Received: 8 September 2022 | Accepted: 30 September 2023

DOI: 10.1002/pan3.10551

PEOPLE AND NATURE | BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

The relation between biodiversity in literature and social and spatial situation of authors: Reflections on the nature–culture entanglement

Lars Langer[1,2,3] | Manuel Burghardt[3] | Roland Borgards[4] | Ronny Richter[1,5,6] | Christian Wirth[1,5,7]

[1]Department of Systematic Botany and Functional Biodiversity, Leipzig University, Leipzig, Germany; [2]Helmholtz-Centre for Environmental Research GmbH (UFZ), Leipzig, Germany; [3]Computational Humanities Group, Leipzig University, Leipzig, Germany; [4]Department of German Literature, Frankfurt am Main, Germany; [5]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany; [6]Geoinformatics and Remote Sensing, Leipzig University, Leipzig, Germany and [7]Max-Planck Institute for Biogeochemistry, Jena, Germany

Correspondence
Lars Langer
Email: lars.langer@uni-leipzig.de

Funding information
Leipzig University; Helmholtz Centre for Environmental Research; German Center for integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Grant/Award Number: DFG FZT 118

Handling Editor: Shonil Bhagwat

Abstract

1. Understanding the nature–culture entanglement by combining the methods of natural sciences and humanities is little approached in neither of the fields. With a specific combination of methods from both digital humanities and ecology, we aimed at identifying several of people's life circumstances that relate to their individual sensitivity towards biodiversity. The circumstances with a strong correlation could be considered and targeted by decision-makers, for example by developing specific education programmes for making people more eco-conscious or adjusting relevant regulations.
2. We applied machine learning techniques onto a database including information about the frequency of biodiversity mentioned in creative literature (BiL) from 1705 to 1969 as response variable related to metadata about the corresponding works and their authors as predictors, including localisation, age, gender and literature genre. The algorithm determined the response's dependency on each predictor, which can be interpreted as the intensity of this particular sensitivity parameter for biodiversity, and which we also related to time.
3. We recognised that gender, age, region and settlement size are predictors significantly correlated to BiL. Statistically, these predictors can be viewed as starting points of the eventual individual level of awareness for biodiversity. For example, authors from villages exhibit a higher BiL than those from cities, which we interpret as a signal for the dependence of awareness for biodiversity on spatial distance from nature, which in turn can be addressed in urban development.
4. Our conclusion is that applying a machine learning technique on literary data yields meaningful results, thereby showing potential for further similar investigations and the combination of methods from natural sciences and humanities to

achieve so far unattainable insights. With our study, these insights could contribute to ecologically based decision-making processes.

KEYWORDS
biodiversity in literature, computational literary studies, corpus study, cultural ecology, environmental humanities, nature's contribution to communication, non-material contribution, random forest analysis

1 | INTRODUCTION

The Earth's biodiversity, a prerequisite for ecosystem functioning, is drastically declining as a result of global anthropogenic influences (Cardinale et al., 2012; IPBES, 2019; Millennium Ecosystem Assessment, 2005; Tilman, 1999). Recent advances in ecology and social sciences confirm that human society fundamentally relies on nature's contributions to people (Díaz et al., 2015, 2018) which are compromised by the loss of biodiversity (IPBES, 2019; Millennium Ecosystem Assessment, 2005; Schmid et al., 2009). To gain awareness of this nature–culture entanglement (Barad, 2007; Bennett, 2010; Haraway, 2016; Haraway & Begelke, 2003) and to advocate on nature's behalf, it appears necessary to mitigate the aggravating biodiversity crisis which impairs ecosystem functioning and thus our livelihood globally (Díaz et al., 2015; IPBES, 2019; Pascual et al., 2017). However, humanity's well-being is correlated not only to material but also to non-material contributions to our recreation, education and other cultural uses (Lautenbach et al., 2019; Pascual et al., 2017; Seppelt et al., 2011). Cultural inspiration and mental health, essential components of well-being, are highly dependent on the interaction between culture and nature (Methorst et al., 2021). Given this close relationship, nature's objects and processes have always been a source of inspiration and essential building blocks for the arts (Montana Hoyos & Fiorentino, 2016; Schmidt-Przewozna, 2008). In the field of literature studies, the humanities recognised the necessity to account for and understand the entanglement of culture and nature by creating the fields of environmental humanities (Emmett & Nye, 2017; Rose et al., 2012; Sörlin, 2012; Vidal & Dias, 2015) and ecocriticism (Buell, 1996; Bühler, 2016; Glotfelty & Fromm, 1996).

In our study, we determine personal and social parameters relating to the propensity of authors of literary texts to use biodiversity terms in their works. We work with the concept of 'situated knowledge', stating that knowledge is always bound to specific circumstances in time and space (Deuber-Mankowsky & Holzhey, 2013; Haraway, 1988; Hinton, 2014), as are scientific findings (Latour, 1999; Rheinberger, 2010). With this analysis, we correlate two relations, namely the relation of authors to their life circumstances shaping their situated knowledge and the relation and interdependence between living nature and literary works as important cultural products. Thereby, we hope to shed light on some conditions that constitute and promote the nature–culture entanglement. By a better understanding of how nature finds its way into literary texts, we also strengthen the role of humanities in the research on biodiversity.

For improving human well-being and developing a sustainable society protecting and using biodiversity, it is essential to understand the determinants of people's valuation of living nature. To this end, first attempts have been made to qualitatively assess (Ainscough et al., 2019; Queiroz et al., 2015) and, more recently, quantitatively investigate cultural products like literature or movies (Celis-Diez et al., 2016; Langer et al., 2021; Prévot-Julliard et al., 2015) with respect to references to living nature. In particular, we apply a measure of mentioned biodiversity in literature (BiL) within a size-normalised work proposed by Langer et al. (2021). Analysing cultural products as a means of assessing people's valuation of nature, Langer et al. (2021) recently argued, sharing the view with Mesoudi (2011), that culture can be viewed as information transferred between individuals and that nature's contributions towards communication is an insightful aspect of non-material contributions. These can be made tangible by diachronically assessing biodiversity awareness within the society, denoting an essential indicator for human attitude towards nature (Langer et al., 2021).

In our study, the cultural products we focus on are literary works. Although this is speculative to some extent, we argue that there is a significant correlation between biodiversity awareness of authors, as a major source of relevant historical data, and the number and diversity of terms related to nature they use in the creative writing process (Langer et al., 2021) which is why we interpret BiL as approximation of biodiversity awareness. Recent advances in the Digital Humanities (Jannidis et al., 1999; Klaussner et al., 2015) show that authors exhibit a unique style of writing, indicating different valuations, preferences and knowledge. This can be viewed as a by-product of their individual development, which is most likely driven by experiences and education, including, for example gender, living location, age and parenthood. We argue that such drivers bring forth a characteristic fingerprint in the author's style and content in all relevant realms of life, hence including the surrounding biodiversity and thus BiL, which can be used to make inferences on the author's biodiversity awareness.

Amongst others, this proxy of awareness may reflect facets of disconnecting humans from nature by spatial distancing (i.e. urbanisation), as well as anthropocentric transformation and exploitative depletion of nature brought forth by the industrialisation. In our study, we take advantage of the data and workflow recently developed by Langer et al. (2021) to determine parameters for the individual sensitivity, as a precondition for awareness, towards nature for a literature corpus. We regard our corpus as a representative entity of cultural products and a manifestation of a kind of preserved

communication that, in contrast to plain text collections like social media posts, is well-described with respect to BiL and sensitivity parameters as introduced below. In the middle of the investigated time period from 1705 to 1969, there was a significant peak of BiL in the corpus. Therefore, resorting to this corpus we can be certain to also cover this peak when investigating the sensitivity for biodiversity. For creating our corpus, we reduced the corpus of Project Gutenberg as represented by the Standardised Project Gutenberg Corpus (SPGC) to literary works in the given time period (1705–1969), yielding a total of nearly 16,000 works of about 4000 authors, which we regard as a large enough corpus to yield interpretable results and to be statistically meaningful. In addition to the SPGC, to each of the works we had assigned an indicator for biological diversity.

However, in order to understand human attitude towards nature, it is necessary to shift the focus from works to individual authors and to analyse determinants of their sensitivity towards biodiversity. So far, BiL has only been assessed diachronically (Langer et al., 2021), but there have been no efforts to assess this sensitivity, as its quantification is computationally challenging and would require extensive and processible metadata characterising various relevant characteristics of authors and their living conditions. To this end, we compiled an extensive database of author- and work-related metadata using a combination of automated and manual approaches, which will also serve as a valuable source for the scientific community in the field of literary studies and digital humanities for future investigations.

As the main focus in our investigation, we want to elicit the correlation between BiL and authors' living circumstances. We opted for five specific circumstances with assumed explanatory power, namely gender, main residence, main region, age and parenthood. Nevertheless, we take several other living circumstances and parameters of their works, such as literary form, genre and the extent of the used vocabulary into account. We consider our results to be highly relevant for culturomics and biodiversity education, for example by understanding what drives people to value their environment and which aspects have to be taken into account by decision-makers, when aiming to increase people's environmental awareness, respect and sustainable behaviour.

For our hypothesis, we included gender of authors as one of the five living circumstances, since all through the time investigated females and males differ notably from each other in their self- and cross-perception, creating different mental and social focal points affecting attitude towards, and perception and awareness of biodiversity (Laslett & Brenner, 1989). One expectation is that women might exhibit a higher BiL as for several minor, but coacting factors, which we detail in the discussion. As a precursor for our investigation, we compared the BiL between the genders (Appendix 1) for an impression of the actual biodiversity within the corpus of our investigation, where we observed a notable difference between male and female authors in the 100 most frequently used taxon labels. We second included the region of the authors' main residence, as different regions of the world underwent different developments, facing different challenges and events, shaping the themes and subjects of authors' writings. Our expectation is that authors are more exposed to nature in regions with stronger ongoing phases of exploration and settlement in the time of investigation, for example North America in the first half of the 19th century, and therefore exhibit a higher BiL. We third included the author's main residence, a reflection of the closeness to nature in their everyday lives, by distinguishing whether they lived in a village, town or city. Our hypothesis specifies that authors living in a more rural setting have a higher awareness and thus a higher BiL than those living in an urban environment. We regard attitude and awareness as dependent on the stages within the individual psychological ontogenesis, which is why we fourth included the age of the author at the time of publication as a relevant predictor. Finally, we expect the circumstance whether an author had children to be correlated to BiL, since parenthood in our current perception likely influences personal life goals and therefore general views and attitudes. Moreover, children tend to have a closer, more direct connection towards their environment, where they give in to their curiosity and which typically leads to parents accompanying them outside more frequently. Alternatively, one could argue that people without children have more freedom to travel and experience the natural world, or that the influence of parenthood particularly on a male person in the beginning of the investigated time period may be negligible. Generally, we would expect a changing influence over time because of the historical variability of the paradigm of parenthood. The implications and limitations of our results have to be assessed carefully, to which end we took multiple perspectives and paradigms in our discussion into account.

2 | METHODS

2.1 | Principal preparation

As a foundation for our research to determine the individual sensitivity for biodiversity on the basis of communication analysis, we selected the Project Gutenberg as our corpus (collection of texts). In a previous study (Langer et al., 2021), we determined several biodiversity indices for each selected literary work in the corpus, like biodiversity richness (number of unique terms for living things), abundance (total number of terms for living things), Shannon diversity and Simpson diversity (both indexes combining richness with a measure of distribution over the different terms). In addition, we determined the lexical richness (the vocabulary used) per literary work to control for both absolute and relative changes in biodiversity richness. All indices were produced via a size-normalising bootstrapping algorithm in order to circumvent biases that potentially were introduced by differences in the length of works as determined by the number of words used. To this end, we averaged the determined indices of repeatedly produced virtual works consisting of 10 randomly sampled 1000-word frames of the respective original work.

For this subsequent analysis, we needed a database containing the relevant sensitivity-related predictors for authors and the response variable. As a response variable, we decided for biodiversity richness per size-normalised work for three reasons. First, richness

is a linear variable so it can easily be compared on different scales, second, it remains comparable to the lexical richness determined as a control variable, and third, it is a meaningful and descriptive measure in itself, placing emphasis on the fact that authors are aware of the existence of a specific quantity of species.

None of the sensitivity-related predictors of our analysis was present in the metadata delivered with the corpus and only one, namely lexical richness, could be directly retrieved from the data. In order to collect comprehensive metadata for this plenitude of authors and literary works we first extracted available data from Wikidata. However, these data were incomplete and often not directly usable for computational processing in the form provided by Wikidata, but required interpretation and categorisation as detailed below. Over a period of 2 years, we complemented and, if necessary, appropriately transformed data to build a comprehensive collection of various metadata. The large number of authors and works made it necessary to restrict the number, extent and time spent for each predictor variable, which is why we primarily used a specific set of additional openly accessible data sources, namely Wikipedia.org, LibraryThing.de, isfdb.org and WorldCat.org. The final corpus, comprising entries of creative literature in English language (including translations from other languages) with sufficient metadata, contains 13,493 records (works) from 2847 Authors.

