Environmental Data Science (2026), 5: e41, 1–20
doi:10.1017/eds.2026.10067

CAMBRIDGE UNIVERSITY PRESS

SURVEY PAPER

Methods for imputing trait and community data in ecological and evolutionary research

Matthew T Turner[1], Georgios Leontidis[2,3], David FRP Burslem[1,4], Justin MJ Travis[5] and Roslyn C Henry[1]

[1]School of Biological Sciences, University of Aberdeen, UK
[2]Department of Physics and Technology, UiT The Arctic University of Norway, Norway
[3]School of Natural and Computing Sciences, University of Aberdeen, UK
[4]Interdisciplinary Institute, University of Aberdeen, UK
[5]Department of Biosciences, Durham University, UK
Corresponding author: Matthew T Turner; Email: r01mt23@abdn.ac.uk

Received: 26 January 2026 UTC; **Revised:** 02 July 2026 UTC; **Accepted:** 18 September 2026 UTC

Keywords: community data; data imputation; deep learning; missing data; species traits

Abstract

Species trait data are widely used in ecology and evolution to approximate ecological niche. Despite this, the data housed in species trait databases are often patchy or biased toward charismatic species and easy-to-study traits. Filling data gaps can be achieved by determining the most likely values via data imputation. We conducted a systematic literature review to identify data imputation methods for species trait data in ecological and evolutionary studies, as well as in data science studies, and compared the two domains to identify whether alternative methods might be suitable for adoption in ecology and evolution. We focus on novel and state-of-the-art deep learning (DL) imputation methods. We use evidence maps to identify missing data contexts where DL alternatives have been shown to be effective in data science. Differences in preferred methods were identified between the two domains. DL methods were rarely used in ecological and evolutionary studies, whereas they were frequently employed in data science. Several DL methods that have not been adopted in ecology and evolution demonstrated state-of-the-art performance when imputing data in contexts that are similar to those faced by ecologists wishing to impute species trait data. We identified a clear gap in the ecological and evolutionary literature in relation to the uptake and use of imputation methods based on DL. We then identify autoencoders, transformer-based architectures, Bayesian neural networks, and feed-forward neural networks as primary architectures worth further investigation for their capacity to impute missing species trait data and potential for multiple imputation.

Impact Statement

Trait data are essential resources in ecological and evolutionary research, although their repositories often have high degrees of missing values. Imputation methods find likely values for these missing data by leveraging phylogenetic and trait covariance structures. Trait imputation methods are often statistical or use machine learning, although deep learning (DL) alternatives have seen increased attention in other fields, outperforming baseline methods such as Random Forest and k-Nearest Neighbors. We present the results of a literature review aimed at identifying DL methods that have not yet been applied to trait data. We identify autoencoders, feed-forward neural networks, and transformers as promising imputers of tabular trait data. We propose further research into the efficacy of these methods in trait data contexts.

This review article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

1. Introduction

1.1. Traditional methods for dealing with missing trait data

Data describing the characteristics of both individual species and entire communities are essential across a variety of research questions and methodologies in ecological and evolutionary studies. Individual species and organismal-level trait data have been used, for example, for the parameterization of process-based models, for predicting how disturbances or climate change affect the persistence and functions of species and ecosystems (Stork et al., 2009; Newbold et al., 2013; Johnson et al., 2021), and for conservation (Penone et al., 2014; Greenberg et al., 2017). Although formal definitions of species traits often limit them to the morphological, physiological, and phenological features of individuals (Violle et al., 2007), additional features describing both functional traits and emergent properties are often utilized within ecological and evolutionary studies (He et al., 2023). To account for these additional features, in this article we use an expanded definition of the term "species trait data" to refer to a broader range of species' features including morphological, demographic, physiological, and phenological data. Data measured at the community or plot level, such as mean canopy height, are used to address many questions in ecology and conservation biology, such as characterizing habitats to improve conservation approaches (Loehle et al., 2015) and incorporating species interactions into macroevolutionary analyses (Weber et al., 2017). Accurate and complete data are essential for accurate predictions and analyses for both types of data, especially when methods are not robust to missing data (Urban et al., 2016, 2022). However, datasets of species traits and community properties (hereafter referred to collectively as trait data) are rarely complete.

