Junge, Henrike

Research Report

From gross to net wages in German administrative data sets

DIW Data Documentation, No. 89

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Junge, Henrike (2017) : From gross to net wages in German administrative data sets, DIW Data Documentation, No. 89, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/161635

DIW BERLIN

89

Data
Documentation

Deutsches Institut für Wirtschaftsforschung 2017

From Gross to Net Wages in German Administrative Data Sets

Henrike Junge

IMPRESSUM

DIW Berlin
Deutsches Institut für Wirtschaftsforschung
Mohrenstr. 58
10117 Berlin
Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
www.diw.de
ISSN 1861-1532

Data Documentation 89

Henrike Junge[1]

From Gross to Net Wages in German Administrative Data Sets

Berlin, May 2017

[1]Financial support by the DFG is gratefully acknowledged (grant HA 7464/1-1). The basis of this paper is the German Socio-Economic Panel (SOEP) and the weakly anonymous Sample of Integrated Labour Market Biographies (SIAB) 1975 - 2014. The SOEP data is provided by the German Institute for Economic Research (DIW Berlin). The SIAB data were accessed on-site at the Research Data Centre (FDZ) of the DIW Berlin and via remote data access at the FDZ provided as part of project fdz 697. Special thanks to Luke Haywood who provided the necessary directions to produce this data documentation.
Contact: Henrike Junge, Guest Researcher DIW Berlin, Department Public Economics, rike.junge@t-online.de

Contents

List of Tables

1 Introduction

This data documentation describes selected ways of how to analyze *net* wages in the weakly anonymous Sample of Integrated Labour Market Biographies 1975-2010 (SIAB) or other administrative data sets provided by the Institute for Employment Research (IAB). Among other things, these data sets provide information about dependently employed persons in Germany and their gross wages over time. Due to the large number of included observations, the data sets are often used in empirical studies of the German labor market. However, net wages are not reported in the data sets and important income tax-relevant characteristics which would allow for a simulation of individual tax burdens are not included. Therefore, labor market studies based on IAB data that want to take effects of (income) taxation into account must impute the tax burdens as a first step of their analyses. This data documentation discusses different ways to perform this imputation and compares their ability to produce reliable predictions.

2 Defining Net Wages

Net wages are defined as the part of the gross labor income from dependent work that remains after taxes are deducted. In Germany in particular, the calculation of net wages based on gross wages requires knowledge about social security contributions, income taxes, and solidarity taxes.[2]

2.1 Social Security Contributions

Social security contributions in Germany mainly encompass contributions to old-age, health, unemployment, and long-term care insurance (the latter since 1996) and are formally paid to an approximate equal share by employers and employees. They are deducted directly at the source. Individual social security contributions can be easily calculated based on gross wages by using statutorily fixed contribution rates and contribution assessment ceilings. Since 2003, there are special rules for the determination of social security contributions for employees holding a so-called *midi job*, i.e. a job with a monthly wage above the mini job threshold but remaining below 800 Euro. Persons holding a *mini* job (with monthly earnings below a certain threshold (e.g. 400 Euro between 2003 and 2006) do not have to pay contributions to the statutory insurance system at all.

2.2 Taxes

In addition to the deduction of social security contributions, gross labor incomes from dependently employed persons in Germany are further reduced by taxes. These mainly

[2] Additional taxes, like church taxes, are exempted from the analysis as they are paid voluntarily.

encompass income taxes as well as solidarity taxes.

2.2.1 Income Taxes

Wages in Germany are subject to a wage payroll tax (the so-called "Lohnsteuer" with progressive tax rates) that is deducted at source. However, the payroll tax on labor income only constitutes an initial payment to the general personal income tax. Until 2009, the income tax system in Germany was based on the comprehensive income taxation paradigm.[3] This principle implies that income from different sources[4] is summed up and taxed without differentiating between the sources. Consequently, the tax burden on the factor labor is influenced by the existence of other income sources. Hence, in order to determine the income tax burden on labor, it is not sufficient to calculate payroll taxes on wages. Instead, the share of the overall income tax burden borne by labor income has to be determined. In the simulation used in this analysis, it is assumed that the total income tax burden is allocated across the different gross income sources according to their shares in total gross income. The latter is defined as the plain sum of all income sources without taking any deductions or tax-free amounts into account. For example, consider a case where gross labor income makes up 80% of the overall gross income of an individual. According to the aforementioned assumption, 80% of the overall income tax burden of this individual is borne by labor income and 20% of it is borne by the remaining income sources. As a result, the same average income tax rate applies to all income sources. The approach is, hence, based on the assumption that the formally prescribed comprehensive income taxation paradigm expresses the actual taxation of different income sources.

