Algorithmic identification of Ph.D. thesis-related publications: a proof-of-concept study

Paul Donner

Received: 17 January 2022 / Accepted: 29 July 2022 / Published online: 18 August 2022 © The Author(s) 2022

Abstract
In this study we propose and evaluate a method to automatically identify the journal publications that are related to a Ph.D. thesis using bibliographical data of both items. We build a manually curated ground truth dataset from German cumulative doctoral theses that explicitly list the included publications, which we match with records in the Scopus database. We then test supervised classification methods on the task of identifying the correct associated publications among high numbers of potential candidates using features of the thesis and publication records. The results indicate that this approach results in good match quality in general and with the best results attained by the “random forest” classification algorithm.

Keywords Cumulative dissertation · Doctoral thesis · Supervised classification · Early career researchers · Publication-based dissertation

JEL Classification C38 · I23

Introduction
Ph.D. candidates are active researchers and in many disciplines they publish their results in the form of articles in journals, contributions in conference proceedings, or chapters in edited books in addition to the dissertation. Because of the sheer number of Ph.D. candidates, they clearly contribute considerably to the total knowledge production. However, the extent of this contribution is currently not known with any degree of accuracy. The scale of research activity of Ph.D. candidates and their contribution to published knowledge production has been identified as a desideratum by science policy actors in Germany (Consortium for the National Report on Junior Scholars, 2017). Monographic Ph.D. theses, although not indexed in classic citation index databases, have already been shown to be amenable to citation analysis (Donner, 2021a). Being able to also automatically identify

* Paul Donner
donner@dzhw.eu

1 Dept. 2 ‘Research System and Science Dynamics’, German Centre for Higher Education Research and Science Studies (DZHW), Schützenstr, 6a, 10117 Berlin, Germany
publications related to Ph.D. theses (with a known error rate) on a large scale would be another valuable tool for studies of scientific careers and for informing science policy.

In this study we address the primary research question of whether it is feasible to automatically identify publications related to cumulative Ph.D. theses in a literature database using only bibliographic information of both the dissertations and the publications. As this is a proof-of-concept study, it is restricted to investigating the proposition that thesis-related publications can be identified automatically and does not extend to finding the best possible specific method to do so nor to applying the approach on a large-scale dataset.

In this work, we distinguish between two types of Ph.D. theses, cumulative theses (also called publication-based theses) and stand-alone theses (also called monographic theses). The difference is that cumulative theses comprise one or more previously published or submitted works such as journal articles, book chapters, or conference proceedings publications while stand-alone theses do not.

We describe the construction and evaluation of statistical prediction models for identifying publications related to Ph.D. theses produced at universities in Germany. To this end, we collected a sample of German Ph.D. theses that contained information on publications on which the theses were based or of which the theses consisted. We then experimentally evaluated the ability of supervised classification methods to distinguish related publications from other candidate publications which were superficially similar to the theses in the sense that they were also authored by persons with the same or very similar names and were published around the time of the publication of the dissertation. While data from dissertations of German universities is used, the methods are also applicable to data from other countries. Secondary research questions investigated in this study are:

- Instead of sophisticated machine learning models, might simple deterministic rules possibly be sufficient for the identification of cumulative dissertation papers? This question is motivated by the approach of Larivière (2012), detailed further below.
- What is the trade-off between precision and recall in the results of classification models for this task?
- Which variables are important predictors in this task?

Following this introduction, we proceed to summarize the state of research on identifying thesis-related publications in the next section. We then describe the methods and data of our study and present and discuss the obtained results.

Prior work

Echeverria et al. (2015) study what they call “derivative works” of British medical Ph.D. theses. Such works are defined as being produced by the same author as a thesis, sharing text and content with the thesis, and being published during or shortly after the doctoral studies. They worked with a corpus of 51 theses and 199 candidate articles by the same authors, of which 30 were considered derivative because they were explicitly mentioned in the theses. They measured text similarity across the IMRaD structure sections of the publications with a commercial plagiarism detection software. They found that the Discussion section of articles provides the most accurate similarity scores. However, they did not consider the similarity of the full-texts of thesis and candidate article as a whole. Thesis authors very often were first authors of the derivative works and supervisors appeared as co-authors.
A study on the automated identification of publications resulting from doctoral work in Brazilian nuclear physics was conducted by Zamudio Igami et al. (2014). They calculated the optimal number of common keywords from a controlled disciplinary vocabulary to link authors’ articles based on the same research project as the Ph.D. theses with a validation by the original authors using a corpus of 401 theses and 2211 candidate journal articles of the same authors. The results of the co-word analysis were validated with survey responses of the thesis authors. Using the optimal common keywords figure of ≥ 3 they achieve a Precision of 0.87 in the training sample, but no independent test sample was used.