Missingness can be caused by three general mechanisms (Table 1), which are important when considering how to handle data gaps (Fielding et al., 2008; Gomer and Yuan, 2023). However, identifying which mechanism applies to a given dataset is often challenging (Galimard et al., 2018). In ecological and evolutionary studies, and trait data repositories, missing species trait data are often the result of various origins of bias (Table 2). These origins of bias mean that species trait data are often comprised of values that are missing under various mechanisms, including missing at random (MAR) and missing not at random (MNAR) data (González-Suárez et al., 2012; Penone et al., 2014).

Within ecological and evolutionary studies, preprocessing data techniques are often employed to handle missing data (Penone et al., 2014; Taugourdeau et al., 2014). Listwise deletion, also referred to as complete case analysis, is the process of removing samples with missing data. While quickly and easily implemented, listwise deletion reduces sample sizes and biases the results of studies using real-world MAR or MNAR data (Myers, 2011; Zhang, 2016).

An alternative is data imputation, whereby missing values are replaced with plausible estimates derived from the observed data. Imputation approaches can be broadly divided into single imputation,

Table 1. *Explanations of missingness mechanisms*

Missingness mechanism	Explanation
MCAR	There is no relationship between the probability of missing data and species trait values, that is, each trait for each species is equally likely to be missing.
MAR	The likelihood of a value being missing is related to value of observed features of the dataset. For instance, data are MAR if the probability of a missing value for natal dispersal distance depends on the known value of another trait, such as average adult body mass
MNAR	The probability of missing data is related to unknown data, that is, either the value of the missing data itself or to data not included in the dataset. For example, if a species' body mass directly influences the likelihood of its own value being missing.

Note: MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.

Table 2. *Origins of bias in species trait data*

Bias origin	Examples
Taxonomy	Data for amphibians and reptiles are scarcer than for birds and mammals (Etard et al., 2020)
Geospatial	Data availability often increases as species range size increases and for species that are more accessible to study (Culumber et al., 2019).
Difficult to study species	By their nature, rare species often suffer from missing data, yet trait data for rare and vulnerable species are particularly valuable (Etard et al., 2020).
Difficult to study traits	Easier to study species traits, such as body mass, typically have greater coverage in trait databases than more complex and challenging to study traits (González-Suárez et al., 2012), such as dispersal behavior (Nathan, 2001; Whitmee and Orme, 2013) or lifespan (Nakagawa and Freckleton, 2008).

which produces one completed dataset, and multiple imputation, which generates several completed datasets that reflect uncertainty in the missing values. Single imputation remains widely used but has been criticized for treating imputed values as if they were observed data, leading to underestimated uncertainty and overly confident statistical inferences, a problem termed the *single imputation fallacy* (Blomberg and Todorov, 2025). In contrast, multiple imputation explicitly propagates uncertainty by drawing imputations from the posterior predictive distribution of the missing data conditional on the observed data. Each completed dataset is analyzed separately, and parameter estimates are subsequently combined using Rubin's rules (Rubin, 1987). As a result, multiple imputation generally produces more reliable estimates of both parameters and their associated uncertainty.

In ecological and evolutionary studies, imputation is typically not an end in itself. Rather, it is used to enable downstream analyses and hypothesis testing that would otherwise be limited in the presence of missing data. Consequently, the usefulness of an imputation method depends not only on its ability to recover missing values accurately but also on its capacity to preserve biologically meaningful relationships and support valid statistical inference. This distinction is illustrated by Blomberg and Todorov (2025), who compared complete-case analysis, single imputation, and multiple imputation by evaluating downstream regression models. Specifically, they assessed bias in regression coefficients and the accuracy of estimated standard errors relative to analyses of complete data. Their results showed that multiple imputation yielded confidence intervals that more frequently recovered the true parameter values, whereas single imputation often produced overconfident estimates and, in some cases, performed worse than complete-case analysis. These findings highlight the importance of accounting for imputation uncertainty when the ultimate goal is biological inference rather than prediction alone.

