Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ-Methodenreport 01/2016 EN

Methodological aspects of labour market data

Imputation of the Working Time Information for the Employment Register Data

Johannes Ludsteck,
Ulrich Thomsen

Bundesagentur für Arbeit

Imputation of the Working Time Information for the Employment Register Data

Johannes Ludsteck (IAB)
Ulrich Thomsen (IAB)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Zusammenfassung

2011 wurde das der Beschäftigungsstatistik zugrundeliegende Datenerhebungsverfahren umgestellt. Dies führte zeitweise zu einem ungewöhnlich hohen Anteil an fehlenden Informationen im Tätigkeitsschlüssel, der bis zu 25 % erreichte. Wir untersuchen die daraus resultierenden Auswirkungen auf Lohnanalysen und entwickeln ein einfaches Modell zur Imputation der fehlenden Informationen. Hierzu vergleichen wir die Prognosefähigkeit von Logit- und linearen Wahrscheinlichkeitsmodellen, sowie von Klassifikationsbäume und fast-gesättigte nichtparametrische Probit-Modellen.

Abstract

Due to changes in the data collection method in 2011, German Employment Register data temporarily exhibit an uncharacteristically high share of missing information (25 %) regarding the job activity profile (Tätigkeitsschlüssel). We assess the consequences of this on wage analyses and develop a simple model to impute the missings. To this aim, we compare the predictive power of Logit and linear probability models, classification trees and nonparametric probit models.

Keywords: German Employment Register, missing observations, imputation

Acknowledgements: We thank Christian Gaggermeier, Jörg Drechsler, Stefan Seth, Thomas Frank, Ursula Jaenichen, Holger Meinken, Philipp vom Berge, Manfred Antoni and Stephanie Eckman for helpful comments and suggestions.

1 Introduction

The collection method for variables concerning the job activity profile (Tätigkeitsschlüssel) in the German employment register was changed in 2011 (see Bertat et al. (October 2013) for further details). The affected variables are occupation, working time, schooling and occupational qualification, indicators for fixed-term contracts and temporary work. Some of them are essential for most kinds of descriptive or causal statistical analyses that use these data. Unfortunately the change of the data collection procedure was associated with a temporary but huge increase of missings on these variables. So this change has severe consequences on the analytic potential of that data.

In this report we try to assess the consequences of missing information of the working time variable and to impute the missings. We focus on the *categorical* variable 'working time'. Under the new reporting scheme, it has two levels, full time and part time. Under the old reporting scheme, the latter category was divided into part time with at least half of the standard full time hours, and part time with less that half of the full time hours. There is another variable 'working hours' in the register data, intended to supply the statutory accident insurance agencies with information on the spell-specific average working hours of the employees. Its usability for statistical analyses is, however, limited since it contains large numbers of missings and the reporting may be dominated by heuristic rules (e.g. Vollarbeiterrichtwert).[1]

2 Incidence and evolution of the missing working time information

To get an idea of the extent of the problems caused by missing information, look at Table 4 in the Appendix. Since reliable wage regressions based on the register data have to be restricted to full timers only, shares of working time missings exceeding 20 percent render reliable wage analyses for these samples impossible. In order to tackle this problem, we impute the missing working time info. This will be described in detail in section 4 below.

Another problem that existed already before but was brought to light by the change of the reporting scheme concerns the reliability of the working time info. As can be seen from Table 1, the share of part-time spells of women was considerably greater in 2011 for reports submitted according to the new scheme. For men we observe the contrary. Further inspection of the data suggests that transitions from full-time work to part-time work which took place within the same establishment and the same employment relationship were overlooked by employers when filling in the social security records.

We will inspect this hypothesis using a simple differences-in-differences regression design in the next section.

[1] See e.g. the reporting advice of the Bayerische Landesunfallkasse (see this Link) concerning the working time: "Bitte geben Sie hier die tatsächlich geleisteten Arbeitsstunden an, in denen das Entgelt erzielt wurde. Ist dies nicht möglich, können Sie auch die Sollarbeitszeit angeben, eine gewissenhafte Schätzung vornehmen oder den Vollarbeiterrichtwert (für Meldezeiträume im Jahr 2012 = 1.600 Stunden, für zurückliegende Meldezeiträume im Jahr 2011 = 1.570 Stunden) heranziehen."

Table 1: Numbers of observations and Part-Time Shares in 2011, detailed for several Subsamples of the IAB-dataset of Employment Data 'Beschäftigtenhistorik'(BeH).

Sex	Reporting Scheme	Missing	Nr. Spells	Share Part-Time	Person Years	Share Part-Time
Men	Old	No	8,275,617	0.344	2,602,200	0.317
		Yes	30,600	–	10,790	–
	New	No	13,534,355	0.250	11,208,969	0.219
		Yes	4,756,108	–	3,918,291	–
Women	Old	No	7,967,891	0.576	2,640,082	0.565
		Yes	39,269	–	15,318	–
	New	No	13,921,465	0.603	11,487,333	0.591
		Yes	3,623,323	–	2,983,180	–

Notes: 'Nr. Spells' represents the raw number of spells, 'Person Years' the duration- weighted number of spells (weights are defined as the spell duration in days divided by 365). The part-time shares are related to the respective columns to their left.
Source: BeH

3 Evaluating the effects of the revisions

To demonstrate the impact of this reporting bias on wage analyses, we compute wage quantiles based on spells reported according to the new and the old reporting scheme.

