Exploiting Data to its Fullest

Machine Learning and Small Area Estimation

December 2023

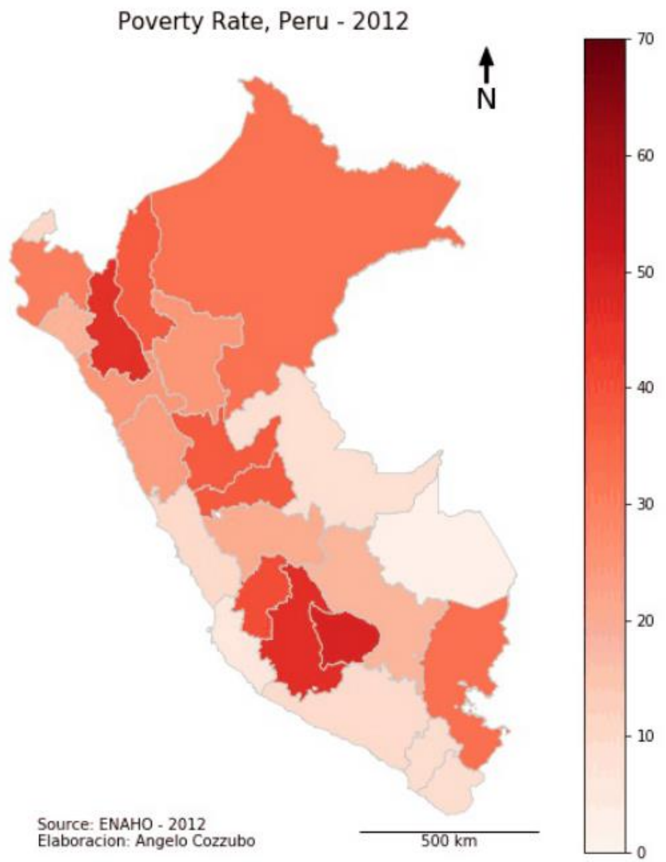
Angelo Cozzubo (International Programs)

Carolina Franco (Statistics and Data Science)

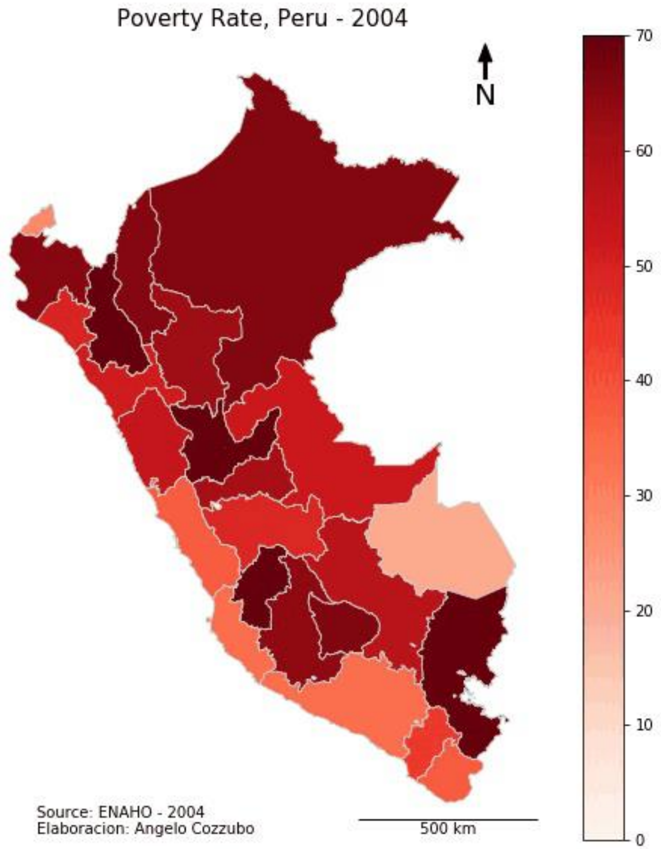
NORC LABS

Exploiting Data to its Fullest

Traditional national surveys provide broad estimates

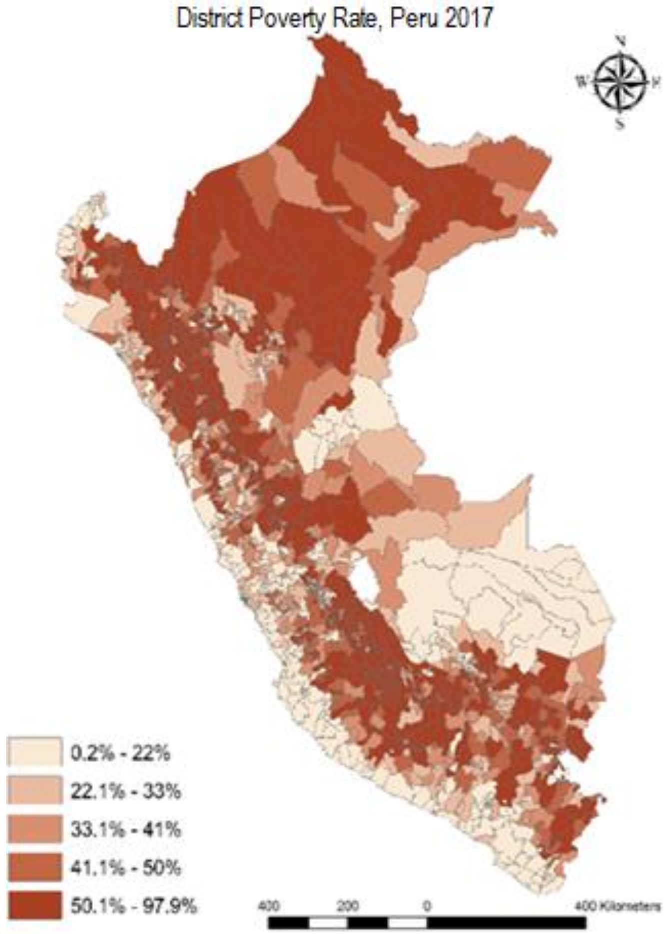


We may be lucky and even have many waves of data



Why stop here?

