SSOAR

Open Access Repository

www.ssoar.info

A new approach for disclosure control in the IAB Establishment Panel: multiple imputation for a better data access

Drechsler, Jörg; Dundler, Agnes; Bender, Stefan; Rässler, Susanne; Zwick, Thomas

Veröffentlichungsversion / Published Version
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
SSG Sozialwissenschaften, USB Köln

Empfohlene Zitierung / Suggested Citation:
Drechsler, J., Dundler, A., Bender, S., Rässler, S., & Zwick, T. (2007). *A new approach for disclosure control in the IAB Establishment Panel: multiple imputation for a better data access.* (IAB Discussion Paper: Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung, 11/2007). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit (IAB). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-320989

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

IAB DiscussionPaper

Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung

No. 11/2007

A New Approach for Disclosure Control in the IAB Establishment Panel

Multiple Imputation for a Better Data Access

Jörg Drechsler, Agnes Dundler, Stefan Bender,
Susanne Rässler, and Thomas Zwick

Bundesagentur für Arbeit

A New Approach for Disclosure Control in the IAB Establishment Panel

Multiple Imputation for a Better Data Access[1]

Jörg Drechsler, Agnes Dundler*, Stefan Bender*,*
Susanne Rässler, and Thomas Zwick***

** Institute for Employment Research (IAB), Regensburger Straße 104,*
90478 Nürnberg, Germany
Joerg.Drechsler@iab.de, Agnes.Dundler@iab.de,
Stefan.Bender@iab.de, Susanne.Raessler@iab.de
*** Centre for European Economic Research (ZEW), L 7, 1,*
68161 Mannheim, Germany
zwick@zew.detor

Auch mit seiner neuen Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

Also with its new series "IAB Discussion Paper" the research institute of the German Federal Employment Agency wants to intensify dialogue with external science. By the rapid spreading of research results via Internet still before printing criticism shall be stimulated and quality shall be ensured.

[1] The research provided in this paper is part of the project “Wirtschaftsstatistische Paneldaten und faktische Anonymisierung“ financed by the Federal Ministry for Education and Research (BMBF) and conducted by the following institutes: Federal Statistical Office Germany, Statistical Offices of the Länder, Institute for Applied Economic Research (IAW), Centre for European Economic Research (ZEW), Institute for Employment Research (IAB). For more information about this project see for instance Ronning and Rosemann (2006) or Ronning et al. (2005). We thank our project partners and the participants of the “UNECE Conference on Data Editing and Imputation”, 25.09-27.09.2006 in Bonn and “The Conference on Privacy in Statistical Databases ’06”, 13.12.-15.12.2006 in Rome, and especially J.M. Abowd, T.E. Raghunathan, D.B. Rubin and J.P. Reiter for their helpful comments on the paper.

Contents

Abstract

For micro-datasets considered for release as scientific or public use files, statistical agencies have to face the dilemma of guaranteeing the confidentiality of survey respondents on the one hand and offering sufficiently detailed data on the other hand. For that reason a variety of methods to guarantee disclosure control is discussed in the literature. In this paper, we present an application of Rubin's (1993) idea to generate synthetic datasets from existing confidential survey data for public release. We use a set of variables from the 1997 wave of the German IAB Establishment Panel and evaluate the quality of the approach by comparing results from an analysis by Zwick (2005) with the original data with the results we achieve for the same analysis run on the dataset after the imputation procedure. The comparison shows that valid inferences can be obtained using the synthetic datasets in this context, while confidentiality is guaranteed for the survey participants.

JEL-Classification: C11, C13, C49, C53

Keywords: confidentiality; multiple imputation; statistical disclosure control; IAB Establishment Panel; synthetic datasets

1 Introduction

In recent years, the public demand for micro data increased dramatically. But statistical agencies face the dilemma that, although they might be willing to provide all the information required, a release of the datasets might not be possible for confidentiality reasons. The natural interest of enabling as much research as possible with the collected data has to stand back behind the confidentiality guaranteed to the survey respondent: Once the confidentiality is in doubt, potential respondents might be less willing to provide sensitive information, might give wrong answers on purpose or might even be unwilling to participate at all - with devastating consequences for the quality of the data collected (Lane 2005).

For that reason, a variety of methods for disclosure control has been developed to provide as much information to the public as possible, while satisfying the disclosure restrictions needed to maintain the quality of the collected data (Willenborg and de Waal, 2001, Abowd and Lane, 2004). Especially for German establishment datasets a broad literature on perturbation techniques with different approaches can be found (for example Brand 2000, Brand 2002, Brand et al. 1999, Gottschalk 2005, Rosemann 2006). However, information loss is a disadvantage for some of these approaches, while for others, the analyst needs to know the techniques used for perturbation or some special software is necessary to achieve valid inferences.

This paper discusses an application of Rubin's (1993) approach to generate synthetic datasets to a panel of establishments in Germany (the IAB Establishment Panel)[1]. Rubin suggests to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, several simple random samples from these fully imputed datasets are released to the public. Because all imputed values are random draws from the posterior predictive distribution of the missing values given the ob-

[1] A slightly modified approach suggested by Little (1993), where only sensitive variables or variables that bear a high risk of disclosure are replaced, has been adopted for some datasets in the US (see for example Abowd and Woodcock, 2001 or Kennickell, 1997).

served values, disclosure of sensitive information is impossible, especially if the released dataset doesn't contain any real data.

For our application, we use a set of variables from the 1997 wave of the IAB Establishment Panel and compare results from a regression run by Zwick (2005) on the original panel with results achieved with the synthetic datasets. We demonstrate that valid statistical inferences can be obtained in this context, while for an intruder, who is interested in the true answers given by a single respondent, the synthetic datasets don't provide any useful information.

The remainder of this paper is organized as follows: Section 2 provides a short overview of the multiple imputation framework and its modifications for disclosure control. Section 3 introduces the two datasets used. Section 4 describes the application of the synthetic data approach to the IAB Establishment Panel. Section 5 evaluates this approach by comparing results from an analysis by Zwick (2005) with the original data with results achieved for the same analysis run on the dataset after the imputation procedure. Section 6 discusses the possible disclosure risk that remains when releasing the synthetic data. The paper concludes with some final remarks.

2 Multiple Imputation

2.1 Multiple Imputation for Missing Data

Missing data is a common problem in surveys. To avoid information loss by using only completely observed records, several imputation techniques have been suggested. Multiple imputation, introduced by Rubin (1978) and discussed in detail in Rubin (1987, 2004), is an approach that retains the advantages of imputation while allowing the uncertainty due to imputation to be directly assessed. With multiple imputation, the missing values in a dataset are replaced by $m > 1$ simulated versions, generated according to a probability distribution for the true values given the observed data. More precisely, let Y_{obs} be the observed and Y_{mis} the missing part of a dataset Y, with $Y=(Y_{mis},Y_{obs})$, then missing values are drawn from the Bayesian posterior predictive distribution of $(Y_{mis}|Y_{obs})$, or an approximation thereof. Typically, m is small, such as $m = 5$. Each of the imputed (and thus completed) datasets is first analyzed by standard methods de-

signed for complete data; the results of the *m* analyses are then combined in a completely generic way to produce estimates, confidence intervals and tests that reflect the missing-data uncertainty. In this paper, we discuss analysis with scalar parameters only, for multidimensional quantities see Little and Rubin (2002, Section 10.2).

