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

FDZ-Methodenreport

09/2018
EN

Methodological aspects of labour market data

Skilled or unskilled – Improving the information on qualification for employee data in the IAB Employee Biography

Ulrich Thomsen,
Johannes Ludsteck,
Alexandra Schmucker

Bundesagentur für Arbeit

Skilled or unskilled – Improving the information on qualification for employee data in the IAB Employee Biography

Ulrich Thomsen, Johannes Ludsteck, Alexandra Schmucker (IAB)

Documentation version: DOI: 10.5164/IAB.FDZM.1809.en.v1

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

Abstract

This FDZ-Methodenreport describes the implementation of a deterministic imputation model for education in the Employee History (Beschäftigtenhistorik (BeH)) of the Institute for Employment Research (IAB). The imputation model was originally invented by Fitzenberger, Osikominu, and Völter, but now was adapted for the current Employee History. First the extent of the data quality problem is sketched out and different variants of the solving imputation model are evaluated.

Then the implementation of the selected variant in the Employee History is described in detail. Afterwards some descriptive graphs show the resulting imputed data. In order to give the user a good impression about the goodness of the selected imputation model the quality of the imputed information is tested with some regression models. The regression results indicate, which variables in the Employee History have an influence on the quality of the imputed values. This allows an assessment of its usability for certain subgroups in the data.

Zusammenfassung

Dieser FDZ-Methodenreport beschreibt die Umsetzung eines von Fitzenberger, Osikominu und Völter entwickelten deterministischen Imputationsregelwerks für die Ausbildungsinformation in der aktuellen Beschäftigtenhistorik des IAB. Zunächst wird der Umfang des Datenqualitätsproblems skizziert und dann unterschiedliche Imputationsverfahren für die Ausbildungsinformation evaluiert.

Anschließend wird die Implementierung des ausgewählten Imputationsverfahrens in der Beschäftigtenhistorik des IAB detailliert beschrieben. Das Ergebnis des Imputationsverfahrens wird einerseits deskriptiv präsentiert und andererseits seine Güte mittels Regressionsmodellen getestet. Dabei wird gezeigt, welche Merkmale der Beschäftigten Auswirkungen auf die Qualität der imputierten Information haben. Damit wird eine Einschätzung der Nutzbarkeit für verschiedene Untergruppen unter den Beschäftigten ermöglicht.

Keywords: German Employment Register, missing observations, imputation, education.

Acknowledgements: We thank Johann Eppelsheimer for helpful comments and suggestions.

1 Introduction and outline

The level of qualification (education and vocational training) is an important information for labor market research. In Germany one of the central sources for that information are the employment register data. The register data are derived from the social security notifications all establishments throughout Germany with at least one employee subject to social security have to send to the German Pension Fund. The register data are used to build up the official statistics of the German Employment register as well as the Employee History (Beschäftigtenhistorik (BeH)) of the Institute for Employment Research (IAB). Missing information reduce the usefulness of these register data for research in many different ways. Having as complete data as possible is central for the analytic potential of the data. Therefore missing data have to be imputed as far as possible.

In this report[1] we first present the amount of missing values in the qualification information in the data. As there are many ways to impute the missing information, some well-known procedures are presented and their results evaluated using data from a different data source. The finally selected imputation procedure is then explained in its implementation details. In the closing section the results of the final imputation model are presented and its strengths and weaknesses discussed.

2 Problem range of the missing qualification information

When looking at individuals' education over a longer period of time, inconsistencies between different educational levels might occur. So a person might have a higher qualification in the data of the past year than in the current year. As educational downgrading is implausible for most cases, one has to deal with those inconsistent information over time.

But missing information might be an even bigger problem, because the number of employment notifications with missing information on qualification has grown substantially over time. This phenomenon is not distributed equally among the different groups of employees in Germany, but highly disproportionately. Especially people in marginal part-time employment are affected to a larger degree than people in standard employment (see Table 1). This problem becomes much more urgent in 1999, because then a new recording scheme for the social security notifications was introduced. Since then, marginal part-time employment must also be reported. The adoption of the Occupation Code 2010 in the notification procedure caused a further rise of the share of missing values, so that in 2011 half of all notifications show a missing value for the qualification.

[1] Part 2 and 3 of this report were already presented in the appendix (Section 8.1) of the FDZ-Datenreport 03/2016 (EN) (Schmucker et al. 2016)

Table 1: Share of missing values in the qualification variable for selected years

Year	Employment status		
	Total	Marginal	Non-marginal
1975	10%		10%
1985	7%		7%
1995	10%		10%
1999	19%	52%	14%
2000	21%	52%	15%
2005	26%	58%	16%
2010	31%	62%	21%
2011	49%	63%	45%
2012	42%	62%	35%
2014	39%	60%	32%
2016	38%	60%	32%

Source: 2% random sample of the BeH 10.02.01-171117, own calculations

The adoption of the Occupation Code 2010 had another side effect on the reported qualification information. After its introduction in 2011 the employers no longer report educational and vocational qualifications in a combined variable but separately according to school qualifications (none, lower secondary, intermediate secondary, upper secondary) and vocational education and training (none, recognised vocational training, master craftsman, bachelor degree, diploma, doctorate). This split makes it possible to record qualification more precisely, even though the differentiation between university of applied sciences (Fachhochschule) and traditional university is lost. But a problem occurs when looking at longer time series in the data. Then it is necessary to make the recording methods compatible. This can be done by matching each combination of values from the new qualification variables with the closest possible qualification level according to the old occupation code (see Table 2). In that way we are perpetuating the old variable to the time of the new Occupational Code 2010.

