Heinisch, Dominik; Koenig, Johannes; Otto, Anne

Working Paper

The IAB-INCHER project of earned doctorates (IIPED): A supervised machine learning approach to identify doctorate recipients in the German integrated employment biography data

IAB-Discussion Paper, No. 13/2019

Provided in Cooperation with:
Institute for Employment Research (IAB)

Suggested Citation: Heinisch, Dominik; Koenig, Johannes; Otto, Anne (2019) : The IAB-INCHER project of earned doctorates (IIPED): A supervised machine learning approach to identify doctorate recipients in the German integrated employment biography data, IAB-Discussion Paper, No. 13/2019, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg

INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

IAB-DISCUSSION PAPER

Articles on labour market issues

13|2019 The IAB-INCHER project of earned doctorates (IIPED): A supervised machine learning approach to identify doctorate recipients in the German integrated employment biography data.

Dominik Heinisch, Johannes Koenig, Anne Otto

ISSN 2195-2663

The IAB-INCHER project of earned doctorates (IIPED): A supervised machine learning approach to identify doctorate recipients in the German integrated employment biography data

Dominik Heinisch (University of Kassel and INCHER-Kassel (Germany)),
Johannes Koenig (University of Kassel and INCHER-Kassel (Germany)),
Anne Otto (Institute for Employment Research (IAB) Rhineland-Palatinate-Saarland (Germany))

Content

1 Record Linkage of Integrated Employment Biography Data 7

2 Data Sources 8

2.1 Doctorate Recipients Data of the German National Library (DNB) 9

2.2 Integrated Employment Biographies (IEB) 10

3 Identifying Doctorate Recipients in the German Labour Market Data 11

3.1 Problem Description 11

3.2 Pre-processing and Record Linkage 15

4 Application 21

5 Limitations 24

6 Conclusions 25

List of graphics

Figure 1: Overview of the data processing and record linkage procedure 15
Figure 2: Evaluation of different machine learning algorithm and model specifications 20
Figure 3: Employment status over time before/after graduation 23
Figure 4: Employment status over time before/after graduation separate for male and female doctorate recipients 24

List of tables

Table 1: Illustration of the DNB data 9
Table 2: Illustration of the IEB data 11
Table 3: Variables for machine learning 18
Table 4: Illustration of DNB-IAB record linkage 18
Table 5: Descriptive statistics for the classification variables in the synthetic training and evaluation data separated for true-negative and true-positive 19
Table 6: Classification results – best parameter settings (on training dataset) 21
Table 7: Evaluation of the classification results – best parameter settings 21
Table 8: Additional quality assessment 22

Appendix

Table A 1: Distributions for multiple matches of the synthetic training- and evaluation dataset and of the full (matched) dataset 28
Table A 2: Descriptive statistics for synthetic training- and evaluation dataset 28
Table A 3: Descriptive statistics for full (matched) dataset 29

Figure B 1: Successfully identified doctorate recipients by graduation year 30
Figure B 2: Successfully identified doctorate recipients by subject field 30

Abstract

Only scarce information is available on doctorate recipients' career outcomes in Germany (BuWiN 2013). With the current information base, graduate students cannot make an informed decision whether to start a doctorate (Benderly 2018, Blank 2017). Administrative labour market data could provide the necessary information, is however incomplete in this respect. In this paper, we describe the record linkage of two datasets to close this information gap: data on doctorate recipients collected in the catalogue of the German National Library (DNB), and the German labour market biographies (IEB) from the German Institute of Employment Research. We use a machine learning based methodology, which 1) improves the record linkage of datasets without unique identifiers, and 2) evaluates the quality of the record linkage. The machine learning algorithms are trained on a synthetic training and evaluation dataset. In an exemplary analysis we compare the employment status of female and male doctorate recipients in Germany.

Zusammenfassung

Es gibt bislang nur wenige wissenschaftliche Studien, welche das Karriereauskommen von Promovierten in Deutschland untersuchen (BuWiN 2013). Daher bildet die empirische Evidenz zum jetzigen Stand für Absolventen keine hinreichende Informationsgrundlage, um eine wohlüberlegte Entscheidung für oder gegen eine Promotion zu treffen (Benderly 2018; Blank 2017). Administrative Daten zu individuellen Karriereauskommen könnten diese Informationslücke schließen. Jedoch sind die derzeitig verfügbaren Datenquellen in dieser Hinsicht unvollständig. In diesem Beitrag verknüpfen wir Daten zu Promovierten die im Katalog der Deutschen Nationalbibliothek (DNB) gesammelt wurden, mit den Integrierten Erwerbsbiografien (IEB) des Instituts für Arbeitsmarkt- und Berufsforschung (IAB). Wir verwenden Methoden des maschinellen Lernens, die es erlauben 1) Datensätze ohne eindeutige Identifier zu verknüpfen und 2) die Qualität des verknüpften Datensatzes zu bewerten. Die Algorithmen werden auf einem synthetischen Trainings- und Testdatensatz trainiert. In einer beispielhaften Analyse werden die Karriereauskommen von weiblichen und männlichen Promovierten miteinander verglichen.

Keywords

PhD, doctorate recipients, labour market trajectories, administrative data, record linkage, machine learning, supervised learning

JEL-Klassifikation

C81, E24, I20

Acknowledgements

We thank Guido Bünstorf and the entire WISKIDZ-Team, Rasmus Bode, Tom Hanika, Andreas Rehs, and Igor Asanov for their valuable and constructive suggestions during the planning and development of this research work, as well as Judith Heinisch for her helpful comments. We gratefully acknowledge support from the German Federal Ministry of Education and Research (BMBF) under grant number 16FWN016. Moreover, we are very grateful for the great and excellent data support of the staff of the IT department of the Institute for Employment Research (IAB) for this project.

1 Record Linkage of Integrated Employment Biography Data

In recent years, the availability of comprehensive new administrative datasets on individual labour market biographies has enabled numerous studies in economics and other social sciences covering a wide range of labour market topics. However, administrative labour market records comprise a limited set of variables, thus narrowing the scope of potential research questions that can be addressed. Only scarce information is available about career outcomes of doctorate recipients in Germany (BuWiN 2013). This holds particularly for those doctorate recipients who pursue careers in the non-academic sector. Knowing more about their labour market biographies is not only important for universities and policy makers. Without knowledge about potential career outcomes, students cannot make an informed decision whether to start doctoral training (Benderly 2018, Blank 2017).

The objective of the IAB-INCHER project of earned doctorates (IIPED) is to construct a comprehensive dataset on labour market biographies of German doctorate recipients. The Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB) cover labour market records of about 80 percent of the German workforce. They comprise detailed individual-level information on socio-demographic characteristics, qualification levels, and job characteristics, however no information about earned doctoral degrees. This information is provided by the catalogue of the German National Library (DNB). The DNB covers almost all German universities' doctorate recipients from 1970 to 2015. The DNB only provides sufficient information for conventional record linkage (e.g. exact dates of birth) for a minority of individuals. To be able to link both datasets on a large scale, we apply a record linkage procedure that utilizes supervised machine learning algorithms, which are trained on a synthetic training and evaluation dataset.

Numerous prior studies have used record linkage methods (Schnell 2013) to supplement administrative labour market data. In many cases, the record linkage could be based on unique identifiers available in both datasets (e.g. name-surname combination, exact birth date, sex). If identifiers are incomplete or not fully reliable, more advanced "Merge Toolboxes" are available, which i.e. utilize string-comparison functions to calculate similarities between key words (e.g. name of the employer) in both datasets (Schnell et al. 2004). Even if conventional approaches are able to successfully link two datasets, a proper evaluation of the linked dataset's quality (in terms of recall and precision) would be advisable, rather than only reporting the number of final matched entities. Multiple matches between entries are another problem our approach is able to take into account.

To overcome the limitations of existing record linkage methods, we develop and assess a set of supervised machine learning algorithms. This approach has several advantages: First, it is not restricted to data with high quality identifiers. Second, the quality of the linked dataset is assessable and comparable across different algorithms, as well as to conventional record linkage approaches. Third, our approach is applicable under strict data security requirements and ensures the strict anonymity of individual records, which are indispensable requirements in any use of social security data in Germany. Fourth, we utilize a synthetic training and evaluation dataset, which allows

us to evaluate the quality of the record linkage in the absence of external training and evaluation data.