2.2 | Sensitivity parameters

As predictor variables for our model, we selected five core variables. The variables are descriptions of an author's situation that we quantified either categorically or continuously, and for which we detail our hypotheses below. In order to disentangle signals from multiple possible predictors influencing BiL, we included a number of covariates in our model that represent additional author-related as well as work-related characteristics (e.g. biological background in education or occupation, literature form, vocabulary and genre), on which we take a closer look in our discussion. To distinguish the individual correlation between each predictor and BiL, we performed a random forest (RF) analysis (Breiman, 2001) on the predictor dataset formed from our metadata collection with biological richness declared as response variable. This method yielded numerous unambiguous correlations, confirming, respectively, contradicting intuitive hypotheses about the trend and the magnitude of the selected predictors' impacts.

In addition to our core variables allowing us to estimate sensitivity parameters, we considered several covariates as controls, which we expect to significantly correlate with BiL. These investigated predictors of BiL are either author-related or work-related (Figure 1).

For each predictor variable, we defined its type (categorical or integer) and a set of relevant levels for the categorical variables. Gender was available as a binary variable, either male or female, if not, we derived the gender information on the grounds of an author's portrait or forename. To determine the main region of the authors, we first collected information about the regions the authors evidently lived in for 2 years or more in (a) their childhood and adolescence up to the age of 20, and (b) their adulthood, each prioritised by the length of their stay. We distinguished four categories for regions: British Islands, Mainland Europe, North America and Other, whereby the latter, containing mainly Australia, New Zealand, South Africa, China and unassigned entries, as well as singlets or doublets from Brazil, Uruguay, Mexico, the Philippines and Japan, was conflated because (a) of the comparably small numbers of works in the corpus from the other continents and subcontinents and (b) to avoid a possible contamination of the other regions. In the collection of an author's region, either we selected the most prioritised region in both childhood and adulthood or, if there was no matching region, the most prioritised region in the adulthood as the main region. The control parameter for 'migrating' is a count of the different regions, thereby rudimentarily capturing authors' exposure to different environments during their lifetime. In addition, we selected the residence with the longest stay and categorised it as a 'village' when it does not have a university and less than 10,000 inhabitants, a 'town' with either one university or between 10,000 and 200,000 inhabitants or a 'city' with more than one university or more than 200,000 inhabitants as of 2019. For assigning these categories, we had to rely on modern data as it would have been too laborious to collect the historical data needed to cover all authors. We reflect on this limitation in the discussion.

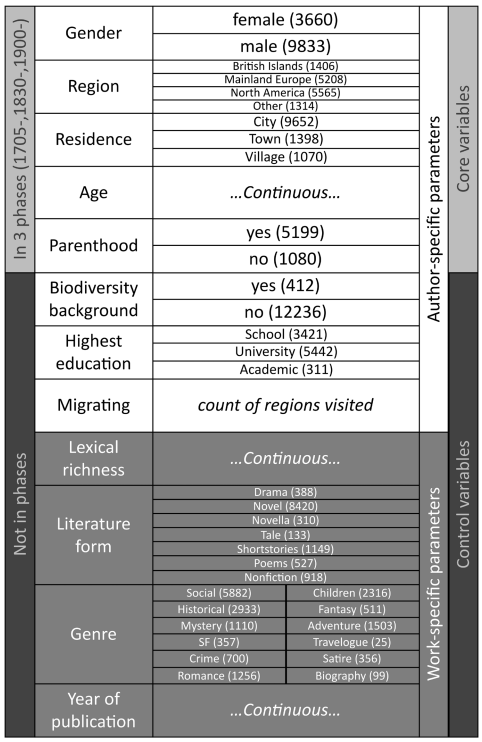

FIGURE 1 Overview of investigated predictors and their categories. All parameters were investigated over the complete time period, and the core variables were also investigated in dependence on the phases according to the work's publication date. The parameters are categorised as either author-specific or work-specific. In brackets, we present the number of records within that specific category.

We calculated the age of the author at the publication date as the difference between the year of publication and the year of birth. The categorical variable for parenthood is a binary flag based on the presence or absence of children. We classified the highest level of an author's education according to three categories: 'school', including autodidactic approaches, 'university' and a higher 'academic' degree, like doctor or professor. This is as simplified inclusion of education and mental aptitude as a predictor for biodiversity awareness and valuation. Finally, we included a binary flag for 'biodiversity background', which was set, when we could find a reference to either a direct occupation or education towards biodiversity-related fields as, for example biologists or an indirect or implied connection as, for example physician or teacher.

Work-related control variables included the lexical richness expressed in a work, which is the count of unique words (types/vocabulary) per size-normalised work, in order to make it directly comparable to the response variable, the biodiversity richness. We included the literature form (literary genre) with the possible categories: 'dramatic art', 'novel', 'tale', 'novella', 'short story', 'lyric art' and 'creative non-fiction' according to available categorisations in the source material. The genre as indicated in our source material is the only parameter, where multiple entries were possible. After some cleaning and conflation of close genres such as humour and satire or detective and crime fiction, each work could fall into the categories 'social fiction', 'children's literature', 'fantasy', 'adventure', 'science fiction', 'mystery', 'crime fiction', 'historical fiction', 'mystery', 'travelogue', 'biography' and 'satire'. Finally, we included the year of publication as a control variable that regards historically fixed events and processes not covered by the other predictor variables, such as wars, social trends or developments in media, technology and world view.

In several cases, some information about authors or their works could not be obtained during our metadata collection process. We marked the respective entries as NA (not available) so that they were not regarded in the analysis. The total proportion of NAs in the dataset is about 5%. The composition of the individual parameters is shown in Figure 2.

Under the influence of different epochs and historical processes, the predictors might diachronically exhibit different influences onto BiL that potentially outweigh the influence of the predictors as determined for the whole investigated period. In that case, the general trends observed have to be interpreted in a very different light. In order to gain insights into the time-dependence of the selected sensitivity parameters, the works in the corpus were categorised into three historical phases ('1', '2' and '3') depending on their publication date. Phase 1 is defined to contain works between 1705 and 1829, Phase 2 between 1830 and 1899 and Phase 3 between 1900 and 1969. These phases were defined to reflect periods characterised not by literary epochs, but by distinct levels of biodiversity richness according to Langer et al. (2021), yet also aim to optimise subcorpus size in order to produce a sufficiently large sample size for the RF analysis, which is why we extended the first phase relative to the two others. With these conditions, Phase 1 contains 659, Phase 2 contains 5476, and Phase 3 contains 7358 records. Subsequently, we analysed the sensitivity parameters generally, by performing a RF analysis on the complete corpus, as well as for the phases separately. In this way, we can show the diachronic robustness of our parameters, or else reveal a potential time-dependence of our parameters and are able to identify possible sensitivity trends.

2.3 | The RF analysis

RF is a machine learning algorithm (Breiman, 2001) with nonlinear fitting capabilities, allowing continuous and categorical predictor variables. Due to its ability to handle complex data, the minimised risk of overfitting and the small number of model parameters that have to be specified, RF is widely used in the scientific literature to solve complex regression and classification tasks (Antoniadis et al., 2021; Ao et al., 2019; Li et al., 2018; Richter et al., 2021). RF constructs an ensemble of randomised binary decision trees, where each tree models the relationship between the response and the predictor variables based on a set of simple decision rules (Breiman, 2001). These rules (variable thresholds) are defined for each node per tree allowing only for a small, randomly selected subset of the potential predictors. This procedure successively splits the data into an increasing number of homogeneous subgroups. The thresholds are defined by minimising a cost function (e.g. mean squared error), between the actual value of the response variable and the predicted value that would result from a specific split (Hutengs & Vohland, 2016). Each single tree is generated using the bagging method that randomly selects samples with replacement from the original data set (Breiman, 2001; Chan & Paelinckx, 2008). The remaining samples, approximately 37% of the total sample size, are referred to as OOB (out-of-bag). Although each resulting single tree might be somewhat overtrained, the randomness introduced into the training process encourages the trees to give independent estimates, which can be averaged to achieve accurate and robust results (Chen et al., 2019; Lindner et al., 2015), outperforming typical regression models like multiple linear regression and producing better receiver operating characteristic curves (Rodriguez-Galiano et al., 2015). To estimate the model accuracy, a tree predicts its OOB data unseen in the model training. Breiman (2001) argues that as the number of trees increases, the OOB errors always converge and overfitting is not a problem, because of the Strong Law of Large Numbers (Feller, 1968). To achieve accurate predictions (low OOB error) and robust estimates of the underlying relationships, we followed the suggestion of Naegeli de Torres et al. (2019) and set the number of trees to grow to a comparably large number of 2000 and kept the number of predictors to test for each split at the default value (p/3, with p being the total number of predictors), however, at the cost of a higher computational time. RF is natively able to handle nominal as well as incomplete data within a diverse dataset. Missing data (Figure 2) were imputed prior to the RF analysis by means of the median value of an individual predictor. For the analysis, we used the package randomForestSRC (Ishwaran & Kogalur, 2021)

PEOPLE AND NATURE | BRITISH ECOLOGICAL SOCIETY

FIGURE 2 Corpus composition showing the distribution of categories of individual parameters within the phases.

implemented in R (R Core Team, 2021), which is particularly suited to deal with patchy data, as it allows for imputation, and provides additional analytic methods, such as interaction analysis or determining the partial dependence with standard error.

In our analysis, we resort to a less used output of the algorithm. To further analyse how author- or work-related predictors affect the response variable, partial dependence was calculated using the partial.rfsrc function. Partial dependence gives a quantitative depiction of the dependence of a variable on the response as predicted by the trained RF (Friedman, 2001). It indicates the effect of each variable on the response variable after taking the average effect of all other variables in the model into account (Elith et al., 2008), thereby giving an estimate of the isolated correlation of the specific predictor and the response variable. The partial dependence function is defined as:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}\left(x, x_C^{(i)}\right),$$

where $x_C^{(i)}$ represents the value for all other variables other than x for individual i and $\hat{f}$ is the predicted value. We determined the partial dependence of biodiversity richness on each of the predictor variables in both the complete corpus and within the three phases.

To show that all selected predictors had a minimal covariance and a mostly independent impact on BiL, we started with a large set of approximately 50 predictors and gradually eliminated negligible and less meaningful predictors from the RF analysis after repeated iterations. With our final set of predictors and within the respective grown forest, we performed an interaction analysis in order to determine pairwise interactions using the find.interaction.rfsrc function. The resulting interaction matrix can be found in Appendix 2.

3 | RESULTS

We determined the individual sensitivity towards biodiversity as the partial dependency of biodiversity richness on each of the target and control variables in both the complete corpus and, in individual RF analyses, within the three historical phases calculating the resulting richness given as count of the taxon labels mentioned per normalised work, in short: 'labels', for each of the possible predictor values. Additionally, the percentage of variation explained by the RF model with the selected predictors is given.

3.1 | RF results for the complete corpus

According to our RF analysis, we could explain approximately 34% (R^2 in internal validation) of the variation within the biodiversity richness in literature with our predictor variables. The partial dependencies of the individual predictor variables (Figures 3 and 4) show a generally small interquartile range, as indicated by the whiskers of the boxplots, in comparison with the categorical variation of the respective variable, demonstrating the significance of these findings. Looking at the whiskers of any of the boxes in Figures 3–5, we found that the second and the third quartile lie within a biodiversity richness range of 0.25 labels, whereas medians of the individual levels of the respective categories may vary by more than 1.5 labels in Figures 3 and 4 and by more than 3.0 labels in Figure 5. We see an equally distinct range of variation of about 3 labels with an author's age as well as with the year of publication (Figure 6). We found a strong influence of lexical richness predicting values of biodiversity richness between 11 and 21 labels (Figure 7). Our analysis of interactions between the predictors (see Appendix 2) revealed no strong or otherwise conspicuous correlation. Moreover, we conducted an analysis of subcorpora for three phases according to a work's year of publication. Importantly, this confirmed the consistency of the influence of our core variables on BiL. In the following, we detail the results for selected predictor variables.