3.1 Inspecting the causes of the revisions

As mentioned above, the change of the reporting scheme revealed biased reporting of the working time. This seems to be caused by underreporting of working time changes that occurred mainly within employment relationships, i.e. when an employee moved from full-time to part-time employment while remaining in the same job. The change of the reporting scheme included a new mapping of working time categories to numeric codes. This prevented employers from just copying the information from the preceding report and by that induced them to revise the content of the records.

We check this hypothesis by comparing the incidence of working time changes simultaneously along two dimensions: First, the difference between establishment stayers and establishment movers. Second, the difference between spell sequence pairs in which both spells were reported according to the same reporting scheme and pairs in which the first spell was reported according to the old and the second spell according to the new scheme. If systematic underreporting of working time switches occurs due to copy-and-paste practices, working time status changes should be relatively more frequent than status persistence if two successive spells are issued according to different reporting schemes.

To perform these comparisons, we identify three groups of spell sequence pairs. The first one contains pairs of non-overlapping successive spells within the same establishments (represented by indicator I_i^{SE} below). The second one contains non-overlapping pairs of successive spells reflecting establishment moves (i.e. the establishment ID of the second spell differs from that of the first spell), and the third contains overlapping spells related to

different establishments (representing multiple employment). To speed up the computations, this analysis is based on a (stratified[2]) 5 percent sample of all persons with at least one spell in the period 2010-2012.

We run the following linear probability regression model[3] for these spells

$$I_i^{FP} = I_i^{RSC}\, b_{RS} + I_i^{SE}\, b_{SE} + I_i^{RSC} \times I_i^{SE}\, b_{\times} + x_i\, \beta + \epsilon_i$$

where I_i^{FP} is an indicator for a switch from full-time to part-time work, I_i^{RSC} indicates a reporting scheme change, x_i contains controls (month dummies, year dummies, dummies for age groups ($\geq 25, \geq 35, \geq 45, \geq 55$) and a gender dummy) and ϵ a residual term. In the difference-in-difference design, heterogeneity between the reporting schemes is captured by I_i^{RSC}, heterogeneity between establishment stayers and movers is captured by I_i^{SE} and the revision effect is represented by the interaction term $I_i^{RSC} \times I_i^{SE}$.

The following Table 2 shows the effects of the central dummies I_i^{SE}, I_i^{RSC} and $I_i^{RSC} \times I_i^{SE}$. The effects are measured by comparing the predicted probabilities with the respective dummy switched on and off. The coefficients show that transitions from full-time to part time work are more frequent if two successive spells are reported by different employers. We also find more transitions associated with a change of the reporting scheme. Finally, the positive interaction effect shows that reporting scheme changes exert stronger effects on continued employment relations, suggesting that revisions would have been otherwise omitted.

Table 2: Linear Probability Regression explaining Transitions from Full-Time to Part Time Work: Effects of the Diff-In-Diff Design Dummies.

Variable	Meaning	Off(0)	On(1)	Difference On(1) - Off(0)
I_i^{SE}	Same Establishment	0.198	0.022	-0.177
I_i^{RSC}	Rep. Scheme Change	0.065	0.072	0.007
$I_i^{RSC} \times I_i^{SE}$	Interaction Term	0.063	0.109	0.045

Source: 5 percent sample from the BeH (2010-2012).

Notes: Size of the regression sample: 7,564,639, Pseudo-R^2: 0.1547. Further control variables: 3 year dummies, 11 month dummies, 4 age group dummies and a gender dummy. The P-values of all reported coefficients and differences are smaller than 0.001 and therefore omitted.

Legend: The entries in columns 3 and 4 are predicted probabilities. Column 5 contains the differences between rows 4 and 3.

[2] The strata are federal state, gender and employment status.

[3] The corresponding results from a probit model are almost identical. We use the linear probability model here for sake of simplicity.

3.2 Assessing the effects of the reporting bias on the wage distribution

As already mentioned above, the full-time part-time information in the register data is of paramount importance for wage analyses since the register data do not contain reliable working hours information. A simple strategy to assess the effect of the reporting bias on the wage distribution is to compute time series of several wage quantiles of full-time employees. The results of this exercise are visualized in figure 1. The graphs show percentage changes (compared to the base period 2000) of inflation-adjusted[4] daily wages at several quantiles for four subsamples of full-timers. Note that the differences of level between the quantiles are 'swallowed' by this presentation format. This is accepted in order to allow the visualization of the (almost) entire range of the wage distribution while making small changes across time still visible.

The graphs have two eye-catching features. First, all graphs show pronounced upward kinks in the conversion year 2011. Second, significant distortions of the wage structure persisting after the transition year 2011 are present for female workers only.