Even though unique identifiers are absent in both datasets, the final linked dataset meets high quality standards in terms of precision and recall. All tested supervised machine learning algorithms outperform heuristic (rule-based) approaches. Achieving a high recall rate not only allows researchers to address questions requiring larger and more complete samples, it also enables differentiations among subgroups. In addition, as the algorithm uses multiple features to predict true positive matches, it is less likely to introduce bias into the sample. While the synthetic test and evaluation dataset might by itself act as a source of bias, we do not find any distortions on observables. Depending on the parameter settings, the quality of the linked datasets can vary for each algorithm, which highlights the necessity of independent training and test data for selecting the best parameter specifications.

The obtained linked dataset allows us to investigate the labour market trajectories of German doctorate recipients from 1975 to 2015 before, during, and after their graduation. As a practical application we use the final dataset to analyze the employment status of doctorate recipients at different point of times in their career. In particular, we analyze gender specific differences in the share of full-time and part-time employment during doctorate recipients' careers. We find that few doctorate recipients are unemployed after graduation. However, a substantial share of female doctorate recipients works part-time. While female and male doctorate recipients show similar employment patterns during their graduation period, the share of part-time and full-time employed females diverges after that.

The paper is structured as follows: In Section 2, the datasets of the record linkage approach are described. Section 3 presents the supervised machine learning algorithms in detail, as well as the underlying assumptions and data requirements. The classification problem is discussed in the context of the administrative data. Section 4 describes our implementation and evaluates the different approaches we tested. In Section 5, the linked dataset is used to investigate the employment status of doctorate recipients over time. In Section 6, we discuss some limitations of the proposed approach and draw implications for further research.

2 Data Sources

In this section, we introduce the two datasets which are integrated by the record linkage: The Integrated Employment Biographies (German: Integrierte Erwerbsbiographien, IEB) and the dataset of doctorate recipients from the German National Library (Deutsche Nationalbibliothek, DNB). Both datasets provide a nearly complete picture of the corresponding populations: The German workforce (subject to social security payments) is represented in the IEB; and doctorate recipients who graduated from German universities are represented in the DNB. As a result, the DNB data provide a suitable supplement for the IEB, where information about tertiary education is incomplete. Both datasets are collected by public institutions following standardized procedures and regularities in the data preparation process, which makes them highly reliable and suitable for research purposes. While the DNB data have been merged via record linkage only with publication

data (Heinisch and Buenstorf 2018), the IEB have been merged via record linkage with a number of external micro databases in the past (see e.g. Antoni and Seth 2012, Dorner et al. 2014, Wydra-Somaggio 2015, Teichert et al. 2018)

2.1 Doctorate Recipients Data of the German National Library (DNB)

The DNB catalogue covers the (almost) entire population of individuals who completed doctoral training at German universities – doctorate recipients, which encompasses about 1 million authors of dissertations.[1] Two peculiarities cause the DNB catalogue to cover the almost entire population of doctorate recipients from German universities. First, all German publications (published in Germany or by Germans) are hold by the German National Library, which is "entrusted with the task of collecting, permanently archiving, bibliographically classifying and making available to the general public all German and German-language publications from 1913" (DNB 2018). According to §§ 14 to 16 of the Act on the German National Library, media works are to be delivered to the library if a holder of the original distribution right has their registered office, a permanent establishment, or the main place of residence in Germany. Second, in Germany doctoral students are obliged to publish their theses in order to be awarded a doctorate from a German university, and thesis publications are tracked by the German National Library.[2]

Publications originating from universities are collected in the publication series H (Hochschulschriften). Within this series a separate note provides additional information on the type of publication, the year of submission, and the corresponding university name. Since data is selected by librarians for the purpose of archiving and classifying these publications, bibliographic information is documented with a high degree of accuracy. The coverage is (almost) complete for all years and disciplines. From 1995 to 1997 onwards, the DNB created the Personennormdatei, a dataset comprising all authors as separated entities. This additional catalogue improves the information available on authors. Beginning in 1997, the year of birth is recorded for the majority of authors in the dataset, as well as additional information on authors' nationality. However, most of these variables cannot be used as identifiers (variables) for the linkage procedure, because the coverage rates vary strongly over time. A stylized example of the DNB data is provided in Table 1.

Table 1: Illustration of the DNB data

dnb_id	name	surname	birth_year	female	nationality	uni_name	publication_year	subject
87640472	Marta	Musterfrau	NA	female	German	Kiel	2010	Economics
12342124	Max	Maulwurf	1979	male	German	Jena	2008	Medicine
07986678	Martin	Mustermann	NA	male	Italian	Kassel	1993	Engineering

Source: own example; note: the table provides fictitious examples of the DNB dataset.

[1] The German National Library makes its data accessible under the Creative Commons Zero license (CCO 1.0).

[2] The DNB dataset has been used for various analyses, e.g. Buenstorf and Geissler (2014) studied advisor effects based on laser-related dissertations, and Heinisch und Bünstorf (2018) identified the doctoral advisors of doctorate recipients. Both studies confirm the high reliability and completeness of the DNB data.

2.2 Integrated Employment Biographies (IEB)

The IEB unites data from five different historic data sources, each capturing a different segment of the German social security system.[3] It contains detailed information on all individuals who are liable to social insurance contributions in Germany, i.e. employees, unemployed individuals, job seekers, recipients of social benefits and participants in active labour market programs. Civil servants and self-employed, family workers, and doctorate candidates financed solely by scholarships etc. are not part of the social security system and therefore not reported in the IEB. Taken together, the data cover approximately 80 percent of the German workforce.

The IEB data comprises starting and ending dates of all spells (i.e. episodes of unemployment, benefit receipt, employment) for each individual (see vom Berge et al. 2013). Additionally, for each individual a range of sociodemographic characteristics is documented (e.g. sex, date of birth, nationality, qualification level), job features (type of employment, occupation, industry affiliation, region of workplace). While, although incomplete, information of obtained vocational training certificates, or bachelor and master degrees is part of the IEB, no information on doctoral degrees exists. Information is available on a daily basis from 1975 to the most current year for West Germany, and from 1993 for East Germany. Hence, the IEB enables labour market biographies of individuals in the public and the private sector to be tracked over time.

The IEB data is highly reliable for all variables that are directly relevant for social insurance contributions. However, some information in the data, i.e. information on secondary schooling, is less reliable as it is transmitted by the employer solely for statistical purposes (Fitzenberger et al. 2005). Furthermore, some variables contain missing values, which vary over time (see e.g. Antoni et al. 2016). Confidential information, which would make individuals identifiable (e.g. name and address), is not accessible for researchers (Schnell 2013). An anonymized system-independent individual identifier links social security registers and administrative data of the Federal Employment Agency (Dorner et al. 2014).[4] Table 2 shows a fictious example of the pre-processed IEB data.

[3] These five data sources are: the Employee History, Benefit Recipient History, Unemployment Benefit II Recipient History, Participants-in-Measures History, and the Jobseeker History.

[4] The IEB and its scientific use file have been extensively discussed in the past. See for example: Dorner et al. (2010) for a brief discussion of the IEB, Oberschachtsiek et al. (2008) for a more detailed description of the IEB sample, and Zimmermann et al. (2007) for the scientific use file.

Table 2: Illustration of the IEB data

iab_id	employ-ment	begin_date	end_date	place_ occ	school_de-gree	apprentice-ship	class_econ_activity
92240472	Mini-Job	01/01/1996	31/12/1996	Kiel	A level	No qualifica-tion	49.32 Taxi operation
92240472	Part-time	01/01/1997	31/12/1997	Kiel	A level	university de-gree	85.42 Tertiary educa-tion
92240472	Part-time	01/01/1998	31/12/1998	Kiel	A level	university de-gree	85.42 Tertiary educa-tion
92240472	Unem-ployed	01/01/1999	31/01/1999	Kiel	A level	university de-gree	
92240472	Full-time	01/02/1999	31/12/1999	Berlin	A level	university de-gree	72.11 Research and experimental development on bio-technology
92240472	Full-time	01/01/2000	31/12/2000	Berlin	A level	university de-gree	72.11 Research and experimental development on bio-technology
32134444	Mini-Job	01/06/2003	31/08/2003	Buxde-hude	No qualifi-cation	No qualifica-tion	55.20 Holiday and other short-stay accommodation
32134444	Mini-Job	01/07/2004	31/09/2004	Jena	Primary School	No qualifica-tion	55.10 Hotels and simi-lar accommodation
32134444	Part-time	01/01/2007	31/12/2007	Jena	A level	university de-gree	86.10 Hospital activi-ties
32134444	Full-time	01/01/2008	31/12/2008	Halle	A level	university de-gree	86.10 Hospital activi-ties
20347523	Part-time	01/08/1980	31/12/1980	Frank-furt	Primary School	vocational training	4.11 Central banking
20347523	Full-time	01/01/1981	31/12/1981	Frank-furt	Primary School	vocational training	66.11 Administration of financial markets

Source: own example; note: the table provides fictitious examples of the IEB dataset.

