Disaggregation of economic indicators to identify left-behind areas

Alberto Díaz-Dapena, Ana Viñuela, Esteban Fernández Vázquez, Fernando Rubiera-Morollón y Elena Lasarte Navamuel

Abstract

The identification of left-behind areas is increasingly becoming a central topic of Regional Economics and the public debate. From an academic perspective, the existence of these areas and their evolution raises a problem of territorial inequality. The identification of key variables could allow researchers to identify what could cause the stagnation of certain places in relative terms to others. These places with a weak economy, do not only matter from a perspective of inequalities, but they could also point towards a problem of incorrect incentives, creating an inefficiency in the allocation of resources. This topic is also increasingly becoming central in the public debate. If the Government tries to modify the incentives of the population to agglomerate in certain places, it could easily need a high public investment in these areas to compensate their handicap. Consequently, there is an intense debate about the necessity of these resources. Citizens in declining areas could demand a higher intervention from the public authorities, like subsidies for traditional sectors, incentives for workers to remain there or integration policies. On the other hand, those citizens living in dynamic areas do not want to suffer a higher fiscal burden because of this. The evaluation of these policies (or lack of them) can easily escalate to the national political debate, with political parties approaching this issue from different perspectives.

From an empirical perspective, identifying these areas to evaluate their situation and create guidelines for policy makers is extremely challenging. Most of the time, there areas are far too small for Statistical Institutes. Even if they want to provide with information to the researchers, their estimations could easily fall into confidentiality problems. As a result, there are no official figures about them in most of the official databases. Hence, making an empirical analysis becomes almost impossible.

Given this lack of data, our research aims to obtain local estimations of economic indicators for Spain, Italy, France and United Kingdom—countries where the necessary microdata from national census is available. The main dependent variable in this analysis is the percentage of households in risk of poverty or social exclusion. According to Eurostat, this variable is defined as 'being at risk of poverty, or severely materially and socially deprived or living in a household with a very low work intensity'. The income of the households is the other variable of interest in the analysis. Thanks to our disaggregation technique it is also possible to compare figures of income from different countries, which in this moment is only available in a few of them. However, other available variables for households in the EU-SILC can easily be disaggregated through the methodology of this research.

Our estimations are obtained thanks to the extrapolation of economic relationships from microdata of the European Income and Living Conditions Survey over the microdata of the national census. This combination of both surveys allows to mix a database where the dependent variable is available, but location cannot be observed, with a wider database, with the same independent variables, information about the location of individuals, but no information about the variable of interest.

The methodology is based on the Tarozzi and Deaton (2009) or Fernandez-Vazquez et al. (2020). The main idea of this process is that it is possible to find a vector of parameters that defines the relationship between a set of independent variables and the variables of interest. However, to be able to make a correct disaggregation of the variables certain requirements must be fulfilled. Firstly, there should be a common group of variables in the microdata of the household survey—where the variable of interest is available—and the census (or other survey with a higher level of disaggregation). One of the most common problems in this regard is the heterogeneity in the definition of the variables in both databases. Even when the same variables are available in both sources of data, there might not be suitable if the intervals or the question in the form is not the same. In addition, even if the variable is the same in both databases, the selection of the variables should try to be as homogeneous as possible between countries. This creates a trade-off between having the model with the highest level of accuracy and being easy to compare the results between countries.

Given the limitations in term of variables, the estimations in this paper are based on two different groups. The first one is refers to the characteristics of the household head, like age, the working status or education level. The second group is the type of household in terms of number of members by age or number of workers.

A second problem is that there might be a different distribution of the independent variables between both sources of data. As a result, the predicted values of the dependent variables may not match with the aggregates that are already available. To solve this issue is necessary to update the predictions with this information, obtaining the optimum prediction which is also coherent with the aggregates. Our approach is based on entropy econometric techniques allowing consistent indicators with the available information coming from official sources.

As indicated in Golan (2017) the entropy econometric approach introduces no assumptions about the distribution of the variables, such as a normal distribution. The optimization process in the entropy econometric approach modifies the estimations as little as possible to make them consistent with the aggregates. The lack of assumptions of this methodology—apart from the restrictions coming from the statistical institutes—makes this approach extremely flexible without having to justify any additional assumption about the distribution of the interest variable. Hence, any deviation from the predictions should come from fulfilling the restriction about the aggregates.

As indicated in Golan (2017), the main restriction of an econometric model in a matrix notation becomes:

$$\frac{1}{n}X^T Y = \frac{1}{n}X^T [P + U]$$

This equation states the same cross-moments in the dependent variable (Y) and the predictions with the independent variables in X. In addition, the restrictions can easily be modified for different scenarios. Fernandez-Vazquez et al. (2020) modifies this restriction where the dependent variable is only observable in the Household Survey (s) and the parameters are the more suitable to be applied over the census microdata (subindex c). This restriction ensures the same cross-moments in the census and the household survey. Consequently, aggregates in both datasets should also be equivalent.

Once the estimations at a household level are obtained, microdata are aggregated again to a local level to be distributed to other researchers. The advantage of this approach is that the local estimates are not produced over the original household survey, which usually have a limitation in terms of representation due to lack of sample. This problem is reduced when the predictions are made over a census database, which allows a higher level of disaggregation in the predictions. The database is scheduled to be released on the website of the EXIT project. In addition to the mean estimations at a local level, this technique can also be used to explore the heterogeneity of the dependent variable within a local entity between groups of the population—like age. In addition, the estimations can also be used to measure the importance of these new variables with other economic variables at a microdata level, measuring differences between rural and urban areas.

References

Fernandez-Vazquez, E., Diaz Dapena, A., Rubiera-Morollon, F., Viñuela, A., (2020) Spatial Disaggregation of Social Indicators: An Info-Metrics Approach. *Social Indicators Research*, 152(2), 809–821.

Golan, A., (2017) *Foundations of info-metrics: modeling and inference with imperfect information*, Oxford University Press.

Tarozzi, A., Deaton, A., (2009) Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *The Review of Economics and Statistics*, 91(4), 773–792.