In addition to the comprehensive income taxation paradigm, the German income tax system is characterized by various additional features. In particular, the income tax function is not applied to the plain sum of income but to the so-called taxable income. Determining this individual taxable income is a complex task as it depends on many factors, including various household characteristics. In particular, the German taxation system incorporates a great amount of tax allowances and special taxation rules for certain groups (e.g. (single) parents, elderly people, or self-employed) and allows for deducting special and extraordinary expenses like those of a provident nature. Consequently, the resulting taxable incomes are usually much smaller than overall gross incomes.

Once taxable income is determined, it is inserted into the income tax function that yields the individual tax burden. The German tax schedule has a linear-progressive functional form and incorporates four to five tariff zones for the period under consideration.[5] Income

[3] Since 2009, capital income is excluded from the general personal income taxation and taxed separately.

[4] Income from all sources encompass income from agriculture and forestry, from business operations, from self-employed work, from employed work, from capital (until 2009), from letting property and miscellaneous income.

[5] Since 2007, there is a fifth zone for especially large incomes.

in the first tax bracket is completely exempt from income tax payments. The tax schedule implies marginal and average tax rates that increase continuously with the taxable income for most incomes.[6]

A further important particularity related to the tax function comes from the possibility of joint income taxation for married couples ('Ehegattensplitting"). According to this rule, married couples can decide to be assessed jointly for income taxation. In this case, joint taxable income is calculated including advantages with respect to certain possible deductions (expenses of a provident nature). The joint taxable income is subsequently divided by two in order to determine the individual average and marginal tax rate that is, as a result, identical for both partners. The procedure has especially large effects in partnerships with one main earner. In this situation, the partner who contributes the major part to the joint income ends up with a tax rate that is lower in comparison to the one he/she would face in case of individual taxation. The partner contributing the minor part to the joint income, contrarily, is usually taxed at a higher rate in comparison to the rate that would result from individual taxation. For the analysis, it is assumed that married couples always decide to be taxed jointly.

An important result of the various rules and exemptions related to the determination of individual taxable incomes is reflected by individual tax burdens that can be very different for persons with comparable labor incomes depending on several personal and household characteristics.

2.2.2 Solidarity taxes

Since the unification of Germany, an additional tax has been levied on all incomes to support the less developed East German regions. This so-called "solidarity tax" is determined as a fixed fraction of the overall income tax burden whereby individuals with income tax burdens below a certain threshold are exempt from the additional tax payments. Hence, the calculation of solidarity taxes is based on the same principles as the calculation of incomes taxes.

3 Problems of Determining Net Wages in IAB Data

Next to daily gross wages, the SIAB and other data sets provided by the IAB contain information about some personal characteristics (gender, age, education) and several variables related to firm characteristics. However, net wages are not reported in the SIAB. Making a few standard assumptions, one component of the tax burden on labor, namely social security contributions, can be calculated quite easily given only individual gross

[6]For taxable incomes above the top income in the tax function, the marginal tax rate stays constant.

wages.[7] The income tax component of the total tax burden on labor as well as solidarity taxes, however, cannot be simulated directly in the data set as important tax-relevant information like family status and the existence of, and number of, children or different sources of income in addition to labor income are not reported in the SIAB. Consequently, for studies aiming to analyze net wages within this data set, we require a function to transform observed gross wages y into net wages y^*. For this purpose, various alternatives exist of which three candidate functions (g, f, and f')

1. Approaches relying on labor income only

   - **Approach 1a:** Use the legal tax code for singles plus a approximated "taxable labor income" in order to impute missing income tax rates:

   $$y_{it}^* = g(y_{it}^{tax}), \tag{1}$$

   with y_{it}^* being the individual net wage at time t, $g()$ being the legal tax code for singles and y_{it}^{tax} denoting an approximate level of individual taxable labor income at time t.