3 Identifying Doctorate Recipients in the German Labour Market Data

3.1 Problem Description

In this section, we describe the general record linkage problem first, and then expand on it in terms of its applicability to social security data, where researchers have to deal with large volumes of highly sensitive data. The record linkage procedure aims at identifying as many entries in both datasets, which belong to the same entity. This target function is optimized under the constraint of keeping the number of incorrect matched entries as low as possible. To achieve this target, a two-step procedure is applied: First, entries of both datasets are matched by using an imperfect identifier (i.e. the names of individuals). Second, false matched combinations are eleminated. Figure 1 presents an overview of the record linkage approach described in this section.

The first step aims to match as many entries as possible of both datasets, which might belong to one entity. In other words: in the first step the datasets are actually linked. This can be achieved i.e. by exact string matching between entries' names, or by calculating distances between the entries' names using a fuzzy string matching algorithm. The second step aims to identify as reliably as possible true linked entities among the matched entry pairs. In other words: in the second step correctly linked entries which belong to one entity are filtered from incorrectly matched entries. As social security data comprises large volumes of data with many homonyms (in our case the entire German workforce) the filtering of true positive matched entries is a more serious problem, in particular, as incorrectly spelled names are less frequent in administrative data. Therefore, the paper is primarily focused on improving the second step of the record linkage procedure.

The linked entries of both datasets by a specific identifier will result into 0-to-n possible combinations of matched entries, of which 0-to-1 combinations truly belong to one entity. In those cases, where multiple entries match into one entity, many-to-many (n-to-m) matched entries occur. Identifying the true matched entities, in a set of n-to-m matched entries, can be described as a classification problem. The following description of the classification problem is based on Gareth et al. (2013) and Bishop (2006). Formally, the classification task is to find a function $f(X)$ that correctly classifies two matched entries of both datasets as one entity. With a quantitative response variable $Y \in c(Same, Different)$ and using a set of p different predictors:

$$X = (X_1, X_2, \dots, X_p)$$

$$Y = f(X) + \epsilon$$

where ϵ is the error term.

In practice, there are numerous restrictions that complicate the estimation of the classification function f: Unique entity identifiers (or keys) and reliable predictors such as combinations of name, birthday, and birthplace may be lacking. Even if the available data are generally of high quality, information may be imprecise, misreported, or incomplete for individual entries. And even in cases where reliable predictors exist, privacy requirements may restrict the number of predictor variables X that are accessible to researchers.

If the reliability of a single or multiple predictors cannot be ascertained, or if only a set of weak predictors is available, machine learning algorithms can improve the record linkage quality. Machine learning algorithms have been applied to a number of record linkage problems and several solutions are available (see e.g. Christen 2012b). In this paper, we use machine learning algorithms to solve the classification problem described above in accurately filtering true matched entries. In this case the classification problem can be described as the best combination of available input variables X that predict $\hat{Y}$:

$$\hat{Y} = \hat{f}(X),$$

with $\hat{Y}$ as classification output and $\hat{f}$ as our estimation equation for the classification function f. The accuracy of $\hat{Y}$ depends on two aspects as the following equation shows: the reducible and irreducible error:

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = [f(X) - \hat{f}(X)]^2 + Var(\epsilon)$$

The reducible error $[f(X) - \hat{f}(X)]^2$ results from $\hat{f}$ not being a perfect estimation for f. As the name implies, the reducible error can be reduced by more sophisticated statistical learning methods or

by increasing the input variables' X predictive power. In contrast, the irreducible error $Var(\epsilon)$ would persist even if $\hat{f}$ were a perfect approximation of f. The set of input variables X entering into function f cannot predict ϵ by definition as they result from errors in measuring X. A suitable classification procedure identifies the best functional relation of X in $\hat{f}$ that approximates f, by minimizing the reducible error $\left[f(X) - \hat{f}(X)\right]^2$.

Solving classification problems is a traditional field of application for machine learning techniques. Machine learning algorithms can help to find suitable approximations of the classification function f (Christen 2012a). However, these approaches have not found much use in research using administrative labour marked data. Record linkage procedures used in this context have mostly been based on heuristic approaches. Data are linked by calculating similarities between names (see Schnell 2013) and "rules based" heuristics, e.g. information on whether two entries originate from the same or different regions. Applying heuristic approaches requires high-quality data. Even then, heuristic approaches do not exploit the full potential of the data because they do not use the optimal functional form of $\hat{f}$ or the best representation of X.

Evaluation Measures

Three different metrics are commonly used to evaluate the classification performance of machine learning algorithms: accuracy, precision, and recall. The measures are calculated by using a confusion matrix, which categorizes the predicted and real classes into four groups: True-positive matched pairs (TP) give the number of real matched entries, which belong to one entity, that were accurately predicted as belonging to one entety by the used algorithm. False-positive matched pairs (FP) give the number of wrong matched entries, which do not represent the same entity but are falsely predicted as such. True-negative matched pairs (TN) and false-negative matched pairs (FN) give the number of pairs correctly and incorrectly classified as representing two different entities. A confusion matrix can be calculated for any kind of classifier. It also provides the basis for calculating accuracy, precision, recall and F1 measures allowing the overall quality of the matching to be characterized. Accuracy is defined as the ratio of correct predictions to all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy alone is insufficient to assess matching quality, especially when large data sets are linked and large numbers of true-negative matched entries are expected (this is frequently the case in labour market research, as many observations share the same name-surname combination). Algorithms that always predict $\hat{Y} = c(Different)$ and as a result do no link any entries of both datasets would have an accuracy of nearly 100 percent. Precision and recall are the two most frequently used metrics that overcome this problem. Precision is a measure of exactness in predicting two matched entries to be the same. In other words, precision measures how many of the linked entities are correctly identified.

$$Precision = \frac{TP}{TP + FP}$$

Recall provides a measure of the ability to identify true linked entities in both data sets. In other words, recall measures how many of the true matches between entities could be identified.

$$Recall = \frac{TP}{TP + FN}$$

Finally, the F1 statistic is calculated as the harmonic average of recall and precision.

Supervised Learning

A wide selection of sophisticated classification algorithms is available to estimate $\hat{f}(X)$. These can broadly be categorized into deterministic, probabilistic and (machine) learning based approaches (Christen 2012b). Higher predictive power can be expected for supervised machine learning techniques. Supervised (machine) learning algorithms require training data to approximate the best representation of f by a specific representation of the input variables X. A wide variety of machine learning algorithms have been developed, and the choice of specific algorithms involves a trade-off between classification quality and computational demands. In addition, not all algorithms are implemented in statistical software packages available in the settings where administrative data may be accessed.[5] Reflecting these considerations, our approach utilizes three well-known machine learning algorithms: regularized logistic regressions, AdaBoost, and Random Forests.[6]

A regularized logistic regression estimates a logistic regression model with an additional penalty term to avoid overfitting. It requires ex ante specification of both the penalty parameter and a threshold probability value above which estimated matches are classified as belonging to the same entity. The Random Forest algorithm uses decision trees for classification. By randomly selecting a set of m variables a specific number of n decision trees is constructed. Each decision tree uses these m variables to split the dataset specific thresholds to classify the data into matches and non-matches. A sequence of multiple splits divides the data into distinct decision regions. A majority vote over the n decision trees decides on the class of each entry in the matched dataset. The number of randomly drawn variables (m) and the number of trees (n) have to be specified ex ante. AdaBoost is a boosting method developed for binary outcome variables. Similar to Random Forests, it is based on decision trees, but the classifiers are trained sequentially. After each iteration, the classification output is weighted by its classification success, giving a higher weight to misclassified matched entries in the next iteration. After converging, all decision trees give a majority vote on the matched entries class. The number of iterations and weights have to be set as parameters ex ante.