In the largest scale study to date, Larivière (2012) investigated the publications of Québec Ph.D. students in terms of peer reviewed periodical publication output and citation impact in comparison to other publications from Québec without Ph.D. candidate co-authors. In this study, a list of all Ph.D. students’ names was matched against author names of Web of Science articles with at least one Québec address. This list needed to be validated manually and automatically. Matched articles were retained only if they were published during the Ph.D. phase of the author or the year after completion of the doctoral studies. It was found that during the observation period 2000–2007, the percentage of doctoral students contributing to at least one publication was 10% in the social sciences, 4% in the arts and humanities, while in health and natural sciences/engineering the figures were 64% and 40%, respectively. This difference was unlikely to be caused by the different coverage across the domains of research. The identified Ph.D.-co-authored papers account for about 30% of Québec’s universities’ papers in health and the natural sciences / engineering, 19% in the social sciences, and 13% in the arts and humanities. Another finding is that doctoral candidates of the earlier observation years who had completed their Ph.D. by the time of data collection had published more papers than those who had not yet completed it. This strongly suggests that two indicators of successful Ph.D.’s, timely completion and publication output, systematically co-vary. Based on the automatic part of this study’s Ph.D. student publication identification approach, we also investigate a similar deterministic rule-based approach in our study, using this as a baseline for comparison with machine learning algorithms. However, unlike Larivière (2012), we do not intellectually post-process the automatically obtained results, so a direct comparison between the two methods is not appropriate.

Breimer and Mikhailidis (1993) studied Swedish and UK medical theses which included publications and found that “when British theses contain papers, the candidate is commonly (66%) the first or sole author” while the figure for Swedish publications is 75% (accepted + submitted papers). They also found that thesis papers from the two countries had an obvious tendency to be published in journals from the respective country or region (Scandinavia). An update for later Swedish biomedical theses found that the rate of first or sole authorship remained at 75% (Breimer, 1996).

These studies show that author names, author position, publication year proximity, and textual similarity are useful in deciding if publications are closely related to Ph.D. theses. They also leave some opportunity for improvement and generalization of methods of Ph.D. thesis publication identification. Building on these results, the present study expands upon this prior research by using a sample from one country and all disciplines, by testing state-of-the-art classification algorithms which makes a fully automated processing possible, and by using new and improved predictor features.
Method and data

Data sample and task challenges.

The starting point for this study is a dataset of records on German Ph.D. theses obtained from the German National Library, which is chartered with collecting all Ph.D. dissertations completed at German universities. There is currently no other comprehensive national scale data source on Ph.D. theses. The National Library collects theses in all available formats and indexes them with a controlled vocabulary and a subject classification. As the data source to be searched for the related publications, we use Scopus. Having identified related records in a citation indexing database allows to directly apply citation analysis methods on the publications once identified. This choice has the disadvantage of incomplete coverage because the Scopus editorial committees are deliberately selective in their choice of sources to include in the database. We will investigate below to what degree the publications of the sample dissertations are covered in Scopus.

The task at hand, the identification of publications related to German Ph.D. theses, is accompanied by some challenges. The bibliographic information on theses is limited to author name, university, publication year, title, controlled vocabulary terms, and classification notations of the thesis. Note that we are restricted to using the bibliographic data and not the full-texts of theses, as the latter are not currently universally available in digital format with standardized metadata. Thesis bibliographic data will be used to filter possible related publications in a publication database, in this case Scopus. In this process, for each dissertation record, all publication records in the database need to be compared because there is no prior list of the thesis authors’ publications. Therefore the sought method needs to be fully automatic and scale well to large numbers of records. Another difficulty is that it should be possible to compare the text similarity of thesis and publications for all combinations of German and English language text, as there are many cases in which the textual components of thesis and candidate publication records are in different languages. Ideally, the method should be able to thus exclude publications of thesis authors which are not related to their thesis as these should have lower content similarity because they would presumably cover projects they worked on other than their thesis project.

