OXFORD

BioScience, 2024, **0**, 1–9
https://doi.org/10.1093/biosci/biae042
Advance access publication date: 0 2024
Overview Article

The omission of seed banks in demography as an example of bias in ecology

Federico López-Borghesi and Pedro F. Quintana-Ascencio

Federico López-Borghesi (federico.borghesi@ucf.edu) is a postdoctoral research fellow and Pedro F. Quintana-Ascencio (pedro.quintana-ascencio@ucf.edu) is a professor in the Department of Biology of the University of Central Florida, in Orlando, Florida, in the United States.

Abstract

Despite enthusiasm for big data in the life sciences, challenges arise because of biases and incomplete data. Demographic studies often overlook dormant life stages, which can skew inferences. They also tend to focus on few populations and short time spans. We assessed omissions of seed banks in demographic studies, exploring trends across life forms, climates, and taxonomic groups. We compared 172 species (192 cases) with independent seed bank and demographic studies. Approximately 25% of the demographic studies excluded known seed bank stages. The probability of omissions was lower for annuals and shrubs and higher for perennial herbs. We found no evidence that ecoregion or phylogeny explained these omissions. Modeling choices and study designs may explain patterns of seed bank omissions. Considering more populations reduced the chance of omissions. Omissions raise concerns for ecological analyses using databases. Leveraging large data is important, but we must be careful to understand their biases and limitations.

Keywords: demographic studies, big data, seed banks, biases, environmental variation

Rapidly advancing technology has radically increased the volume and accessibility of data across a diverse range of fields. Within the business sector, in particular, big-data approaches have empowered companies to monitor products in great detail (e.g., Praveen et al. 2020), gauge the effectiveness of marketing with precision (e.g., Aljumah et al. 2021), and even begin to forecast customer behavior (e.g., Kachamas et al. 2019). Naturally, these achievements have sparked enthusiasm for the potential application of similar approaches to an ever-growing array of disciplines, including the life sciences (e.g., Bouwmeester et al. 2019, Gobeyn et al. 2019). However, the attributes of data accessible to life science research contrasts markedly with what is available to businesses. The crucial distinction is one of data completeness. Although, in most business operations, data sets include complete lists of transactions, website visitations, and so on, in the realm of life sciences we must contend with incompletely collected data representing historical biases.

The recent development and proliferation of technology has enabled researchers to collect, store, and analyze vast amounts of ecological data more efficiently than ever before. Remote sensing, GPS tracking, sensor networks, and other data collection methods have provided a wealth of information about various ecological phenomena. Informatics and communication technologies have made it possible to store and manage large data repositories. In ecology, this has led to a proliferation of databases compiling and standardizing a broad range of information, including species traits, biodiversity measurements, animal movement, and other phenomena (e.g., Edwards et al. 2000, Iversen et al. 2017, Kays et al. 2022). These databases can contribute to a deeper understanding of ecosystems, biodiversity, and the impact of human activities on the natural world. The rise of big data and data analytics has enabled researchers to extract meaningful insights from large and complex ecological data sets. Machine learning and advanced statistical techniques can help uncover patterns, relationships, and trends that might not be apparent through traditional methods (Schmaljohann et al. 2012, Flack et al. 2016, McCormack and Iversen 2019). Furthermore, ecological databases have been integrated into decision support systems that aid in making informed choices related to such things as land use, conservation prioritization, and resource management. Despite the benefits, there are challenges associated with ecological databases, such as data quality control, standardization, and data privacy concerns. Ensuring that data are accurate, reliable, and up to date is crucial for maintaining the integrity of these databases.

We should be cautious of conclusions derived from analyses of incomplete data sets with undisclosed and sometimes unidentifiable biases. The potential risk of biases on big data analysis has already been described in health research (e.g., Kaplan et al. 2014) and in the social sciences (e.g., Raub 2018). These considerations should not dissuade researchers from engaging with data-driven inquiries. Instead, they should serve as a reminder to be critical of the sources of the data and test the assumptions behind them.

Seed banks as an example of bias in data

Within population ecology, researchers have long been aware of several biases and limitations in the application of demographic models. Most notably, dormant life stages—particularly those that are hard to detect—have been ignored with relatively high frequency (Doak et al. 2002, Nguyen et al. 2019). This type of life stages can occur across a broad taxonomic range and includes hibernating animals, dormant cysts and eggs, and, within plants, dormant individuals, and dormant propagules in the soil and within fruits and cones. Furthermore, demographic studies rarely explore the range of conditions necessary to quantify the full

Received: January 16, 2024. **Revised:** April 2, 2024. **Accepted:** April 16, 2024

impact of dormancy. Within a population, individuals go through various life stages, starting from propagules and progressing to reproductive adults. Each stage plays a distinct role in the population's persistence and is affected by different environmental factors. Failing to consider any of these stages can lead to inaccurate inferences (Caswell 2000; e.g., López-Borghesi et al. 2023). A large portion of plant species possess dormant seed banks that can potentially buffer against adverse environmental conditions. The demographic patterns of seed banks can be extremely challenging to quantify, particularly over the relatively short time frames typical of field studies. Under specific conditions, usually as populations recover from disturbances, the relevance of the seed banks becomes more apparent.

Seed banks play a key role in maintaining the demographic and genetic viability of numerous plant species. They may serve as a buffer against environmental adversity, provide a record of the vegetation history, and wield significant influence over future vegetation composition, particularly after disturbances (Warr et al. 1993). The soil seed bank is composed of seeds that remain viable in and on the soil for an extended period of time, sometimes spanning multiple years (Simpson et al. 1989). They are often categorized as either persistent and transient on the basis of whether the seed remains viable for longer than a year. Thompson and colleagues (1998) proposed to further divide the persistent seed bank into short term (lasting 1–5 years) and long term (lasting over 5 years). Regardless of the classification, only a fraction of a seed bank undergoes germination at any time; this may represent a bet-hedging mechanism that reduces the risk during unfavorable years (Venable 2007). Seed banks can potentially prevent the extinction of a population even after the aboveground individuals have been eradicated (Stöcklin and Fischer 1999). Seed dormancy has also been found to promote species coexistence in fluctuating environments (Nathan and Muller-Landau 2000). Given the significant implications to population dynamics, overlooking the seed bank from demographic models could be very problematic (Kalisz and McPeek 1992). Despite this, past reviews of the literature have shown that seed banks are excluded without justification in over 40% of published demographic studies (Doak et al. 2002, Nguyen et al. 2019).