In our approach these machine learning algorithms are tested against a heuristic (rule-based) classification. The latter classifies pairs of entries by comparing one or several variables, e.g. whether an individual was employed in the university region.[7] For the heuristic classification approach the number of variables considered in classification needs to be specified ex ante.

The common objective of all these approaches is to develop a function $\hat{f}$ that accurately separates the spaces of same versus different entities in both datasets. Applying different model specifications enables us to select from a range of models with different properties. The aim of this task is to find an optimum between precision and recall, i.e. to link as many entries of both datasets as possible (high recall) while minimizing the number of false classification decisions (high precision).

[5] The administrative data used can only be used on secured machines available at IAB. More advanced methods such as multilayer neuronal networks are computationally intense and their application is not technically feasible in our case.

[6] All algorithms used are available as R Packages. We used the programming language R Version: 3.3.2 (R Core Team 2017) and the following R packages: For AdaBoost the package ada (Culp et al. 2006), for regularized logistic regressions the package glmnet (Friedman et al. 2010), for Random Forest the package randomForest (Liaw and Wiener 2002).

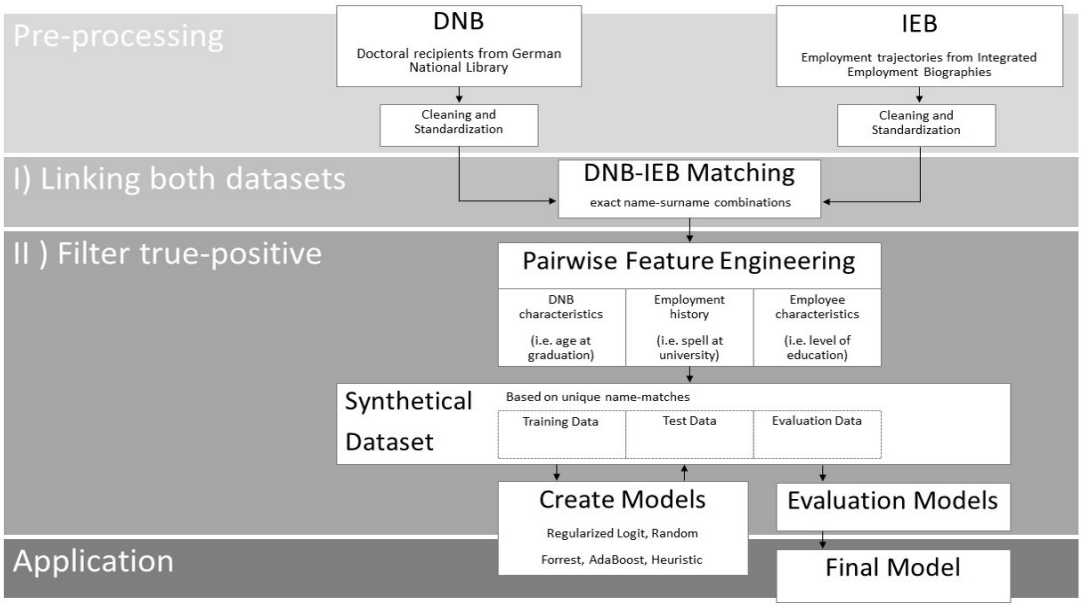
[7] Variables for training the algorithm are presented in Table 1. While for the heuristic classification approach we generated all possible combinations of the following variables. As a result, we get a number of possible decisions where only one of these variables, up to all of these variables need to take the value 1.

Data Pre-processing and Training Sample:

Overfitting is a serious risk when the best algorithm is selected. Overfitting means that the prediction function $\hat{f}$ follows the error term $Var(\epsilon)$, generating estimates for f that are as close as possible to the observed training data, but do not allow accurate estimates for new observations outside the training data. In this case, the trained algorithm is useless as the trained model is an exact representation of the training data but cannot be generalized to other data. This would fail the task of finding a function $\hat{f}$ that predicts our outcome variable Y as well as possible: $Y \approx \hat{f}(X)$ for any observation.

To overcome overfitting, out-of-bag predictions are used to evaluate the algorithms' classification success. Out-of-bag predictions require an independent dataset that has not been used in training the algorithms. The training data are split into several datasets that are specifically used for, first, training, second, identification of the right parameters, and, third, evaluation. For training and evaluation data are required for which true outcomes of the quantitative response variable $Y \in c(Same, Different)$ are known to the researcher.

Figure 1: Overview of the data processing and record linkage procedure



Source: own illustration.

3.2 Pre-processing and Record Linkage

In this section, we discuss the application of the record linkage procedure described in subsection 3.1 to identify dissertation authors from the DNB dataset in the IEB dataset.

Data Pre-Processing:

Even though both datasets are of a high quality, several pre-processing steps were required before the actual record linkage. First, the data on dissertation authors were downloaded from publication series H (Hochschulschriften) in the DNB online catalog. Then we excluded all authors with

incomplete name information (e.g. entries with missing first name or surname), as well as duplicated entries related to the same dissertation. The cleaned dissertation dataset includes 984,359 doctorate recipients.

In a second step, the DNB dataset is merged with the IEB data based on exact name-surname combinations.[8] Unlike other datasets (i.e. patent data), both datasets are of high quality regarding the spelling of names, including the spelling of German umlauts. We therefore used a naïve string matching algorithm to minimize the number of false-positive matched pairs. For 787,065 doctorate recipients at least one individual with the same name-surname combination was identified in the IEB. That some names of doctorate recipients do not match with any entry in the IEB may be explained by the fact that they include only individuals covered by the German social security system, but not others such as civil servants or students receiving scholarships (see above). Moreover, some doctorate recipients with very common name-surname-combinations (e.g. "Werner Müller") were matched to more than 300 IEB entities and had to be excluded for data privacy purposes. The final dataset is further limited to doctorate recipients who graduated between 1975 and 2015. East German doctorate recipients graduating before 1990 also had to be excluded because reliable IEB employment spells are only available for East Germany beginning in 1993. To save computational power and reduce the number of false-positive matched pairs, we deleted all individual matched pairs aged below 20 in the year of submission.

Generation of Synthetic Test and Training Data

Supervised (machine) learning algorithms require training data to approximate the best predictive model. As a result, for training and evaluation of the algorithm a set of reliable observations is necessary where matched entries belonging to one entity (true-positive matches) can be distinguished from false-positive matched entries (true-negative matches). Several strategies can be applied to identify a "gold standard" sample that can be used to train and evaluate the algorithm (Christen 2012a). An ideal solution would require surveying a selection of doctorate recipients asking about their realized career paths, or asking them to identify which career trajectory belongs to them among all the matched entries. The responses would provide the "gold standard" dataset, which can be generalized to predict other matched entries. However, data security and practical reasons make this infeasible. First, social security data is subject to stringent data privacy requirements. The data are strictly anonymized, and contacting individuals based on their private addresses is restricted as well. Second, even if individuals could be directly asked, mistakes as well as low response rates might reduce the representativeness of the sample obtained.

Therefore, we create a synthetic training and evaluation dataset from the available data. One important aspect in creating a synthetic training and evaluation dataset is its representativeness of the overall (matched) population. It should contain the same variables, which should moreover follow a similar frequency distribution and similar error characteristics. In our approach, we use name-surname combinations, as we believe the frequencies of name-surname combinations are independent of the variables used as classifiers.

[8] For data security reasons, this step is conducted by the Data- and IT-Management (DIM) Department of the IAB. In all further processing steps, the data is processed in a completely anonymized form. The execution of all processing steps is required to take place on the secured server infrastructure provided by the IAB.

Our true-positive matches ($Y \in c(Same)$) are based on unique name-surname combination, i.e. doctorate recipients whose name-surname combination appears only once in both the IEB and the DNB datasets. Since both datasets cover the underlying populations almost completely, these matched entries are expected to belong to the same entity.[9] For our true-negative matches ($Y \in c(Different)$), we merged the same DNB entry with a random set of entries from the IEB dataset. As the name of an individual is highly gender-dependent, we limit the randomly matched sample to entries with the same name but different surname. This procedure leads to a large number of wrongly matched entries. To specify a representative number of true-negative matched entries, we follow the overall distribution of matched entries and randomly draw a similar number of matched entries for each wrongly matched DNB entry. Using this strategy, we obtain a synthetic training and evaluation dataset, for which the true matching status is known and which is representative of the overall matched population.

