Does aggregation preserve communities?

Adeline Decuyper[1], Yérali Gandica[1,2,*], Christophe Cloquet[3], Isabelle Thomas[1] and Jean-Charles Delvenne[1,2]

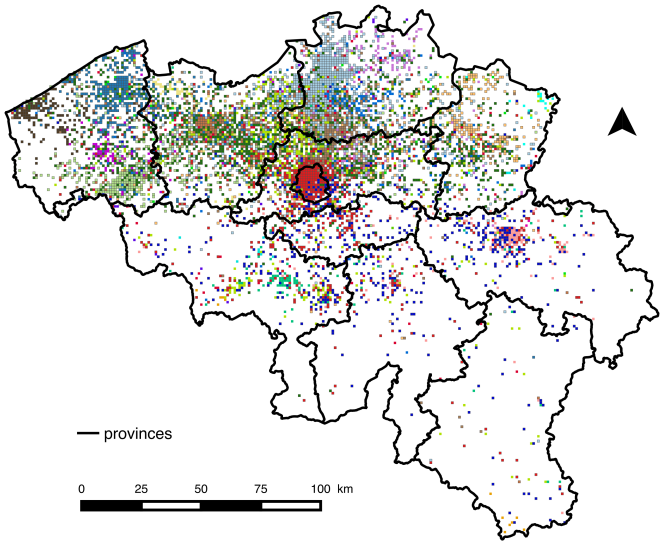
1 Center for Operations Research and Econometrics (CORE), Université catholique de Louvain, Louvain-la-Neuve, Belgium.
2 Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université catholique de Louvain, Louvain-la-Neuve, Belgium
3 Poppy, Rue Van Bortonne, 7, 1090 Jette, Belgium.

The ecological fallacy refers to the statistical bias caused by the aggregation of individuals into categories. In geography, particular form of such fallacy is called the Modifiable Areal Unit Problem (MAUP). MAUP affects results when individual-based measures of spatial phenomena are aggregated into any administrative units, e.g. districts or municipality. The aggregation can also be related not to a geographical context, but according to any individual or social category, for instance age, economical income or the intensity of any kind of social contacts. Some other reasons can be privacy concerns, where the datasets may be only accessed by researchers after any kind of uncontrollable and even unknown way of aggregation.

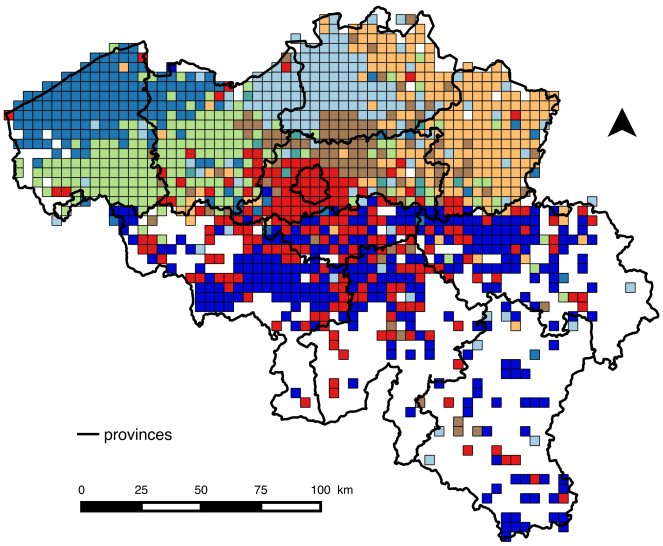
We are interested in analyzing the impact of this fallacy. We focus in one of the most common task in network science, community detection. We measure quantitatively the impact of node aggregation on the community structure in networks and we introduce the aggregability index, predicting quantitatively the robustness of the community structure to a graph, in order to place it into a given aggregation class. We show that some community detection methods are more suitable than others when computing communities on aggregated networks.

We illustrate our methodology on a dataset of geolocalized tweets in Belgium, and mobile phone data from one provider in Belgium. By means of the normalized mutual information we have shown that only the phone calls data preserves the community structure of the fine-grain level. We show that our proposed index is able to predict that the Twitter dataset is highly sensitive to aggregation, while the mobile phone dataset is robust against aggregability.

In order to analyze the effect of aggregating data, we systematically increase the scale of spatial aggregation. As an example, we show in Fig. 1 the result for communities detected in the Twitter network aggregated data among two different sizes of square cells. It is shown in Fig. 1-a) communities detected in networks aggregated by square cells whose side measures 1km and in Fig. 1-b in cell sizes of 4km. We can see that as the aggregated area increases, some communities of non-geographical close people (as the light green community having people in separated provinces in Fig. 1-a), were forced to merge into geographical closed communities (light green in Fig. 1-b). Further explanations about this mechanism, how to measure it and how to be able to detect it, is the material proposed for the talk.



(a) Network with aggregated cells of side 1km



(b) Network with aggregated cells of side 4km

Figure 1: Communities detected in the Twitter network, aggregated into square cells of two different sizes.