   - **Approach 1b:** Build a simple prediction function within an auxiliary data set that contains individual tax burdens or allows for a simulation of them; as predictor variables use only labor income and time variables; apply the prediction function to the SIAB (regression imputation):

   $$y_{it}^* = f(y_{it}, t) \tag{2}$$

   with $f()$ being a function predicted from the auxiliary data set depending on individual gross wage y_{it} and time t.

2. Approach relying on labor income plus additional variables

   - **Approach 2:** Build a more complex prediction function within an auxiliary data set that contains individual tax burdens or allows for a simulation of them; as predictor variables use labor income and time variables, but also other SIAB variables like gender, age, education, occupational status, industry and interactions, in order to better approximate tax-relevant household characteristics; apply the prediction function to the SIAB (regression imputation):

   $$y_{it}^* = f'(y_{it}, t, X_{it}) \tag{3}$$

   with $f'()$ being a function predicted from the auxiliary data set depending on the individual gross wage y_{it}, time t and a vector of additional predictor variables X_{it}.

[7] Standard assumptions are that all persons make use of the statutory health insurance (excluding private insurances). Some exceptional rules, i.e. for childless individuals are ignored (see e.g. Gunselmann 2014).

4 Approaches to Impute Missing Income Tax Rates

In the following, the different approaches to impute missing (average) income tax rates on labor income outlined above are discussed in more detail. As an auxiliary data set, the German Socio-Economic Panel (SOEP) is used. It contains most tax-relevant information and stems from the same population as the IAB data. Within this data set, individual income tax burdens are simulated for each individual using a simple micro-simulation program.[8]

Approach 1a

For the imputation of income tax burdens on labor income in the SIAB according to Approach 1a, a "taxable labor income" is needed that is inserted in the legal tax code for singles. The legal taxable income (based on all income sources) is determined by subtracting individual deductions (e.g. expenses of a provident nature, professional expenses, ...) from the individual gross income (in case of married couples, a joint taxable income is determined). Following from this, the level of taxable *labor* income should be determined by subtracting a certain amount of deductions from the gross wage. In order to determine this amount of deductions, the following **assumption** is made:

All types of deductions, i.e. the difference between gross income and taxable income, are allocated among the sources of gross income according to their share in total income. For example, if labor income makes up 80% of the total gross income of an individual, 80% of the overall deductions that can be written off by the individual are attributed to labor income according to this assumption. In case of joint income taxation for married couples, the joint amount of deductions is allocated among the joint sources of gross income according to their shares in total income. The assumption implies that the same "deduction rate" (share of deductions on income source) applies to all income sources. The so-defined deduction rate can be easily determined in the SOEP after the simulation of individual income tax burdens. In order to determine the level of "taxable labor income" in the SIAB, an average of these individual deduction rates (relation of total deductions to gross income) across all individuals in all years is applied to the reported gross wages.

In detail, the procedure works as follows:

- In the SOEP, determine for each taxable unit j (individual or married couple) the share of total deductions (D_{jt}) on total gross income (GI_{jt}) in each year t ($\alpha_{jt} = \frac{D_{jt}}{GI_{jt}}$).

[8] A description of this program is available on request. As explained in section 1.2.1, within the micro-simulation it is assumed that the total amount of income taxes is allocated across the different income sources according to their shares in total gross income. In addition, married couples are assumed to be taxed jointly yielding an equal tax rate for both spouses. Of course, any other micro-simulation model, e.g. the STSM (Steiner et al. 2012) can also be used. However, for the time period under consideration the STSM was not available.

- Use the mean value (across all taxable units and over time) of this share ($\overline{\alpha_{jt}} = \frac{1}{N^*T}\sum_{j,t}\alpha_{jt}$ with N^* being the total number of taxable units and T being the number of periods in a balanced panel) in order to determine an average amount of deductions on labor income ($\overline{\alpha_{jt}}y_{it}$) which can be used to derive a taxable labor income y_{it}^{tax} for each individual i in the SIAB.

$$y_{it}^{tax} = y_{it} - \overline{\alpha_{jt}}y_{it} \tag{4}$$

- Insert the taxable labor income into the legal tax function for singles in order to obtain imputed income tax burdens on labor income.

The approach has the advantage that it is easy to implement. However, due to the usage of the tax code for singles and the application of only one average deduction rate for all individuals it depicts an extreme simplification of the highly complex taxation system in Germany that could result in very inaccurat approximations.

