Exploiting Data to its Fullest

Venture Fund Project

(LABO.10.35)

Angelo Cozzubo

Carolina Franco

Zachary Scheffler

Anna Solovyeva

Abstract. This document is the write-up report of the Venture Fund project Exploiting data to its Fullest, selected by the 2022 Venture Fund Call for Proposals. The project was developed between March and July 2022, and its objective is to combine Machine Learning and Small Area Estimation techniques for addressing the covariate selection task. This document is accompanied by an online repository with the replication material, which includes the R code and the anonymized datasets.

June 2022

1 Introduction

Why small area estimation?

Small area estimation (SAE) is a series of techniques to improve traditional survey estimates for various domains by borrowing strength from other sources or exploiting relationships between the domains of interest. It is a well-established sub-discipline within statistics with a vast literature (see, for instance, Rao and Molina 2015). Yet, it continues to grow and evolve as the need for producing statistics at low levels of aggregation increases.

SAE already has a significant role in survey statistics as the US and other nations use it to produce critical official statistics. For instance, in the US, SAE techniques have been used to allocate federal resources for children in poverty[1] and to enforce the Voting Rights Act, Section 203, which dictates some jurisdictions are required to print voting materials in more than one language[2]. These methods have also been employed to estimate health insurance coverage across geographies and sociodemographic groups[3] and to produce monthly unemployment estimates at the sub-national level[4]. Other countries such as Peru, the UK, Chile, the Netherlands, etc., are also using SAE to generate official local statistics.

In fact, the United Nations has recently launched a website to encourage developing countries to adopt SAE to help achieve Sustainable Development Goals (SDGs), such as ending poverty and hunger. "The SDG indicator framework has included an overarching principle of data disaggregation: SDG indicators should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability, and geographic location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics [...]. As sound statistical methods are vital to

[1] https://www.census.gov/programs-surveys/saipe.html

[2] https://www2.census.gov/programs-surveys/decennial/rdo/about/voting-rights-determination/2021_Section203/Sec203_ExecSummary2021_v3.pdf

[3] https://www.census.gov/programs-surveys/sahie.html

[4] https://www.bls.gov/lau/

overcome this challenge, Small Area Estimation (SAE) constitutes an important topic in the way forward"[5].

Why is SAE growing in importance? When properly utilized and when good auxiliary information is identified, small area estimation can dramatically increase measures of uncertainty relative to direct survey estimates. This often implies that more small areas (domains) pass the quality publication thresholds imposed by national statistical offices, making it possible to publish statistics that would otherwise be suppressed, and to produce statistics at a lower level of aggregation than would be possible with survey direct estimation methods alone.

Starting a new SAE initiative can be challenging, and it is recommended that an SAE team include both a technical expert in SAE and, if possible, a subject matter expert on the topics of interest. That being said, the potential of using SAE techniques to obtain improved estimates from survey data is substantial. For an introduction to small area estimation, see Erciulescu et al. (2021) or Rao and Molina (2015).

Typically, we seek to use small area estimation techniques when we have a survey that measures the quantity of interest approximately unbiasedly but the sample size is not sufficient to estimate all the quantities of interest using design-based survey estimates with the desired level of accuracy. For instance, we may have a survey designed only for national estimates, but we would like to produce information at the state level or for specific demographic groups. The idea of SAE then is to use models to "borrow strength."

Borrowing strength is typically done by using covariates from administrative records, censuses, commercial sources, etc. One can also borrow strength via spatial models that exploit known geographic proximity metrics across areas or via temporal models that exploit the repeated nature of a survey. Alternatively, one can combine estimates from various surveys. Each of these possibilities has nuances, so it is good to consult with a technical expert when embarking on an SAE problem. Our focus here will be on using covariates from administrative records and censuses.

[5] See https://unstats.un.org/wiki/display/SAE4SDG

Additionally, we will explore the adjacency relationships among areas using spatial models to borrow strength from neighbors.

Why Machine Learning?

Machine learning (ML) is also a rapidly evolving field that has experienced a boom in the past several years as computational capacities increase. ML emphasizes prediction fueled by computational science. There has been some research on combining the ideas behind machine learning and small area estimation, but the research so far is limited (e.g., Ren et al. 2020, Brioniecki et al. 2022). Because small area estimation involves prediction, we explored ways in which these two disciplines could be married to provide potential value to NORC.