Many imputation methods are available, varying in complexity (see Supporting information A for descriptions of common methods). Simple imputation methods include replacing a missing value with the feature (e.g., trait) average, or using the "last observation carried forward" in longitudinal datasets, where the missing value is replaced by the most recent observed value. Similar to listwise deletion, these simple imputation methods can result in biased results and poor accuracy when data are not MCAR, that is, the data are classified as MAR or MNAR (Fielding et al., 2008; Zhang, 2016).

More complex imputation methods draw on additional information available within the dataset to estimate a likely value. By leveraging relationships among variables or similarities between samples, these methods can better preserve feature distributions and inter-variable correlations (Gromski et al., 2014; Zhang, 2016; May et al., 2023). For instance, k-nearest neighbors (kNN) imputation identifies samples that are most similar to the one with missing data to inform the imputation (Zhang, 2012). As closely related species tend to have similar trait values, the addition of phylogenetic information in these more advanced methods, either as a phylogenetic tree or in tabular format, can also boost model

performance (Penone et al., 2014; Johnson et al., 2021). Though relatively well performing, and more robust than simpler imputation methods (Zhang, 2016), the efficacy of advanced methods is dependent on the specific context of the missingness of the data. Conditions including the missingness type (MCAR, MAR, and MNAR), proportion of missing data, data type (such as numerical, categorical, longitudinal, image, text, or mixed), and the complexity of the missing data all influence their performance (Hunt, 2017; Johnson et al., 2021).

In their study on missing value handling in trait data, Johnson et al. (2021) highlight that conventional imputation methods are not always the most appropriate solutions for missing trait data. These conventional methods, including multiple imputation by chained equations (MICE) (once via regression and again via random forest [RF]), phylogenetic imputation via phylopars (Goolsby et al., 2017), and Bayesian hierarchical probabilistic matrix factorization, all see a fall in accuracy when realistic missingness biases (MAR and MNAR) are included within simulated trait datasets and as the proportion of missingness increases. These findings underscore the urgent need to improve methods for imputing missing trait data.

1.2. Could data science and AI improve our efforts at imputation?

Over the past few decades, we have transitioned into the era of big data (Kitchin and Lauriault, 2015; Duan et al., 2019). Large datasets and databases are becoming more common and tools for handling such data are now essential. Data science is a discipline that focuses on the development of data analysis and management techniques (Dhar, 2013), and as such data science studies are a rich source of alternative imputation methods. It is possible that some of these alternative imputation methods could outperform the conventional trait data imputation methods used in ecological and evolutionary studies.

Artificial Intelligence (AI) methods, particularly Machine Learning (ML), are a common feature of state-of-the-art imputation methods in data science (Luo, 2022; Adhikari et al., 2023). AI and ML are renowned for their capacity to handle and analyze large and complex data efficiently (Zhuang et al., 2017). Consequently, modern AI and ML methods have garnered attention for their capacity to effectively process and impute data (Irwin et al., 2021). Deep learning (DL) is a branch of ML which uses neural networks (see Supporting information B for a more in-depth explanation of neural networks), such as AEs or recurrent neural networks (RNN), to "learn" complex relationships between data features (LeCun et al., 2015). By accurately representing these complex feature relationships, deep neural networks have the capability to conduct highly complex data analysis tasks and consequently show promise for data imputation (Duan et al., 2016; Cardoso Pereira et al., 2020; Fang and Wang, 2020).

Despite its promise, the suitability of DL for data imputation remains contentious. A recent review has shown that DL architectures, including generative adversarial imputation networks (GAIN) and variational autoencoders (VAE), may be outperformed by conventional methods such as RF and MICE when assessed using root mean square error (RMSE) and overall accuracy (Sun et al., 2023). Additionally, two studies (Elezaj et al., 2018; Xie et al., 2024) found that DL methods fail to outperform conventional methods in terms of not only RMSE but also in the preservation of variable distributions and mean absolute percentage error (MAPE). Conversely, many DL methods have been shown to perform well for data imputation tasks, outperforming simple statistical and ML alternatives (Cappelletti et al., 2020; Lin et al., 2022). Reviews of DL architectures, such as autoencoders (AEs), have highlighted the effectiveness of AE-based methods in surpassing conventional techniques (Cardoso Pereira et al., 2020). Moreover, the potential of DL has been recognized across various domains, including healthcare, where it has shown promise for improving imputation performance (Liu et al., 2023).

