# Drivers of Variation in Synonym Numbers of Angiosperm Species Names

**Authors:**

Petra Führding-Potschkat[1]*, Patrick Weigelt[1], Holger Kreft[1], Stefanie M. Ickert-Bond[2],

*Corresponding author, *Biodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences, University of Göttingen, Büsgenweg 1, 37077 Göttingen*. fuehrding@gmail.com

[1]*Biodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences, University of Göttingen, Büsgenweg 1, 37077 Göttingen, Germany*. hkreft@uni-goettingen.de

[2]*Frontier Botany, University of Alaska Fairbanks, 1962 Yukon Drive, Fairbanks, AK 99775-6969, USA*. smickertbond@alaska.edu

## Abstract

Synonyms are part of the scientific progression in taxonomy and nomenclature and reflect the evolving knowledge about species based on revisionary systematics. However, synonyms frequently cause problems in biodiversity repositories, so understanding the causes of the variation of botanical synonyms is essential. Recent studies attribute variation in synonyms to intrinsic and extrinsic drivers, such as nomenclature, taxonomic group membership (e.g., of orchids), and the age of the accepted name. Here, we examine the drivers of the synonyms for a large global subset of all angiosperms. Across 137,378 accepted names of 193 angiosperm families and 5,019 genera present in 355 botanical countries and regions worldwide, range size, the age of the accepted name, and insularity (insular or mainland occurrence, or occurrence on both) emerged as drivers with a positive effect on angiosperm synonyms. After accounting for these three factors, the residual differences in the number of botanical continents and the interaction between insularity and the range size became less significant. The combined multi-predictor model explained about 41% of the global variation in angiosperm synonymy (96%, including the random effects of the families, genera, and the presence patterns of accepted species on one or more botanical continents). We suggest that geographic distance between taxonomists enables wide-ranging species and species with insular distributions to accumulate more synonyms. Also, the age of an accepted name plays a vital role in synonym accumulation. Our results can help to set priorities in revising floras and checklists and to resolve synonymy problems in biodiversity databases, likely leading to more realistic global species numbers. As the drivers may also impact other plant taxa, the study likely has implications for a wider range of families and genera.

## Introduction

Taxonomy aims at identifying, characterizing, and classifying living organisms and thereby sets the foundation for hypothesis-driven research in ecology, biogeography, and conservation biology (e.g., Isaac et al., 2004, Wilson, 2004, Thomson et al., 2018). Taxonomists use morphological, genetic, behavioral, and biochemical characters to identify and describe taxa following specific nomenclatural principles and rules. The principle of priority (Turland et al., 2018), essential to naming organisms, states that the accepted name is the earliest validly published name for a given species; younger names are considered synonyms if more than one name describes the same species. Synonyms may emerge for different reasons, for instance, from different taxonomists interpreting and classifying interspecific variation differently; the two resulting philosophies are referred to as 'splitting' and 'lumping'. If a splitter and a lumper

classify species of the same genus, the former will usually recognize more species than the latter.

It was suggested that if multiple names exist for the same species, these were not solely caused by altered taxonomic relationships, e.g., that the natural variation of a species was unknown, and various forms of the same species were given different names (e.g., Mori, 2013). For example, it was speculated that taxonomists might show preferences toward attractive taxa and that this would increase synonym numbers (Pillon & Chase, 2006, Lughadha et al., 2016). An uneven distribution of synonymy among families and high concentrations in a few large families like Asteraceae, Orchidaceae, and Poaceae was detected by Lughadha et al. (2016). Other studies explored cases of taxonomists describing unknowingly and independently the same species more than once (e.g., Valdecasas et al., 2008, Joppa et al., 2011, Ickert-Bond et al., 2019). For example, due to the geographic distance between taxonomists, wide-ranging species may accumulate more synonyms (e.g., Baselga et al., 2010, Mori, 2013, Fenneman, 2017). This assumption might also be realistic for continents separated by large bodies of water that expand the range, like the Americas and Africa. Also, species with island distributions might accumulate more synonyms because of a higher number of endemic species (Kier et al., 2009) and complex distributional ranges or because of a researcher's assumption that a species discovered on an island is endemic. Other studies proposed that the time passed since the original description of the accepted name plays a crucial role in the accumulation of synonyms (Alroy, 2002, Baselga et al., 2010, Joppa et al., 2011). Finally, some taxonomists noted that other taxonomists were creating species' names as if to 'retain a place in posterity' through authorship of taxa (Bruun, 1950, Pillon & Chase, 2006, Dubois, 2008, Evenhuis, 2008).

Synonyms are an integral part of the natural progression of taxonomy and nomenclature and reflect the ever-changing knowledge about species (Valdecasas et al., 2008, Mori, 2013). Revealing synonyms helps to deepen our understanding of organisms by better understanding otherwise hidden properties of organisms (Holman, 1987). Recent studies, however, showed that the degree of synonymy is quite substantial for some taxa. In some insect groups, the observed ratio of synonyms to accepted names plus synonyms (synonymy rate) exceeds 50% (Gaston & Mound, 1993, Wells et al., 2019). Similarly, it was estimated that around 66% of all published seed plant names are synonyms (Wortley & Scotland, 2004).

Taxonomic uncertainties resulting from the inconsistent treatment of species' delineation and synonymy represent a major challenge for integrating biodiversity data in public data repositories and may lead to erroneous results (Alroy, 2002, Gotelli, 2004, Dubois, 2008, Jansen

& Dengler, 2010). For instance, unresolved synonyms artificially increase the number of names in biodiversity repositories. Synonyms also confuse taxonomy when, for example, it is difficult to recognize whether a species' name in a repository is simply an alias of a more common species (Gaston & Mound, 1993). The same applies to a synonym that cannot correctly relate to an accepted parent name. When taxonomic sources do not consistently identify a scientific name as a synonym, the likelihood for misinterpretation in checklists and other floristic and faunistic treatments increases (Gotelli, 2004, Jansen & Dengler, 2010, Meyer et al., 2016). As a result, thousands of floras and checklists used worldwide are rarely congruent in their taxonomy (Dubois, 2008, Jansen & Dengler, 2010).

Here, we analyzed the variation of synonym numbers in angiosperm names worldwide and tested five competing but not mutually exclusive hypotheses contributing to synonymy (Table 1). We examined the variation in synonym numbers across families and genera. Furthermore, we explored the variation in synonymy across botanical continents where the species were distributed, species' insularity (defined as a species occurrence on islands, the mainland, or both), and the species' range sizes. Finally, we tested the age of the accepted name as a proxy for the time passed since the description of an accepted name. Our results can be used to identify plant taxa that may have an increased probability of unidentified and unresolved synonyms, and to set priorities in revising checklists, floras, or biodiversity databases. The identified name discrepancies can also be further tracked for negative effects across related floras, checklists, and repositories. The outcome of this study likely has implications for a broader range of plant families and genera beyond those examined in the current study.

## Material and Methods

### Data cleaning and preparation of the analysis file

On February 13, 2020, we retrieved 537,000 seed plant name records of 270 families from the World Checklist of Selected Plant Families (hereafter: WCSP; WCSP, 2020), including species' accepted names, synonyms, and publication information. We removed 27,538 non-angiosperm names (Stevens, 2016), 12,927 erroneous, and 7,974 unplaced names (both categories were already flagged by the WCSP). 161,392 accepted names and 340,271 synonyms from 193 angiosperm families remained. An additional 18,262 lower-level names (e.g., subspecies) with 27,453 synonyms were removed, leaving 143,130 accepted names and 312,791 synonyms for a total of 455,921 angiosperm names. We also removed 24,542 synonyms containing nonsensical publication year values (e.g., 0 (zero), 3- or 5-digit years, multiple years, and comments) that could not be matched to an accepted name. For 475 accepted

names with nonsensical years (1,577 assigned synonyms), it was impossible to derive the age of the accepted names even from the oldest synonym. We used parent-dependant relationship information to link the remaining 279,694 synonyms (dependants) to their respective accepted parent names (142,655 records) and counted them (synonym number: *synNum*; hereafter, variable names in italics). The *synNum* served as the response variable during hypothesis testing. The publication year was unavailable for 3,694 accepted names (1.1%). In this case, we used the oldest synonym. The oldest publications date to 1753 (Linnaeus, 1753) and end in 2019 (spanning a total of 267 years). Therefore, we calculated the *age of an accepted name* by subtracting the publication year from 2020. For further analyses, we used the *full accepted name*, *family* (predictor variable of hypothesis H1a, Table 1), *genus* (H1b), *age of an accepted name* (H5), and the *synNum*.

In addition, we used occurrence information for 143,130 accepted seed plants at the species level (in one or more of the 378 TDWG countries and regions, hereafter: TDWG entity, level 3, indicated by 1, presence, and 0, absence; Brummitt, 2001, WCSP, 2020a). From this, second, WCSP file (hereafter: occurrence file) and a spatial polygon (TDWG, 2021), we prepared the predictor variables *botanical continent, where a species is present* (hereafter: "*BC*"; H2a) and *number of botanical continents on which a species occurs* (hereafter: *BCNum*, H2b). In addition, we established the predictor variable *insularity* of a species and computed the *range size* by summing the areas of the respective TDWG units. We considered eight botanical continents (by TDWG Level 1 code, Brummitt, 2001), excluding Antarctica, to avoid the bias of a large continent with very few species (WCSP, 2020a. If a species was reported in a TDWG unit (given in the occurrence file for each accepted name), we marked the corresponding botanical continent ('1', presence and '0', absence). Furthermore, we summed up the *BCNum*. We concatenated the species' occurrences on the eight botanical continents into an eight-digit *BC* string (presence-absence patterns). The TDWG continent number was the position number in the string, determined from left to right. Examples for presence-absence patterns were, e.g., for the presence in South America only: '00000001', and presence in Europe, Africa, and South America: '11000001'. We determined the *insularity* of a species by their respective TDWG classification (Brummitt, 2001). For example, Australia and its continental subunits (e.g., Western Australia, Queensland) were classified as mainland, and Tasmania as an island. Depending on the determined species insularity type, we set *insularity* to 'I' (islands), 'M' (mainland), or 'A' (island and mainland) (factor with three levels). We regarded a species' *range size* as a proxy for the physical distance between taxonomists. As an estimate for *range size*, we computed the sum of all country areas where a species was reported.

We merged the continent-related explanatory variables (the presence-absence string *BC* and *BCNum*), *insularity*, and the total *range size* to the initial part of the analysis file, achieving a final set of nine variants of five putative drivers of synonym numbers in angiosperms. Resulting from the merge, we identified 2,058 accepted names with 6,592 synonyms that were not associated with a TDWG unit or Antarctica. We also identified 3,219 records of accepted species with 11,278 synonyms that had not all predictor variables filled with values and therefore had to be removed. The data cleaning process resulted in 137,378 accepted angiosperm names with 261,824 synonyms.

**Statistical modelling**

Collinearity among predictor variables was tested using the *R* package *rstatix* (Kassambara, 2020) and visualized using the *GGally* package (Schloerke et al., 2018; Appendix, Figures A1(a) to (f)). We examined skewness and kurtosis of the data using the package *moments* (Komsta et al., 2015, Appendix, Table A2), and nested, multilevel structures with *lmerTest* (Kuznetsova et al., 2017). Structural details of the data were visualized using *ggplot2* (Wickham, 2016, e.g., Appendix, Table A2) and the *ggpubr* function *ggdensity* (Kassambara, 2020a, Appendix, Table A2(a), Density diagram).

We used generalized linear mixed effects models (GLMM) to examine the drivers of synonym numbers. We analyzed the linear relationships of *synNum*, including interactions of explanatory variables and assessed variable performances using *R* packages *ROCR* (Sing et al., 2015) and *performance* (Lüdecke et al., 2021). We natural log-transformed *range size* to approximate its observed distribution to a normal distribution. The explanatory variables were standardized (z-transformation, using the *rescale* function) to improve the linearity and comparability of coefficient estimates. We analyzed the suitable error distribution for the count data and the appropriate link functions (Garson, 2013). Frequent issues to be handled in count data are zero-inflation (e.g., Hartig, 2019) and overdispersion (causing incorrect standard errors, e.g., Bell & Grunwald, 2011, Meyer, 2021). In terms of error distribution, Poisson, Poisson/zero inflation, and negative binomial, Poisson/zero inflation, the employed logit link function provided the best-fitting models (Tlhaloganyang & Sakia, 2020, Appendix, Table A2).

The variables *range size* and *age of an accepted name*, *insularity*, and *BCNum* were used as fixed factors in the GLMM model. The other variables, *family* (193 levels), *genus* (5,019 levels), and *BC* (217 levels) were used as random factors (McGill, 2015, Appendix, Figure A3). All variables showed significant effects (SE < 0.013, *p*-values < 0.001) in the GLMM analyses,

suggesting they were predictive (Bell et al., 2019). In addition to the single predictor variables, we tested how the interaction of species occurring on islands, the mainland, or both related to their range size (hypotheses H3 and H4) influences the accumulation of synonyms (Hox et al., 2017, partial correlation analysis: R-package *ppcor*, Kim 2015; $p = 0$).