Approach 1b

Approach 1b relies on regression imputation: within the auxiliary data set, simulated average tax rates on labor income are regressed on a set of predictor variables available in both data sets. For Approach 1b, these variables only include labor income variables and time variables, whereby the latter are supposed to capture changes in the tax schedule. The estimated function is then applied to the SIAB data to impute missing average income tax rates on labor income.

With respect to the functional form of the prediction function, a fractional logit model is chosen that (partly) accounts for the special characteristics of the dependent variable. The model reads:

$$l_i(\beta) = y_i \log[\Lambda(x_i\beta)] + (1 - y_i)log[1 - \Lambda(x_i\beta)] \tag{5}$$

where $\Lambda(\cdot) = exp(\cdot)/[1 + exp(\cdot)]$ is the logistic cumulative distribution function and the observations $\{(x_i, y_i) : i, \ldots, N\}$ are an independent sequence of observations with N being equal to the sample size. Further it holds that $0 \leq y_i \leq 1$. Estimations for β are obtained by maximizing the Bernoulli log-likelihood function. The fractional logit specification prevents predictions below zero. However, the functional form does not ensure that predicted tax rates lie above maximum tax rate levels. Predictions above legally fixed rates are treated with a censoring rule.[9]

As predictor variables, the following variables are included:

[9]Obvious miss-predictions in the SIAB, i.e. individual average tax rates that lie above legal thresholds, are cut to the maximum rates that have been simulated in the SOEP in the respective period.

- Yearly gross wage in current Euros (in levels plus a squared and a cubic term)
- Period dummies in order to capture changes in the tax schedule (e.g. the period from 1990 to 2009 is divided into 11 time intervals)
- Low labor income dummy in order to capture the tax-free zone of the German income tax schedule: The threshold value for the low labor income dummy is chosen such that it includes at least 95% of the persons who have a simulated tax rate smaller than 1% in the SOEP. The rule was tested against alternative specifications. It is applied separately for each taxation period. As resulting values do not differ strongly between some periods, some years can be summarized. This leads to the following definition of the low income dummy:

  1990-1995: yearly gross labor income below 9,000 Euro

  1996-1999: yearly gross labor income below 15,400 Euro

  2000-2009: yearly gross labor income below 18,000 Euro
- Interactions between yearly gross wage terms and period dummies

As only labor income and time variables are included in the set of predictor variables, the **assumption** is made that gross labor income is sufficient to approximate individual income tax rates. The assumption is supported by the fact that gross labor income makes up the largest share of household income for the majority of persons (see ILO 2015) and therefore strongly influences individual tax burdens. In addition, it is correlated with other tax-relevant characteristics (e.g. family status, children, ... (see e.g. Hill 1979)), which turns it into the most important available predictor variable for individual income tax rates.

Approach 2

Approach 2 relies basically on the same procedure as Approach 1b, i.e. on regression imputation. However, in addition to labor income and time variables, additional variables from the SIAB (age, gender, education, occupational status, industry) are incorporated into the set of predictor variables. Again, a fractional logit model is chosen and obvious miss-predictions, i.e. individual average tax rates that lie above legal thresholds, are cut to the maximum rates that have been simulated in the SOEP in the respective period.

The approach is based on the **assumption** that the additional variables available in the SIAB can help to capture tax-relevant characteristics like family status, the number of children, other income sources, etc. They are therefore supposed to improve the imputation from Approach 1b with tries to approximate variations in individual tax rates only based on gross labor income (and time).

Choice of predictor variables

The following variables are identified as possible predictor variables:

- Yearly gross wage in current Euros (in levels plus a squared and a cubic term)
- Low income dummy in order to capture the tax-free zone of the German income tax schedule (see Approach 1b for a definition)
- Period dummies in order to capture changes in the tax schedule (the period from 1990 to 2009 is divided into 11 time intervals)
- Male dummy
- Age in years
- Education

  1=Less than High School

  2=High School

  3=More than High School

- Occupational status

  1=Unskilled worker

  2=Skilled worker

  3=Master craftsman/foreman

  4=Employee

- Industry

  1=Agriculture

  2=Energy

  3=Mining

  4=Manufacturing

  5=Construction

  6=Trade

  7=Transport

  8=Bank/Insurance

  9=Services

- **Interactions** between the variables

As the inclusion of all these variables and their interaction terms would result in a large number of regressors and could result in the over-specification of the model, some simple statistical selection methods are applied to choose a suitable subset of predictors. For this purpose, all possible predictor variables (including most interactions)[10] are tested separately for their significance in explaining the individual average tax rates on labor incomes in fractional logit specifications.[11]