To evaluate the effectiveness of state-of-the-art DL methods for imputing species trait data, it is essential to first identify which imputation techniques are currently used in ecological and evolutionary research. In parallel, we must assess which specific DL architectures, and their extensions, are most suitable for these data types and research contexts. The broader field of data science likely offers a wide range of modern imputation methods that have yet to be applied in ecological or evolutionary studies. As

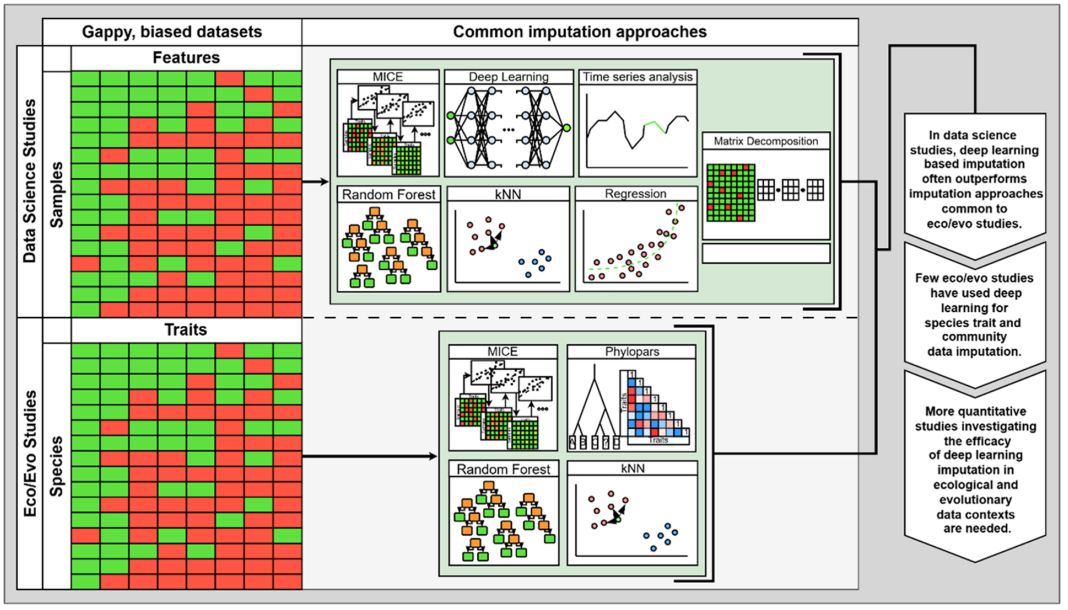


Figure 1. *High degrees of biased missing data are common in datasets used by both data science studies and ecological and evolutionary studies. Imputation methods can fill these gaps by estimating the most likely value for missing data. Data science studies employ a wider variety of imputation methods than those used for species trait data imputation, including the persistent use of deep learning. These deep learning methods can outperform the methods that are used most often for species trait data imputation, and as such, studies into their efficacy in ecological and evolutionary contexts are recommended. kNN = k Nearest Neighbor; MICE = Multiple Imputation by Chained Equations.*

such, DL methods represent a valuable opportunity to explore powerful alternatives for improving trait data imputation in these domains.

To this end, we conducted a systematic review of imputation methods used in ecological, evolutionary, and data science studies. By examining the distinct domains, we identify state-of-the-art DL imputation architectures from data science that hold promise for species trait data imputation. We compare the context of current applications of DL in terms of missingness proportion, missingness type, and data type with that of the equivalent in species trait datasets. Such comparisons allow for parallels to be drawn, and limitations identified. We conclude by deriving recommendations for the utilization of DL imputation methods for ecological and evolutionary studies (Figure 1).