In some cases, excluding a seed bank stage from demographic models might be justified by empirical evidence (Logfet et al. 2017) or by the known absence of dormancy within the species. However, in most cases, the omissions do not seem to be justified (Doak et al. 2002). These omissions might be the consequence of study limitations but can have profound consequences—for instance, in predicting the long-term success of restoration efforts (López-Borghesi et al. 2023). In an already established population, a failure to include the seed bank can lead to an underestimation of viability or a failure to capture its ability to recover from disturbances.

Besides the unjustified exclusion of seed banks, demographic studies often fail to include the variability in environmental conditions necessary to estimate their impact. Most demographic studies are short term (more than 3 years), and they concentrate on few areas (1–2 years) and in conditions where aboveground individuals are numerous enough to facilitate data collection (Salguero-Gómez et al. 2015). This trend challenges the possibility of documenting more realistic long-term dynamics and extrapolation to other areas (Crone et al. 2011). The limited variation represented in these studies may obscure the long-term effects of seed bank in the life history of the species. For example, conditions that increase survival and longevity of established individuals but reduce recruitment may reduce the expression of seed banks. Low levels of disturbance are not conducive to regeneration from seeds. Higher levels of disturbances create gaps in the vegetation and tend to promote increased recruitment from the seed bank (Fenner and Thompson 2005). As the intensity of disturbances increases, the role of the seed bank becomes more prominent. Indeed, current theory stipulates that areas with higher levels of disturbance contain a higher proportion of species with persistent seed banks (Grime 2006). However, it is important to remember that belowground dynamics tend to require longer time spans than aboveground ones (Ma et al. 2020). When evaluating the conclusions of a demographic analysis, then we should consider whether it accounts for the breadth of conditions necessary to observe the role of the seed bank.

As we stated earlier, seed bank dynamics can be difficult to study. Properly accounting for them in demographic models, however, is essential to obtain accurate projections—an issue of increasing importance under conditions of rapid environmental change and disruption of disturbance patterns. Regardless, it appears that published demographic models often assume an absence of persistent seed banks or downplay their importance. In the present study, we assess the rate at which published demographic studies omitted seed bank data when these data were available through a second independent study. In particular, we address the following questions: Are omissions of seed banks more pervasive in the demography of certain life forms, climates, or taxonomic group? Are omissions less likely in demographic studies with longer duration and evaluating greater number of conditions? The omission of seed banks in demographic studies has been documented before (e.g., Doak et al. 2002 and Nguyen et al. 2019), but the possible causes of its prevalence were not investigated enough. To address this, we document the occurrence or lack of occurrence of seed banks only for species with independent available assessments. We argue that the omission of seed dormancy and probably other vital rates can be explained by human oversight of effects temporal and spatial heterogeneity on population dynamics likely prompted by systemic limitations of graduate and funding programs. Although it can be argued that negative impact of these omissions may vary among studies depending on their original objectives, the use of demographic models lacking accurate accounts of the organism's life history in data conglomerates most probably misinform overall patterns.

Compiling data: Seed bank studies and plant demography

We compared the inclusion of seed bank stages in plant demographic models with studies independently documenting their occurrence for each focal species. We assembled a list of species with both types of information available. We began by finding common elements across two existing databases: the COMPADRE Plant Matrix Database and the LEDA Traitbase. Like the databases listed before, these two repositories compile and standardize information from hundreds of studies and have been used to explore global patterns and advance ecological knowledge.

The COMPADRE Database is an open-source online repository currently hosting over 8900 matrix population models for more than 790 plant species worldwide (Salguero-Gómez et al. 2015). The elements of these matrices classify individuals into different stages, including a propagule class, which allows to readily identify studies that include seedbanks. This database has already contributed significantly to expanding our understanding of population ecology. Salguero-Gómez and colleagues (2016) leveraged data from COMPADRE to develop a framework for predicting

plant life-history strategies on the basis of growth form, habitat characteristics, and phylogenetic relationships. This seminal work found that much of the variation in these strategies can be explained by whether the organism is fast or slow growing (fast–slow continuum) and its reproductive strategy. Multiple studies have since expanded on these findings, exploring the interplay of life history strategies with other factors, including responses to climate (Compagnoni et al. 2021) and another usually neglected ecological process—seed dispersal (Beckman et al. 2018). This framework has even been used to test theoretical questions about demographic processes, such as the effects of temporal autocorrelation (Paniw et al. 2018). These are just a few examples of how data hosted in COMPADRE has helped expand our understanding of life-history strategies and their evolution.

The LEDA Traitbase compiles life-history traits of plant species from Northwest Europe (Kleyer et al. 2008). It contains information on 26 plant traits related to persistence, regeneration, and dispersal for over 3000 species. Among the regeneration trait data, it contains over 40,000 observations on seed bank type and seed longevity collected via a mixture of literature compilations and field experiments. Following a strict protocol, the LEDA Traitbase classifies seed banks into long-term persistent (more than 5 years), short-term persistent (1–5 years), and transient (less than 1year). This repository has also had a profound impact in furthering our ecological knowledge. For instance, Fry and colleagues (2018) leveraged data from LEDA and other databases to aid in a field-based experiment to show the link between belowground functional traits—that is, root architecture and depth—and soil function during grassland restoration. Studies using data on seed characteristics have also increased rapidly with the advent of databases such as LEDA (e.g., Jiménez-Alfaro et al. 2016).

By combining the information available in COMPADRE and LEDA, we were able to assemble an initial list of 111 species that have both see bank data and demographic models. This list was expanded to 172 through a systematic search of the literature (for a list of species, see the supplemental material). We followed the same criteria as the LEDA Traitbase to incorporate new seed bank studies (Kleyer et al. 2008). In addition, we extracted data from the databases and literature regarding the estimated seed longevity, growth form (e.g., woody shrub), and climate (e.g., tropical) for each species. Because some species possess multiple published demographic studies, our final list contained 192 instances comparing seed bank measurements with their representation in matrix population models. For each of the corresponding demographic studies, we reviewed the manuscripts to retrieve information on the duration of the study, the range of environmental conditions (e.g., type of habitat, sites, and experimental treatments), and the number of individual matrices.

