estimates is between 30 and 50 percent smaller than the direct estimates. For unemployment, the difference is even starker;
for both male and female unemployment, the small area estimate's simulated absolute deviation is less than half the size of the direct estimate's absolute deviation. The standard deviation across simulations is also markedly lower for SAE, across all four outcomes, with the difference again largest for unemployment. Finally, rank correlations are about 0.1 higher for the small area estimates than the direct estimates for LFP, 0.08 higher for male unemployment, and about the same for female unemployment.

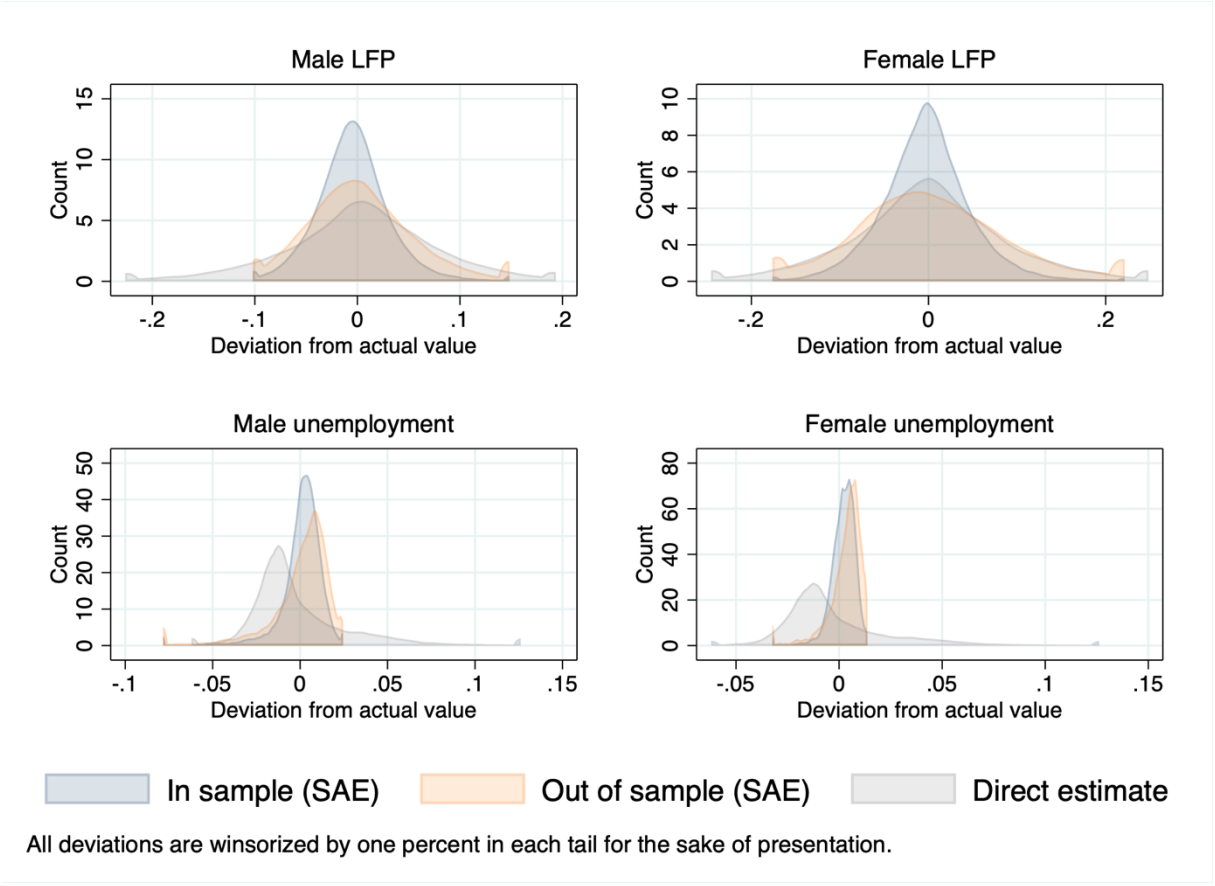

Figure 5 - In sample SAE, out of sample SAE, and direct estimates
Deviation from truth across 100 simulations

Figure 5 plots the deviations from census values across all the simulations. Both male and female unemployment have a mass below zero for the direct estimates, implying an underestimate of the actual value. Moreover, there is a long right tail for the direct estimate. This is even more pronounced in smaller municipalities, for whom the sample size tends to be much smaller (since fewer AGEBs are selected in each simulation). Figure 6 plots the same kernel density estimates, except only for municipalities below the median adult (12+) population. The underestimate for

unemployment is more pronounced here, with very little mass at zero, a large mass below zero, and a long right tail.

Figure 6 - In sample SAE, out of sample SAE, and direct estimates
Municipalities below median population

We present one final graph to better understand the variability in the direct and small area estimates across simulations. Figure 7 plots each municipality across the x axis. The y axis is the deviation from truth, with the bars showing the range (maximum value to minimum value) across the 100 simulations for each municipality. The results are stark; the range for direct estimates are markedly larger than for small area estimates. This is especially true for unemployment, with most municipalities having some very large deviations in the direct estimates, but very few showing similar deviations with the small area estimates.