As a result of these simple tests, the entire set of pre-chosen variables as well as all interactions that have been taken into account can be considered to be suitable predictors from a statistical point of view. Summing up categorical interaction variables in one term, the set of possible predictor variables encompasses m=33 terms.[12] Due to the extensive amount of possible combinations (2^{33}) of these terms, not all of them can be tested within the scope of this analysis. Instead, a backward selection procedure is applied using both simple p-value rules and two different specifications of the Akaike information criterion (AIC), namely the original version of the AIC as well as a corrected specification, AICc.[13] The backwards approach examines at most $m(m+1)/2$ of all possible models (Lindsey et al. 2010) and is hence computationally much less costly than an exhaustive model search algorithm.[14]

In the selection procedure, first the full model is estimated. Subsequently, based on the chosen criterion, the predictor whose elimination yields the largest improvement in the chosen criterion is dropped. In case of the p-value criterion, the least significant variable (or interaction) is dropped if it is less significant than the chosen significant

[10]The square and cubic terms of nominal labor income are interacted with the period dummy in order to capture effects of changes in the tax function related to the nominal level of income but not with the remaining variables. The low income dummy is not interacted with other variables.

[11]If interactions are tested, the models always also include the main effects. If interactions between categorical variables are tested, they are always tested for joint significance (e.g. all interactions between all education levels combined with all industry categories are tested jointly). This proceeding prevents the fragmentation of the categorical interactions, which would make them extremely difficult to interpret.

[12]One term is defined as a variable or as an interaction between two variables. In case of interactions including one or two categorical variables, a term summarizes all interactions between all categories (e.g. interaction between education and industry is defined as one term here but includes $3x9 = 27$ variables.)

[13]The AIC takes the goodness of fit of the model as well as its complexity into account and chooses the model in a set of models that yields the smallest expected Kullback-Leibler information loss. The explanatory power of the model is measured by the maximized log likelihood of the predictor coefficients. The complexity penalization comes from the addition of the number of predictors. The formula for the AIC reads

$$AIC = -2lnL + 2k \tag{6}$$

with lnL being the maximized log-likelihood of the model and k being the number of predictors. Comparing models with respect to their AIC values, the one with the smallest value has the best trade-off between goodness of fit and complexity. The corrected version of the AIC (AICc) corrects for small sample sizes and reduces the risk of overfitting in large samples.

[14]Selections procedures of this manner certainly carry limitations as they require an *ex ante* choice of the selection criterion. Further, an extensive search over all possible models is generally preferred over procedures that only test a subset of them. Regarding the aim of performing out-of-sample predictions, a model selection procedure that tests as many possible models based on cross-validation would be presumably the preferred approach. However, such an approach involves high computational costs and was not performed for this analysis.

level. Subsequently, the same rule is applied to the reduced model and repeated until the elimination of another variable does not lead to an improvement of the criterion. Again, in case of the p-value, the procedure stops when all variables are statistically significant at the chosen level. The backwards procedure is conducted with two different significance levels for the p-value (5 and 15%) as well as with the AIC and the corrected version AICc.[15].

The selected functions based on the different criteria are very similar and only differ with respect to some interaction categories included. In fact, the AIC and AICc yield the same model. In a last step, the models chosen based on the different criteria are compared using cross-validation methods. These can be considered to mimic an out-of-sample prediction most closely. Specifically, a 50-fold cross-validation is applied, implying that the sample is randomly divided into 50 parts and the specified model is estimated including 49 parts of the sample. Subsequently, the estimated equation is used to predict values of the dependent variable for the 50th part that was excluded within the estimation. Predicted values for these observations are compared to true values calculating the Root Mean Squared Error (RMSE). The procedure is repeated 50-times, always excluding a different one of the 50 parts from the estimation. Finally, an average RMSE is calculated (averaging over all 50 steps), which is used to compare models with respect to their predictive power. In this comparison, the model chosen by the AIC/AICc performs *slightly* better than the ones chosen by the p-values and is therefore selected for the further analysis.