Data analysis and visualization

All generalized linear regressions presented in this article were fitted using Bayesian inference. We used binomial distributions for binary data and negative binomial distributions for count data. We conducted all statistical analyses using R version 4.2.3 (R Core Team 2019) and version 2.21.0 of Stan (Carpenter et al. 2017, Stan Development Team 2018). We used the "geiger" R package (Harmon et al. 2007) to calculate Pagel's λ to search for potential phylogenetic patterns in seed bank type. For to calculate the D statistic used for estimating the phylogenetic pattern in literature disagreements, we used the "caper" package in R (Orme et al. 2013). All visualizations were performed using base R language with the exception of phylogenetic trees, which were constructed using the "diversitree" package (FitzJohn 2012).

Patterns of omission: Demographic studies without known seed bank stages

We used the information from studies on seed bank formation as our assumption for the presence of this stage. The comparison between the corresponding elements in these data bases can have four outcomes: a species has a seed bank, and it is included in the matrix (73 positive agreement cases in our study); the species has no seed bank, and the matrix doesn't include it (60 cases of negative agreement); a seed bank isn't included in the matrix for a species that has it (omission in 49 cases; i.e., the subject of this article); and a seed bank is included for a species that doesn't have it (unexpected inclusion in nine cases). We found that seed bank was omitted in approximately 25% demographic studies for species with evidence of seed banks. Although this is lower than the proportion of unjustified seed bank omissions found in other studies (Doak et al. 2002, Nguyen et al. 2019), it still represents a large bias in demographic data. We recognize that incorporating seed banks can be a complex task, mainly because of reliance in supplementary experiments needed to estimate factors such as seed survival and germinability (Lesica and Steele 1994). Although sometimes these exclusions might be justified, as in the case of clonal species (Logofet et al. 2017); in most cases, these omissions can have both theoretical (Nguyen et al. 2019) and practical consequences (López-Borghesi et al. 2023). When evaluating the potential effect of these biases on large-scale analysis, an important question to make is whether specific patterns exist in the exclusion of seed banks.

Ecoregion and growth forms

A previous study of 10,170 angiosperm plant species representing 305 families and covering all world biomes found that 82% of the studied species had dormant seeds and concluded that seed dormancy most likely maximizes plant recruitment in habitats with variation in environmental suitability (Rosbakh et al. 2023). After cautioning of a relatively low predictive power of the studied climatic variables, these authors suggest that physiological dormancy may be prevalent in dry biomes with high temperature seasonality, whereas physical dormancy can be frequent in biomes with high seasonal temperature and precipitation fluctuations. Nondormancy was likely common in stable, warm, and wetter climates. Surprisingly, Rosbakh and colleagues (2023) did not find an effect of pyroclimate but recognized that the binary character used to assess dormancy (0, absent; 1, present) did not include fire severity or frequency that may have obscured any patterns, and they encouraged further work.

Considering these findings, our expectations were that omissions of seed banks in demographic models would be less common for species growing in harsh environments (e.g., deserts) and in environments more prone to disturbances (e.g., shrublands and grasslands) because is more likely that they will be observed and less common in relatively more stable environments (e.g., tropical and temperate forests) because dormancy is less prevalent. Likewise, omissions may be more common for demographic models of species with slower life cycles (e.g., perennial species) because long term studies may be necessary to document recruitment. We evaluated the probability of seed bank omission across ecoregions and across growth forms.

It is worth noticing that our data might be biased, because different ecoregions and growth forms were unequally represented in the compiled list—with about 51% of species from temperate mixed forests and 67% classified as herbaceous perennials. For this analysis, it was necessary to remove *montane* from ecoregions,

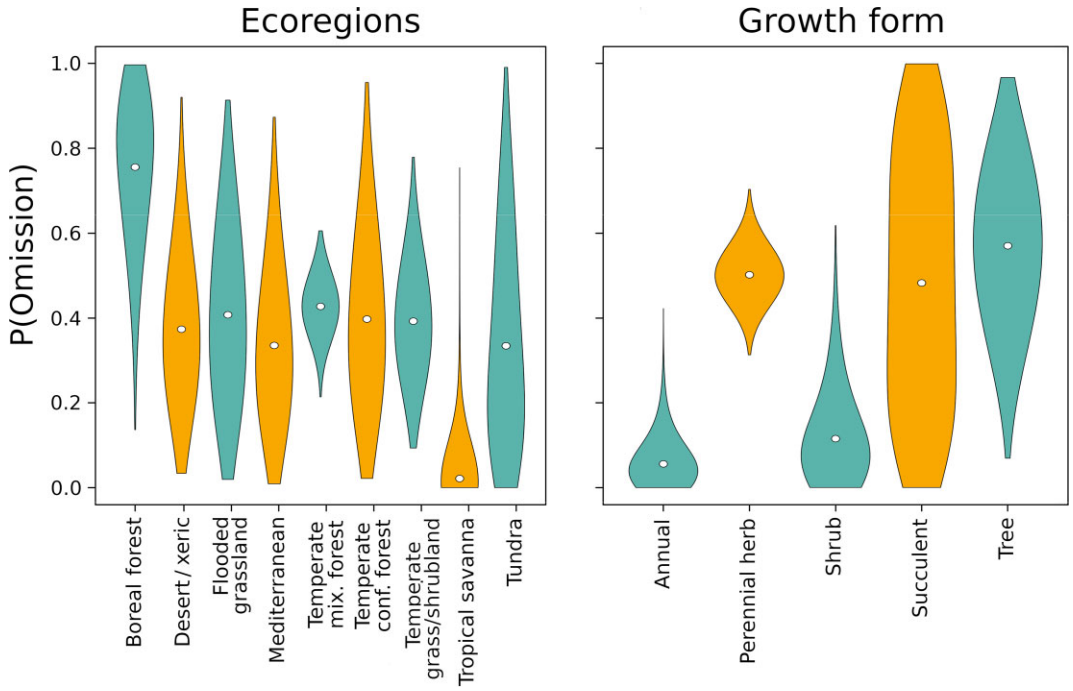

Figure 1. Violin plots of the posterior predicted distributions of the probability of omission between seed bank studies and their inclusion in demographic models for studies across different ecoregions (left) and plants with different growth forms (right).

as well as *palms* from growth forms, because they were heavily underrepresented (McHugh 2013).