To understand the procedure of analyzing multiply imputed datasets, think of an analyst interested in an unknown scalar parameter θ, where θ could be e.g. the mean of a variable, the correlation coefficient between two variables or a regression coefficient in a linear regression.

Inferences for this parameter for datasets with no missing values usually are based on a point estimate $\hat{\theta}$, an estimate for the variance of $\hat{\theta}$, $\hat{V}$ and a normal or Student's *t* reference distribution. For analysis of the imputed datasets, let $\hat{\theta}_i$ and $\hat{V}_i$ for $i = 1,...,m$ be the point and variance estimates for each of the *m* completed datasets. To achieve a final estimate over all imputations, these estimates have to be combined using the combining rules first described by Rubin (1978).

For the point estimate, the final estimate simply is the average of the *m* point estimates $\hat{\theta}_{MI} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$ with $i = 1,...,m$. Its variance is estimated by $T = W + (1+m^{-1})B$, where $W = m^{-1}\sum_{l=1}^{m}\hat{V}_l$ is the "within-imputation" variance $B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta}_{MI})^2$ is the "between-imputation" variance, and the factor $(1+m^{-1})$ reflects the fact that only a finite number of completed-data estimates $\hat{\theta}_i$, $i = 1,...,m$ is averaged together to obtain the final point estimate. The quantity $\hat{\gamma} = (1+m^{-1})B/T$ estimates the fraction of information about θ that is missing due to nonresponse.

Inferences from multiply imputed data are based on $\hat{\theta}_{MI}$, *T*, and a Student's *t* reference distribution. Thus, for example, interval estimates for θ have the form $\hat{\theta}_{MI} \pm t(1-\alpha/2)\sqrt{T}$, where $t(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the *t* distribution. Rubin and Schenker (1986) provided the approximate value $v_{RS} = (m-1)\hat{\gamma}^{-2}$ for the degrees of freedom of the *t* distribution, under the assumption that with complete data, a normal reference distribution

would have been appropriate (that is, the complete data would have had large degrees of freedom). Barnard and Rubin (1999) relaxed the assumption of Rubin and Schenker (1986) to allow for a *t* reference distribution with complete data, and suggested the value $\nu_{BR} = (\nu_{RS}^{-1} + \hat{\nu}_{obs}^{-1})^{-1}$ for the degrees of freedom in the multiple-imputation analysis, where $\hat{\nu}_{obs} = (1-\hat{\gamma})(\nu_{com})(\nu_{com}+1)/(\nu_{com}+3)$ and ν_{com} denotes the complete-data degrees of freedom.

2.2 Fully Synthetic Datasets

In 1993, Rubin suggested to create fully synthetic datasets based on the multiple imputation framework. His idea was, to treat all units in the population that have not been selected in the sample as missing data, impute them according to the multiple imputation approach and draw simple random samples from these imputed populations for release to the public.

For illustration, think of a dataset of size *n*, sampled from a population of size *N*. Suppose further, the imputer has information about some variables *X* for the whole population, for example from census records, and only the information from the survey respondents for the remaining variables *Y*. Let Y_{inc} be the observed part of the population and Y_{exc} the nonsampled units of *Y*. For simplicity, assume that there are no item-missing data in the observed dataset.

Now the synthetic datasets can be generated in two steps: First, construct *m* imputed synthetic populations by drawing Y_{exc} *m* times independently from the posterior predictive distribution $f(Y_{exc}|X,Y_{inc})$ for the *N-n* unobserved values of *Y*. If the released data should contain no real data for *Y*, all *N* values can be drawn from this distribution. Second, make simple random draws from these populations and release them to the public. The second step is necessary as it might not be feasible to release *m* whole populations for the simple matter of data-size. In practice, it is not mandatory to generate complete-data populations. The imputer can make random draws from *X* in a first step and only impute values of *Y* for the drawn *X*.

The analysis of the *m* simulated datasets follows the same lines as the analysis after multiple imputation (MI) for missing values in regular datasets (see Section 2.1). However, the calculation of the total variance

slightly differs from the calculation of the total variance in MI settings for treating missing data:

$$\hat{\text{var}}(\hat{\theta}_{MI}) = T_f = \frac{m+1}{m}B - W$$

This difference is due to the additional sampling from the synthetic units for fully synthetic datasets. Hence, the variance *B* between the datasets already reflects the variance within each imputation. For a formal justification see Raghunathan et al. (2003).

If *m* is large, inferences can be based on normal distributions. For moderate *m*, a *t* reference distribution is more adequate. The degrees of freedom are given by

$$\nu_f = (m-1)(1-r^{-1})^2 \text{ where } r = \frac{(1+m^{-1})B}{W}.$$

A disadvantage of this variance estimate is that it can become negative. For that reason, Reiter (2002) suggests a slightly modified variance estimator that is always positive:

$$T_f^* = \max(0, T_f) + \delta(\frac{n_{syn}}{n}W)$$, where $\delta=1$ if $T_f<0$, and $\delta=0$ otherwise.

Here, n_{syn} is the number of observations in the released datasets sampled from the synthetic population.

3 The Datasets[2]

For the imputation of the IAB Establishment Panel, we use additional information from the German Social Security Data. In the following Section both datasets will be described in detail.

3.1 The German Social Security Data

The German employment register contains information on all employees covered by social security. The basis of the German Social Security Data (GSSD) is the integrated notification procedure for the health, pension and

[2] This chapter follows the description given in Alda, Bender & Gartner (2005).

unemployment insurances, which was introduced in January 1973.[3] This procedure requires employers to notify the social security agencies about all employees covered by social security.

As by definition the German Social Security Data only includes employees covered by social security - civil servants and unpaid family workers for example are not included - approx. 80% of the German workforce[4] are represented. However, the degree of coverage varies considerably across the occupations and the industries.

The notifications of the GSSD include for every employee, among other things, the workplace and the establishment identification number. We use this number to match the selected establishment characteristics aggregated from the employment register with the IAB Establishment Panel. As we use the 1997 wave of the panel, data are taken from the register for June, 30th 1997 (see Figure 5 in the Appendix for all characteristics used).