It is worth noting that there do appear to be some municipalities that are consistently estimated poorly by the small area estimation. For both male and female LFP, we see a few municipalities that are always overestimated. This of course does not happen with the direct estimates, which are a random sample and thus generally cross zero.

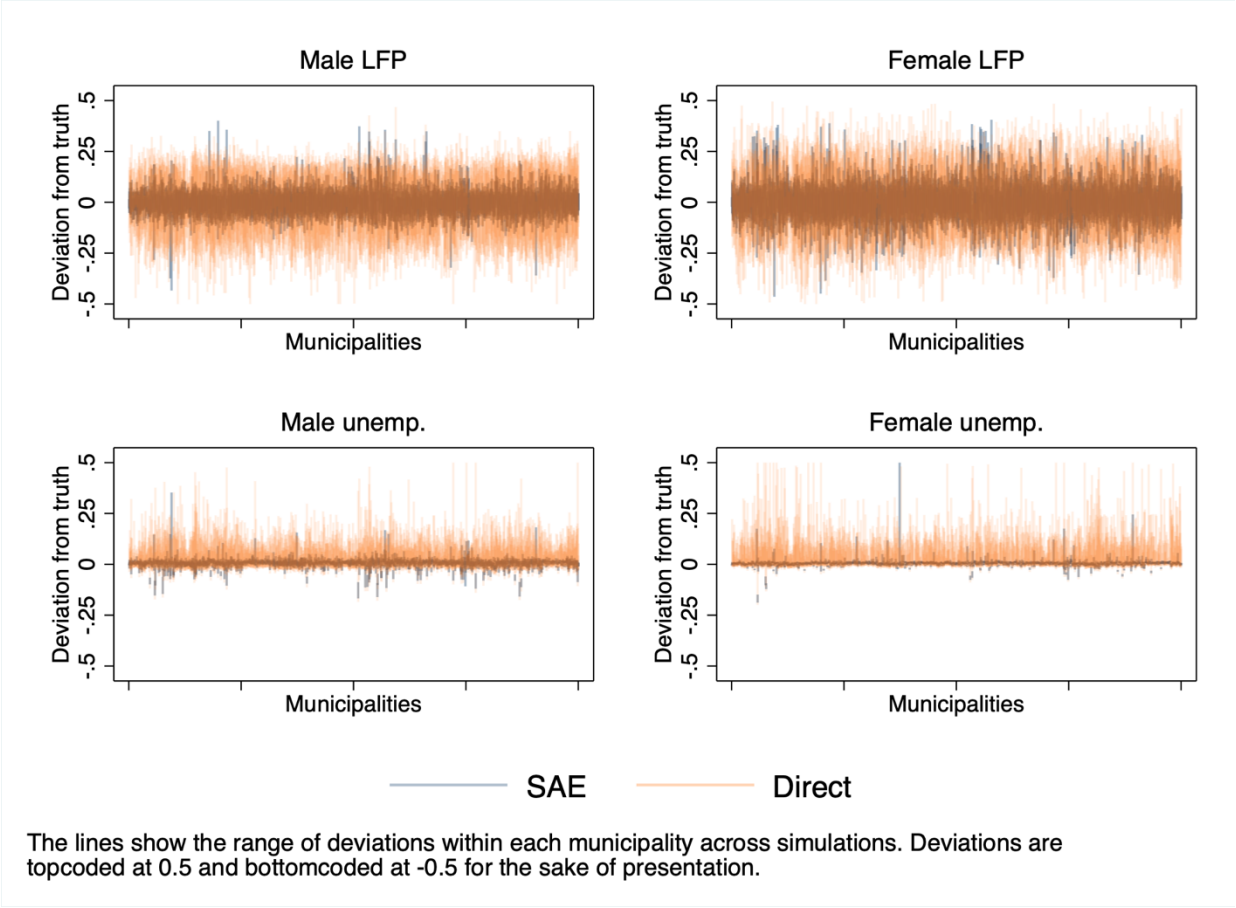

Figure 7 - Range of deviation from truth across simulations

A natural question is whether we can say anything about the characteristics of the municipalities for which predictions are overestimated. One obvious characteristic is population. We look at municipalities for which the minimum deviation is above 0.1, meaning that the entirety of the range lies quite a bit above zero and that we are consistently overestimating LFP for these municipalities. For male LFP, there are 33 such municipalities. The average number of adults in these 33 municipalities is 3,223. This is in stark contrast to the other municipalities, with a mean of 49,618 adults. Similarly, for female LFP, there are 83 such municipalities, with a mean adult population of 3,898, compared to a mean of 51,071 for the municipalities that are more accurately estimated.

