**Microsimulations of income inequalities and potentials of data fusion methods to account for social disaggregation**

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

# Introduction

The overall aim of this project is to examine spatial income inequality and to disentangle its determinants over time in Germany on a small regional scale. Not only individual income inequality but also regional disparities gain increasing attention in the public debate in Germany. It is an important aim of the federal government to guarantee equal living conditions [1]. To design appropriate policy measures, it is essential to study the extent and form of regional disparities especially with regard to economic and demographic differences and developments. Microsimulations provide a useful method for that purpose. At the NTTS 2021, the research idea and preliminary results will be presented.

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. We use the taxpayer panel (TPP) – an income register recording the total population of more than 54 million taxpayers in Germany from 2001 to 2014. The data set covers all municipalities and reveals the full income distribution from the bottom to the top. However, due to the purpose of the administrative data, important covariates covering the social disaggregation like education, family structures, occupation, and working time, are not available.

The German Microcensus (MC), which is a representative 1% sample, covers such socio-demographic variables. Yet, the income variables are known to face several drawbacks, i.e. self-response bias, top censoring, reports of classified or heaped data [2]. Therefore, we seek to combine information from the TPP and the MC in order to benefit from the strength of both studies – the reliable income details within the taxpayer data and the social disaggregation variables observed in the MC.

This requires sophisticated and performant data fusion methods. The objective of a data fusion, also known as statistical matching, is to jointly analyse variables from (at least) two different data sources, where each of the data sources originally served a different purpose [3]. The aim is to perform an analysis containing variables from some study A, here the TPP, in combination with additional variables from a different study B, in our case the MC.

Therefore, our research objectives are two-fold: First, we apply different methods to model income inequalities and choose the method with best predictive performance for microsimulation purposes. We estimate these preliminary income models only on the taxpayer data. Second, we seek to address the question of how these income models can be improved by statistically matching the taxpayer data with the Microcensus data in order to incorporate variables concerning the social disaggregation in the underlying income models. Precisely speaking, the models developed in the first step are extended by including further variables added to the TPP via data fusion methods. This includes a discussion about possibly appropriate data fusion methods that are to be evaluated within a simulation study.

# Methods

## Income prediction models

First, we apply conventional linear regression methods and mixed model approaches to estimate a tax unit’s income based on the taxpayer panel 2012-14. We compare the following four main estimation approaches:

1. Conventional models of the form: with *Y* and *X* being the vector of incomes and the explanatory variables of individual *i* at time *t*, respectively, being the vector of coefficients and the error term .
2. Panel models with random / fixed effects :
3. Additive models with or without random effects :
4. Distributional regression (GAMLSS):

with distribution parameters und of the following distributions Gaussian, Lognormal (LGN), Dagum, Singh-Maddala (SM), Generalized Pareto (GP), Generalized Beta [4].

## Data harmonisation for data fusion purposes

However, as already pointed out, these models only rely on the taxpayer data where no information concerning the social disaggregation is available. Therefore, we seek to match data from the largest survey sample in Germany, i.e. the Microcensus, with the extensive administrative taxpayer panel data. We harmonise the income concepts and the observation units from both data sets and appropriately define the same subpopulation of taxpayers.

The tax return and survey data differ in three main aspects: First, the frames of the two data sets are not the same. While the TPP includes only taxpayers (and for the years before 2012 only tax-filers), the MC covers the whole population. Second, the income-sharing units in the tax data are tax units, whereas the MC records individual, family and household income. Last but not least, the income concept in the tax data is the total amount of earnings defined by tax law as opposed to disposable income in the MC. To make the two data sources comparable, we reconcile these differences as follows.

In order to ensure that the frames of both data sources cover the same population, we restrict the MC to the subpopulation of taxpayers, i.e. all non-taxpayers are excluded from the analysis. We construct the income tax units, e.g. single persons or married couples, from the tax data in the MC based on legal status. As household affiliations are unknown in the tax data, the opposite direction is not feasible, i.e. it is impossible to construct MC household compositions in the TPP. Finally, we construct net incomes in the tax data by calculating economic gross incomes from the individual income components and subtracting all taxes, transfers, and social security contributions.

