Statistical Matching of EU-SILC and MC Environment for Analysing Environmental Conditions and Behaviour in Dependence of Household Income

Alexandra Wegscheider-Pichler  
Statistics Austria, Guglgasse 13, 1110 Vienna, AUT

Alois Haslinger  
Statistics Austria, Guglgasse 13, 1110 Vienna, AUT

The project “environmental conditions and behaviour depending on income and purchasing power” analysed the correlation between environmental impacts (e.g. noise, dust) or environmental behaviour (purchase of organic food, usage of public transport) and household income. Statistical matching was used to add income variables from EU-SILC to the the Micro Census (MC) environment dataset 2011. For each respondent of the MC environment a donor from the data from EU-SILC with minimal distance was found. Several matching options were tested to find the best option to link the variable “total disposable household income” to the environmental data. The matching options differed in the used distance function, which depended on the selected matching variables and their weight. The analysis provided an insight into advantages or problems of the method of statistical matching. Analyses of the environmental responses by household income terciles revealed significant differences in the assessment of quality of life, disturbance by noise, purchase of ecological products or the use of public transport.

Keywords: statistical matching, distance function, environmental conditions, environmental behavior

Introduction

The project “environmental conditions / behaviour in dependence of household income” allowed for the first time to analyse how different household income groups in Austria are affected by environmental conditions or differ in their environmental behavior. This by using official data sources [project report Wegscheider-Pichler, 2014.]

The Micro Census "Environmental conditions and environmental behaviour" (MC environment), which is collected together with the Micro Census Labour Force Survey - MC LFS, [Statistics Austria 2013], contains extensive data concerning ecological issues. These environmental characteristics are regularly analysed by

Corresponding author: Alexandra Wegscheider-Pichler, Mag., Unit Analysis, Statistics Austria. Alois Haslinger. DI, Methods, Statistics Austria.
different socio-demographic factors like gender, education or age. The socio-demographic variables are collected within the MC LFS and used for analysis of the MC environment as well. An influence of income on environmental behavior and burden is commonly expected but could not be confirmed yet as the variable “household income” is not part of the MC survey. If education and employment status are 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, 2011, p.77].

Statistical matching allowed to add income variables from EU-SILC - EU Statistics on Income and Living Conditions [Statistics Austria 2012] to the data of the MC environment 2011. This enabled to analyse the correlation between environmental behavior (e.g. use of public transport) as well as environmental impacts (e.g. noise, dust) with overall household income of the subjects interviewed.

One focus of the project was the comprehensive presentation of the used method for the statistical matching process. Advantages or problems of the method of statistical matching are described.

**Method**

Statistical matching is “a model-based approach for providing joint information on variables and indicators collected through multiple sources (surveys drawn from the same population)” [Eurostat, 2013, p10]. Although the units in the concerned data sets refer to the same population, they are not overlapping. This means that synthetically statistical information is formed, based on variables from at least two different data sources. One or more variables in the donor data set (“matched variables”) are integrated in the recipient data set. For each observation case in the recipient data set a so-called statistical twin in the donor data set has to be found. For this process so called “matching variables” are used. These matching variables consist of pre-selected socio-demographic variables, e.g. gender, age, employment status or residential area of the respondents. They must be correlated with the matched variable and be present in both data files.

The advantage of this approach is to generate additional analysis capabilities from existing data sources (in this example, EU-SILC and MC environment) without much additional cost or respondent burden [Eurostat 2013, p10]. Nevertheless, values obtained by this method do not represent real observations. They depend strongly on the selected matching variables to connect the data files and their relevance for the matched variable.

In the presented study income variables from EU-SILC 2011 were inserted to the data-set of the MC environment 2011 by statistical matching. Great attention was put on the selection and adjustment of the matching variables used to link the two data files. An extensive review of the different matching options was conducted in the project.

**Selection and Adjustment of Matching Variables**

Comparability and homogeneity of the variables used are essential for the quality of the statistical
matching process [see Eurostat, 2013, p.13]. The matching variables (see table 1) had to be correlated with the matched variable “household income” and had to be present in both data files in a way that allowed a consistent use. Table 1 gives an overview of the used matching variables and their correlation with the matched variable “total disposable household income” in EU-SILC. All characteristics are significantly correlated (at a level of 0.000) with the household income, the only exception being “provinces”. Since the latter is essential for sampling and extrapolation of both data sets, it was - in this first approach - left in the model.