It seems that municipalities with less accurate predictions are, on average, much smaller municipalities. Moreover, if we look at simple unweighted correlations, smaller municipalities are also less accurately predicted in general across the simulations. For example, the correlation between the small area estimates for male LFP and the actual census value for municipalities above the median adult population is 0.560, while the correlation for municipalities below the median is just 0.317. We see similar differences for female LFP (0.677 vs. 0.532), male unemployment (0.399 vs. 0.157), and female unemployment (0.286 vs. 0.040). If we weight by the relevant population value, the differences are even larger.

One interesting area of future research will be to determine whether these smaller municipalities are simply more idiosyncratic and, thus, harder to predict, or whether they differ structurally in observables ways. If the latter is true, different types of data or methods might improve estimation for the smallest municipalities. In other words, is it possible to improve the estimation of these municipalities, or are there simply weaker systematic relationships between geospatial data and labor market statistics for these areas?

B. Area-level model

An alternative method for combining survey data and geospatial data is to obtain municipal estimates is an area-level model (Fay and Herriot, 1979). This method is convenient and well-known, but suffers from a few limitations in this context. The main shortcoming is that it requires aggregating the auxiliary data to the municipal level, which discards variation across AGEBs within municipalities. In addition, the method faces challenges to accurately generate estimates of municipal variance, which is an important input into the model. This is particularly true when there is a significant number of zero values, as is the case for unemployment rates in this context.

To better understand whether these potential limitations are important in practice, we examine the results of two area-level modeling exercises. Fay-Herriot models can either be assumed to be linear models or the dependent variable can be transformed prior to estimation. For ratios such as labor force participation and unemployment rates, it is common to transform the variable using an arcsine transformation (Halbmeier et al, 2019). Use of such a transformation adds additional complexity to the procedure, but ensures that predicted participation or unemployment rates lie between zero and one. In addition, using the arcsine transformation is more consistent with the AGEB-level model discussed above. For these reasons, we use the transformed model as our main results, but display the results of a linear model in Table A4.

We select a model for each outcome by applying LASSO to the full set of municipal candidate geospatial variables, again selecting lambda to minimize BIC. When using the arcsine transformation, the dependent variable is the survey-weighted estimate of the transformed mean outcome variable of interest in each municipality. When using the linear model, no transformation is applied prior to the LASSO. We then calculate the sampling variance of each outcome using the Horvitz-Thompson approximation, as recommended by Halbmeier et al (2019), and fit the model using the restricted information maximum likelihood option of Stata's fayherriot package. The R^2s of the transformed models are quite low: 0.06 and 0.19 for male and female LFP, and only 0.10 and 0.06 for male and female unemployment. Even after reducing the noise associated with the unemployment measure by aggregating to the municipality level, the geospatial variables are weak predictors of unemployment rates and male LFP, although they predict cross-municipal variation in female LFP somewhat better.

Table 9 displays the results from the transformed Fay-Herriot (FH) area-level model.[25] We estimate the FH using an arcsine transformation of the target variable, just like with the sub-area

[25] The table omits MSE because the State fayherriot command used to estimate the FH model does not output standard errors. Instead, it outputs confidence intervals derived from a bootstrap procedure. Because the confidence intervals are asymmetric, calculation of the MSE is not straightforward.

(AGEB-level) model. Across all statistics and labor market indicators -- except one -- the area-level model performs worse than the sub-area model. The exception is the estimated rank correlation for female unemployment, which the AGEB-level model predicts poorly due to the large number of zeroes in the unemployment indicator. The area-level model often performs better than the direct estimates (with absolute deviation, for example), especially for unemployment. Table A4 in the appendix shows the comparison between the transformed and linear Fay-Herriot model. The transformed area-level model generates much more accurate estimates of uncertainty for unemployment and male LFP than the linear area-level model, as the linear model greatly underestimates mean squared error for these indicators, leading to very low coverage rates. However, the results on accuracy are mixed. For example, when examining correlations, the arcsine estimates are more accurate for male LFP but less accurate than the linear model for female LFP. The results suggest that the linear F-H is preferred to the transformed F-H model in the case of female LFP, when the prediction model is better. However, when considering male and female LFP, the AGEB-level model generates more accurate estimates than both area-level model according to all criteria.

Table 9 – Comparison of direct, AGEB-level, and municipal-level (FH)

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
Coverage rate				
Direct	0.971	0.853	0.409	0.215
AGEB-level model	0.988	0.990	0.952	0.976
Transformed area-level model	0.887	0.823	0.804	0.842
Estimated Relative Bias (median)				
Direct	0.039	-1.017	-4.363	26.429
	(0.306)	(-1.288)	(-59.069)	(-100.000)
AGEB-level model	-0.507	1.423	30.732	99.948
	(-0.622)	(-0.726)	(9.651)	(40.522)
Transformed Area-level model	1.382	4.058	-37.108	-53.035
	(0.717)	(-1.784)	(-51.308)	(-76.510)
Estimated mean absolute error (median)				
Direct	0.061	0.071	0.022	0.017
	(0.046)	(0.054)	(0.015)	(0.009)
AGEB-level model	0.028	0.040	0.009	0.006
	(0.022)	(0.028)	(0.006)	(0.005)
Transformed Area-level model	0.034	0.053	0.015	0.008
	(0.025)	(0.040)	(0.010)	(0.006)
Estimated rank correlation				

