The University of Chicago

MACS 30123 Final Project: Small Area Vulnerability Map

Angelo Cozzubo

11th December 2020

1 Research question

In the last decade, Peru has gone through rapid economic growth, accompanied by a dramatic reduction of poverty by 37 percentage points. However, the recent economic slowdown has led to wonder if that growth has allowed the country to consolidate a middle class, freed from the risk deprivation, or if those households who left poverty would return to it with an unfavorable macroeconomic context or negative shock as the COVID-19 (see Figure 1).

Our study aims to fill a gap in the empirical literature concerning the household's vulnerability to poverty in Peru using a dynamic approach. The vulnerability concept is defined as the ex-ante probability of a non-poor household falling into poverty next year. Following the methodological proposal of [3,5,6], we estimate this probability and build vulnerability lines which, in complement with poverty lines, allow us to exactly decompose the population into poor, vulnerable, and non-vulnerable ("middle class"). Finally, this decomposition is computed using the census data, which let me derive the district level's vulnerability incidence.

2 Methodology

To obtain a district vulnerability map, I will combine small area estimation and machine learning techniques to implement the vulnerability as an ex-ante risk framework. This application will be conducted using cloud computation.

2.1 Econometric strategy

The quantitative strategy employs three databases:

- Enaho pool biannual panels 2007-2019: as a panel, this dataset contains the label to predict and is used for modeling. This label takes the value of 1 for households that fall into poverty and 0 for those that remained non-poor. Complementary features from nationally representative health surveys and local government census were merged at the region at the district level, respectively.

- Enaho cross-section 2007-2019: this database is almost five times bigger in sample (~372000 households) than the biannual panels (~70000) and contains the same variables except for the label. This dataset will be used to compute the vulnerability line in monetary levels[1] (PEN).
- Poverty Map 2018: this dataset corresponds to the National Household Census 2017, for which the expenditure level of each household was estimated [7]. I will use the pre-computed poverty lines and the estimated vulnerability to decompose the population into three groups: poor, vulnerable, and non-vulnerable.

Using the panel dataset, I estimate four models to obtain the positive labels' best predictive power. For this, I employ logistic regression, Gradient Boosted Trees, Multilayer Perceptrons, and Random Forest as classifiers with an ambitious grid search and 5-fold Cross-Validation (CV) approach to select the best predictor model for households that will fall into poverty. The CV uses the area under the PR curve as evaluation, given that the label variable is imbalanced. The best model is chosen using several criteria as the recall and the TPR, since we are interested in predicting the "ones."

With the best model chosen, I make an out-of-sample prediction in the cross-section database and obtain the predicted probability of falling into poverty for each household. Using this prediction, I compute the vulnerability line as the average expenditure of households with a likelihood of falling of $\widehat{P_i}$. This threshold in the probability is selected using the proportion of households that fall into poverty for the whole period from the transition matrices previously computed in Table 1, as recommended by [1]. The threshold is set to 9% with a caliper of ± 1, and this average gives us the vulnerability line in PEN.

Using the predicted expenditure variable from the Poverty Map, I define the vulnerable population as a household whose expenditure level is above the poverty line but below the vulnerability one. Similarly, poor households will be the ones with expenditure below the poverty line and the non-vulnerable those with expenditure above both lines. Finally, the vulnerability and poverty incidences are computed as the proportion of households in each district with that condition.

2.2 Cloud computing approach

The characteristics of the computing approach were the following:

- Machine learning classification pipeline

[1] All monetary values were deflated spatial and temporarily to Lima2018 levels.

- The modeling was done in the distributed analytics engine PySpark using an AWS EMR notebook with eight x5large nodes.
- Given that the databases were not so big, the descriptive steps and the final incidence computation were done locally using the distributed engine Dask.
- All the steps are scalable for more computation and memory-intensive applications of the same model.

3 Results

We can summarize the results in the following bullets

- Given that the poverty line in PEN Lima2018 is 432.2, the computed vulnerability line is ~1.67 times the poverty line.
- The vulnerable households are distributed differently than the poor ones in the Peruvian geography. As a striking result, districts in the jungle with low poverty incidence have a considerably high incidence of vulnerability (see Figure 3)
- The poverty and vulnerability incidence in Peru seems to have a positive quadratic relationship. As we can see from Figure 2, there are (i) wealthy districts with low vulnerability and low poverty, (ii) poor districts with almost no vulnerable households, and (iii) districts that combine poor and vulnerable populations.
- These results should be considered to design and implement social policy tailored to vulnerable populations in Peru. Moreover, the non-overlapping in the territory makes it necessary to improve the National Targeting Algorithm (SINAFO) in socially include vulnerable households.

4 References

[1] Cruces, G., Lanjouw, P., Lucchetti, L., Perova, E., Vakis, R., & Viollaz, M. (2011). Intra-generational mobility and repeated cross-sections: a three-country validation exercise. The World Bank.