Classification Variables

Three types of variables are created that are used as classifications. The first set of variables contains information on entries in the IEB dataset (e.g. an employment spell at a university); the second one contains information on entries in the DNB dataset (e.g. the year of submission), and the third one contains information calculated from both datasets (e.g. the lag between dissertation submission and the first employment spell). Table 3 gives an overview of the classification variables X, which are used to predict $\hat{Y}$. In Table 5 a stylized sample illustrates the final dataset. Table A 1, Table A 2, Table A 3(in the Appendix) provide descriptive statistics for an assessment of the representativeness of the synthetic training dataset and the full (matched) population.

[9] We performed a number of plausibility checks, which provided support to our conjecture. For example on an aggregated level, we investigated the career paths of this unique name-surname combinations for different subjects, gender and years and compared their career paths to known career paths of doctorate recipients from previous studies (e.g. the BuWiN 2017). The identified career trajectories indicate plausibility of these matches on an aggregated level.

Table 3: Variables for machine learning

Name	Description	Source
spell_research	Dummy, value one if individual has/had a spell at a university or research institute. European statistical classification for economic activities was used. Values were extended by record linkage for research institutions and universities.	IEB
spell_hospital	Dummy, value one if individual has a spell in a hospital/ medical practice. European statistical classification for economic activities was used.	IEB
prop_educ	Dummy, value one if education of individual belongs to university entrance qualification.	IEB
age_sub	Continuous, age in submission year.	IEB/DNB
right_age	Dummy, value one if individual is between 26 and 40 years old in submission year. Used for heuristic approach instead of age_sub.	IEB/DNB
same_ror_y5	Dummy, value one if individual was 5 years before/after graduation employed in university region.	IEB/DNB
first_spell_before	Continuous, first year in IEB subtracted from year of submission.	IEB/DNB
right_first_spell_before	Dummy, value 1 if first_spell_before is between -10 and 5. Used for heuristic approach instead of first_spell_before.	IEB/DNB
year_diss	Continuous, year of submission.	DNB
eastern	Dummy, value one if individual graduated in new federal states.	DNB
social science	Dummy, value one if individual graduated in social science.	DNB
natural science	Dummy, value one if individual graduated in natural science.	DNB
engineering	Dummy, value one if individual graduated in engineering.	DNB
medicine	Dummy, value one if individual graduated in medicine.	DNB
law/economics	Dummy, value one if individual graduated in economics/business studies/law.	DNB
nbr	Continuous, number of common namesakes in IEB Data.	IEB

Source: own classification.

Table 4: Illustration of DNB-IAB record linkage

dnb_id	iab_id	spell_research	spell_hospital	prop_educ	age_sub	same_ror_y5	first_spell_before	year_diss	eastern	social s.	natural s.	engineering	law/economics	medicine	nbr
12342124	92240472	1	0	1	40	0	-11	2007	1	0	0	0	0	1	3
12342124	32134444	0	1	1	29	1	-5	2007	1	0	0	0	0	1	3
12342124	20347523	0	0	0	45	0	-27	2007	1	0	0	0	0	1	3
87640472	08898092	0	0	0	66	0	5	2010	0	0	0	0	0	0	2
87640472	90980983	1	0	1	31	1	-10	2010	0	0	0	0	1	0	2

Source: own illustration; note: the table shows the stylized IAB-DNB linkage in fictitious examples

Table 6 reports the general descriptive statistics for the classification variables separately for the true-positive and true-negative matched entries in the synthetic training and evaluation dataset. For example, about 63.68 percent of the individuals in the true-positive sample had one employment spell at a university or other research institution (spell_research), as compared to 6.57 percent of the individuals in the true-negative sample, indicating high predictive power of the spell_research variable. This synthetic training and evaluation dataset contains some 50,000 matched doctorate recipients with up to 300 potential matched IEB entries. We divided this dataset into two equal parts: a training dataset and an evaluation dataset. A block randomization was applied to divide the dataset into the two subsets. A block randomization is a technique, which reduces bias and balances the allocation of individuals into different subsets. This increases the probability that each subset contains an equal number of multiple matched entries.

Table 5: Descriptive statistics for the classification variables in the synthetic training and evaluation data separated for true-negative and true-positive

Variable	Same	Median	Mean	Min	Max
spell_research	1	1	0.6368	0	1
spell_research	0	0	0.0657	0	1
spell_hospital	1	0	0.3745	0	1
spell_hospital	0	0	0.1008	0	1
prop_educ	1	1	0.9507	0	1
prop_educ	0	0	0.3238	0	1
age_sub	1	31	32.5199	20	91
age_sub	0	36	37.8844	20	102
right_age	1	1	0.8996	0	1
right_age	0	0	0.4546	0	1
same_ror_y5	1	1	0.7297	0	1
same_ror_y5	0	0	0.0156	0	1
first_spell_before	1	-6	-6.9672	-40	37
first_spell_before	0	-11	-11.4541	-45	39
right_first_spell_before	1	1	0.7112	0	1
right_first_spell_before	0	0	0.4242	0	1

Note: Descriptive statistics on the distribution of features used to classify true-positive matched entries in the IEB and DNB data in the synthetic training and evaluation dataset. The data are split into two samples: True-positive matches based on unique name-surname combinations and true-negative matches based on entries with the same name, but different surname. The true-positive matches are indicated by “Same” = 1.

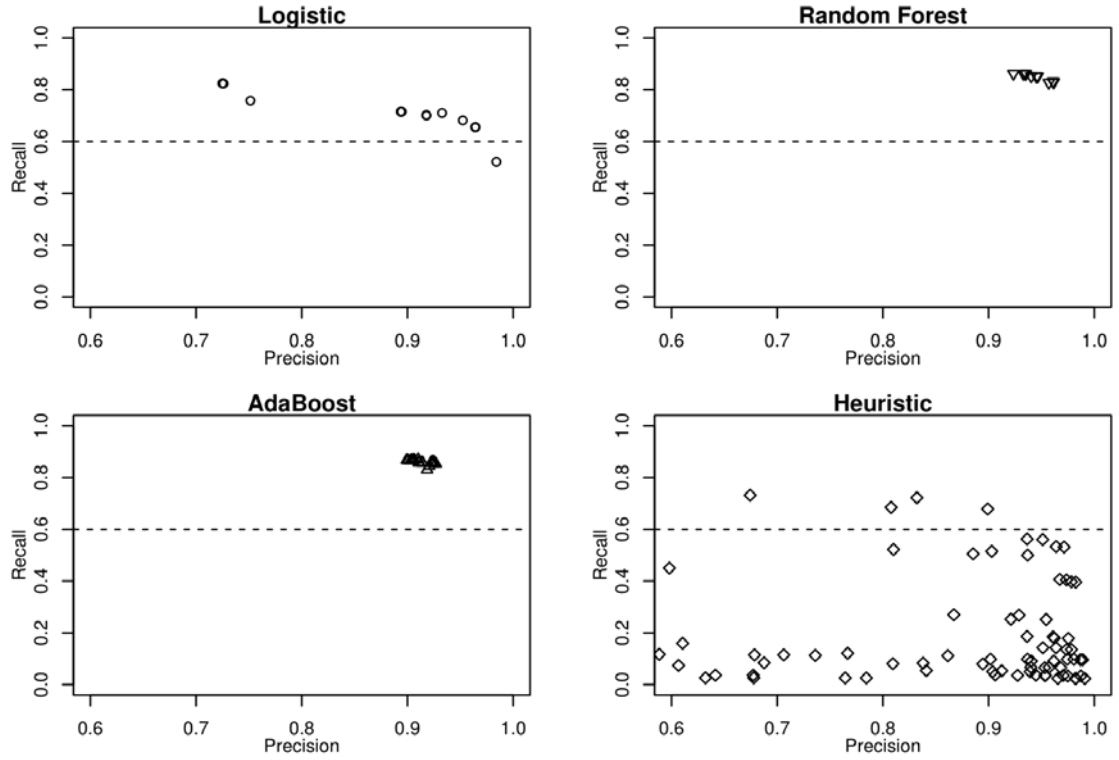
Source: own calculations.

Model Selection and Evaluation

For model selection, each classification algorithm was trained and tested for various parameter specifications. Algorithms were trained on three quarters of the training dataset end evaluated (by

recall and precision) on the remaining quarter. Results are shown in Figure 2. Figure 2 shows the recall-precision curve separately for alternative classification algorithms and model specifications.Table 6 shows the best training results for our evaluation measures.