The chosen prediction function contains almost all possible variable groups, namely 31 of them. This yields a total number of 215 regressors plus a constant whereby some categories of categorical variables or interaction terms including categorical variables are dropped due to multicollinearity.[16]

[15]The backwards procedure using the AIC/AICc criterion could be only performed for a linear specification.

[16]Within the final prediction model, the predictor variables are interacted with the taxation period dummies in order to capture time-varying effects. This implies that the time-varying effects of the variables are restricted to the periods between which a change in the tax function takes place. For the variables next to labor income, this approach lacks a solid theoretical foundation: for example, there is no clear reason why the effect of education should change between taxation period 3 and taxation period 4 and not within different time intervals. However, as income from employment can be considered to be the most important predictor variable, the period dummies were selected accordingly. They capture the time-varying tax effects of this variable in the best possible way. Interactions of the chosen period dummies and the other variables still capture time-varying effects. In order resolve remaining doubts concerning the choice of the taxation period dummies, some tests against alternative specifications are performed. In particular, the interactions of the variables next to labor income with the period dummies are tested against simple interactions with a linear time trend that would prevent the somehow arbitrary partitioning of the years into period dummies. As a further test, a full set of year dummies instead of the taxation period dummies is added to the model and interacted with all variables. Finally, a last test is concerned with the circumstance that the first taxation period with an unchanged tax function lasts from 1990 to 1995. This results in a dummy encompassing 6 years whereby all other dummies only include 1 to 2 years. The finally chosen prediction model is therefore tested against a specification that divides the respective period into two smaller ones. However, in terms of cross-validation selection criteria (RMSE), the model using the taxation period dummies is preferred over all tested alternatives.

5 Empirical Application and Comparison of Approaches

The empirical application of the discussed approaches focuses on a sample of full-time employees in West Germany, aged between 20 and 60 for the years between 1990 and 2009. This sample can be identified in the SIAB data set as well in the auxiliary data set (SOEP).

The three different approaches described above are compared by evaluating their ability to produce reliable predictions of individual income tax rates on labor income. In particular, the different methods are applied to the SOEP data, which allows for a comparison of resulting predictions of the models with the true (simulated) individual tax rates on labor income. The prediction functions for Approaches 1b and 2 are estimated using sampling weights provided in the SOEP. The following evaluation criteria are used for the comparison:

- Applicable to all approaches:
  - Comparison of simple Root Mean Squared Errors (RMSE) and Mean Absolute Errors (MAE).[17] Smaller values of the criterion are preferred.
  - Simple R-squared as proposed by Papke and Wooldridge (1996), which is defined as $R^2 = 1 - \frac{SSR}{SST}$ with SST being the total sum of squares of y and SST being the sum of squared residuals based on unweighted residuals $\hat{u} = y - \hat{y}$. Larger values of the criterion are preferred.
- Applicable to regression imputation approaches (1a, 2):
  - Akaike Information Criterion (AIC), here defined as $(-2lnL+2k)/N$ with N denoting the sample size, lnL being the maximized log-likelihood of the model and k being the number of predictors. Smaller values of the AIC denote a better trade-off between accuracy and complexity.
  - Cross-validation procedure (50-fold)[18]: Comparing RMSE, Mean Absolute Errors (MAE) and a pseudo-R-squared (square of the correlation coefficient of the predicted and actual values of the dependent variable)

Table 1 summarizes the evaluation criteria for the different approaches. Approaches 1a and 1b produce the same Mean Absolute Error while Approach 2 clearly yields the smallest value of this criterion. Comparing the average Root Mean Square Errors and simple R^2 values of all three models reveals that Approach 1a performs worst according to these criteria. This is not surprising given the strong assumptions made for the calculation of a hypothetical "taxable labor income" and the usage of the tax code for singled for all individuals. Overall, the regression imputation approaches appear to be more successful

[17] The first one being more sensitive to outliers.

[18] See the section about Approach 2 for a description of the cross-validation procedure.

than Approach 1a in predicting average income tax rates on labor income.

Approach 2 outperforms Approach 1b according to most criteria including the cross-validation criteria. In fact, Approach 1b is only preferable according to the Akaike Information Criterion. This finding can be most probably explained by the higher complexity of Approach 2 (the larger number of predictor variables) which is penalized within this criterion.