[2] Dang, H. A., & Lanjouw, P. (2013). Measuring poverty dynamics with synthetic panels based on cross-sections. The World Bank.

[3] Dang, H. A. H., & Lanjouw, P. (2014). Welfare dynamics measurement: Two definitions of a vulnerability line and their empirical application. The World Bank.

[4] Dang, H. A. H., & Lanjouw, P. (2016). Toward a new definition of shared prosperity: A dynamic perspective from three countries. In Inequality and Growth: Patterns and policy (pp. 151-171). Palgrave Macmillan, London.

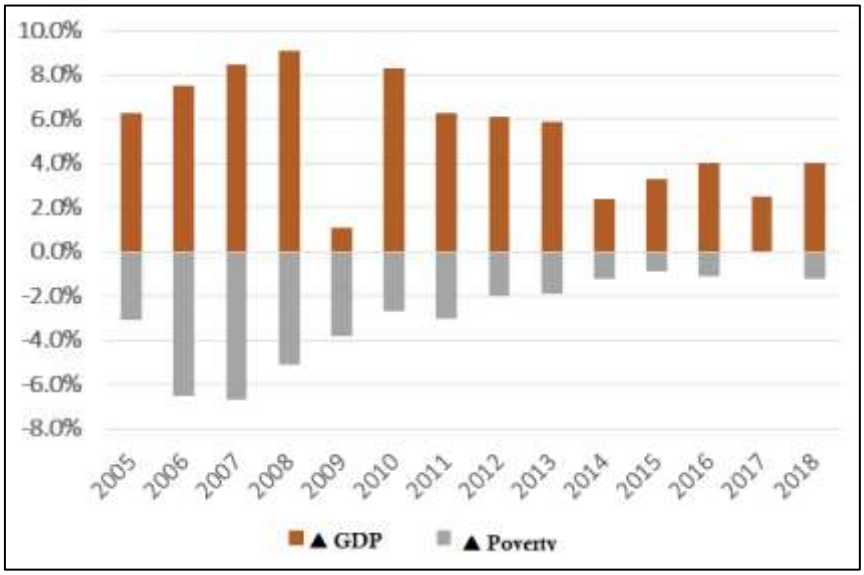
[5] Herrera, J., & Cozzubo, A. (2016). La Vulnerabilidad de los hogares a la pobreza en el Perú, 2004-2014. DDD429, Departamento de Economia, PUCP,

[6] López-Calva, L. F., & Ortiz-Juarez, E. (2014). A vulnerability approach to the definition of the middle class. The Journal of Economic Inequality, 12(1), 23-47.

[7] INEI (2020). Mapa de pobreza provincial y distrital 2018. Instituto Nacional de Estadistica e Informatica.

5 Appendix

Figure 1. GDP growth and poverty reduction, 2004-2018



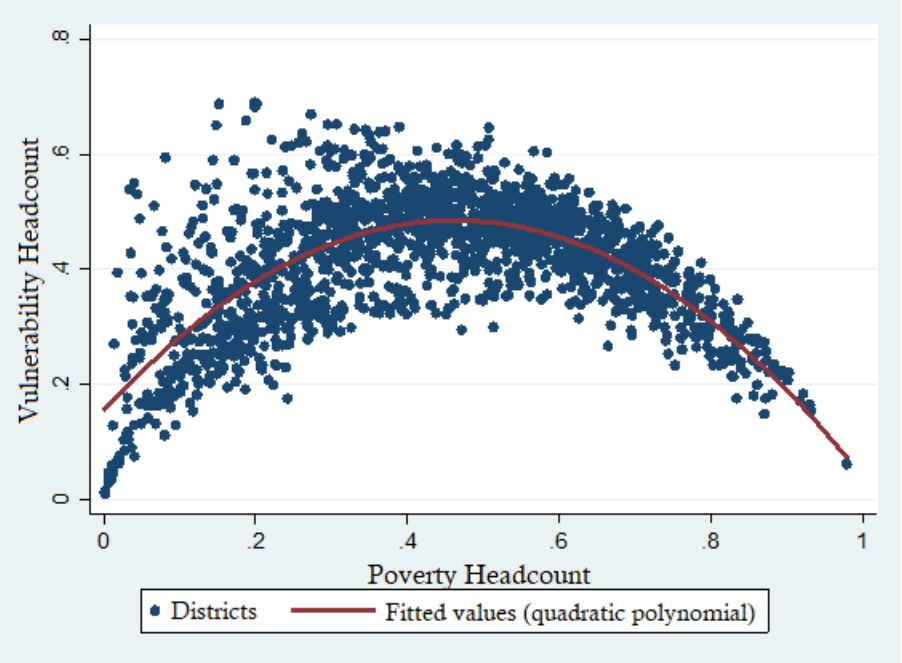
Source: Enaho 2004-2018. Compiled by authors.