Table 2: Assigning the information from the new occupation code to the old occupation code

New qualification variables		Old qualification variable	
Highest general-education school qualification	Highest vocational qualification	Highest general education and vocational qualification	
IF (AND operation)		THEN	
Classification code/contents	Classification code/contents	Classification code	Contents
2 Lower secondary school cert.	1 No vocational qualification	1	Lower secondary school, intermediate secondary school cert. or equivalent school education, no vocational qualification
3 Intermediate secondary school cert. or equivalent qualification	1 No vocational qualification	1	
2 Lower secondary school cert.	2 Recognised vocational qualification	2	Lower secondary school, intermediate secondary school cert. or equivalent school education, with a voca-
3 Intermediate secondary school cert. or equivalent qualification	2 Recognised vocational qualification	2	

New qualification variables		Old qualification variable	
Highest general-education school qualification	Highest vocational qualification	Highest general education and vocational qualification	
IF (AND operation)		THEN	
Classification code/contents	Classification code/contents	Classification code	Contents
2 Lower secondary school cert.	3 Master craftsman / technician or equivalent qualification	2	tional qualification (completed training in a skilled or semi-skilled occupation, qualification from a full-time vocational school / technical college)
3 Intermediate secondary school cert. or equivalent qualification	3 Master craftsman / technician or equivalent qualification	2	
4 Upper secondary school leaving certificate (general or subject-specific)	1 No vocational qualification	3	Upper secondary school leaving certificate (general or subject-specific aptitude for higher education), no vocational qualification
4 Upper secondary school leaving certificate (general or subject-specific)	2 Recognised vocational qualification	4	Upper secondary school leaving certificate (general or subject-specific aptitude for higher education), with vocational qualification (completed training in a skilled or semi-skilled occupation, qualification from a full-time vocational school / technical college)
4 Upper secondary school leaving certificate (general or subject-specific)	3 Master craftsman / technician or equivalent qualification	4	
1 No school qualifications	4 Bachelor degree	6	University degree
2 Lower secondary school cert.	4 Bachelor degree	6	
3 Intermediate secondary school cert. or equivalent qualification	4 Bachelor degree	6	
4 Upper secondary school leaving certificate (general or subject-specific)	4 Bachelor degree	6	
9 Qualification unknown	4 Bachelor degree	6	
1 No school qualifications	5 Diploma/ Magister/ Master degree/ state examination	6	
2 Lower secondary school cert.	5 Diploma/ Magister/ Master's degree/ state examination	6	
3 Intermediate secondary school cert. or equivalent qualification	5 Diploma/ Magister/ Master's degree/ state examination	6	
4 Upper secondary school leaving certificate (general or subject-specific)	5 Diploma/ Magister/ Master's degree/ state examination	6	
9 Qualification unknown	5 Diploma/ Magister/ Master's degree/ state examination	6	
1 No school qualifications	6 Doctorate	6	
2 Lower secondary school cert.	6 Doctorate	6	
3 Intermediate secondary school cert. or equivalent qualification	6 Doctorate	6	
4 Upper secondary school leaving certificate (general or subject-specific)	6 Doctorate	6	
9 Qualification unknown	6 Doctorate	6	
1 No school qualifications	1 No vocational qualification	7	Education and training unknown, no details possible
9 Qualification unknown	1 No vocational qualification	7	

New qualification variables		Old qualification variable	
Highest general-education school qualification	Highest vocational qualification	Highest general education and vocational qualification	
IF (AND operation)		THEN	
Classification code/contents	Classification code/contents	Classification code	Contents
1 No school qualifications	2 Recognised vocational qualification	7	
9 Qualification unknown	2 Recognised vocational qualification	7	
1 No school qualifications	3 Master craftsman / technician or equivalent qualification	7	
9 Qualification unknown	3 Master craftsman / technician or equivalent qualification	7	
1 No school qualifications	9 Qualification unknown	7	
2 Lower secondary school cert.	9 Qualification unknown	7	
3 Intermediate secondary school cert. or equivalent qualification	9 Qualification unknown	7	
4 Upper secondary school leaving certificate (general or subject-specific)	9 Qualification unknown	7	
9 Qualification unknown	9 Qualification unknown	7	

3 Evaluating different imputation procedures

Using the described matching between new and old qualification variables enables the generation of a long time series for the qualification variable in the Employee History (Beschäftigtenhistorik (BeH)) – but the problem of the missing data still remains. In order to solve this problem an imputation of the missing values is needed. There are several possibilities to do that, but for the data sets of the IAB, the deterministic replacement rule that was suggested by Fitzenberger et al. (2005 and 2006) is one of the best known. In order to find an optimum procedure, which is applied as standard imputation for the Employee History dataset, we compare several deterministic and other stochastic imputation algorithms in advance. Based on their performance we declare the best model as standard.