5.1 Limitations due to Out-Of-Sample-Predictions

The aim of the analysis is to apply the prediction functions built in the SOEP to another data set (here the SIAB). However, the quality of the imputation methods can only be evaluated with respect to their ability to predict the variable of interest within the auxiliary data set. In contrast, it is obviously impossible to assess the quality of the predictions in the SIAB where the true values of the predicted variable (individual tax rates) are unknown. In theory, the methods applied in a sufficiently large random sample from a certain population should produce similarly reliable predictions in other random samples that stem from the same population. However, in practice, if data sets are based on the same population but stem from different data sources, many factors like different definitions of variables in data sets, measurement errors or incomplete random sampling can lead to deviations from this ideal. Before applying the discussed methods to other data sets, the respective data sets hence have to be compared concerning their composition and their distribution of predictor variables. In general, though, the risk of miss-predictions due to differences between data sets and variables increases with the number of variables included in the prediction function.

Given these considerations, the superiority of Approach 2 according to the evaluation criteria appears less convincing: even though Approach 2 dominates Approach 1b according to most criteria applied here, the large amount of terms included in the chosen prediction model in Approach 2 represents a weakness of the approach given the ultimate purpose of producing out-of-sample predictions within a different data set. In addition, the gain in predictive power that is obtained by adding the additional variables to the set of predictor variables in Approach 2 is rather small in relation to the implied increase in complexity.

6 Conclusion

This data documentation discusses different ways of determining net wages based on gross wages in German administrative data sets provided by the IAB with a focus on the SIAB data set. Whereas social security contributions can be simulated directly in the data sets, income tax burdens on labor income (as well as solidarity tax burdens) must be imputed as important tax-relevant characteristics are missing. Three different

Table 1: Evaluation of Predictions

Approach	RMSE	MAE	R^2	AIC	Cross-validation			# predictor variables[a]
					RMSE	MAE	pseudo R^2	
1a	0.045	0.035	0.581					1
1b	0.044	0.035	0.591	0.587	0.045	0.036	0.589	45
2	0.037	0.027	0.638	0.603	0.041	0.032	0.630	207

Evaluation of imputation approaches 1a, 1b and 2 in the SOEP; sample: full-time dependently employed persons in West Germany, aged between 20 and 60, 1990-2009.

[a]Due to otherwise perfect multicollinearity some reference terms of categorical variables and interactions including categorical variables are dropped.

approaches are discussed that rely on different assumptions and that vary with respect to their degree of complexity. The different procedures are derived and tested in an auxiliary data set (SOEP) that allows for a comparison of predicted and true (simulated) income tax rates. Various evaluation criteria are applied to assess the quality of the produced predictions. The results reveal that the simplest approach performs worst. The procedure tries to approximate a taxable labor income that is subsequently inserted into the legal tax code for singles. Instead, the tested approaches relying on the method of regression imputation yield better results. The two different models that are compared in this respect only differ with respect to the number of included predictor variables. The more complex prediction function that includes various variables in addition to gross labor income is preferred according to most criteria. However, the gain in predictive power is rather small while the approach implies a significantly higher degree of model complexity (a larger number of included regressors). It follows that the additional explanatory variables used in addition to labor income, which are supposed to approximate tax-relevant household characteristics, do not bear strong predictive power. In addition, the larger number of predictor variables included in the more complex model is connected to a higher risk of miss-predictions in case of performing out-of-sample predictions within a different data set. The regression imputation approach relying solely on labor income and time variables can be hence considered as a solution to the imputation problem that implies a good trade-off between achievable precision and limited complexity.

References

Gunselmann, I. (2014). Programmierbeispiele zur Umrechnung des Brutto-in ein Netto-Tagesentgelt für die administrativen Daten des FDZ. *FDZ Methodenreport 01/2014*.

Hill, M. S. (1979). The wage effects of marital status and children. *Journal of Human Resources 14*, 579–594.

ILO (2015). Global Wage Report 2014/15: Wages and income inequality. *International Labour Office*.

Lindsey, C., S. Sheather, et al. (2010). Variable selection in linear regression. *Stata Journal 10*(4), 650–669.

Papke, L. and J. Wooldridge (1996). Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates. *Journal of Applied Econometrics 11*, 619–632.

Steiner, V., K. Wrohlich, P. Haan, and J. Geyer (2012). Documentation of the tax-benefit microsimulation model STSM: Version 2012. *DIW Data Documentation*.