Our application and data

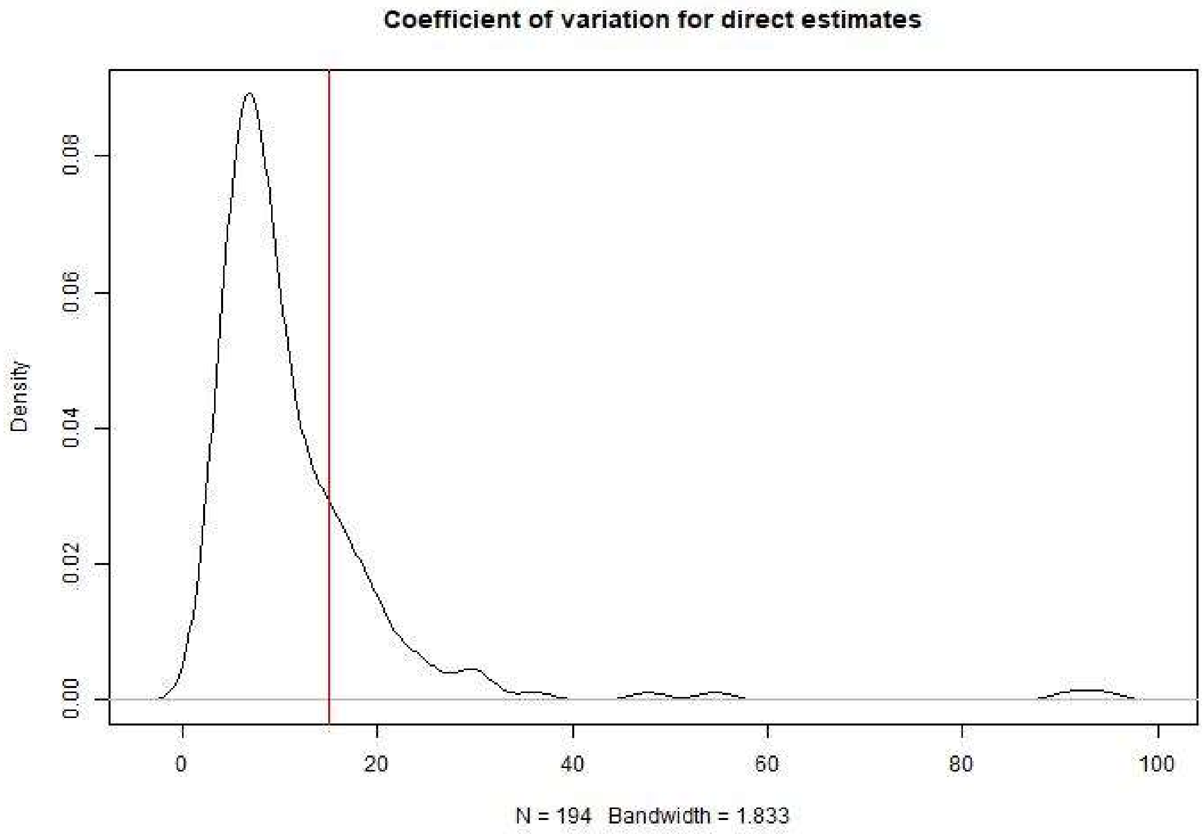
Our application is to estimate anemia for children under five years in Peru. The main focus is to obtain disaggregated estimates at the second subnational level (provinces) where official survey data has not been published yet. Anemia is a pressing problem in Peru, but the lack of local estimates measuring its prevalence makes it challenging to plan local interventions. The National Statistical Office only publishes yearly estimates at the regional level due to reliability standards. There are 25 regions in Peru and 196 provinces embedded within these regions.

The primary survey used to measure anemia is called the Demographic and Health Survey (Endes, for Encuesta Demográfica y de Salud Familiar in Spanish). For our analysis, we pooled data collected by Endes for 2017-2019.

We began our analysis by computing the direct survey estimates from Endes. Specifically, we calculated Horvitz-Thompson estimates using the R package survey (Lumley, 2021). The National Statistical Office in Peru has as a publication standard requirement that estimates only be published if their coefficient of variation (CV) is less than 15%. When we computed the direct survey-weighted estimates at the province level, we found that 42 out of the 194 estimates do not meet that threshold.

Hence, 42 provinces will not have official information about the anemia prevalence. Figure 1 illustrates the distribution of CVs across provinces, with a vertical line at the abovementioned threshold, showing that a substantial number of provinces do not meet it. This motivates the need for small area estimation techniques to improve upon the estimates.

Figure 1 - Kernel density, coefficient of variations direct estimators



Note: The red line indicates the suppression threshold used by the Peruvian National Statistical Office (15%).

Our goals

We had three main goals in this strategic initiative. The first one is to show the benefits that can be attained using small area estimation techniques to improve inference for anemia prevalence among children 0-5 in Peru. We had the support of the Peruvian Statistical Office and the potential for eventually creating a small area estimation program that can be used for official statistics, which would create positive publicity for NORC.

A second goal is to explore which variable selection methodology would be best for small area estimation problems where many potential covariates are available. Unlike other small area problems, where relevant auxiliary information is hard to come by, here we had a wealth of covariates to explore. These came from the census and various administrative records. In fact, we had 500 covariates to choose from. Hence, we wanted to examine different ways to perform variable selection to identify the most promising covariates. We needed a quick and efficient method to handle many covariates with a short running time. Hence, we drew various techniques from the Machine Learning literature, as well as more traditional ways to select covariates such as expert opinions or stepwise selection. This type of situation, where many alternative covariates are available, is becoming more common in this age of information proliferation.

Our third goal was to test a particular Machine Learning Technique and package that had recently appeared in the literature called Auto MrP. This method combines a modeling technique called multilevel regression with post-stratification with several machine learning techniques via Bayesian model averaging. Because it turned out that this methodology is not very promising, we discuss the results on this in the Appendix.

2 Basic SAE models utilized

We focus primarily on using the Fay-Herriot model (1979) and a spatial extension. For a more complete treatment of the anemia application in Peru, we recommend that other model forms be considered. For instance, one may consider the Binomial Logit Normal Model used in Franco and Bell (2013, 2015, 2022), which naturally handles zeros and potential skewness. In our application, the number of zeros is relatively few, and the proportions are not extreme, so the Fay-Herriot model's normality assumptions are somewhat reasonable. To meet our goals within the given budget constraints, we needed a model that was easy to apply and had a very fast running time. The Fay-

Herriot model, and its spatial extensions, can be implemented via the sae package in R (Molina and Marhuenda, 2020).

The Fay Herriot (FH) model is one of the most widely used small area estimation models, perhaps due to its simplicity. It can be expressed as

$$\widehat{Y_i} = \theta_i + e_i$$

$$\theta_i = x_i'\beta + u_i$$

Where $\widehat{Y_i}$ is the vector of direct estimates for the provinces, e_i is the vector of sampling errors, x_i' is the matrix of explanatory variables, and u_i is the vector of area random effects independent of e_i. The first level of the model describes the uncertainty due to sampling, given that we do not observe the area level quantity but a noisy estimate of it. The quantity e_i is the direct estimator's sampling variance, usually assumed for identifiability. In practice, e_i needs to be estimated from the microdata. Smoothing might be advisable (see Franco and Bell, 2022), though we do not pursue it here.