Evaluating the quality of an imputation procedure is done by comparing imputed information with an information that is considered to be the “true” information. The goodness-of-fit between the true and the imputed information indicates the quality of the imputation model. Instead of comparing already known and imputed information for the same cases in the same data set, we decided to select a cross-validation approach. So we looked for a second data source with the needed information for the validation. With the survey “Working and Learning in a Changing World” (Arbeiten und Lernen im Wandel - ALWA) (Antoni et al., 2010) we found a good complementing data source for the comparison. The ALWA survey can be regarded as a very good reference point, because the qualification information were collected in detailed personal interviews and were already checked for consistency during the interviews.

The evaluation is done by first using each considered imputation procedures to impute the qualification information for individuals from the Employee History (Beschäftigtenhistorik (BeH)). Then the imputed values are compared with the qualification levels from the retrospective ALWA survey. The comparison is conducted at individual level. Because of the elaborated collection of the qualification information in the ALWA-survey, the information reported there is taken as the gold-standard or "true" information.

The evaluation compares the following procedures: IP1, IP2, IP3, IP0, IPS1, IPS2, IPS3.

IP1: For each individual, in each observation the information on qualification is replaced with the highest qualification level reported so far. Note, however, that the qualifications "3 – upper secondary school leaving certificate" and "2 – completed vocational training" are not ordered in this sense. For this reason a "2" qualification level is not updated to a "3" (and vice versa). With this extrapolation onto later observations, missing values at the beginning of an employment biography are not imputed. That is done in a second step in which the (chronologically) first valid qualification level is extrapolated onto the preceding observations. Because acquiring qualifications takes time, this extrapolation is only conducted up certain minimum age limits.

IP2 proceeds in the same way as IP1, treating education and training information as valid and extrapolating it onto subsequent observations only if it is repeated identically at least three times.

IP3 also differs from IP1 due to a narrower definition of valid education and training information. Reporting establishments are classified as unreliable (and their education and training information is regarded as missing) if they revise[2] the education and training qualifications repeatedly (at least twice).

IP0, like IP1, fills missing values with the education and training information from preceding observations, but only overwrites missing values. The backward extrapolation is conducted like in IP1. A consistency check that guarantees a monotonically increasing qualification profile (as in IP1, IP2 and IP3) is not conducted here. IP0 was introduced as a conservative minimum variant.

See Fitzenberger et al. (2005 and 2006) for a detailed account and further discussion.

The stochastic imputation procedures IPS1, IPS2, IPS3 are based on a multinomial logit model. It explains the current qualification level using the information from temporally adjacent (lagged and future) observations. As the temporally adjacent information may also be missing, the mode of all the respective person's observations in the past and the mode of all the observations in the future are additionally included in the model. Further predictors are age, wages, category of employment (person group aggregated into 5 classes), gender and (grouped) nationalities.

[2] Revisions in this sense are changes from a higher to a lower qualification.

IPS1 imputes all observations, including observations with valid education and training details. IPS2 only imputes missing values. IPS3 is generated from IPS2 by applying IP1 to the imputed accounts. (This means that IPS3 generates monotonically increasing qualification profiles.)

Outline of the evaluation procedure

As mentioned above, the procedures are evaluated by comparing the imputed education and training biographies generated from the Employee History (BeH) with education and training profiles reconstructed from the retrospective ALWA survey ("true qualification level"). The comparison is based on the full Employee History (BeH) employment biographies of all the participants in the ALWA survey who have at least one Employee History (BeH) observation.[3] The time when a qualification was acquired may differ from the time when it is first reported in the Employee History (BeH); either because the person did not take up a job until after gaining the qualification or because the qualification was not reported correctly by the employer. For this reason, the Employee History (BeH) biographies are combined with the ALWA biographies and cut in such a way as to generate parallel, non-overlapping observations. (Within these observations the qualification level is constant). Then the qualifications from the Employee History (BeH) are compared with the information from the ALWA on a day-to-day basis.

The comparison is based on a binary measure of non-agreement. It has the value 0 when the qualifications from the Employee History (BeH) correspond to those in the ALWA, and 1 otherwise. The non-agreement is calculated as the mean across all observations and persons in the comparison sample, weighted with the durations. Table 3 shows the results of the evaluation.

Table 3: Percentage deviations between imputed qualifications from the BeH and the corresponding results from the ALWA survey

Distance measure	Variable	Incl. missing values	Excl. missing values
Yes(1)/No(0)	**Qualification**	34.64	30.23
	IP0	28.67	30.16
	IP1	24.94	26.16
	IP2	25.51	26.83
	IP3	26.16	27.18
	IPS1	38.48	38.00
	IPS2	31.76	30.23
	IPS3	26.22	26.66

Source: Combined data of ALWA-ADIAB and Employee History, own calculation.

Interpretation of Table 3: The figure 34.64 (top left) indicates that the education and training details from the BeH do not agree with those in the ALWA survey in 34.64% of all observations (when weighted with durations in the evaluation). The figure 30.23 in the next column to the right shows that the non-agreement decreases to 30.23% when no observations in which the

[3] A small number of individuals who did not conduct their interview in German were excluded.

original variable on education and vocational training qualifications is missing are taken into account in the comparison. The Table shows, that IP1 clearly performs best.

The table also shows clearly that the stochastic imputations exhibit larger distances from the ALWA data than the deterministic imputation models do. This result reinforces the decision to use a comparably simple and transparent deterministic model. So the IP1-variante is used for imputing the employment data.