Direct	0.471	0.623	0.363	0.324
AGEB-level model	0.541	0.750	0.497	0.264
Transformed Area-level model	0.444	0.652	0.294	0.373
Estimated Pearson correlation				
Direct	0.493	0.659	0.418	0.385
AGEB-level model	0.553	0.770	0.415	0.614
Transformed Area-level model	0.405	0.619	0.209	0.384

Table A4 in the appendix also presents a comparison of an area-level model using the transformed target variable and an area-level model using a linear (untransformed) model. The transformed model generally performs better than the linear model, with the possible exception of some female LFP statistics (coverage rates, absolute deviation, and correlations). However, the AGEB-level model remains the best choice in this context.

5. Conclusion

This paper considers the extent to which combining simulated sample data with publicly available geospatial data improves state and municipal estimates of male and female labor force participation and unemployment rates. Results are compared against the full 2020 census. The small area estimation procedure greatly improves the accuracy and precision of state level estimates for all four indicators, as well as municipal estimates of male and female labor force participation. This method can therefore be used to obtain more granular information on municipalities with low LFP, which can lead to a better understanding of the causes of low female LFP and potential policies to address it.

The method does not work nearly as well for estimating urban unemployment, because the values for unemployment rates are very low and zero for many clusters in the sample. In this setting, linear mixed models do not perform well. Although rank correlations and mean absolute deviations both improve, the simple correlation with the census value falls and mean relative bias is very high. In all cases except for one, estimates from a model specified at the AGEB level generate more accurate predictions than those obtained from an area-level model, with the lone exception being when considering the rank correlation for estimated female unemployment rates. Not surprisingly, using a hypothetical enumeration of all households in selected clusters dramatically improves the accuracy of the estimates. But in this case the small area estimates offer very minor improvements in accuracy and precision over direct estimates from the fully enumerated sample.

Non-linear models such as two-part models may be an appealing alternative for small area estimation of low-probability events in a sample. In the two-part model, the first part models the probability of a positive unemployment rate and the second part models the unemployment rate conditional on a positive rate. This is feasible to incorporate in a small area estimation framework, but significant effort would be required to implement it in existing software.

Another important area for further work is to continue to experiment with additional forms of geographically comprehensive auxiliary data to better predict labor market outcomes. This could include, for example, information on building footprints or the presence of different types of businesses. New types of geospatial data are being released each year that can be incorporated into this type of approach.

There is room for further work on model diagnostics. The results of this study demonstrate that R^2 can be a misleading metric for assessing the usefulness of the model, since even models with relatively low R^2s can contribute important information to reduce bias in the predictions (Marhuenda et al., 2017). At the same time, model R^2s are much higher in the full enumeration sample, but the small area estimates contribute little additional information. An important research agenda for further work is to better understand which diagnostic indicators, in cases where census data is not available, can provide a rough assessment of the gains from incorporating geospatial data.

Another open question is how optimal sample design changes in the presence of free, predictive geospatial data that can be linked to surveys. The evidence presented above indicates that estimates for sampled municipalities are substantially more accurate than non-sampled municipalities, suggesting that surveys should try to cover all target areas if possible. It is not clear, however, how the benefit of small area estimation relates to the sampling structure, and what this implies for optimally structuring household surveys.

Importantly, it remains to be seen how well the results here will generalize to other contexts. As such, it will be important to conduct studies to evaluate the methodology in other settings before applying it generally. This is easier said than done, however. In order to validate the methodology, it is best to use a census – or an unusually large survey –as a measure of ground truth. Furthermore, either the survey or ground truth must have either geolocated enumeration areas or, as in this case, highly disaggregated geographic identifiers that can be matched with a shapefile.

A final important issue pertains to the appropriate weighting strategy for the conditional random effect models relied upon here and in other small area estimation applications. We make three points regarding weights. The first relates to the choice of whether to give equal weight to each municipality or to weight by population when evaluating the estimates. We consider only the former, as is common in the literature on small area estimation. However, in some policy contexts, there is a strong argument that it is more important to generate accurate estimates for more populous areas than less populous areas. This is a decision best made based on political rather than technical grounds.

Second, properly weighing the estimates of a sub-area model such as this one is not trivial. In this case, the weights for each AGEB in the sample were normalized by dividing each weight by its municipal average. This ensured that each municipality was given weight proportional to its sample size, while appropriately giving more weight to more populous AGEBs within municipalities. The latter is important because population weights are used to aggregate AGEB-level estimates to municipalities. Furthermore, in most sub-area models, it would be beneficial to adjust the weights further to correct for heteroscedasticity, to the extent that the number of

observations used to construct the sub-area average used as the dependent variable varies across observations.

Finally, there are different methods for accounting for weights in conditional random effect models. We used a particular method implemented by Pinheiro et al. (2021), although other methods have also been proposed and implemented.[26] Establishing the pros and cons of different approaches to incorporating weights in small area estimation models is an important topic for further research.

