Potential applications of emergent class models

Fall 2022

In this short note, I present the Louf emergent class model[1] and provide some early testing to assess its relevance.

Consider a given space divided into T parts. The population is distributed according to a scale into categories (classes). Let $n_\alpha(t)$ denote the population belonging to class α residing in part t. Two simple measures are then constructed: representation $r_\alpha(t)$, which gives the ratio of the weight of category α in region t compared to its weight in the entire space; and an exposure measure $E_{\alpha\beta}$, which can be interpreted as an index of closeness between two classes in space (it is the weighted average, by the population of class α, of the representation of β):

$$r_\alpha(t) = \frac{n_\alpha(t)/n(t)}{N_\alpha/N} \tag{1}$$

$$E_{\alpha\beta} = \frac{1}{N_\alpha}\sum_{t=1}^{T} n_\alpha(t) r_\beta(t) \tag{2}$$

In a space where classes are randomly located, a perfectly non-segregated space, we get for all classes $\mathbb{E}[r_\alpha(t)] = 1$. A class is said to be over-represented in a given space if $r_\alpha(t) > 1 + 2.57\sigma_\alpha(t)$ where:

$$\sigma_\alpha(t) = \sqrt{\frac{1}{N_\alpha}\left(\frac{N}{n(t)} - 1\right)} \tag{3}$$

Similarly, in the ideally non-segregated space, $\mathbb{E}[E_{\alpha\beta}] = 1$. If $E_{\alpha\beta} > 1$, there is attraction between the two classes, repulsion if $E_{\alpha\beta} < 1$. To be more precise, the confidence interval is obtained using Chebyshev's inequality:

$$\mathbb{P}(|\mathbb{E}[E_{\alpha\beta}] - E_{\alpha\beta}| \geq 10 \times \sqrt{\mathbb{V}E_{\alpha\beta}}) \leq 0.1 \tag{4}$$

If there are only two categories in space, the exposure reaches its theoretical maximum $E_{\alpha\beta}^{max} = \frac{N^2}{4N_\alpha N_\beta}$. Another important property of exposure is its symmetry: $E_{\alpha\beta} = E_{\beta\alpha}$

To proceed with aggregation, we start with the finest class subdivision proposed by the data. In the original article, it is the income scale of households in the Community Survey of the Census Bureau (2014), which proposes a scale with 16 classes (from less than $10,000 per year to over $200,000 per year). The authors construct a 16×16 matrix for all values of $E_{\alpha\beta}$. The iteration then proceeds as follows:

- The exposure values are normalized with respect to their maximum; denote these normalized values as $\tilde{E}_{\alpha\beta} = 1 + \frac{E_{\alpha\beta}-1}{E_{\alpha\beta}^{max}-1}$;
- We isolate in the matrix the pair ψ, ω for which $\tilde{E}$ reaches its maximum;
- We merge classes ψ and ω;
- We repeat until the aggregated classes exhibit no aggregation dynamics, i.e., until the exposure coefficients in the matrix are all less than $1 + 10 \times \sqrt{\mathbb{V}E_{\alpha\beta}}$.

On American data, the authors isolate, through iteration, three income classes: a popular block ($< $50,000) representing 59% of the population, an affluent block ($> 0,000) weighing 29%, and a small intermediate class (11%).

[1] Louf, R. Barthelemy, M., 2016. "Patterns of Residential Segregation," PLOS ONE, Public Library of Science, vol. 11(6), pages 1-20, June.

Out of curiosity, I replicated the method on income data declared to the DGFiP. These data provide a breakdown by income brackets for most municipalities with more than 1000 inhabitants. Recent data are divided into 8 brackets; I preferred an older base, that of 2010, where the division was into 12 brackets (from less than 9,400 euros for the first to over 97,501 euros for the last). It includes 4218 municipalities representing 26 million tax households (about two-thirds of French taxpayers). The matrix of exposure values obtained in the first iteration is plotted below:

$$\begin{pmatrix} 1,108 & 1,034 & 1,028 & 1,015 & 1,001 & 0,991 & 0,977 & 0,959 & 0,934 & 0,901 & 0,864 & 0,836 \\ & 1,049 & 1,039 & 1,037 & 1,026 & 1,015 & 0,996 & 0,986 & 0,972 & 0,944 & 0,892 & 0,832 \\ & & 1,034 & 1,034 & 1,024 & 1,014 & 0,997 & 0,988 & 0,978 & 0,954 & 0,904 & 0,837 \\ & & & 1,039 & 1,029 & 1,019 & 1,002 & 0,994 & 0,984 & 0,963 & 0,912 & 0,84 \\ & & & & 1,029 & 1,021 & 1,009 & 1 & 0,992 & 0,977 & 0,935 & 0,867 \\ & & & & & 1,023 & 1,013 & 1,006 & 1 & 0,99 & 0,96 & 0,899 \\ & & & & & & 1,019 & 1,014 & 1,012 & 1,013 & 1,004 & 0,965 \\ & & & & & & & 1,023 & 1,028 & 1,036 & 1,04 & 1,023 \\ & & & & & & & & 1,051 & 1,073 & 1,088 & 1,082 \\ & & & & & & & & & 1,128 & 1,175 & 1,184 \\ & & & & & & & & & & 1,348 & 1,549 \\ & & & & & & & & & & & 2,557 \end{pmatrix}$$

Overall, the attraction forces at the ends of the income hierarchy seem weaker than in the American data. In the aforementioned article, the factor $E_{\alpha\beta}$ between the last and the penultimate class is 2.3, compared to 1.549 for our data. Our iteration proceeds in the following order: we aggregate brackets 10 and 11 into an upper block, which eventually encompasses brackets 9, 12, and 8 in order. We then merge brackets 1 and 2, which subsequently aggregate brackets 3 to 5 in order. At the very end of the iteration, bracket 7 is attached to the upper group, bracket 6 to the lower group. In summary, the iteration ends with two groups, a popular group (56% of the sample) and an affluent group (with the limit being 18,750 euros in annual declared income to the tax authorities). Their $E_{\alpha\beta}$ is 0.953, their isolation $E_{\alpha\alpha}$, respectively 1.037 and 1.06, is significantly lower than American levels, although the comparison must be made with caution (the authors work at a finer level, the city neighborhood, which would be equivalent to IRIS in INSEE data).

What interests me more here is not so much the chosen index (which after all gives very macro and not necessarily intuitive results) but the method used, the idea of emerging classes from an endogenous process, by iteration-grouping until reaching a threshold of internal homogeneity. The problem for such a research project would therefore be less obtaining the data than constructing a finer index.