# Area Classification on Dimension Reduced Micro-level Survey Data

Mingze Huang

11/19/2020

## Background

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.