4 Applying IP1 - A detailed description of the implementation

4.1 Some general comments on applying IP1 on the IAB Employee History

In contrast to the original imputation developed by Fitzenberger, Osikuminu, Völter (2005) that was developed for the IAB employment 1% subsample, the imputation described below is applied on the IAB Employee History. It uses the rules of IP1, but the different data set results in some diverging details in the implementation. In the following we describe these divergences.

First, the age is computed for the starting date of each employment episode in the data. Compared with this, Fitzenberger et al. could only use the year of birth.

Second, the complete data set of the IAB-Employee History contains more than 1.6 Billion employment episodes spanning from 1975 up to date. In this long time span in the value of the qualification variable there appears one more value than in the original imputation. This is the value “0 – without training” (keine Ausbildung). It is considered the lowest of all valid values.

Third, episodes with a remuneration information of zero DM / € (but not missing value) are treated as normal episodes in the imputation procedure. Typically employees on maternal leave or longer periods of employment interruption due to illness have a remuneration of zero DM / €. Here the qualification information is considered as valid as in the episodes before, so that the episodes are treated normally in the imputation.

Fourth, the biggest breach with the original Fitzenberger et al. IP1 rules is the non-usage of the employment status (e.g. unskilled worker, skilled worker, master craftsman). The background of this decision is the changing reporting scheme of the Occupational Code. The new Occupational Code 2010 no longer reports the employment status in a way that can be used to derive information on the qualification level. So we decided not to use the employment status anymore.

Having explained the largest differences to the original IP1, we can move on to the description of the detailed implementation in the following chapters.

4.2 Prerequisites for the implementation

Please note, that all computational steps described below are done for each person and its employment biography separately. The coding was done with the statistical program package SAS.

Before the final imputation procedure IP1 can be applied, some basic prerequisites have to be fulfilled. These are the following:

First, just one variable is imputed for the complete time series. Therefore the above mentioned matching of the two new qualification variables with the closest possible qualification level according to the old occupation code is used to construct a comprehensive qualification variable for the Employee History (BeH). There is one exception from that principle. The value 0 meaning "without training" (keine Ausbildung) exists only in the Occupational Codes before 2010. After its introduction in 2011, the value 0 cannot be reproduced from above mentioned matching of the two new qualification variables. So in the final data the value 0 exists only from 1975 to 2011.

Second, the employment episodes of an employed person are chronologically sorted, starting with the beginning of the first employment episode and ending with the last one. If two employment episodes start at the same point of time, the one with a higher qualification is sorted to the first place (sorting criterion: qualification descending).

Third, the age of an employed person is computed once per original episode. It is calculated for the day the original employment notification begins.

Fourth, the imputation procedure is computed on the basis of episode splitted data. Episode splitting means the following:

> *"If observations overlap within an account, these observations are replaced by artificial observations with new dates so that completely parallel periods and non-overlapping periods are created. This increases the number of observations" (Antoni et al., 2016, p. 26f.)*[4]

After the imputation is finished, the episode splitting is reversed, restoring the original data.

4.3 First step of data editing: Writing forward

First, all valid values in the original data are taken to fill up the imputed variable. An exception is made for episodes in which the person is aged under 18. For employees being so young it is assumed, that they cannot have finished a formal vocational training. Therefore all qualification levels between the value 2 and 6 are coded as being 1 in the imputed variable.

Second, for each splitted episode of the person's employment biography, the imputed information on qualification is replaced with the highest qualification level reported so far. So going sequentially through the ordered episodes, every ascending qualification variable results in a higher value in the imputed variable for the current and all following episodes as long as there is no further ascending qualification level.

Third: There is an exception for the described rule of sequential processing. The qualifications "3 – upper secondary school leaving certificate" and "2 – completed vocational training" are

[4] For further information on episode splitting see FDZ-Datenreport SIAB 04/16 (EN) (Antoni et al., 2016, p. 26f.)

not hierarchically ordered. For this reason a qualification level “2” is not updated to a “3” (and vice versa). Instead the combination of the two mentioned values in two consecutive episodes results in a new value for the imputed variable in the second episode. The new value is a combination of the two not hierarchically ordered values. It is the value “4 – upper secondary school leaving with completed vocational training”. Note, that for this comparison, the imputed qualification variable is as well taken into account as the original information and that a gap between the two consecutive episodes is allowed too.

4.4 Second step of data editing: Writing backward

In the first step of writing the valid information forward onto later observations, missing values at the beginning of an employment biography are not imputed. That is done in a second step in which the (chronologically) first valid qualification level is extrapolated onto the preceding observations. In this step information only is imputed, if the original information is a missing value. Finally, because acquiring a qualification takes time, this extrapolation is only conducted up to the following minimum age limits suggested by Fitzenberger, Osikuminu, Völter (2005):

Qualification:	Minimum age:
University	29
University of applied sciences	27
Upper secondary cert. and voc. Training	23
Upper secondary cert., no voc. Training	21
Voc. training, no upper secondary school cert.	20

4.5 Reversing the episode splitting procedure

As the last step of the imputation procedure, the episode splitting is reversed and the original episodes are restored. Therefore one has to make a selection, which of the imputed values for the qualification information in the different split episodes remain for the original episode. If an original episode had more than one split episodes and imputed qualification information, then the highest imputed value of the qualification levels of any split episodes is kept.

But here too, no rule without exception: For employment episodes, with employees aged under 18, the maximum qualification level still is value “1 – Lower secondary school, intermediate secondary school cert. or equivalent school education, no vocational qualification”.