Our results for categorical variables show the isolated impact of core (Figure 3) and control (Figures 4 and 5) predictor variables (all other variables fixed at their mean) on biodiversity richness predicted by the RF regression model. Works of female authors (15.1) included 0.9 more labels than works of male authors (Figure 3, left). The results for the main region (Figure 3, left centre) show a biodiversity richness below 13.8 labels for the median for the British Isles, 14.7 labels for continental Europe and more

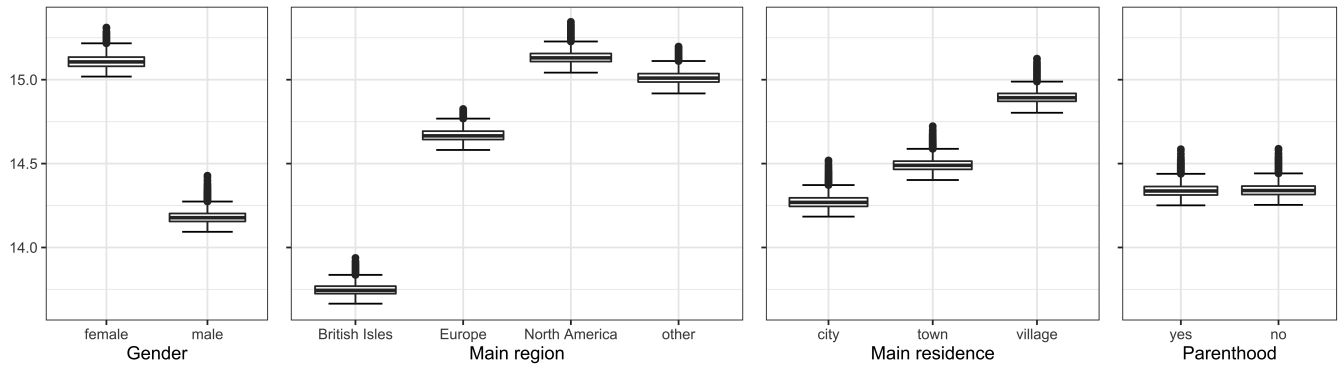

FIGURE 3 Isolated impact, as represented by the box plots of the corresponding partial dependency, of categorical core predictor variables is shown as a predicted resulting biodiversity richness against gender (left), main region (left centre), main residence (right centre) and presence of children (right) over the investigated period.

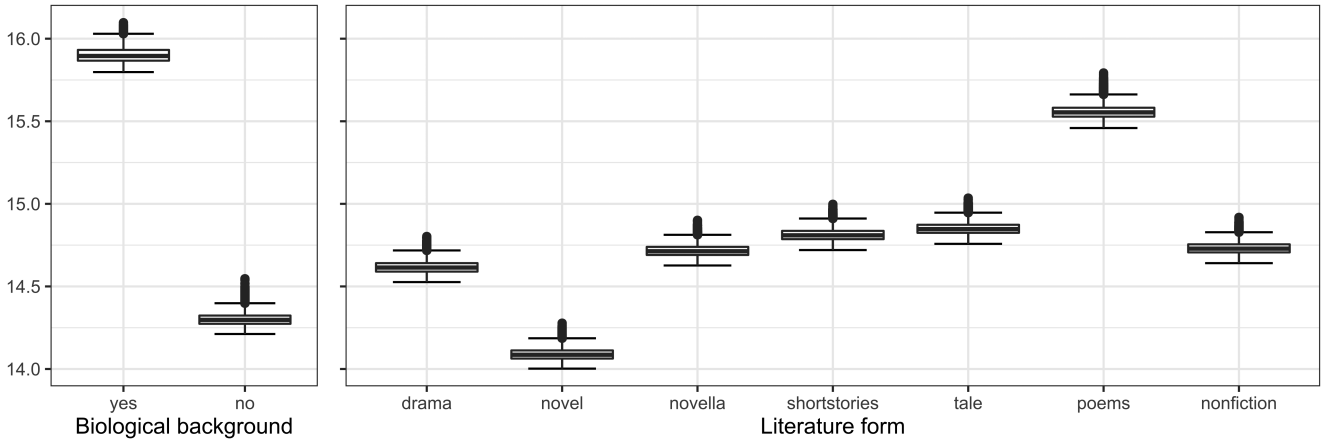

FIGURE 4 Isolated impact, as represented by the box plots of the corresponding partial dependency, of categorical control predictor variables is shown as a predicted resulting biodiversity richness against academic or occupational biological interest (left) and literature form (right) over the investigated period.

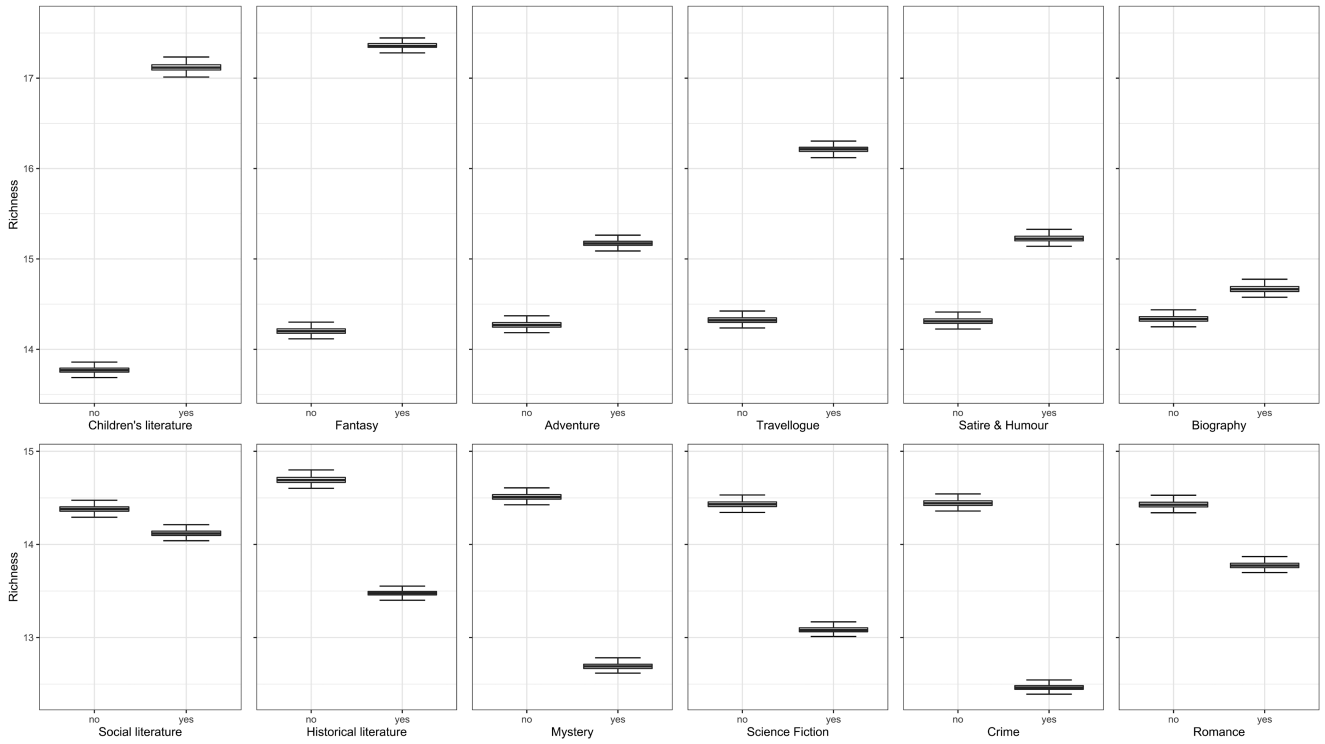

FIGURE 5 Isolated impact, as represented by the box plots of the corresponding partial dependency, of the predictor variable genre that is presented as binary variables for each category. Each genre is shown with a predicted biodiversity richness depending on the presence or absence of an association of work with the genre. The top row shows genres that yielded an increased biodiversity richness, whereas the bottom row shows the genres with a decreased biodiversity richness. Each row is sorted in descending order according to the number of works in this genre.

than 15.1 labels for Northern America. All other regions yielded comparable medians around 15.0 labels. Authors that live in an urban area (Figure 3, right centre) tend to the lowest biodiversity richness with a median below 14.3 labels, whereas authors from smaller towns or rural areas incorporated a slightly higher biodiversity richness of 14.5 or 14.9 labels, respectively. While there was a clear difference in biodiversity richness for authors with academic or occupational biological background (Figure 4, left) compared with authors without such a background, 15.9, respectively, 14.3 labels, we found no isolated impact of parenthood (Figure 3, right) on the resulting biodiversity richness. Concerning the literature form (Figure 4, right), we yielded a variation of about 1.5 labels from novels (14.1) to poems (15.6) and all other forms between 14.6 and 14.8 labels.

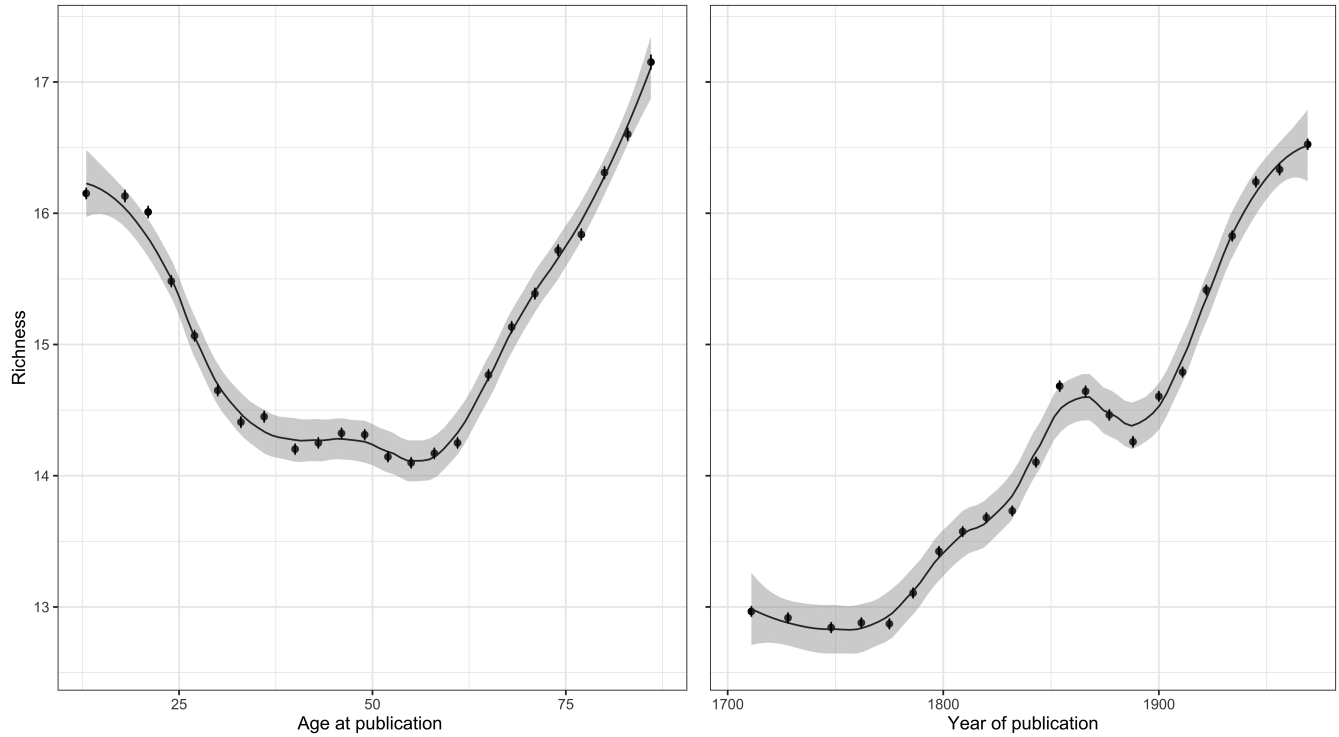

FIGURE 6 Partial dependencies of the time-related continual predictor variables, showing notable variation in age at publication (left) and year of publication (right), including a moving regression with confidence interval (99%). Whiskers of data points show the standard error of the partial dependency approximation by the random forest model.