Figure 2: Evaluation of different machine learning algorithm and model specifications



Source: own recall-precision plots for estimated algorithms under different tuning parameters.

All algorithms achieve satisfactory classification results and would generally be applicable. The heuristic approach also achieves sufficiently high values in terms of precision. In some specifications it outperforms most of the more advanced and computationally demanding algorithms.[10] However, the more computational demanding algorithms outperform the heuristic approach in that they reach comparable rates of precision but achieve substantially higher recall. Depending on parameter settings, the classification success of the specific algorithms varies substantially (e.g. results for the logit model vary from a recall/precision of 0.5683/0.8805 to 0.9840/0.5219). This illustrates the advantage of using a supervised learning approach as it allows the evaluation of the record linkage quality not only by how many individuals are linked, but also by the achieved quality of linked entities.

[10] For example, one heuristic classified matched entries as belonging to the same entity if a matched IEB entry had a spell in a hospital/doctor's office, or a spell at a university/research institute, one spell in the university region at least -5/5 years after submission, aged between 25 and 40 at submission, and a labour marked entry at least 10 years before or at least 5 after submission. This heuristic reached a precision of 0.9889. However, while being very precise the heuristic is only able to link a very selective sample of doctorate recipients with the IEB dataset with a recall of 0.0962.

Table 6: Classification results – best parameter settings (on training dataset)

Model	+1 (best parameter)				+1 (min recall 0.6)			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Logistic	0.9328	0.7099	0.8062	0.9860	0.9644	0.6558	0.7807	0.9848
Random Forest	0.9457	0.8520	0.8964	0.9919	0.9616	0.8287	0.8902	0.9916
AdaBoost	0.9246	0.8602	0.8912	0.9914	0.9268	0.8534	0.8886	0.9912
Heuristic	0.8991	0.6786	0.7734	0.9826	0.8991	0.6786	0.7734	0.9826

Source: own calculations.

We next selected those specifications of the algorithms that achieved the highest average values in recall and precision and those with the highest precision and a recall of at least 0.6. For the evaluation we took the best parametrized models and trained them again on the full training dataset. Then we evaluated the trained models on the evaluation dataset. Table 7 shows the further evaluation results. All models show qualitatively similar results. The Random Forest algorithm outperforms the other algorithm. The best performing algorithm was then used to classify true-positive matched entries in the full (matched) dataset.

Based on the approach outlined above, the Random Forest algorithm identifies 552,459 individuals as $\hat{Y} = c(Same)$. If the Random Forest algorithm identifies more than one entry in the IEB that matches one entry in the DNB (or vice versa), then we decided to exclude respective cases from the final dataset. Hence, the final dataset for the IAB-INCHER project of earned doctorates (IIPED) consists of a total of 447,606 doctorate recipients, and the overall matching quote amounts to 45.47 percent.

Table 7: Evaluation of the classification results – best parameter settings

Model	+1 (best parameter)			
	Precision	Recall	F1	Accuracy
Logistic	0.9410	0.7018	0.8040	0.9847
Random Forest	0.9584	0.8337	0.8917	0.9910
AdaBoost	0.9196	0.8605	0.8891	0.9904
Heuristic	0.9110	0.6742	0.7749	0.9825

Source: own calculations.

4 Application

In this section, we evaluate data from the IAB-INCHER project of earned doctorates (IIPED) in two ways. First, we assess how representative the linked dataset is of the total population of doctorate recipients in Germany. Second, we present an exemplary analysis of the employment status of female and male doctorate recipients over time. This example is used to check whether the empirical results obtained with the linked dataset are consistent with existing empirical evidence. In doing

so, we explore whether the data can be used to analyze research questions related to the labour market biographies of doctorate recipients in Germany.

The Labour Market Sample of Doctorate Recipients

Figure B 1 depicts the share of linked doctorate recipients in the total population of doctorate recipients over time. This share increases strongly from 34.51 percent in the starting year 1975 to 61.70 percent in 2015. For doctorate recipients in the period before/after the German reunification the matching quote lies at 39.61 percent and 57.43 percent respectively. At 33.08 percent, the share of female doctorate recipients in the merged database is comparable to the 33.51 percent share in the population of doctorate recipients received from the DNB. Reliable information on domestic and foreign doctorate recipients is available for selected years in the DNB catalogue. In 2013, the share of domestic doctorate recipients in the DNB was 85.37 percent, while the respective share in the merged database is 87.62 percent, indicating that domestic-born doctorate recipients are slightly overrepresented. Figure B 2 illustrates average shares of merged doctorate recipients by discipline over the entire observation period. Overall average matching rates vary across fields, with values ranging from 42.81 percent for sports to 60.88 percent for sciences and mathematics.

As additional evidence of matching quality, we compared variables in both datasets (IEB and DNB) that were not employed in the matching procedure. Table 8 depicts the consistency of linked entries for year of birth and gender, which were both not used as classification variables because of limited coverage in the DNB dataset. Both variables indicate a high accuracy of our record linkage procedure on an aggregate level. Nevertheless, in some cases the identified linked entries were not correctly matched.

Table 8: Additional quality assessment

	Same value in IEB and DNB data	Different value in IEB and DNB data
year of birth	95.33%	4.67%
gender	99.08%	0.92%

Source: own calculations.

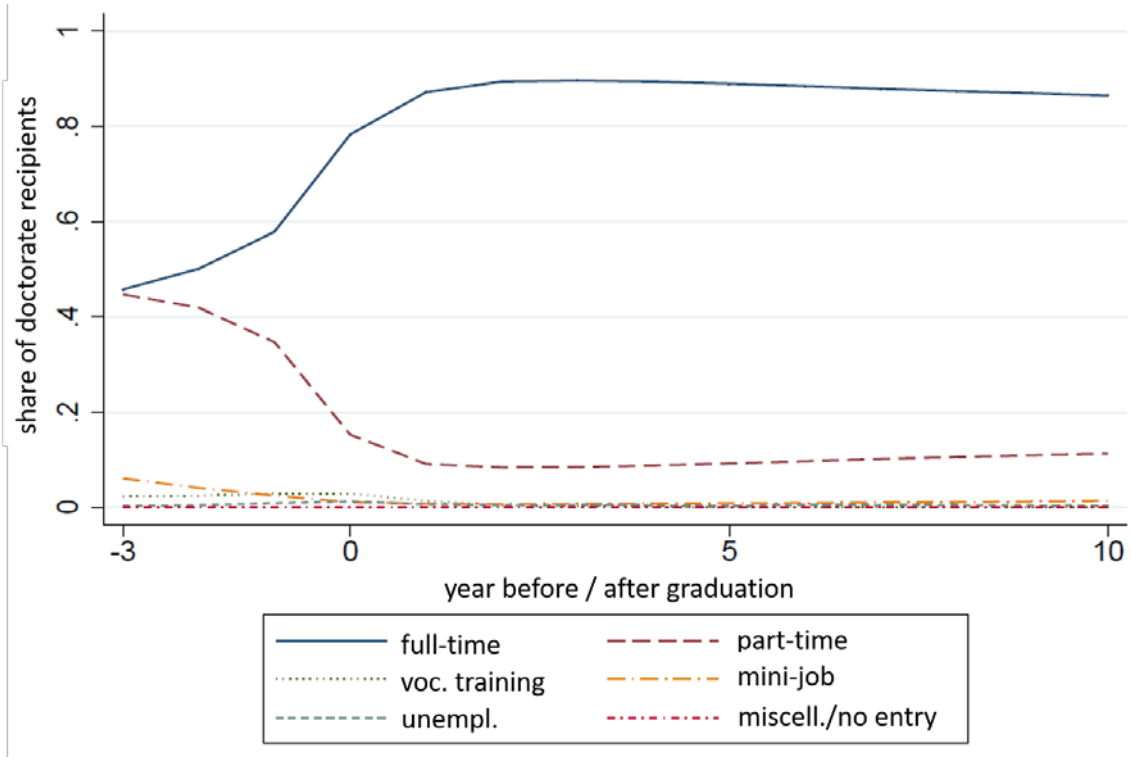
The Employment Status of Doctorate Recipients

We now investigate how the employment status of doctorate recipients change before, during, and after their doctoral studies. We differentiate among five types of employment status: full-time job, part-time job, mini-job,[11] vocational training and unemployment. Figure 3 shows the employment status of all linked doctorate recipients in the final dataset at different points in time throughout their careers. As the exact date of graduation is unknown, our point of reference (year zero) is the final day of the year the dissertation was published. Most doctorate recipients hold full- or part-time positions, with only small shares of graduates being unemployed, in vocational training or holding mini-jobs at any point in time. Doctoral students are often employed in part-time positions at universities or public research organizations. The shares of part-time employment range be-

[11] The monthly income in a mini-job does not exceed € 450. This job type is not subject to social security contributions.

tween 44.71 percent and 34.70 percent three to one year before graduation, whereas post-submission employment changes from part-time to full-time positions in academia, other parts of the public sector or in the private sector.