The matching variables were present in both data sets. However, configuration and definition between micro-census and EU-SILC not always coincided in detail. An extensive variables reconfiguration in both data files was necessary to have the variables in a harmonised and coherent form.

Table 1

Matching Variables and Correlation with total disposable household income (matched variable)

| Matching Variables – number | Matching Variables – name                          | Correlation coefficient¹) with “total disposable household income“ |
|-----------------------------|----------------------------------------------------|---------------------------------------------------------------|
| V1                          | Gender                                             | -0.085                                                        |
| V2                          | Age                                                | -0.178                                                        |
| V3                          | Occupational status (Type of activity)             | 0.320                                                         |
| V4                          | Economically active or not                         | -0.286                                                        |
| V5                          | Working part-time / full-time                      | 0.284                                                         |
| V6                          | Highest level of education                         | 0.254                                                         |
| V7                          | Provinces (NUTS 2)                                 | 0.011                                                         |
| V8                          | Eurostat- degree of urbanization                   | 0.047                                                         |
| V9                          | Household size                                     | 0.421                                                         |
| V10                         | Number of apartments in the building               | -0.180                                                        |
| V11                         | Ownership or not to the apartment                  | -0.345                                                        |
| V12                         | Country of birth                                   | -0.125                                                        |
| V13                         | Citizenship                                        | -0.098                                                        |
| V14                         | Income of the employed (net)                       | 0.397                                                         |
| V14a (alternative to V14)   | Household income of the employed (net)             | 0.714                                                         |

Source: EU-SILC 2011. ¹) The nominal and ordinal scaled variables (V1, V3, V8, V11-V13) were correlated by Spearman’s-Roh with household income. For the metric scaled variables age (V2), household size (V9), Number of apartments in the building (V10) and income of the employed (V14 and V14a) a Pearson-correlation was carried out.

Used income variables (see table 2)

For this project total household income (instead of personal income) was considered. Decisions concerning e.g. the purchase of organic food or the affection by noise should be influenced by the total household income and not by the personal income of a single household member. Therefore as matched variable the “total disposable household income” was defined. The variable was taken from the EU-SILC 2011 dataset to be inserted in the MC data file with statistical matching. Additionally “net income of the employed” was used for two purposes: 1) to improve the matching process itself and 2) to evaluate the matching process. See table 2 for the income variables in use:
Table 2

Income variables in EU-SILC and MC LFS

| MC LFS (Micro Census Labour Force Survey) 2011 | “EU-SILC (EU-Statistics on Income and Living Conditions) 2011” |
|-----------------------------------------------|------------------------------------------------------------|
| Included income variables (from administrative data) | Included income variables (survey data) |
| Income of the employed (net) | Household net income of the employed |
| Matching variable V14 | Matching variable V14a |
| Matching variable V14a | Matching variable V14 |
| Matching variable V14a | Matched variable |
| Total disposable household income |

Source: Statistics Austria.

As already mentioned the MC does not collect variables on income directly but the “income of the employed (net)” was enclosed subsequently through administrative data in the MC LFS [see Baierl et al., 2011]. The “income of the employed” provides information for individuals with employment in the survey period (about 50% of respondents). However, it does not cover further important income components, such as unemployment compensation, pensions or aids. The item therefore is not suitable for analysing environmental conditions and environmental behaviour in dependence of income. However, it could be included as matching variable, which led to a distinct improvement of the data linking. Furthermore the matching options were evaluated including or not including this variable.

The variable “household net income of the employed” (net income of all employed persons of a household aggregated) was calculated in both original data files MC LFS and EU-SILC. Thus for every single person of the household the variable “household net income” was assigned. This variable was used for the matching process as well.

The variable “total disposable household income” consists of various components. Income from employment or self-employment of all household members, property income, unemployment benefits, pensions or allowances may be essential components of the total disposable income of a household. The total disposable household income can not be calculated from the data of the MC LFS. However all this components are available through the data set of EU-SILC as being the most relevant data set on income and social living conditions in the European Union. The income variable represents the “matched variable”.