3.2 The IAB Establishment Panel

The IAB Establishment Panel[5] is based on the employment statistics aggregated via the establishment number as of 30 June of each year. Consequently the panel only includes establishments with at least one employee covered by social security. The sample is drawn following the principle of optimum stratification. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region, and 16 classes for the industry[6]. These cells are also used for weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. For the first wave, 4,265 establishments were interviewed in Western Germany in the third quarter of 1993. Since then the Establishment Panel has been conducted annually – since 1996 with over 4,700 establishments in Eastern Germany in addition. The re-

---

3 On the structure of the insurance number and on the data office of the pension insurance providers cf. Steeger (2000).

4 An overview of the data is given in Bender, Hass, and Klose (2000), a detailed description can be found in Bender, Hilzendegen, Rohwer, and Rudolph (1996).

5 The approach and structure of the establishment panel are described for example by Bellmann (2002) and Kölling (2000).

6 From 2000 onwards 20 industry classes are used.

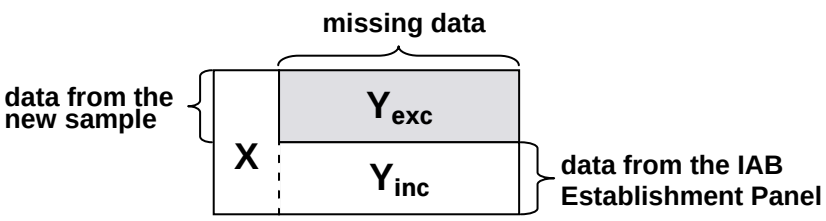
sponse rate of units that have been interviewed repeatedly is over 80%. Each year, the panel is accompanied by supplementary samples and follow-up samples to include new or reviving establishments and to compensate for panel mortality. The list of questions contains detailed information about the firms' personnel structure, development and personnel policy. An overview of available information in 1997 is listed in the Appendix, Figure 5.

4 Application to the IAB Establishment Panel

4.1 Generating Synthetic datasets

In a first step, we only impute values for a set of variables from the 1997 wave of the IAB Establishment Panel. As it is not feasible to impute values for the millions of establishments contained in the German Social Security Data for 1997, we sample from this frame, using the same sampling design as for the IAB Establishment Panel: Stratification by establishment size, region and industry (see Table 4 in the Appendix for an example). Every stratum contains the same number of units as the observed data from the 1997 wave of the Establishment Panel. We gain further information by adding variables from the German Social Security Data and matching these variables to the observations in the Establishment Panel via establishment identification number. After matching, every dataset is structured as follows: Let N be the total number of units in the newly generated dataset, that is the number of units in the sample n_s plus the number of units in the panel n_p, $N=n_s+n_p$. Let X be the matrix of variables with information for all observations in N. Then X consists of the variables establishment size, region and industry and the variables added from the German Social Security Data (see Figure 5 in the Appendix). Let Y be the selected variables from the Establishment Panel, with $Y=(Y_{inc}, Y_{exc})$, where Y_{inc} are the observed values from the Establishment Panel and Y_{exc} are the hypothetic missing data for the newly drawn values in X (see Figure 1).

Fig. 1. The full MI approach for the IAB Establishment Panel



Now, values for the missing data can be imputed as outlined in Section 2 by drawing Y_{exc} m times independently from the posterior predictive distribution $f(Y_{exc}|X,Y_{inc})$ for the N-n_p unobserved values of Y.

After the imputation procedure, all observations from the Establishment Panel are omitted and only the imputed values are kept for analysis. Results from this analysis can be compared with the results achieved with the real data.

4.2 Drawing a New Sample from the German Social Security Data

Due to panel mortality a supplementary sample has to be drawn for the IAB Establishment Panel every year. In the 1997 wave, this supplementary sample primarily consisted of newly founded establishments because in that year the questionnaire had a focus on new foundations. Therefore, start-ups are overrepresented in the sample. Arguably, answers from these establishments differ systematically from the answers provided by establishments existing for several years. Drawing a new sample without taking this oversampling into account could lead to a sample after imputation that differs substantially from that in the Establishment Panel.

For simplicity reasons, we define establishments not included in the German Social Security Data before July 1995 as new foundations and delete them from the sampling frame and the Establishment Panel. For the 1997 wave of the Establishment Panel, this means a reduction from 8,850 to 7,610 observations. In a later stadium of the project, we will analyse the influence of new foundations on answers given in the survey.

Additionally, we have to make sure that every establishment in the survey is also represented in the German Social Security Data for that year. Merging the two datasets using the establishment identification number reveals that 278 units from the panel are not included in the employment statistics. These units are also omitted leading to a final sample of 7,332 observations.

Furthermore, we have to verify that the stratum parameters size, industry and region match in both datasets. Merging indicates that there are some

differences between the two records. If the datasets differ, values from the employment statistics are adopted.

Cross tabulation of the stratum parameters for the 7,332 observations in our sample provides a matrix containing the number of observations for each stratum. For example, one cell of the matrix contains companies specialized in investment goods that are located in Berlin-West with 20 to 49 employees (see Table 4 in the Appendix). Now, a new dataset can be generated easily by drawing establishments from the German Social Security Data according to this matrix.

5 Comparison Between the Original and the Imputed Dataset

5.1 A regression by Zwick (2005) as a means of evaluation

To evaluate the quality of the synthetic data, we compare analytic results achieved with the original data with results from the imputed data. Basis is an analysis by Thomas Zwick: 'Continuing Vocational Training Forms and Establishment Productivity in Germany' published in the German Economic Review, Vol. 6(2), pp. 155-184 in 2005.

Zwick analyses the productivity effects of different continuing vocational training forms in Germany. He argues that vocational training is one of the most important measures to gain and keep productivity in a firm. For his analysis he uses the waves 1997 to 2001 from the IAB Establishment Panel.

In 1997 and 1999 the Establishment Panel included the following additional question that was asked if the establishment did support continuous vocational training in the first part of 1997 or 1999 respectively: 'For which of the following internal or external measures were employees exempted from work or were costs completely or partly taken over by the establishment?' Possible answers were: formal internal training, formal external training, seminars and talks, training on the job, participation at seminars and talks, job rotation, self-induced learning, quality circles, and additional continuous vocational training. Zwick examines the productivity effects of these training forms and demonstrates that formal external training, formal internal training and quality circles do have a positive im-

pact on productivity. Especially for formal external courses the productivity effect can be measured even two years after the training.

To detect why some firms offer vocational training and others not, Zwick runs a probit regression using the 1997 wave of the Establishment Panel. The regression (see Table 5 in the Appendix for details) shows that establishments increase training if they expect to lose workers. One reason could be that the market for skilled labour in Germany is small and establishments have difficulties in finding new skilled workers. Furthermore, establishments tend to offer more training if high qualification needs are expected. This is also the case if establishments give a higher priority to additional apprenticeship training and continuing vocational training efforts instead of hiring externally qualified employees when they have vacancies for skilled jobs. Larger establishments tend to qualify employees more often because they usually have own training departments and can therefore train workers more efficiently. For firms with a high share of qualified employees, state-of-the-art technical equipment or investments in information and communication technology (IT) it is also essential to offer more training. Collective wage agreements are often associated with fringe benefits such as training, while works councils usually attach high importance to continuing vocational training. Therefore both have a positive effect on the amount of training offered.