The second level, often called the linking model, explains the relationship between the underlying population quantity of interest and the covariates used to attempt to describe it. The area random effect is often called the model error and attempts to capture what cannot be explained by the covariates. Typically, small area models can be fit using empirical Bayes, or Hierarchical Bayes approaches. The former is frequentist, and the latter is fully Bayesian. The sae package we use implements an Empirical Bayes approach.

The Fay-Herriot model, and other similar area-level models, have the property that the model predictions are very similar to the corresponding direct estimators for domains with large sample sizes. Hence, the covariates from auxiliary data play a more prominent role in areas with small sample sizes but do not substantially change the estimates for large domains. This property is highly desirable as areas with large sample sizes have very good direct estimators, so it is attractive for the model to produce similar estimates in such cases.

The spatial extension of the model (SFH) assumes a spatial auto-regressive (SAR) structure for the model error. We can express it by extending the random effects as following

$$u_i = \rho W u_i + \eta_i$$

where ρ denotes the autoregressive parameter, W is the standardized queen proximity matrix and $\eta_i \sim N(0_i, AI_i)$ for A unknown. This extension takes advantage of information about the proximity among different small areas. Our spatial analysis showed that the data are spatially correlated. We applied a Moran's I test and a Geary's C test to test this. Both tests were statistically significant, suggesting promise for the spatial approach. We also compared the SFH model with a traditional FH model in terms of AIC, and the SFH performed better than the FH model. Hence, for the rest of our analysis, we use the SFH model.

3 Variable selection in SAE

Variable selection in SAE usually involves several techniques. This may include comparisons based on AIC, BIC, DIC, or related likelihood-based methods. Other commonly used techniques are hypothesis testing (e.g., significance levels and p-values) and residual analysis. In addition, cross-validation techniques can be used.

In practice, experts are often consulted to help identify relevant covariates. One example is the implementation of small area poverty estimates for Chilean comunas. Similar to our study case here about Peru, in this case, the Chilean government also had a wealth of administrative data within the government system. The authors of this study decided to narrow down the list of potential predictors by the opinion of subject matter experts. The final list was selected using a combination of stepwise selection, residual analysis, and other model diagnostics available in Stata (Casas-Corderos, Encina, Lahiri, 2015). For a discussion on the characteristics of good covariates, see Erciulescu et al. (2021).

Combining variable selection techniques from machine learning with small area estimation models is beginning to receive some attention in the literature. For example, Ren et al. (2020) used the

LASSO plus an analysis of the correlation between the variables to pre-select covariates for an SAE program using the Program for the International Assessment of Adult Competencies (PIAAC). The techniques they used for variable selection ignore the random effects. However, the second phase of their analysis considers the sampling variance, where further refinement is done via cross-validation. It is well known that ignoring the random effects in variable selection does not lead to the "ideal" variable selection criterion (see Lahiri and Suntornchost, (2015)).

4 Preliminary analysis

We began our analysis with several summaries of the data. We decided to examine the Fay-Herriot model for our analysis because it is fast and easy to implement. It is perhaps the most frequently used and studied model in the small area estimation literature. Our preliminary analysis suggested that the data departs slightly from normality, so for future research, we recommend also examining other model forms such as the Binomial Logit Normal model (e.g., Franco and Bell, 2022). Nonetheless, the Fay-Herriot model yields design-consistent predictions and is somewhat robust to model misspecification, especially when the proportions are not in the extremes. One potential pitfall of using the FH model for proportions is obtaining estimates outside the [0,1] range. This issue did not occur in our application.

5 Variable selection techniques explored

LASSO—The main idea behind LASSO is to minimize an objective function while including a penalty term to avoid overfitting. The objective function is a distance metric between the predictions and observations. We used the R package glmmLasso to implement this (Groll, 2022) and included an area-level random effect as in the FH model. It was not possible to have a random variable representing the sampling variance because the software does not allow including an effect with known variance. Including the sampling variance as an unknown parameter would make the model unidentifiable. Given that the sampling variance was ignored in the procedure, we filtered out direct estimators that are highly variable (CV>15) as such observations are inaccurate and could unduly affect the fit. Furthermore, to avoid LASSO selecting too many covariates aiming in the same direction, we applied a correlation filter to avoid having multiple highly correlated parameters.

Sparce PCA—In our application, we have many potential covariates relative to the number of observations, which can lead to overfitting with LASSO. Hence, we also attempted to use Sparse Principal Component Analysis (PCA). The idea behind PCA is dimensionality reduction—that is, the majority of the information contained in the covariates can be captured by a smaller number of variables that are created as linear combinations of the original covariates. Sparse PCA attempts to overcome the limitations of PCA when the number of parameters is larger than the number of observations. In traditional PCA, each principal component is a linear combination of all the available covariates. For the algorithm's sparse extension, the optimization problem is reformulated by imposing a LASSO constraint on the regression coefficient. Hence, sparse PCA (sPCA) aims to find sparse weight vectors ("loadings"), making that the contributions of many of the variables are set to zero. In this sense, the sPCA makes a selection of relevant variables. We used the package *sparsepca* (Erichson et al.,2018) to implement it. Unfortunately, this technique does not include random effects.

Stepwise regression—This model selection method consists of several iterations where a variable is removed or added at each iteration, based on a criterium such as a p-value threshold, AIC, etc. We implemented this using the StepReg package in R (Li et al., 2022). As we had more covariates than observations, the backward selection method was not an option as the first step would involve fitting a model with more variables than rows. Instead, we used bidirectional stepwise regression, which is similar to forward selection but somewhat more flexible by allowing the removal of covariates as you move forward in the looping process.