Overview Statistical Matching Options

For each statistical unit of the recipient data set a so-called statistical twin in the donor data set was selected, which was as similar as possible concerning the used matching variables. For the statistical matching a distance-based procedure was used. For each respondent of the MC environment a donor from the data from EU-SILC with minimal distance was assigned. In case of several donors with the same distance one was randomly selected. Each respondent from EU-SILC was accepted only once as donor for the MC. This was feasible as in 2011 the sample of EU-SILC comprised of nearly 11,500 cases and the sample of MC
environment being considerably smaller with round 7,000 cases.

Five different matching options were computed to find the best option to link the variable “total disposable household income”. They differ according to the used distance function which depends on the selected matching variables and the used weight (see table 3).

Table 3
Statistical matching options

| Used matching variables | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 |
|-------------------------|---------|---------|---------|---------|---------|
| Defined as nominal      | V1, V3 - V13 | V1, V3 - V13 | V1, V3 - V13 | V1, V3 - V13 | V1, V3 - V13 |
| Defined as metric       | V2       | V2, V14 | V2, V14 | V2, V14a | V2, V14a |
| Differend weight        | no       | no     | yes    | yes     | yes     |

Source: Statistics Austria.

Statistical matching option 1

Data matching with the matching variables V1 - V13 (see table 1), all the features have the same weight in the distance function. The variable income of the employed (V14) is not considered. The distance between a data set of the Micro Census LFS and EU-SILC is calculated as:

\[ \sum_{i=1,3,\ldots,13} 1 \cdot d(v_i) + |z_{2,MC} - z_{2,SILC}| \]

where \( d(v_i)=0 \), if the characteristics of the feature \( v_i \) are identical in MC respectively EU-SILC and 1 otherwise.

\( z_{2,\ldots} \) denotes the “1” standardized variance of the variable “age”.

Statistical matching option 2

As option 1 but additionally income of the employed (V14) is considered. The distance between a data set of the Micro Census LFS and EU-SILC is calculated as:

\[ \sum_{i=1,3,\ldots,13} 1 \cdot d(v_i) + |z_{2,MC} - z_{2,SILC}| + |z_{14,MC} - z_{14,SILC}| \]

where \( d(v_i)=0 \), if the characteristics of the feature \( v_i \) are identical in MC respectively EU-SILC and 1 otherwise.

\( z_{2,\ldots} \) denotes the “1” standardized variance of the variable “age”.

\( z_{14,\ldots} \) denotes the “1” standardized variance of the numeric variable “income of the employed”

Statistical matching option 3

Data matching with the matching variables V1 - V14 (see table 1), the features got different weight in the distance function. This means that e.g. household size could receive a higher weights than gender in the distance function.

For weighting the correlation coefficients of the socio-demographic variables with the total household
Income (see Table 1) were used. Tertiles of this correlation coefficients were calculated and the matching variables were assigned to the three groups. “Occupational status (type of activity) (V3)”, “household size (V9)”, “ownership or not to the apartment (V11)” and “income of the employed (net) (V14/V14a)” formed the closest correlation with household income and got a triple weight in the distance function. The variables “age (V2)”, “economically active or not (V4)”, “working part-time / full-time (V5)”, “highest level of education (V6)” and “number of apartments in the building (V10)” were weighted by factor two. All other variables (V1, V7, V8, V12, V13) got the factor 1 for weight in the distance function.

Distance between a data set of the Micro Census LFS and EU-SILC is calculated accordingly:

\[
\sum_{i=1,7,8,12,13} d(v_i) + \sum_{i=4,5,6,10} 2 d(v_i) + \sum_{i=3,9,11} 3 d(v_i) + 2|z_{2,MC} - z_{2,SILC}| + 3|z_{14,MC} - z_{14,SILC}|
\]

where \(d(v_i) = 0\), if the characteristics of the feature \(v_i\) are identic in MC respectively EU-SILC and 1 otherwise.

\(z_{2,...}\) denotes the “1” standardized variance of the variable “age”.