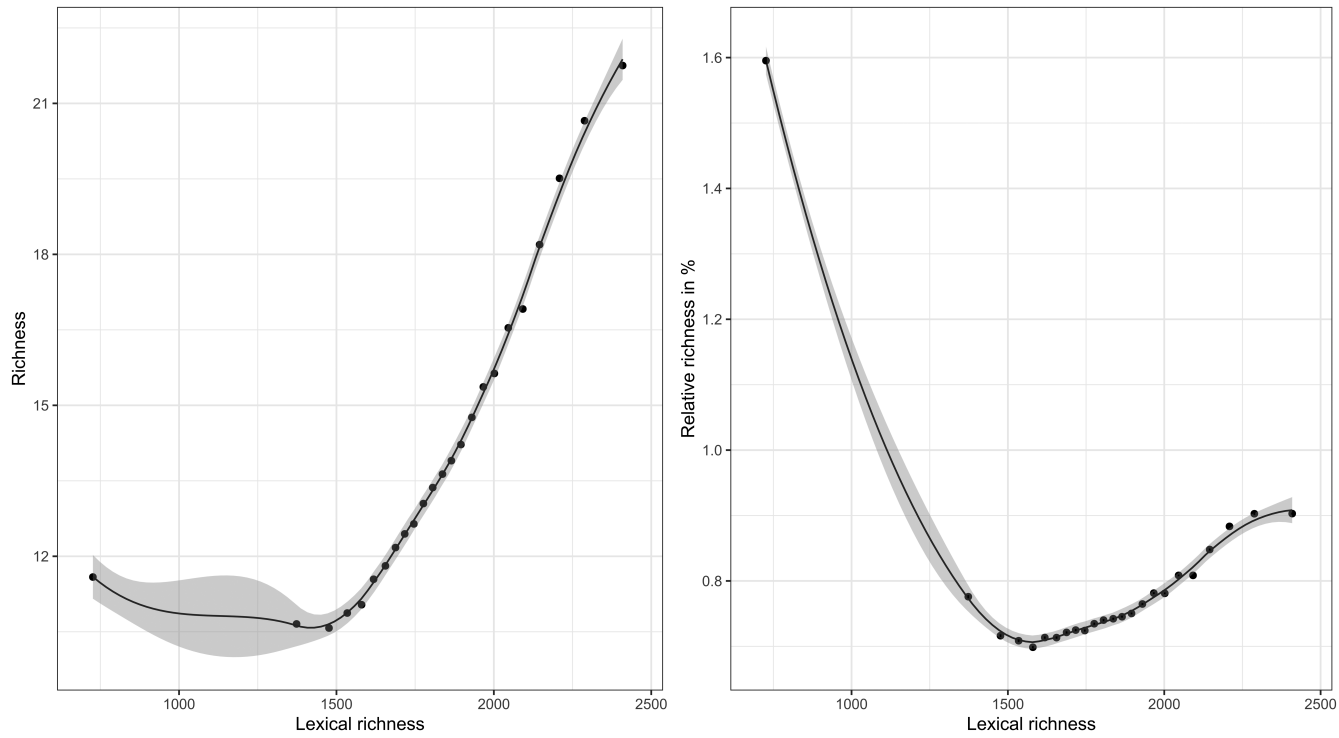

FIGURE 7 Partial dependencies of the vocabulary-related continuous predictor variable lexical richness in relation to the absolute (left) and relative (right) biodiversity richness, including a moving regression.

The predictor variable for the genre of a work is to be handled differently, as a work is potentially attributed to several genres simultaneously. Hence, we show our results separately in Figure 5, indicating the modelled biodiversity richness either for a present or absent attribution of works to a genre, making genre a binary variable for each category. We yielded the highest differences for

the genres children's literature and fantasy, where an attribution resulted in a biodiversity richness median being increased by more than 3 labels to over 17 labels. The genres adventure, humour and travelogue yielded an increase of 1 label, respectively, 2 for the latter. Also, biographies, though only yielding an increase of 0.4 labels, seem to be notably distinctive. We found the highest decrease for an attribution of the work to mystery and crime, where the median difference approaches two labels. For science fiction and historical fiction, the decrease is 1.3 and 1.2 labels, respectively. Romance literature is attributed to a decrease in biodiversity richness by 0.7 labels. Finally, an attribution to social fiction resulted in a distinctive decrease of 0.3 labels; however, we like to highlight that this is a complex genre with a number of subgenres.

In Figure 6, we present the results of the continuous predictor variables, namely the author's age at the time of publication (left) and the year of publication (right). The whiskers of the data points give an impression of the comparably small standard error between the partial dependencies approximated with the grown trees of the RF model. With a variation in biodiversity richness between 14.1 and 17.1 labels, the author's age has an influence on biodiversity richness comparable to our categorical variables. We found that especially works of authors younger than 25 and older than 70 years contain a high biodiversity richness (around 16 and 17 labels, respectively; Figure 6, left). Works of middle-aged authors between 35 and 60 years of age have a biodiversity richness below 14.5 labels.

The generally positive trend of biodiversity richness with the year of publication (Figure 6, right), increasing from about 13 to 16.5 labels within the investigated time window, can be attributed to time-related processes and events, which are not covered by the other predictor variables. We regard this as an essential control variable, since it aims to filter out the impact of general historical incidents and developments.

We observed a slightly decreasing trend of biodiversity richness for works with a lexical richness below 1500 unique words per 10,000 words (Figure 7, left). The correlation between lexical richness and biodiversity richness as per their partial dependency within the RF model was highly robust. As a result, the whiskers, depicting the standard error between all grown trees of the RF model, are smaller than can be made adequately visible. Note that this subcorpus largely contains children's literature, which was characterised by a high biodiversity richness (see Figure 5, top left). The steep increase of biodiversity richness (from about 11 to 21 labels) with lexical richness for works with a lexical richness >1500 unique words shows that with an increasing general vocabulary also the vocabulary for biodiversity increases. This strong influence highlights the importance of lexical richness as a control variable in order to isolate the individual impacts of the other predictor variables. For a closer look, we show the relative biodiversity richness in Figure 7 (right), revealing an increase of relative biodiversity richness from 0.7 to 0.9% with lexical richness above 1500 unique words per 10,000 words. The relative biodiversity richness for a lexical richness on the low end is with 1.6% about twice as high, most likely revealing the high preference for biodiversity in children's literature.

3.2 | RF results for the subcorpora of the three phases

The corpus was divided into three subcorpora according to their year of publication. Phases 1, 2 and 3 contain works published in the time windows 1705–1829, 1830–1899 and 1900–1969, respectively. The RF models for the individual phases yielded R^2s in internal validation of 48%, 36% and 32%, respectively. The core variables mostly showed no distinguishing features between the phases (see Appendix 3), exhibiting merely minor variation along the y-axis with approximately the same basic pattern. The few exceptions can be attributed to the small number of data points within the subcorpus, such as the small number of female authors in the already smaller subcorpus of the first phase. In conclusion, the results exhibit stability of the influence of predictors on BiL through time.

4 | DISCUSSION

For our analysis, we drew material and methods produced in recent advances in digital humanities, ecology and literary studies. This combination allowed us to detect relationships between biodiversity in literature (BiL) and individual characteristics of authors and their works. By applying a RF regression, we developed a model that was able to explain a significant fraction (34%) of the variation of BiL in dependence on selected predictor variables. From our model, we yielded distinctive results for the relevant partial dependencies, indicating personal characteristics of authors relating to the occurrence of biodiversity in their works. Apart from parenthood, all predictors (see Figure 1) exhibit a considerable impact on the size-normalised biodiversity richness that mostly falls in line with intuitive expectations regarding the degree of urbanisation, cultural differentiation and individual development, thereby confirming our hypotheses in almost all elements. Absolute differences may appear small, when comparing individual partial dependencies, but are highly significant, facilitated by our large sample size. Differences also add up by combination, meaning that, for example a low vocabulary crime novel written by a 40-year-old male author from 1750s central London may end up with half of the average biodiversity, statistically. With our diachronic assessment, we also recognised that sensitivity for biodiversity is subject to historical change on the grounds of societal and cultural development. With our interaction analysis, we also investigated the overlap of the predictors in their effect on BiL, confirming a typically low, reinforcing, or non-systematic correlation between predictors.

4.1 | Limitations to interpreting our results

Before we assess the validity of our results with respect to our desired inferences on the sensitivity towards biodiversity, we have to assess the limitations of our analysis and critically reflect on the current viewpoints on the relevant discourses in the involved scientific fields. Although we argue that the correlation of a variable with

biodiversity richness reflects a sensitivity for biodiversity, there are several debatable steps we perform in the logical chain of reaching this understanding. Our data in focus is the biodiversity richness within a size-normalised work, which we relate to the awareness for biodiversity of an author, that is assuming that authors with a higher awareness and knowledge of biodiversity also use more biodiversity terms in their works. In literary theory, researchers tend to the poststructuralist paradigm that it is not expedient to generalise interpretations of a text in terms of a supposed intention of its author (Barthes, 1967). We agree that the perceived meaning of a text correlates with the experiences of the reader and further the text may be reinterpreted in the light of alternative circumstances. However, authors equally are subject to their experiences, or more general their input received and processed, causing them to develop each in their idiosyncratic way and giving rise to a unique individual (Pol, 2006). Considering statistically meaningful numbers of people, we expect certain combinations of inputs to yield non-random results. In particular, we argue that people statistically are more aware of entities and topics, when exposed to them more often and, moreover, that they tend to work (e.g. by making analogies, communicating with or about it, correlating it with emotions like familiarity and creating art) with them more readily with increased awareness and hence exposure. The input received determines the moral compass and shapes what a person deems important or pleasant. In this way, each person exhibits a unique behaviour and each author a unique writing style (Jannidis et al., 1999). Hence, while an intention of a text may not be successfully retrieved from it and would potentially be opposed by numerous reinterpretations, we can statistically correlate selected text measures to authors' living circumstances (Klaussner et al., 2015). This goes along with the mentioned concept of situated knowledge widely accepted in the environmental humanities. Hence, with our data we do not aim to reconstruct the intention of an author, but the situation which frames the production of this text. With this concept of 'situated literature', we can trace the correlations between specific characteristics of a text and specific elements of the situation in which this text was conceived. For our particular study, we paraphrase that, looking at a large number of individuals, authors appear to be sensitive for biodiversity based on a number of parameters, which we aimed to cover in our investigation. Naturally, such a correlation cannot produce meaningful insights or would not arise when dealing with small numbers of authors or even individuals.

We are aware that the clientele of an author does not represent an average or ordinary citizen, which is to be considered when interpreting our results by accounting for the deviation of the class structure of our corpus from the real world. However, when taking into account that the proportion of a middle class is much larger now, than in the time of industrialisation (Hudson, 2015), this deviation now can be expected to be much smaller than back then. As a result, the mentioned bias can be regarded as comparatively small, when considering the validity of the determined sensitivity parameters in present times.

Although the applied algorithm typically is used for creating predictive models, we (1) do not claim to be able to predict future trends with our results, as they merely represent the statistics of a historical correlation, and (2) cannot impose a specific relationship, such as a polynomial or exponential, between any of the sensitivity parameters and a measure for BiL.

4.2 | Sensitivity parameters

In our investigation, we focussed on the analysis and discussion of variables that can be subject to political, ethical, urbanistic and economical influences. We expected our core variables to exhibit a changing influence over time, which is why we aimed to disentangle time-related effects from basic effects of our core variables. Hence, we also assessed those diachronically in order to gather insights to what extent sensitivity for biodiversity is subject to change with societal and cultural processes. We decided on three phases to be investigated, which we defined on the grounds of chronological and corpus-compositional parameters. However, our diachronic analysis indicated a comparably consistent effect of the core variables on BiL. In the following, we detail our findings and reflect on possible causes that lead to the observed correlation.

4.2.1 | Gender

We observed a higher BiL for female authors than for male. Gender is an especially complex parameter as it contains all social differences correlated to gender, which, particularly in the past, are substantial in amount and extent (Becker & Kortendiek, 2010). Gender studies recognise that in the investigated timespan, women were subordinate to men, for example as per their political and social standing or by the traditional division of labour (Bauhardt, 2010). Social scientists agree that less privileged individuals automatically suffer more from environmental changes as they cannot be sufficiently mitigated with their deficient social and economic potency. Hence, women in our time period might have experienced a closer relationship with nature than men (Weller, 2019). An author can be regarded as having been more privileged than the average citizen, which is why the environmental impact is likely to be reduced for female authors; nevertheless, it is probable that they were the recipients of a transfer of responsibility and identity of working-class women onto women in general. Another position that emerged in the gender studies is that women, as they potentially can bear children, are more closely attached to natural processes, potentially increasing sensibility for nature (Bauhardt, 2010). Historically, the genders have also been attributed to other binary relationships, giving rise to the notion that men are closer to, for example culture or mind, whereas women were assumed to be closer to nature or flesh, respectively (Gaard, 1993; Holland-Cunz, 2014; Weller, 2019). We see ourselves not in a position to argue for a specific position as a main driver for the different BiL. However, it becomes clear that both genders had distinguished society-imposed perceptions of the genders and their roles in culture and nature that inevitably lead to a difference in attitude and awareness and accordingly to a statistically meaningful difference in BiL.

4.2.2 | Main region

To consider the main region of an author encompasses two major time-related features relevant in our investigation, which is the progression of the industrialisation within a region as well as the sociocultural development. While the latter is difficult to present in numbers, for the former we convey an impression of the key stages of the region's development relevant to our arguments via exemplary events below. Bear in mind that we had to conflate a number of culturally diverse peoples into each region, which means that there may be additional cultural effects on BiL that can only be entangled with a less Western-dominated distribution within a more diverse corpus. We observed a higher BiL for North America over Mainland Europe and for Mainland Europe over the British Isles. Each region experienced fundamentally different developments (see Table 1 for manufacturing output as an example) as a result of different natural conditions, historical premises, social processes and influential extraordinary individuals.