Expert's opinion—We interviewed four experts on anemia in Peru and asked them to select the variables they think are most predictive. We included this as selection criteria, as this method is done in practice as an initial variable selection filter (e.g., Casas-Corderos, et al. (2015)). Our experts included Kathy Vilcarromero (Health Indicators Analyst, Peru Ministry of Economics and Finance); Oriana Salomon (Health Consultant at Videnza & Università degli Studi di Padova), Elard Amaya (University of San Andres and University of San Ignacio de Loyola), and Pedro Francke (Former Minister of Economy and Finance and Full Professor at PUCP). We fit a Spatial FH model using the variables selected by all the experts (intersection criteria).

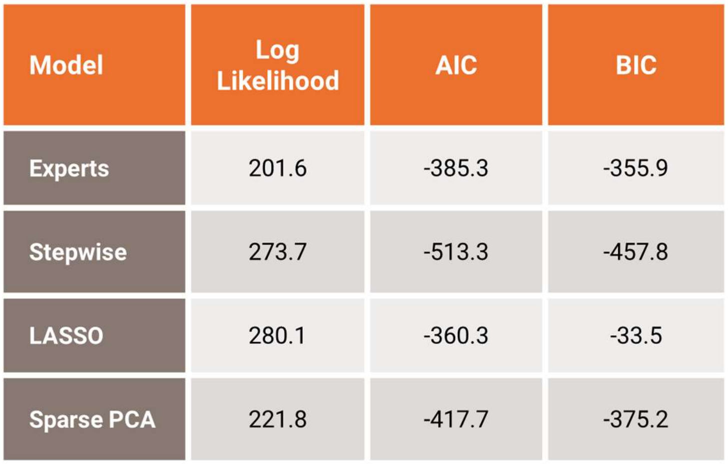
6 Results of variable selection

We conducted various analyses to shed some light on which variable selection methodology to use for the anemia application. First, we examined the model estimates and their mean squared errors (MSE) for each optimal model from the four variable selection techniques. We should note that the standard errors are not directly comparable among the models because they use a different number of parameters. Nonetheless, it is interesting to see if different selection methods yield different results, and the magnitude of the difference is of practical relevance. Overall, the estimates were quite similar for the different versions of the Spatial FH when using the four variable selection methods. However, the prediction MSEs differed quite a bit, as did the reductions in variance over the direct estimators. The median improvements were 35%, 28%, 24%, and 12% for the stepwise, SPCA, experts, and LASSO; respectively.

Furthermore, as we mentioned earlier, using the threshold of a coefficient of variation above (CV>15), 42 of the direct estimators were suppressed. Using the same threshold, the number of suppressed estimates from our models was reduced to 3, 5, 13, and 25 for the stepwise, SPCA, experts, and LASSO; respectively. Note that variances alone are not an indication that one model is better than the other, as this ignores the fact that these models have different numbers of parameters.

As other tools to compare the results, we used the Akaike Information Criteria (AIC) and the Bayesian information Criteria (BIC). These methods are based on the loglikelihood but include a penalization for a larger number of parameters. The results of the comparison are given in the table below. We see that the stepwise regression model has the lowest AIC, followed by sPCA, the expert, and the LASSO. The BIC also shows the same trends. The LASSO has the highest log likelihood, but this is not a good measure given that it has many parameters (it selected 97 covariates). The number of covariates for the expert's selection is 7; for stepwise is 14; while for sPCA, only the first ten components are included.