We found some evidence that ecoregion affected the frequency of seed bank omissions in plant demography. A simple visualization of the model output shows that, except for boreal forest and tropical savannas, omissions occur in similar proportions across different ecoregions (see figure 1). Probability of omissions was more prevalent in boreal forests and less prevalent in tropical savannas, an ecosystem with frequent fires that may promote seed banks.

There was clear support for differences in omissions across growth forms. Examination of the probability of omission among the data shows annual species and shrubs had low probability of omission, whereas perennial herbs had higher probability of omission (figure 1). We did not find clear patterns for trees and succulents.

Phylogenetic signals

Certain seed traits are known to be phylogenetically conserved, including some dormancy mechanisms (Holye et al. 2015) and morphometric characteristics (e.g., Süngü Şeker et al. 2021). Because the physical characteristics of the seeds could make them harder to detect in the soil, we hypothesized that seed bank omissions in demographic studies might be more prevalent within certain taxa. To measure the phylogenetic signal in our data, we used a phylogenetic tree that was cropped from a tree provided by the COMPADRE team (Salguero-Gómez et al. 2016). This tree covers 163 of the 172 species being analyzed, allowing for estimating whether omissions of seedbanks in demographic studies are more prevalent in certain clades. The expectation is that omissions may be more prevalent in groups that are more difficult to study because of the size, shape, and color of the seeds.

It was first necessary to determine if species displaying seed persistence were phylogenetically clustered, biasing the analysis of discrepancies in the literature. To achieve this goal, we used the Pagel's λ index—a multiplier of internal branches (off-diagonal elements of a variance–covariance matrix) designed to find the best fit for the data (Pagel 1999). This index can take values between 0 (phylogenetic independence) and 1 (evolution according to Brownian motion). Pagel's λ was chosen because it minimizes the risk of type I error, so that it provides a strong method for detecting phylogenetic independence (Münkemüller et al. 2012).

The temporal persistence of seed banks (classified as long or short term, persistent or not persistent) showed an extremely weak phylogenetic signal (Pagel's $\lambda << .0001$). In other words, the formation of persistent seed banks does not seem to be associated with particular taxonomic groups (see figure 2). Although this might be surprising, because the literature suggests seed dormancy mechanisms are phylogenetically conserved, persistence in the soil can be achieved through multiple mechanisms (Long et al. 2015). Honda (2008) claimed that seed longevity of nondormant species has been underestimated and his analysis indicated that dormancy may not be essential for the formation of seed banks. In addition, conservation of traits might happen at a smaller phylogenetic scale than the scope of our study.

In order to test whether agreements between literature sources are more prevalent in certain taxa, we used the *D* statistic developed by Fritz and Purvis (2010). This approach uses the sum of differences between sister clades to estimate the strength of the phylogenetic signal in binary traits. Low numbers of the *D* statistic indicate clumping expected under Brownian motion, whereas larger values indicate random association. We ran 20,000 permutations to estimate the potential *D* value under phylogenetic randomness (randomly shuffling the tips) and under a Brownian threshold model (traits evolve along the phylogeny) and compared the observed value with these baselines.

A discrepancy (or a lack of agreement) between seed bank studies and demographic models also seemed to follow a random