5 Results and effects

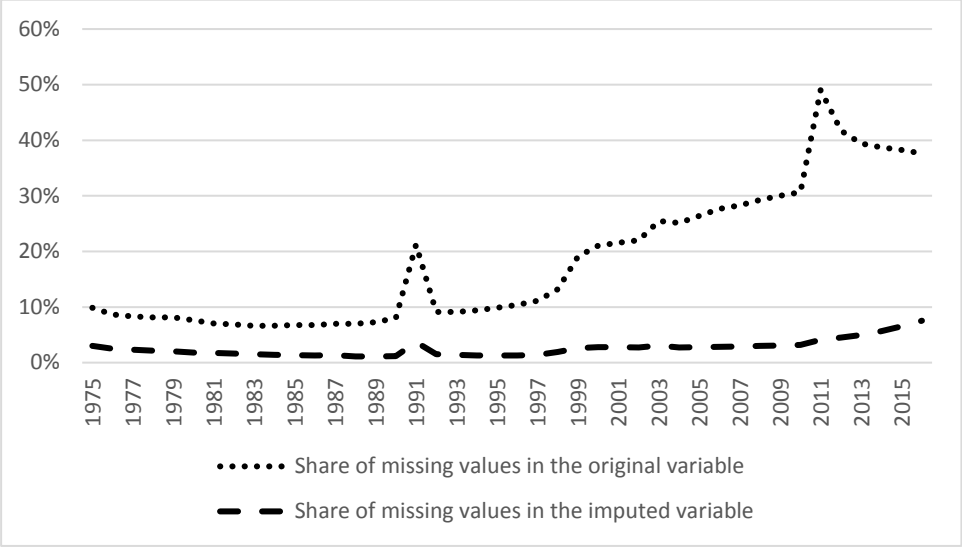
5.1 A first glance: Results on a descriptive level

In figure 1 we present the shares of missing values in the original and in the imputed variable from 1975 until 2016. The original variable shows remarkable increases in the years 1991, 1999 and 2011 which reflect changes in the reporting for the social security notifications. In 1991 the employees in eastern Germany were included in the social security system. But in the beginning many reports were incomplete, in particular the information on education and occupation was missing. In 1999 the compulsory registration was extended to the marginal

employees. As these jobs often only require unskilled tasks the education degree is not relevant for the employers and hence the information is not recorded and not reported in the Social Security notifications. Finally, the introduction of the new occupation code in 2011 lead to a rising number of missing values in the education variable since 2011.

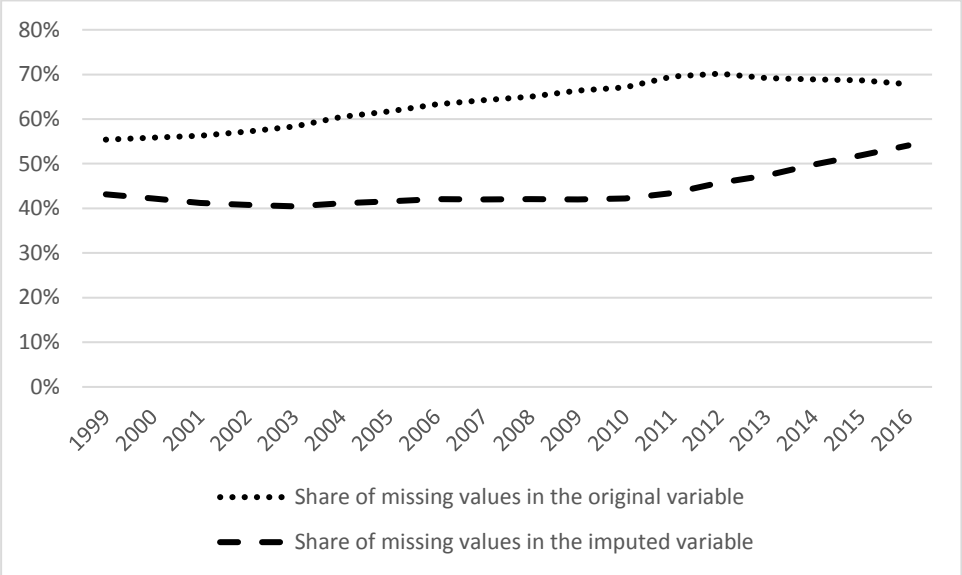
The applied imputation technique was successful as the share of missing values could be reduced to less than 4% in most years. Except the last five years show higher rates of remaining missing values. The reasons for this are very short employment biographies at the end of the observation period which do not provide sufficient information to replace the missing values. E.g. the original education variable of employees which started working in 2016 shows 60 percent missing values, but the imputed variable still has 56 percent missing values. Another limitation of our procedure affects employees which provide only information on marginal employment in our data. As valid information on education often can be obtained from reports on standard employment episodes this is not possible for this group. Hence (as figure 2 shows) the shares of missing values in the original variable are quite high between 56 and 71 percent and could be reduced only to 41 to 55 percent. Fortunately, only 15 percent of the employees in the BeH (1999-2016) belong to this group.