We can use Small Area Estimation (SAE)

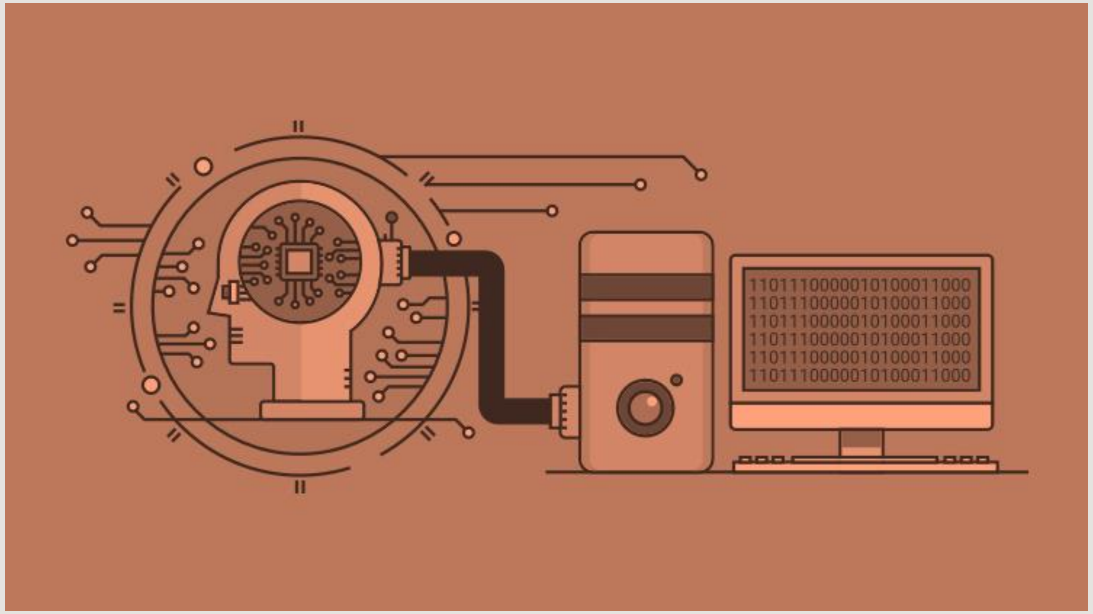


What is Small Area Estimation (SAE)?

- **Problem:** surveys often cannot accurately estimate all quantities of interest through “traditional” methods
- **Goal:** Estimating quantities for geographic or demographic subdivisions with small or no sample size.
- **SAE:** modeling techniques to borrow strength from additional information as admin. records, censuses, neighbors, etc.
- **Results:** ↓ uncertainty of survey estimates!
- **Impact:** Allows publication of local-level indicators that would otherwise be suppressed
- **Applications:** Disease mapping, insurance coverage, poverty mapping, unemployment at local levels, etc. etc. etc.!

And again, why stop in traditional SAE?

Since SAE is a predictive task → potential to exploit Machine Learning



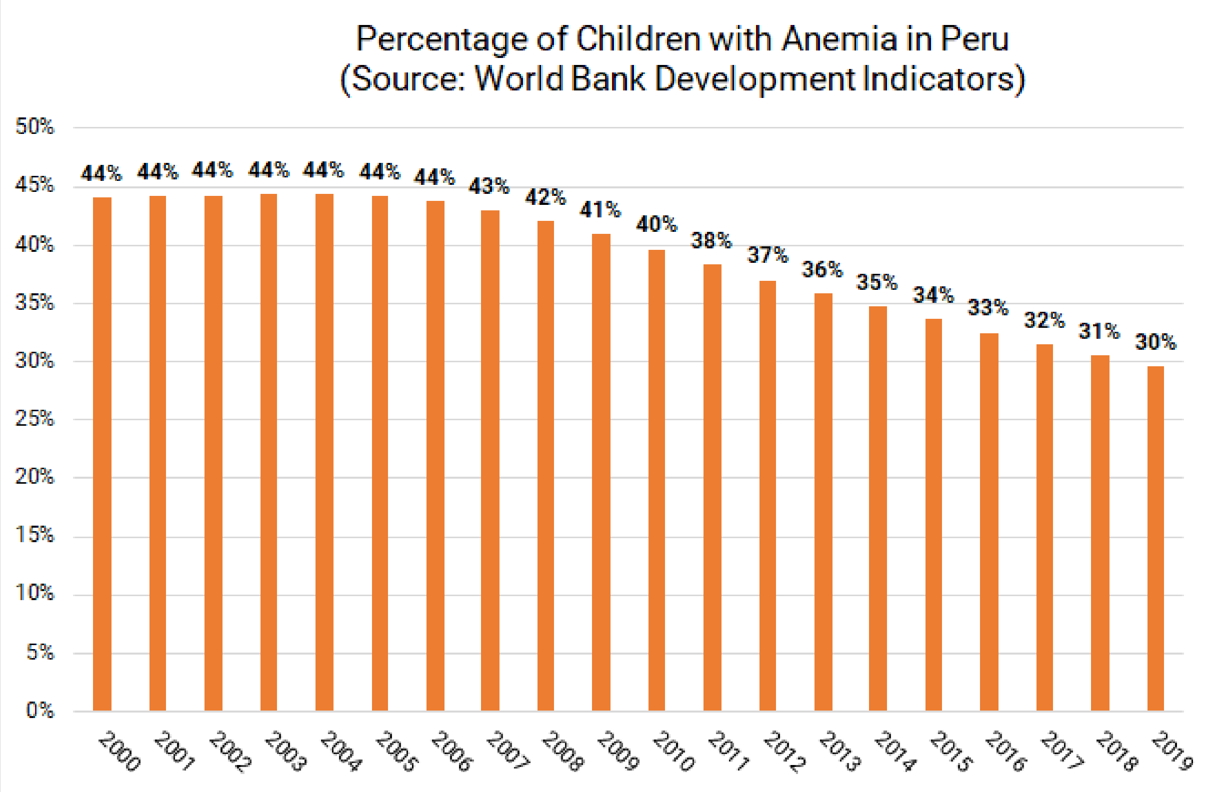
NORC LABS

Our implementation

- ✓ SAE of anemia prevalence in Peru
- ✓ Pressing problem: local estimates not available
- ✓ Access to non-public rich datasets. **558 clean covariates**
- ✓ Special challenge: $K > N$
- ✓ Pool several waves of data

Anemia in Peru

- 2018 → National Plan to Combat Child Anemia
- Official estimates → Only at the regional level.
- Hard for policymakers to plan local interventions
- We built an anemia prevalence map
  - New SAE-ML approaches
  - Province-level estimates



Data Source: World Health Organization, Global Health Observatory Data Repository/World Health Statistics.Accessed via World Bank Development Indicators

Exploiting Data to its Fullest with Machine Learning

National Statistics Offices do not publish estimates with high uncertainty: UNRELIABLE

Objective → Reduce uncertainty of provincial-level estimates

- We used area level SAE models → Fay Herriot model
- Model *borrow strength* from administrative records and census covariates to reduce the uncertainty
- Spatial Fay Herriot: also borrows strength from neighbors
- We explore techniques to find the “best” set of covariates
  - Expert’s opinion 
  - Stepwise selection 
  - LASSO 
  - Sparse PCA 

The process

Survey data preparation

Data wrangle, clean & pool. Peruvian DHS 2016-2019

Direct Estimates

Estimates (and variance) of Anemia at the Province Level

- Normality tests
- Spatial tests
- CV > 15

Covariates pre-processing

Cleaning & merge: Admin. registry, censuses, etc.