With our results, we confirm our hypothesis that authors from regions with delayed phases of exploration and settlement exhibit a higher BiL. Please note that with our method we cannot ascertain this correlation to be causal, but consider it as highly probable. We argue that especially the status of industrialisation including urbanisation and land use change (Reeder & Rodger, 2001) can be regarded as proxies for the disconnection from nature and constitute the most probable causes for a lower BiL. In particular, the British Isles were the forerunners of industrialisation (Perkin, 1969), inventing most of the key technology, starting off with machines promoting clothes' production and operating their first steam-driven spinning machine by 1764 and exhibiting the lowest BiL. Mainland Europe consists of several early industrialised countries like Germany (first spinning machine in 1784, followed by machine-assisted drainage of the shafts of a copper shale mine in 1985) and France (borrowed first machines from the British in the 1780s), but on average the industrialisation was delayed in comparison to the British Isles. North America was still in a phase of exploration and settlement, for example populating California during the goldrush 1848–1855, during the 19th century (Mountjoy, 2009). Whereas industrialisation in some states on the east coast was only slightly delayed (the first cotton separation machine started 1793) with respect to European countries, on average, regarding the central and western parts, there was still a strong exposure to nature, following the comparably late discovery of the Rocky Mountains by 1754 and the west coast by 1793, while European industrialisation was already ongoing. Accordingly, bearing the promising new lands as well as the challenging, threatening wilderness, for example during and in the wake of migration treks such as the migration along the westward expansion trails setting off only in 1829, as the mental poles of this exposure in mind, authors from North America exhibited a notably higher BiL. Nevertheless, differences in BiL do not only reflect the state of industrialisation and closeness to nature, but BiL, as correlated to people's attitude, is also subject to polity, economy and cultural trends (Kühne & Bruns, 2015; cf. French vs. English horticulture), the influence of which may not be entangled easily from each other and requires further investigation.

4.2.3 | Main residence

We observed a higher BiL with decreasing size of the author's settlement of residence. Whereas the main region is an indirect correlation of space with BiL, as it does not account for author-specific local conditions, the main residence represents a direct correlation of an author's geographic environment with BiL. With our results, we confirm our hypothesis that authors living in a rural environment produce a higher BiL than those living in an urban setting. This observation is in line with the intuitive as well as researched supposition that people in an environment closer to nature have a stronger tendency to regard living beings as key players in their lives (Fischer et al., 2015; Mehring et al., 2017). As a result, these people are more likely to resort to living beings when it comes to, for example comparisons, threats, adventure, daily business or the notion of freedom. In contrast, cities teeming with technical inventions, growing factories, economic strife and social challenges appear to be obvious sources for topics with an inherently low degree of biodiversity other than a limited number of domesticated animals and plants. Hence, we argue that the geographical exposure to nature is a pivotal factor for BiL as supported by our results. The categorisation of settlements as city, town or village was based on concurrent population sizes and central institutions upon the assumption that the statistical relation to a settlement size remains significant over three centuries. As a result, analysing individual authors on the grounds of this data can be misleading.

4.2.4 | Age

We observed a high BiL for authors below the age of 25 and above the age of 70, and the lowest values are correlated with the age between 35 and 60. Naturally, age as a number has no particular meaning, but is a proxy for (1) distinctive phases in somebody's life that can be associated with particular inputs (McMullen & Smiles, 2016)

TABLE 1 Development of economic power of the relevant regions in comparison by the relative share of the world manufacturing output (Kennedy, 1987).

	1750	1800	1830	1860	1880	1900
United Kingdom	1.9	4.3	9.5	19.9	22.9	18.5
Mainland Europe	21.3	23.8	24.7	33.3	38.4	43.5
USA	0.1	0.8	2.4	7.2	14.7	23.6
Rest of the world	76.7	71.1	63.4	39.6	24	14.4

as well as (2) stages of ego development (Loevinger, 1966) accounting for individual perception filters and attitudes. Our results show notable differences in BiL in relation to age and hence confirm our hypothesis that age is a relevant statistical predictor for BiL. We suppose that young authors exhibit a higher BiL, since they, in their time of childhood and adolescence, (1) tended to explore the vicinity, finding places and discovering details an adult would pass by unnoticing, (2) might have projected their lives upon purposeful life goals with adventurous travels and (3) were generally less privileged in comparison to adults, which entailed a slightly stronger dependence on natural conditions (Weller, 2019). We suppose that middle-aged authors exhibit a lower BiL, as they are more likely to be in a life phase that involves the sustenance and challenges concerning family members, work or social as well as economic structure. Especially by the changes in industrialisation, these topics are inherently poor of biological references and concerns (Brown & Harrison, 1978). We suppose that elderly authors exhibit a higher BiL, as (1) they returned to nature, seeking solutions for unsolved problems, for example by comparing the current affairs to the more natural old days, (2) generally incorporate a higher level of knowledge and education into their literature and (3) are stronger dependent on natural conditions once again. All these correlations still are subject to cultural differences, which is why also age cannot be used to predicting an individual production of BiL.

4.2.5 | Parenthood

We observed no notable differences in BiL in correlation with the presence or absence of children. This contradicts our hypothesis that parenthood is an important parameter of sensitivity towards biodiversity. We assume that there are several factors that might produce individual impacts, but fail to produce a coherent signal. One expectation was that people with children are more involved in their play with basal natural objects and also educate them to some minimal extent with respect to nature (Weller, 2019). However, a large-scale attitude towards a child-near education might be a phenomenon coming up only in the later 20th century, which is especially true for men as a significant fraction of authors of our corpus. As an opposing driver, we expect childless adults to be more spatially flexible spatially and to be more venturesome, thereby potentially gathering more nature-close experiences. Generally, parenthood changes people's attitudes towards life in unpredictive ways, leading in total to an obscured signal for this parameter. We suggest further investigations taking other factors and combinations such as early or late parenthood into consideration.

4.2.6 | Control variables

In order to disentangle the signals of our core variables from other influences, we included a number of control variables, which we generally observed to have a notable correlation to BiL. This suggests that without accounting for these covariates, their signals would have been transferred onto our core variables thereby overriding their intrinsic signals. Given the overall strong signals of the control variables, we may assume that this approach greatly improved the accuracy of our model. With our controls, we also provided further insights into several variables relevant in other scientific fields and questions. We controlled for the bias of a biological background by training or occupation, for the level of education and for the publication date of the work as a signal of general historical events and processes not included in the individual's or work's variables, for all of which we confirmed the hypothesis to be a relevant signal. We regard the rise of BiL with the isolated predictor variable of the publication date as a sign of biodiversity-promoting historical events and processes, possibly as a result of increasing quality and extent of education, life sciences and information transfer. In this light, it is even more noteworthy that BiL declined over time during the industrialisation (Langer et al., 2021), and from this contrast, we can deduce that our selection of the other predictor variables indeed covers and explains the parameters counteracting the rise of BiL by historical processes. We controlled for the migratory nature of some authors, but the results were not sufficiently conclusive to make sensible interpretations.

While we regard work-specific parameters as control variables in our study, we nevertheless can assume a correlation between the attitudes of persons and their predisposition towards certain kinds of literature, or more generally topics and communication forms (Haraway, 1988). This means that there is an indirect, less tangible connection between authors and the parameters concerning their works. Looking at the partial dependencies of these parameters, we confirmed our hypothesis that the literature form and the genres of a work have a notable influence on BiL. Both correlations should be investigated further; however, we like to present a first assumption in their regard.

The diverse literature forms have fundamentally different structures and literary functions and thus account for diverse themes and subjects, biodiversity amongst them, in different fashions. Poems, with the tendency towards the shortest literature form within our corpus, contain densely packed interpretations and information, as whole worlds and world views have to be evoked with sparse usage of words. To this end, poems are drawing from rhetorical devices, such as comparisons and metaphors, which potentially transfer meaning in an abbreviated way, and that make use of complex systems and processes, to which purpose nature, particularly life appears to be unequally well suited. With biological entities and processes, people created historically grown and nurtured pointers towards typically unrelated topics, such as character traits, actions or locations. Hence, we expect shorter works to contain a higher concentration of BiL than longer works, such as novels.

The genre of a literary text tends to play a vital role for setting and theme, predefining to some extent the relevance of living nature for the work. Social fiction, for instance, might concentrate statistically more on social injustice with the result of the low observed

BiL, fantasy or adventure might invoke biodiverse nature-near experiences, and science fiction on the other hand focusses on technological challenges and inventions rather than biodiversity. Counterintuitively, romances and historical fiction do not make use of a higher BiL, apparently drawing the reader to the more psychological, respectively social issues, but humorous literature in turn seems to resort to animals and plants more than average, most probably by means of comparisons and verbal caricatures. Furthermore, we correlate the high BiL in children's literature to (1) the purpose of education, by itself, but also by the demand to learn nature's benefits and threats, (2) the stimulation of children with anthropomorphic entities, mostly animals, (3) the placement of the works subject in a world familiar to children with their curious explorative nature and in their comparatively nature-near daily lives and (4) the generally nonlinear development of vocabulary across different fields of world and life, promoting vocabulary in fields most relevant to the current life situation, which includes living nature for children.

Finally, we confirmed our hypothesis that a higher lexical richness correlates with a higher biodiversity richness. We argue that authors ascribe some minimal relevance to the biosphere, which is why the richness of the general vocabulary used in their works also affects the used biodiversity vocabulary. With our results of the relative biodiversity vocabulary, we have an indication that biodiversity becomes even more relevant with a higher general vocabulary. However, we also observed a higher BiL in work with a lexical richness below 1000. This we attribute to a change in paradigm, as works with low lexical richness almost exclusively contains children's literature, which, to promote understanding in younger children and toddlers, draw from a simpler vocabulary, while retaining a high biodiversity content as discussed above.

In general, biodiversity minima in our results are always notably above zero, demonstrating that a minimum of biodiversity seems to be reasonable or even necessary for an effective communication. This inference is supported by the observation that the development of a higher vocabulary over-proportionally results in a higher biodiversity richness.

4.3 | Relevance of the study

Our assessment confirms that sensitivity for biodiversity, as an expression of the relation between authors in their social and geographical situation and living nature in literary works, is subject to change with time and hence the occurred societal and cultural processes, supposing that they can be targeted by respective measures. Determining the relationship between the proposed predictors and BiL enables us to argue for a stronger consideration of these factors, as part of nature's immaterial contributions (Díaz et al., 2015; Pascual et al., 2017), in decision-making. Depending on the predictor and the desired outcome decision-making can be amended in multiple areas, for example polity, urban development, promoting family orientation, by regarding these sensitivity parameters as potential entry or pivot points to biodiversity education. The dependence on decision-making processes makes the parameters especially relevant for reaching policy targets related to increasing biodiversity awareness as demanded in the so far underachieved Aichi Target 1 (Convention on Biological Diversity, 2011).

First studies in culturomics (Michel et al., 2011) showed that cultural change can be quantified, particularly with cultural products in the form of written text. From the viewpoint of the humanities, our study emphasises the nature–culture entanglement by making statistical relationships between several parameters of the real world and biological entities in the creation of art visible (Langer et al., 2022). Thereby, we reinforce the necessity of an advocate for nature (Bühler, 2016), not only for reasons of material resources but also in terms such as aesthetics, valuation and education. Instating ecocriticism, or a similar paradigm, as such an advocate ensures the consideration of less tangible facets of nature and promotes the reappraisal of its relationship with culture via an array of interdisciplinary methods and may discuss the preferences of nature as a counterpart to humanity's wants.

4.4 | Methodological considerations in resource creation

Further limits to interpretations of our results arise by the process of parameterising circumstances of life. These are inherently complex, intertwined, underlie temporal variation and are not always available for every individual author. Nevertheless, we had to reduce them into few categories, such as the three categories for the residence of an author, for them to be processible efficiently. This simplification imposes limits on the explanatory power of our investigation, and it probably contributed to some reduction in the predictability of our model to 34%.

Automatically retrieving metadata concerning the living circumstances of authors as well as the classification of works, we were presented with incomplete metadata. To our knowledge, there was no other adequately applicable metadata collection, which is why we opted for manually closing most gaps in our metadata, while being somewhat limited by missing freely accessible data. Since a RF regression can be adjusted to deal well with incomplete data and because we were able to reach a generally high level of completion, we are comfortable to assume that our final metadata is well suited to the undergone analysis. Moreover, our metadata collection can be used as a valuable resource in similar investigations utilising either the Gutenberg Project or overlapping corpora.