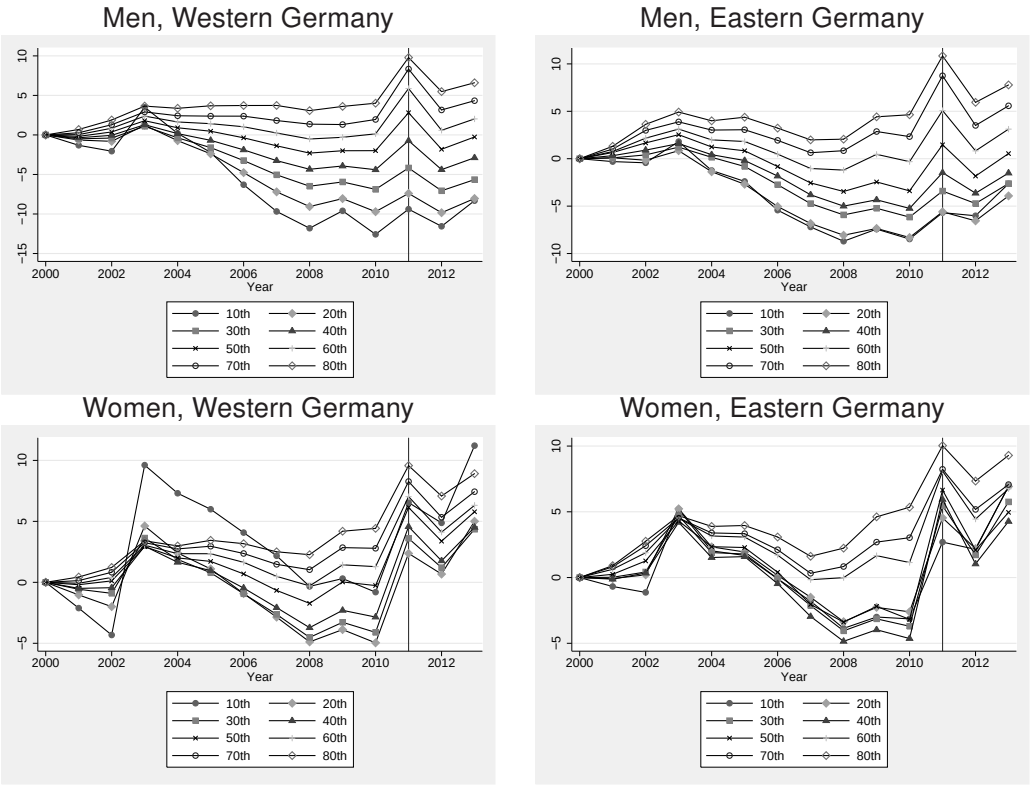
The second feature can be explained by the fact that part-time work plays a much greater role for female workers. As can be seen from Table 1, the duration-weighted share of part-timers exceeds 50 percent for women irrespective of whether the reporting was done within the old or the new scheme. The part-time shares are considerably lower for men for whom they remain always below 35 percent.

To understand the first feature we have to consider how the full timers are selected to compute the wage quantiles. After restricting the base sample to all employees subject to social security contributions[5], full timers are identified using the variable 'employment status' and 'working time' for spells reported according to the old and new scheme, respectively. Spells reported according to the new scheme that lack working time information cannot be assigned to the full- or part-time sample and must be dropped from the sample. This implies that wages from spells lacking working time information are lower on average. This explanation is confirmed by the fact that the pronounced increases across all quantiles are almost reversed in 2012 when the share of working time missings approaches zero.

Starting with Fitzenberger (1999), a considerable number of papers[6] has documented a secular increase of wage inequality in Germany that started in the mid-nineties. Since most of these papers are based on the register data, it is of paramount interest to assess whether their results are affected and biased by the misreporting documented here. To look at this issue at a finer grid, the 80/50 and 50/20 quantile ratios are plotted in figure 2. As is evident already at a glance, reporting bias generates visible changes of the big picture only in two cases: First, we see a considerable drop of the ratio from about 1.58 in 2010 to 1.52 in 2012 for the upper part of the wage distribution (80/50 ratio) for women in Eastern Germany (see the graph with triangle markers on the left hand side). However, compared to 2000 or 2005, there is still a significant increase of the ratio even after the revisions took

[4] We use the consumer price index to deflate the wages.
[5] identified by the variable 'person group' being equal to 101
[6] See e.g. Antonczyk/Fitzenberger/DeLeire (March 2010) for a still incomplete list of contributions.

Figure 1: Time Series of Wage Quantiles at Cutoff Dates, 30th of June)



All graphs give **growth rates** of the respective wage quantiles **compared to the base year 2000.**

place. Second, the 50/20 wage quantile ratio returns to its 2005 value after the break for women in western Germany (see the graph with square markers on the right hand side). Note, however, that neither of these breaks or kinks reverts the trends observed between 2000 and 2010 completely.

It is important to note that neither of the main conclusions of the empirical papers cited above are affected significantly by these revisions. Either these papers focus on men's wage distribution or they put disclaimers on the results for women.[7]

To inspect and quantify our explanation for the kink in more detail, we run simple linear probability models explaining the indicator variable 'working time is missing' by wages and some controls.[8] The sample of this analysis is restricted to the period during which reporting was possible according to both schemes (2011-01 to 2011-11). We allow for non-linearities in a simple but flexible way by recoding wages into a factor variable. Wages are partitioned into 8 quantile groups of equal size. An additional 9th group contains all top-coded (censored) observations.

[7] As is well known, the greater importance of part-time work for the female subsample creates problems of comparability.

[8] Further controls are the employee group, indicators for spell-overlap, indicators for job continuation (previous and/or next spell are reported by the same employer), the duration of the spell, the duration of the gap between the current and the preceding, and the current and the succeeding spell), age, log establishment size, and indicators for eastern German regions and foreigners. We omit the coefficient estimates of the controls to save space. They can be obtained from the corresponding author upon request.