We fitted multi-level regression models using the *R* package *glmmTMB*, which minimized overdispersion and zero inflation (Bolker, 2016; see: Table A2). We used three distinct goodness-of-fit measures for the model selection: the Akaike information criterion (AIC; Burnham & Anderson, 2004), the root-mean-square error (RMSE), and the marginal and conditional pseudo-$R^2$ (Nagakawa & Schielzeth, 2013, Johnson, 2014, Schielzeth et al., 2020). We computed models of the individual predictors in all possible combinations (Stoffel et al., 2021; predictors: four fixed factors, one interaction, and three random factors). The possible combinations were determined by the mandatory specifications of the used algorithm. At least one random factor was compulsory for *glmmTMB*. The computations delivered the $R^2$ proportions of the fixed and random factors for the models (as the conditional and marginal $R^2$s). We decomposed the $R^2$s per explanatory variable as described in the computation procedure of the R packages *PartR2* (Stoffel et al., 2021) and *rptR* (Stoffel et al., 2017). We selected four models, all with an almost identical AIC at a stable minimum and a maximized pseudo-$R^2$ (model selection criteria; Myung, 2000. Appendix, Table A4(a) and (b)). We evaluated the model performances with the packages *jtools* (Long 2017), *sjPlot* (Lüdecke, 2021), and residual information. We also analyzed the models with the *DHARMa* diagnostics package (Hartig, 2020). We performed a Kolmogorov-Smirnov (KS) test (normal distribution of the residuals), an overdispersion, and an outlier test. The *p*-values were calculated for each model (Appendix, Figure A5, Appendix, Table A4(a) based on 500 replications.

While we counted *synNum* per accepted species, we computed a *synonymy rate* (*synRate*) from the *sum of accepted species* and their collective *synNum*s (both from the counted, hereafter: observed, and predicted by the GLMM model) of a given group (*synRate* = (*synNum*/(*sum of accepted names* + *synNum*)) * 100 [%]; Lughadha et al., 2016). The predicted values were reverted from the natural log using the R *exp* function. The *synRate*s allowed a species richness-independent ranking of each categorical predictor level based on the observed or predicted *synNum* (predicted: from the model) they accumulated or computed for their accepted names (*family*, *genus* and *BC*). The predicted *synNum* were higher than the observed *synNum*. For example, we extracted the observed and predicted synonyms per botanical continent using the variable *BC* as a presence indicator. We summed the synonym numbers per botanical continent

according and analyzed the variation between observed and predicted synonym numbers (a) per botanical continent and (b) across botanical continents (Figure 1).

For the data retrievals, manipulations, analyses, and modeling in this study, we employed *R* Studio and *R* versions 3.0.2-3.2.1 (*R* core team, 2013).

## Results

### Data basis for model fitting

The data cleaning exhibited out-of-scope species records, i.e., non-angiosperms (27,538 records, including 12,928 erroneous records and 176 unplaced records, respectively marked by the WCSP; 5.1%, based on the initial 537,000 WCSP seed plant records), unplaced angiosperm names (7,799 records, 1.5%), subspecific angiosperm names including their synonyms (45,715 records, 8.5%), and species, occurring on the continent of Antarctica (1,608 records, 0.3%). Among the species of interest, we also found records lacking correct values for essential variables. This category contained 2,052 accepted names and synonyms where no oldest name was available (leaving the *publishing year* and, subsequently, the *age of an accepted name* variable empty; 0.4%). This category comprised erroneous synonym records containing nonsensical data in variables essential for our study (24,542 records, 4.6%, e.g., zeroes, text in the *publication year*) that could not be matched to an accepted name. This category also included 1,661 accepted names with 5,381 synonyms that were not assigned to a BC (1.3%) and 3,219 accepted names with 17,863 synonyms that were missing proper values in one or more of the relevant predictor variables (3.9%). During data cleaning, we removed a total of 137,798 records (25.7%, summed from the individual percentages).

We ultimately obtained 137,378 accepted names (25.6%) with a total of 261,824 synonyms (48.8%) present in 355 TDWG units. The *synNum* varied strongly between zero and 377, mean *synNum* was 1.904, median *synNum* was 1, and the distribution was strongly right-skewed (skewness coefficient: 17.58) with a steep kurtosis (685.65). Natural log-transforming the *synNum* led to a slightly right-skewed, approximated normal distribution (skewness coefficient: 1.33, kurtosis: 4.68, Appendix, Table A2(a)). 68,979 of 137,378 accepted angiosperm names (50.2%) had no synonyms, while five names accumulated more than three hundred synonyms each since 1753. The five accepted species with the highest synonym numbers were *Mentha arvensis* L. (377 synonyms), *Sorghum bicolor* (L.) Moench (344 synonyms), *Pandanus tectorius* Parkinson *ex* Du Roi (321 synonyms), *Oryza sativa* L. (320 synonyms), and *Mentha*

*aquatica* L. (302 synonyms). Table 2 lists the top-fifteen accepted names with the highest synonym numbers among angiosperms available in the WCSP.

The synonym numbers differed significantly across families and genera. Synonym numbers per family varied from no synonyms (in sixteen out of 193 families) to more than 40,000 in the Poaceae (47,443 synonyms) and Orchidaceae (43,839 synonyms). Cannaceae exhibited the highest *synRate* (95.2%) of all families for the twelve accepted names and 238 synonyms (*synNum* [mean]: 19.83). The relatively small family Potamogetonaceae took second place (88.2%) with 106 accepted names and 790 synonyms (*synNum* [mean]: 7.19). Large families such as the Poaceae ranked 12th with a *synRate* of 80.4% (*synNum* [mean]: 4.11). The Orchidaceae ranked 99$^{th}$ with a *synRate* of 60.3% (*synNum* [mean]: 1.5). (Details: Table A6(a)). At the generic level, *synNum* varied by four orders of magnitude, ranging from zero synonyms (in a total of 578 genera out of 5,019) up to more than 4,000 (*Carex*, *synNum* [mean]: 2.42) and 3,000 (*Dendrobium*, *synNum* [mean]: 1.95, *Euphorbia*, *synNum* [mean]: 2.46, and *Cyperus*, *synNum* [mean]: 3.54). The highest *synRates* were found for *Ricinus* (*R. communis*, one accepted name and 212 synonyms) and *Phillyrea* (two accepted names and 247 synonyms). Both had a *synRate* of more than 99% (Details: Table A6(b)).

The *synRates* also varied among the botanical continents (Figure 1, Table 3). Europe, Pacific, and North America emerged as the continents with the highest *synRates* from observed synonym numbers (90.7%, *synNum* [mean]: 9.79; 85.6%, *synNum* [mean]: 5.96; 84.2%, *synNum* [mean]: 5.31).

**Drivers of synonym numbers**

Collinearity among the explanatory variables was generally low and highest between the numerical variables *range size* and *age of the accepted name* (absolute Pearson correlations of 0.35). Repeated testing of different error distributions and the *DHARMa* diagnostics showed that the best model performances were obtained using the *glmmTMB* package, the Poisson/zero-inflation error distribution, and the logit-link function (Appendix, Table B2). Iterative GLMM analyses resulted in four fitted models of very similar model parameters, performing nearly equally well. As a result, the AIC values and the conditional and marginal $R^2$s of the four models were also similar (Appendix, Table A4(a)), ranging from 4.880E+05 to 4.882E+05. The RSME of 4.005 revealed a high predictive accuracy with a quantified average error of 4%. The conditional $R^2$s ranged from 0.958 to 0.964, and the share of the fixed factors (marginal $R^2$) ranged from 0.396 to 0.414. Only Model 4 met the equidispersion requirement (conditional $R^2$

of 0.989, fixed factors: 0.414). According to *DHARMa* diagnostics (Appendix, Figure A5), the models did not show significant zero inflation (i.e., given the fitted model, the expected and modeled zeroes were in the same range, Hartig, 2019). Thus, we selected Model 4 as the final model to explain the combined drivers of synonym numbers.

The combined multi-predictor GLMM model 4 explained about 41% of the global variation in angiosperm synonym numbers (96% including the random effects; Table 3). The model included the range size (explaining 21.0% of the variation in synonym numbers), the age of an accepted name (11.6%), and insularity (5.6%) as main predictors (Table 3). We observed root-mean standard errors (RMSE) between 4 to 5% suggesting that the variables were highly predictive. Range size had a positive effect on the accumulation of synonym numbers. The larger the range size of an accepted species, the more synonyms it accumulated (Rank 1, Figure 2B – with insularity, Table 3). The age of an accepted species had a positive effect on the species' accumulated synonym numbers (Figure 2C): The more time had passed since the description of a species name, the more synonyms it accumulated. (Rank 2, Figure 2C, Table 3). The three insularity types showed a positive effect on species' accumulated synonym numbers, albeit to different extents, as displayed in the regression lines with varying points of intersection and slopes (Rank 3, Figure 2B). Species found on islands had a significantly lower *synRate* than those found on the mainland or even both islands and the mainland (99 percent confidence interval: $p < 2.2e\text{-}16$). Species observed only on islands showed a *synRate* of 51.0% (*synNum* [mean]: 1.04), and species only present on the mainland showed a *synRate* of 58.2% (*synNum* [mean]: 1.39). Yet, species present in both showed a *synRate* of 89.0% (*synNum* [mean]: 8.09). The working residuals (Hardin & Hilbe, 2007) varied somewhat for the range size and the age of the accepted name, and they varied slightly within the insularity based on the range size (Figures A7a-c). For the age of an accepted name, the working residuals varied only slightly. We also found differences for the *BCnum* and the interaction of the range size and insularity (Table 3). The number of botanical continents on which a species is present had a positive effect on accumulated synonym numbers, but showed only weak effects on global synonym numbers (Rank 4, Figure 2A, Table 3). The interaction of insularity and the range size showed very weak effects (Rank 5, Figure 2B, Table 3). The botanical continent's *synRates* (predicted synonym numbers from the patterns, split per botanical continent: *BC*) confirmed the ranking from the observed synonym numbers, but were higher. For Europe, a predicted *synRate* of 96.5% was computed, followed by Pacific with 94.6%, and North America with 93.3%. The observed *synRate* of South America increased from 67.8% to a predicted *synRate*

of 86.7%, similar to Asia-Tropical, where the observed *synRate* increased from 69.9% to 87.6% (Figure 1).

## Discussion

In this study, we analysed geographical and taxonomical patterns and drivers of synonymy of 137,378 accepted angiosperm names and 261,824 synonyms from 5,019 genera and 193 families on eight botanical continents. We examined five competing but not mutually exclusive hypotheses of synonym numbers (Hypotheses H1 to H5, Table 1). We observed a large variation in synonym numbers in the used global subset of angiosperms ranging from zero (about 50% were accepted names without synonyms) to 377 synonyms. Variation in synonym numbers was associated with all drivers investigated, which positively affected the accumulation of synonym numbers, but range size, the age of an accepted name, and insularity emerged as the primary drivers. Together, these three drivers explain about 41% of the global variation in angiosperm synonymy, the results are presented in order of their relative importance, below.

### Drivers of synonym numbers

Among all analyzed factors, the range size (H4) emerged as the driver with the highest predictive power for the accumulation of synonyms among the three primary drivers. This finding supports the hypothesis that widespread angiosperm species collect more synonyms than range-restricted species as the the geographical distance between taxonomists is large (Baselga et al., 2010, Fenneman, 2017, Figure 2B).

The age of an accepted name (H5) served as the proxy for the time that passed since the publication of an accepted angiosperm name. This variable also positively affected synonym numbers and ranked second in predictive power. The result corroborates the "historical accumulation of names" hypothesis which states that the more time had passed since the description of a species' name, the more synonyms it accumulated (e.g., Baselga et al., 2010, Joppa et al., 2011, Figure 2C).

In our analyses, the three insularity types positively affected the accumulated synonym numbers (H3: rank 3). However, we found differences between the types' accumulation extent (Figure 2B). Computed from counted *synNum*, the *synRate* of species present on islands and the mainland show 89.0%, compared to the *synRates* of species restricted to islands (51.0%) and the mainland (58.2%). Computed from predicted *synNum*, the *synRate* of species present on islands and the mainland still shows 63.1%, while the *synRates* of islands and mainland species

drop to 2.2% and 6.3%, respectively. The results are probably due to extended species' ranges, and the ranges increased complexity.

Synonym numbers were unevenly distributed among the studied families and genera and differed significantly (Hypothesis H1). The rank positions of families and genera (by *synRate*) may hint at particular taxa being more notable than others. For example, some families are morphologically difficult (e.g., Poaceae), others tend to produce hybrids (e.g., Betulaceae). Also, the attractiveness of a taxon may have a decisive impact on taxonomists' motivation in general (Henrich & Gil-White, 2000, Pimm & Joppa, 2015, Jensen, 2019). However, attractiveness is subjective and difficult to quantify. Thus, our results cannot support findings in previous studies which suggested that particular families, like Orchidaceae, accumulate more synonyms due to being more attractive to researchers than others (Pillon & Chase, 2006, Lughadha et al., 2016) (Tables A6(a) and (b). Yet, the attractiveness of taxonomic study objects in the selection process of researchers (e.g., due to specific pollination mechanisms, ecology, and horticultural value, Heß, 1990, Lughadha et al., 2016) may warrant a study on their consequences on research biases.

The expectation that species in particular botanical continents and continent combinations will accumulate synonyms more frequently was confirmed. However, the botanical continent proved to be a contradicting driver when comparing the predictive power of the continent patterns to the number of continents a species is present. With almost 32%, the botanical continent accounted for a dominant proportion as a random effect (H2a, continent patterns: high conditional pseudo R2 share). The number of continents, a species is present, had a positive effect on the accumulation of synonyms, albeit with very low predictive power, accounting only for 3% of the global variation (H2b, number of continents: very low marginal pseudo R2 as a fixed effect).