Figure 1: Share of missing values in the original and imputed qualification variable



Source: 2% random sample of the BeH 10.02.01-171117, own calculations

Figure 2: Share of missing values in the original and imputed qualification variable for employees with only marginal part-time jobs



Source: 2% random sample of the BeH 10.02.01-171117, own calculations

5.2 Digging deeper: Detailed effects of the imputation and its quality

On a first glance it became clear that the imputation procedure has varying success for different subgroups of employees. Therefore a detailed view on the quality of the imputed information is advisable. For a detailed evaluation of the quality in the different subgroups again the ALWA-ADIAB survey data[5] are used as reference point and gold standard. These data are combined with the IEB data filtering for the subset of employment time periods.

Like in the selection of the imputation procedure the detailed evaluation is done by comparing the imputed qualification information with the reported qualification levels from the retrospective ALWA survey. Again the comparison is conducted at the individual level of the employee in the Employee History who for the same time period was questioned in the ALWA survey.

To give an impression of the general effects the imputation has, in a first step a direct comparison between the imputed values and the ALWA-information is presented. In the following cross table we present an overview of the IP1- and the corresponding ALWA-information in the combined data. As the number of the comparable ALWA and Employee History time periods is limited, the shares of a matching or differing information is more meaningful. Therefore in Table 4 we present the unweighted shares of the two variables:[6]

[5] For a description of that dataset see FDZ-Datenreport 5/2011 (de) (Antoni et al. 2011).

[6] The cross tables with the original numbers of the corresponding spells can be found in Appendix Table A1, as well as the numbers and shares of the two variables weighted by the duration of the time period. The weighting does not have a large effect.

Table 4: Unweighted Shares of ALWA- and corresponding IP1-education information

IP1	Education from ALWA						
	Missing	No Voc. Train.	Voc. Train.	USSC *	USSC* + Voc.Train.	FH**	Univer-sity
No Voc. Train.	0.000	**0.715**	0.258	0.014	0.009	0.004	0.000
Voc.Train.	.	0.048	**0.887**	0.001	0.041	0.019	0.004
USSC*	0.000	0.208	0.130	**0.419**	0.168	0.031	0.043
USSC* + Voc.Train.	.	0.021	0.356	0.026	**0.515**	0.056	0.026
FH**	.	0.003	0.199	0.008	0.172	**0.436**	0.183
University	0.000	0.007	0.072	0.018	0.072	0.089	**0.742**

USSC* = Upper Secondary School leaving Certificate | FH** = University of Applied Sciences

Source: Combined data of ALWA-ADIAB and Employee History, own calculation.

The table should be read as follows. In the first row the shares of spells with *No Vocational Training* in IP1 with the corresponding ALWA-information are given. So 71.5 percent of the spells have a matching information of *No Vocational Training* in both variables, whereas 25.8 percent show the higher category of *Vocational Training* in the ALWA-data. The shares are calculated for the evaluation of the imputed variable and therefore add up to 1 in each row.

The lower bound of the fitting rates can be found in the IP1-category of *Upper Secondary School leaving Certificate* with no further training. Here just nearly 42 percent of the corresponding time periods show a match. In the ALWA-data this group is often reported to be lower skilled, as the second largest share is the ALWA category of *No Vocational Training*. As this educational level refers to a school leaving certificate, but not to information about completing a training, it is hard to grasp in the operational practice. Therefore the quality of that special kind of educational information can be problematic.

The upper bound of the fitting rates can be found in the educational category of completed *Vocational Training*. Nearly nine out of ten cases with this educational category in IP1-variable report a matching qualification in the ALWA-survey data. As *Vocational Training* is the standard educational category in Germany these high rates are not surprising. The educational category often is associated with classical (industrial) standard employment and can be used as something like the default educational attainment.

Going one step further the quality of the imputed variable can be analysed in certain contexts. For research just considering one subgroup of employees it might be helpful to know, if variable in the dataset influence the quality of the imputation procedure. By looking at those influences, the reader might learn, if for a subgroup defined by a certain variable, the imputed information works better or worse. For this analytical task, the results of simple regression models are presented (see Table 5). For that we use different meaningful variables and regress them on the compared IP1-ALWA-value. The target variable is just true (zero) and false (one) of IP1 compared to the gold standard of ALWA, differentiated by sex.

As a starting point for this data quality assessment we present a regression model that uses the original information of ALWA as a predictor. That model shows generally for which level of (true) education the imputation works better or worse.

Table 5: Regression model "ALWA-information"

ALWA educational level	Bin Men	Bin Women
Vocational Training	-0.134***	-0.182***
Upper Secondary School Leaving Certificate	0.367***	0.314***
Upper Secondary School Leaving Certificate +Vocational Training	0.310***	0.147***
University of applied sciences	0.291***	0.343***
University	0.008	0.038***
Constant	0.222***	0.304***

Source: Combined data of ALWA-ADIAB and Employee History, own calculation.

The reference category is *No Vocational Training*. Looking at the first cell on the left, giving the information for *Voc. Training* for the group of men and the true / false difference in educational information the number of -0.134*** is given. This means, that having a Vocational Training degree in the ALWA-data reduces the probability of having a wrong information in IP1 by 13.4 percent.[7] In contrast to that positive influence stands the *Upper Secondary School leaving Certificate* with no further vocational training. Belonging to this group reduces the probability of a matching IP1-information for men for nearly 37 percent. This is not surprising as already the cross table of the shares of fitting information showed the lowest rates for exactly this group. So, in general the imputation procedure works better for persons with the standard educational level of *Vocational Training* and worse for other groups.