Figure 2: Time Series of Wage Quantile Ratios (at Cutoff Dates, 30th of June)

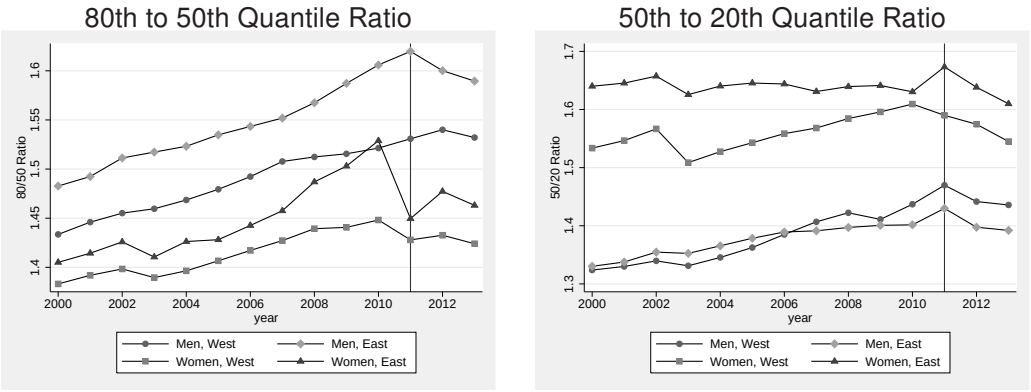
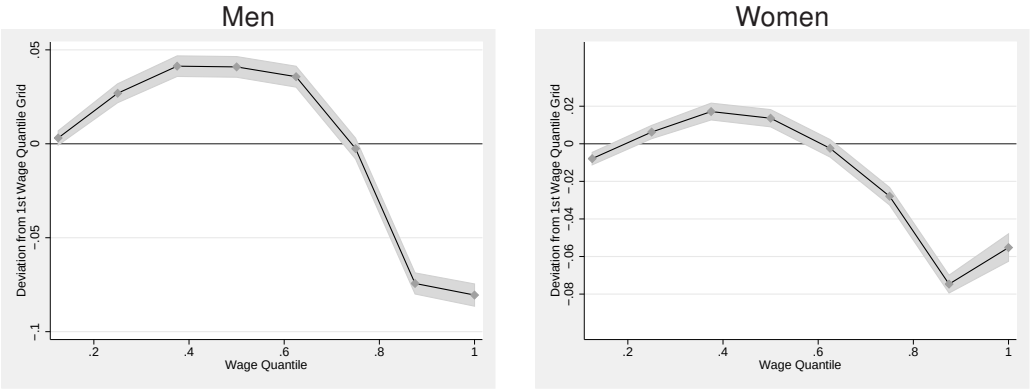


Figure 3: Relationship between Wages and predicted Prob(Working Time is Missing)



Legend: The graph gives the predicted missings incidence of the respective wage quantile groups, measured as deviations from the missing incidence of the lowest quantile group (quantiles 0 to 0.125). The predicted probabilities are taken from a linear probability model described in the text above. The shaded areas represent 99 percent confidence intervals.

Figure 3 shows the profile of the relationship between wages and the incidence of missings from this regression. The graphs are to be interpreted as predicted probabilities of missing working time information for all quantile groups, compared to the lowest group (which contains all wages below quantile 0.125). The incidence attains a maximum at roughly 0.4 and is clearly below average for all quantiles above the median. Note also that the graphs drop considerably on the right hand side of 0.6 and that the missings incidence of the two topmost quantile groups is roughly 8 percent lower than that of the lowest group. In summary this confirms that missings are more prevalent among spells with lower wages.

4 The imputation procedure

The main purpose of this project is to impute the missing working time information by employing spells with complete information. After experimenting with several alternative modelling approaches (standard Logit/Probit models, near-saturated nonparametric Probit regressions and linear probability models) we opted for classification trees for two reasons.

First, they delivered slightly better in-sample and cross validation fit rates. Second, (we see this as an important advantage compared to traditional Logit/Probit models), they are built using automatic variable selection algorithms, including the selection of interactions of arbitrary order.

The trees are generated by recursive binary splitting of the estimation sample such that – for every splitting step – the deviation of the dependent variable between the two subsamples is maximized (and – as a logic consequence – the within-sample heterogeneity is minimized). The algorithm allows for general forms of interactions between the (discrete) predictors by considering – for every predictor – all possible combinations of its values within one recursion step. To explain this assume that age (in years) is available in three categories, $[< 30]$, $[30 - 49]$ and $[\geq 50]$. The algorithm forms all possible *binary* splits

1. $[< 30] \cup [30 - 49] \quad versus \quad [\geq 50]$
2. $[< 30] \quad versus \quad [30 - 49] \cup [\geq 50]$
3. $[< 30] \cup [\geq 50] \quad versus \quad [30 - 49]$.

If, for example, the heterogeneity of working time is greatest for partition [2.], then the data set is split according to it. In the following recursive step, the remaining possible splits [1.] and [3.] are tested against each other in order to determine further splits of the data. These tests are performed *separately* for the two parts of the data set produced in the preceding step. If other predictors are available, all possible combinations of their categories are formed and interacted with those above. If, for example, gender would be used as a second predictor, the 5 categories $[< 30]$, $[30 - 49]$, $[\geq 50], F, M$ would be available in total which could be combined to 5 binary splits.

Once the tree is generated, imputed values are obtained as

$$\widehat{I_i^{PT}} := I(S_i^{PT} > u_i)$$

where u_i is a pseudo random draw from a uniform distribution with range $[0; 1]$, S_i^{PT} is the part-time share of the tree-leaf[9] to which the observation i belongs, and $I(\cdot)$ represents the indicator function, taking on value unity if its argument is true and zero otherwise.

The imputation model is based on all observations that are submitted using the new reporting schemes during 2011 and 2012 and contain complete information on all predictors (to be described below) and the working time. Note that all variables obtained from the same key-variable mapping as the working time (these are occupation, schooling and occupational qualification, indicators for fixed-term contracts and temporary work) are missing whenever the working time is missing. Consequently they cannot be used a predictors. However, our imputation results suggest that the available predictors are quite powerful. The most important ones are the lags *(LPt)* and leads *(FPT)* of the working time, the person group *(Personengruppe, PG)*, (real) wage percentiles *(Wage)*, the duration of spells

[9] Leafs of the tree represent the final data set splits.