Table 1 - Goodness of fit measures

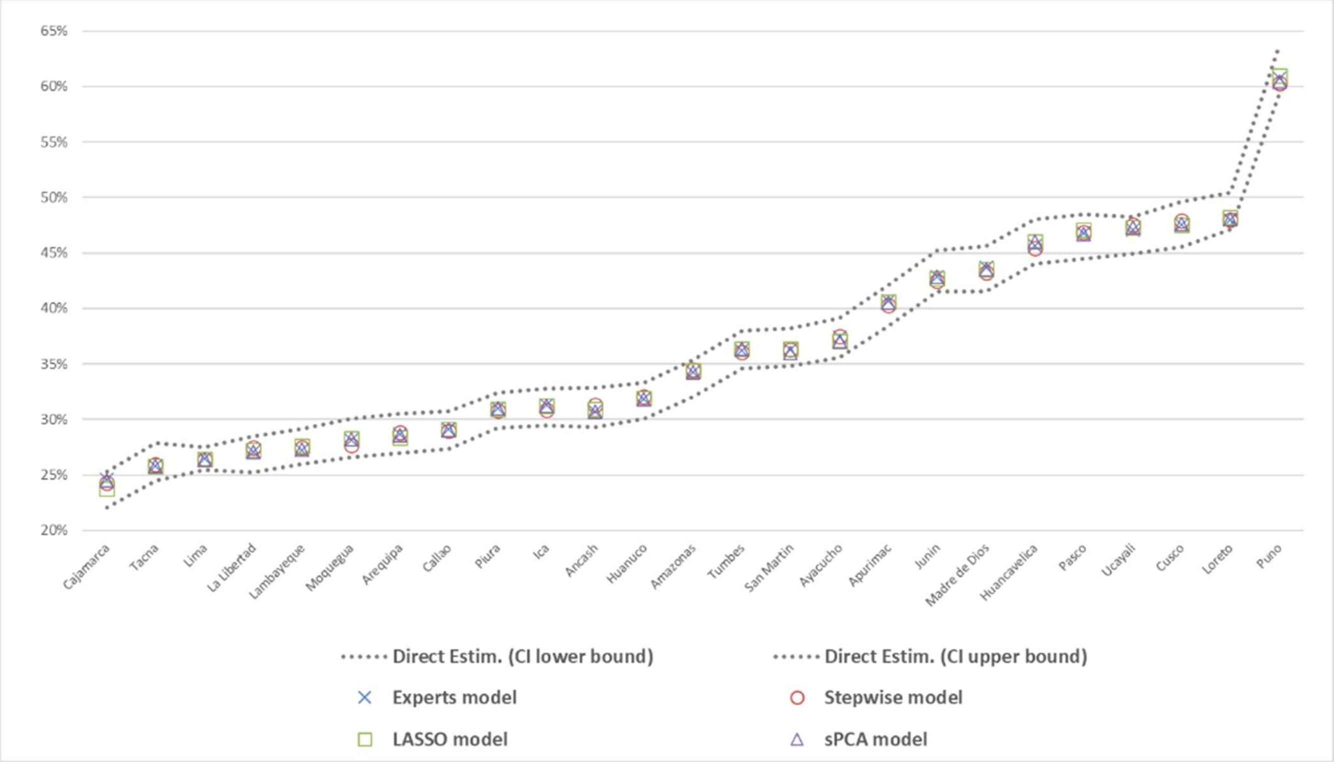


Model	Log Likelihood	AIC	BIC
Experts	201.6	-385.3	-355.9
Stepwise	273.7	-513.3	-457.8
LASSO	280.1	-360.3	-33.5
Sparse PCA	221.8	-417.7	-375.2

Note: AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

We aggregated our model estimates at the first subnational level (regions) as a third model testing. The idea behind this process is that the regional estimates have a large sample size and should be reasonably accurate. Moreover, the Endes survey is designed to be representative of the regions. Therefore, we hope the model estimators add up to a similar quantity. The figure below shows these aggregates for our four models. We see that the regional estimates for all models are pretty close to each other. This behavior is not surprising because, as we pointed out above, the model estimates are very similar among the four models. As we can see, all the estimates fall within the 95% confidence bounds of the direct estimators, which speaks in favor of the spatial FH results.

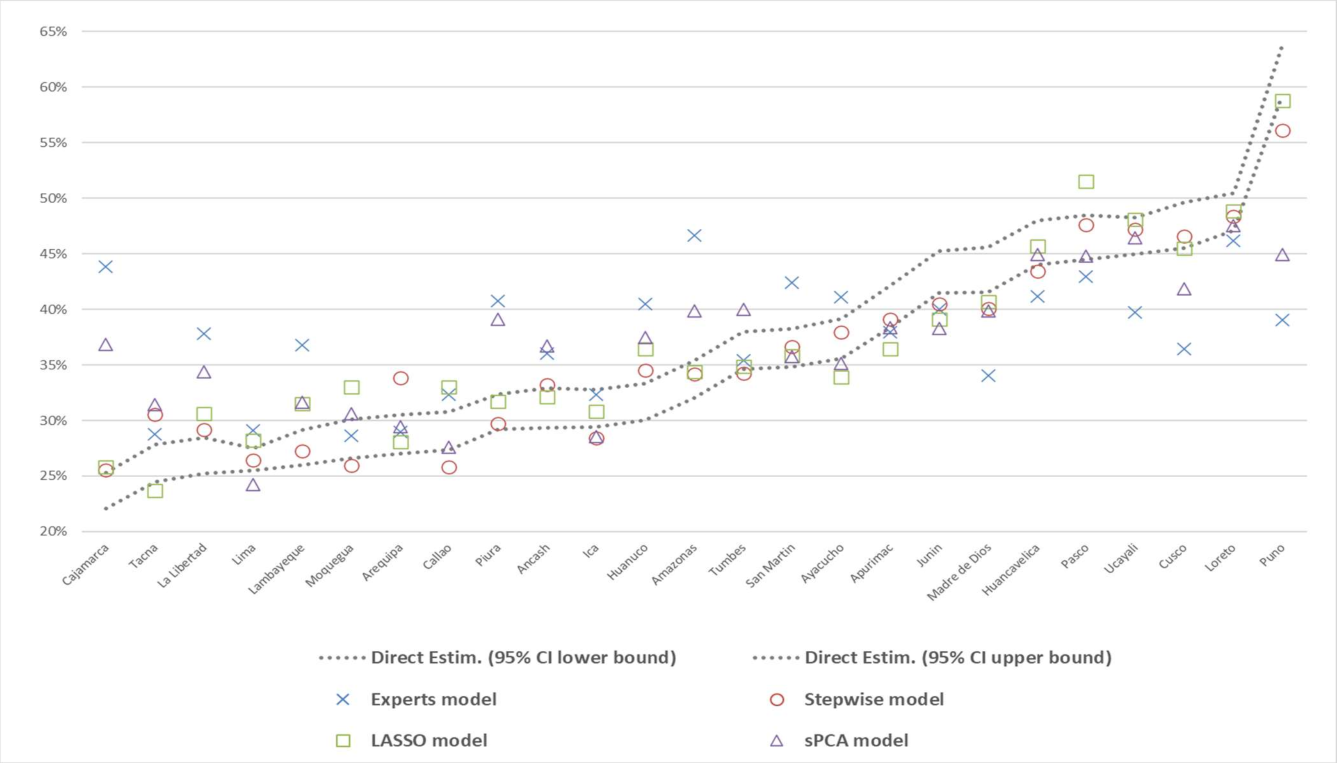
Figure 2 - Survey regional estimates vs model aggregate estimates



It is also of interest to aggregate the synthetic estimates—those that come directly from the regression function $(x_i{}'\hat{\beta})$ without putting any weight on the direct estimators. Again, ideally, these estimators would aggregate to something close to the direct estimators for larger levels of aggregation as the regions. The figure below shows that some of the models' aggregate estimates tend to fall inside the direct estimator's confidence intervals more consistently than others. The stepwise regression model tends to be within or close to the confidence bounds more frequently than the other methods.