[26] See, for example, Skarke and Kreutzmann (2021), who implement the correction for informative sampling proposed by Guadaramma et al.(2018).

References

Arora, V., Lahiri, P., & Mukherjee, K. (1997). Empirical Bayes estimation of finite population means from complex surveys. *Journal of the American Statistical Association*, *92*(440), 1555-1562.

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.

Belotti, F., Deb, P., Manning, W. G., & Norton, E. C. (2015). twopm: Two-part models. *The Stata Journal*, *15*(1), 3-20.

Butar, F. B. (1997). *Empirical Bayes methods in survey sampling*. The University of Nebraska-Lincoln.

Butar, F. B., & Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, *112*(1-2), 63-76.

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

Chambers, R., Salvati, N., & Tzavidis, N. (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 453-479.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119-127.

Erciulescu, A. L., Cruze, N. B., & Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(1), 283-303.

Engstrom, R., Hersh, J., Newhouse, D., Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being, *The World Bank Economic Review*, forthcoming;, lhab015, https://doi.org/10.1093/wber/lhab015

Esteban, M. D., Lombardía, M. J., López-Vizcaíno, E., Morales, D., & Pérez, A. (2020). Small area estimation of proportions under area-level compositional mixed models. *TEST*, *29*(3), 793-818.

Ghosh, M., & Lahiri, P. (1992). A hierarchical Bayes approach to small area estimation with auxiliary information. In *Bayesian Analysis in Statistics and Econometrics* (pp. 107-125). Springer, New York, NY.

Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., ... & Zhou, Y. (2020). Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. *Remote Sensing of Environment*, *236*, 111510.

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.

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.

Jiang, J., & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, *53*(2), 217-243.

Inchauste,M. G; Tavares,P; Reteguis, N; Moreno Herrera,L; Arceo-Gómez, E; Ríos Cázares, A; Santillán, A; Cadena,Kiyomi E.; Iacovone,L; Saucedo Carranza,C; Anderson,M. Mexico - Gender Assessment (English). Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/377311556867098027/Mexico-Gender-Assessment"

Klasen, S. (2019). What explains uneven female labor force participation levels and trends in developing countries?. *The World Bank Research Observer*, *34*(2), 161-197.

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*.

Li, Y., & Lahiri, P. (2019). A simple adaptation of variable selection software for regression models to select variables in nested error regression models. *Sankhya B*, *81*(2), 302-317.

Lobell, D. B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T., & Murray, S. (2020). Eyes in the Sky, Boots on the Ground: Assessing Satellite-and Ground-Based Approaches to Crop Yield Measurement and Analysis. *American Journal of Agricultural Economics*, *102*(1), 202-219.

Lobell, D. B., Di Tommaso, S., Burke, M., & Kilic, T. (2021). Twice Is Nice: The Benefits of Two Ground Measures for Evaluating the Accuracy of Satellite-Based Sustainability Estimates. *Remote Sensing*, *13*(16), 3160.

López-Vizcaíno, E., Lombardía, M. J., & Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical modelling*, *13*(2), 153-178.

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. N. K. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 1111-1136.

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2020). Small area estimation of non-monetary poverty with geospatial data.

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).

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. N. K. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 1111-1136.

Molina, I., Saei, A., & José Lombardía, M. (2007). Small area estimates of labour force participation under a multinomial logit mixed model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(4), 975-1000.

Molina, I. and Y Marhuenda (2015). R package SAE: Methodology. Available at: https://cran.r-project.org/web/packages/sae/vignettes/sae_methodology.pdf

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team. "nlme: Linear and nonlinear mixed effects models. R package version 3.1-145." (2020).

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.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(4), 805-827.

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.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, *16*(1), 155-160.

Ugarte, M. D., Goicoa, T., Militino, A. F., & Sagaseta-López, M. (2009). Estimating unemployment in very small areas. *SORT-Statistics and Operations Research Transactions*, 49-70.

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., & Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, *115*(14), 3529-3537.

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.

Zhang, Y., Li, R., & Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*(489), 312-323.

Appendix

Table A1 - Post-lasso regression coefficients for simple random sample models

	(1) Female LFP b/se	(2) Male LFP b/se	(3) Female Unemp b/se	(4) Male Unemp b/se
Baja California	0.1064**	-0.0115	-0.0111	-0.0144
Baja California Sur	0.0495**	-0.0263	0.0025	0.0019
Campeche	0.0072	0.0225	0.0040	-0.0065
Coahuila de Zaragoza	-0.0318	-0.0515**	0.0031	0.0242*
Colima	0.0654**	0.0122	-0.0088	0.0012
Chiapas	0.0248	0.0239	0.0478**	0.0203
Chihuahua	0.0157	-0.0363**	-0.0057	0.0096
Ciudad de México	0.0417	0.0172	-0.0008	-0.0203
Durango	-0.0180	-0.0330**	-0.0088	0.0380**
Guanajuato	0.0144	-0.0024	-0.0048	0.0104
Guerrero	0.0854**	0.0092	0.0006	0.0108
Hidalgo	0.0439**	-0.0085	-0.0020	-0.0099
Jalisco	0.0517**	0.0222	-0.0227**	-0.0266*
México	-0.0075	-0.0008	-0.0015	0.0060
Michoacán de Ocampo	0.0518**	0.0336*	-0.0080	0.0030
Morelos	0.0819**	0.0141	0.0277*	0.0156
Nayarit	0.0994**	0.0113	-0.0115	-0.0163
Nuevo León	-0.0037	-0.0123	-0.0073	-0.0003
Oaxaca	0.0917**	-0.0013	-0.0108	0.0020
Puebla	-0.0005	0.0289*	-0.0031	0.0054