(DSp), The duration of the gap between the current and the previous spell *DGap*, the establishment size *(ES)* and age *(Age)*. To account for the fact that the predictive power of the leads and lags of the working time (i.e. the working time of predecessor and successor spells) depend on whether they refer to the same job, they are interacted with job continuation dummies *(LSEs and FSEs)* indicating whether the preceding or succeeding spell is reported by the same establishment Our (slightly aggregated) version of the employment status variable *(Personengruppe, PG)* comprises 5 broad groups: regularly employed persons, minor jobs, apprentices, persons in partial retirement and a small residual category. Some continuous predictors (wages, the duration of spells and the duration of a block containing overlapped spells) are converted into ordinal variables. For example, wages are partitioned into eight quantile range groups (plus one for the censored wages), and age is split into groups spanning 5 years. For other variables, e.g. a spell's share in the total duration of an overlap block *(SDObl)*, the cutoff limits for the grouping are heuristically determined such that the between-group heterogeneity of the dependent variable is maximized. See the technical appendix for a complete list of variables and their definitions.

The trees were fitted using the gnu R package rpart by **?** with complexity parameter 0.0001. The quality of the imputation procedure is assessed heuristically in two directions. First, we compute the in-sample-fit rates between the imputed and the original variable. The rates can be read from Table 3.[10] Second, we re-render the time series quantile graphs from Figure 1 above after filling the missings with the imputed part-time information and recomputing the quantiles based on the imputed part-time variable. The resulting graphs which suggest a clear improvement by the imputation can be inspected in appendix figure 4.

The reader is referred to the Appendix Figures 5 and 6 in order to get an impression of the estimation results. Clearly, the algorithm does not yield coefficients in the traditional sense. The imputation rules (and the structure of interactions) can be read from the tree figures instead.

Table 3: Weighted in-sample fit rates (percent) for respective the subsamples (spell durations are used as weights).

	Women	Men
Eastern Germany	95.3	96.9
Western Germany	96.4	97.7

Legend: All fit rates relate to the final predictions, i.e. those obtained by setting the predicted value for a spell to 'part-time' if the part-time share of the relevant tree-leaf (to which the spell belongs) exceeds the value of a draw from a uniform $U[0;1]$ distribution.
Source: BeH, own computations.

[10] We also computed 5-fold cross-validation fit rates in the explorative steps of the project. We abandoned this additional computational burden since they were quite similar to and highly correlated with the the in-sample rates.

5 Corrections to be done by the user

Our imputation design follows the principle to reproduce the information contained in the base data as well as possible. To this aim we consciously refrain from correcting some unplausible observations. For example, there are some rare cases of persons in mini jobs registered as full-timers. Depending on the objective of the analysis, some users may wish to clean such cases. Obvious candidates for such corrections are persons in mini-jobs and apprentices (an apprenticeship training in part-time employment was in principle impossible before the reform in 2005, see §8 BBIG and §27 HwO.)

6 Conclusion

Though the change of the data collection procedure for the German Employment Register (Beschäftigtenstatistik) brought about a temporary and large increase of missing information, it also delivered useful information on how reporting errors may accumulate over time. Using simple regression models we show that employers seem to miss to report changes of the working time in continuing jobs. We assess the effects of the reporting errors on wage analyses in a loose way by considering time series of wage quantiles. At least if the analysis is restricted to men and suspiciously low wages (i.e. wages not exceeding the double of the minor job threshold (Geringfügigkeitsgrenze)) are dropped from the full-time sample, the reporting errors seem to induced only minor bias. Since the trend increase of part-time observed over the last decade can be expected to continue in the future, the reporting errors may become substantial and accumulate over time. We thus should try to improve the reporting procedure, preferably by supporting the employers.

Finally, we develop simple parametric and nonparametric imputation procedures yielding in-sample fit rates of roughly 90 percent. They will be checked and possibly improved in future versions of the paper.

Variable definitions

General information related to several of the following variable definitions:

1. The code '5' is used to indicate that a state or the value of a variable cannot be defined sensibly. The latter is the case e.g. for the initial entry of a worker into the labour market. Then an appropriate definition of the duration of the gap between the current and the preceding spell is not available. Note that code '5' delivers information to the regression model since an indicator variable (or interaction term) is generated whenever it appears as a category.

2. The prefixes **L** and **F** stand for lags and leads (**F**uture values).

3. Prefix **D** indicates durations.

Definitions of all used variables

1. **Overlap Indicator** LOl, FOl**:** Takes on value unity if a spell is overlapped with at least one other spell and zero otherwise. The imputation model contains also the first lag LOl and the first lead FOl of this indicator. The lead and the lag are assigned to a third value (5) for spells without preceding or succeeding spells (i.e. the first and last spell of a working biography).

2. **Lag and Lead of the part-time indicator** LPT, FPT**:** The part-time information of the preceding and succeeding spell, respectively. These indicators take on value 5 if a spell is overlapped by at least one other spell since the predictive power and meaning of the lag and a lead are different for overlay spells.

3. **Employer Continuation Indicators** $LSEs, FSEs$ (**S**ame**E**mployer)**:** Indicate whether the preceding or succeeding spell was reported by the same establishment. The indicators are set to 5 for overlap spells or cases where the succeeding spell is not observed yet.