In order to develop, test, and validate a method that fulfills these requirements we created a manually curated gold standard dataset of German Ph.D. theses and their related publications. We used three approaches to build this dataset. First, we downloaded thesis full-texts from the German National Library corpus using URLs in the dataset. While we were able to download many full-texts, many URLs turned out to be outdated and unreachable. We searched the obtained texts for keywords and phrases indicating a cumulative thesis. Second, we found a small number of records containing the phrase “kumulative Dissertation” in the title of the record in the National Library corpus. These were mostly from a single university. Third, we randomly sampled universities and searched their online publication repositories for dissertations containing keywords or phrases indicating cumulative theses. This sample is therefore only a convenience sample and not representative. Bibliographic records for publications referenced in the theses written by the thesis authors and indicated as being part of the thesis or appearing as chapters were extracted and manually matched with a snapshot of the Scopus bibliometric database from Spring 2019 by

1 This section repeats material from Donner (2021b)
assigning the Scopus item identifier to the extracted publication records. We chose Scopus because many thesis-related publications are published in the German language and Scopus covered more German-language literature than Web of Science. For the period from 1996 to 2017, Scopus contained around 694,000 German-language publication records and Web of Science around 500,000 records.

We screened 1181 doctoral thesis records from 77 universities from the publication years 1996 to 2018. 449 of the records were identified as cumulative doctoral theses but 21 of them did not have any Scopus-covered publications. 732 theses were identified as stand-alone theses without any incorporated publications. There were 1499 pairs of theses and Scopus-contained publications out of 1946 thesis-publication pairs in total. The Scopus coverage of this dataset’s thesis-associated publications is thus approximately 77%.

We divided the identification task into two stages. In the first stage, we narrow down the number of candidate matches by applying deterministic heuristics in order to achieve a feasibly low number of possible matches for the more sophisticated second stage. The second stage is a classification model which assigns each first-stage candidate publication a probability value for being a constitutive publication of a given thesis based on various bibliographic features.

### Filtering of candidate matches

For the first stage, the goal is to rule out as many unlikely matches as possible while discarding as few actual matches as possible with relatively simple rules. We only briefly summarize this stage, as a detailed description is already available in Donner (2021b). Scopus records are filtered (a) by author name similarity for all authors affiliated with Germany and (b) publication year range. Author name matching was implemented to be tolerant to minor differences in spelling and name part order. Candidate publications published from up to four years before to nine years after thesis publication year were retained. Preliminary analysis showed that these limits provide a reasonable balance between erroneously excluding too many relevant publications and unnecessarily inflating the number of retained candidate matches per thesis. This method is able to find candidates with an error rate of about 3% missed true candidates and retrieved about 1500 candidates per thesis with a large variability in the number of candidates per thesis record. Nevertheless, this stage reduced the dataset to 1176 thesis records, of which 419 have Scopus-covered publications while 757 theses have no Scopus-covered publications, leaving 1448 positive matches. Much of this loss is because of name changes.

To characterize this part of the final dataset some more, we make the following observations. Of the 419 theses with Scopus-covered publications, 204 have English

| Thesis language | Article language |
|-----------------|-----------------|
| German          | 161             |
| English         | 7               |

| Thesis language | Article language |
|-----------------|-----------------|
| German          | 585             |
| English         | 695             |
titles and 215 have German titles. Table 1 shows the relationship between thesis and constituent articles title languages for the 1448 thesis-article pairs in the final dataset.