- One-hot encoding
- Missings filter
- Zero-variance filter
- Standardization

Modelling

Compute four rival models: estimates, uncertainty & quality of fit

- Spatial matrix
- Specifics of each selection technique
- Comparable results

Results report

Evaluation & results plot. Model adjustments

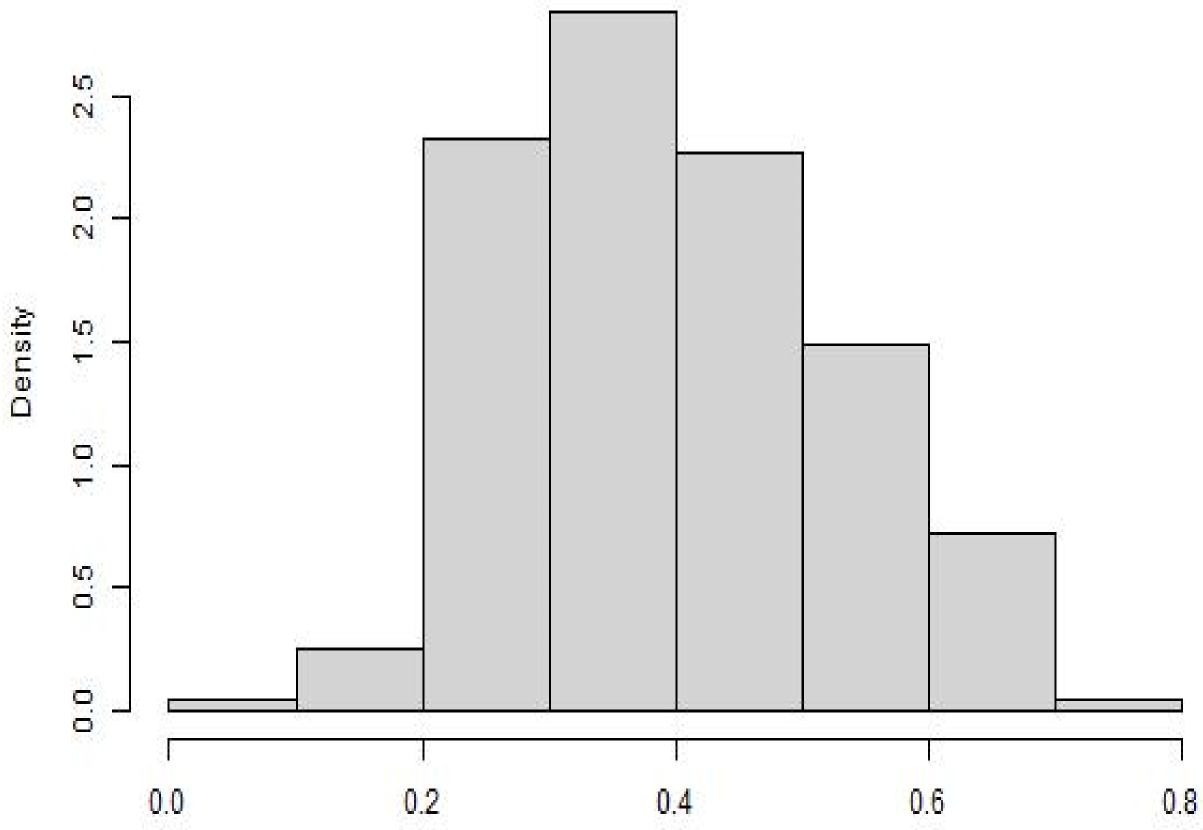
- Convergence plots
- CV and sample size
- Improve of variance
- Choropleths. AIC.

Additional tests & write up

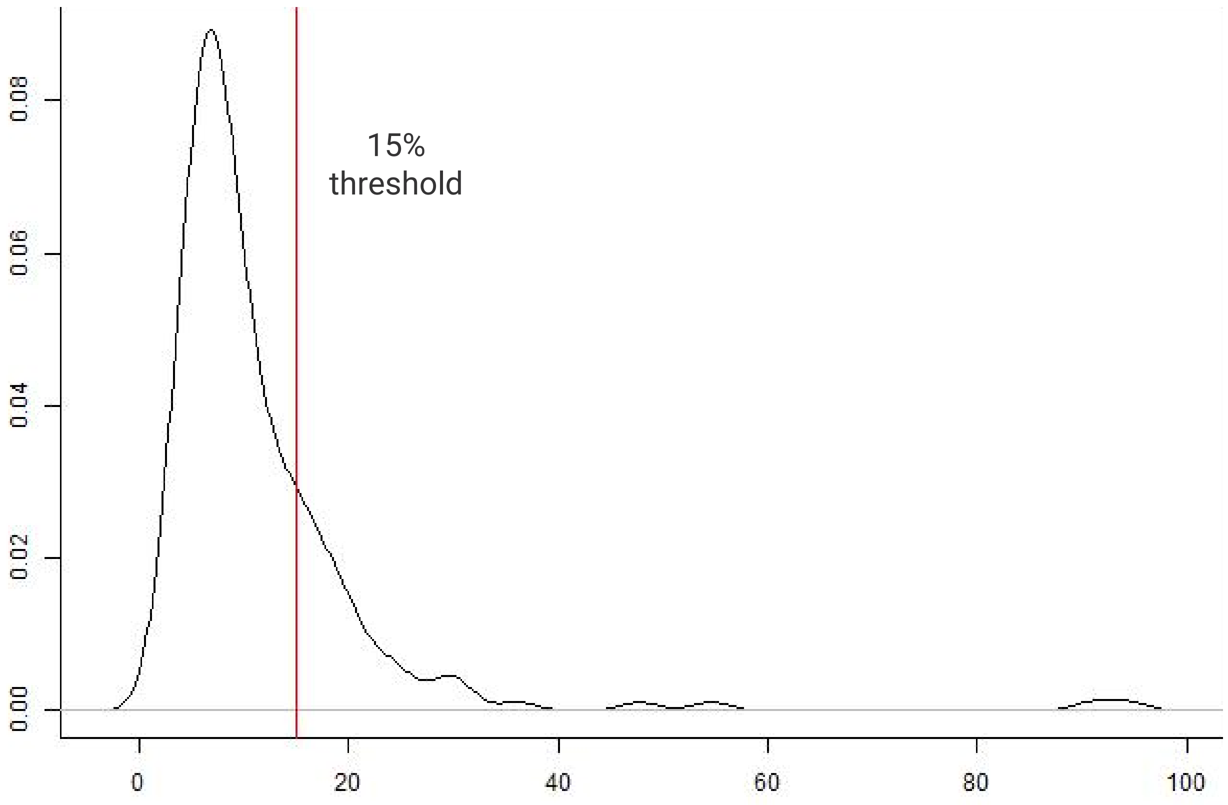
Compare model and synth. estimates to higher level of aggregation