Figure 3: Employment status over time before/after graduation



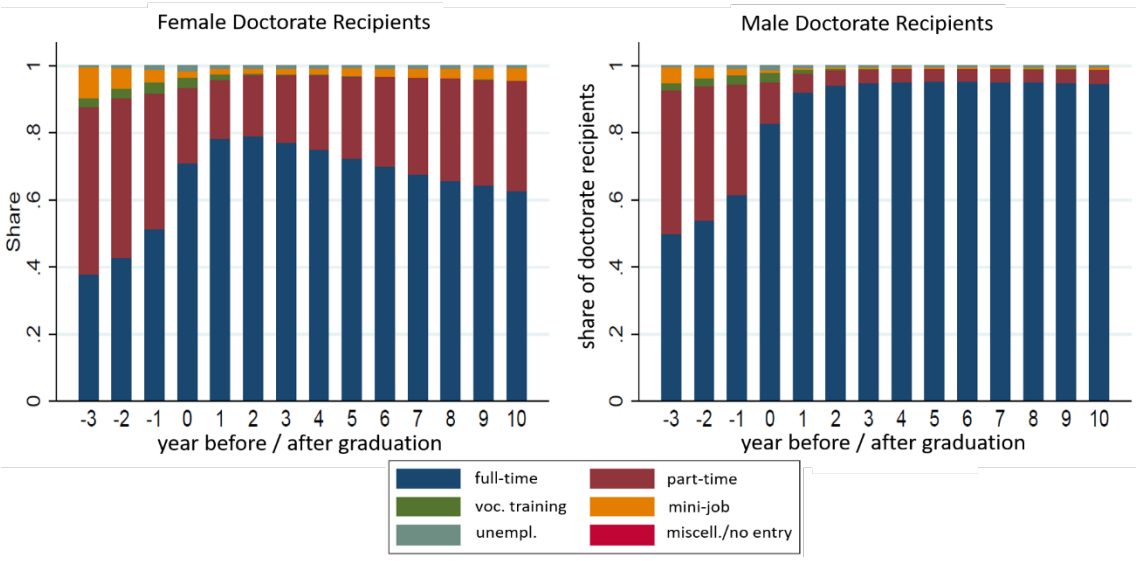
Source: own calculations.

The share of full-time jobs increases from 78.29 percent in year zero to a maximum of 89.59 percent three years later, and then diminishes to 86.46 percent in year ten after graduation. In turn, the share of part-time employment increases from 8.50 percent three years after to 11.28 percent then years after. This change can be explained with male and female doctorate recipients following different career patterns over time (see Figure 4).[12] While the majority of male graduates constantly work full time after their doctorate education, a larger share of women also have a part-time position after graduation. This gender-specific full-time gap increases over time. While 94.34 percent of men are full-time employed ten years after graduation, the corresponding share among female doctorate recipients' declines to 62.51 percent after ten years. These results are in line with existing evidence on gender-specific employment patterns, where female part-time employment is often attributed to an uneven distribution of family-related responsibilities such as childcare and care of elderly family members among men and women (Wanger 2015).

[12] For the analyses, we used a sample of the full-linked dataset, which is restricted in the following way: Since data were collected for administrative purpose, we had to correct some spell information in the data (see Kaul et al. 2016) to construct the sample for the subsequent analysis. Further, we dropped unreliable very short (un-) employment episodes (below seven days). For the analysis, we use information on all graduates at the end of a given year (December 31) for 3 years prior to and 10 years after the publication year of the dissertation.

These results indicate the data from IAB-INCHER project of earned doctorates (IIPED) is representative of the overall doctorate recipient population who enters the German labour market, particularly in more recent cohorts. The exemplary analysis of doctorate recipients' employment status over time is in line with previous findings. This dataset can therefore be employed to study a wide range of research questions related to the post-doctoral careers of doctorate recipients.

Figure 4: **Employment status over time before/after graduation separate for male and female doctorate recipients (source: own calculations)**



Source: own calculations.

5 Limitations

As shown above, machine learning provides a suitable approach to overcome limitations of traditional record linkage methods. However, machine learning comes with limitations of its own, which are in the focus of this section. Most importantly, as noted above, the linkage is based on a synthetic training and evaluation dataset. Here, unique name-surname combinations were merged with individuals sharing the same name but a different surname to receive the true-negative sample of matched entries. While this method allows us to create a database for training the algorithm that is as close as possible to the original database, this method is biased, if characteristics of surnames are dependent on (some of) the classification variables. Moreover, we carefully controlled the plausibility of the linked data for the unique name-surname combinations. Nevertheless, this check was only possible at an aggregated level of different disciplines and years before and after graduation. While the results were comparable to other findings about labour market trajectories of doctorate recipients at these aggregate levels of analysis (for example to information of the BuWiN (2013, 2017)), the chosen approach could nevertheless lead to misclassifications in individual cases. In addition, the algorithm was used only for doctorate recipients with equal or less than 300 namesakes. Even if it is expected, that the algorithm would work sufficiently

well for more than 300 potential matches for each entity, more linkage variables X would be advisable for training the function $\hat{f}$ for a precise classification.

Moreover, the IEB does not capture individuals who are not liable to social security contributions (e.g. civil servants, self-employed individuals, and family workers). Therefore, the final database may be biased towards those doctorate recipients who are part of the German social security system. For instance, certain occupations like physicians and lawyers are traditionally self-employed or employed as civil servants (e.g. pastors, teachers). These graduate groups are underrepresented in the database. Furthermore, the DNB only contains a records of published doctoral theses for German universities, while foreign doctorate recipients from non-German universities are not covered.

6 Conclusions

In this paper we describe our approach using machine learning techniques to link two sets of administrative data: a list of all German doctorate recipients collected in the catalog of the German National Library (DNB), and the Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB). Linking these datasets was motivated by the interest to study labour market trajectories of German doctorate recipients at different stages of their career. We show that supervised machine learning algorithms can be fruitfully applied to the linkage of social security data with other data.

The proposed method has several advantages over traditional methods. On the one hand its application is not restricted to micro data with overall high quality (where e.g. name-surname combinations and exact birth dates, or social security numbers, are available as unique identifiers). In addition, the quality of the matching algorithm can be assessed and compared to simple heuristics. At the same time, the approach is applicable in contexts with strong privacy requirements, as is the case for anonymous social security data.

Bearing in mind a number of limitations, an evaluation of the method provides the following insights, which may help inform further work: First, a supervised machine learning algorithm can be used for classifying individuals in administrative data. Second, in our specific application simple heuristics (as have been used in prior record linkage approaches for German social security data) reach sufficiently high rates of precision. However, machine learning algorithms combine comparably high precision with drastically improved recall. Third, dependent on the tuning parameters used, each algorithm can have a number of potential classification outcomes. This indicates the need to evaluate results from different algorithms.

The final database allows us to investigate the labour market trajectories of German doctorate recipients before, during and after their graduation from 1975 up to 2015. A first evaluation of the database provides the following insights: while only a few doctorate recipients are unemployed, we find a substantial share of female doctorate recipients working part time. While female and male doctorate recipients show similar employment states during their graduation period, shares of part-time and full-time employment diverge over the career paths of men and women.

References

Antoni, M. and Seth, S. (2012). ALWA-ADIAB–Linked individual survey and administrative data for substantive and methodological research. In: Schmollers Jahrbuch, 132(1), 141–146.

Antoni, M., Ganzer, A. and vom Berge, P. (2016). Sample of integrated labour market biographies (SIAB) 1975-2014. FDZ-Datenreport, 4/2016.

Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.

Buenstorf, G. and Geissler, M. (2014). Like doktorvater, like son? Tracing role model learning in the evolution of German laser research. in: Jahrbücher für Nationalökonomie und Statistik, 234(2-3), 158–184.

Benderly, B. L. (2018). A trend toward transparency for Ph.D. career outcomes? Science. https://doi.org/10.1126/science.caredit.aat5250.