4. **Age (categorical):** Age of the person in years, partitioned into intervals of length 5:

Class	1	2	3	4	5	6	7	8	9	10	11
Age	≤ 20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65	>65

5. **Duration of the spell (categorical)** DSp**:** Duration of the spell in days, partitioned using the (inner) interval limits: 5, 15, 31, 90, 180.

6. **Duration of the gap between the current and the preceding spell (categorical)** $DGap$**:** Partitioned using the (inner) interval limits: 5, 10, 30, 90, 180, 366.

7. **Wage (categorical):** The (real) wage variable is partitioned into 8 quantile range groups. An additional 9th group contains the top-coded observations. To allow for heterogeneity along region and sex, the wage quantiles are computed separately for the 4 cells obtained by interacting Eastern and Western Germany and Men and Women.

8. **Employee group** PG (**P**erson**G**roup)**:** The employee group is slightly aggregated according to the following map:

Code	Meaning	Aggregated Code
101	Empl. subj. to soc. sec. contributions without further special characteristics	1
140	Sailors subj. to soc. sec. contributions without further special characteristics	1
102	Apprentices without further special characteristics	2
121	Apprentices, wage below minor employment threshold	2
122	Apprentices, external training	2
141	Sailor apprentices	2
103	Persons in partial retirement	3
142	Sailors in partial retirement	3
109	Persons in Minor Jobs	4
209	Persons in household-related minor jobs	4
$*$	All other	5

The last category $*$ contains more than 10 groups which amount, however, to less than 3 percent of all spells in total.

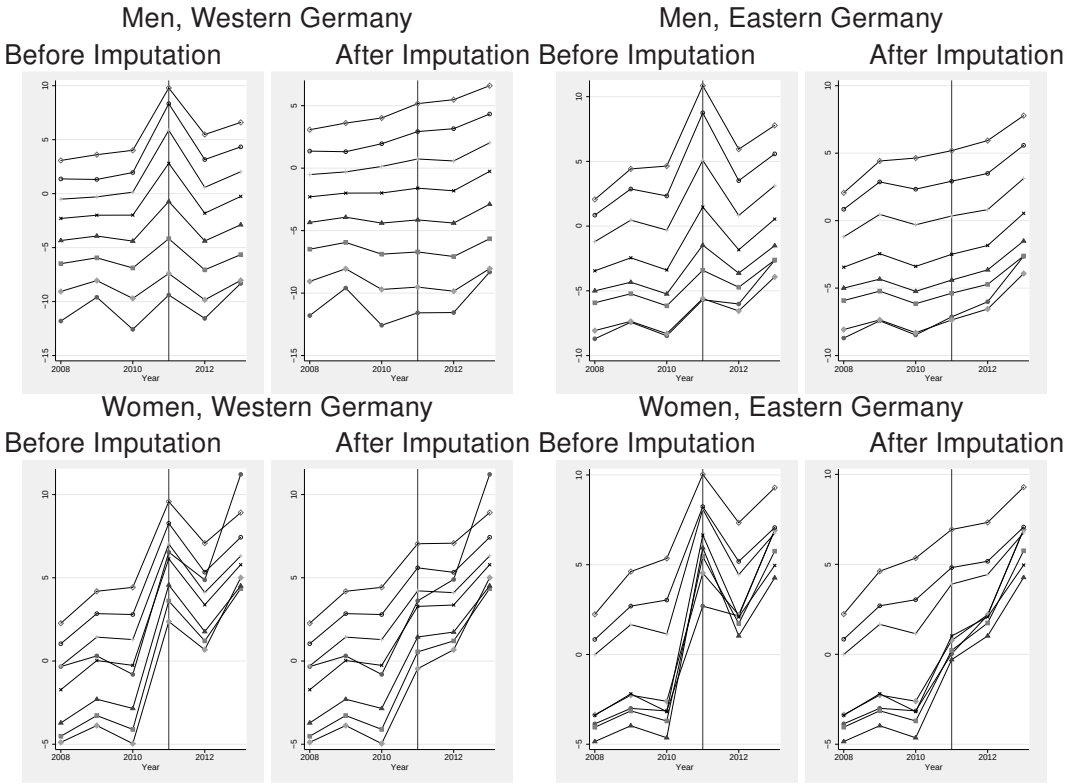
9. **Establishment size** ES**.** Weighted sum of spells of an establishment in a year.

$$es_{et} = \sum_{i \in E_{et}} w_i \times dur_i / 365.$$

where E represents all ID's of the persons employed in year t in establishment e, dur_i denotes the duration of the spell in days, and

$$w_i = \begin{cases} 1 & \text{full-time spell} \\ 0.5 & \text{part-time spell} \\ 0.75 & \text{working time is missing} \end{cases}$$

Figure 4: Time Series of Wage Quantiles at Cutoff Dates, (30th of June). Quantiles Based on Raw Working Time **Versus** Quantiles based on the Imputed Working Time



All graphs give **growth rates** of the respective wage quantiles **compared to the base year 2000.** (See Figure 1 for further info.)