Identification of matches with decision criteria and classification algorithms

In the second stage, the filtered candidates of the first stage are classified into matches and non-matches, that is, publications related to theses and unrelated ones, using a number of decision criteria (features) tested in several classification algorithms. For this purpose, the collected thesis and extracted publication records, marked as correct matches plus all other identified candidates for the sampled theses from the first stage, marked as non-matches, are split into a training and a test subset. As the scale of data proved too large for model training with the available computational resources we undersampled negative observations by randomly selecting 10% of the non-match candidates per thesis while retaining all matches. The resulting dataset consisted of 179,151 observations of which 1448 were positives, that is, thesis-candidate matches. We employed five-fold cross-validation, based on the dissertation records. The list of dissertation records was split into five approximately equally sized randomly grouped subsets. Each fold contained 235 or 236 thesis records and on the order of 30,000 to 40,000 observations (thesis-candidate publication pairs). Classification algorithms were trained on four of the subsets and evaluated on the remaining subset, for the five combinations of test and training data. A number of generally well-performing classification algorithms were evaluated on the sample.

Classification features

The classification features that we constructed can be grouped into a number of categories and we proceed by discussing these categories and the individual criteria in turn. As noted earlier, the features are limited to information extracted from bibliographic metadata and do not include full texts and citation data. This limitation was imposed by the availability of data. Should such information be available, it would certainly improve the obtainable results.

Content similarity criteria

We take it as given that Ph.D. theses and associated publications are similar, that is, they treat the same or very closely related topics. The first criterion in this category is classification agreement. The German National Library classifies works according to a German version of the Dewey Decimal Classification since 2007, and prior to that applied its own “Sachgruppen” classification. The official mappings were used to unify the data in the Ph.D. thesis records across these two systems and we further simplified the system to 40 classes by grouping closely related classes. We then constructed a mapping to the higher level of the Scopus All Science Journal Classification (26 fields, using the first two digits of the 4-digit codes) consisting of 48 relations. The bivariate criterion variable “classification agreement” is set to 1 if any of the Ph.D. records classes and any of the ASJC codes of the candidate publications match according to the mapping table, else it is 0. It might be an option to also include the specific classification class of the
thesis. However, we refrained from doing so as there are not enough observations for all classes to make this a reliable predictor.

The second criterion is textual similarity. As this is a considerably complex task in its own right, a separate study for the selection of suitable methods was carried out in Donner (2021b) and the reader is referred to this study for a detailed description of the methods. To summarize, the difficulties relate to the sparsity of the textual information, the bilingual nature of the texts, and domain specificity of the texts. Evaluation of several approaches using ground truth data revealed that classic language-agnostic Vector Space Model cosine similarity and language-aware Random Indexing cosine similarity, in both cases with tf-idf weighting, stopword removal, and stemming, performed best for this task. We computed text similarity scores between dissertation titles plus indexing terms and candidate publication titles plus abstracts for these two methods as variables “VSM cosine” and “RI cosine.”

The Vector Space Model similarity relies on exact term matching and is very reliable if the texts contain the same terms. Thus this method has an inherent bias against cross-language matching, that is, cases in which thesis text data and publication text data are in different languages. Random Indexing similarity is not affected by this problem. This method can project terms from different languages into the same low-dimensional embedding space which needs to be learned by supervised training from a sufficiently large bilingual text corpus. In this learned embedding space semantically similar terms, irrespective of language, have small distances while unrelated terms have large distances. Thus RI similarity can not only provide reliable similarity values for texts from different languages but also for semantically related short texts which have no matching terms in common. We include the VSM method despite its drawbacks in cross-language cases because of its good performance in same-language cases, which constitute a large fraction of all cases, and in cases with exact term matches.

**Author criteria**

In the pre-selection step we used a relatively tolerant author name similarity function to include as many plausible records as possible and minimize the false negative rate. As it is more likely that the name of the thesis author and that of the publication author refer to the same person if there is no or only a very small difference between the name text strings, we calculate given name dissimilarity and family name dissimilarity using the Levenshtein distance function. Because given names were quite often abbreviated to initials in the publication data, we compared these cases to equally abbreviated versions of the thesis author names. Family names are normally not abbreviated, so they are compared separately without the optional abbreviation step. The Levenshtein function returns the number of character changes needed to transform one name into the other, thus higher values mean more different names. The variables are called “last name Levenshtein distance” and “first name Levenshtein distance.” We also include the presumed author’s position in the author byline of the publication (“author position”) and the total number of authors (“author count”).