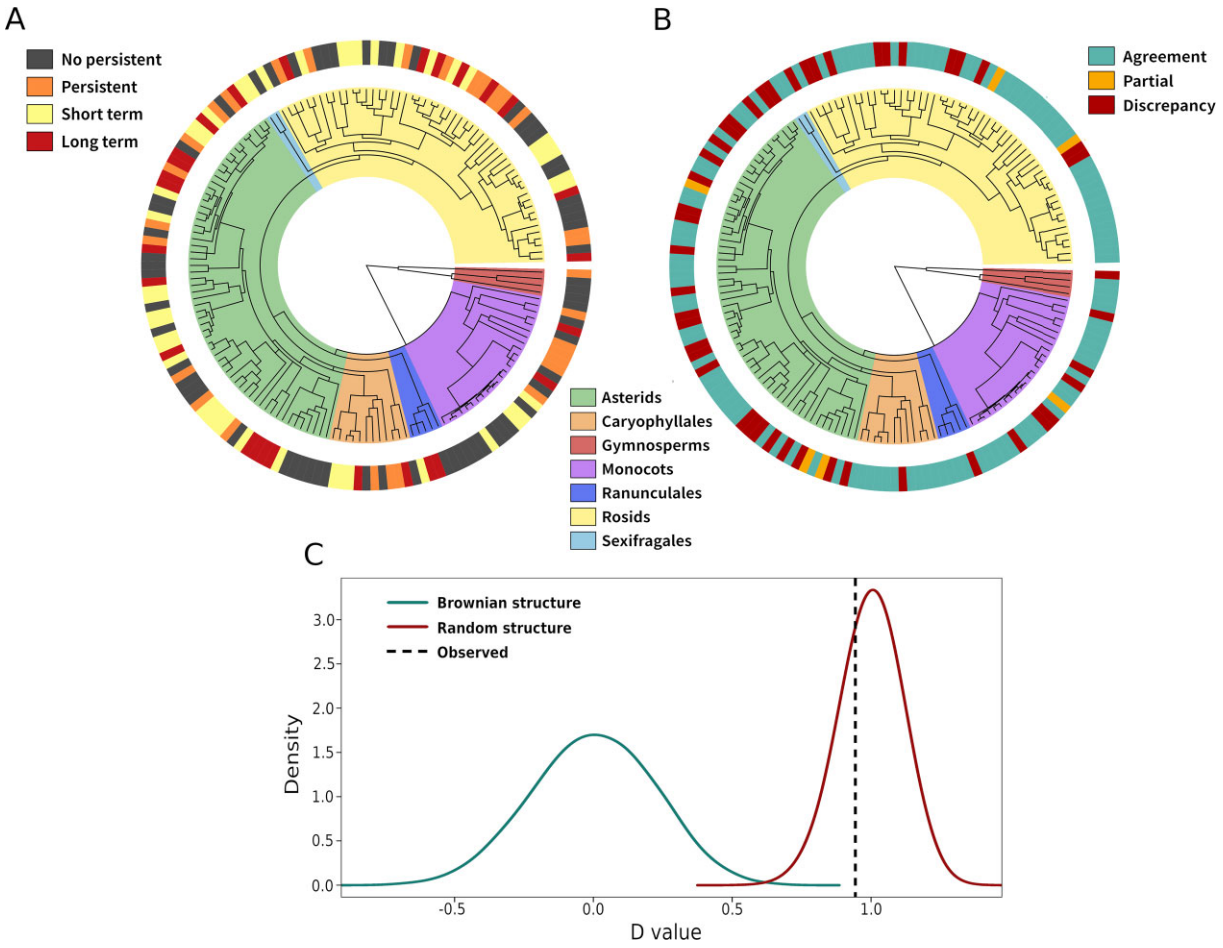

Figure 2. (A) Distribution of different types of seed bank (transient, persistent, short term, and long term) identified for the 162 study species represented in the phylogenetic tree. (B) The distribution of agreements and discrepancies between the representation of seed banks in demographic models and published direct studies of the seed bank—partial represents species where more than one demographic study was conducted by only part of them agreed with the seed bank literature. (C) The result of a ***D*** statistic test to measure the phylogenetic signal in the distribution of agreements and discrepancies—the distribution of the random structure hypothesis and the Brownian structure hypothesis are the result of 20,000 permutations.

phylogenetic structure (estimated $D = .944$). A visual review of the distribution of disagreements in the phylogenetic tree (figure 2) shows that they occur sporadically across every major taxonomic group, including basal groups such as gymnosperms. As seen in figure 2, the estimated value for observed data falls just below the mean value of the probability distribution $E(D)$ for the permutations with random structure—and clearly above the probability distribution $E(D)$ following Brownian motion.

The lack of evidence of phylogenetic signal in the discrepancy between literature sources suggests that omissions of seed bank in plant demography are likely related to modeling choices and study design. Although, in some cases, these limitations are related to sampling complexity (Crone et al. 2011), it is often related to issues of temporal and spatial scale of the study. Those that focus on shorter intervals or only include more favorable environmental conditions may be more prone to overlook the presence of seed banks.

Constructed matrices with variable spatial and temporal extent and heterogeneity

Environmental variation is essential for describing and understanding life history traits (Quintana-Ascencio 2023). Seeds are highly responsive to changes in environmental conditions, and these variations can play a major role in shaping seed bank dynamics. Some species might be limited by their life history traits—such as those growing in ephemeral habitats (e.g., Husband and Barrett 1998). Other species might be limited by the occurrence of the proper growing conditions or the lack of niches for seeds to recruit. Often, disturbances are needed to generate those conditions and promote increased recruitment from the seed bank (Fenner and Thompson 2005). Although changes in aboveground plants tend to occur rapidly, longer periods are needed to trigger shifts within the seed bank (Ma et al. 2020).

Decisions on the inclusion of the seed bank, the length of study, or the range of environmental conditions occur during the design stages, often responding to research objectives or to material limitations. Researchers might not be compelled to evaluate complex vital rates, such as seed banks, which are difficult to study, particularly when they require specific conditions to be elicited and may not be critical for the objectives of the study. The conditions under which the studies are implemented may meet these particular objectives but may not be enough to properly characterize the life history of the species, reducing the value of these data in comparative studies across organisms.

It can be proposed that studies presenting multiple matrix population models encompass a greater range of conditions and are therefore more likely to reveal the role of the seed bank. The number of matrices used in a study can aid in capturing the required variation to reveal seed bank dynamics. Using multiple

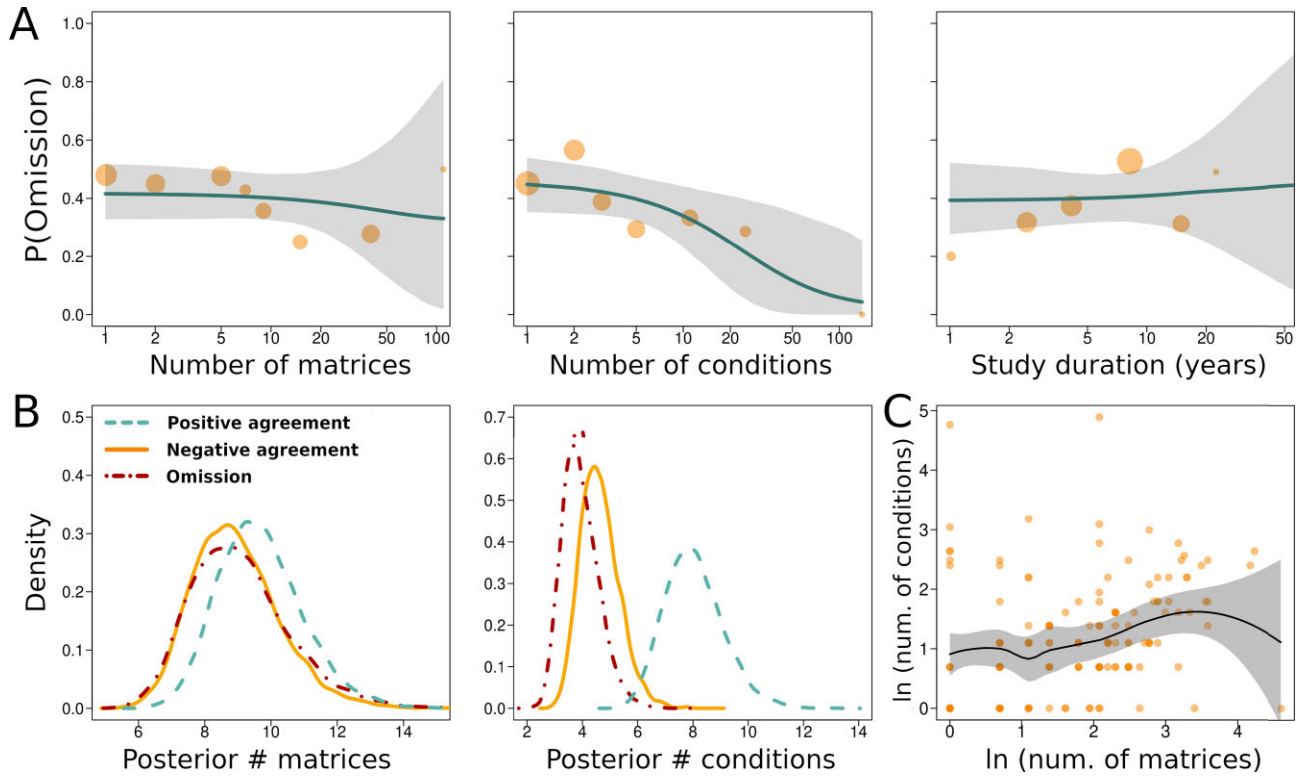

Figure 3. (A) the top three panels show the results of logistic regressions aimed at estimating changes in the probability of omission as a function to the number matrices presented by a study (left), the number of conditions included (middle), and the duration of the study (right). (B) The posterior distribution of negative binomial models estimating the number of matrices (left) and the number of conditions (middle) associated with demographic modeling literature with positive agreements (both with seed bank), negative agreements (both without seed bank), and with seed bank omission. (C) A Loess curve showing the relationship between the number of matrices constructed in a study (natural log transformed) and the number of conditions explored (also natural log transformed).