Anemia – 194 direct provincial estimates

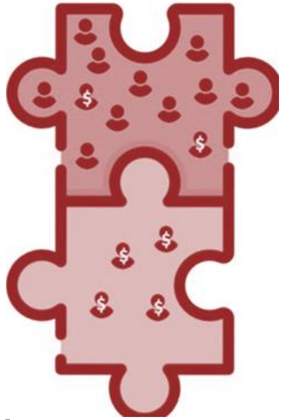
Anemia Incidence



Coefficients of Variation



The (Spatial) Fay Herriot Model



Canonical area level model (1979). For a population characteristic θ_i (anemia)

$$\widehat{Y}_i = \theta_i + e_i$$
$$\theta_i = x_i'\beta + u_i$$

- $\widehat{Y}_i$: vector of direct HT estimates, $\forall$ i provinces
- e_i: *vector sampling errors, independent of* u_i
- x_i' : matrix of explanatory variables
- u_i: vector of area effects

Spatial extension. Borrows strength from neighbors

- u_i follows Spatial Autoregressive process

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

where ρ denoting the autoregression parameter, W a standardized queen proximity matrix and $\eta_i \sim N(0_i, AI_i)$ for A unknown

Rival models for covariate selection

Experts

Zoom interviews to Peruvian experts on health topics

Intersection criteria for the predictors

7 predictors

Stepwise

Bidirectional step method.

AIC and significance criteria

14 predictors

LASSO

Model with Random Effects

Hyperparameter by GridSearch. Correlation filter

97 predictors

Sparse PCA

PCA decomp. Selection main components

1-in-20 criteria. 80% of variance explained

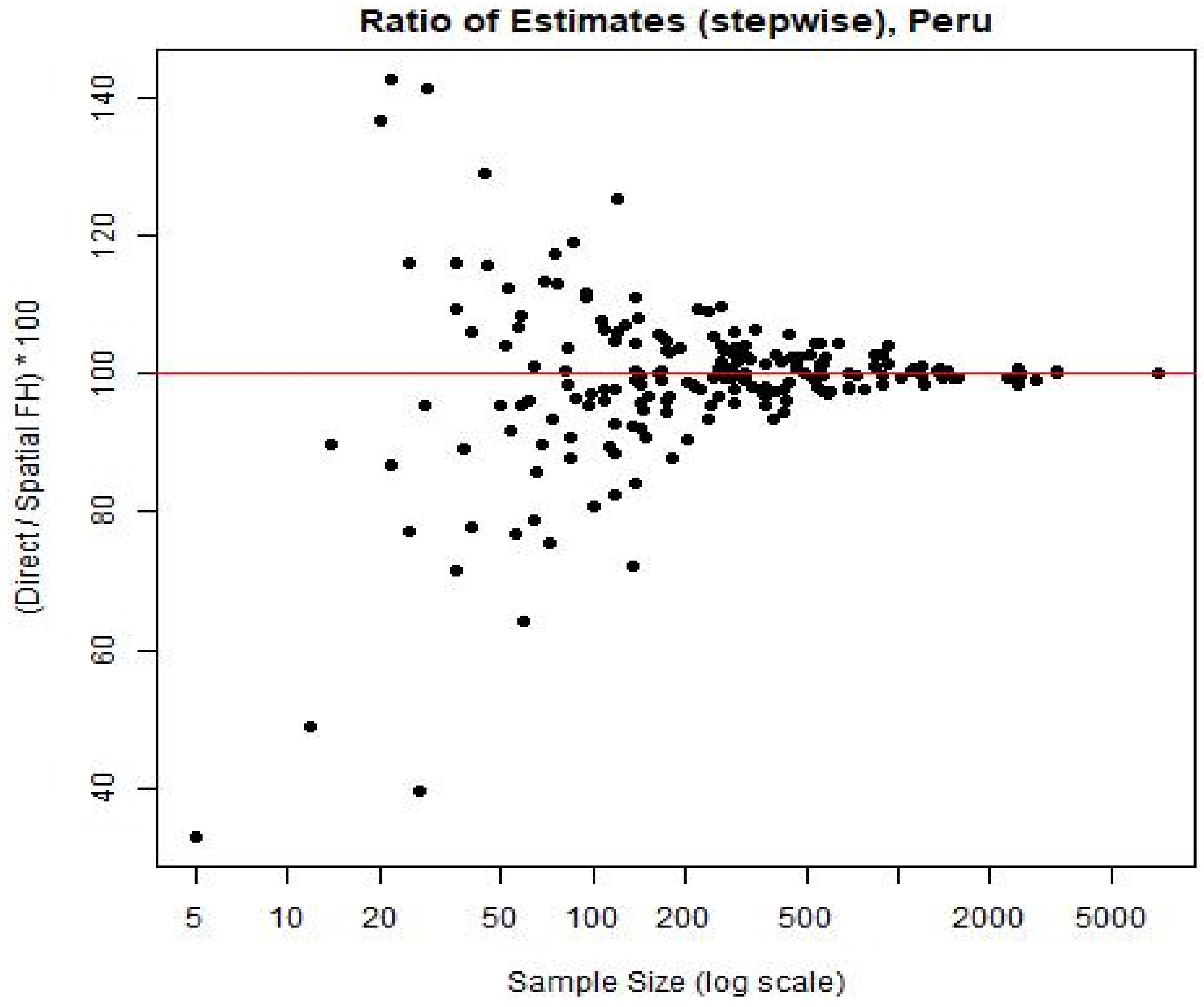
10 predictors (components)

