An aggregate learning approach for interpretable semi-supervised population prediction and disaggregation using ancillary data*

Guillaume Derval[1][0000−0002−6700−3519], Frédéric Docquier[2], and Pierre Schaus[1][0000−0002−3153−8941]

[1] ICTEAM, UCLouvain, Louvain-la-Neuve, Belgium
[2] IRES, UCLouvain, Louvain-la-Neuve, Belgium
{first}.{last}@uclouvain.be

Most countries periodically organize rounds of censuses of their population at a granularity that differs from country to country. The level of disaggregation is often governed by the administrative division of the country. Although census data are usually considered as accurate in terms of population counts and characteristics, the spatial granularity, that is sometimes in the order of hundreds of square kilometers, is too coarse for evaluating local policy reforms or for making informed decisions about health and well-being of people, economic and environmental interventions, security, etc. For example, fine-grained, high-resolution mappings of the distribution of the population are required to assess the number of people living at or near sea level, near hospitals, in the vicinity or airports and highways, in conflict areas, etc. They are also needed to understand how population movements react to various types of shocks such as natural disasters, conflicts, plant creation and closures, etc.

Multiple methods can be used to produce gridded data sets (also called rasters), with pixels of a relatively small scale compared to the administrative units of the countries. Gridded Population of the World (GPW) [1] provides a gridded dataset of the whole world by (mostly) redistributing the population in a given census unit uniformly on the census unit surface.

More advanced and successful models rely on ancillary and remotely sensed data, and are trained using machine learning techniques. These data can include information sensed by satellite or data provided by NGOs and governments.

All methods in the literature are converted into standard supervised regression learning tasks. Supervised regression learning aims to predict an output value associated with a particular input vector. In its standard form, the training set contains an individual output value for each input vector. Unfortunately, the disaggregation problem does not directly fit into this supervised regression learning framework since the prediction function is not directly available for input vectors. In a disaggregation problem, the input consists of a partition of the training set (the pixels of each unit) and for each partition, the sum of the output values is constrained by the census count. This framework is exactly the one

* This paper is an extended abstract of the paper with the same title published at ECML-PKDD 2019.

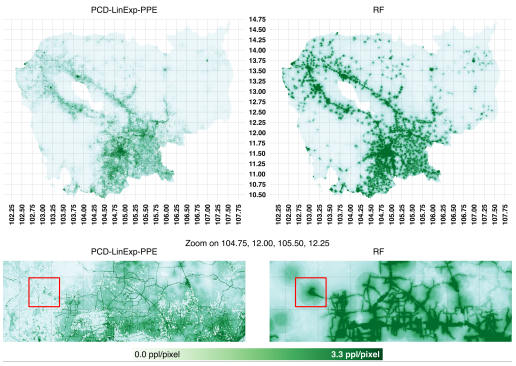

Fig. 1. Cambodian population maps obtained by a specific version of PCD and by RF. The bottom maps are a zoom of a specific, moderately populated region of Cambodia. The red box highlights a region where RF produces seemingly artificial results: it creates a circle around a (non-displayed) hospital, and saturates near the road network.

introduced as the aggregated output learning problem by [2]. The PCD method introduced in this paper conceives the formulation of the disaggregation problem as an aggregated output learning problem. PCD is able to train the model based on a much larger training set composed of pixels. This approach is parameterized by the error function that the user seeks to minimize.

As a case study, we experiment it on Cambodia using various sets of remotely sensed/ancillary data. Our main result, the disaggregated map of Cambodia, is depicted on the left panel of Fig 1 (on the right panel is shown the state of the art, the RF method [3]). The paper also discusses methodological issues raised by existing approaches. In particular, we demonstrate that the previously used error metrics are biased when available census data involves administrative units with highly heterogeneous surfaces and population densities. We propose alternative metrics that better reflect the accuracy properties that should be fulfilled by a sound disaggregation approach. We then present the results for Cambodia and compare methods using various error metrics, providing statistical evidence that PCD-LinExp generates the most accurate results.

References

1. Center for International Earth Science Information Network - CIESIN - Columbia University: Gridded population of the world, version 4 (gpwv4): Population density, revision 10 (20180711 2017), https://doi.org/10.7927/H4DZ068D
2. Musicant, D.R., Christensen, J.M., Olson, J.F.: Supervised learning by training on aggregate outputs. In: Seventh IEEE International Conference on Data Mining (ICDM 2007). pp. 252–261. IEEE (2007)
3. Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J.: Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PLOS ONE **10**(2), 1–22 (02 2015). https://doi.org/10.1371/journal.pone.0107042