Blank, R., Daniels, R. J., Gilliland, G., Gutmann, A., Hawgood, S., Hrabowski, F. A. and Schlissel, M. S. (2017). A new data effort to inform career choices in biomedicine. In: Science, 358(6369), 1388–1389.

Christen, P. (2012a). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. New York: Springer.

Christen, P. (2012b). A survey of indexing techniques for scalable record linkage and deduplication. In: IEEE transactions on knowledge and data engineering, 24(9), 1537–1555.

Culp, M., Johnson, K. and Michailidis, G. (2006). ada: An r package for stochastic boosting. Journal of Statistical Software, 17(2), 9.

Deutsche Nationalbibliothek (DNB). (2018). The German National Library in brief. (http://www.dnb.de/EN/Wir/ueberblick) (13.11.2018).

Dorner, M., Bender, S., Harhoff, D., Hoisl, K. and Scioch, P. (2014). The MPI-IC-IAB-Inventor data 2002 (MIID 2002): Record-linkage of patent register data with labor market biography data of the IAB. FDZ Methodenreport, 06/2014.

Dorner, M., Heining, J., Jacobebbinghaus, P. and Seth, S. (2010). The sample of integrated labour market biographies. In: Schmollers Jahrbuch, 130(4), 599-608.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. In: Journal of Statistical Software, 33(1), 1–22.

Gareth, J., Witten, D., Hastie, T. and Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: Springer.

Heinisch, D. P. and Buenstorf, G. (2018). The next generation (plus one): an analysis of doctoral students' academic fecundity based on a novel approach to advisor identification. In: Scientometrics, 117(1), 351–380.

Kaul, A. Neu, N., Otto, A. and Schieler, M. (2016). Karrierestart, Mobilität und Löhne von Absolventen der Informatik. IAB-Regional - Berichte und Analysen aus dem Regionalen Forschungsnetz. 03/2016.

Konsortium Bundesbericht Wissenschaftlicher Nachwuchs (BuWiN) (2013). Bundesbericht Wissenschaftlicher Nachwuchs 2013. Bielefeld: W. Bertelsmann Verlag.

Konsortium Bundesbericht Wissenschaftlicher Nachwuchs (BuWiN) (2017). Bundesbericht Wissenschaftlicher Nachwuchs 2017. Bielefeld: W. Bertelsmann Verlag.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. In: R News, 2(3), 18–22.

Oberschachtsiek, D., Scioch, P., Seysen, C. and Heining, J. (2008). Stichprobe der Integrierten Erwerbsbiografien. Handbuch für die IEBS in der Fassung 2008. FDZ Datenreport, 01/2018.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Schnell, R., Bachteler, T. and Bender, S. (2004). A toolbox for record linkage. In: Austrian Journal of Statistics, 33(1), 125–133.

Schnell, R. (2013). Getting big data but avoiding big brother. German Record Linkage Center Working Paper Series, 2/2013.

Teichert, C., Niebuhr, A., Otto, A. and Rossen, A. (2018). Graduate migration in Germany: New evidence from an event history analysis. IAB-Discussion Paper, 3/2018.

Vom Berge, P., König, M. and Seth, S. (2013). Sample of integrated labour market biographies (SIAB) 1975-2010. FDZ Datenreport, 1/2013.

Wanger. S. (2015). Traditionelle Erwerbs- und Arbeitszeitmuster sind nach wie vor verbreitet. IAB-Kurzbericht, 4/2015.

Wydra-Somaggio, G. (2015). Das Ausbildungspanel Saarland: Dokumentation der Datenaufbereitung. IAB-Regional. IAB Rheinland-Pfalz-Saarland, 3/2015.

Zimmermann, R., Kaimer, S. and Oberschachtsiek, D. (2007). Dokumentation des" Scientific Use Files der Integrierten Erwerbsbiographien"(IEBS-SUF V1), Version 1.0.

A Assessment of the Training Dataset

To assess the representativeness of the synthetic training- and evaluation dataset, we present descriptive statistics for both datasets. Results for the number of multiples matched entries per entity can be seen in Table A1. Table A2 shows descriptive statistics of the variable distributions for the synthetic training- and evaluation dataset. Table A3 shows descriptive statistics of the variable distribution for the full (matched) dataset.

Table A 1: Distributions for multiple matches of the synthetic training- and evaluation dataset and of the full (matched) dataset

Feature	Min	1stQ	Median	Mean	3rdQ	Max
Artificial training/ evaluation dataset	1	1	4	22.0889	20	296
Full (matched) dataset	1	1	4	22.4841	20	299

Source: own calculations.

Table A 2: Descriptive statistics for synthetic training- and evaluation dataset

Feature	Median	Mean	Min	Max
spell_research	0	0.0911	0	1
spell_hospital	0	0.1130	0	1
prop_educ	0	0.3517	0	1
age_sub	35	37.6489	20	102
same_ror_y5	0	0.0475	0	1
first_spell_before	-11	-11.2539	-45	39
year_diss	2001	2000	1975	2015
eastern	0	0.1658	0	1
nbr	90	103.1859	1	296
social science	0	0.1048	0	1
natural science	0	0.2564	0	1
engineering	0	0.0833	0	1
medicine	0	0.4001	0	1
law/economics	0	0.1187	0	1

Source: own calculations.

Table A 3: Descriptive statistics for full (matched) dataset

Feature	Median	Mean	Min	Max
spell_research	0	0.0846	0	1
spell_hospital	0	0.0964	0	1
prop_educ	0	0.3319	0	1
age_sub	35	37.3718	20	115
same_ror_y5	0	0.0573	0	1
first_spell_before	-10	-10.2697	-62	40
year_diss	1999	1998	1975	2015
eastern	0	0.1677	0	1
nbr	94	106.8058	1	299
social science	0	0.0855	0	1
natural science	0	0.2550	0	1
engineering	0	0.0882	0	1
medicine	0	0.4171	0	1
law/economics	0	0.1118	0	1

Source: own calculations.

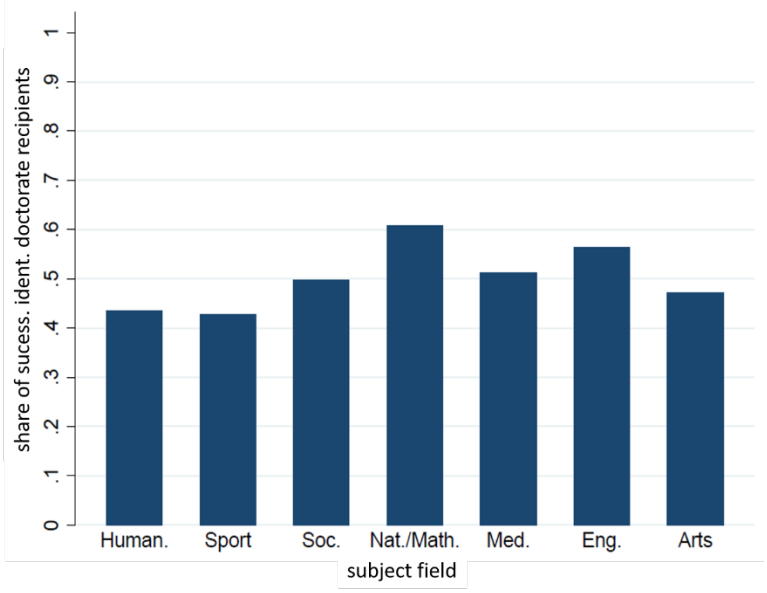
B Assessment of Merged Dataset

To check the quality of the matched IIPED data the following figures have been created.

Figure B 1: Successfully identified doctorate recipients by graduation year

Source: own calculations.

Figure B 2: Successfully identified doctorate recipients by subject field



Source: own calculations.

Imprint

IAB-Discussion Paper 13|2019

Publication date

23 May 2019

Editorial address

Institute for Employment Research (IAB)
of the Federal Employment Agency (BA)
Regensburger Straße 104
90478 Nuremberg
Germany

Download

http://doku.iab.de/discussionpapers/2019/dp1319.pdf

All publications in the series "IAB-Discussion Paper" can be downloaded from

https://www.iab.de/en/publikationen/discussionpaper.aspx

Website

www.iab.de

ISSN

2195-2663

For further inquiries contact the author

Anne Otto
Phone +49 681 849-207
E-Mail Anne.Otto@iab.de