We computed Spatial Fay Herriot models

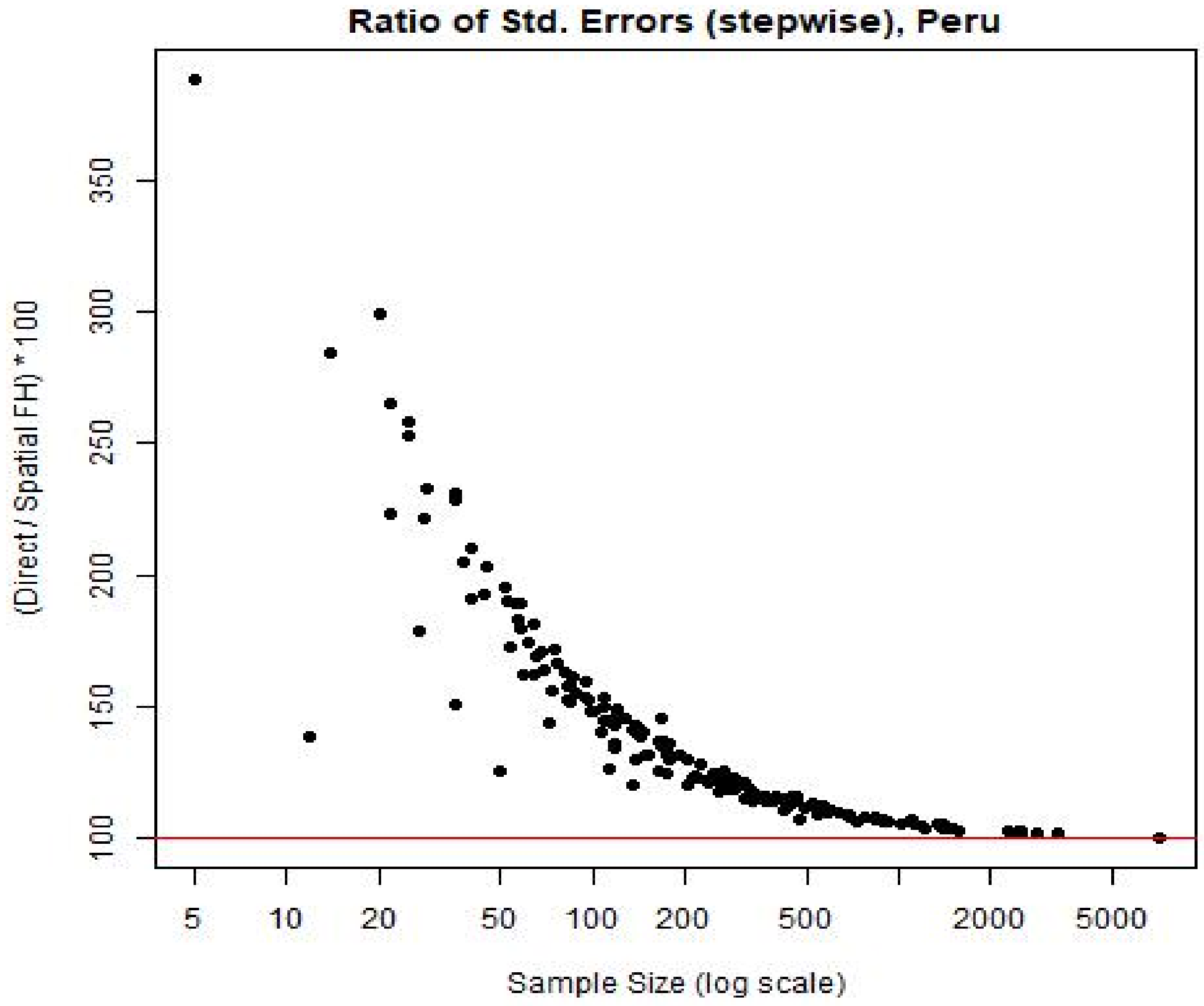
Use set of predictors chosen by each technique