\(z_{14,...}\) denotes the “1” standardized variance of the numeric variable “income of the employed”

### Statistical matching option 4

Data matching as in option 3, but the variables “highest level of education”, “number of apartments in the building” and “household size” are treated as numeric values additionally. The distance between a data set of the Micro Census LFS and EU-SILC is calculated as:

\[
\sum_{i=1,7,8,12,13} d(v_i) + \sum_{i=4,5,6,10} 2 d(v_i) + \sum_{i=3,11} 3 d(v_i) + 2 \sum_{i=2,6,10} |z_{i,MC} - z_{i,SILC}| + 3 \sum_{i=9,14} |z_{i,MC} - z_{i,SILC}|
\]

where \(d(v_i) = 0\), if the characteristics of the feature \(v_i\) are identic in MC respectively EU-SILC and 1 otherwise.

\(z_{2,...}\) denotes the “1” standardized variance of the variable “age”.

\(z_{6,...}\) denotes the “1” standardized variance of the variable “highest level of education”.

\(z_{9,...}\) denotes the “1” standardized variance of the variable “household size”.

\(z_{10,...}\) denotes the “1” standardized variance of the variable “number of apartments in the building” and

\(z_{14,...}\) denotes the “1” standardized variance of the numeric variable “income of the employed”

### Statistical matching option 5

Data matching with the matching variables V1 - V14a (see Table 1), the features again got different weight in the distance function. The variable V14a “household income of the employed” replaces V14. This means that for each person in the two data sets a household income was calculated through aggregation of incomes of every employed person living in the same household. “Highest level of education” and “number of apartments in the building” are treated as nominal values.

The distance between a data set of the Micro Census LFS and EU-SILC is calculated as:
Statistical Matching of EU-SILC and MC Environment for Analysing Environmental Conditions and Behaviour in Dependence of Household Income

\[
\sum_{i=1,7,8,12,13} d(v_i) + \sum_{i=4,5,6,10} 2 \cdot d(v_i) + \sum_{i=3,11} 3 \cdot d(v_i) + 2 \left| z_{2,MC} - z_{2,SILC} \right| + 3 \sum_{i=9,14} d(z_{i,MC} - z_{i,SILC})
\]

where \(d(v_i)=0\), if the characteristics of the feature \(v_i\) are identical in MC respectively EU-SILC and 1 otherwise.

\(z_{2,\ldots}\) denotes the “1” standardized variance of the variable “age”.

\(z_{9,\ldots}\) denotes the “1” standardized variance of the variable “household size”.

\(z_{14a,\ldots}\) denotes the “1” standardized variance of the numeric variable “household income of the employed”.

**Evaluation of the Matching Process**

Statistical matching imputed income information from EU-SILC to the data set MC environment. However, the matching procedure itself delivers no information about the quality of imputation. Data validation is therefore an important supplement to the matching process. Asmah [2010] distinguishes between external and internal evaluation.

External evaluation is performed by analyzing the recipient data set with the matched values. The methodological approach is verified if the analytic results correspond with results from a profound literature research.

Internal evaluation checks whether the imputed values in MC environment reflect the information of the donor data set. Following, some results of the internal evaluation are described.

As mentioned the “total disposable household income” was determined as matched variable. It was seen as most suitable to jointly analyse environmental conditions and environmental behavior of groups of people with different income. The main objective of the evaluation, therefore, was to find the best way to impute the variable “total disposable household income” from EU-SILC into the data set of MC environment.

The matched variable "total disposable household income" (in options 1 – 5) was correlated to the aggregate MC variables V14 “income of the employed (net)” and V14a "household net income of the employed". The correlation coefficients are displayed in table 4, for the different income variables see also table 2. Option 3 and option 5 show the highest fit with V14 “income of the employed (net)” at the individual level.

For V14a "household net income of the employed" option 5 with its household perspective achieved by far the highest correlation coefficient (0.642). For evaluation it has to be considered that no unemployment compensation, pensions or social transfers (such as family allowances) are included in the “household net income of the employed”. In the donor data set EU-SILC V14a "household net income of the employed" and the “total disposable household income” correlate just slightly higher with a correlation coefficient of 0.713. The - by statistical matching - achieved correlation should thus be regarded as high.