Querétaro	0.0596**	0.0483**	-0.0129	0.0058
Quintana Roo	0.0659**	0.0464*	0.0055	-0.0115
San Luis Potosí	0.0318*	-0.0109	-0.0032	0.0078
Sinaloa	0.0385	-0.0368*	-0.0077	-0.0162
Sonora	0.0463*	-0.0466**	0.0012	0.0335*
Tabasco	0.0014	0.0062	0.0153	0.0425*
Tamaulipas	0.0129	-0.0045	0.0089	0.0187
Tlaxcala	0.0222	0.0312*	0.0174	0.0450**
Veracruz	0.0350*	0.0031	0.0070	0.0261*
Yucatán	0.0219	0.0233	-0.0163	-0.0171
Zacatecas	-0.0235	-0.0499**	-0.0104	0.0241
Share of AGEB classified as cropland	-0.0003			
Share of AGEB classified as urban	-0.0004**			
Mean NDVI	-0.2772**			
Median NO2	57.5187			
Number of untagged roads	0.0012			
Number of platforms/bus stops	0.0001			
Number of residential roads	0.0064*			
Number of secondary roads	0.0030	-0.0067**		
Number of service roads	-0.0009			
Number of unclassified roads	-0.0025		-0.0004	
Length of untagged highways	0.0001			
Length of footways	0.0008	-0.0031**		-0.0005

Length of living streets	0.0011			
Length of paths/trails	-0.0007	0.0004		
Length of primary highways	0.0008			-0.0003
Length of primary link highways	0.0003			
Length of tertiary link roads	0.0027			
Number of mini roundabouts	0.0122			
Number of emergency escape ramps in municipality	-0.0001			
Number of roads in municipality	0.0000			
Number of trunk roads in municipality	0.0000			
Total length of crossings in municipality	0.0009**			
Total length of motorways in municipality	-0.0000**			
Number of untagged highway points in municipality	0.0016	0.0046		
Number of crossings in municipality	-0.0017			
Number of give way signs in municipality	-0.0257*			
Number of residential highways in municipality	0.0025			
Number of townhalls in municipality	-0.0032	-0.0037		
Number of public marketplaces in municipality	-0.0022			
Number of places of worship in municipality	0.0067			
Number of fountains in municipality	0.0028			
Number of miscellaneous companies in municipality	-0.0030			
Mean nighttime lights in municipality	-0.1097	-0.0013		
Median nighttime lights in municipality	0.1446*		0.0031	

Mean year of switch to impervious surface in municipality	-0.0024**			
Share of bare land cover in municipality	0.0009	0.0109*		
Share of tree land in municipality	0.0019			
Standard deviation of vegetation index in municipality	0.4751**			
Earliest year of switch to impervious surface in municipality	0.0018	0.0070**		
Latest year of switch to impervious surface in municipality	0.0094**	0.0078**		0.0010
Log of geographic area of municipality	0.0121**			
Number of steps in municipality		0.0019**		
Number of schools in municipality		0.0014**		
Number of bus stations in municipality		-0.0395		
Median year of switch to impervious surface in municipality		-0.0029	-0.0020	
Share of shrub land in municipality		-0.0006		
Minimum nighttime lights in municipality		-0.0006		
Number of secondary link roads		-0.0080**		
Number of trunk roads		0.0000		
Share of urban land in municipality		-0.0006		
Std Dev of No2 levels in municipality		-0.0100*		
Intercept		-0.0013		
Number of observations	7,733	7,733	7,727	7,733
R-squared	0.1258	0.0623	0.0189	0.0220

Dependent variable is transformed rate, transformed using arcsine transformation. * 0.05 ** 0.01

Table A2 – Shapley decompositions of R^2: post-lasso regression results

	Female LFP	Male LFP	Female Unemp.	Male Unemp.
State dummies	25.3%	27.1%	62.9%	73.4%
Area of municipality	11.5%		0.5%	
Land Cover	4.6%	27.3%	7.7%	1.2%
Vegetation Index	10.2%		0.7%	2.3%
Night-time lights	11.3%	1.6%	14.0%	14.9%
Pollution				2.0%
Year of switching to impervious surface	9.0%	7.7%	2.5%	2.4%
Population	0.9%	6.2%	1.4%	
Highway counts	5.7%	3.5%	3.9%	2.7%
Highway lengths	4.2%	7.6%	3.9%	1.6%
Points of interest counts	9.5%	13.7%	0.4%	
Amenity counts	7.9%	5.4%	2.2%	