**Timing criteria**

Thesis-related publications can be assumed to be published in a certain time frame before and after thesis publication as the research reported in these papers is completed at the time of thesis publication. Candidate thesis-related publications published longer before or
after thesis publication year are less likely to be true constitutive publications than those published closer in time to the thesis. We therefore construct a variable for the difference between dissertation publication year and article publication year and, to account for non-linearities, another variable for the square of this difference ("publication year difference," "squared publication year difference"). Furthermore, as thesis publications are typically published at the beginning of a researcher's career, we use the difference between dissertation publication year and candidate publication author’s first article’s publication year, as obtained by selecting the year of the first publication of the author as identified by the Scopus Author ID system. Again we also include the square of the number ("first publication year difference," "squared first publication year difference"). As cumulative theses were becoming increasingly more common over the course of the period covered in the dataset, thus increasing the baseline probability that theses from the later part might have any associated publications compared to theses from the early part of the time period, we also include the publication year of the thesis as a predictor variable ("thesis publication year").

### Location and language criteria

Using the cleaned and mapped university data from the thesis records in conjunction with processed affiliation data for German universities (Donner et al., 2020; Winterhager et al., 2014) a binary variable was constructed for university agreement between the thesis and the publication affiliation ("university match"). We further included a variable indicating the presence of any German address in the publication affiliations and a variable for German language publications ("German address," "German language"). While we stipulated above in the first selection step that only names with a German affiliation are considered, it is still possible that other publications of persons with such names are chosen as candidates which themselves do not have German affiliations. Thus the idea behind this and the

| Variable                                | Mean for non-matches | Mean for matches |
|-----------------------------------------|----------------------|------------------|
| classification agreement                | 0.15                 | 0.48             |
| VSM cosine                              | 0.00                 | 0.15             |
| RI cosine                               | 0.17                 | 0.41             |
| last name Levenshtein distance          | 2.89                 | 0.21             |
| first name Levenshtein distance         | 1.45                 | 0.88             |
| author position                         | 197.14               | 1.44             |
| author count                            | 377.79               | 4.62             |
| publication year difference             | 1.54                 | 1.07             |
| squared publication year difference      | 19.43                | 3.83             |
| first publication year difference        | 10.28                | 3.28             |
| squared first publication year difference| 141.53               | 16.65            |
| thesis publication year                 | 2011.03              | 2011.50          |
| university match                        | 0.04                 | 0.78             |
| German address                          | 0.65                 | 0.99             |
| German language                         | 0.06                 | 0.12             |
German language variable is to better distinguish publications from Germany which are intrinsically much more likely to be relevant than others.

Table 2 gives a summary of the average values for the criteria variables for candidate publications which are non-matches and those which are matches.

**Classification algorithms**

We tested a number of supervised classification algorithms in the R programming environment. In this context it must be stressed that the purpose of this study is not primarily to identify the single of best-performing algorithm in this task but to investigate if automatic classification algorithms as a class can succeed at the task of identifying constitutive publications of cumulative dissertations given the restrictions of the available data. The use of several fundamentally different algorithms serves two purposes. First, to make the results more robust such that a possible failure of one specific algorithm to perform the task would not lead to the conclusion that no classification algorithm might work well. And second because it is generally difficult to predict which algorithm will work best for a new task a comparative evaluation scenario was considered appropriate. We also used a deterministic baseline procedure as a basic model to compare the algorithms to. In this model we use only the publication year difference, sum of given name dissimilarity and family name dissimilarity, and university agreement as criteria, all with fixed permissible ranges.

The following classification algorithms were tested:

- Logistic regression, a generalized linear model with a binary outcome. We use the base R implementation in function glm(). Logistic regression models the probability of a binary outcome variable taking the value 1 by a linear combination of a set of binary or continuous predictor variables whose parameters are estimated.
- Random forest classification (Ho, 1995). We used the implementation offered in the randomForest package (Liaw & Wiener, 2002). The random forest method trains a number of different decision trees to predict the outcome variable based on the predictor variables and aggregates the results of these trees to improve the prediction over the result any given single decision tree would achieve.
- Single hidden layer neural network classification from the nnet package (Venables & Ripley, 2002). In the context of supervised classification, neural networks can be considered a non-linear, high-parameter extension of regression functions which gives them great flexibility (Venables & Ripley, 2002, p. 243).
- Extreme gradient boosting (Chen & Guestrin, 2016) as implemented in the xgboost package (Chen et al., 2021). Extreme gradient boosting, like random forest classification, learns decision trees and uses an aggregate of the prediction of many trees as its final output.