In addition, an OLS regression was performed on the basis of the donor data set EU-SILC: with household income as the dependent variable and the independent variables V1 – V14a. The beta coefficients from this
regression (not-standardized) were used to calculate a “Regression household income” in the recipient data set MC environment. Then again a correlation was carried out, comparing the “Regression household income” with the matched variable "Total disposable household income” in options 1 – 5 (see table 5).

Option 5 of the statistical matching process is correlated highest with the MC Regression household income. This depends on the fact that option 5 takes a household perspective. A coefficient of 0.720 suggests a high interrelation.

Table 4

| EU-SILC Total disposable household income = matched variable | Correlation with MC “Income of the employed (net) (V14)” | Correlation with MC "Household net income of the employed” (V14a) |
|-----------------------------------------------------------|----------------------------------------------------------|---------------------------------------------------------------|
| Option 1                                                  | .209**                                                   | .295**                                                        |
| Option 2                                                  | .315**                                                   | .368**                                                        |
| Option 3                                                  | .329**                                                   | .404**                                                        |
| Option 4                                                  | .319**                                                   | .401**                                                        |
| **Option 5**                                              | .328**                                                   | .642**                                                        |

Source: EU-SILC 2011 and MC Environment 2011. All characteristics of table 4 are significantly correlated (at a level of 0.000, Pearson).

Table 5

| EU-SILC Total disposable household income = matched variable | Correlation with MC regression household income          |
|-----------------------------------------------------------|----------------------------------------------------------|
| Option 1                                                  | .380**                                                   |
| Option 2                                                  | .450**                                                   |
| Option 3                                                  | .501**                                                   |
| Option 4                                                  | .498**                                                   |
| **Option 5**                                              | .720**                                                   |

Source: EU-SILC 2011 and MC Environment 2011. All characteristics of table 4 are significantly correlated (at a level of 0.000, Pearson).

The results of the evaluation process confirm the expectations: option 1 (with no income variable used for matching procedure) displays significantly lower correlations than the options with income variable in use. Option 5 has the highest correlation with both the MC "Household net income of the employed" (table 4) as well as the MC regression household income (table 5). Consequently, for data analysis the statistical matching option 5 was chosen, which takes the household perspective by using the "household net income of the employed (V14a)” as additional matching variable.

**Brief Comparison: Statistical Matching Versus Regression**

As an alternative to statistical matching in order to include the “total disposable household income” of
EU-SILC into the MC environment data set, the “regression household income” – as described above – could have been used for analyzing the environmental data. So what are the advantages of the method statistical matching?

Table 6 displays the distribution of household income according to statistical matching (option 5) versus regression analysis: demonstrating the main difference between the two methods: The edges of the distribution are not presented by regression. While by statistical matching outliers or extreme values of the distribution are mapped (perhaps assigned incorrectly), regression analysis neglects very small or very high values. The displayed values according to statistical matching therefore better correspond with the original distribution in EU-SILC.

It should be noted that the above-described disadvantage of statistical matching, not being based on actual observations, also applies to the calculation of the regression household income. Again, the values obtained via regression are only as good as the independent variables used.

Table 6

Comparison of income data by statistical matching and regression

| Statistical matching - Total disposable household income | Regression household income |
|--------------------------------------------------------|-----------------------------|

Source: EU-SILC 2011 and MC Environment 2011.

Data Analysis

As mentioned an external evaluation of the matching process is performed by analyzing the recipient data set with the matched values. In the best case these outcomes can be verified by literature results. Following a selection of the analytic results is listed.

A distinction of responses by household income revealed significant differences in the assessment of quality of life: While 56.1% of respondents with high household income defined their quality of life as high, this was true for only 36.9% of those with low household income. Persons with low household income stated to
nearly 7%, that their quality of life was poor or very poor; this information was reported by only 1.5% of the group with a high household income. 44.9% of persons with medium household income rated their quality of life as very good, 2.5% as poor or very poor.

The exposure to noise differs also depending on the total disposable household: 43.6% persons with low household income were affected by noise, while this applied to only 35.7% of high household income earners. The exposure to odor or dust, was highest for the lowest income tercile as well.