Overall, the stepwise selection for this application seems to perform best based on the metrics we considered. Of course, many more metrics can be computed, but the results suggest that stepwise regression is a reliable variable selection method. Hence, Section 7 presents the results obtained from Stepwise Regression.

Figure 3 - Survey regional estimates vs model synthetic estimates

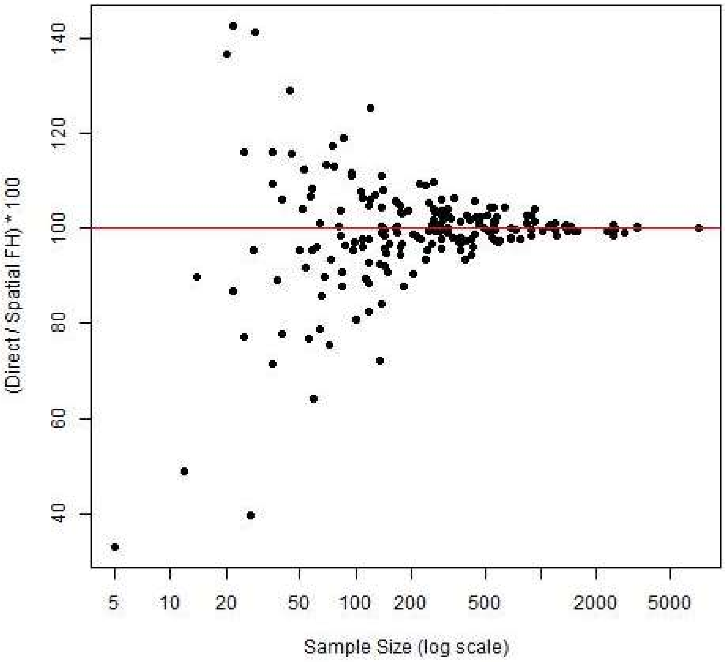


7 Benefits of SAE for the Anemia application

As previously mentioned, using stepwise regression for variable selection in the Spatial FH context seems to perform best under various metrics. Here, we present more information on the results of applying the spatial FH model to the Endes anemia data using a stepwise selection of covariates. Figure 4 below plots the direct and model estimates ratio against the sample size. We noticed that as the sample size increases, the model estimators converge to the direct estimators. This scenario is expected behavior for this model and, as discussed previously, a desirable property.

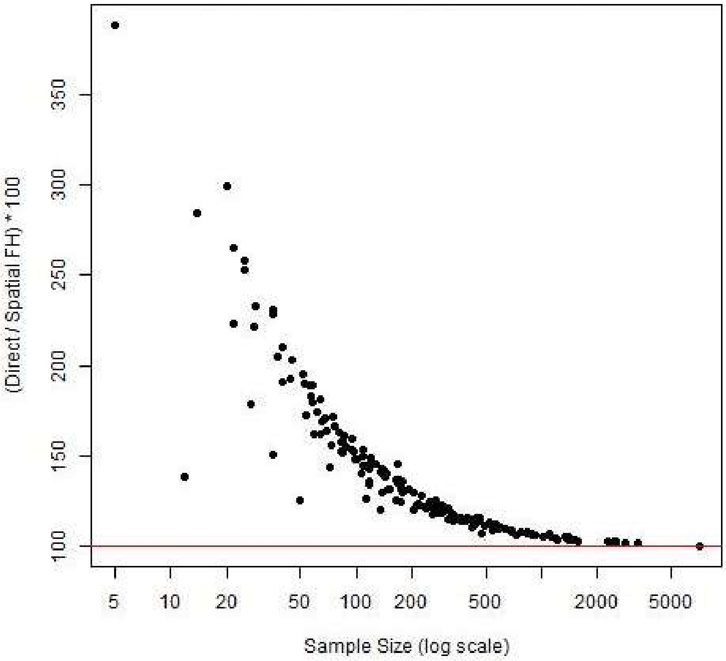
In Figure 5, we plot the ratio of standard errors between the direct and the model estimates. A ratio equal to one means the two standard errors are identical, and ratios greater than one implies that the direct estimators have higher standard errors than the model predictors. We observe that in all cases, the model estimators have lower standard errors than the direct. Once again, the difference between the standard errors diminishes with larger sample sizes, as the model predictors converge to the direct estimators.

Figure 4 - Ratio of estimates, Provincial Anemia Prevalence (stepwise model)



Note: FH = Fay-Herriot model. X-axis in logarithmic scale.

Figure 5 - Ratio of Standard Errors, Provincial Anemia Prevalence (stepwise model)