Objective: Improve the variance of the province-level Anemia estimates

Results (I) – Convergence to direct estimate & reduction of variance

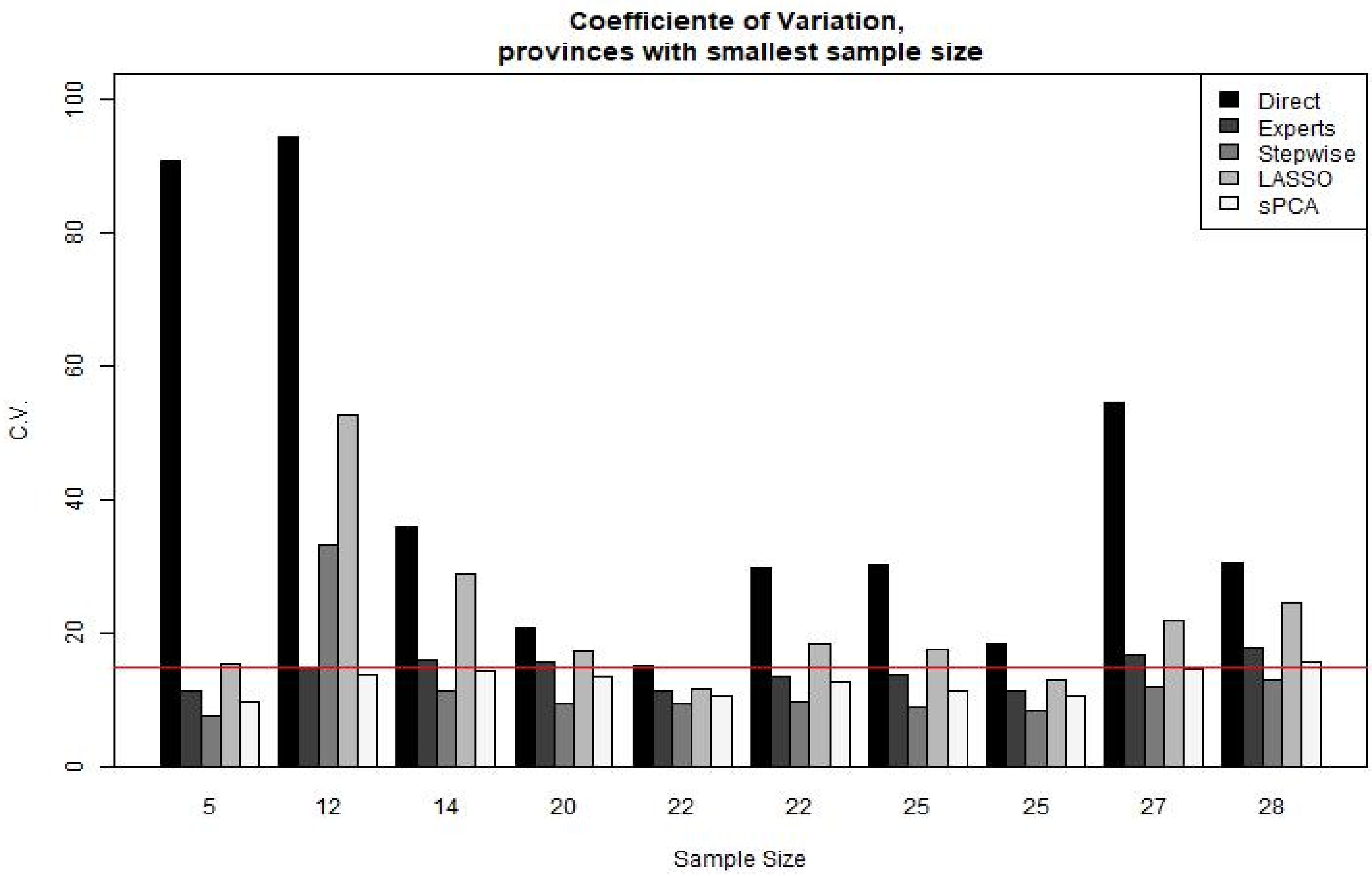


Note: FH = Fay-Harriot model. X-axis in logarithmic scale. Compiled by authors.



Note: FH = Fay-Harriot model. X-axis in logarithmic scale. Compiled by authors.

Results (II) – All variable selection methods helped to reduce the CVs/variances. But some were more effective.



Median variance reduction percentage for each selection method:

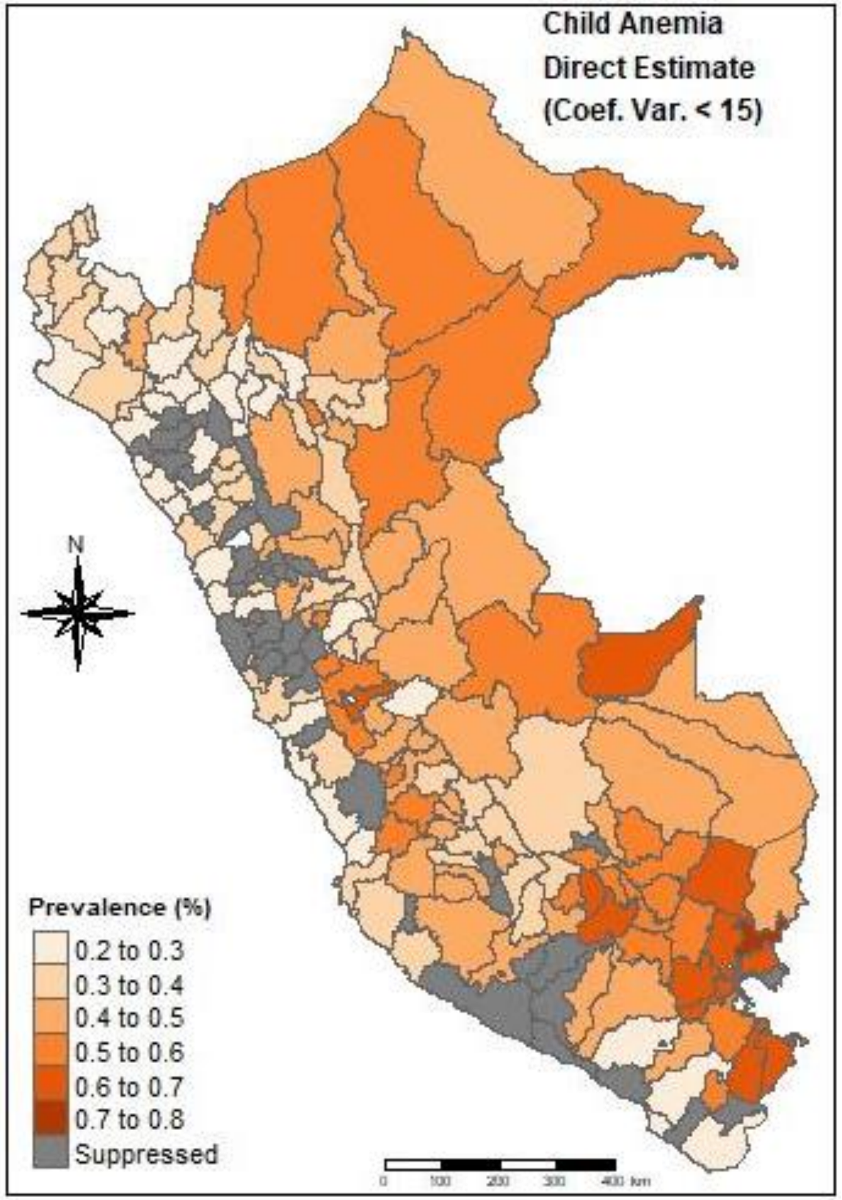

Experts 24%

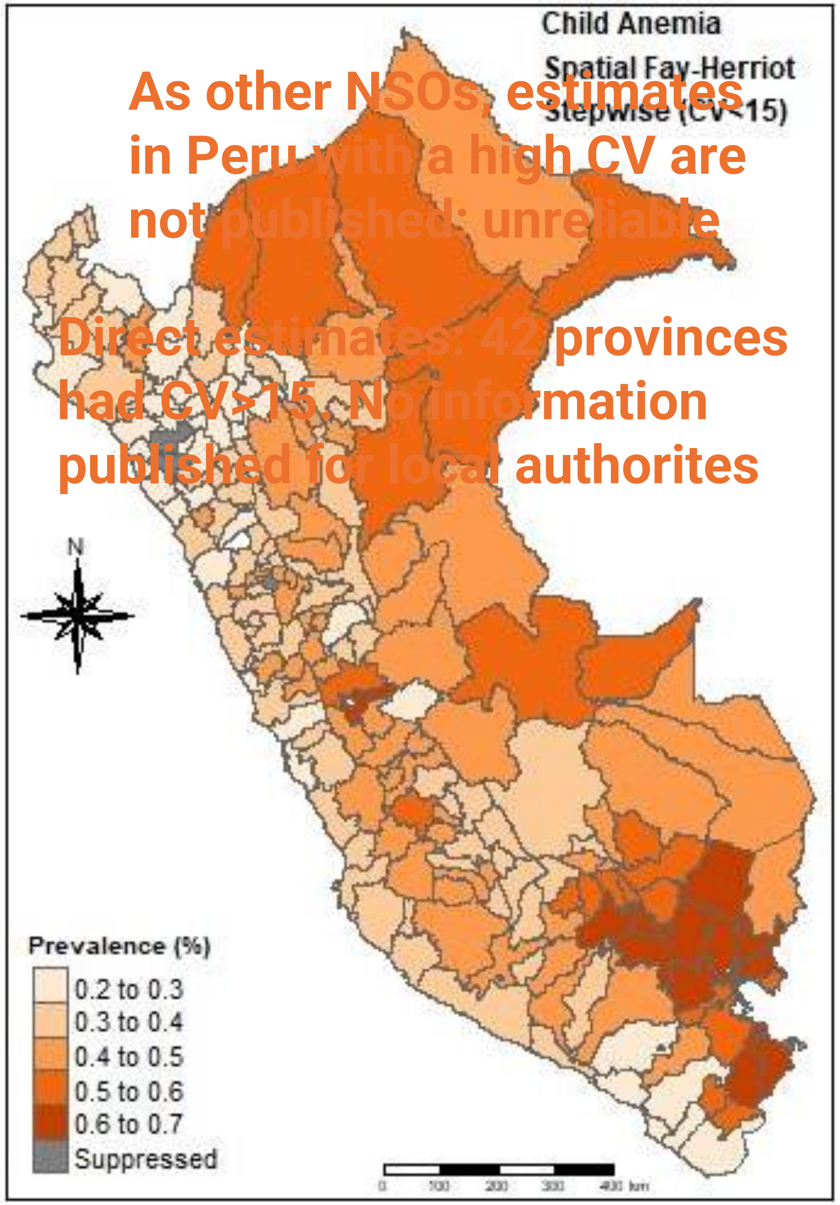

Stepwise 35%


LASSO 12%


sPCA 28%

Results (III) – Recover estimates for 20% of the provinces





As other NSOs, estimates in Peru with a high CV are not published: unreliable

Direct estimates: 42 provinces had CV>15. No information published for local authorites

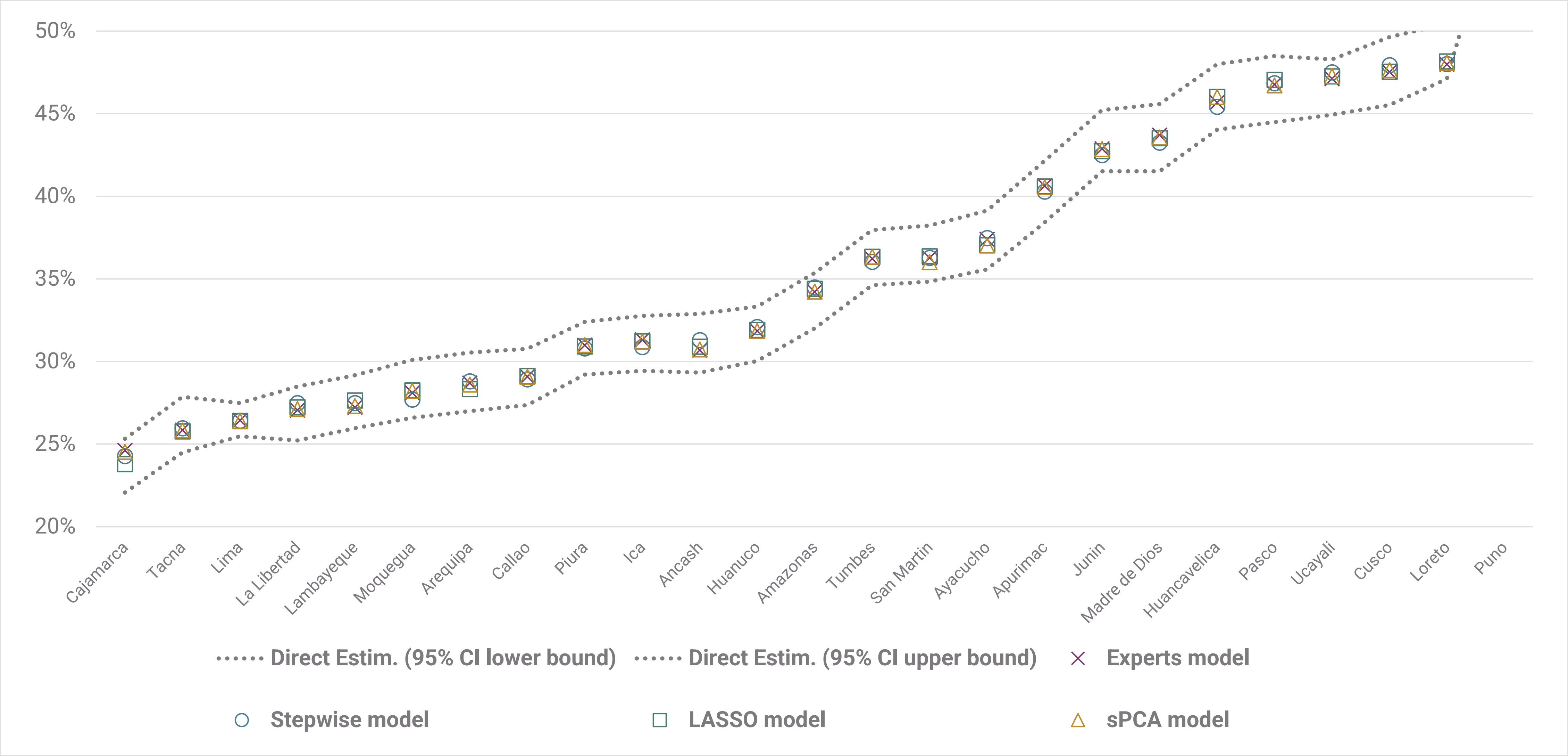
By SAE methods, suppressed estimates are reduced to