In the regression, Zwick uses two variables (investment in IT and the co-determination of the employees) that are only included in the 1998 wave of the Establishment Panel. Moreover, he excludes some observations based on information from other years. As we impute only the 1997 wave eliminating newly founded establishments, we have to rerun the regression, using all observations except for newly founded establishments and deleting the two variables which are not part of the 1997 wave. Results from this regression are given in Table 6 in the Appendix and it is evident that the new regression differs only slightly from the original regression. All the variables significant in Zwick's analysis are still significant. Only for the variable "high number of maternity leaves expected", the significance level decreases from 1% to 5%.

5.2 Results from the Fully Synthetic Datasets

For his analysis, Zwick runs the regression only on units with no missing values for the regression variables, loosing all the information on establishments that did not respond to all questions used. This might lead to biased estimates if the assumption of a missing pattern that is completely at random (see for example Rubin 1987) does not hold.

For that reason, we compare the regression results from the synthetic datasets that by definition have no missing values, with the results, Zwick would have achieved if he would have run his regression on a dataset with all the missing values multiply imputed.

To create the synthetic datasets we draw ten new samples from the German Social Security Data as described in Section 4.2 and impute every sample ten times using chained equations as implemented in the software IVEware by Raghunathan, Solenberger and Hoewyk. For the imputation procedure we use 26 variables from the GSSD and reduce the number of panel variables to be imputed to 48 to avoid multicollinearity problems. Comparing results from Zwick`s regression run on the original data and on the synthetic data are presented in Table 1.

All estimates are very close to the estimates from the real data and except for the variable “high number of maternity leaves expected”, for which the significance level decreases to 5% in the synthetic data, remain significant on the same level when using the synthetic data. For all the variables excluding the dummy variable that indicates establishments with 200 to 499 employees, the “true” value from the original dataset lies in the 95% confidence interval of the estimates from the synthetic datasets. This establishment size variable together with the dummy variable for establishments with more than 1,000 employees are the only two variables, for which the absolute deviation between the estimates from the two datasets is higher than 0.1 (0.152 and 0.187 respectively). Obviously Zwick would have come to the same conclusions in his analysis, if he would have used the synthetic data instead of the real data.

Table 1: Comparison between the regression coefficients from the real data and the coefficients from the synthetic data

Exogenous variables	Coeff.from org. data	Coeff.from synth. data	$\beta_{syn} - \beta_{org}$
Redundancies expected	0.250***	0.251***	0.001
Many employees are expected to be on maternity leave	0.266**	0.244*	-0.021
High qualification need exp.	0.648***	0.625***	-0.023
Apprenticeship training reaction on skill shortages	0.113*	0.147*	0.034
Training reaction on skill shortages	0.527***	0.523***	-0.004
Establishment size 20-199	0.686***	0.645***	-0.041
Establishment size 200-499	1.355***	1.203***	-0.152
Establishment size 500-999	1.347***	1.340***	-0.007
Establishment size 1000 +	1.964***	1.778***	-0.187
Share of qualified employees	0.778***	0.820***	0.043
State-of-the-art technical equipment	0.169***	0.168***	-0.001
Collective wage agreement	0.254***	0.313***	0.059
Apprenticeship training	0.484***	0.406***	-0.078
Pseudo R^2	0.32	0.30	
Number of observations	7,332	7,332	
15 sector dummies and East Germany dummy	Yes		

Notes: *** Significant at the 0.1% level, ** Significant at the 1% level, * Significant at the 5% level; the standard errors are heteroscedasticity-corrected.

Source: IAB Establishment Panel 1997 without newly founded establishments and establishments not represented in the employment statistics of the German Federal Employment Agency; regression according to Zwick (2005)

A closer look at the variables used in the analysis further confirms the good quality of the imputation results. Table 2 compares the means for these variables in both datasets. For most of them, the relative deviation of the means is lower than five percent. Only the variable that indicates if many employees are expected to be on maternity leave shows a deviation, that is more than 10%, but one has to bear in mind the low percentage of establishments that expect this to happen (7.37% in the original data). Therefore a relative deviation of 14.34% stems from an absolute deviation that is lower than 0.01. In general the absolute deviation is very low, never higher than 0.05, once more underlining the good results achievable with the synthetic data.

Table 2. Comparison between the means from the real data and the means from the synthetic data for the variables used by Zwick

Regression variables	survey mean	synthetic data mean	relative deviation	absolute deviation
training yes/no	0.7070	0.7109	0.55%	0.0039
Redundancies expected	0.2239	0.2223	-0.75%	-0.0017
Many employees are expected to be on maternity leave	0.0645	0.0737	14.34%	0.0092
High qualification need exp.	0.1550	0.1551	0.02%	0.0001
Apprenticeship training reaction on skill shortages	0.3619	0.3655	1.00%	0.0036
Training reaction on skill shortages	0.4494	0.4678	4.10%	0.0184
Establishment size 20-199	0.3973	0.4043	1.77%	0.0070
Establishment size 200-499	0.1348	0.1439	6.78%	0.0091
Establishment size 500-999	0.0745	0.0769	3.30%	0.0025
Establishment size 1,000 +	0.0942	0.0977	3.71%	0.0035
Share of qualified employees	0.6740	0.6271	-6.96%	-0.0469
State-of-the-art technical equipment	0.6512	0.6861	5.35%	0.0349
Collective wage agreement	0.7643	0.7535	-1.41%	-0.0108
Apprenticeship training	0.6141	0.6247	1.73%	0.0106

These results indicate that valid statistical inferences can be achieved using the synthetic datasets, but is the confidentiality of the survey respondents guaranteed? In our case disclosure of potentially sensitive information is possible, when the following two conditions are fulfilled:

1. An establishment is included in the original dataset and in at least one of the newly drawn samples.

2. The original values and the imputed values for this establishment are nearly the same.

6 Assessing the Disclosure Risk

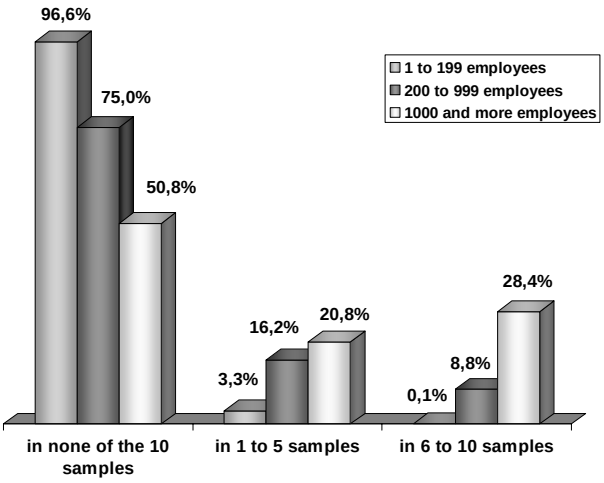
Re-identification of survey respondents can be achieved by intruders if they link external datasets (for example publicly available business or credit information databases) containing specific characteristics and names with the confidential survey data, hoping to get a single match. To determine the disclosure risk in our setting it is necessary to find out, how many of the establishments from the original IAB Establishment Panel (wave 1997) are also contained in the synthetic datasets and how close the imputed values of these establishments are to the original ones.