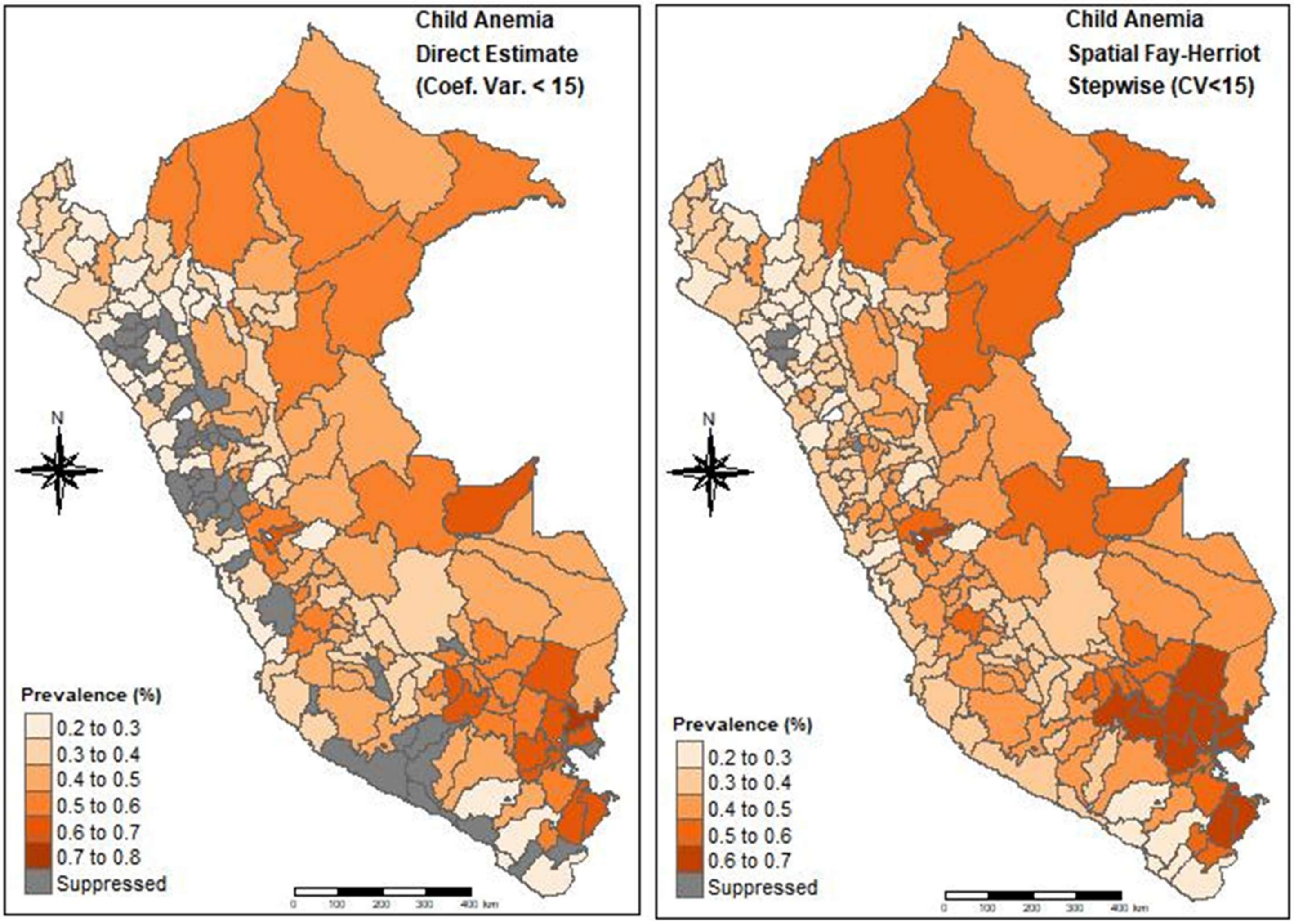
Note: FH = Fay-Herriot model. The X-axis is on logarithmic scale.

Anemia Maps for children 0-4 in Peru

First, we present a map of Peru using direct estimates. The grey areas represent provinces that had to be suppressed because of not passing the CV threshold. As mentioned before, 42 provinces do not meet the quality threshold. The map is a striking reminder that many provinces have their estimates suppressed, and no official information is available to policymakers.

Next, we create the same map using the Spatial Fay Herriot model estimates, using Stepwise Regression for variable selection. Using the SFH model, we have recovered 39 out of the 42 provinces where we had previously suppressed estimates. We estimate that 30,000 children with anemia live in these "recovered provinces," which targeting mechanisms would have ignored. Furthermore, for most province estimates, we have reduced the standard errors substantially.

Figure 6 - Anemia choropleths, survey and model estimates



Note: Right map presents the direct survey-weighted estimates. Left map presents the Spatial Fay-Herriot model estimates. In both maps, estimates with a coefficient of variation larger than 15 are suppressed.

8 Conclusions

Our methodology achieved substantial variance reductions for the estimates of proportions of children with anemia in provinces, which could have a tangible impact on intervention policies by allowing fewer children with anemia to be "left behind." In addition, the results yielded interesting insights into variable selection in small area estimation, showing that stepwise selection is a promising methodology, but also suggesting that sparce sPCA might yield good results when the number of potential covariates available is very large.

Future research on variable selection might explore ways to incorporate the effect of the sampling variance on the proposed methods and, in some cases, the random effect. Furthermore, a promising avenue of research relevant to the anemia application is the study of categorical and ordinal small area estimation models, which would allow capturing different severities of anemia. In addition, it would be interesting to study other SAE model forms, such as binomial logit or multinomial logit (in the case of many anemia categories), since this might be more appropriate for proportions than a normality-based model. Nonetheless, like other area models, the Spatial FH model has certain robustness in that it is design-consistent when the direct estimator used for the model is also design-consistent. Furthermore, as the anemia proportion for children within provinces is not close to zero or one, the normality assumptions are somewhat reasonable.

9 References

Broniecki, Philipp; Leemann, Lucas; Wüest, Reto (2022). Improved multilevel regression with post-stratification through machine Learning (autoMrP). The Journal of Politics, 84(1):597-601.

Casas-Cordero C., Encina, J. and Lahiri, P. (2015). Poverty mapping for the Chilean comunas. In Analysis of Poverty Data by Small Area Estimation (M. Pratesi, ed.). Wiley, New York.

Erciulescu, A., Franco, C., and Lahiri, P. (2021). Use of administrative records in small area estimation. Chun, A. Y. and Larsen, M. (Eds.) Administrative records for survey methodology. New York: Wiley.

Erichson, B. M., Zheng, P., Aravkin, S. (2018). Sparcepca: Sparse Principal Component Analysis (SPCA), version 0.1.2. https://cran.r-project.org/web/packages/sparsepca/sparsepca.pdf

Franco, C. and Bell, W. R. (2022). Using American Community Survey data to improve estimates from smaller US surveys through bivariate small area estimation models. Journal of Survey Statistics and Methodology, 10, 1, 225-247

Franco, C. and Bell, W. R.(2015). Borrowing information over time in binomial/logit normal models for small area estimation. Statistics in Transition (new series)and Survey Methodology, joint issue on Small Area Estimation, 16, 563--584, available at http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/volume-16-number-4-december-2015/.