Changing the perspective from looking at the validating ALWA-information to the dataset internal variables, we analyze the influence of different variables in the BeH on the quality of imputed IP1-information. We do the regression models separately for each variable, because most research populations will focus on one or two variables as filter criterion, but not on all of them.[8] The following tables show the results of the different regression models.

[7] The three stars indicate a high significance of that result on the 99.9 percent-level. Two stars indicate a level of significance of 99 percent, one star a level of 95 percent.

[8] In the Appendix Table A2 you can find a complete regression model, regressing all independent variables on the data quality measures. Again two models were run and are presented.

Table 6: Regression models of Employee History variables

	Bin Men	Bin Women
Age		
25-34	0.031***	0.001
35-44	0.032***	0.000
45 and more	0.047***	0.037***
Constant	0.196***	0.260***
Reference Category: Below 25		
Establishment Size		
10-49	-0.006	0.028***
50-99	0.008	0.021***
100-199	0.017**	-0.001
200-499	-0.003	-0.020***
500-999	0.024***	0.002
>=1000	-0.042***	0.002
Constant	0.226***	0.259***
Reference Category: Below 10 employees		
Nationality		
Western Europe	0.088***	0.106***
Eastern Europe	0.145***	0.255***
Turkey	0.273***	0.134***
All Other	0.119***	0.105***
Constant	0.217***	0.261***
Reference Category: Germany		
Wage		
2nd Quintile	-0.043***	-0.011**
3rd Quintile	-0.037***	0.011*
4th Quintile	-0.029***	0.013*
5th Quintile	0.025***	-0.004
Constant	0.233***	0.264***
Reference Category: 1st Quintile		
Person group: Category of Employment	**Bin Men**	**Bin Women**
Apprentice	-0.031***	0.011*
Marginal employment	0.123***	0.017**
All Other	0.142***	0.136***
Constant	0.220***	0.261***
Reference Category: Standard Employment		
Reason for notification		
Deregistration due to change of health insurance	-0.034**	-0.014
Deregistration due to change of contribution group	0.029**	-0.014
Deregistration for other reasons	-0.013	0.009
Annual notification	-0.044***	-0.035***
All other	-0.020	-0.042***
Constant	0.257***	0.294***
Reference Category: Deregistration due to end of employment		

Source: Combined data of ALWA-ADIAB and Employee History, own calculation.

When looking at the numbers a clear pattern can be observed. The most important information concerning the quality of IP1 is nationality. Especially the factor Turkey has a significant and

strong influence on the binary target variables. This is even more pronounced for males than for females as the Turkish nationality drives up the rate of non-matches by 27.3 percent in the male subpopulation and “just” 13.4 percent in the female one. Interestingly the finding flips for the category of Eastern Europe nationality, where for the female group with 25.5 percent a stronger negative influence can be found than for the male one with just 14.5 percent.

The other variables mostly lag such a clear pattern for both sexes. For example we can see a strong effect of the person group on the quality of IP1, but just for the male subgroup whereas the female subgroup shows a more fixed picture. This is a finding that applies to many of the independent variables. The pattern is clearer for the male population than for the female one.

One could image that the size of the establishment plays a key role for the quality of the IP1 information, because in larger companies the human resource department is much more specialized and professional than in small and medium sized companies. Our regression results show, that against our expectation, this is not the case. The size of the establishment shows nearly no influence on the quality of the imputed variable, that itself is based on the educational information reported by the companies. This finding holds true for both sexes.

To sum up the results of the presented regressions, we now know more about the quality of the IP1-variable. Two findings are especially clear and can be considered when using the data.

First, the data quality of the imputed IP1 variable is good for persons with a vocational training and worse for those just having a higher school leaving certificate without a further vocational degree. This might result from the fact, that the former educational level represents a standard level and dominant group and the latter a much smaller group and exotic one. So the deviation from the norm is not reported very well by the companies and therefore not very well imputed.

Second the nationality of the employee is a strong predictor for the quality of IP1. Being non-German decreases data quality of the educational information. Be especially aware of data quality issues when focusing on male subpopulation having a Turkish nationality and female subpopulation having an Eastern European nationality.

6 Conclusion

In this paper we presented a detailed description of a method of improving the qualification information for employees in German Employee History data. We showed the strengths and weaknesses of the selected model in detail.

The imputation procedure works very well for employees being standard employment over a longer period of time. It comes to its limits when the individuals deviate form that employment pattern. So for persons in marginal employment or for very young persons with only a short and perforated employment career the selected imputation model only brings marginal improvement in the qualification information.

Therefore we have to find another way of improving the needed information for these subgroups. One classical way to do that would be enriching the data base for the people we lag enough information in the Employment History. Luckily there are more datasets focusing especially on the mentioned fringes of employment. A good example for that could be data on job seekers in Germany. So a next step forward to better qualification information should be the integration of further data sources into the imputation procedure instead of using another, perhaps more sophisticated, imputation algorithm.

7 References

Antoni, Manfred; Drasch, Katrin; Kleinert, Corinna; Matthes, Britta; Ruland, Michael; Trahms, Annette (2010): Arbeiten und Lernen im Wandel. Teil I: Überblick über die Studie. FDZ-Methodenreport, 05/2010 (de), Nürnberg.

Antoni, Manfred; Ganzer, Andreas; vom Berge, Philipp (2016): Sample of integrated labour market biographies (SIAB) 1975-2014. FDZ-Datenreport, 04/2016 (en), Nürnberg.