matrices allows researchers to incorporate a broader range of data points and conditions, leading to a more nuanced description. We assessed how the number of matrices in a study affected the probability of omission of seed banks. We found that probability of omission remained constant—at around .4, for studies with fewer than five sites, then decreased to .3 in studies with 10–50 matrices (figure 3). This slight declining trend in probability of omission was not supported for studies with more than 50 matrices. In these last studies, matrices often represent replicates of experimental treatments and not necessarily explore any additional environmental heterogeneity. We did not find a consistent association between number of matrices and number of conditions studied (a proxy of environmental heterogeneity). Agreement between literature sources is not necessarily the same for species that form seed banks (positive agreement) as for species that don't (negative agreement). Therefore, we compared the number of matrices in studies representing each type of agreement with those exhibiting omission (there was not enough data to evaluate unexpected inclusions). All three groups showed similar average number of matrices (figure 3), with studies showing positive agreement having slightly higher values than studies with negative agreement and omissions (9.7 versus 9.1). Partly, this neutral relationship might be explained by a lack of actual relationship between the matrices constructed in a study and the heterogeneity of the environment represented (figure 3). In some cases, matrices might represent replicates of theoretical manipulations or short-term experimental treatments (e.g., Sletvold and Rydgren 2007), rather than actual differences among environments.

The length of a study might influence the ability to detect the seed bank in a population and gauge its relevance. In simple terms, longer studies provide increased opportunities to observe the germination and establishment of plants from the seed bank. Over longer time frames, different environmental conditions and disturbances may occur, revealing how seed banks respond to a wider range of factors. Longer-term studies enable researchers to track shifts in populations, including those caused by slow successional processes, and to capture recruitment from long-lived seeds. As with number of matrices, we evaluated the variation in omissions across studies with different lengths. We did not find evidence that study length changed the probability of omissions.

Alternatively, the study design could include sufficient spatial variation in habitat or experimental manipulation to cover a wide range of environmental conditions in a shorter period of time. By explicitly including a range of environmental conditions, successional stages, or population structures, studies can increase the likelihood of capturing important events in seed bank dynamics. We evaluated the probability of omission as a function of the number of conditions represented by a study. We considered experimental treatments, habitats, populations, and sites as representing different conditions (the number of plots was not included). The probability of omission clearly decreased as a function of the number of conditions (figure 3). Again, we compared the number of conditions in studies with positive agreement, negative agreement in the inclusion of seed banks, and probability of omission. In this case, the number of conditions in studies showing positive agreement was much higher on average than in

studies exhibiting negative agreement (8.05 and 4.02, respectively; see figure 3). Studies with omissions had the lowest number of conditions. These results support the notion that studies including greater variation of conditions might be less prone to seed bank omissions in demographic models.

In short, increasing the number of environmental conditions represented can provide a more comprehensive and accurate understanding of seed banks in particular and life histories in general. The role of environmental variation in elucidating seed bank and other stages dynamics cannot be overstated. We suggest that future meta-analysis and studies seeking to leverage these data bases should account for environmental heterogeneity. A review of 355 demographic papers by Crone and colleagues (2011) similarly concluded that predicting capabilities of demographic studies will depend on better understanding on population environmental drivers, habitat heterogeneity, and longer studies.

Is it justified?

We explored the differences in documentation regarding the inclusion of seed banks between demographic studies with positive agreement, negative agreement, and discrepancy with seed bank studies. We considered whether each study presented documentation for its decision—either by providing a literature reference or by including seed banks in the experimental design. We estimated differences in the probability of providing documentation. The results showed that studies with positive agreement were 36% more likely to provide documentation than those exhibiting omissions and 31% more likely than studies with negative agreement. The lower incidence of documentation in negative agreements may be due to the higher number of asexually reproducing plants in that group.

Conclusions

The potential limitations arising from seed bank omissions in plant demography (Kalisz and McPeek 1992, Nguyen et al. 2019) might affect our ability to leverage databases for robust analyses of ecological patterns. Our study was focused on the interaction between seed bank studies and plant demographic models as a way to unveil biases in the omission of seed bank stages within matrix population models. These omissions are particularly pertinent given the role that seedbanks play in maintaining population dynamics and genetic diversity.

Surprisingly, no strong patterns emerged in terms of ecoregions forms that could explain the prevalence of seed bank omissions. Herbaceous perennials, particularly from temperate regions, were heavily overrepresented in the demographic literature. This life form had the highest frequency of unjustified omissions. Furthermore, the lack of phylogenetic signal in the discrepancies between seed bank studies and demographic models suggest that the source of bias likely arises from study limitations, particularly spatial scales, and factors that hinder accurate representation of environmental heterogeneity.

Several attempts have been made to find generalizations in seed persistence rates as functions of more easily measured characteristics. A seminal paper by Thompson and colleagues (1993) successfully described the correlation between seed size and seed shape with persistence in the soil for plant species. Such results were replicated in works performed in other geographic regions, such as Argentina (Funes et al. 1999). However, the pattern was not maintained in other works, most notably for Australia species (Leishman and Westoby 1998) and New Zealand species (Moles et al. 2000). Likewise, Rees (1993) and Honda (2008) were able to link optimal seed dormancy rates to adult longevity and other life history traits. More recently, machine learning algorithms have been developed to predict seed bank persistence on the basis of seed traits (Rosbakh et al. 2022, Tang and Li 2023). Although these models are promising, they might suffer from the same limitations described for big-data approaches.