Franco, C., and W. R. Bell (2013), "Applying Bivariate Binomial/Logit Normal Models to Small Area Estimation," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 690–702

Fay R.E. and Herriot R. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, 269–277.

Groll, A. (2022). glmmLasso:Variable Selection for Generalized Lienar Mixed Models by L1-Penalized Estimation, version 1.6.1. https://cran.r-project.org/web/packages/glmmLasso/index.html

Gelman, A. Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. Survey Methodology. 23 2127-136

Lahiri, P. and Suntornchost, J. (2015). Variable Selection for Linear Mixed Models with Applications in Small Area Estimation. Sankhya B. 77. 10.1007/s13571-015-0096-0.

Li, J., La, X, Cheng, K.., Liu, W. (2022) StepReg: Stepwise Regression Analysis. Version 1.4.3. https://cran.r-project.org/web//packages/StepReg/StepReg.pdf

Lumley, T. (2021) Survey: Analysis of Complex Survey Samples. Version 4.1.1 https://cran.r-project.org/web/packages/survey/survey.pdf

Molina and Marhuena (2020). Sae: Small area estimation. R package version 1.3. https://cran.r-project.org/web/packages/sae/sae.pdf

Pfeffermann, D. and Sverchkov, M. (2007). "Small-area estimation under informative probability sampling of areas and within the selected areas." Journal of the American Statistical Association, 102, 480, 1427–1439.

Rao, J. N. K., and Molina, I. Small-Area Estimation. Wiley, 2nd edition, 2015.

Ren, W., Li. J., Erciulescu, A., Krenzke, T., Mohadjer, L. (2020). A Variable Selection Method for Small Area Estimation Modeling of the Proficiency of Adult Competency. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 924-956

Appendix: Our investigations of auto MrP

A second line of research we pursued relates to a new methodology called Auto MrP which has been proposed in the political science literature. Auto MrP (Broniecki et al., 2022) is a machine-learning extension of Multilevel Regression with Post-Stratification MrP (Gelman and Little, 1997). Next, we will explain MrP, and later discuss the extension.

Multilevel Regression with Post-Stratification

The idea behind MrP is to divide the population into categories or ideal types. The categories are defined by a set of unit-level variables, such as age, education, gender, etc. Then, a unit-level model mixed effects model (logistic regression) is fit predict the probability interest for the ideal types (e.g., the probability of having anemia for a person in each ideal type of category, which is assumed to be the same for all individuals in each category). The models include an area-level random effect, possibly other random effects, and possibly fixed effects. These ideal-type predictions are then multiplied by the population size of the categories for each geography of interest from a census (Gelman and Little (1997) use PUMS data instead and ignore their variability).

Gelman and Little (1997) assume that conditional on the R explanatory variables, the non-response is ignorable within each category. There is also an assumption of non-informative sampling within each category. This means that the sample has the same distribution as the population with respect to the variables of interest. When inclusion in the sample is subject to selection bias, this condition may not hold. For instance, the design-variables may be correlated with the quantity of interest, or the quantities of interest themselves can be included as the design-variables.

To successfully apply MrP, the categories should be defined such that it is reasonable to assume Simple Random Sampling (SRS) within each category. This means that the categories should include all the information to construct survey weights, as well any other variables that might be informative about the response. If these conditions do not hold, the resulting estimates might be biased (see for instance Pfeffermann and Scherkov (2007), or Rao and Molina (2015)).

Auto MrP

Auto MrP uses ensemble Bayesian model averaging to combine several models, most of which rely on Machine Learning. Specifically the classifiers are (i) MrP with best subset selection of covariates (II) MrP with best subsect selection of Principal Components (iii) MrP with LASSO (iv) gradient boosting (v) support vector machine. The authors combine the predictions from individual classifiers using a super-learner. Each classifier is assigned a weight that determines its contribution to the final predictions based on a training set that is new (e.g., different than that used for the classifiers). We will summarize each below

Best subset—fits separate models for each combination of candidate variables, and choose the one with smallest out of sample MSE

PCA—Explained in more detail above. Reduces a set of possibly correlated covariates into a smaller set of uncorrelated covariates that are linear combination of the former. After applying PCA, the authors apply the best subset to the PCAS,

LASSO —Explained in more detail above. To avoid over-parametrization, includes a penalty that shrinks the coefficient estimates of covariates

Gradient Tree Boosting—in this paper, this involves classification tress in the form of simple rules involving the unit and area level covariates.

Support Vector Machine—In the training stage, a non-linear decision boundary (kernel) is constructed, and each response is classified as belonging in one of two categories. The authors use radial kernel and cross validation to choose the optimal values of the two training parameters.

Our evaluation of MrP

Our analysis of the Auto MrP methodology revealed that it has several limitations. First, there is a shortcoming of MrP itself when it comes to our application (and many others that share similar features). MrP relies on having all survey-design variables available for constructing the categories available or having non-informative sampling. This is often not feasible in household surveys, as this information is not available, and the sampling is often informative. The Endes survey we used has informative sampling, and we did not have all the design variables at our disposal.

In addition, the Auto MrP package did not perform well with larger sample sizes than those the developers tested. The developers had tested sample sizes of up to 5,000 observations. In contrast, our dataset had 90,000 observations.

Furthermore, the proposed way of computing confidence intervals via bootstrap was not well justified either via theory or empirical evidence. It is also computationally intensive.

Lastly, the authors only tested their methodology with 7 covariates, while we had over 500.