Figure A1 - Quantile-quantile plots of residuals using simple random sample

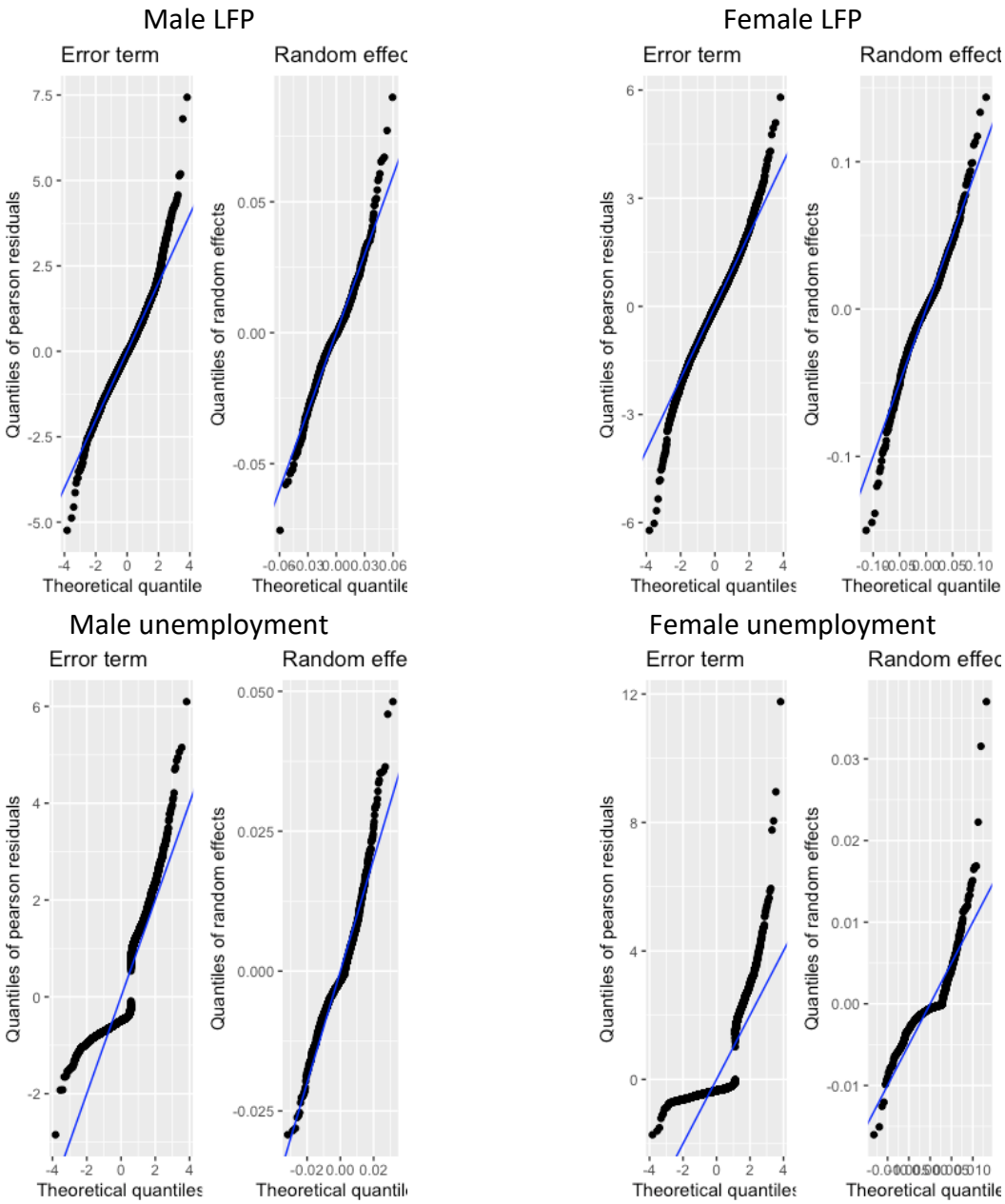

Figure A2 - Quantile-quantile plots of residuals using full enumeration sample

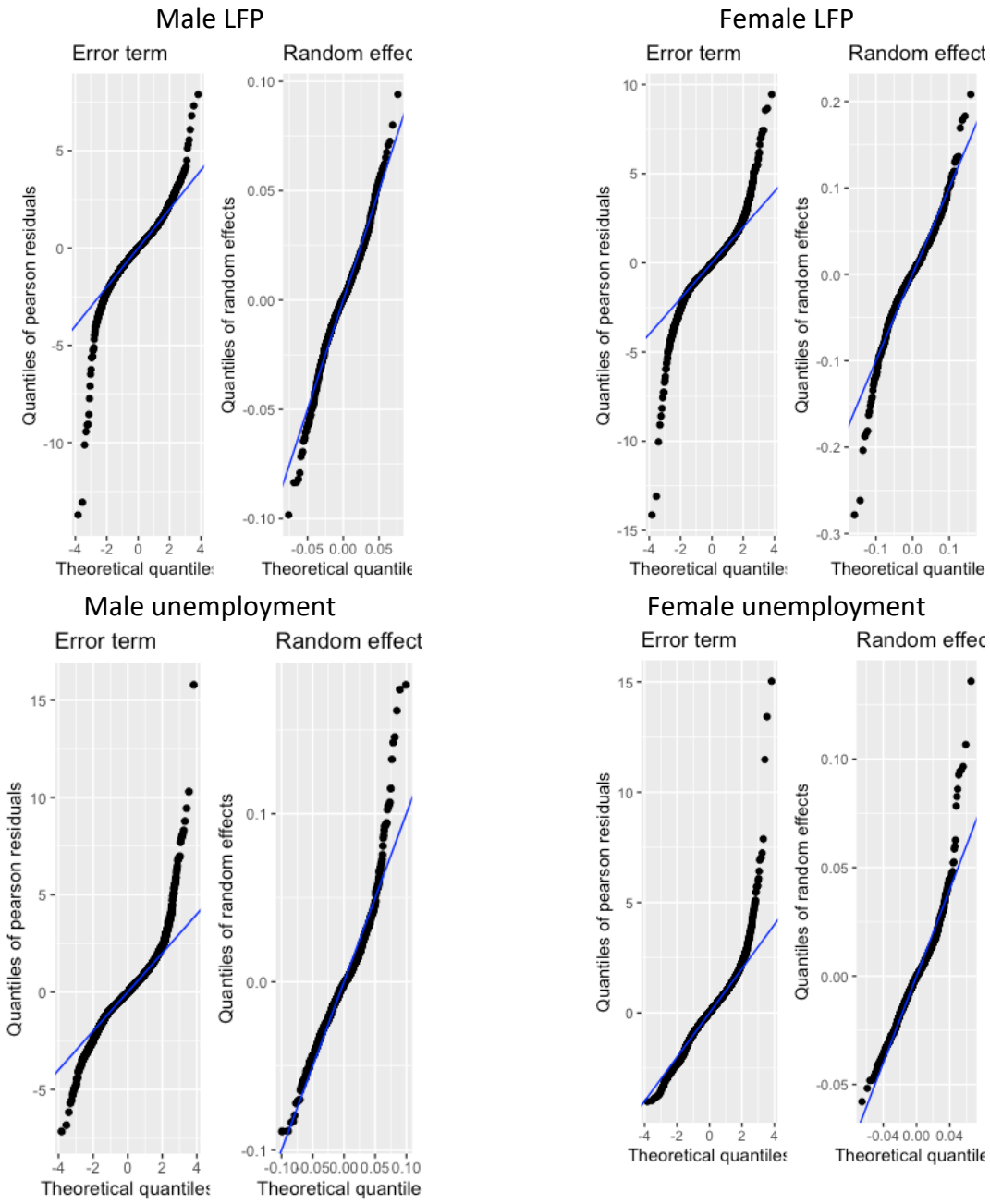

Table A3 – Aggregating municipality point estimates to state level

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
Estimated mean absolute error				
Direct	0.0179	0.0193	0.0048	0.0032
SAE – State level	0.0088	0.0073	0.0023	0.0024
SAE – Muni aggregation	0.0091	0.0117	0.0023	0.0023
Median estimated relative bias				
Direct	2.1164	-3.0573	-10.7269	-12.3980
SAE – State level	-0.0757	-0.4322	1.5223	15.1544
SAE – Muni aggregation	-0.1279	-1.9771	-0.1216	14.1461
Estimated rank correlation				
Direct	0.5652	0.7933	0.7911	0.6635
SAE – State level	0.7185	0.9498	0.8867	0.7529
SAE – Muni aggregation	0.7243	0.9380	0.8904	0.7775
Estimated Pearson correlation				
Direct	0.6048	0.8141	0.7528	0.6506
SAE – State level	0.7614	0.9583	0.8698	0.5467
SAE – Muni aggregation	0.7441	0.9506	0.8884	0.5652
States	32	32	32	32

Table A4 – Fay-Herriot with arcsine transformation

	(1) Male LFP	(2) Female LFP	(3) Male Unemp	(4) Female Unemp
Coverage rates				
F-H linear model	0.396	0.909	0.001	0.174
F-H arcsine transformation	0.887	0.823	0.804	0.842
Median relative bias				
F-H linear model	1.435	-6.127	-100.000	-100.000
F-H arcsine transformation	0.717	-1.784	-51.308	-76.510
Median absolute deviation				
F-H linear model	0.025	0.039	0.021	0.008
F-H arcsine transformation	0.025	0.040	0.010	0.006
Estimated rank correlation				
F-H linear model	0.319	0.723	0.193	0.356
F-H arcsine transformation	0.444	0.652	0.294	0.373
Estimated Pearson correlation				
F-H linear model	0.343	0.708	-0.012	0.155
F-H arcsine transformation	0.405	0.619	0.209	0.384