Synonym numbers varied systematically by botanical continent (Hypothesis 2a). Europe, Pacific, and North America emerged as the continents with the highest *synRate*s based on observed synonym numbers (from nearly 85% to more than 90%). The *synRate*s from predicted synonym numbers confirmed these results, although predicted synonym numbers were slightly higher (by about 10%) as compared to observed synonym numbers. Contrary to this overall trend, the observed *synRate* of South America and Asia-Tropical, however, increased by 15 to 20% (Figure 1). Overall, these results are consistent with the notion that numbers of invalid, infraspecific, and hybrid names are significantly higher in Europe than in surrounding areas, which coincides with the high number of systematists working there (Pillon & Chase, 2006).

(Hypothesis 2b). Also, for the Eupelmidae (family of parasitic wasps), it was found that the larger their species range size and the more western a Eupelmid species was located, the earlier a species was described both in Afrotropical and in the Palearctic biogeographical regions (Baselga et al., 2010).

Considering species with high synonym numbers, it is striking that mostly their range is recorded across a higher number of botanical continents. In addition, some of these species are native only to one or a few continents. The botanical continent's predictive power was possibly influenced by such species introduced to new continents. The extension of species' range sizes due to cultivation or invasiveness may have created new opportunities for species to accumulate additional synonyms outside their native range. For example, *Mentha arvensis* with 377 synonyms was a taxon in our analysis that accumulated the highest number of synonyms. The species is native to Africa and Asia-Tropical, and was introduced to Europe in the 16th century as a pharmaceutical (Roy et al. 2020). *M. arvensis* was introduced to at least ten countries (GBIF, 2022), and recorded for seven out of eight continents by the WCSP (WCSP, 2020a). Its European relative *M. aquatica* (302 synonyms) is likewise pharmaceutically significant. It was introduced to at least seven countries (GBIF, 2022), and recorded for five out of eight continents by the WCSP (WCSP, 2020a). Today known as one of the most important crops worldwide (Dial 2012), *Sorghum bicolor* (344 synonyms) is an even more extreme example than *M. arvensis*, having been introduced to 54 countries or islands on eight out of eight continents (WCSP 2020a, GBIF 2022). *S. bicolor* originated in the savannahs of north-eastern Africa (De Wet & Harlan, 1971). Effects from such events were not considered in the models. Taking all findings into account, it may be interesting to investigate the role of the botanical continent on the accumulation of synonyms further.

## Conclusion

In our study, we identified range size, the age of an accepted name, and insularity as the main drivers that positively affected the global variation of synonym numbers. Residual differences in the number of botanical continents and the interaction of insularity and the range size became less significant. Our combined multi-predictor model explained about 41% of the global variation in angiosperm synonymy. Four main interpretations emerged from the study. First, the geographic distance between taxonomists caused widespread and insular species to accumulate more synonyms. Also, the time passed since the publication of an accepted species played a dominant role – a trend that is expected by chance. Second, the rank positions of families and genera may hint at particular taxa being more appealing than others. Thus, the

attractiveness of taxonomic study objects in the selection process of researchers and the associated research bias may warrant further study. Third, the predictive power of the continent patterns (high) and the number of continents a species is present (low) contradict each other. Also, the artificial extension of species on the botanical continents due to cultivation or invasiveness needs more attention. Therefore, it may be interesting to further explore the botanical continent's role. Fourth and finally, the outcome of this study likely has implications for a wider range of plant families and genera and might also extend to other groups of organisms.

# References


Alroy, J. (2002). How many named species are valid? *Proceedings of the National Academy of Sciences* **99**: 3706–3711. DOI: https://doi.org/10.1073/pnas.062691099.

Baselga, A., Lobo, J. M., Hortal, J., Jiménez-Valverde, A., & Gómez, J. F. (2010). Assessing alpha and beta taxonomy in eupelmid wasps: determinants of the probability of describing good species and synonyms. *Journal of Zoological Systematics* **48**: 40–49. DOI: https://doi.org/ 10.1111/j.1439-0469.2009.00523.x.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using *lme4. Journal of Statistical Software* **67**: 1–48. DOI:10.18637/jss.v067.i01.

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: making an informed choice. *Quality & Quantity* **53**: 1051–1074. DOI: https://doi.org/10.1007/s11135-018-0802-x.

Bell, M. L., & Grunwald, G. K. (2011). Small sample estimation properties of longitudinal count models. *Journal of Statistical Computation and Simulation* **81**: 1067–1079. DOI: https://doi.org/10.1080/0094965100367 4144.

Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., & Blanchet, G. (2015). Package *spdep*. *The Comprehensive R Archive Network*. [Accessed 2020 November 08]. Available from: https://www.yumpu.com/en/document/view/9283478/package-spdep-the-comprehensive-r-archive-network.

Bolker, B. (2016). *Getting started with the* glmmTMB *package*. *R* Foundation for Statistical Computing, Vienna, Austria. [Accessed 2020 December 02]. Available from: cran.uni-muenster.de.

Brummitt, R.K. (2001). *World Geographical Scheme for Recording Plant Distributions*. Plant Taxonomic Database Standards No. 2, Edition 2. Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh. [Accessed 2017 October 10]. Available from: biofund.org.mz.

Bruun A. F. (1950). The Systema Naturae of the twentieth century. *Science* **112**: 342–343. DOI: https://doi.org/ 10.1126/science.112.2908.342.b.

Burnham, K. P., &Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* **33**: 261–304. DOI: https://doi.org/10.1177%2F0049124104268644.

De Wet, J. M. J., & Harlan, J. R. (1972). The origin and domestication of *Sorghum bicolor*. *Economic Botany* **25**: 128–135. [Accessed 2020 December 15]. Available from: https://www.jstor.org/stable/4253238.

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). *Addressing big data issues in scientific data infrastructure*. International conference on collaboration technologies and systems (CTS). IEEE, 2013. [Accessed 2017 October 10]. Available from: https://ieeexplore.ieee.org/abstract/document/6567203.

Dial, H.L. (2012). Plant guide for *sorghum* (*Sorghum bicolor* L.). USDA-Natural Resources Conservation Service, Tucson Plant Materials Center, Tucson, AZ. Accessed on: 2022-07-11. Available at: https://plants.usda.gov/pdf/pg_sobi_2.

Dubois, A. (2008). A partial but radical solution to the problem of nomenclatural taxonomic inflation and synonymy load. *Biological Journal of the Linnean Society* **93**: 857–863. DOI: https://doi.org/10.1111/j.1095-8312.2007.00900.x.

Evenhuis, N. L. (2008). The "Mihi itch" – a brief history. *Zootaxa* **1890**: 59–68. DOI: https://doi.org/10.11646/zootaxa.1890.1.3.

Fenneman, J. (2017). *Synonyms Explained: Why Plants Sometimes Have Other Scientific Names*. Electronic Atlas of the Flora of British Columbia (eflora.bc.ca). [Accessed 2019 June 13]. Available from: http://ibis.geog.ubc.ca/biodiversity/eflora/VascularPlantSynonymy.html.

Garson, G. D. (2013). *Hierarchical linear modeling: Guide and applications*. Sage.

Gaston, K. J., & Mound, L. A. (1993). Taxonomy, hypothesis testing and the biodiversity crisis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **251**: 139–142. DOI: https://doi.org/10.1098/rspb.1993.0020.

GBIF.org. (2022), GBIF Home Page, search facility. [Accessed 2022 January 07]. Available from: https://www.gbif.org.

Gotelli, N. J. (2004). A taxonomic wish–list for community ecology. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**: 585–597. DOI: https://doi.org/10.1098/rstb.2003.1443.

Hardin, J. W., & Hilbe, J. (2007). *Generalized linear models and extensions*. Stata press. City?

Hartig, F. (2019). GLMM for unbalanced zero inflated data. [Accessed 2021 December 08]. Available from:https://stats.stackexchange.com/questions/396336/r-glmm-for-unbalanced-zero-inflated-data-glmmtmb.

Hartig, F. (2020). *DHARMa*: residual diagnostics for hierarchical (multi-level/mixed) regression models. *R* package version 0.3.3. [Accessed 2020 December 02]. Available from: https://cran.r-project.org/web/packages/DHARMa/index.html.

Henrich, J., & Gil-White, F. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behaviour* **22**: 165–196. DOI: https://doi.org/10.1016/S1090-5138(00)00071-4.

Heß, D. (1990). *Die Blüte*. Stuttgart: Ulmer.

Holman, E. W. (1987). Recognizability of sexual and asexual species of rotifers. *Systematic Zoology* **36**: 381–386. DOI: https://doi.org/10.2307/2413402.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications,* 3rd Ed. Routledge, New York. DOI: https://doi.org/10.4324/9781315650982.

Ickert-Bond, S. M., Murray, D., Oliver, M. G., Berrios, H. K., & Webb, C. O. (2019). The *Claytonia arctica* complex in Alaska—Analyzing a Beringian taxonomic puzzle using taxonomic concepts. *Annals of the Missouri Botanical Garden* **104**: 478–494. DOI: https://doi.org/10.3417/2019491.

Isaac, N. J., Mallet, J., & Mace, G. M. (2004). Taxonomic inflation: its influence on macroecology and conservation. *Trends in Ecology & Evolution* **19**: 464–469. DOI: https://doi.org/10.1016/j.tree.2004.06.004.

Jansen, F., & Dengler, J. (2010). Plant names in vegetation databases – a neglected source of bias. *Journal of Vegetation Science* **21**: 1179–1186. DOI: https://doi.org/10.1111/j.1654-1103.2010.01209.x.

Jensen, S. (2019). *Ausländerstudium in Deutschland: die Attraktivität deutscher Hochschulen für ausländische Studierende.* Springer-Verlag. City?

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's *R* (2) GLMM to random slopes models. *Methods in Ecology and Evolution* **5**: 944–946. DOI: https://doi.org/10.1111/2041-210X.12225.

Joppa, L. N., Roberts, D. L., & Pimm, S. L. (2010). How many species of flowering plants are there? *Proceedings of the Royal Society B: Biological Sciences* **278**: 554–559. DOI: https://doi.org/10.1098/rspb.2010.1004.

Kassambara, A. (2020). *rstatix*: Pipe-friendly framework for basic statistical tests. *R* package version 0.6.0. [Accessed 2021 April 19]. Available from: https://rpkgs.datanovia.com/rstatix/.

Kassambara, A. (2020a). Package *ggpubr*. *R* package version 0.1, 6. [Accessed 2021 April 19]. Available from: cran.microsoft.com.

Kier, G., Kreft, H., Lee, T. M., Jetz, W., Ibisch, P. L., Nowicki, C., Mutke, J. & Barthlott, W. (2009). A global assessment of endemism and species richness across island and mainland regions. *Proceedings of the National Academy of Sciences* **106**: 9322–9327. DOI: https://doi.org/10.1073/ pnas.0810306106.

Kim, S. (2015). *ppcor*: an *R* package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods* **22**: 665. DOI: https://doi.org/10.5351%2FCSAM. 2015.22.6.665.

Komsta, L., & Novomestky, L. (2015). *moments*, cumulants, skewness, kurtosis and related tests. *R* package version 0.14. [Accessed 2021 April 19]. Available from: http://cran.r-project.org/package=moments.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). *lmerTest* package: tests in linear mixed effects models. *Journal of Statistical Software* **82**: 1–26. DOI: https://doi.org/10.18637/jss.v082.i13.

Linnaeus, C. (1753). *Species Plantarum. Vol. 1*. London.

Long, J. A. (2017). Package *jtools*. [Accessed 2021 April 17]. Available from: cran.microsoft.com.

Lüdecke, D. (2021). Package *sjPlot*. [Accessed 2021 June 07]. Available from: mran.revolutionanalytics.com.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). *performance*: An *R* package for assessment, comparison and testing of statistical models. *Journal of Open Source Software* **6**. [pdf]. DOI: https://doi.org/10.21105/joss.03139.

Lughadha, E. N., Govaerts, R., Belyaeva, I., Black, N., Lindon, H., Allkin, R., McGill, R. E. & Nicolson, N. (2016). Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa* **272**: 82–88. DOI: https://doi.org/10.11646/phytotaxa.272.1.5.

McGill, B. (2015). Is it a fixed or random effect? [Accessed 2021 October 11]. Available from: https:// dynamicecology.wordpress.com/2015/11/04/is-it-a-fixed-or-random-effect/.

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology letters* **19**: 992–1006. DOI: https://doi.org/10.1111/ele.12624.

Meyer, J. (2021). Overdispersion in Count Models: Fit the Model to the Data, Don't Fit the Data to the Model. https://www.theanalysisfactor.com/overdispersion-in-count-models-fit-the-model-to-the-data-dont-fit-the-data-to-the-model/.

Mori, S. A. (2013). *Plant talk – The Shifting Science of Botanical Nomenclature, I and II*. [Accessed 2022 January 10]. Available from: https://www.nybg.org/blogs/plant-talk/tag/scott-mori/

Myung, I. J. (2000). The importance of complexity in model selection [Special issue]. *Journal of Mathematical Psychology* **44**: 37. DOI: https://doi.org/10.1006/jmps.1999.1283.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* **4**: 133–142. DOI: https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Nixon, C. G. (2016). *How Valuable is that Plant Species: Application of a Method for Enumerating the Contribution of Selected Plant Species to New Zealand's GDP*. Ministry for Primary Industries.

Pillon, Y., & Chase, M. W. (2007). Taxonomic exaggeration and its effects on orchid conservation. *Conservation Biology* **21**: 263–265. DOI: https://doi.org/10.1111/j.1523-1739.2006.00573.x.

Pimm, S.L, & Joppa, L.N. (2015). How many plant species are there, where are they, and at what rate are they going extinct? *Annals of the Missouri Botanical Garden* **100**: 170–176. DOI: https://doi.org/10.3417/ 2012018.