2. Materials and methods

2.1. Systematic review of the literature

The structure of the literature review followed the Protocol and Reporting result with Search, Appraisal, Synthesis, and Analysis framework (PSALSAR) methodology set out by Mengist et al. (2020). To define the scope and aims of the review, we used the Population, Intervention, Comparison, Outcome(s), and Context (PICOC) protocol. We searched "Web of Science" (Primary collection and Citation Connection databases) and "Scopus" using strings of combinations of the following keywords/phrases: "Trait data," "Imputation," "Missing data," "Ecology," "Evolution," "Parameterization," "Data science," and "Gap filling." Certain keyword combinations were omitted because they resulted in large returns (e.g., "Missing data" AND "Ecology") without any specific relevance to data imputation. Following this process, 13 search strings were chosen:

1. "Trait data" AND "Imputation"
2. "Trait data" AND "Missing data"
3. "Trait data" AND "Missing data" AND "Ecology"
4. "Trait data" AND "Missing data" AND "Evolution"
5. "Imputation" AND "Ecology"
6. "Imputation" AND "Ecology" AND "Trait data"
7. "Imputation" AND "Evolution"
8. "Imputation" AND "Evolution" AND "Trait data"
9. "Imputation" AND "Data science"
10. "Imputation" AND "Missing data" AND "Data science"
11. "Trait data" AND "Parameterization"
12. "Data science" AND "Parameterization"
13. "Gap filling" AND "Missing data"

Studies were evaluated against a set of criteria (Table 3) and either included or excluded in the final review. Studies were deemed pertinent to data science based on the journal in which they were published. Those published in journals specific to computer science, data science, and mathematics were retained, whereas those published in journals tangentially related to these domains, such as the *Journal of Medical Statistics and Informatics*, were considered too far removed from these focal fields and were thus omitted. Data from the retained studies were then extracted, including article title, study domain, publication year, imputation method(s) used, type of data imputed, missingness type, and missingness proportion. The scope of this review is limited by our explicit search for studies that define their problem as a missing data problem, and our focus on use of the term "imputation" to define their missing data handling approach. In some contexts, methods that could be considered imputation, such as the reconstruction of temporal patterns, are not viewed as imputation problems by ecologists and evolutionary biologists, and as such, may be overlooked in our search. Additionally, alternative terms, such as "interpolation" or "gap-filling," have also been used historically, and as such, it is likely that some studies were not discovered by our search terms. Despite these limitations, we consider this review to be a reliable representation of the current state of missing data handling within both domains because the term "imputation" is currently prominent in both ecological/evolutionary and data science studies (Table 4).

2.2. Data analysis

2.2.1. Comparing domains

Comparisons were made between the preferences for imputation methods used in ecology and evolution studies and those in data science studies. Studies making comparisons between several imputation methods as opposed to focusing on a single method were omitted from this section of the study, as it was determined that they did not constitute the application or evaluation of an individual method.

Table 3. *Inclusion criteria for articles in the systematic review*

Criteria	Decision
The article is written in English.	Inclusion
The article is published in a peer-reviewed scientific journal.	Inclusion
The article pertains to the imputation of species trait data, functional trait data, emergent properties of populations or communities, OR data science generally.	Inclusion
The article includes the evaluation of at least one imputation method.	Inclusion
The article is primary research.	Inclusion
The full text of the article is available.	Inclusion

Table 4. *Replication statement covering the scale of inference, scale at which the factor of interest is applied, and number of replicates at the appropriate scale for our analyses*

Scale of inference	Scale at which the factor of interest is applied	Number of replicates at the appropriate scale
Domain of study.	Individual articles.	177 ecological/evolutionary articles, 160 data science papers.
Domain of study.	Individual articles with one main method.	156 ecological/evolutionary articles, 123 data science papers.
Deep learning studies in data science.	Individual articles.	96 applications of DL across 37 data science studies

Imputation methods were categorized into groups based on a methodological classification. For example, a generalized least squares method which incorporates phylogenetic information would be categorized as a form of regression. This yielded the following 19 categories: AEs, Average Imputation, Bayesian statistics, Decompositions, Feed-forward neural networks (FFNN), Generative adversarial networks (GAN), kNN, MICE, Phylogenetic imputation, Probabilistic, Regression, RF, RNN, Time series analysis, Other DL, Other ML, Other Statistical, Hybrid, and Other methods. Of these AE, GAN, RNN, FFNN, and Other DL were DL architectures.