The random forest, neural network, and extreme gradient boosting algorithms were parameter-tuned on the first of five cross-validation folds according to their manuals before training proper.
Results

Algorithm comparison

The four tested classification algorithms are evaluated by the correctness of their predictions about out-of-training-sample observations in fivefold cross-validation. That means that for each method, five models are trained on approximately 80% of the data and the trained models are used to predict the response variable (match or no match) of the remaining data, the fold’s test sample, to evaluate the performance. All methods return match probability scores in the range from 0.0 (sure non-match) to 1.0 (sure match) as predictions. In Fig. 1 we show for each of the four algorithms the curves for precision and recall calculated across 99 thresholds (from 0.01 to 0.99 in 0.01 increments) of the probability scores. Precision and recall are the classic evaluation metrics of information retrieval. Precision is the fraction of true positives within retrieved objects; recall is the fraction of retrieved objects among all relevant objects. As we systematically increase the threshold for the prediction probability required to accept a candidate as a match, the number of accepted candidates diminishes but the set of retained candidates becomes increasingly free of false positives. In other words, with increasing threshold, precision monotonically increases, while recall decreases. For each method and threshold there are five values for precision and recall, one from each cross-validation fold. The figures show the averages across folds at each threshold as wide lines and the minimum and maximum scores as dotted lines. We find that all algorithms give broadly similar results.

Fig. 1 Precision and Recall over prediction probability thresholds
Each algorithm is able to attain 0.8 recall at 0.8 precision at some match probability threshold value. Importantly, the ranges between minimum and maximum prediction and recall values across all threshold levels for all methods are relatively small, indicating stable models and sufficient training data.

Table 3 shows some more detailed numerical results for the single prediction probability threshold value of 0.5, an obvious crossover point to set. It can be seen that the performance across the algorithms is comparable in terms of variability, as shown by the ranges of precision and recall across the five folds. The specific point values are also similar albeit not identical due to the differing performance profiles of the methods across the range of threshold values.

Figure 2 shows the 5-way average precision and recall curves for the methods in juxtaposition and also includes additionally the precision/recall scores for a deterministic rule classification approach as a baseline. By varying acceptance ranges for a few basic
variables similar to those in Larivière (2012) by trial and error we arrived at the best combination according to F1 measure with precision of 0.45 and recall of 0.65. This is obtained at these values: “publication year difference” in [-1, 4], (“last name Levenshtein distance” + “first name Levenshtein distance”) < 3 and “university match” = true. It can be seen in Fig. 2 that the performance of all classification algorithms is superior to such a simple baseline. Moreover, all classification algorithms perform quite similarly. The random forest method shows some slight advantage in that it can attain higher precision values at given low recall values compared to the other methods or at the same high precision values has higher recall than other methods, which can be seen in the Fig. 2 where its curve is positioned more towards the top and right than the others. We interpret these results as that it is not the specific classification algorithm which makes a decisive difference for this task but that is more the nature of the data and the size of the training data that determine the results.

**Predictor variable analysis**

Having achieved generally satisfactory classification performance and having identified random forest as the best performing method, we now consider the subsidiary research question on which are the best predictors to identify cumulative thesis publications. Table 4 shows the output of the variable importance function (R package randomForest) for the random forest model trained on the fifth cross-validation fold, specifically the

| Predictor variable | Mean decrease accuracy |
|--------------------|-----------------------|
| University match   | 77.0                  |
| Author position    | 66.2                  |
| Last name Levenshtein distance | 65.1          |
| Publication year difference | 58.5            |
| VSM cosine         | 48.8                  |
| Thesis publication year | 46.5              |
| RI cosine          | 40.8                  |
| First publication year difference | 37.5              |
| Squared first publication year difference | 36.8          |
| Author count       | 36.2                  |
| Squared publication year difference | 33.4          |
| First name Levenshtein distance | 29.4          |
| German address     | 21.0                  |
| Classification agreement | 11.6         |
| German language    | 5.6                   |