*R* Core Team. (2013). *R*: A language and environment for statistical computing. *R* Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/

Roy, D., Alderman, D., Anastasiu, P., Arianoutsou, M., Augustin, S., Bacher, S., Başnou, C., Beisel, J., Bertolino, S., Bonesi, L., Bretagnolle, F., Chapuis, J. L., Chauvel, B., Chiron, F., Clergeau, P., Cooper, J., Cunha, T., Delipetrou, P., Desprez-Loustau, M., Détaint, M., Devin, S., Didžiulis, V., Essl, F., Galil, B. S., Genovesi, P., Gherardi, F., Gollasch, S., Hejda, M., Hulme, P. E., Josefsson, M., Kark, S., Kauhala, K., Kenis, M., Klotz, S., Kobelt, M., Kühn, I., Lambdon, P. W., Larsson, T., Lopez-Vaamonde, C., Lorvelec, O., Marchante, H., Minchin, D., Nentwig, W., Occhipinti-Ambrogi, A., Olenin, S., Olenina, I., Ovcharenko, I., Panov, V. E., Pascal, M., Pergl, J., Perglová, I., Pino, J., Pyšek, P., Rabitsch, W., Rasplus, J., Rathod, B., Roques, A., Roy, H., Sauvard, D., Scalera, R., Shiganova, T. A., Shirley, S., Shwartz, A., Solarz, W., Vilà, M., Winter, M., Yésou, P., Zaiko, A., Adriaens, T., Desmet, P., & Reyserhove, (2020). *DAISIE - Inventory of alien invasive species in Europe. Version 1.7*. Research Institute for Nature and Forest (INBO). Checklist dataset. [Accessed 2022 March 15]. Available from: https:// doi.org/10.15468/ybwd3x, accessed via GBIF.org.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L.Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution.* **11**: 1141–1152. DOI: https://doi.org/10.1111/2041-210X.13434

Schloerke, B., Crowley, J., & Cook, D. (2018). Package '*GGally*'. *Extension to 'ggplot2.* [Accessed 2022 January 12]. Available from: cran.microsoft.com

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2015). Package *ROCR*. Visualizing the performance of scoring classifiers. *Bioinformatics*. DOI: https://doi.org/10.1093/bioinformatics/bti623.

Steinbart, P. J., & Nath, R. (1992). Problems and issues in the management of international data communications networks: the experiences of American companies. *MIS quarterly*: 55–76. DOI: https://doi.org/10.2307/249701.

Stevens, P. F. (2016). *Angiosperm Phylogeny Website*. Version 13. [Accessed 2017 January 16]. Available from: http://www.mobot.org/MOBOT/research/APweb/.

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2017). *rptR*: repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution* **8**: 1639. DOI: https://doi.org/10.1111/2041-210X.12797

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2021). *partR2*: Partitioning R2 in generalized linear mixed models. *Bioinformatics and Genomics* **9**: e11414. DOI: https://doi.org/10.7717/peerj.11414.

TDWG. (2021). *Biodiversity Information Standards (TDWG) - SpatialPolygonsDataFrame*. [Accessed 2020 May 17]. Available from: www.tdwg.org/standards/109/tdwg_lv3.

Thomson, S. A., Pyle, R. L., Ahyong, S. T., Alonso-Zarazaga, M., Ammirati, J., Araya, J. F., Ascher, J. S., Audisio, T. S., Azevedo-Santos, V. M., Bailly, N., J. Baker, W. J., Balke, M., Barclay, M. V. L., Barrett, R. L., Benine, R. C., Bickerstaff, J. R. M., Bouchard, P., Bour, R., Bourgoin, T., … & Zhou, H. Z. (2018). Taxonomy based on science is necessary for global conservation. *PLoS One* **16**: e2005075. DOI: https://doi.org/10.1371/ journal.pbio.2005075.

Tlhaloganyang, B. P., & Sakia, R. M. (2020). Zero inflated Poisson distribution in equidispersed data with excessive zeros. *Research Journal of Mathematics and Statistics* **8**: 31–34. [Accessed 2022 February 26]. Available from: www.iscamaths.com.

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T. W., McNeill, J., Monro, A. M., Prado, J., Price, M. J. & Smith, G. F. (eds.). (2018). *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017*. Regnum Vegetabile 159. Koeltz Botanical Books, Glashütten. DOI https://doi.org/10.12705/Code.2018

Valdecasas, A. G., Williams, D., & Wheeler, Q. D. (2008). Integrative taxonomy then and now: a response to Dayrat (2005). *Biological Journal of the Linnean Society* **93**: 211–216. DOI: https://doi.org/10.1111/j.1095-8312.2007.00919.x.

WCSP. (2020). *World Checklist of Selected Plant Families*. Facilitated by the Royal Botanic Gardens, Kew. [Accessed 2017 August 7]. Available from: http://wcsp.science. kew.org/.

WCSP. (2020a). *World Checklist of Selected Plant Families*. Facilitated by the Royal Botanic Gardens, Kew. [Accessed 2017 August 7]. Available from: http://wcsp.science. kew.org/.

Wells, A., Johanson, K. A., & Dostine, P. (2019). Why are so many species based on a single specimen? *Zoosymposia* **14**: 32–38. DOI: https://doi.org/10.11646/zoosymposia.14.1.5.

Wickham, H. (2016). ggplot2: *Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilson, E. O. (2004). Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 739–739. DOI: https://doi.org/10.1098/rstb.2003.1440.

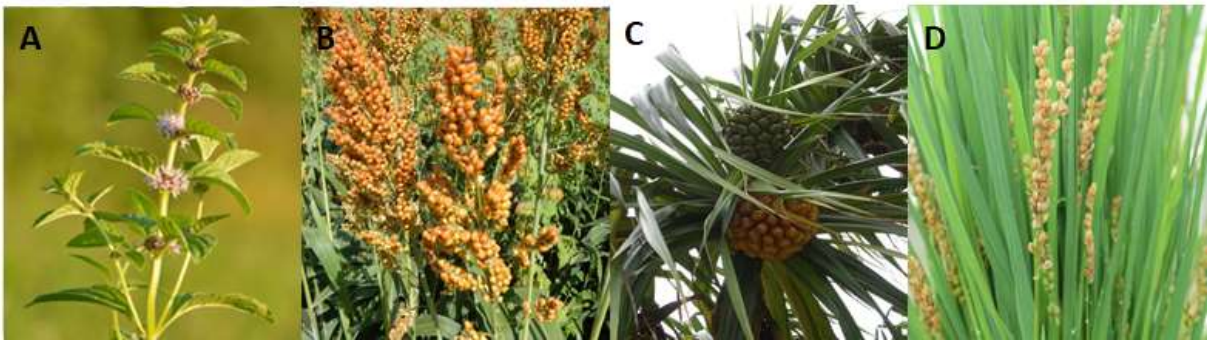
Wortley, A. H., & Scotland, R. W. (2004). Synonymy, sampling and seed plant numbers. *Taxon* **53**: 478–480. DOI: https://doi.org/10.2307/4135625.

## Tables

***Table 1.*** **Hypotheses summary: Drivers of synonym numbers, affecting the variation in synonym numbers and synonymy rates (*synRate*).** Species which are affected by the drivers described in the hypotheses below are expected to accumulate synonyms more frequently.

| |
|---|
| **H1. Synonym number and *synRate* vary among families or genera.** Species belonging to particular angiosperm families and genera – regardless of the higher taxonomic level's species number – have an increased probability of being described as different species (Pillon & Chase, 2006, Lughadha et al., 2016). Thus, we expected species of these families and genera to accumulate synonyms more frequently.<br>Explanatory variables: *family* and *genus* (both categorical). |
| **H2a. Synonym number and *synRate* vary across the botanical continents / continent combinations**. Species present in specific continents and continent combinations have an increased probability of being described as different species. We expected species being present in particular botanical continents and continent combinations to accumulate synonyms more frequently.<br>Explanatory variable: *occurrence on TDWG botanical continent(s)* (except Antarctica) (eight-digit variable, binary: Y = 1 / N = 0).<br>**H2b. Synonym number and *SynRate* vary with the number of botanical continents where species are present**. Species present on more than one botanical continent have an increased probability of being described as different species. We, thus, expected species present on many botanical continents to accumulate synonyms more frequently than species occurring on only one or few continents (extension of H4, below).<br>Explanatory variable: *number of botanical continents a species occurs* (numerical: 1 to 8). |
| **H3. Synonym number and *synRate* vary with the insularity of a species**. Species present on islands have an increased probability of being described as different species than species occurring on the mainland. We expected species present on islands to accumulate synonyms more frequently than species occurring on the mainland only.<br>Explanatory variable: *insularity* of a species, on islands only, on the mainland only, and both on islands & the mainland (using the TDWG classification of the respective botanical country as island or mainland). |
| **H4. Synonym number and *synRate* vary among species range sizes.** Wide-ranging species are more likely to be described as different species (proxy for the geographic distance between taxonomists) than species with small ranges. We expected species with large ranges to accumulate synonyms more frequently than species with small ranges (Baselga et al., 2010, Fenneman, 2017).<br>Explanatory variable: *range size*, computed as the sum of TDWG countries where a species occurs (Source: TDWG shapefile data frame). |
| **H5. Synonym number and *synRate* vary with the age of a species' accepted name.** Species' accepted names validly published a long time ago had more time to accumulate synonyms than recently published accepted names (Alroy, 2002, Baselga et al., 2010). We expected early published names to accumulate synonyms more frequently than recently published names.<br>Explanatory variable: *age of a species' accepted name* or – if not available or younger than the first published synonym – its oldest synonym. |

***Table 2.*** **Summary of the fifteen accepted species names with the highest synonym numbers among the angiosperms studied.** Images A to D show the four species with the highest synonym numbers per species name. (Images: A, *Mentha arvensis*; B, *Sorghum bicolor*; C, *Pandanus tectorius*; D, *Oryza sativa*). Column pubYear (Publication year: For each scientific name marked with *, the publication year was determined using the oldest synonym in the absence of the publication year of the accepted name. (Images: A: Ivar Leidus, B: Forest & Kim Starr. C: Judgefloro. D: C.T. Johansson. Creative commons licences: A, B, and D: CC BY-SA 3.0; C: CC BY-SA 4.0.)




| Family | Scientific name | synNum | pubYear | BotCont | Human use |
|---|---|---|---|---|---|
| Lamiaceae | *Mentha arvensis* L. | 377 | 1753 | 1 to 5,7,8 | medicinal, spice |
| Poaceae | * *Sorghum bicolor* (L.) Moench | 344 | 1753 | 1 to 8 | staple food (crop) |
| Pandanaceae | *Pandanus tectorius* Parkinson ex Du Roi | 321 | 1774 | 2 to 8 | food, building |
| Poaceae | *Oryza sativa* L. | 320 | 1753 | 1 to 8 | staple food (crop) |
| Lamiaceae | *Mentha aquatica* L. | 302 | 1753 | 1 to 3,7,8 | medicinal, spice |
| Asparagaceae | * *Cordyline fruticosa* (L.) A.Chev. | 233 | 1754 | 2 to 8 | ornamental gardening |
| Poaceae | *Festuca rubra* L. | 222 | 1753 | 1 to 8 | ornamental gardening |
| Euphorbiaceae | *Ricinus communis* L. | 212 | 1753 | 1 to 8 | medicinal |
| Poaceae | *Agrostis stolonifera* L. | 209 | 1753 | 1 to 8 | ornamental gardening |
| Rubiaceae | *Kadua affinis* Cham. & Schltdl. | 200 | 1829 | 6 | ornamental gardening |
| Oleaceae | *Phillyrea latifolia* L. | 187 | 1753 | 1 to 3 | ornamental gardening |
| Campanulaceae | *Campanula rotundifolia* L. | 179 | 1753 | 1,3,5,7,8 | ornamental gardening |
| Myrtaceae | * Myrcia splendens (Sw.) DC. | 170 | 1788 | 7,8 | medicinal, fruits, timber |
| Poaceae | *Festuca ovina* L. | 168 | 1753 | 1 to 4,7,8 | - |
| Cannaceae | *Canna indica* L. | 166 | 1753 | 1 to 8 | ornamental gardening |

BotCont, botanical continent: 1 = Europe, 2 = Africa, 3 = AsiaTemperate, 4 = AsiaTropical, 5 = Australasia, 6 = Pacific , 7 = Northern America, 8 = Southern America. synNum: synonym number.