The overall difference in methodological preference between domains was identified using Fisher's exact tests. Studies were then divided into four time brackets: "1990–1999," "2000–2009," "2010–2019," and "2020 Onwards." Where methods used overlapped in these time brackets, Fisher's exact tests were employed to compare the two domains. Comparisons were then made between time brackets within domains to identify trends in preferred methods within individual domains.

2.2.2. Reviewing the application of deep learning for imputation across data contexts

Three evidence maps were created to summarize the suitability of state-of-the-art DL architectures for imputing species trait data. Individual maps were created based on studies identified using the keywords "data type," "missingness type," and "missingness proportion," which we define collectively as the "data context."

To illustrate the suitability of state-of-the-art methods in terms of data type, studies were generalized into six types: numerical, categorical, longitudinal, image, text, and mixed. The missingness proportion was categorized into brackets of 10% from 0% to 100%, and the missingness type was classified as either MCAR, MAR, or MNAR. Here, a separate set of categorizations for observed DL imputation architectures was developed based on their generalized architecture, in part following the outline from Liu et al. (2023). Overall, seven DL architectures were identified: graph neural networks (GNNs), FFNNs, RNNs, AEs, GANs, hybrid models including architectures such as long short-term memory (LSTM) with attention and a transformer, and "Other" models such as deep Gaussian mixture models.

Map generation was conducted using only data science studies, as DL methods used within ecological and evolutionary studies inherently already indicate their suitability. Articles reviewing the efficacy of multiple methods were excluded from the analyses comparing the two domains of study, but they were included in these evidence maps as they offer results that highlight the suitability of these alternative methods under different data contexts. We then consider this with respect to the type of missing data commonly found in empirical trait data. By doing so, we explore the suitability of DL methods for imputation under different types of missingness.

All data comparisons were conducted in R version 4.4.0 using RStudio version 2024.04.1 + 748 (R Core Team, 2025).

3. Results

3.1. Systematic review of the literature

A total of 2387 articles were identified by the initial search (1508 from the Web of Science and 849 from Scopus), which reduced to 337 following applications of the evaluation criteria in Table 3. Of these studies, 160 were classified as data science articles and 177 classified as ecological and evolutionary articles. A list of data sources used in this study can be found in the "Full Data" sheet of our Data files on our Github.

3.2. Comparing domains

Retaining only studies which focus on a single imputation method yielded a subset of 279 articles, 156 of which were classified as ecological/evolutionary studies, and the remaining 123 studies were classified as data science.

A strong difference was observed in the preferred imputation methods overall between the two domains ($p < 0.0005$) (Figure 2). This difference was maintained within each time bracket comparison, with the domains either having no method overlap, as in the brackets of "1990–1999" and "2000–2009," or Fisher's exact tests indicating substantial differences between the two in the case of brackets "2010–2019" and "2020 Onwards" ($p < 0.005$).

In terms of comparing between contiguous time brackets within domains, Fisher's exact tests indicated little difference in method preference between "1990–1999" and "2000–2009" and between "2000–2009" and "2010–2019" in ecological and evolutionary studies (p-values > 0.05), whereas a substantial difference between "2010–2019" and "2020 Onwards" was indicated by a p-value of < 0.0005. For data science studies, Fisher's exact tests indicated a moderate difference in preferred imputation methods between "2000–2009" and "2010–2019" ($p = 0.081$), whereas results for between "2010–2019" and "2020 Onwards" were consistent with the null hypothesis ($p = 0.388$).