As described in Section 5.2 we draw ten new samples from the sampling frame and impute every sample 10 times, ending up with 100 imputed datasets that have to be examined. 61.0 percent of the establishments included in the original survey do not occur in any of the 10 new drawn samples. 14.9 percent are contained in one of the 10 samples while only 5.5 percent can be found more than five times (see Table 3). Larger establishments have a higher probability of inclusion in the original survey (for some of the cells of the stratification matrix this probability is close to one). Since we use the same sampling design for drawing new establishments for our synthetic datasets, this means that larger establishments also have a higher probability to be included in the original survey and in at least one of the new samples. Keeping that in mind, having only 25% of establishments between 200-999 employees and 49% of establishments with 1000+ employees in at least one of the new samples is a very good result in terms of data confidentiality (see Figure 2).

Table 3: Establishments from the IAB-Establishment Panel that also occur in at least one of the new samples

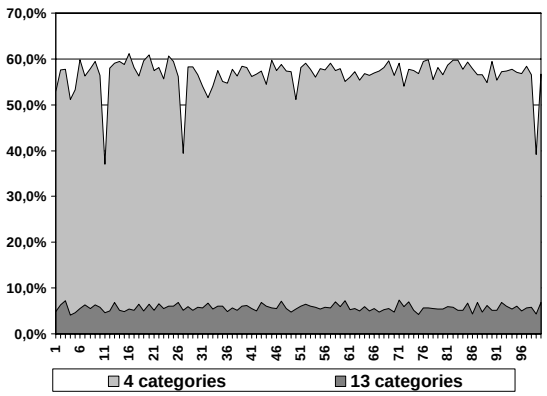
Occurrence in the sample(s)	Number of Records	Percentage
None	4,469	61.0%
1	1,091	14.9%
2	535	7.3%
3	362	4.9%
4	275	3.8%
5	199	2.7%
6	144	2.0%
7	89	1.2%
8	53	0.7%
9	32	0.4%
10	83	1.1%
Total	7,332	100%

Fig. 2: Occurrence of establishments already included in the original survey by establishment size



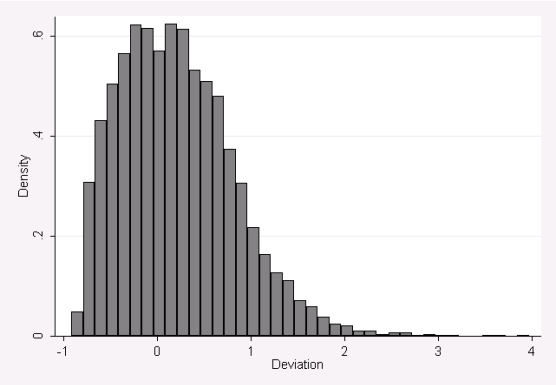
The second step of our evaluation takes a closer look at the establishments from the survey that appear at least once in the newly drawn samples. Using only these establishments the differences between original and imputed values can be detected. Binary variables tend to have a matching rate between 60 percent and 90 percent. Multiple response questions with few categories show a high rate of identical answers in the total item block, too. But with an increase in the number of categories this rate decreases rapidly. For example, for an imputed multiple response variable consisting of 4 categories, the probability of having the same values for all 4 categories is round 57 percent. This probability decreases to round 6 percent if the number of categories climbs up to 13 (see Figure 3).

Fig. 3: Multiple response questions (identical answers in whole item block)



Imputed numeric variables always differ more or less from the original value. To evaluate the uncertainty for an intruder wanting to identify an establishment using the imputed data, we examine the variable *establishment size* for the 83 establishments that appear in all 100 datasets. The average relative difference between the imputed and the original values is 21%. A plot of the distribution of the relative difference shows that there are outliers for which the imputed values are two, three or even four times higher than the original ones (see Figure 4). Thus, for an intruder who wants to identify an establishment using his knowledge of the true size of the establishment, the imputed variable *establishment size* will hardly be of any use.

Fig. 4: Histogram of the relative difference between original and imputed values for the variable establishment size



Observations	8,300
Mean	.2072634
Standard Deviation	.6101745

Summing up the second step, we find that for establishments, which are represented in both datasets, up to 90 percent of some imputed binary variables are identical to the original values. But just one binary variable won't be sufficient to identify a single establishment. Using more binary variables, the risk of identical values will decrease quickly. If, for example, we assume the intruder needs five binary variables for identification and the variables are independently distributed, the risk will be $0.9^5=0.59$. Still, this only holds, if the establishment she or he is looking for is really included in the synthetic data which is very unlikely to begin with. Normally an intruder needs variables with more information than just two categories for a successful re-identification. But as shown for the variable

establishment size, the chance of identifying an establishment by combining information from numeric and categorical variables is almost zero.

7 Concluding Remarks

In this paper we discuss an application of Rubin's (1993) approach to generate fully synthetic datasets to the German IAB Establishment Panel. Releasing these synthetic datasets has the advantage that for an intruder, who is interested in the true values from a single respondent, the synthetic data is useless since fully synthetic datasets don't contain any real values. For researchers however, the datasets still provide all the required information, since their main interest lies on aggregated information like (sub)population means, correlations, variances or information from regressions run on the data. If the imputation model is carefully selected, the correlation structure from the original data is preserved and inference for the synthetic data is the same as for the real data.

For evaluation, we use a typical state-of-the-art analysis by Zwick (2005) on the 1997 wave of the IAB Establishment Panel and compare the results he achieved with the original data with results, the synthetic datasets would have provided. We find that the regression coefficients are almost identical and Zwick would have drawn the same conclusions in his paper if he would have used the synthetic datasets. Some descriptive comparisons of the means of Zwick`s regression variables from the original and from the synthetic datasets further emphasize the good quality of the imputation results.

From the data protection perspective, we show that generating synthetic datasets is an appropriate way of guaranteeing confidentiality. In our setting an intruder has to face two levels of uncertainty: For most establishments, the probability that the establishments of interest are included in the imputed datasets is very low and if they are included, there is no guarantee that the imputed values are (near) the original ones.

Disclosure control to some extent naturally leads to information loss, since the data has to be manipulated in some way. In our paper, we are able to demonstrate that multiple imputation for disclosure control can maintain inference for descriptive as well as for regression analysis. Still, the quality of the synthetic data strongly depends on the imputation model, so gen-

erating imputations only for selected variables decreases the risk of biased estimates. For that reason we will apply the partially synthetic approach to the IAB Establishment Panel in a next step.