## Relevant data fusion methods

Concerning the data fusion implementation, we match the taxpayer data with the Microcensus by adding the information from some socio-demographic variables observed in the Microcensus (i.e. the donor data) to the taxpayer data (i.e. the recipient data). To meet this objective, we test different data fusion techniques including one traditional matching method (NND) as well as two alternative Statistical Learning approaches:

* Nearest Neighbour Donor (NND): This is a very traditional data fusion algorithm that relies on covariate-based Nearest Neighbour matching. NND algorithms match data on maximal similar observations with regard to some preselected common variables observed in both the TPP and the MC [5]. Usually, a certain distance metric is applied to identify maximal similar observations, e.g. the Gower distance [6].
* Decision trees via recursive binary splitting: The basic idea is that input variables, in this case the common variables, are successively split in optimal (binary) subgroups according to a minimum classification error rate [7]. The classification trees are estimated within the donor data set, i.e. the Microcensus. This model is subsequently used to predict the missing socio-demographic values in the recipient file, i.e. the TPP.
* Decision trees via Random Forest: This approach is related to the decision trees described above but relies on multiple decision trees via bootstrapping. Its advantage compared to conventional trees is the induced variance reduction through decorrelating (avoiding too strong dominance of a highly correlated predictor) [8].

# Results

## Income predictions

We find that the use of structured additive distributional regression enhances predictive performance of income estimates for microsimulations by providing estimates of the full income distribution conditional on the included covariates. The Dagum distribution provides a particularly good fit to the taxpayer panel data.

Applying the income models to the Microcensus data with its less accurate income data, we see that the explanatory power is rising by adding certain socio-demographic variables. For example, by adding education within the linear model, the adjusted R² is rising from 0.08 to 0.24. In case of the distributional regression (GAMLSS) following the Dagum distribution, the R² is rising from 0.22 (without education) to 0.33 (including education).

## Evaluation of data fusion methods

Therefore, applying convincing data fusion methods in order to match the TPP and MC data seems promising. The precise reproduction of the correlation structures between the income variables from the TPP and the socio-demographic variables from the MC is one key objective of our data fusion research, as this is relevant for model-based income estimation. Preliminary results obtained from a simulation study suggest that alternative Statistical Learning methods, such as Random Forest, are able to adequately reproduce the correlations between the income observed in the taxpayer data and the socio-demographic variables observed in the Microcensus. However, these preliminary results need to be further evaluated within a comprehensive simulation study based on the TPP and the MC data situation.

# Conclusions

When using the taxpayer panel data, the income distribution can best be modelled using distributional regression based on the Dagum distribution. Data fusion methods, especially Statistical Learning methods like Random Forest, seems a promising approach to allow for social disaggregation of the incomes and possibly enhance predictive performance of the models. New insights on this research question shall be presented at NTTS.

# References

[1] Bundesministerium für Arbeit und Soziales (BMAS) (2017): Lebenslagen in Deutschland. Der Fünfte Armuts- und Reichtumsbericht der Bundesregierung (August).

[2] Angel, Stefan; Disslbacher, Franziska; Humer, Stefan; Schnetzer, Matthias (2019): What did you really earn last year? Explaining measurement error in survey income data. In Journal of the Royal Statistical Society: Series A 182 (4), pp. 1411–1437. DOI: 10.1111/rssa.12463.

[3] Rässler, S. (2002): Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. New York: Springer-Verlag, Inc.

[4] Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, 23(7):1–46.

[5] D'Orazio, M. / Di Zio, M. / Scanu, M. (2006): Statistical Matching. Theory and Practice. Chichester: John Wiley & Sons, Ltd.

[6] Gower, J. C. (1971): A General Coefficients of Similarity and Some of Its Properties. In: Biometrics, Vol. 27, No. 4 (Dec. 1971), pp. 857-871.

[7] Therneau, T. M. / Atkinson, E. J. / Mayo Foundation (2019): An Introduction to Recursive Partitioning Using the RPART Routines. April 11, 2019, URL: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

[8] James, Gareth / Witten, Daniela / Hastie, Trevor / Tibshirani, Robert (2013): An Introduction to Statistical Learning. New York: Springer Science+Business Media.