Table 4: Percentage Shares of Missing Working Time Information by Reporting Scheme and Month

Month	Spells RS2003	Spells RS2010	% Missing RS2003	% Missing RS2010
2010-02	33,379,260	0	0.13	–
2010-03	33,496,106	0	0.13	–
2010-04	33,769,692	0	0.13	–
2010-05	34,007,788	0	0.13	–
2010-06	34,187,520	0	0.13	–
2010-07	34,154,228	0	0.12	–
2010-08	34,313,328	0	0.12	–
2010-09	34,593,456	0	0.13	–
2010-10	34,576,804	0	0.14	–
2010-11	34,842,536	0	0.13	–
2010-12	34,697,544	0	0.13	–
2011-01	9,159,076	24,940,916	0.39	30.18
2011-02	8,631,592	25,531,504	0.41	30.16
2011-03	8,154,146	26,187,366	0.43	30.22
2011-04	7,576,040	26,973,384	0.46	30.22
2011-05	7,063,052	27,695,552	0.50	30.20
2011-06	6,494,528	28,462,988	0.52	30.18
2011-07	5,485,699	29,399,804	0.49	30.17
2011-08	4,424,368	30,663,552	0.54	30.45
2011-09	3,384,760	32,004,462	0.69	30.90
2011-10	2,176,062	33,199,282	0.89	30.84
2011-11	1,229,112	34,422,304	1.13	30.69
2011-12	0	35,467,304	–	30.49
2012-01	0	34,835,816	–	3.37
2012-02	0	34,743,812	–	2.75
2012-03	0	34,884,240	–	2.16
2012-04	0	35,042,400	–	1.47
2012-05	0	35,245,708	–	0.97
2012-06	0	35,284,648	–	0.81
2012-07	0	35,198,352	–	0.80
2012-08	0	35,353,380	–	0.79
2012-09	0	35,606,384	–	0.80
2012-10	0	35,665,760	–	0.81
2012-11	0	35,799,468	–	0.81
2012-12	0	35,636,960	–	0.81
$*--*$	3,878,480	1,681,474	0.62	12.51
Total	443,675,168	443,675,168	0.18	11.36

Notes: All observation numbers and shares refer to stocks at the 15th of each month. $*--*$ refers to spells that cannot be assigned to a cutoff date (15th of month).
Source: BeH

Table 5: Observation Numbers and Part-Time Shares in 2011, detailed for several Subsamples of the BeH.

Year Year	Sex	Reporting Scheme	Missing	Nr. Spells	Share Part-Time	Person Years	Share Part-Time
2010	Men	Old	No	26,099,104	0.240	17,200,832	0.195
			Yes	43,908	–	22,860	–
		New	Yes	2,134	–	1,289	–
	Women	Old	No	25,289,688	0.541	16,756,787	0.529
			Yes	33,226	–	16,352	–
		New	Yes	3,791	–	2,754	–
2011	Men	Old	No	8,275,617	0.344	2,602,200	0.317
			Yes	30,600	–	10,790	–
		New	No	13,534,355	0.250	11,208,969	0.219
			Yes	4,756,108	–	3,918,291	–
	Women	Old	No	7,967,891	0.576	2,640,082	0.565
			Yes	39,269	–	15,318	–
		New	No	13,921,465	0.603	11,487,333	0.591
			Yes	3,623,323	–	2,983,180	–
2012	Men	New	No	25,304,698	0.266	17,560,098	0.206
			Yes	735,272	–	133,239	–
	Women	New	No	24,371,732	0.594	16,708,548	0.572
			Yes	780,329	–	308,501	–

Notes: 'Nr. Spells' represents the raw number of spells, 'Person Years' the duration- weighted number of spells (weights are defined as the spell duration in days divided by 365). The part-time shares are related to the respective columns to their left.
Source: BeH

Figure 5: Estimated Classification Tree, Western German Women (top) and Men (bottom)

