Area Classification on Dimension Reduced Micro-level Survey Data

Mingze Huang

November 27, 2020

Github Account: https://github.com/mingzehuang

Backgroud

There are some area classification for UK National Statistics (Vickers and Rees 2007). They simply use K-means methods to classify UK statistical areas based on 41 census variable. For some reason, there is no classification on US census data, although there are much more census data in US.

The classification for statistical areas may not be a difficult problem in Statistics but essentially very important in social-economic research. Nowadays most of causal inference or policy evaluation papers in Microeconomics are based on random control trial (RCT) such as difference-in-difference approach. However, the first thing to proceed difference-in-difference is to select control group and treatment group. That requires the similarity between control group and treatment group except for treatment (policy implementation). In practice, microeconomic researchers just pick the adjacent areas as control group and treatment group and includes some covariates potentially affect outcome not through the treatment. I think the better way is to classify the statistical areas into different clusters then select control group and treatment group within the same cluster.

Microeconomic research are shifting from aggregate level to regional level in recent years. Regional comparison on the response of specific macro shocks also need to cluster areas into several groups for comparison.

Functionality

The structure of my package I think should be suitable for multi-level multi-purpose area classification. The input argument will be survey data with lots of area observations (n) and many variables (social-economic features). If the dimensions of features are small, the package is going to proceed classification directly, otherwise it will do sparse PCA first, then go to classification. If the data are hierachical, the sparse PCA will have penalty term for group LASSO and exclusive LASSO (the magnitude of penalty coefficient can be customized by user to meet their interpretation purpose). For classification, if the data has pre-classified labels for training purpose, the package is going to do supervised classification on training data then apply the model to unlabeled area. If the data have no labels, it will directly go to unsupervised learning to classify by features. Users should be able to adjust the weights on different features in classification. The unsupervised learning I'm going to include at least K-means, the supervised learning I'm going to include at least multinomial logistic regression. I'm going to include more methods gradually.

Eventually it should be output clusters and principle components for each clusters for users by customization. The interface shiny page would be also included in package so that it will generate a dashboard shows different layers, and top influential components which determine the cluster for a specific area.

References

Vickers, Dan, and Phil Rees. 2007. "Creating the Uk National Statistics 2001 Output Area Classification." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170 (2): 379–403.