- **Experts 13**
- **Stepwise 3**
- **LASSO 25**
- **sPCA 5**

Additional tests – Quality of fit

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

Additional tests – Model estimates $\left(\hat{Y}_i^{FH}\right)$ at the region level



Additional tests – Synthetic estimates $(x_i'\hat{\beta})$ at the region level

Remember the FH model

$$\widehat{Y}_i = \theta_i + e_i$$
$$\theta_i = x_i'\beta + u_i$$

And the best linear predictor of θ_i

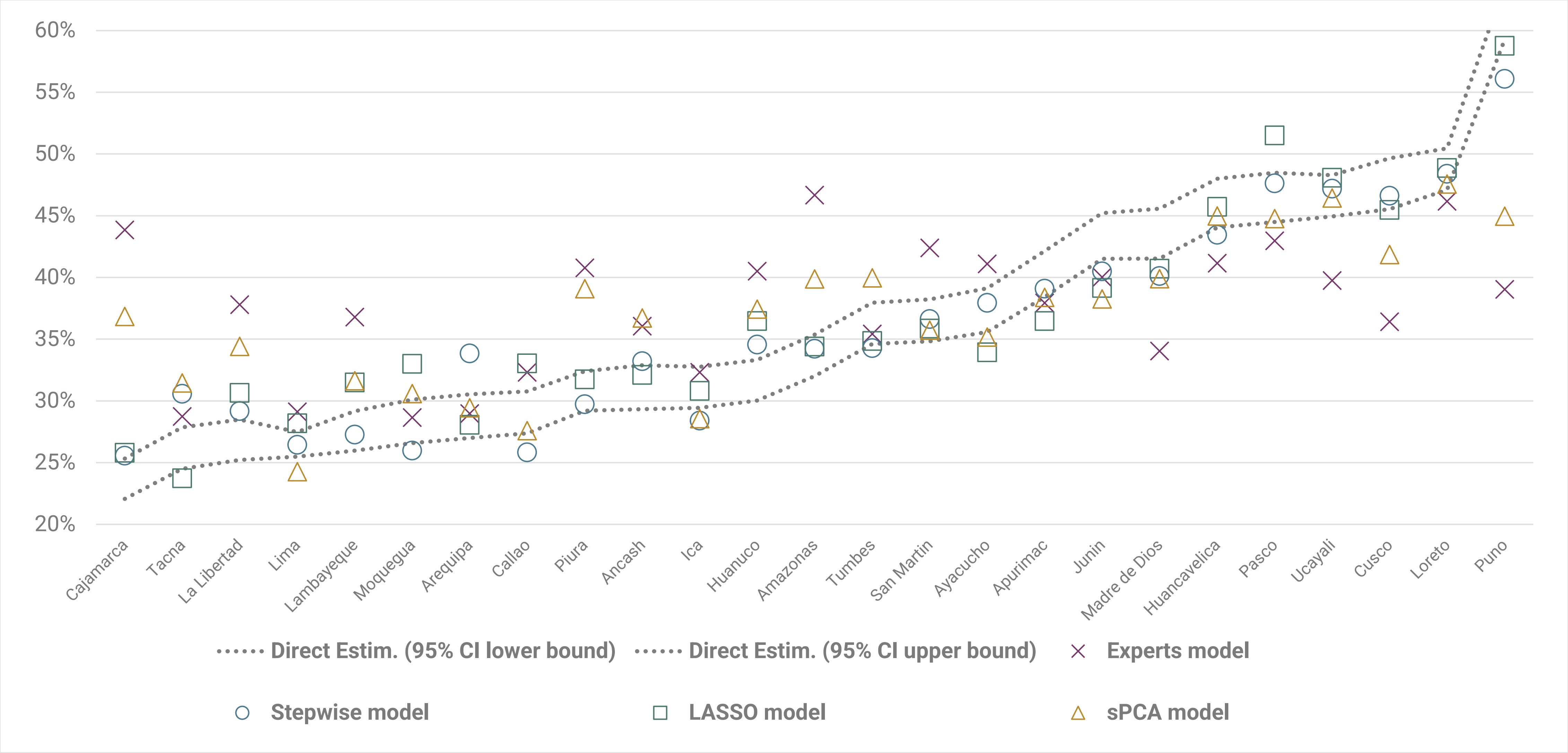
$$\widehat{\theta}_i = (1-\gamma_i)\widehat{Y}_i + \gamma_i x_i'\hat{\beta}$$

Then, FH estimator is a weighted linear combination of

- Direct estimator: $\widehat{Y}_i$
- Synthetic estimator: $x_i'\hat{\beta}$

NORC LABS

Additional tests – Synthetic estimates $(x_p'\hat{\beta})$ at the region level



Results by the Numbers

39 out of 42

Suppressed provincial estimates were recovered

~33,000

Anemic children in recovered provinces

35%

Median variance reduction

NORC LABS

Takeaways

SAE modelling to reduce uncertainty of local estimates

We studied alternative methods for covariate selection from a large pool of candidates (+500)

Stepwise model outperformed other methods based on our metrics

Borrow strength from administrative records, censuses, and neighbors' data

We applied our models to the child anemia problem in Peru. Great uncertainty reduction

Tackled an unresolved statistical problem in Peru. <u>Opportunity for other health applications.</u>

Questions?

Thank you.

Angelo Cozzubo
Data Scientist I
cozzubo-angelo@norc.org

Appendix

Variables selected by the Stepwise method

From Administritative Records

- Total children under 3 years with anemia Percentage 2018
- Children under 3 years with severe anemia Percentage 2018
- Children under 5 years with mild anemia Percentage 2018
- Children under 5 years with severe anemia Percentage 2018
- Percentage of students in public school who only achieved undemanding tasks
- Percentage of students in a private school who achieved a partial learning objectives

From Population Census

- Household does not use manure for cooking
- House walls made of adobe or quincha
- Household does not have a refrigerator or freezer
- Household has a gas stove
- Household has a cell phone
- Household does not have a sound system
- Household has a motorcycle
- Household does not have an electric iron