5 | CONCLUSION AND PROSPECT

For our analysis, we drew material and methods produced in recent advances in digital humanities, ecology and literature studies to provide further elements for consideration, debate and reflection as evolving in discourses of environmental humanities, ecocriticism, environmental protection and sustainable development, as well as respective

decision-making. We assessed the validity of several intuitive hypotheses concerning the correlation of BiL with the circumstances of authors' lives, arguing for several variables to be included in quantifying sensitivity towards biodiversity. The correlation of the selected parameters and BiL was discussed regarding the most probable historical factors and processes, also suggesting a negative correlation of the ongoing industrialisation with distancing from nature and with BiL.

We highlight that our core variables appear robust through time. In particular, they should be considered in biodiversity education by targeting individual sub-populations specifically with measures counteracting limiting factors. Since gender differences are largely shaped by society, this parameter changes with society and can therefore be regarded in regulations, awarding programmes or similar measures by decision-makers, for example, by accounting for the gender gap. The correlation between region and residence and BiL is an indicator for a direct dependence of attitude towards nature on actual surroundings and may be used by decision-makers, such as economists navigating economic pursuits and urbanists shaping the living environments of the future and to approach desired goals, such as Aichi Target 1. Similarly, age-dependent campaigns or educational, cultural and activity offers can be adapted for optimising the promotion of biodiversity awareness. The sensitivity parameters found could be fine-tuned and brought up to date investigating contemporary text sources, such as newspapers, blogs or surveys.

For further analyses to promote understanding of the relationship between nature and culture, especially situated biological knowledge, several methods might shed light onto individual facets. We regard it as relevant to target more specific relationships by, for example, assessing the sentiment and frequency differences between wild and domestic animals and plants, between local and exotic species and between different taxonomic units. We also see potential in correlating Latent Dirichlet Allocation modelled topics to the taxa occurring within these topics providing information about the closeness of nature to individual topics.

AUTHOR CONTRIBUTIONS

Christian Wirth and Lars Langer proposed the study. Lars Langer, Christian Wirth and Manuel Burghardt conceptualised the study. Lars Langer and Ronny Richter developed the methodology. Lars Langer gathered and processed the data and conducted the analysis. Lars Langer, Christian Wirth, Manuel Burghardt, Roland Borgards and Ronny Richter interpreted the results. Lars Langer managed the project and wrote the manuscript. Lars Langer, Christian Wirth, Manuel Burghardt, Roland Borgards and Ronny Richter contributed substantially to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

We thank N. Kowalewski and F. Küchler, who helped to collect comprehensive amounts of metadata. We also thank the people in the Department of Systematic Botany and Functional Biodiversity at Leipzig University for their constant support and feedback throughout the project. This research was supported by the Helmholtz Centre for Environmental Research, the German Center for integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig (DFG FZT 118) and Leipzig University. We acknowledge support from Leipzig University for Open Access Publishing. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The Standardized Project Gutenberg Corpus used in this study is available in the repository pgcorpus/Gutenberg at GitHub (https://github.com/pgcorpus/gutenberg) or at Zenodo (https://zenodo.org/record/2422561) as a permanent record (Gerlach & Font-Clos, 2018). The metadata collected in order to enable the random forest analysis is available as Dataverse dataset (https://doi.org/10.7910/DVN/7POSNW).

ORCID

Lars Langer https://orcid.org/0000-0002-1076-2936
Manuel Burghardt https://orcid.org/0000-0003-1354-9089
Ronny Richter https://orcid.org/0000-0002-8728-7918
Christian Wirth https://orcid.org/0000-0003-2604-8056

REFERENCES

Ainscough, J., de Vries Lentsch, A., Metzger, M., Rounsevell, M., Schröter, M., Delbaere, B., de Groot, R., & Staes, J. (2019). Navigating pluralism: Understanding perceptions of the ecosystem services concept. *Ecosystem Services*, *36*, 100892. https://doi.org/10.1016/j.ecoser.2019.01.004

Antoniadis, A., Lambert-Lacroix, S., & Poggi, J.-M. (2021). Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, *206*, 107312. https://doi.org/10.1016/j.ress.2020.107312

Ao, Y., Li, H., Zhu, L., Ali, S., & Yang, Z. (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering*, *174*, 776–789. https://doi.org/10.1016/j.petrol.2018.11.067

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press.

Barthes, R. (1967). *The death of the author*. Aspen.

Bauhardt, C. (2010). Ökologiekritik: Das Mensch-Natur-Verhältnis aus der Geschlechterperspektive. In R. Becker & B. Kortendiek (Eds.), *Handbuch Frauen- und Geschlechterforschung* (pp. 322–327). VS Verlag für Sozialwissenschaften.

Becker, R., & Kortendiek, B. (Eds.). (2010). *Handbuch Frauen- und Geschlechterforschung*. https://doi.org/10.1007/978-3-531-92041-2

Bennett, J. (2010). *Vibrant matter—A political ecology of things*. https://doi.org/10.1215/9780822391623

Breiman, L. (2001). *Random forests—Random features* (Vol. 5). In Technical report 547. https://statistics.berkeley.edu/tech-reports/567

Brown, D., & Harrison, M. J. (1978). *A sociology of industrialisation: An introduction*. https://doi.org/10.1007/978-1-349-15924-6

Buell, L. (1996). *The environmental imagination—Thoreau, nature writing, and the formation of American culture*. Belknap Press.

Bühler, B. (2016). *Ecocriticism*. https://doi.org/10.1007/978-3-476-05489-0

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., … Naeem, S. (2012). Biodiversity loss and its impact on

humanity. *Nature*, *486*(7401), 59–67. https://doi.org/10.1038/nature11148

Celis-Diez, J. L., Díaz-Forestier, J., Márquez-García, M., Lazzarino, S., Rozzi, R., & Armesto, J. J. (2016). Biodiversity knowledge loss in children's books and textbooks. *Frontiers in Ecology and the Environment*, *14*(8), 408–410. https://doi.org/10.1002/fee.1324

Chan, J. C.-W., & Paelinckx, D. (2008). Evaluation of random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*(6), 2999–3011. https://doi.org/10.1016/j.rse.2008.02.011

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U. A., Katsouyanni, K., Janssen, N. A. H., Martin, R. V., Samoli, E., Schwartz, P. E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., … Hoek, G. (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment International*, *130*, 104934. https://doi.org/10.1016/j.envint.2019.104934

Convention on Biological Diversity. (2011). *Strategic plan for biodiversity 2011–2020, including aichi biodiversity targets*. https://www.cbd.int/sp/

Deuber-Mankowsky, A., & Holzhey, C. F. E. (2013). *Situiertes Wissen und regionale Epistemologie: Zur Aktualität Georges Canguilhems und Donna J. Haraways*. Turia + Kant.

Díaz, S., Demissew, S., Carabias, J., Joly, C., Lonsdale, M., Ash, N., … Zlatanova, D. (2015). The IPBES conceptual framework — Connecting nature and people. *Current Opinion in Environmental Sustainability*, *14*, 1–16. https://doi.org/10.1016/j.cosust.2014.11.002

Díaz, S., Pascual, U., Stenseke, M., Martín-López, B., Watson, R. T., Molnár, Z., … Shirayama, Y. (2018). Assessing nature's contributions to people. *Science*, *359*(6373), 270–272. https://doi.org/10.1126/science.aap8826

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Emmett, R. S., & Nye, D. E. (2017). *The environmental humanities: A critical introduction*. The MIT Press.

Feller, W. (1968). *An introduction to probability theory and its applications* (Vol. 1). Wiley.

Fischer, J., Gardner, T. A., Bennett, E. M., Balvanera, P., Biggs, R., Carpenter, S., … Tenhunen, J. (2015). Advancing sustainability through mainstreaming a social–ecological systems perspective. *Current Opinion in Environmental Sustainability*, *14*, 144–149. https://doi.org/10.1016/j.cosust.2015.06.002

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gaard, G. (Ed.). (1993). *Ecofeminism: Women, animals, Nature*. Temple University Press.

Gerlach, M., & Font-Clos, F. (2018). *Standardized Project Gutenberg corpus*. https://doi.org/10.5281/zenodo.2422561

Glotfelty, C., & Fromm, H. (Eds.). (1996). *The ecocriticism reader—Landmarks in literary ecology*. University of Georgia Press.

Haraway, D. J. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575. https://doi.org/10.2307/3178066

Haraway, D. J. (2016). Staying with the trouble. In *Staying with the trouble*. https://doi.org/10.1215/9780822373780

Haraway, D. J., & Begelke, M. (2003). *The companion species manifesto: Dogs, people, and significant otherness*. Prickly Paradigm Press.

Hinton, P. (2014). 'Situated knowledges' and new materialism(s): Rethinking a politics of location. *Women: A Cultural Review*, *25*(1), 99–113. https://doi.org/10.1080/09574042.2014.901104

Holland-Cunz, B. (2014). *Die Natur der Neuzeit—Eine feministische Einführung*. Verlag Barbara Budrich.

Hudson, N. (2015). Literature and social class in the eighteenth century. In *Oxford handbooks online*. https://doi.org/10.1093/oxfordhb/9780199935338.013.007

Hutengs, C., & Vohland, M. (2016). Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, *178*, 127–141. https://doi.org/10.1016/j.rse.2016.03.006

IPBES. (2019). *Summary for policymakers of the global assessment report on biodiversity and ecosystem services*. (S. Díaz, J. Settele, E. S. Brondízio, H. T. Ngo, M. Guèze, J. Agard, A. Arneth, P. Balvanera, K. A. Brauman, S. H. M. Butchart, K. M. A. Chan, L. A. Garibaldi, K. Ichii, J. Liu, S. M. Subramanian, G. F. Midgley, P. Miloslavich, Z. Molnár, D. Obura, … C. N. Zayas, Eds.). https://doi.org/10.5281/zenodo.3553579

Ishwaran, H., & Kogalur, U. (2021). *Fast unified random forests for survival, regression, and classification (RF-SRC)*. https://cran.r-project.org/package=randomForestSRC

Jannidis, F., Lauer, G., Martínez, M., & Winko, S. (Eds.). (1999). *Rückkehr des Autors—Zur Erneuerung eines umstrittenen Begriffs*. Niemeyer.

Kennedy, P. (1987). *The rise and fall of the great powers*. Random House.

Klaussner, C., Nerbonne, J., & Çöltekin, Ç. (2015). Finding characteristic features in stylometric analysis. *Digital Scholarship in the Humanities*, *30*(Suppl. 1), i114–i129. https://doi.org/10.1093/llc/fqv048

Kühne, O., & Bruns, D. (2015). Gesellschaftliche Transformation und die Entwicklung von Landschaft. In *Transformation und landschaft* (pp. 17–34). https://doi.org/10.1007/978-3-658-00605-1_2

Langer, L., Burghardt, M., Borgards, R., Böhning-Gaese, K., Seppelt, R., & Wirth, C. (2021). The rise and fall of biodiversity in literature: A comprehensive quantification of historical changes in the use of vernacular labels for biological taxa in Western creative literature. *People and Nature*., *3*, 1093–1109.

Langer, L., Burghardt, M., Borgards, R., Köhring, E., & Wirth, C. (2022). Digital environmental humanities—Zum potential von "computational and literary BiodiversityStudies" (CoLiBiS). *Proceedings of the 8th annual conference of the DHd (digital humanities Im Deutschsprachigen Raum) association* (pp. 82–86). https://doi.org/10.5281/zenodo.6304590

Laslett, B., & Brenner, J. (1989). Gender and social reproduction: Historical perspectives. *Annual Review of Sociology*, *15*, 381–404. http://www.jstor.org/stable/2083231

Latour, B. (1999). *Pandora's hope—Essays on the reality of science studies*. Harvard University Press.

Lautenbach, S., Mupepele, A.-C., Dormann, C. F., Lee, H., Schmidt, S., Scholte, S. S. K., … Volk, M. (2019). Blind spots in ecosystem services research and challenges for implementation. *Regional Environmental Change*, *19*(8), 2151–2172. https://doi.org/10.1007/s10113-018-1457-9

Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J. C.-W., van den Bossche, P., … Omar, N. (2018). Random forest regression for online capacity estimation of lithium-ion batteries. *Applied Energy*, *232*, 197–210. https://doi.org/10.1016/j.apenergy.2018.09.182

Lindner, C., Bromiley, P. A., Ionita, M. C., & Cootes, T. F. (2015). Robust and accurate shape model matching using random forest regression-voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1862–1874. https://doi.org/10.1109/TPAMI.2014.2382106

Loevinger, J. (1966). The meaning and measurement of ego development. *American Psychologist*, *21*(3), 195–206. https://doi.org/10.1037/h0023376

McMullen, G., & Smiles, S. (2016). *Late style and its discontents: Essays in art, literature, and music*. Oxford University Press.