Table 1. Transition Matrices, 2007-2019

Panel	% Total Households				% Poor households that escaped	% Non-poor households that fall
	Falls into poverty	Remains poor	Remains non-poor	Escapes poverty		
2007-2008	7.2	23.7	57.6	11.5	32.6	11.2
2008-2009	7.4	21.0	61.5	10.0	32.2	10.8
2009-2010	7.6	18.8	64.4	9.2	33.0	10.6
2010-2011	8.1	16.4	66.3	9.2	35.8	10.9
2011-2012	7.6	13.7	70.6	8.1	37.0	9.8
2012-2013	4.9	11.3	74.5	9.3	45.1	6.2
2013-2014	6.5	10.6	75.4	7.5	41.3	7.9
2014-2015	6.3	10.6	75.7	7.5	41.7	7.6
2015-2016	5.8	9.2	77.0	8.1	46.8	7.0
2016-2017	6.4	8.7	78.2	6.7	43.6	7.6
2017-2018	6.8	8.6	77.8	6.8	44.4	8.0
2018-2019	7.3	8.3	77.2	7.3	46.5	8.6
Total	**6.8**	**13.4**	**71.4**	**8.4**	**38.6**	**8.7**

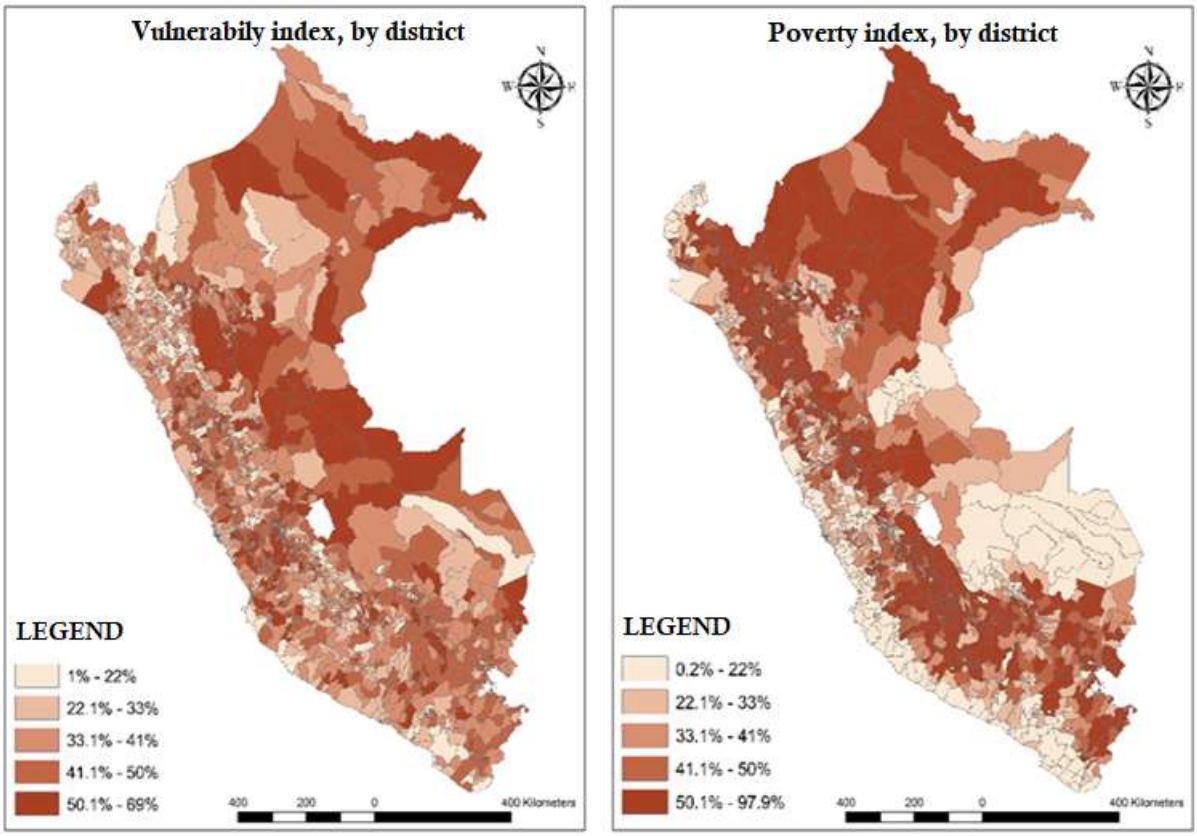
Source: Enaho panel 2004-2018. Compiled by authors.

Figure 2. Poverty and Vulnerability relation, by district 2018



Source: Poverty and Vulnerability Map 2018. Compiled by authors.

Figure 3. Poverty and Vulnerability Maps, by districts 2018



Source: Poverty and Vulnerability Map 2018. Compiled by authors.