# Potentials of data fusion methods to enhance income predictions for microsimulations

**Keywords:** data fusion, income analysis, microsimulation, missing data, statistical matching.

The overall aim of this project is to examine spatial income inequality and to disentangle its determinants in Germany over time on a small regional scale. Microsimulations of spatial inequalities require a comprehensive measurement and profound prediction models for income information. However, in Germany there exist different data sources to model income inequalities with their own particular strengths and drawbacks. While the taxpayer data – an income register recording the total population of all taxpayers in Germany – provides reliable income details, it does not cover variables concerning the social disaggregation that are relevant for income modelling, for example education and working time. The German Microcensus does contain these variables, however, the income information face several shortcomings, i.e. self-response bias, top censoring, reports of classified data. Therefore, we seek to match the taxpayer data with the Microcensus data in order to incorporate variables concerning the social disaggregation in the underlying income models. This requires sophisticated and performant data fusion methods that allow the joint analysis of variables from (at least) two different data sources, where each of the data sources originally served a different purpose. We discuss results from a simulation study in order to investigate different data fusion techniques and evaluate its potentials to improve and enhance income modelling. One central aspect of our study is the suitable incorporation of the available income information within the data fusion process, taking into account that the income information obtained from the Microcensus differs drastically from those observed from the taxpayer data. Our preliminary results show that two data fusion methods, random forest and predictive mean matching, are the most promising methods in our context. Moreover, the suitable incorporation of the available income information further improves the data fusion outcome.