Administrative Data and statistical matching
(EU-SILC and Micro-census Environment)

**Keywords:** Statistical Matching, administrative data, environmental conditions, environmental behaviour, household income

# Introduction

The micro-census special programme "Environmental conditions and environmental behaviour" by Statistics Austria contains widespread data material concerning ecological issues. The influence of income on the collected environmental characteristics is commonly assumed but could not be confirmed with the original data sample because the variable “income” is not a direct part of the micro-census survey. If education and employment status were used as approximations, it can be assumed that the income of households is a crucial factor, for example for the purchase of organic products [Baud - Milota, 2017, p 69].

In 2014 a pilot project “environmental conditions / behaviour and income” used statistical matching to add income variables from EU-SILC 2011 to the data of the micro-census environment 2011 [Wegscheider-Pichler, 2014]. This allowed analysing the correlation between environmental behaviour as well as environmental impacts (e.g. noise, dust) with the overall household income of the people interviewed.

In 2018 a follow up with data of the micro-census environment 2015 was conducted. Unlike the previous project 2014, now much of the income information was generated from administrative data, which increased data validity substantially. Thus, the analyses carried out in the previous report on the subject of "environmental justice" could be examined in more detail.

# Methods

To obtain the most accurate income information possible for the sample of the micro-census environment, income variables from administrative data were linked to the sample data. EU-SILC actually generates more than 85% of the income components from administrative data. This applies, for example, to income from employment, unemployment benefits or pensions. These variables from administrative sources (such as payroll data) can also be directly incorporated into the micro-census environment. To ensure compliance with the secrecy guidelines, the administrative data is assigned to the data set of the micro-census Environment by means of an anonymised 28-digit personal key.

For the remaining income components (like income from self-employment or some housing benefits) income information from EU-SILC 2016 were inserted to the data-set of the micro-census environment 2015 by statistical matching. Some focus was laid on the selection of the variables used to link the two data files (= connecting variables). The comparability and homogeneity of the variables is granted by using a so called “Rich Frame” of variables, conducted by the methodological department for sample surveys of Statistics Austria [for the relevance of harmonised variables see also Eurostat, 2013, p.13].

With the income components of administrative data combined with the income components from EU-SILC, gained by statistical matching, a total household income for 2015 was calculated for further use of analytics evaluations.

## Statistical matching process

The statistical matching of the income components was carried out by a random forest model [see Briemann, 2001]. This is a machine-learning algorithm where multiple (in this case 1500) decision trees are created at random. Each of these decision trees provides an estimate of income, and these estimates are then averaged. The decision trees are created from random samples of data from EU-SILC. Model-based methods, such as the random forest algorithm, use a model estimated on the donor record, transfers it to the recipient record, and estimates values of the searched variables based on it.

In a first run, a regression model was employed for the creation of the decision trees, which uses the logarithm of the household income on the basis of the connecting variables.

During data evaluation, it turned out that, although the overall estimation of income by this method is satisfying, but especially in the lower income deciles there is a large deviation at distribution level from the actual income from EU-SILC. Therefore in a second run, only the households at-risk-of-poverty were estimated, (as a dichotomous variable, which indicates whether the equivalent household income is below the at-risk-of-poverty threshold). This logical variable was then used as an additional regressor to create the decision trees. In this way the rate of persons at risk of poverty could be well transferred to the imputed values, which substantially improved the results.

# Preliminary Results



Figure 1. Assessment of the subjective quality of life by income (terciles)

A distinction of responses by household income revealed significant differences in the assessment of quality of life: While 59.6% of respondents with high household income defined their quality of life as high, this was true for only 38.5% of those with low household income. Persons with low household income stated to nearly 5.3%, that their quality of life was not so good or bad; this information was given only by 1.3% of the group with a high household income. Persons with medium household income rated their quality of life to 48.2% as very good, 2.4% as poor or very poor.

Further findings:

Persons with low household income were more affected by noise disturbance, then high household income earners.

People with high household income answered more frequently to buy "often" or "sometimes" organic food, as people in the middle or low income group.

People with a low household income were least likely to use their car for daily commuting, medium and high income had significant more daily use of car.

# Conclusions

The use of administrative data to generate income components into an existing data set improves the data quality compared to a sole statistical matching. But not all needed data information (here for income components) is available in administrative sources (or not available in time). The advantages of statistical matching are the cost reduction and the reduced burden on respondents [Eurostat, 2013] as well as the timeliness. Nevertheless the variables are synthetically generated “statistical twins” and not actual observations. The values obtained can thus be distorted. They depend highly on the “connecting variables”.

A mixed use of both methods (administrative and matching) therefore seems to be the most useful way to generate a variable as the total household income for an existing data set. This way the advantages of both data generating possibilities can add up and improve the data quality.

For this project, some expectations, described in relevant literature, were confirmed by the data obtained: the assessment of life quality showed a clear correlation with income. Also the connection between environmental behaviour and income has been confirmed for several aspects. The results followed largely the findings of the pilot study in 2014 and thus contribute to a further verification of results.

# References

1. S. Baud and E. Milota, Umweltbedingungen, Umweltverhalten 2015, Ergebnisse des Mikrozensus, Statistics Austria publication, (2017), Vienna.
2. A. Wegscheider-Pichler, Umweltbetroffenheit und –verhalten von Personengruppen abhängig von Einkommen und Kaufkraft, Statistics Austria publication, (2014), Vienna.
3. Eurostat, Statistical matching: a model based approach for data integration, Methodologies and Working papers (2013), ISBN 978-92-79-30355-2 Luxembourg.
4. L. Briemann, Ramdom forests. Machine Learning, 45:5-32, 2001, http://dx.doi.org/10.1023/ A:1010933404324.