***Table 1.*** **Global model of angiosperm synonymy.** Selection conditions of the model were: (1) AIC at a stable minimum, (2) maximized pseudo-$R^2$, (3) *DHARMa* performance tests successful. Result of GLMM of a combined eight-predictor model, by random factors and fixed factors. H1 to H5: Hypotheses (see: Table 1). ***, $p < 0.001$. The table below is an extract of Table A4(b) (Appendix).

| Hypothesis | Combined model | $R^2$ share | $RMSE_{mean}$ | *z* | Variation |
|---|---|---|---|---|---|
| | **Random Factor $R^2$ share** | **0.544** | **-** | | **54.4%** |
| H2(a) | Botanical continents (Presence on particular continents) | 0.302 | 4.857 | - | 30.2% |
| H1 | Genus | 0.223 | 4.404 | - | 22.3% |
| H1 | Family | 0.019 | 4.930 | - | 1.9% |
| | **Fixed Factor $R^2$ (Marg.)** | **0.414** | **-** | | **41.4%** |
| H4 | Range size | 0.201 | 4.621 | 69.7 *** | 20.1% |
| H5 | Age of accepted name | 0.111 | 4.698 | 139.9 *** | 11.1% |
| H3 | Insularity | 0.054 | 4.800 | -31.9 *** | 5.4% |
| H2(b) | No. of botanical continents (a species is present) | 0.029 | 4.763 | 3.6 *** | 2.9% |
| H3*H4 | Range size * Insularity | 0.019 | 5.078 | 17.8 *** | 1.9% |
| | **Total $R^2$ (Cond.)** | **0.958** | **4.005** | | **95.8%** |

## Figures

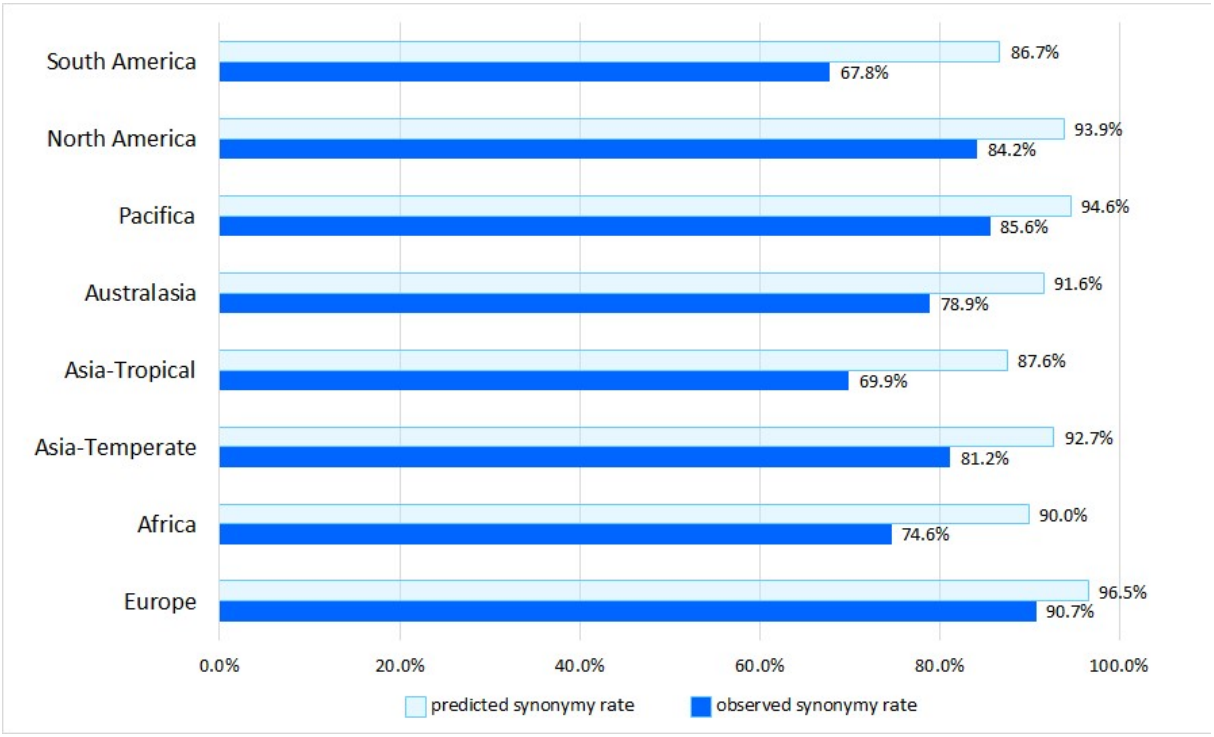



**Figure 1. Variation of synonym numbers across botanical continents (random factor), by synonymy rates (observed: blue, predicted: light blue).** We extracted the observed and predicted synonym numbers per botanical continent using the *BC* (eight-digit presence-absence pattern) as a presence indicator. We summed the synonym numbers per botanical continent and analyzed the variation between observed and predicted (a) per botanical continent and (b) across botanical continents. Observed synonym numbers were counted when linking synonyms to their parent species name. Thus, the observed *synonymy rates* (*synRate*) are derived from this number for the *synRate* formula. The predicted synonym number is derived from the glmmTMB model. Thus, the predicted *synRate* is used in the formula. Both the observed and the predicted *synRates* were computed as: *synRate* = (*synNum*/(accepted names + *synNum*)) * 100 [%] (Lughadha et al. 2016). The *synRate*s allowed for a relative level ranking independent of each continent's absolute synonym number. Europe (90.7%), Pacific (85.6%), and North America (84.2%) emerged as the continents with the highest *synRates* from observed synonym numbers. The predicted *synRates* were even higher. For Europe, a *synRate* of 96.5% was computed, followed by Pacific with 94.6%, and North America with 93.9%. The observed *synRate* of South America increased from 67.8% to a predicted *synRate* of 86.7%, similar to Asia-Tropical, where the observed *synRate* increased from 69.9% to 87.6% (predicted).

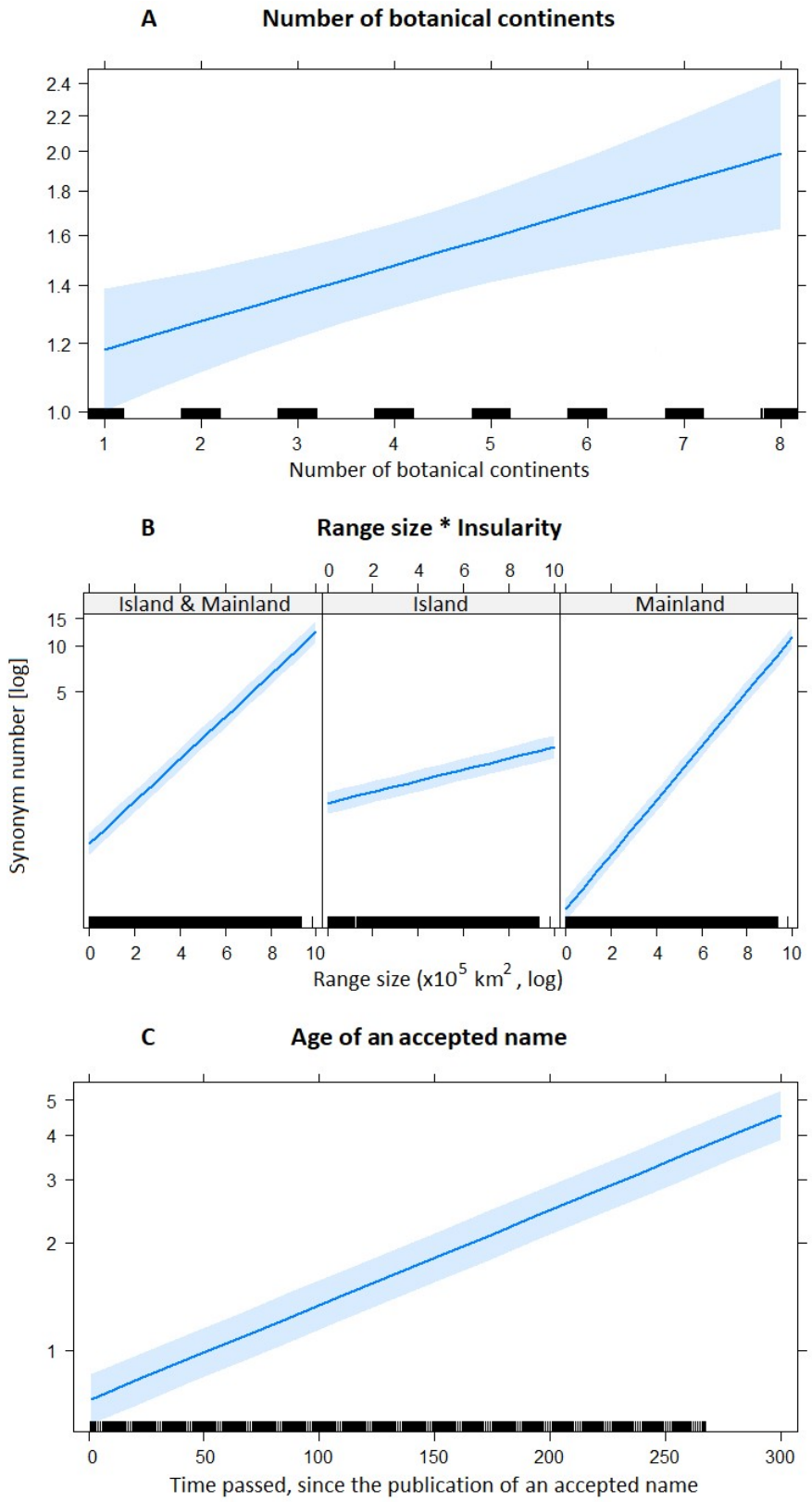



***Figure 2.*** **Variation of synonymy rates (from the predicted fixed factor synonym numbers):** A: Number of botanical continents, a species is present, rank 4. B: "Range size" and "Insularity" (H3/H4), individual predictor ranking: "Insularity", rank 5, "Range size", rank 1, interaction: rank 5. The working residuals vary by the species' insularity. C: "Age of an accepted name" (H5), rank 2. The plots were prepared using the *effects* package (Fox et al. 2016).

# Appendix

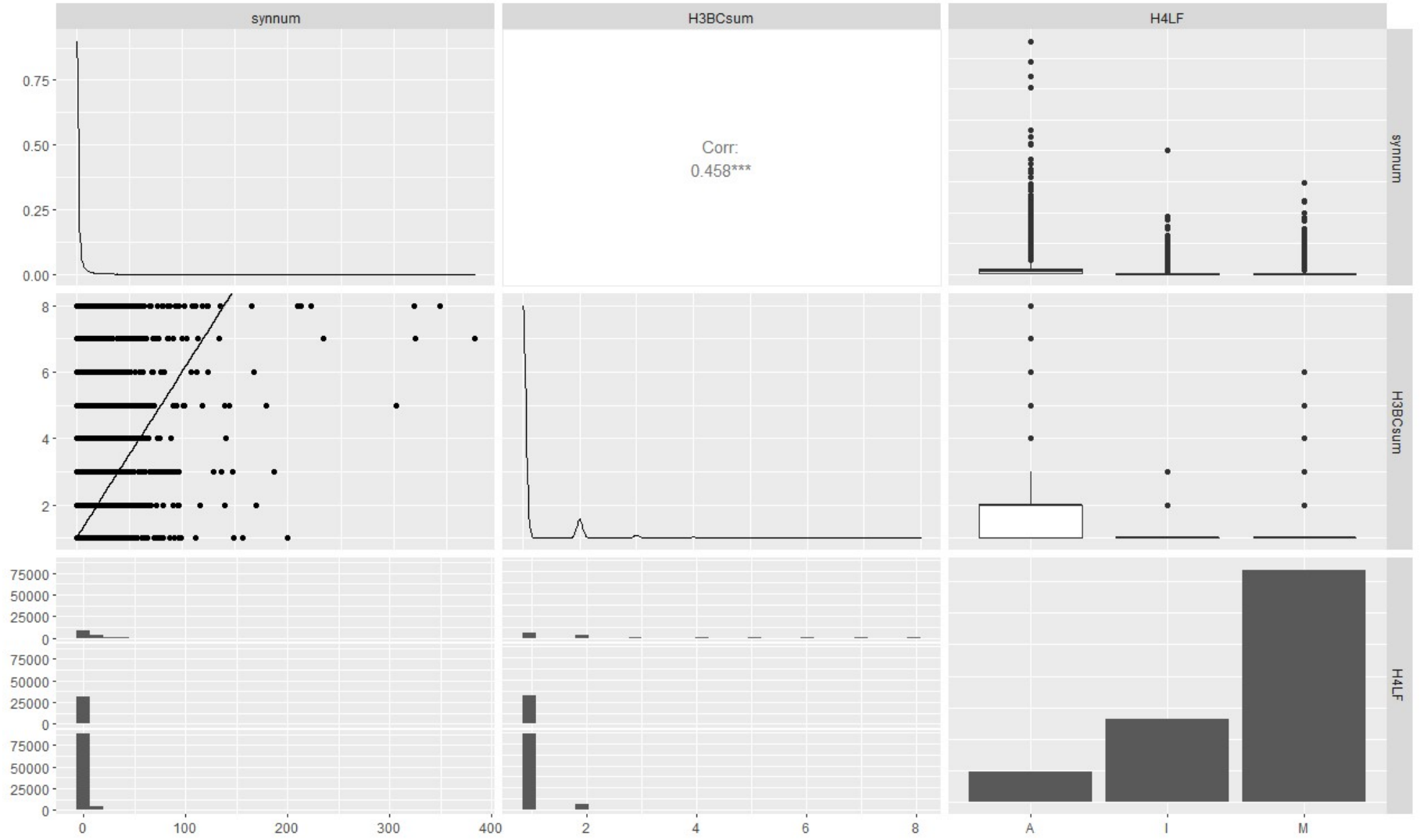



***Figure A1(a).*** **Correlation test of predictor pair "Number of botanical continents present" (H3BCsum) and "Insularity" (H4LF).** The correlation coefficient shows that the predictors are likely uncorrelated (0.458). (Plot: GGally package, ggpairs function, Schloerke et al. 2018).