2 Note that Larivière (2012) did in fact manually post-process their data after basic matching, so this observation is not intended as a criticism of their approach.
permutation-based “mean decrease in classification accuracy” measure. The listed variables are those described above. The results indicate that the most important prediction criteria are whether the thesis and publication are associated with the same university, the author position of the presumed thesis author in the publication author list, the degree of author last name similarity, and the publication year difference. All of these have already been used or at least suggested in the prior literature. All other tested predictors are also contributing to the matching performance, with possibly the exception of “German language” which is the least important criterion. Of considerable importance are the text similarity predictors, the thesis publication year, and the difference between thesis publication year and the author’s first publication year of any Scopus-indexed paper, which can be understood as a proxy for career duration. The inclusion of squared time difference variables also proved useful, indicating some non-linearity in the predictive power of these variables. Based on these results, there is no strong case to exclude any of the tested variables from a final model.

**Trade-off analysis**

When considering the actual application of such a model to new data, it is evident that precision and recall are in a direct trade-off relationship – it is not feasible to attain complete results with perfect correctness. Considering more closely the average data for the random forest model in Figs. 1 and 2 it appears that a generally robust model with balanced precision and recall can be obtained by specifying a prediction probability threshold of 0.45, resulting in precision and recall both at 0.83. On the other hand, in some application scenarios more accuracy (less false positives) may be desirable at the cost of less complete data. Precision of 0.90 is possible with recall of 0.60 (at threshold 0.74). On the other hand, if more complete data is required but the loss in accuracy can be tolerated, then a recall of 0.95 with precision at 0.59 (threshold of 0.10) seems suitable.

**Discussion**

We have carried out a proof-of-concept study attempting to automatically identify publications which are constituent parts of cumulative doctoral theses. We introduced a two-stage process. The first stage identifies possible candidate publications based on a combination of country and names of authors and publication years with high recall. The second stage uses a classification algorithm to specifically identify which candidates are most likely to be thesis-related publications. The two stages were analyzed with purposely collected ground truth data. Evaluation of the classification experiment shows that any of the tested methods can generally predict correct publications with satisfactory performance and all

---

3 Described succinctly by Strobl, Boulesteix, Zeileis, & Hothorn (2007) thus: “The most advanced variable importance measure available in random forests is the ‘permutation accuracy importance’ measure. Its rationale is the following: By randomly permuting the predictor variable $X_j$, its original association with the response $Y$ is broken. When the permuted variable $X_j$, together with the remaining unpermuted predictor variables, is used to predict the response, the prediction accuracy (i.e. the number of observations classified correctly) decreases substantially, if the original variable $X_j$ was associated with the response. Thus, a reasonable measure for variable importance is the difference in prediction accuracy before and after permuting $X_j$.”
methods outperform a baseline heuristic rule-based selection method by a large margin (cf. Larivière (2012)). The random forest algorithm exhibited the overall best performance and its outputs can be used to choose high-recall, high-precision, or balanced precision-recall results. Practically all of the investigated predictor variables, constructed from bibliographic data only, turned out to contribute positively to prediction performance, including some that were not used in any prior studies. In particular, the text similarity measures, traditional VSM cosine similarity and Random Indexing cosine similarity based on a semantic bilingual embedding space, and the thesis publication year were important predictors.

The present study has several limitations that should be addressed and considered for the interpretation of the results. As we only used available bibliographic data and not full-texts and data extractable from full-texts, our study did not include the presence of thesis supervisors as candidate publication co-authors as a predictor. If this data could be obtained at scale it would most likely increase prediction performance as Ph.D. candidates frequently co-author with their supervisors. Comprehensive availability of thesis and publication full-texts would also allow the calculation of textual similarity at the level of complete works, which can be expected to lead to better predictive power of the text similarity variables used in this study and more interestingly, allow for the creation of a variable which indicates that the text of a publication is contained within the text of a thesis, as would be expected for cumulative thesis publications. We did not use any citation-based variables as the references of German theses are not indexed in any citation indexing database. If such data were available, citation links between theses and publications would also be a good predictor for the present task. A more serious limiting factor is the degree of coverage of publication databases of the material included in cumulative theses. For our dataset, Scopus covered some 77% of publications. However, much of the works missing in Scopus were either not formally published (manuscripts, working papers, reports) or appeared in locally oriented serials or books.