Mehring, M., Bernard, B., Hummel, D., Liehr, S., & Lux, A. (2017). Halting biodiversity loss: How social–ecological biodiversity research makes a difference. *International Journal of Biodiversity Science, Ecosystem Services & Management*, *13*(1), 172–180. https://doi.org/10.1080/21513732.2017.1289246

Mesoudi, A. (2011). *Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences*. University of Chicago Press.

Methorst, J., Bonn, A., Marselle, M., Böhning-Gaese, K., & Rehdanz, K. (2021). Species richness is positively related to mental health—A study for Germany. *Landscape and Urban Planning*, *211*, 104084. https://doi.org/10.1016/j.landurbplan.2021.104084

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Millennium Ecosystem Assessment. (2005). *Ecosystems and human well-being: Synthesis* (W. V. Reid, H. A. Mooney, A. Cropper, D. Capistrano, S. R. Carpenter, K. Chopra, P. Dasgupta, T. Dietz, A. K. Duraiappah, R. Hassan, R. Kasperson, R. Leemans, R. M. May, A. J. McMichael, P. Pingali, C. Samper, R. Scholes, R. T. Watson, A. H. Zakri, … M. B. Zurek, Eds.). Island Press.

Montana Hoyos, C., & Fiorentino, C. (2016). Bio-utilization, bio-inspiration, and bio-affiliation in design for sustainability: Biotechnology, biomimicry, and biophilic design. *The International Journal of Designed Objects*, *10*(3), 1–18. https://doi.org/10.18848/2325-1379/CGP/v10i03/1-18

Mountjoy, S. (2009). *Manifest destiny: Westward expansion*. Infobase Publishing.

Naegeli de Torres, F., Richter, R., & Vohland, M. (2019). A multisensoral approach for high-resolution land cover and pasture degradation mapping in the humid tropics: A case study of the fragmented landscape of Rio de Janeiro. *International Journal of Applied Earth Observation and Geoinformation*, *78*, 189–201. https://doi.org/10.1016/j.jag.2019.01.011

Pascual, U., Balvanera, P., Díaz, S., Pataki, G., Roth, E., Stenseke, M., … Yagi, N. (2017). Valuing nature's contributions to people: The IPBES approach. *Current Opinion in Environmental Sustainability*, *26–27*, 7–16. https://doi.org/10.1016/j.cosust.2016.12.006

Perkin, H. J. (1969). *The origins of modern English society 1780–1880*. Routledge & Kegan Paul.

Pol, E. (2006). Blueprints for a history of environmental psychology (I): From first birth to American transition. *Medio Ambiente y Comportamiento Humano*, *7*(2), 95–113. http://mach.webs.ull.es/PDFS/Vol7_2/Vol7_2_e.pdf

Prévot-Julliard, A.-C., Julliard, R., & Clayton, S. (2015). Historical evidence for nature disconnection in a 70-year time series of Disney animated films. *Public Understanding of Science*, *24*(6), 672–680. https://doi.org/10.1177/0963662513519042

Queiroz, A. I., Fernandes, M. L., & Soares, F. (2015). The Portuguese literary wolf. *Digital Scholarship in the Humanities*, *30*(3), 388–404. https://doi.org/10.1093/llc/fqt069

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Reeder, D., & Rodger, R. (2001). Industrialisation and the city economy. In *The Cambridge urban history of Britain* (pp. 551–592). https://doi.org/10.1017/CHOL9780521417075.019

Rheinberger, H.-J. (2010). *On historicizing epistemology: An essay*. Stanford University Press.

Richter, R., Hutengs, C., Wirth, C., Bannehr, L., & Vohland, M. (2021). Detecting tree species effects on forest canopy temperatures with thermal remote sensing: The role of spatial resolution. *Remote Sensing*, *13*(1), 135. https://doi.org/10.3390/rs13010135

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rose, D. B., van Dooren, T., Chrulew, M., Cooke, S., Kearnes, M., & O'Gorman, E. (2012). Thinking through the environment, unsettling the humanities. *Environmental Humanities*, *1*(1), 1–5. https://doi.org/10.1215/22011919-3609940

Schmid, B., Balvanera, P., Cardinale, B., Godbold, J., Pfisterer, A., Raffaelli, D., … Srivastava, D. (2009). Consequences of species loss for ecosystem functioning: Meta-analyses of data from biodiversity experiments. In Biodiversity, ecosystem functioning, and human wellbeing (pp. 14–29) https://doi.org/10.5167/uzh-25528

Schmidt-Przewozna, K. (2008). The nature as inspiration for art. *Journal of Natural Fibers*, *5*(1), 88–93. https://doi.org/10.1080/15440470801901555

Seppelt, R., Dormann, C. F., Eppink, F. V., Lautenbach, S., & Schmidt, S. (2011). A quantitative review of ecosystem service studies: Approaches, shortcomings and the road ahead. *Journal of Applied Ecology*, *48*(3), 630–636. https://doi.org/10.1111/j.1365-2664.2010.01952.x

Sörlin, S. (2012). Environmental humanities: Why should biologists interested in the environment take the humanities seriously? *Bioscience*, *62*(9), 788–789. https://doi.org/10.1525/bio.2012.62.9.2

Tilman, D. (1999). The ecological consequences of changes in biodiversity: A search for general principles. *Ecology*, *80*(5), 1455. https://doi.org/10.2307/176540

Vidal, F., & Dias, N. (Eds.). (2015). *Endangerment, biodiversity and culture*. https://doi.org/10.4324/9781315726823

Weller, I. (2019). Gender & Environment: Geschlechterforschung im Kontext Ökologie, Umwelt, Nachhaltigkeit. In B. Kortendiek, B. Riegraf, & K. Sabisch (Eds.), *Handbuch Interdisziplinäre Geschlechterforschung* (pp. 874–880). Springer VS. https://doi.org/10.1007/978-3-658-12496-0

How to cite this article: Langer, L., Burghardt, M., Borgards, R., Richter, R., & Wirth, C. (2024). The relation between biodiversity in literature and social and spatial situation of authors: Reflections on the nature–culture entanglement. *People and Nature*, *6*, 54–74. https://doi.org/10.1002/pan3.10551

APPENDIX 1

SPECIFIC GENDER-DEPENDENT BIODIVERSITY IN LITERATURE

In order to gain an impression at the difference between the genders, we compared them in their specific BiL usage in the respective 100 most frequently occurring taxon labels (Figure A1). We observed notable differences, like distinct topmost frequent taxon labels, a relatively higher usage of plants by female authors and a generally diverging composition of the top 100 terms.

APPENDIX 2

CORRELATION MATRIX BASED ON THE RF INTERACTION

We determined the pairwise interaction that, similar to calculating pairwise covariance, determines the difference between the individual effects and the joint effect based on predictor importance. In order to enable a comparison between the effects, we show the individual differences relative to the respective maximally possible effect (Figure A2), which yields an interaction of 1 for a reflexive pair, like Year:Year, 0 for a non-interacting pair and negative values for mutually reinforcing effects. We clarify that when speaking of interactions, a value of −1 does not mean a strong but reversed correlation that suggests the

PEOPLE AND NATURE BRITISH ECOLOGICAL SOCIETY

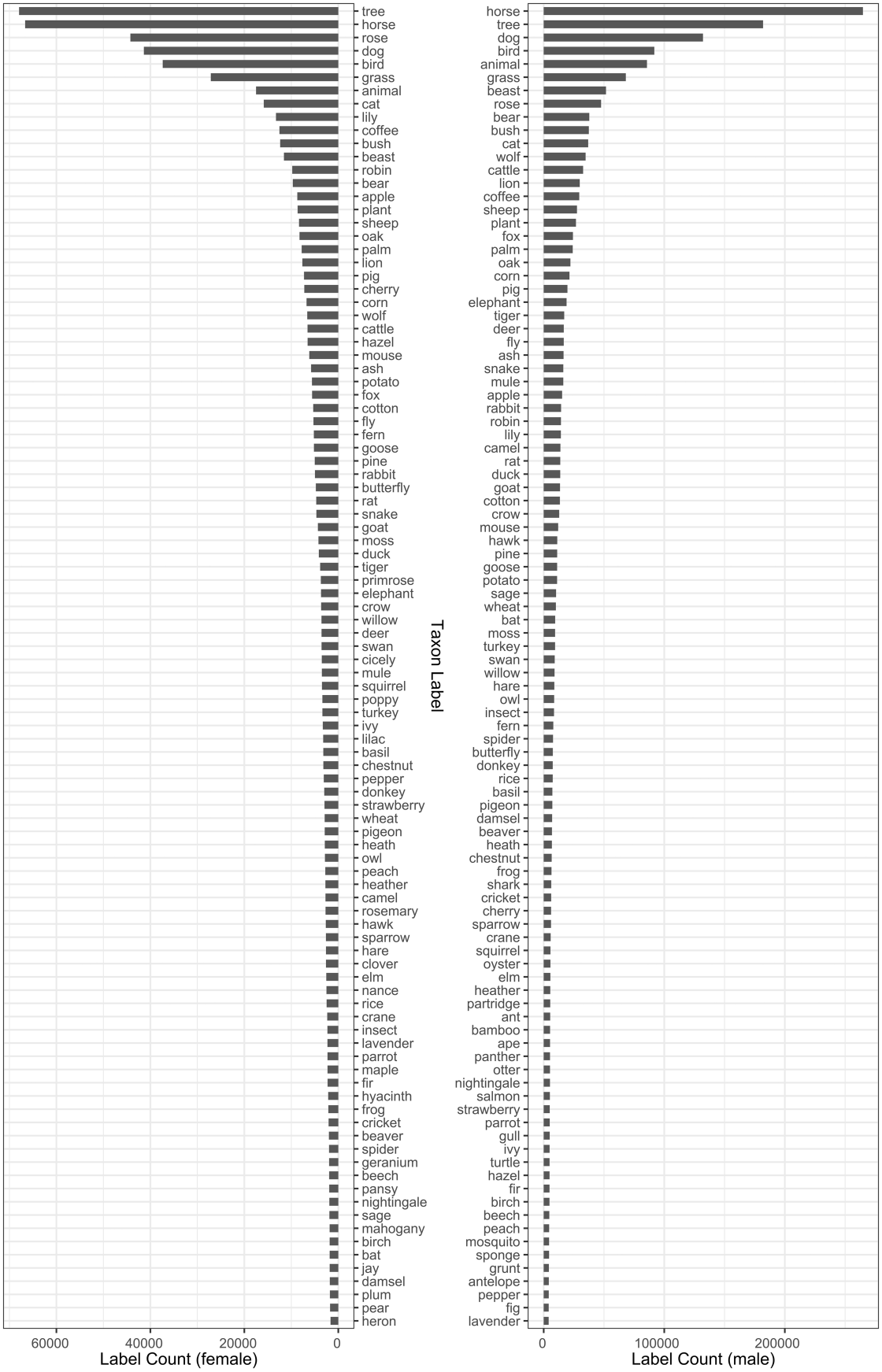

FIGURE A1 Gender-dependent label count of the respective top 100 taxon labels within the corpus. The left side shows the label count for female authors, and the right side shows the label count for male authors.