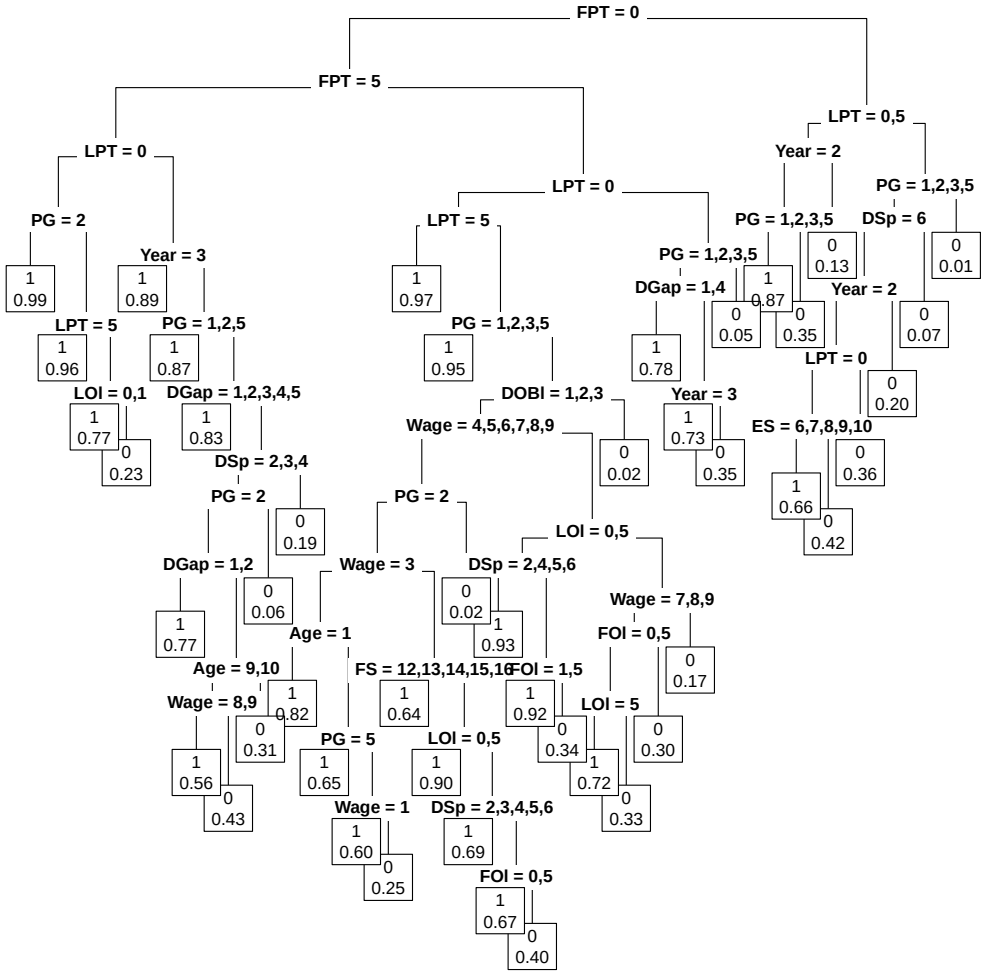
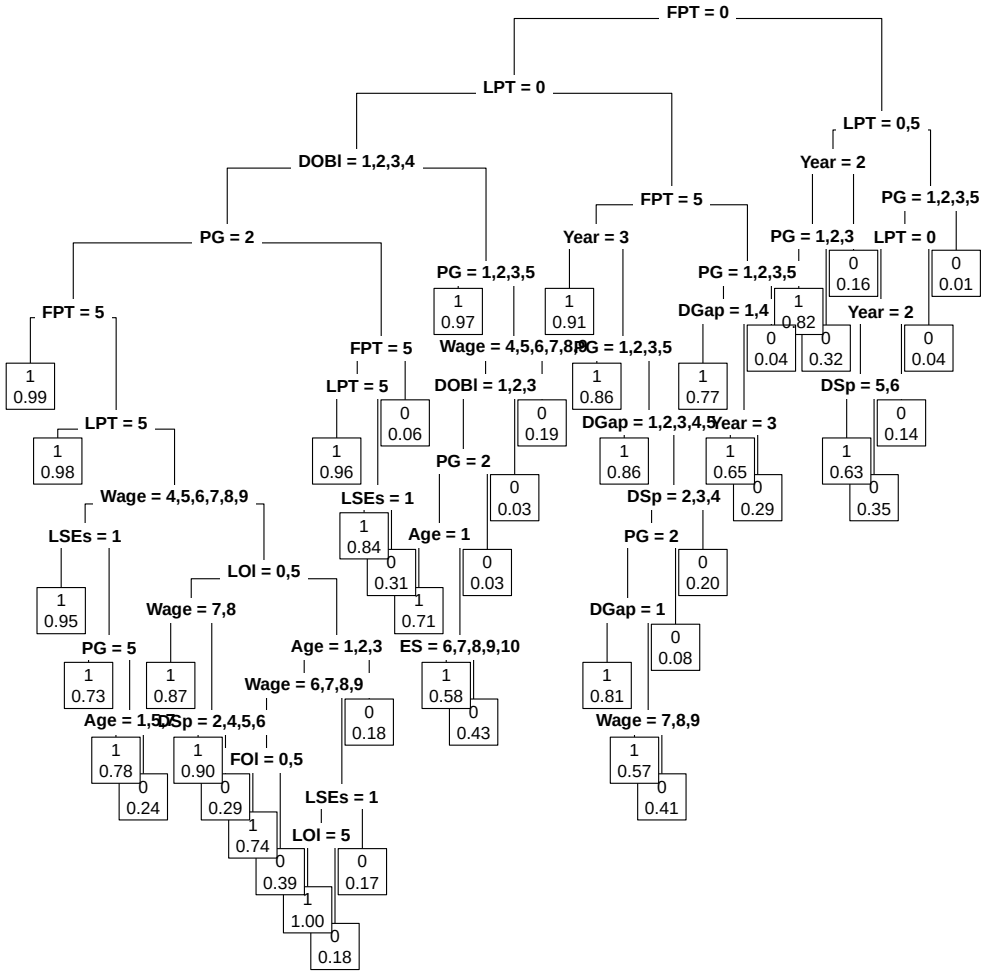


Figure 6: Estimated Classification Tree, Eastern German Women (top) and Men (bottom)



References

Antonczyk, Dirk; Fitzenberger, Bernd; DeLeire, Thomas (March 2010): Polarization and Rising Wage Inequality: Comparing the U.S. and Germany. Discussion paper no. 4842, IZA.

Bertat, Thomas; Dundler, Agnes; Grimm, Christopher; Kiewitt, Jochen; Schomaker, Christine; Schridde, Dr. Henning; Zemann, Dr. Christian (October 2013): Neue Erhebungsinhalte ‘Arbeitszeit’, ‘ausgeübte Tätigkeit’ sowie ‘Schul- und Berufsabschluss’ in der Beschäftigungsstatistik. Methodenbericht, Bundesagentur für Arbeit – Statistik.

Fitzenberger, Bernd (1999): Wages and Employment Across Skill Groups: An Analysis for West Germany. Heidelberg: Physica/Springer.

Imprint

FDZ-Methodenreport 1/2016 (EN)

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Dr. Jörg Heining, Dagmar Theune

Technical production
Dagmar Theune

All rights reserved
Reproduction and distribution in any form, also in parts,
requires the permission of FDZ

Download
http://doku.iab.de/fdz/reporte/2016/MR_01-16_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Johannes Ludsteck
Phone: +49-911-179-4810
Email: Johannes.Ludsteck@iab.de

Ulrich Thomsen
Phone: +49-911-179-1259
Email: Ulrich.Thomsen@iab.de