References

1. Abowd, J.M., Lane, J.: New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. Privacy in Statistical Databases. Springer Verlag, New York (2004) 282-289

2. Abowd, J.M., Woodcock, S.D.: Disclosure limitation in longitudinal linked data. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam (2001) 215-277

3. Abowd, J.M., Woodcock, S.D.: Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data. Privacy in Statistical Databases. Springer Verlag, New York (2004) 290-297

4. Alda, H., Bender, S., Gartner, H.: The Linked Employer-Employee Dataset of the IAB (LIAB). IAB Discussion Paper, No. 6 (2005)

5. Barnard, J., Rubin, D.B.: Small-sample Degrees of Freedom With Multiple Imputation. Biometrika, Vol. 86 (1999) 948-955

6. Bellmann, L.: Das IAB-Betriebspanel - Konzeption und Anwendungsbereiche. Journal of the German Statistical Society, Vol. 86 (2002) 177-188

7. Bender, S., Haas, A., Klose, C.: The IAB Employment Subsample 1975-1995. Journal of Applied Social Science Studies, Vol. 120 (2000) 649-662

8. Bender, S., Hilzendegen, J., Rohwer, G., Rudolph, H.: Die IAB Beschäftigtenstichprobe 1975-1990. Beiträge zur Arbeitsmarkt- und Berufsforschung, No. 197 (1996)

9. Brand, R.: Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos. Beiträge zur Arbeitsmarkt- und Berufsforschung, Bd. 237 (2000)

10. Brand, R.: Masking through Noise Addition. Inference Control in Statistical Databases. Springer Verlag, Berlin Heidelberg (2002) 97-116

11. Brand, R., Bender, S., Kohaut, S.: Possibilities for the creation of a scientific-use file for the IAB-Establishment-Panel. Statistical Data Confidentiality Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Confidentiality Held in Thessaloniki in March 1999. Eurostat, Brüssel 57-74.

12. Gottschalk, S.: Unternehmensdaten zwischen Datenschutz und Analysepotenzial. ZEW Wirtschaftsanalysen, Bd. 76, Nomos Verlag, Baden Baden (2005)

13. Kennickell, A.B.: Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. Record Linkage Techniques. National Academy Press, Washington D.C. (1997) 248-267

14. Kölling, A.: The IAB-Establishment Panel. Journal of Applied Social Science Studies, Vol. 120 (2000) 291-300

15. Lane, J.: Optimizing the Use of Micro-data: An Overview of the Issues. Paper presented at the Joint Statistical Meetings. http://client.norc.org/jole/SOLEweb/Accesstomicrodata%5B1%5D.pdf. (2005)

16. Little, R.J.A.: Statistical Analysis of Masked Data, Journal of Official Statistics, Vol. 9 (1993) 407-426

17. Little, R.J.A., Rubin, D.B.: Statistical Analysis With Missing Data. John Wiley & Sons, Hoboken (2002)

18. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, Vol. 19 (2003) 1-16

19. Reiter, J.P.: Satisfying Disclosure Restrictions with Synthetic Data Sets. Journal of Official Statistics, Vol. 18 (2002) 531-544

20. Ronning, G., Rosemann, M.: Estimation of the Probit Model From Anonymized Micro Data. Work Session on Statistical Data Confidentiality, Geneva, 9-11 November 2005. Monograph of Official Statistics. Eurostat, Luxemburg (2006) 207-216

21. Ronning, G., Rosemann, M., Strotmann H.: Post-Randomization under Test: Estimation of a Probit Model. Journal of Economics and Statistics, Vol. 225 (2005) 544-566

22. Rosemann, M.: Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. IAW (2006)

23. Rubin, D.B.: Multiple Imputation in Sample Surveys - a Phenomenological Baysian Ap-proach to Nonresponse. American Statistical Association Proceedings of the Section on Survey Research Methods (1978) 20-40

24. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York (1987)

25. Rubin, D.B.: Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, Vol. 9 (1993) 462-468

26. Rubin, D.B.: The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys. The American Statistician, Vol. 58 (2004) 298-302

27. Rubin, D.B., Schenker, N.: Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. Journal of the American Statistical Association, Vol. 81 (1986) 366-374

28. Steeger, W.: 25 Jahre Datenstelle der Rentenversicherungsträger (DSRV). Deutsche Rentenversicherung, 10-11/2000, (2000) 648-684

29. Willenborg, L. and de Waal, T.: Elements of Statistical Disclosure Control. Springer-Verlag, New York (2001)

30. Zwick, T.: Continuing Vocational Training Forms and Establishment Productivity in Germany. German Economic Review, Vol. 6(2), (2005) 155-184

Appendix

Fig. 5: Data comparison

Information contained in the IAB Establishment Panel (wave 1997)

Available for establishments in the survey

- number of employees in June 1996
- qualification of the employees
- number of temporary employees
- number of agency workers
- working week (full-time and overtime)
- the firm's commitment to collective agreements
- existence of a works council
- turnover, advance performance and export share
- investment total
- overall wage bill in June 1997
- technological status
- age of the establishment
- legal form and corporate position
- overall company-economic situation
- reorganisation measures
- company further training activities
- additional information on new foundations

Information contained in the German Social Security Data (from 1997)

Available for all German establishments with at least one employee covered by social security

- number of full-time and part-time employees
- short-time employment
- mean of the employees age
- mean of wages from full-time employees
- mean of wages from all employees
- occupation
- schooling and training
- number of employees by gender
- number of German employees

Covered in both datasets

- **establishment number, branch and size**
- **location of the establishment**
- **number of employees in June 1997**

Table 4: Stratification matrix

Federal state	Branch of trade (16 categories)							
	Establishment size[7]	1 Agriculture, forestry	2 Mining and quarrying	3 Raw material processing	4 Investment goods	...	16 Non-profit organization	Total
Berlin-West	1 0-4	0	0	1	1	...	6	42
	2 5-9	2	0	0	2	...	0	25
	3 10-19	1	0	2	4	...	3	35
	4 20-49	0	1	1	4	...	5	29
	5 50-99	0	0	1	3	...	1	13
	6 100-199	1	0	2	2	...	2	31
	...	...	...	...	...	...	...	...
	10 5,000+	0	1	0	0	...	1	5
	Total	4	3	9	28	...	40	275
Berlin-East	1 0-4	0	0	0	0	...	1	52
	2 5-9	0	0	1	6	...	3	45
	...	...	...	...	...	...	...	...
	10 5,000+	0	0	0	0	...	1	1
	Total	3	2	4	30	...	41	303
Brandenburg	1 0-4	5	0	2	7	...	8	96
	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...