Such generalities could be used to predict the formation of a substantial seed bank. However, they have been largely focused in understanding evolutionary trade-offs and do not offer enough detail about governing vital rates as to be usefully applied to demographic models. Because determining seed age structure tends to be difficult in most situations, alternative model structures (such as those based on physiological characterizations) might provide a valuable alternative. Long and colleagues (2015) suggested the use of exposure-resistance models in which seed traits conferring resistance to environmental factors were associated with changes in viability. Such approach could provide a window for more rapid characterizations of seeds and to establish a more robust method for cross-species and cross-population comparisons. Another potential avenue for enhancing seed bank detection is to expand the duration of demographic studies. In addition, our investigation into the inclusion of a diverse range of conditions within a shorter study period offered a promising approach.

Although technological advancements and the availability of ecological databases offer immense opportunities, they also warrant a critical examination of biases and limitations inherent in the data. By addressing these issues and incorporating a comprehensive understanding of seed bank dynamics and other dormant life stages, researchers can move closer to robust ecological insights.

In a rapidly changing world, where environmental disruptions and disturbances are becoming more prevalent, a thorough understanding of vital rate variation among multiple stages and with environmental heterogeneity becomes all the more critical. Our findings underscore the need for continued efforts in refining demographic models and ensuring the comprehensive role of all relevant stages is incorporated. More importantly, as the field of ecology continues to evolve, embracing data-driven approaches while maintaining a critical perspective of the data sources will be essential for disentangling complex ecological patterns and relationships.

Supplemental data

Supplemental material are available at *BIOSCI* online.

Acknowledgments

We thank Ken Fedorka, Chase Mason, Nicolle Beckman, and Dave Jenkins for their support through the process of developing this work. We are also indebted to the editors and the anonymous reviewer for their critical suggestions and corrections that helped greatly improve this article.

Author contributions

Federico López-Borghesi (Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Visualization, Writing – original draft), and Pedro F. Quintana-Ascencio

(Conceptualization, Data curation, Methodology, Validation, Writing – review & editing)

References cited

Aljumah AI, Nuseir MT, Alam MM. 2021. Traditional marketing analytics, big data analytics and big data system quality, and the success of new product development. *Business Process Management Journal* 27: 1108–1125.

Beckman NG, Bullock, JM, Salguero-Gómez, R. 2018. High dispersal ability is related to fast life-history strategies. *Journal of Ecology* 106: 1349–1362.

Bouwmeester H, Schuurink RC, Bleeker PM, Schiestl F. 2019. The role of volatiles in plant communication. *Plant Journal* 100: 892–907.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76: 1–32.

Caswell H. 2000. *Matrix Population Models,* vol. 1. *Sinauer*.

Compagnoni A,, et al. 2021. Herbaceous perennial plants with short generation time have stronger responses to climate anomalies than those with longer generation time. *Nature Communications* 12: 1824.

Crone EE,, et al. 2011. How do plant ecologists use matrix population models? *Ecology Letters* 14: 1–8.

Doak DF, Thomson D, Jules ES. 2002. Population viability analysis for plants: Understanding the demographic consequences of seed banks for Population health. Pages 312–337 in Beissinger SR McCullough DR, eds. *Population Viability Analysis*. The University of Chicago Press.

Edwards JL, Lane MA, Nielsen ES. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* 289: 2312–2314.

Fenner, M, Thompson, K. 2005. *The Ecology of Seeds*. Cambridge University Press.

FitzJohn RG. 2012. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution* 3: 1084–1092.

Flack A,, et al. 2016. Costs of migratory decisions: A comparison across eight white stork populations. *Science Advances* 2: e1500931.

Fritz SA, Purvis A. 2010. Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology* 24: 1042–1051.

Fry EL, Savage J, Hall AL, Oakley S, Pritchard WJ, Ostle NJ, Pywell RF, Bullock JM, Bardgett RD. 2018. Soil multifunctionality and drought resistance are determined by plant structural traits in restoring grassland. *Ecology* 99: 2260–2271.

Funes G, Basconcelo S, Díaz S, Cabido M. 1999. Seed size and shape are good predictors of seed persistence in soil in temperate mountain grasslands of Argentina. *Seed Science Research* 9: 341–345.

Gobeyn S, Mouton AM, Cord AF, Kaim A, Volk M, Goethals PL. 2019. Evolutionary algorithms for species distribution modelling: A review in the context of machine learning. *Ecological Modelling* 392: 179–195.

Grime JP. 2006. *Plant Strategies, Vegetation Processes, and Ecosystem Properties*. Wiley.

Harmon LJ, Weir JT, Brock CD, Glor RE, ChalLenger W. 2008. GEIGER: Investigating evolutionary radiations. *Bioinformatics* 24: 129–131.

Honda Y. 2008. Ecological correlations between the persistence of the soil seed bank and several plant traits, including sees dormancy. *Plant Ecology* 196: 301–309.

Hoyle GL, Steadman KJ, Good RB, McIntosh EJ, Galea LM, Nicotra AB. 2015. Seed germination strategies: An evolutionary trajectory independent of vegetative functional traits. *Frontiers in Plant Science* 6: 731.

Husband BC, Barrett SC. 1998. Spatial and temporal variation in population size of *Eichhornia paniculata* in ephemeral habitats: Implications for metapopulation dynamics. *Journal of Ecology* 86: 1021–1031.

Iversen CM,, et al. 2017. A global fine-root ecology database to address below-ground challenges in plant ecology. *New Phytologist* 215: 15–26.

Jiménez-Alfaro B, Silveira FA, Fidelis A, Poschlod P, Commander LE. 2016. Seed germination traits can contribute better to plant community ecology. *Journal of Vegetation Science* 27: 637–645.

Kachamas P, Akkaradamrongrat S, Sinthupinyo S, Chandrachai A. 2019. Application of artificial intelligent in the prediction of consumer behavior from Facebook posts analysis. *International Journal of Machine Learning and Computing* 9: 91–97.

Kalisz S, McPeek MA. 1992. Demography of an age-structured annual: Resampled projection matrices, elasticity analyses, and seed bank effects. *Ecology* 73: 1082–1093.

Kaplan RM, Chambers DA, Glasgow RE. 2014. Big data and large sample size: A cautionary note on the potential for bias. *Clinical and Translational Science* 7: 342–346.

Kays R, et al. 2022. The Movebank system for studying global animal movement and demography. *Methods in Ecology and Evolution* 13: 419–431.