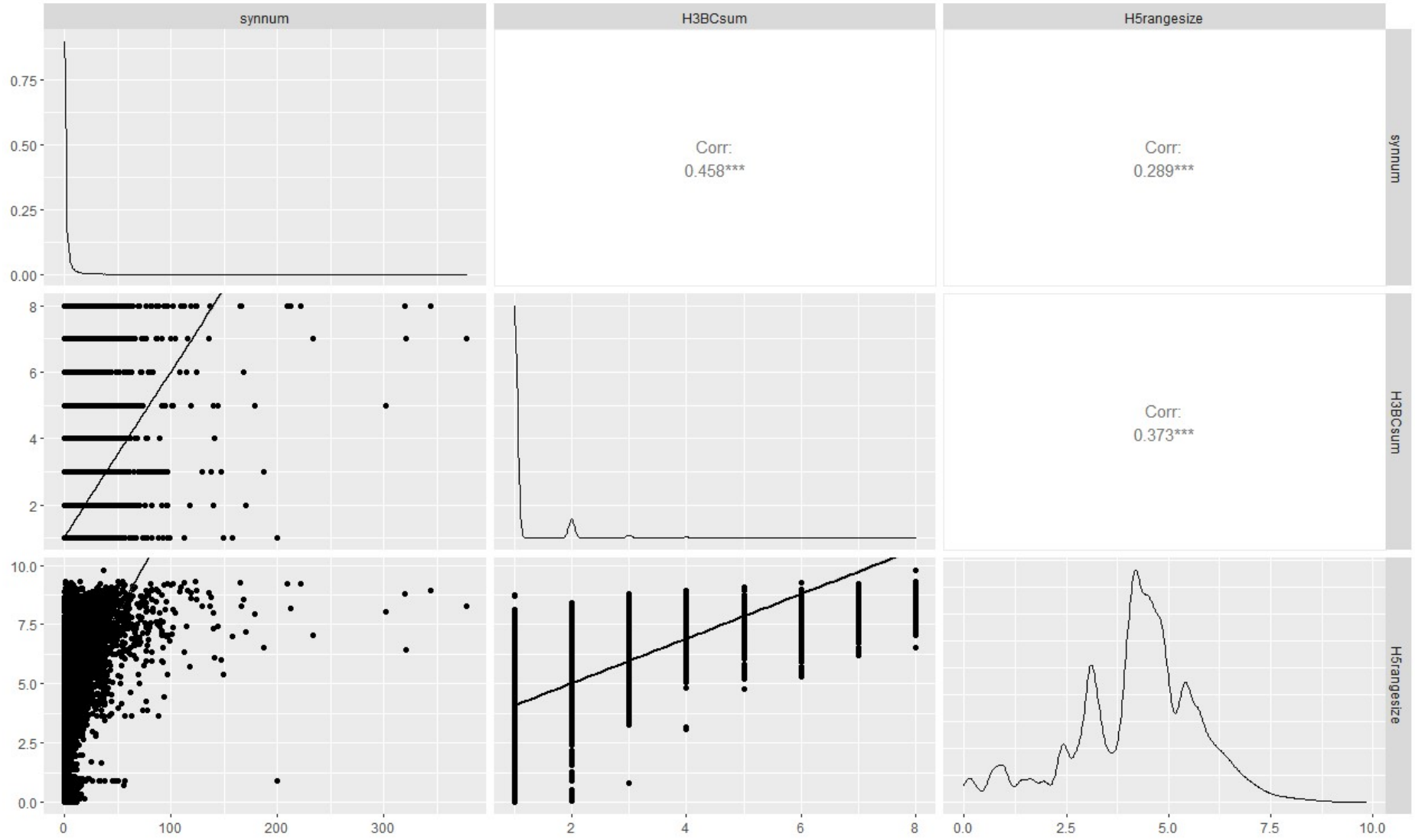



***Figure A1(b).*** **Correlation test of predictor pair "Number of botanical continents present" (H3BCsum) and "Range size" (H5Rangesize).** The correlation coefficients show that the predictors are likely uncorrelated.

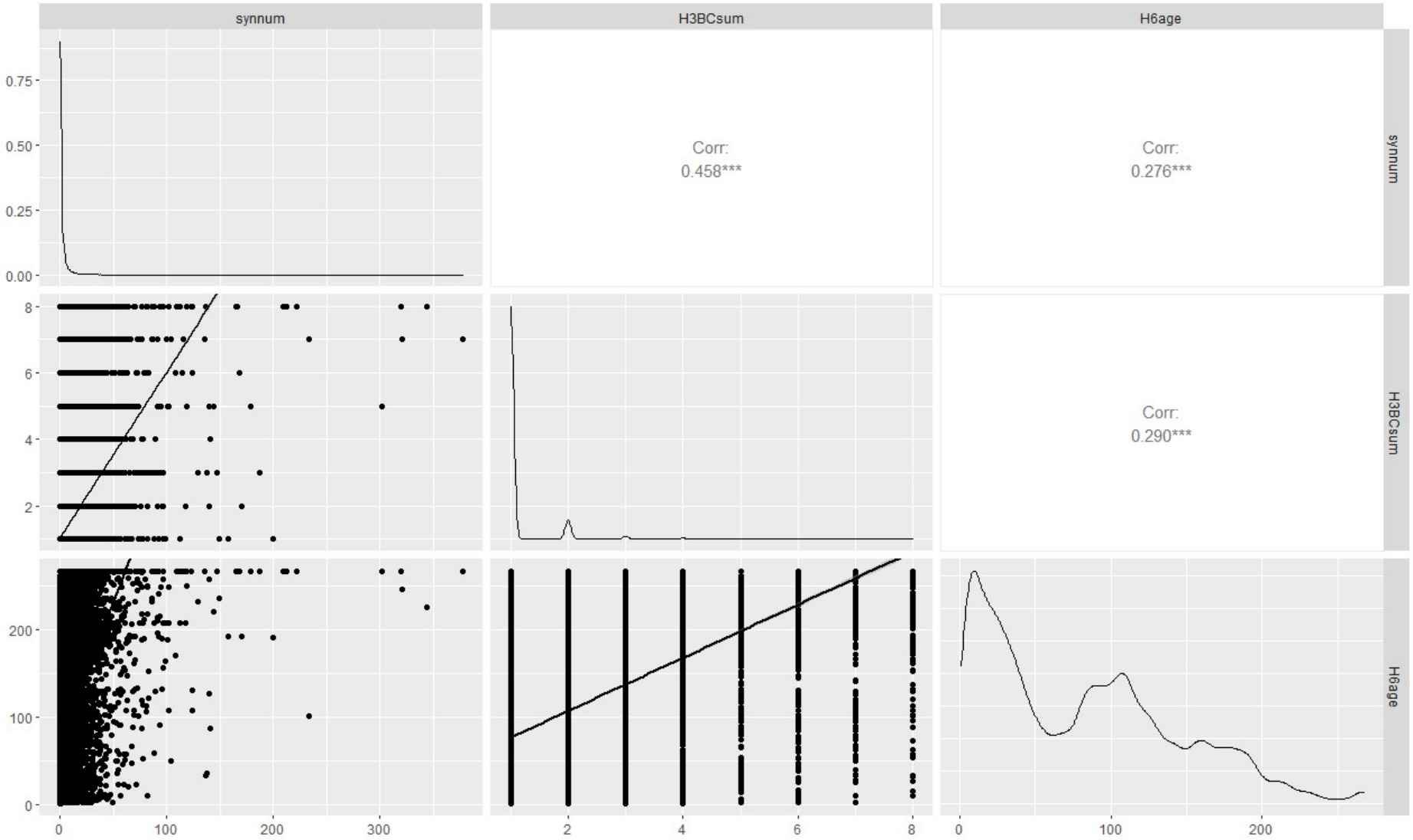



***Figure A1(c).*** **Correlation test of predictor pair "Number of botanical continents present" (H3BCsum) and "Age of a species' name" (H6age, as the proxy for the time passed since the publication of the accepted name).** The correlation coefficients show that the predictors are likely uncorrelated.

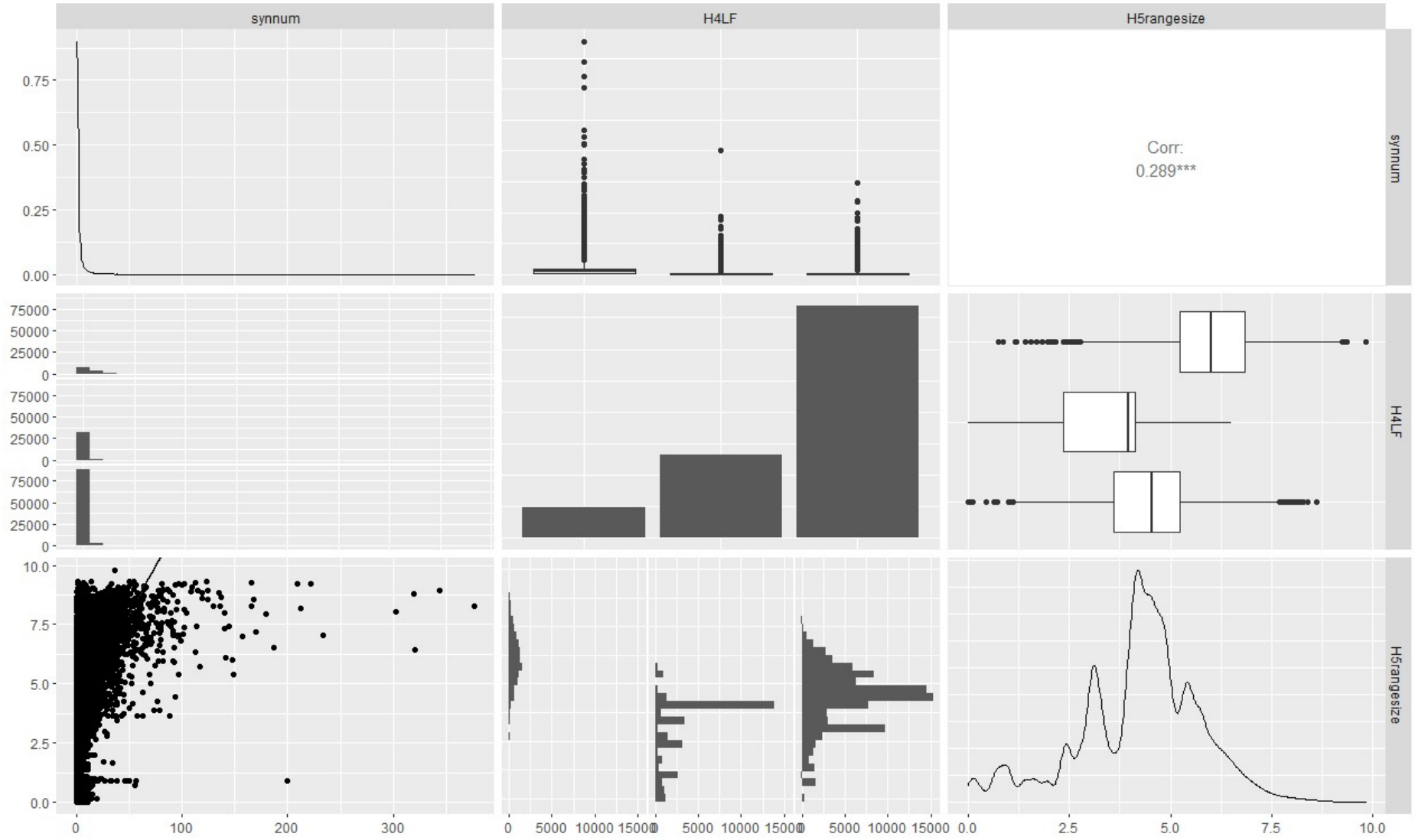



***Figure A1(d).*** **Correlation test of predictor pair "Insularity" (H4LF) and "Range size" (H5rangesize).** The correlation coefficient shows that the predictors are likely uncorrelated.

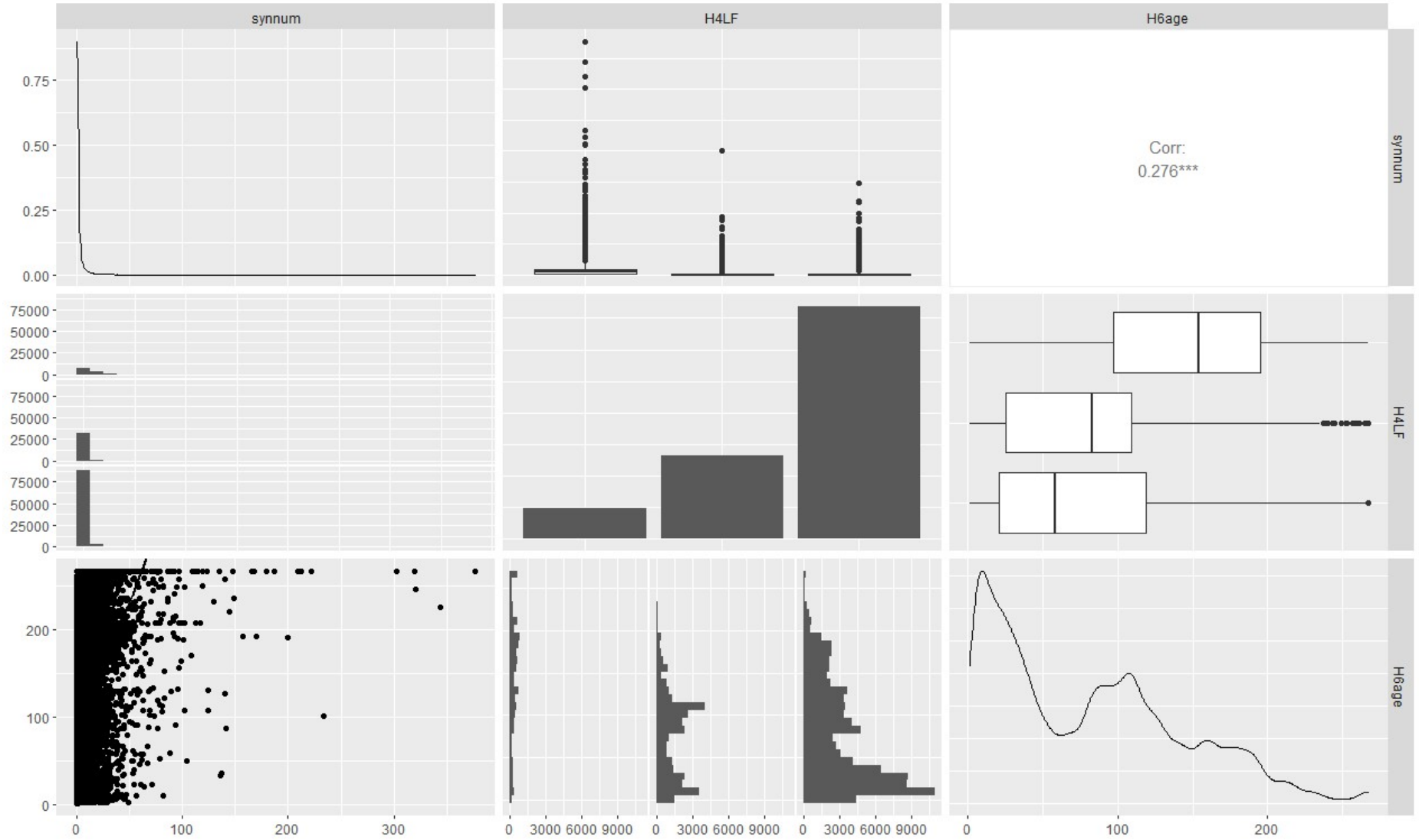



*Figure A1(e).* **Correlation test of predictor pair "Insularity" (H4LF) and "Age of a species' name" (H6age, as the proxy for the time passed since the publication of the accepted name).** The correlation coefficient shows that the predictors are likely uncorrelated.