As for the available data, some other limitations warrant discussion. We chose not to include the discipline of a thesis as a predictor due to low numbers of observations per class. With larger size of the training dataset, this variable could be included to benefit from the different baseline proportions of cumulative versus monograph theses across disciplines. We had to undersample negative observations, which are massively more prevalent than positive cases, due to computational restrictions. In the limited subsample, positive cases were still far more rare than negatives. Yet, the general prediction quality of all tested methods and limited variability across cross-validation folds indicates that this did not negatively affect model training. Another possibly relevant predictor variable which we did not pursue in the present study was whether the candidate publication was published in a local journal (cf. Breimer & Mikhailidis, 1993). We encourage this to be tested in follow-up research.

These concerns notwithstanding, the results show very clearly that publications associated with cumulative publications can successfully be identified automatically at scale with reasonable accuracy. In fact, the nature of the data suggests that the performance might be even better when just publications of Ph.D. graduates in general are desired rather than specifically only their thesis-related publications. It seems likely that Ph.D. graduates are involved with research besides their thesis project. They may appear as co-authors of thematically related publications. Algorithmic approaches like ours would be likely to identify such observations as thesis-related publications even if they are not strictly part of the thesis, thus they would be considered false positives. However, if the research contribution of Ph.D. graduates in general is of interest, they would be relevant cases. Before implementations in practice, the findings from this feasibility study should be supplemented by an improved model which should be estimated from a true random sample of theses stratified by scientific disciplines.
Acknowledgements. Funding was provided by the German Federal Ministry of Education and Research [grant numbers 01PQ16004 and 01PQ17001]. We thank Beatrice Schulz for her help in data collection.

Funding. Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflicts of interest. The author has no relevant financial or non-financial interest to disclose.

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Breimer, L. (1996). Authorship on and usage of published papers in current Swedish biomedical theses. *Scientometrics, 36*(2), 255–258. https://doi.org/10.1007/bf02017318

Breimer, L. H., & Mikhailidis, D. P. (1993). Towards a doctoral thesis through published works. *Biomedicine & Pharmacotherapy, 47*(9), 403–407. https://doi.org/10.1016/0753-3322(93)90106-U

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., … Li, Y. (2021). Xgboost: Extreme gradient boosting. R package version 1.3.2.1.

Consortium for the National Report on Junior Scholars. (2017). *2017 National Report on Junior Scholars. Statistical Data and Research Findings on Doctoral Students and Doctorate Holders in Germany. Overview of Key Results.* Retrieved from https://www.buwin.de/dateien/buwin-2017-keyresults.pdf

Donner, P. (2021a). Citation analysis of Ph.D. theses with data from Scopus and Google Books. *Scientometrics, 126*(12), 9431–9456. https://doi.org/10.1007/s11192-021-04173-w

Donner, P. (2021b). Identifying constitutive articles of cumulative dissertation theses by bilingual text similarity. Evaluation of similarity methods on a new short text task. *Quantitative Science Studies*. https://doi.org/10.1162/qss_a_00152

Donner, P., Rimmert, C., & Van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies, 1*(1), 150–170. https://doi.org/10.1162/qss_a_00013

Echeverria, M., Stuart, D., & Blanke, T. (2015). Medical theses and derivative articles: Dissemination of contents and publication patterns. *Scientometrics, 102*(1), 559–586. https://doi.org/10.1007/s11192-014-1442-0

Ho, T. K. (1995). Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition, 1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics, 90*(2), 463–481. https://doi.org/10.1007/s11192-011-0495-6

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R News, 2*(3), 18–22.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 1–21. https://doi.org/10.1186/1471-2105-8-25

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Winterhager, M., Schwechheimer, H., & Rimmert, C. (2014). Institutionenkodierung als Grundlage für bibliometrische Indikatoren. *Bibliometrie - Praxis Und Forschung, 3*(14), 1–22.

Zamudio Igami, M. P., Bressiani, J. C., & Mugnaini, R. (2014). A new model to identify the productivity of theses in terms of articles using co-word analysis. *Journal of Scientometric Research, 3*(1), 3–14. https://doi.org/10.4103/2320-0057.143660