[7] Number of employees covered by social security

Table 5: Probit estimation to explain if an establishment trains or not from Zwick (2005)

Exogenous variables	Coeffcients	z-Value
Redundancies expected	0.303***	4.72
Many employees are expected to be on maternity leave	0.332**	3.21
High qualification need exp.	0.565***	6.94
Apprenticeship training reaction on skill shortages	0.222***	4.32
Training reaction on skill shortages	0.652***	13.08
Establishment size 20-199	0.616***	12.67
Establishment size 200-499	1.119***	10.47
Establishment size 500-999	1.239***	7.32
Establishment size 1,000 +	1.661***	5.38
Co-determination	0.258***	3.81
Share of qualified employees	0.633***	9.03
State-of-the-art technical equipment	0.199***	4.65
Investor in IT	0.244***	5.29
Collective wage agreement	0.213***	4.82
Apprenticeship training	0.457***	10.01
15 sector dummies and East Germany dummy	Yes	
Pseudo-R^2	0.32	
Number of observations	5,629	

Notes: *** Significant at the 0.1% level, **Significant at the 1% level; the standard errors are heteroscedasticity-corrected.

Source: Zwick (2005), p. 169.

Table 6: Probit estimation to explain if an establishment trains or not after modifications described in Section 5.1

Exogenous variables	Coeffcients	z-Value
Redundancies expected	0.261***	4.58
Many employees are expected to be on maternity leave	0.252*	2.49
High qualification need expected	0.641***	8.10
Apprenticeship training reaction on skill shortages	0.176***	3.40
Training reaction on skill shortages	0.597***	11.91
Establishment size 20-199	0.683***	15.19
Establishment size 200-499	1.351***	15.71
Establishment size 500-999	1.398***	11.75
Establishment size 1,000 +	1.972***	9.15
Share of qualified employees	0.766***	10.28
State-of-the-art technical equipment	0.175***	4.16
Collective wage agreement	0.245***	5.46
Apprenticeship training	0.420***	9.31
15 sector dummies and East Germany dummy	Yes	
Pseudo-R^2	0.32	
Number of observations	6,258	

Notes: *** Significant at the 0.1% level, **Significant at the 1% level; * Significant at the 5% level, the standard errors are heteroscedas ticity-corrected.

Source: IAB Establishment Panel 1997 without newly founded establishments and establishments not represented in the employment statistics of the German Federal Employment Agency; regression according to Zwick (2005).

Recently published

No.	Author(s)	Title	Date
1/2004	Bauer, T. K. Bender, S. Bonin, H.	Dismissal protection and worker flows in small establishments published in: Economica	7/04
2/2004	Achatz, J. Gartner, H. Glück, T.	Bonus oder Bias? : Mechanismen geschlechtsspezifischer Entlohnung published in: Kölner Zeitschrift für Soziologie und Sozialpsychologie 57 (2005), S. 466-493 (revised)	7/04
3/2004	Andrews, M. Schank, T. Upward, R.	Practical estimation methods for linked employer-employee data	8/04
4/2004	Brixy, U. Kohaut, S. Schnabel, C.	Do newly founded firms pay lower wages? : first evidence from Germany published in: Small Business Economics, (2007)	9/04
5/2004	Kölling, A. Rässler, S.	Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models published in: Zeitschrift für ArbeitsmarktForschung 37 (2004), S. 306-318	10/04
6/2004	Stephan, G. Gerlach, K.	Collective contracts, wages and wage dispersion in a multilevel model published as: Wage settlements and wage setting : results from a multi-level model. In: Applied Economics, Vol. 37, No. 20 (2005), S. 2297-2306	10/04
7/2004	Gartner, H. Stephan, G.	How collective contracts and works councils reduce the gender wage gap	12/04
1/2005	Blien, U. Suedekum, J.	Local economic structure and industry development in Germany, 1993-2001	1/05
2/2005	Brixy, U. Kohaut, S. Schnabel, C.	How fast do newly founded firms mature? : empirical analyses on job quality in start-ups published in: Michael Fritsch, Jürgen Schmude (Ed.): Entrepreneurship in the region, New York et al., 2006, S. 95-112	1/05
3/2005	Lechner, M. Miquel, R. Wunsch, C.	Long-run effects of public sector sponsored training in West Germany	1/05
4/2005	Hinz, T. Gartner, H.	Lohnunterschiede zwischen Frauen und Männern in Branchen, Berufen und Betrieben published in: Zeitschrift für Soziologie 34 (2005), S. 22-39, as: Geschlechtsspezifische Lohnunterschiede in Branchen, Berufen und Betrieben	2/05
5/2005	Gartner, H. Rässler, S.	Analyzing the changing gender wage gap based on multiply imputed right censored wages	2/05
6/2005	Alda, H. Bender, S. Gartner, H.	The linked employer-employee dataset of the IAB (LIAB) published as: The linked employer-employee dataset created from the IAB establishment panel and the process-produced data of the IAB (LIAB). In: Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften 125 (2005), S. 327-336 (shortened)	3/05
7/2005	Haas, A. Rothe, T.	Labour market dynamics from a regional perspective : the multi-account system	4/05
8/2005	Caliendo, M. Hujer, R. Thomsen, S. L.	Identifying effect heterogeneity to improve the efficiency of job creation schemes in Germany	4/05

9/2005	Gerlach, K. Stephan, G.	Wage distributions by wage-setting regime published as: Bargaining regimes and wage dispersion. In: Jahrbücher für Nationalökonomie und Statistik, Bd. 226, H. 6 (2006)	4/05
10/2005	Gerlach, K. Stephan, G.	Individual tenure and collective contracts	4/05
11/2005	Blien, U. Hirschenauer, F.	Formula allocation : the regional allocation of budgetary funds for measures of active labour market policy in Germany published in: Economics Bulletin, Vol. 18, no. 7 (2006)	4/05
12/2005	Alda, H. Allaart, P. Bellmann, L.	Churning and institutions : Dutch and German establishments compared with micro-level data	5/05
13/2005	Caliendo, M. Hujer, R. Thomsen, S. L.	Individual employment effects of job creation schemes in Germany with respect to sectoral heterogeneity	5/05
14/2005	Lechner, M. Miquel, R. Wunsch, C.	The curse and blessing of training the unemployed in a changing economy : the case of East Germany after unification	6/05
15/2005	Jensen, U. Rässler, S.	Where have all the data gone? : stochastic production frontiers with multiply imputed German establishment data published in: Zeitschrift für ArbeitsmarktForschung, Jg. 39, H. 2, 2006, S. 277-295	7/05
16/2005	Schnabel, C. Zagelmeyer, S. Kohaut, S.	Collective bargaining structure and its determinants : an empirical analysis with British and German establishment data published in: European Journal of Industrial Relations, Vol. 12, No. 2. S. 165-188	8/05
17/2005	Koch, S. Stephan, G. Walwei, U.	Workfare: Möglichkeiten und Grenzen published in: Zeitschrift für ArbeitsmarktForschung 38 (2005), S. 419-440	8/05
18/2005	Alda, H. Bellmann, L. Gartner, H.	Wage structure and labour mobility in the West German private sector 1993-2000	8/05
19/2005	Eichhorst, W. Konle-Seidl, R.	The interaction of labor market regulation and labor market policies in welfare state reform	9/05
20/2005	Gerlach, K. Stephan, G.	Tarifverträge und betriebliche Entlohnungsstrukturen published in: C. Clemens, M. Heinemann & S. Soretz (Hg.): Auf allen Märkten zu Hause, Marburg 2006, S. 123-143	11/05
21/2005	Fitzenberger, B. Speckesser, S.	Employment effects of the provision of specific professional skills and techniques in Germany	11/05
22/2005	Ludsteck, J. Jacobebbinghaus, P.	Strike activity and centralisation in wage setting	12/05
1/2006	Gerlach, K. Levine, D. Stephan, G. Struck, O.	The acceptability of layoffs and pay cuts : comparing North America with Germany	1/06
2/2006	Ludsteck, J.	Employment effects of centralization in wage setting in a median voter model	2/06
3/2006	Gaggermeier, C.	Pension and children : Pareto improvement with heterogeneous preferences	2/06
4/2006	Binder, J. Schwengler, B.	Korrekturverfahren zur Berechnung der Einkommen über der Beitragsbemessungsgrenze	3/06
5/2006	Brixy, U. Grotz, R.	Regional patterns and determinants of new firm formation and survival in western Germany	4/06
6/2006	Blien, U. Sanner, H.	Structural change and regional employment dynamics	4/06