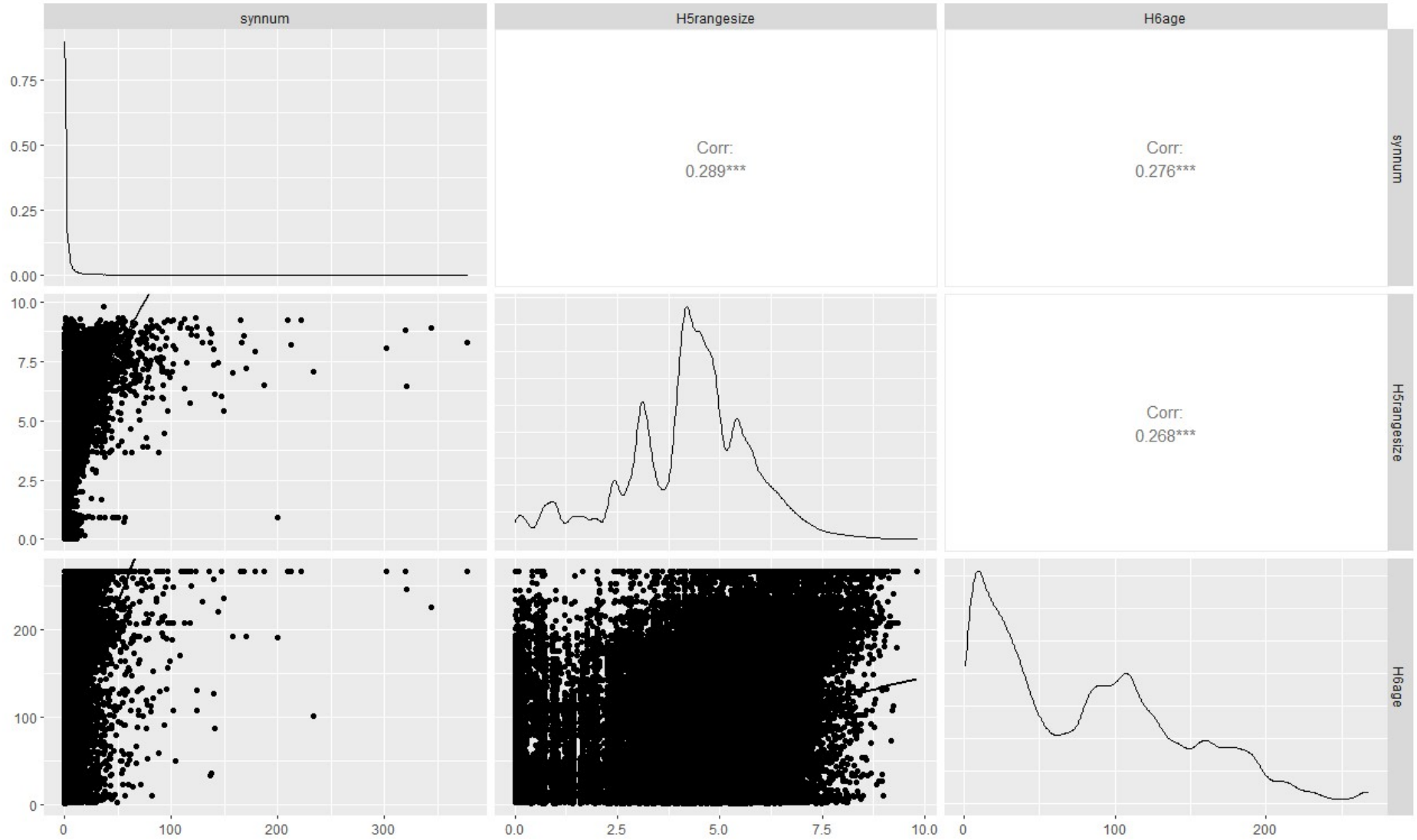
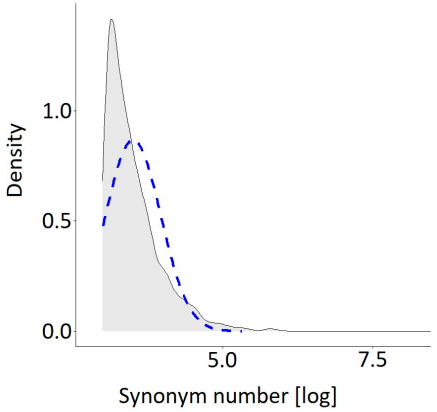



***Figure A1(f).*** **Correlation test of predictor pair "Range size" (H5rangesize) and "Age of a species' name" (H6age, as the proxy for the time passed since the publication of the accepted name).** The correlation coefficients show that the predictors are likely uncorrelated.

***Table B2.*** **(a) Density distribution of the log-transformed synonym number (*synNum*).** Observed distribution: grey; superimposed normal distribution: dashed blue. The *synNum* showed non-normal distribution with positive (right) skewness (skewness coefficient: 1.33, kurtosis: 4.68), indicating zero-inflation in the count data (Density plot: *ggdensity*, Kassambara 2020a). **(b) Analysis of potential count data issues and solutions.** Frequent issues to be handled in count data are zero-inflation and overdispersion. Table: Comparison of three suitable model-fitting methods to best handle the count data issues, using the final model parameters (*lme4*: Bates et al. 2015). The Poisson distribution, that included zero-inflation (*glmmTMB* package, Bolker 2016), showed the optimal model-fitting results in the *DHARMa* diagnostic tests (Hartig 2020).

**(a)**




**(b)**

| **GLMM** | **Poisson** | **Poisson/zi** | **negative binomial** |
|---|---|---|---|
| ***R* package** | *lme4*::glmer | ***glmmTMB*** | *lme4*::glmnb |
| **$R^2$ cond** | 0.939 | **0.958** | 0.705 |
| **$R^2$ marg (fixed)** | 0.398 | **0.414** | 0.361 |
| **AIC (E+05)** | 5.032 | **4.880** | 4.022 |
| **RMSE** | 3.948 | **4.005** | 4.882 |
| ***DHARMa* diagnostics** | | | |
| **KS test: deviation** | p = 0, sign. | **p = 0, sign.** | p = 0, sign. |
| **Dispersion: dev.** | p = 0.032, sign. | **p = 0.504, n. sign.** | p = 0, sign. |
| **Outlier test: dev.** | p = 0.004, n. sign. | **p = 0.454, n. sign.** | p = 0, n. sign. |

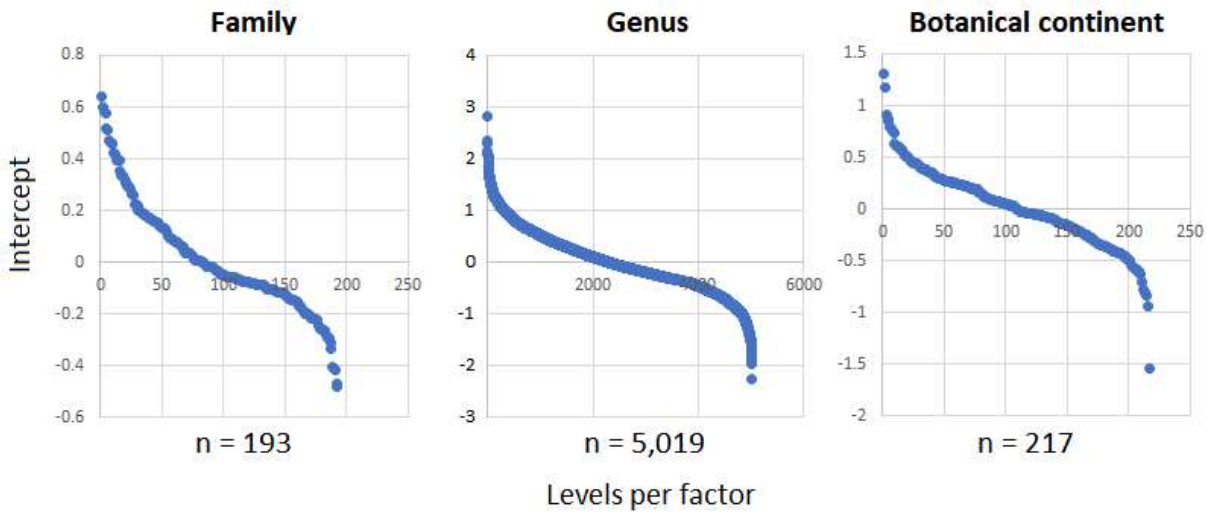



***Figure A3.*** **Random Factor selection: Taxonomic Family and Genus, and Botanical Continents, a Species is Present (Factors).** Each grouping factor per random factor has its own random intercept. We selected these predictor variables with high level numbers as random factor (McGill 2015). The three variables also exhibited low standard errors and very low *p*-values (< 0.001), suggesting that they were predictive.

***Tables B4, (a) and (b).*** **Performance diagnostics and model evaluation. (a)** Determining details of the model, by the explanatory variables, we found that model 4 minimized the Akaike information criterion, increased the pseudo-$R^2$s (Diagnostics: *jtools*, Long 2017, *sjPlot*, Lüdecke 2021), and reduced overdispersion and zero inflation below a significant threshold, as als shown in the final *DHARMa* diagnostic tests (Appendix, Figure A5). **(b)** Four global models of angiosperm synonymy. Selection conditions of the models were: (1) AIC at a stable minimum, and (2) a maximized pseudo-$R^2$. Result of GLMM of a combined nine-predictor model, by random factors and fixed factors. H1 to H5: Hypotheses (see: Table 1). ***, $p < 0.001$.

| (a) | GLMM | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|---|
| | **Performance parameters:** | | | | |
| | $R^2$ cond | 0.964 | 0.964 | 0.958 | **0.958** |
| | $R^2$ marg (fixed factors) | 0.421 | 0.440 | 0.396 | **0.414** |
| | Random factor share | 0.543 | 0.524 | 0.562 | **0.544** |
| | AIC (E+05) | 4.882 | 4.882 | 4.880 | **4.880** |
| | RMSE | 4.005 | 4.005 | 4.005 | **4.005** |
| | ***DHARMa* residual diagnostics:** | | | | |
| | KS test: deviation | p = 0, sign. | p = 0, sign. | p = 0, sign. | **p = 0, sign.** |
| | Dispersion: dev. | p = 0.008, sign. | p = 0.016, sign. | p = 0.252, n. sign. | **p = 0.504, n. sign.** |
| | Outlier test: dev. | p = 0, sign. | p = 0.230, n. sign. | p = 0, sign. | **p = 0.454, n. sign.** |

| (b) | Factor | Model 1 | Variation | Model 2 | Variation | Model 3 | Variation | Model 4 | Variation |
|---|---|---|---|---|---|---|---|---|---|
| | **Random Factor $R^2$ share** | **0.543** | | **0.524** | | **0.562** | | **0.544** | |
| | Botanical continents | 0.312 | 31.2% | 0.301 | 30.1% | 0.311 | 31.1% | 0.302 | 30.2% |
| | Genus of species | 0.231 | 23.1% | 0.223 | 22.3% | 0.231 | 23.1% | 0.223 | 22.3% |
| | Family of species | - | 0.0% | - | 0.0% | 0.020 | 2.0% | 0.019 | 1.9% |
| | **Fixed Factor $R^2$ (Marg.)** | **0.421** | | **0.440** | | **0.396** | | **0.414** | |
| | Range size | 0.214 | 21.4% | 0.215 | 21.5% | 0.202 | 20.2% | 0.201 | 20.1% |
| | Age of accepted name | 0.118 | 11.8% | 0.118 | 11.8% | 0.111 | 11.1% | 0.111 | 11.1% |
| | Insularity | 0.058 | 5.8% | 0.058 | 5.8% | 0.054 | 5.4% | 0.054 | 5.4% |
| | No. inhab. continents | 0.031 | 3.1% | 0.031 | 3.1% | 0.029 | 2.9% | 0.029 | 2.9% |
| | Range size * Insularity | - | 0.0% | 0.018 | 1.8% | - | 0.0% | 0.019 | 1.9% |
| | **Total $R^2$ (Cond.)** | **0.964** | | **0.964** | | **0.958** | | **0.958** | |