Some differences according to income were found in environmental behavior: Respondents with high household income reported more often to buy organic food "often" or "sometimes" than that of the medium or low income group. Environmentally friendly mobility (frequent use of public transport / rare use of car), however, decreased with increasing income. People with a low household income on the other hand were least likely to use their car for daily commuting (26.3%), medium and high income groups reported more than 40% daily use of car.

Most of this analytic results can be verified by other literature sources:

A study by the "Socio-Economic Panel" Germany, for example, displayed a correlation between household income and exposure to air pollution or noise: with decreasing income the exposure increased [see Bolte - Mielck, 2004, p.139ff]. Bialas [2010, p.6] analysed that the income of buyers from organic food is above the average income and the expenditures on transport increased with growing household income according to the household budget survey of Statistics Austria [Kronsteiner-Mann, 2012 p.90].

**Conclusion**

Statistical matching still is a relatively new model-based approach to combine statistical information from at least two sources. Some advantages of the method are the cost reduction and the reduced burden on respondents [Eurostat, 2013]. Nevertheless, the variables are synthetically generated “statistical twins” and not actual observations. The values obtained can thus be distorted because of the matching procedure.

The MC variable "income of the employed", which is subsequently supplied from administrative data, leads to a significant improvement in the matching process. Option 1 of statistical matching - without this variable - comprises for all correlations of table 4 and 5 the lowest correlation coefficient. This is based on the fact that the “income of the employed” constitutes a substantial part of the total disposable household income. This means that a matching process based purely on socio-demographic characteristics such as gender or age - similar option 1 - has to manage with a correspondingly lower correlation as displayed in the actual study.

The statistical matching process depends highly on the “matching variables” used. Much emphasis has to be laid on the comparison of the variables and the data evaluation - than data generation with statistical matching can lead to interesting results.

In this project, the data obtained confirmed some content expectations gained from relevant literature:
Environmental conditions as noise and dust showed a correlation with income. Also the connection between environmental behavior and income has been confirmed for several aspects as using cars versus public transport.

References

Asmah, S. (2010) Statistical Matching – Multiple Imputation und Datenfusion am Beispiel von Daten zu Religiosität und Gesundheit, Bachelorarbeit am Ladislaus von Bortkiewics Chair of Statistics der Wirtschaftswissenschaftlichen Fakultät der Humboldt Universität zu Berlin.

Baierl, A., Gumprecht, D., Gumprecht, N. (2011) Monatliches Nettoeinkommen im Mikrozensus - Konzept, Statistische Nachrichten 7/201: 596-612, Vienna.

Baud, S. and Milota, E. (2013) Umweltbedingungen, Umweltverhalten 2011, Ergebnisse des Mikrozensus, Statistics Austria publication, Vienna. http://www.statistik.at/web_de/statistiken/energie_und_umwelt/umwelt/umweltbedingungen_verhalten/index.html

Bialas, E. (2008) Von der ökologischen Innovation zum Massenmarkt – biologisch-dynamisch erzeugte Lebensmittel, GRIN Verlag ISBN 978-3-640-71015-7, Norderstedt.

Bolte, G., Mielck, A. (Hrsg.) (2004), Umweltgerechtigkeit, die soziale Verteilung von Umweltbelastungen, Juventa Verlag, ISBN 3-7799-1141-8, Weinheim und München.

Eurostat (2013), Statistical matching: a model based approach for data integration, Methodologies and Working papers, ISBN 978-92-79-30355-2 Luxembourg.

Kronsteiner-Mann, Ch. (2012) Verbrauchs ausgaben – Sozialstatistische Ergebnisse der Konsumerhebung, Statistics Austria publication, Vienna.

Statistics Austria (2012), Standard-Documentation – Metainformation EU-SILC 2010, revision: 30.09.2012, http://www.statistik.at/web_de/statistiken/soziales/index.html

Statistics Austria (2013), Standard-Documentation – Metainformation Micro Census Labour Force Survey, revision: 16.05.2013, http://www.statistik.at/web_de/statistiken/arbeitsmarkt/erwerbsstatus/index.html

Wegscheider-Pichler, Umweltbetroffenheit und Umweltverhalten von Personengruppen abhängig von Einkommen und Kaufkraft, Statistics Austria publication, Vienna.