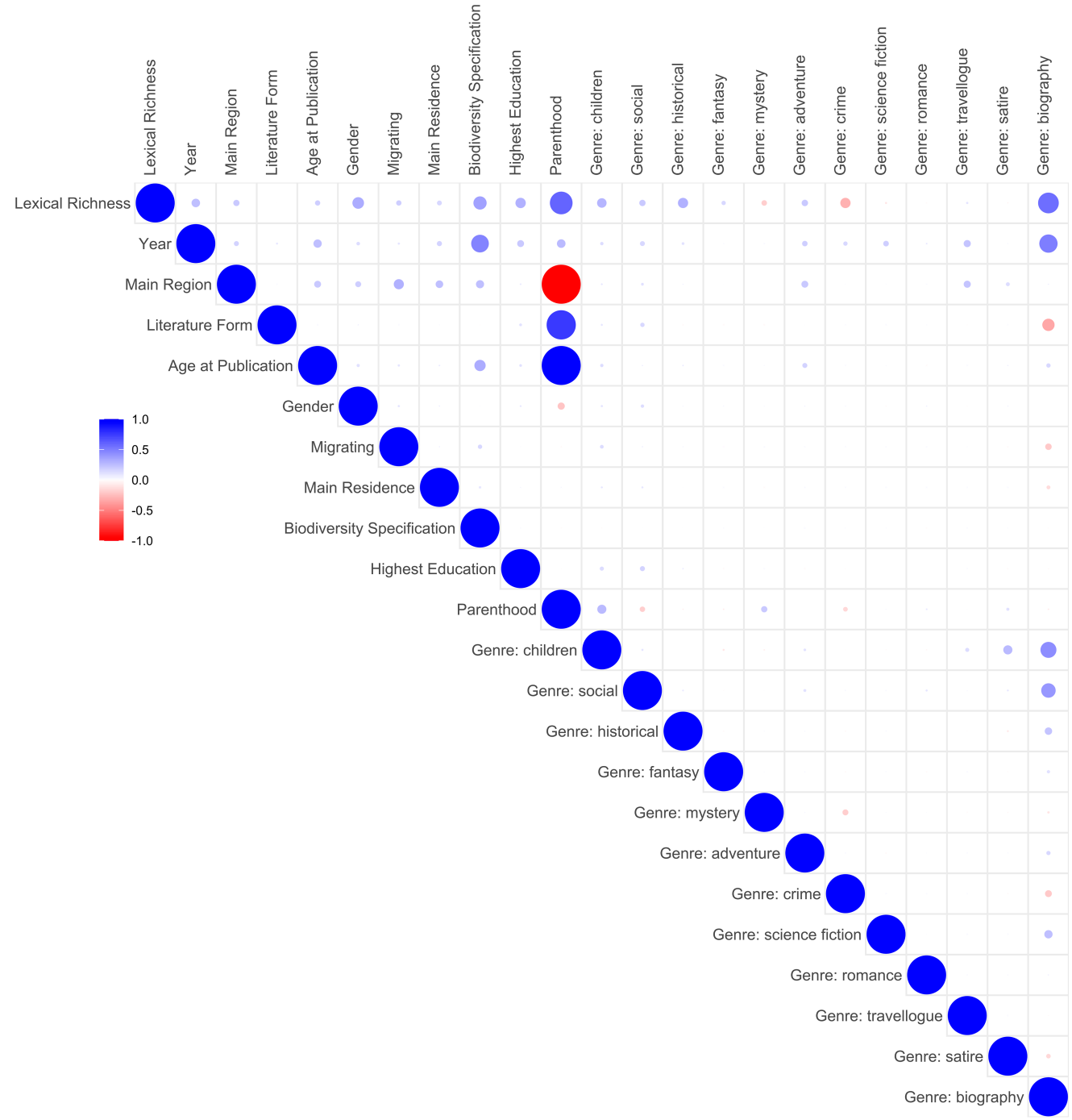

FIGURE A2 Relative interaction matrix, showing the pairwise interaction relative to the added individual interaction. Blue circles indicate an overlapping interaction, whereas red circles indicate a slight reinforcing interaction. Fainter colours and smaller circles indicate a smaller interaction of the respective pair of predictors.

redundancy of one predictor. Instead, values below zero show an increasing meaningfulness of the effect of the pair over the individual predictors, thereby mutually reinforcing their joint expressivity.

We observe three stronger overlapping (blue) interactions between parenthood and literature form, respectively, and age at publication. Looking into these interactions, literature form does not have a notable interaction with age at publication, meaning that parenthood itself appears to be independently affecting BiL and happens to non-systematically affect BiL in a fashion similar to Literature Form or Age at Publication. In addition, the combination of parenthood with main region has a strongly reinforcing effect (red) on the meaningfulness of their impact on BiL. The interactions parenthood with main region and with age at publication suggest further that there simply is an expectable correlation between people of higher ages and specific regional connection to have a higher probability of parenthood. This natural, respectively, cultural correlation does not

mean that these parameters contain different concepts that should be covered when addressing people's living conditions.

Other values for interactions quickly decline to around 0.5 and below with an arithmetic mean of 0.1 and a median below 0.05, both regarding absolute values. As most values are either comparably low or otherwise do not exhibit a systematic tendency of overlapping with other predictors in their expressivity, we regard each individual predictor as a meaningful contribution towards the analysis of the sensitivity for BiL. This is supported by our gradual predictor elimination process (see Section 2.3).

APPENDIX 3

PHASE DEPENDENCE OF THE CORE VARIABLES

In the following, we present the results of our diachronic investigation of the influence of our core predictors (Figures A3–A5).

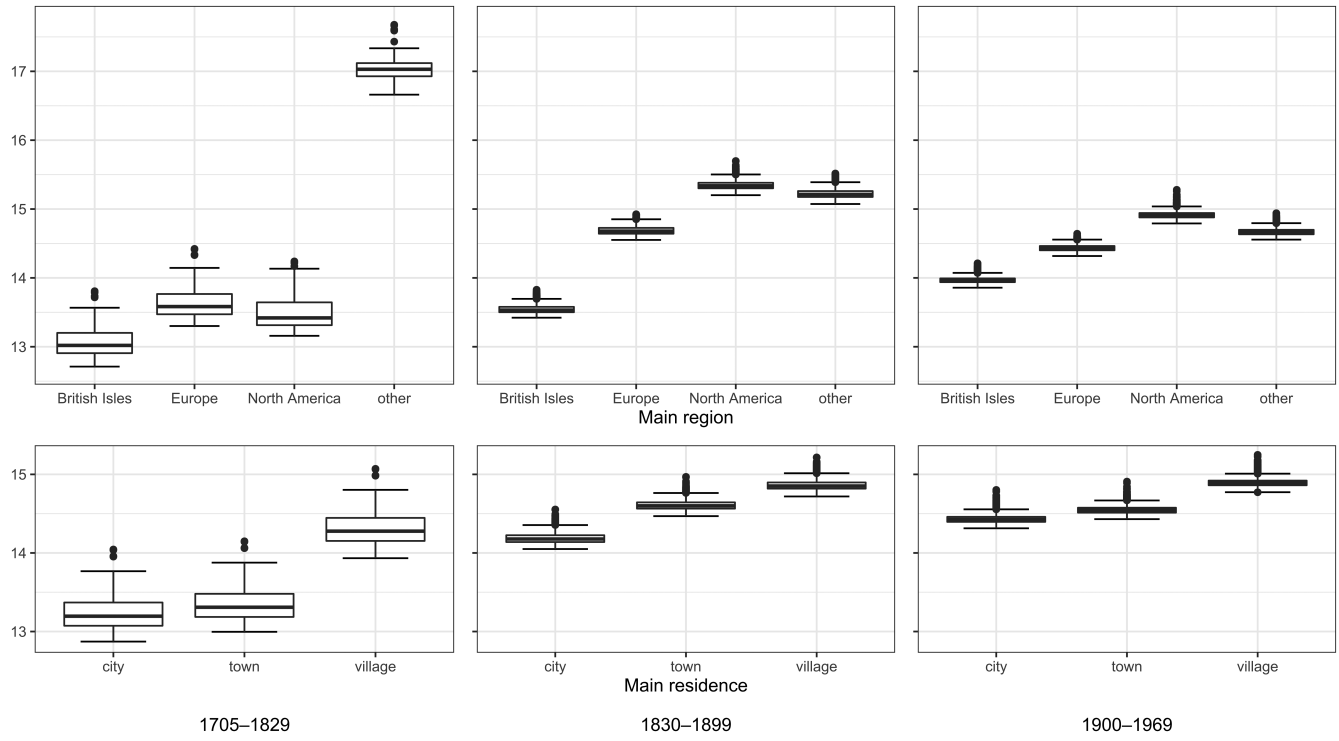

FIGURE A3 Trends for the partial dependencies of spatial predictor variables.

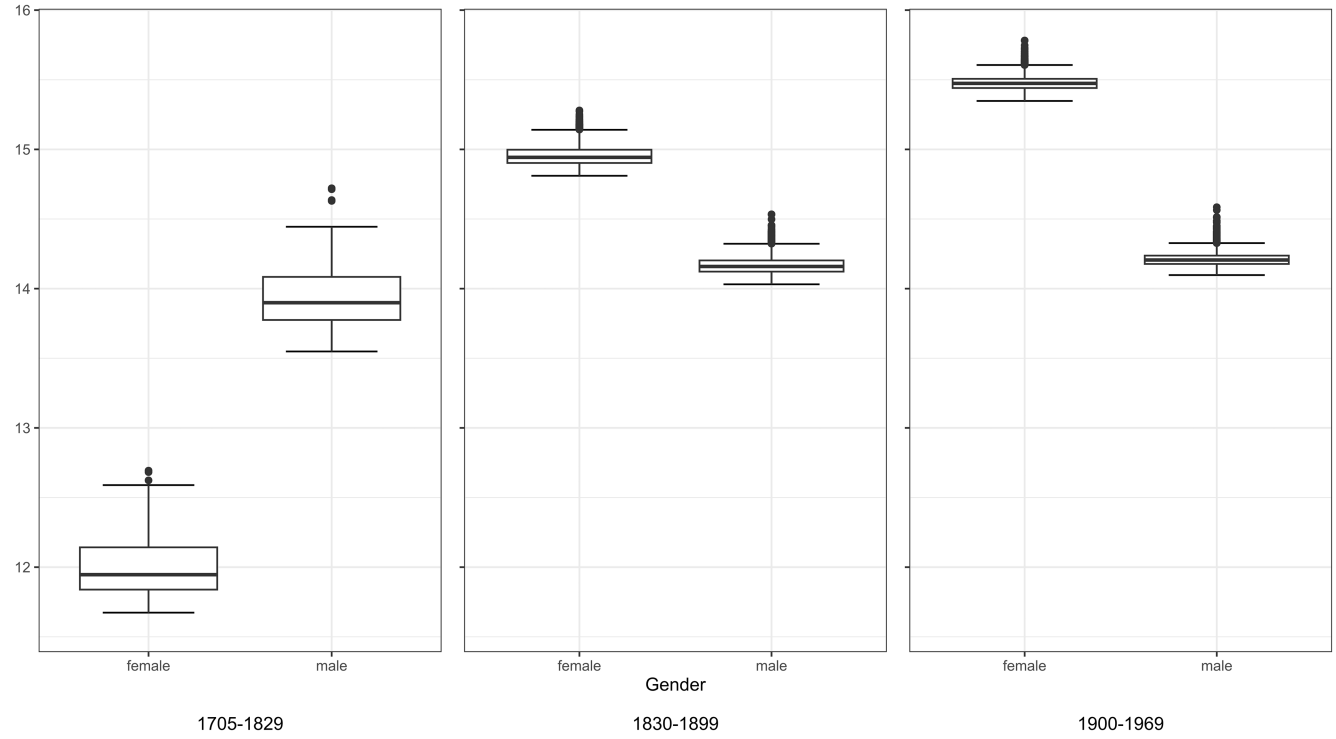

FIGURE A4 Trends for the partial dependencies of the predictor variable for gender.

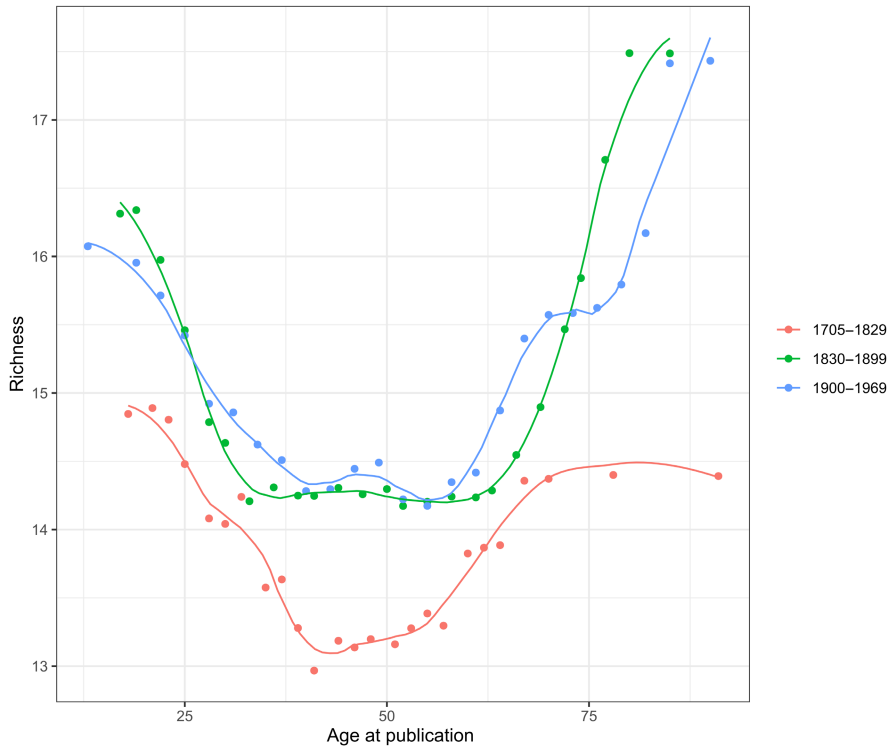

FIGURE A5 Trend for the partial dependency of age as the single core variable with continual character.