7/2006	Stephan, G. Rässler, S. Schewe, T.	Wirkungsanalyse in der Bundesagentur für Arbeit : Konzeption, Datenbasis und ausgewählte Befunde published as: Das TrEffeR-Projekt der Bundesagentur für Arbeit : die Wirkung von Maßnahmen aktiver Arbeitsmarktpolitik. In: Zeitschrift für ArbeitsmarktForschung, Jg. 39, H. 3/4 (2006)	4/06
8/2006	Gash, V. Mertens, A. Romeu Gordo, L.	Are fixed-term jobs bad for your health? : a comparison of West-Germany and Spain	5/06
9/2006	Romeu Gordo, L.	Compression of morbidity and the labor supply of older people	5/06
10/2006	Jahn, E. J. Wagner, T.	Base period, qualifying period and the equilibrium rate of unemployment	6/06
11/2006	Jensen, U. Gartner, H. Rässler, S.	Measuring overeducation with earnings frontiers and multiply imputed censored income data	6/06
12/2006	Meyer, B. Lutz, C. Schnur, P. Zika, G.	National economic policy simulations with global interdependencies : a sensitivity analysis for Germany	7/06
13/2006	Beblo, M. Bender, S. Wolf, E.	The wage effects of entering motherhood : a within-firm matching approach	8/06
14/2006	Niebuhr, A.	Migration and innovation : does cultural diversity matter for regional R&D activity?	8/06
15/2006	Kiesl, H. Rässler, S.	How valid can data fusion be? published in: Journal of Official Statistics, (2006)	8/06
16/2006	Hujer, R. Zeiss, C.	The effects of job creation schemes on the unemployment duration in East Germany	8/06
17/2006	Fitzenberger, B. Osikominu, A. Völter, R.	Get training or wait? : long-run employment effects of training programs for the unemployed in West Germany	9/06
18/2006	Antoni, M. Jahn, E. J.	Do changes in regulation affect employment duration in temporary work agencies?	9/06
19/2006	Fuchs, J. Söhnlein, D.	Effekte alternativer Annahmen auf die prognostizierte Erwerbsbevölkerung	10/06
20/2006	Lechner, M. Wunsch, C.	Active labour market policy in East Germany : waiting for the economy to take off	11/06
21/2006	Kruppe, T.	Die Förderung beruflicher Weiterbildung : eine mikroökonometrische Evaluation der Ergänzung durch das ESF-BA-Programm	11/06
22/2006	Feil, M. Klinger, S. Zika, G.	Sozialabgaben und Beschäftigung : Simulationen mit drei makroökonomischen Modellen	11/06
23/2006	Blien, U. Phan, t. H. V.	A pilot study on the Vietnamese labour market and its social and economic context	11/06
24/2006	Lutz, R.	Was spricht eigentlich gegen eine private Arbeitslosenversicherung?	11/06
25/2006	Jirjahn, U. Pfeifer, C. Tsertsvadze, G.	Mikroökonomische Beschäftigungseffekte des Hamburger Modells zur Beschäftigungsförderung	11/06
26/2006	Rudolph, H.	Indikator gesteuerte Verteilung von Eingliederungsmitteln im SGB II : Erfolgs- und Effizienzkriterien als Leistungsanreiz?	12/06
27/2006	Wolff, J.	How does experience and job mobility determine wage gain in a transition and a non-transition economy? : the case of	12/06

		east and west Germany	
28/2006	Blien, U. Kirchhof, K. Ludewig, O.	Agglomeration effects on labour demand	12/06
29/2006	Blien, U. Hirschenauer, F. Phan, t. H. V.	Model-based classification of regional labour markets : for purposes of labour market policy	12/06
30/2006	Krug, G.	Kombilohn und Reziprozität in Beschäftigungsverhältnissen : eine Analyse im Rahmen des Matching-Ansatzes	12/06
1/2007	Moritz, M. Gröger, M.	The German-Czech border region after the fall of the Iron Curtain: Effects on the labour market : an empirical study using the IAB Employment Sample (IABS)	1/07
2/2007	Hampel, K. Kunz, M. Schanne, N. Wapler, R. Weyh, A.	Regional employment forecasts with spatial interdependencies	1/07
3/2007	Eckey, H.- F. Schwengler, B. Türck, M.	Vergleich von deutschen Arbeitsmarktregionen	1/07
4/2007	Kristen, C. Granato, N.	The educational attainment of the second generation in Germany : social origins and ethnic inequality	1/07
5/2007	Jacob, M. Kleinert, C.	Does unemployment help or hinder becoming independent? : the role of employment status for leaving the parental home	1/07
6/2007	Konle-Seidl, R. Eichhorst, W. Grienberger-Zingerle, M.	Activation policies in Germany : from status protection to basic income support	1/07
7/2007	Lechner, M. Wunsch, C.	Are training programs more effective when unemployment is high?	2/07
8/2007	Hohendanner, C.	Verdrängen Ein-Euro-Jobs sozialversicherungspflichtige Beschäftigung in den Betrieben?	2/07
9/2007	Seibert, H.	Frühe Flexibilisierung? Regionale Mobilität nach der Lehrausbildung in Deutschland zwischen 1977 und 2004	2/07
10/2007	Bernhard, S. Kurz, K.	Familie und Arbeitsmarkt	2/07

Imprint

IABDiscussionPaper
No. 11 / 2007

Editorial address
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit
Weddigenstr. 20-22
D-90478 Nürnberg

Editorial staff
Regina Stoll

Technical completion
Regina Stoll

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

Download of this DiscussionPaper:
http://doku.iab.de/discussionpapers/2007/dp1107.pdf

Website
http://www.iab.de

For further inquiries contact the author:
Stefan Bender, Tel. 0911/179-3082,
or e-mail: stefan.bender@iab.de