Antoni, Manfred; Jacobebbinghaus, Peter; Seth, Stefan (2011): ALWA-Befragungsdaten verknüpft mit administrativen Daten des IAB (ALWA-ADIAB) 1975-2009. FDZ-Datenreport, 05/2011 (de), Nürnberg.

Fitzenberger, Bernd; Osikominu, Aderonke; Völter, Robert (2006): Imputation rules to improve the education variable in the IAB employment subsample. In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, Vol. 126 (3), pp. 405-436.

Fitzenberger, Bernd; Osikominu, Aderonke; Völter, Robert (2005): Imputation rules to improve the education variable in the IAB employment subsample. FDZ-Methodenreport, 03/2005 (en), Nürnberg.

Schmucker, Alexandra; Seth, Stefan; Ludsteck, Johannes; Eberle, Johanna; Ganzer, Andreas (2016): Establishment History Panel 1975-2014. FDZ-Datenreport, 03/2016 (en), Nürnberg.

8 Appendix

Table A1: Comparison of ALWA- and corresponding IP1-education information

Unweighted numbers of ALWA- and corresponding IP1-education information

	Education from ALWA						
IP1	Miss-ing	No Voc. Train.	Voc. Train.	Abitur	Abi+Voc.Train.	FH	Uni
No Voc. Train.	0	14109	5090	278	181	72	6
Voc.Train.	.	3481	64546	58	2990	1380	327
Abitur	0	1732	1081	3481	1398	260	360
Abi+Voc.Train.	.	355	6137	443	8866	966	448
FH	.	22	1639	63	1419	3602	1510
Uni	0	95	985	252	991	1218	10208

Source: Combined data of ALWA-ADIAB and Employee History, own calculation.

Weighted numbers of ALWA- and corresponding IP1-education information

	Education from ALWA						
IP1	Miss-ing	No Voc. Train.	Voc. Train.	Abitur	Abi+Voc.Train.	FH	Uni
No Voc. Train.	0	12532	4023	169	135	60	6
Voc.Train.	.	2590	67582	47	2937	1532	328
Abitur	0	1313	794	2497	951	187	247
Abi+Voc.Train.	.	228	6395	273	8976	978	421
FH	.	13	1784	39	1470	4004	1616
Uni	0	91	1063	216	974	1286	10758

Source: Combined data of ALWA-ADIAB and Employee History, own calculation.

Weighted Shares of ALWA- and corresponding IP1-education information

	Education from ALWA						
IP1	Missing	No Voc. Train.	Voc. Train.	Abitur	Abi+Voc.Train.	FH	Uni
No Voc. Train.	0.000	0.740	0.238	0.010	0.008	0.004	0.000
Voc.Train.	.	0.035	0.901	0.001	0.039	0.020	0.004
Abitur	0.000	0.219	0.133	0.417	0.159	0.031	0.041
Abi+Voc.Train.	.	0.013	0.370	0.016	0.520	0.057	0.024
FH	.	0.001	0.200	0.004	0.165	0.449	0.181
Uni	0.000	0.006	0.074	0.015	0.068	0.089	0.748

Source: Combined data of ALWA-ADIAB and Employee History, own calculation.

Table A2: Full regression models of all target variables and subgroup

Variable	Bin_Men	Bin_Women
ALWA educational level		
Vocational Training	-0.244***	-0.266***
Upper Secondary School Leaving Certificate	0.350***	0.317***
Upper Secondary School Leaving Certificate +Vocational Training	0.203***	0.070***
University of applied sciences	0.183***	0.266***
University	-0.100***	-0.030***
Age		
25-34	0.002	0.001
35-44	0.003	0.010*
45 and more	0.022***	0.047***
Establishment Size		
10-49	-0.011**	0.007
50-99	-0.015**	-0.009
100-199	-0.006	-0.026***
200-499	-0.024***	-0.042***
500-999	-0.010	-0.027***
>=1000	-0.071***	-0.042***
Nationality		
Western Europe	0.087***	0.073***
Eastern Europe	0.116***	0.228***
Turkey	0.269***	0.089**
All Other	0.069**	0.057**
Wage		
2nd Quintile	0.024***	0.017***
3rd Quintile	0.053***	0.028***
4th Quintile	0.059***	0.006
5th Quintile	0.047***	-0.050***
Person group: Category of Employment		
Apprentice	-0.143***	-0.162***
Marginal employment	-0.027*	-0.030***
All Other	-0.067***	-0.053***
Reason for notification		
Deregistration due to change of health insurance	-0.029**	-0.015
Deregistration due to change of contribution group	-0.023*	-0.024**
Deregistration for other reasons	-0.005	0.020
Annual notification	-0.023***	-0.015**
All other	-0.010	-0.024**
Constant	0.322***	0.397***

Source: Combined data of ALWA-ADIAB and Employee History, own calculation.

Imprint

FDZ-Methodenreport 9/2018

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
Dana Müller, Dagmar Theune

Technical production
Dagmar Theune

Download
http://doku.iab.de/fdz/reporte/2018/MR_09-18_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Ulrich Thomsen
Institute for Employment Research (IAB)
Regensburger Str. 104
D-90478 Nürnberg
Phone: +49-911-179-1259
Email: Ulrich.Thomsen@iab.de