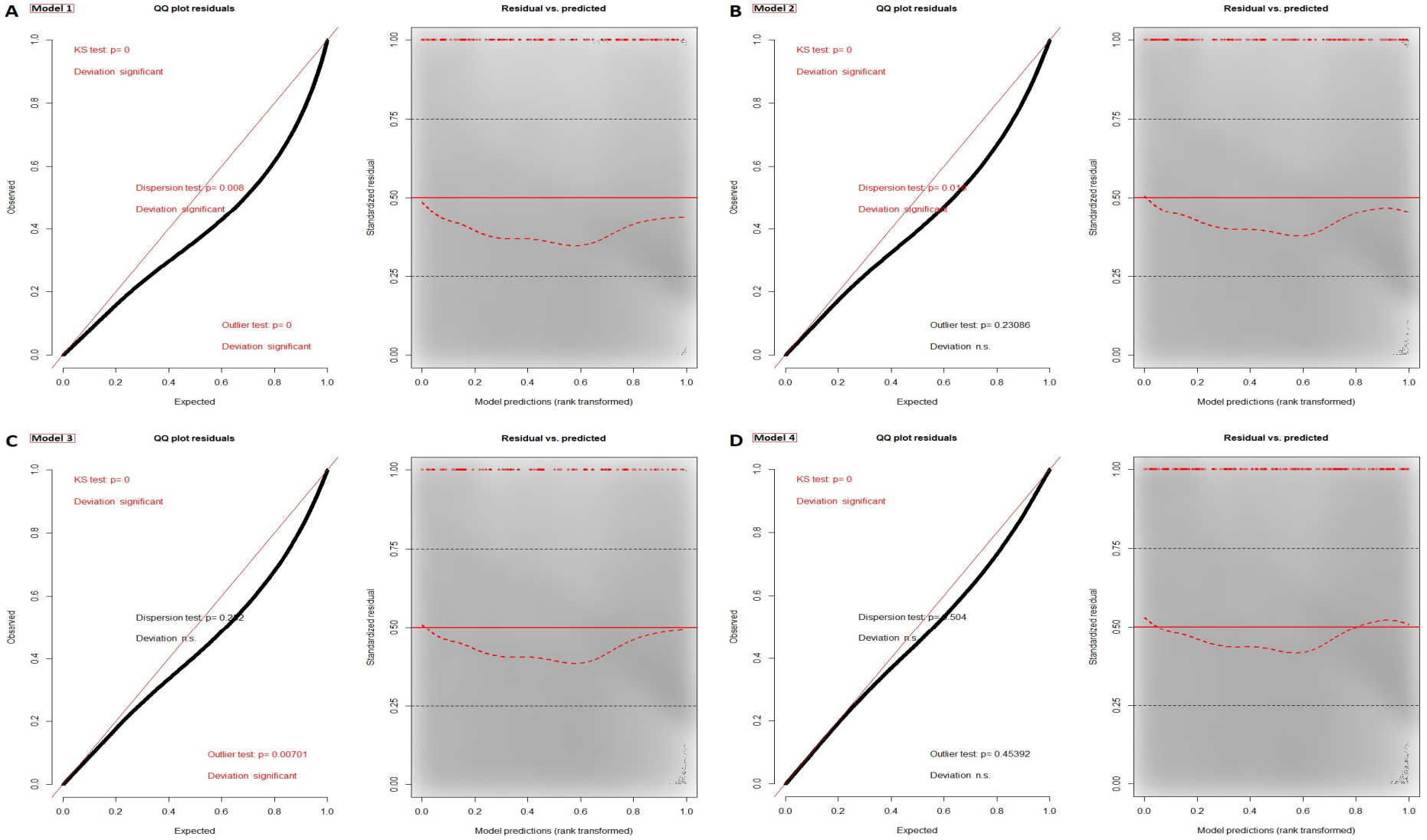



***Figure A5. DHARMa* diagnostic protocols (Hartig 2020) for the four best-performing models 1 to 4.** See also: Table 3 and Appendix, Tables A4, (a) and (b) for further details.

***Table A6(a).* 25 angiosperm families with the highest synRates.** Listed are families with more than 10 accepted names. Large families such as Orchidaceae (28,899 accepted names, rank 99), Rubiaceae (10,796 accepted names, rank 94), and Myrtaceae (5,778 accepted names, rank 96) rank in the middle, due to their synRate. **B6(b). 25 angiosperm genera with the highest synRates.** Notably, many of the genera in the table have only one or a few accepted species (names). Images: Angiosperm families (Table A6(a): A. *Canna generalis,* Pos. 1; B: *Potamogeton gramineus*, Pos.2) and genera (Table A6(b): C. *Rizinus communis,* Pos. 1; D: *Phillyrea angustifolea,* Pos. 2). Abbreviations: accNum: Number of accepted species. synNum: Number of synonyms. synRate: synonymy rate.


***Table B6(a).***

| Pos | family | accNum | synNum | synRate% | Ratio: synnum/accnum |
|---|---|---|---|---|---|
| 1 | Cannaceae | 12 | 238 | 95.20% | 19.8 |
| 3 | Potamogetonaceae | 106 | 790 | 88.20% | 7.5 |
| 5 | Ruppiaceae | 11 | 69 | 86.30% | 6.3 |
| 7 | Irvingiaceae | 12 | 67 | 84.80% | 5.6 |
| 8 | Paeoniaceae | 36 | 187 | 83.90% | 5.2 |
| 10 | Betulaceae | 172 | 846 | 83.10% | 4.9 |
| 11 | Stilbaceae | 21 | 100 | 82.60% | 4.8 |
| 12 | Juncaginaceae | 22 | 99 | 81.80% | 4.5 |
| 13 | Cornaceae | 103 | 463 | 81.80% | 4.5 |
| 14 | Typhaceae | 62 | 272 | 81.40% | 4.4 |
| 15 | Pontederiaceae | 45 | 193 | 81.10% | 4.3 |
| 17 | Alismataceae | 138 | 583 | 80.90% | 4.2 |
| 19 | Poaceae | 11540 | 47443 | 80.40% | 4.1 |
| 20 | Tofieldiaceae | 28 | 114 | 80.30% | 4.1 |
| 24 | Basellaceae | 19 | 76 | 80.00% | 4.0 |
| 26 | Fagaceae | 958 | 3757 | 79.70% | 3.9 |
| 28 | Oleaceae | 619 | 2162 | 77.70% | 3.5 |
| 29 | Plantaginaceae | 57 | 198 | 77.60% | 3.5 |
| 31 | Cymodoceaceae | 18 | 61 | 77.20% | 3.4 |
| 36 | Altingiaceae | 15 | 48 | 76.20% | 3.2 |
| 37 | Pandaceae | 17 | 54 | 76.10% | 3.2 |
| 39 | Juncaceae | 470 | 1460 | 75.60% | 3.1 |
| 40 | Bignoniaceae | 874 | 2710 | 75.60% | 3.1 |
| 41 | Melanthiaceae | 184 | 554 | 75.10% | 3.0 |
| 46 | Nothofagaceae | 38 | 113 | 74.80% | 3.0 |

*Table B6(b).*

| Pos | H1family | genus | recnum | synNum | synRate |
|---|---|---|---|---|---|
| 1 | Euphorbiaceae | Ricinus | 1 | 212 | 99.5% |
| 2 | Oleaceae | Phillyrea | 2 | 247 | 99.2% |
| 3 | Poaceae | Avenula | 1 | 83 | 98.8% |
| 4 | Arecaceae | Cocos | 1 | 56 | 98.2% |
| 5 | Apocynaceae | Nerium | 1 | 45 | 97.8% |
| 6 | Campanulaceae | Platycodon | 1 | 41 | 97.6% |
| 7 | Poaceae | Arctophila | 1 | 41 | 97.6% |
| 8 | Poaceae | Molinia | 2 | 77 | 97.5% |
| 9 | Lamiaceae | Mentha | 24 | 889 | 97.4% |
| 10 | Poaceae | Apluda | 1 | 36 | 97.3% |
| 11 | Poaceae | Taeniatherum | 1 | 34 | 97.1% |
| 12 | Poaceae | Vulpiella | 1 | 34 | 97.1% |
| 13 | Poaceae | Sasaella | 11 | 341 | 96.9% |
| 14 | Poaceae | Oplismenus | 7 | 212 | 96.8% |
| 15 | Myrtaceae | Blepharocalyx | 4 | 120 | 96.8% |
| 16 | Araceae | Pistia | 1 | 29 | 96.7% |
| 17 | Poaceae | Vahlodea | 1 | 27 | 96.4% |
| 18 | Potamogetonaceae | Stuckenia | 7 | 176 | 96.2% |
| 19 | Hydrocharitaceae | Hydrilla | 1 | 25 | 96.2% |
| 20 | Asparagaceae | Eustrephus | 1 | 25 | 96.2% |
| 21 | Potamogetonaceae | Groenlandia | 1 | 25 | 96.2% |
| 22 | Apocynaceae | Apocynum | 4 | 97 | 96.0% |
| 23 | Poaceae | Trachypogon | 4 | 97 | 96.0% |
| 24 | Poaceae | Dupontia | 1 | 24 | 96.0% |
| 25 | Poaceae | Ampelodesmos | 1 | 24 | 96.0% |

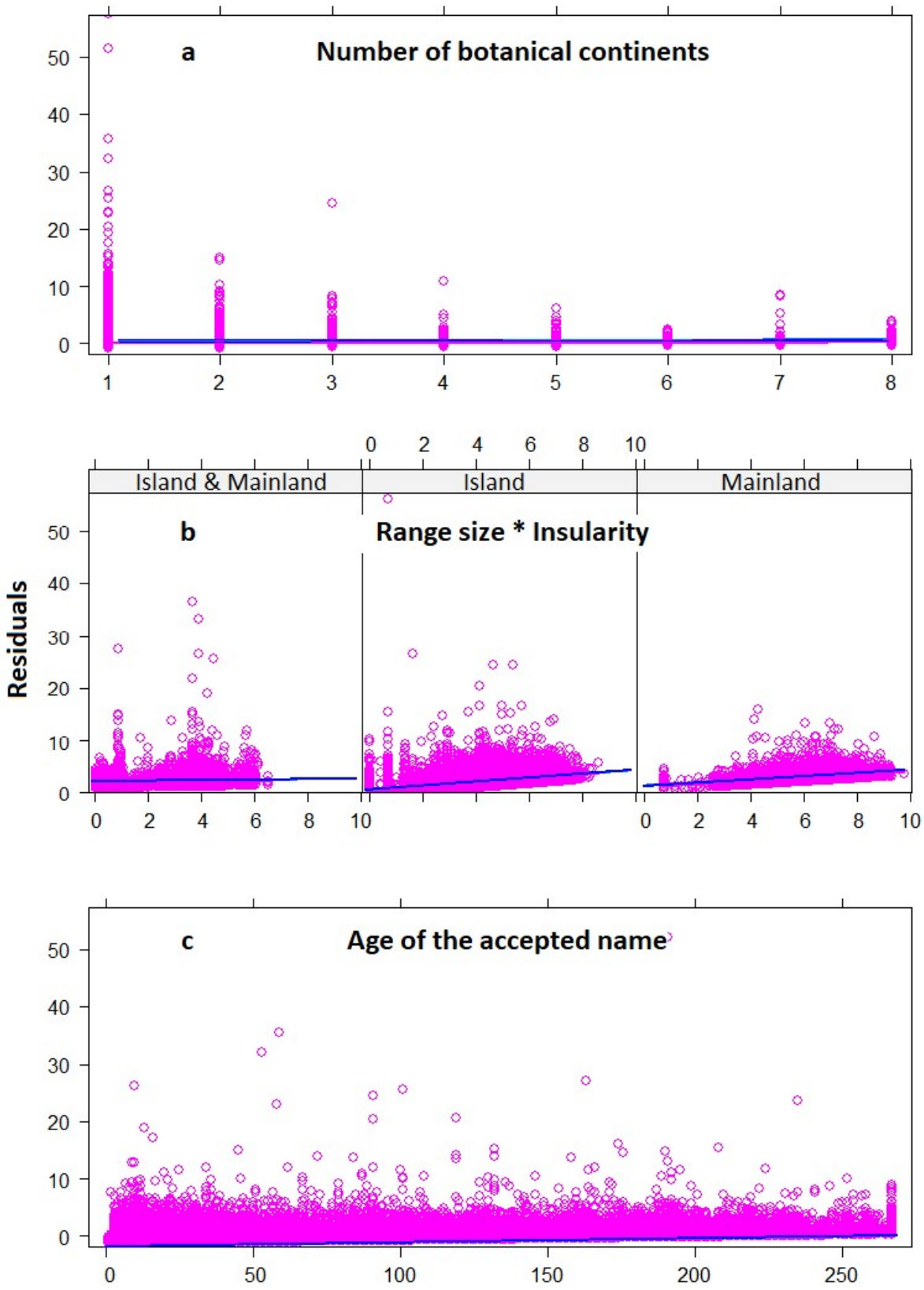



***Figure A7*. Working residuals of the global model of angiosperm synonymy**. (a) The working residuals vary with the *number of botanical continents, a species is present* (variable *BCNum*; (b), the working residuals vary within the different *insularities* and the *range size*, respectively, (c) The working residuals vary with the *age of an accepted name*. Details regarding the predictor rankings, see Table 3). All residual plots support the confidence intervals of the predicted regression lines. The plots were prepared, using the *effects* package (Fox et al. 2016).