Kleyer M,, et al. 2008. The LEDA Traitbase: A database of life-history traits of the Northwest European flora. *Journal of Ecology* 96: 1266–1274.

Leishman MR, Westoby M. 1998. Seed size and shape are not related to persistence in soil in Australia in the same way as in Britain. *Functional Ecology* 12: 480–485.

Lesica P,Steele BM. 1994. Prolonged dormancy in vascular plants and implications for monitoring studies. *Natural Areas Journal* 14: 209–212.

Logofet DO, Ulanova NG, Belova IN. 2017. From uncertainty to an exact number: Developing a method to estimate the fitness of a clonal species with polyvariant ontogeny. *Biology Bulletin Reviews* 7: 387–402.

Long RL, Gorecki MJ, Renton M, Scott JK, Colville L, Goggin DE, Finch-Savage WE. 2015. The ecophysiology of seed persistence: A mechanistic view of the journey to germination or demise. *Biological Reviews* 90: 31–59.

López-Borghesi, F, Koontz SM, Smith SA, Haller Crate SJ, Quintana-Ascencio PF, Menges ES. 2023. Leveraging projection models to evaluate long-term dynamics of scrub mint translocations. *Conservation Science and Practice* 5: e12947.

Ma M, Collins SL, Du G. 2020. Direct and indirect effects of temperature and precipitation on alpine seed banks in the Tibetan Plateau. *Ecological Applications* 30: e02096.

McCormack ML, Iversen CM. 2019. Physical and functional constraints on viable belowground acquisition strategies. *Frontiers in Plant Science* 10: 1215.

McHugh ML. 2013. The chi-square test of independence. *Biochemia Medica* 23: 143–149.

Moles AT, Hodson DW, Webb CJ. 2000. Seed size and shape and persistence in the soil in the New Zealand flora. *Oikos* 89: 541–545.

Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, Schiffers K, Thuiller W. 2012. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution* 3: 743–756.

Nathan R, Muller-Landau HC. 2000. Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology and Evolution* 15: 278–285.

Nguyen V, Buckley YM, Salguero-Gomez R, Wardle GM. 2019. Consequences of neglecting cryptic life stages from demographic models. *Ecological Modelling* 408: 108723.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, Pearse W. 2013. The Caper Package: Comparative Analysis of Phylogenetics and Evolution in R. R Package, vers. 5. R Foundation for Statistical Computing.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401: 877.

Paniw M, Ozgul A, Salguero-Gómez R. 2018. Interactive life-history traits predict sensitivity of plants and animals to temporal autocorrelation. *Ecology Letters* 21: 275–286.

Praveen KB, Kumar P, Prateek J, Pragathi G, Madhuri J. 2020. Inventory management using machine learning. *International Journal of Engineering Research and Technology* 9: 866–869.

Quintana-Ascencio PF. 2023. The importance of heterogeneity. *Proceedings of the National Academy of Sciences* 120: e2314786120. https://doi.org/10.1073/pnas.2314786120

R Core Team. 2019. R: A Language and Environmental for Statistical Computing. R Foundation for Statistical Computing. www.R-project.org.

Raub M. 2018. Bots, bias, and big data: Artificial intelligence, algorithmic bias, and disparate impact liability in hiring practices. *Arkansas Law Review* 71: 529.

Rees M. 1993. Trade-offs among dispersal strategies in British plants. *Nature* 366: 150.

Rosbakh S, Pichler M, Poschlod P. 2022. Machine-learning algorithms predict soil seed bank persistence from easily available traits. *Applied Vegetation Science* 25: e12660.

Rosbakh S, Carta A, Fernández-Pascual E, Phartyal SS, Dayrell RL, Mattana E, Saatkamp A, Vanderlook F, Baskin J, Baskin C. (2023). Global seed dormancy patterns are driven by macroclimate but not fire regime. *New Phytologist* 240: 555–564.

Salguero-Gómez R,, et al. 2015. The compadre Plant Matrix Database: An open online repository for plant demography. *Journal of Ecology* 103: 202–218.

Salguero-Gómez R, Jones OR, Jongejans E, Blomberg SP, Hodgson DJ, Mbeau-Ache C, Zuldema PA, de Kroon H, Buckley YM. 2016. Fast–slow continuum and reproductive strategies structure plant life-history variation worldwide. *Proceedings of the National Academy of Sciences* 113: 230–235.

Schmaljohann H, Fox JW, Bairlein F. 2012. Phenotypic response to environmental cues, orientation and migration costs in songbirds flying halfway around the world. *Animal Behaviour* 84: 623–640.

Simpson RL, Leck MA, Parker VT. 1989. Seed banks: General concepts and methodological issues. Pages 3–8 in Leck MA, Parker VT Simpson RL, eds. *Ecology of Soil Seed Banks*. London Academic Press.

Sletvold N, Rydgren K. 2007. Population dynamics in *Digitalis purpurea*: The interaction of disturbance and seed bank dynamics. *Journal of Ecology* 95: 1346–1359.

Stan Development Team. 2018. Stan Modeling Language UsersGuide and Reference Manual, vers. 2.18.0. http://mc-stan.org.

Stöcklin J, Fischer M. 1999. Plants with longer-lived seeds have lower local extinction rates in grassland remnants 1950–1985. *Oecologia* 120: 539–543.

Süngü Şeker Ş, Akbulut MK, Şenel G. 2021. Seed morphometry and ultrastructure studies on some Turkish orchids (Orchidaceae). *Microscopy Research and Technique* 84: 2409–2420.

Tang Y, Li H. 2023. Comparing the performance of machine learning methods in predicting soil seed bank persistence. *Ecological Informatics* 77: 102188.

Thompson K, Band SR, Hodgson JG. 1993. Seed size and shape predict persistence in soil. *Functional Ecology* 7: 236–241.

Thompson K, Bakker JP, Bekker RM, Hodgson JG. 1998. Ecological correlates of seed persistence in soil in the north-west European flora. *Journal of Ecology* 86: 163–169.

Venable DL. 2007. Bet hedging in a guild of desert annuals. *Ecology* 88: 1086–1090.

Warr SJ, Thompson K, Kent M. 1993. Seed banks as a neglected area of biogeographic research: A review of literature and sampling techniques. *Progress in Physical Geography: Earth and Environment* 17: 329–347.

Received: January 16, 2024. **Revised:** April 2, 2024. **Accepted:** April 16, 2024