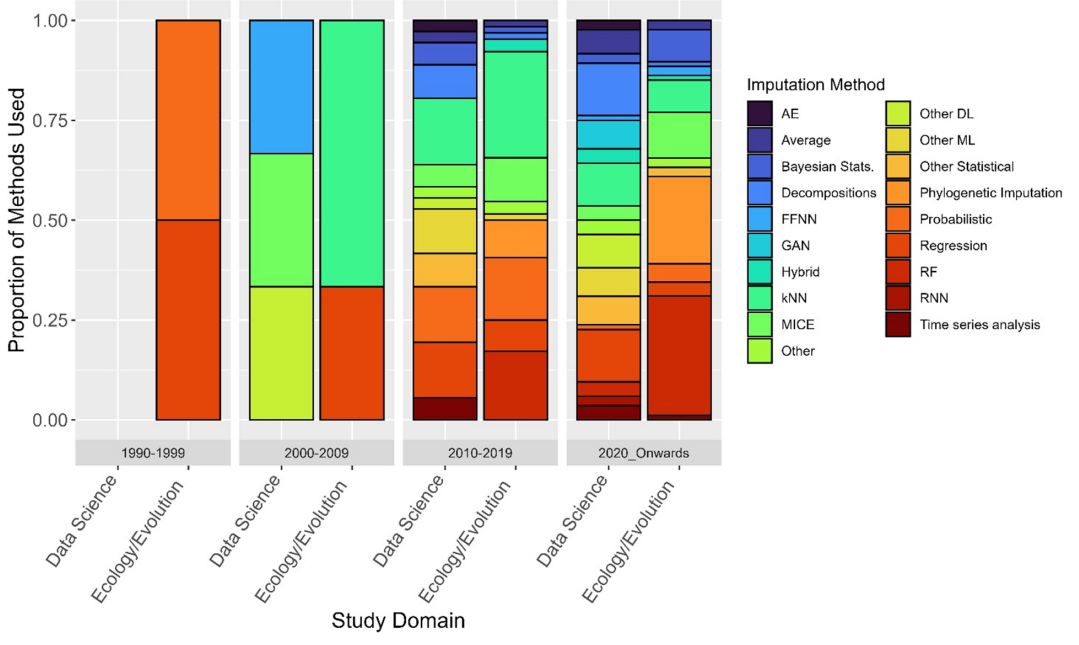


Figure 2. *Imputation methods used in both data science and ecology/evolution studies by time bracket. AE = Autoencoder; FFNN = Feed-Forward Neural Network; GAN = Generative Adversarial Network; kNN = k-Nearest-Neighbors; MICE = Multiple Imputation by Chained Equations; MLP = Multilayer Perceptron; Other DL = Other Deep Learning; Other ML = Other Machine Learning; RF = Random Forest; RNN = Recurrent Neural Network.*

ML methods have consistently dominated species trait data imputation in ecological and evolutionary studies across time brackets, with RF and kNN being the most commonly used methods (23.7% and 16.7% of total methods, respectively). Following these methods in popularity was the trait-data-specific method of phylogenetic imputation (16.0%). DL methods were only applied twice, via FFNNs, accounting for only 1.3% of total imputation applications.

In data science studies, although ML methods were still commonly used, DL was far more prevalent. Here, explicitly ML methods accounted for 22.8% total (including kNN, RF, and other ML). The most commonly used individual categories within data science were regression, at 13.0% of the methods used, and kNN methods, at 12.2%. DL methods in data science accounted for 20.3% of the methods used, including applications of AEs, FFNNs, GANs, RNNs, Hybrid models, and Other DL architectures.

3.3. Reviewing the application of deep learning for imputation across data contexts

A total of 96 applications of DL methods across 37 data science studies were identified. Table 5 shows the evidence maps generated for missingness proportion, missingness type, and data type associated with each instance of applied DL methods. All studies provided the type of data being imputed, 28 provided the proportion of data missing in their imputation trials, and 20 referred to the specific type of the missingness

Table 5. *Evidence maps of missing data contexts (missingness proportion, missingness type, and data type) provided within data science studies using deep learning imputation methods*

		GNN	FFNN	RNN	AE	GAN	Hybrid	Other
Missingness proportion	0–10%	0	5	7	5	11	6	8
	11–20%	3	3	7	4	13	7	8
	21–30%	0	5	8	7	13	8	9
	31–40%	0	2	8	1	11	5	5
	41–50%	0	4	8	5	10	8	7
	51–60%	0	1	4	2	7	3	4
	61–70%	0	2	3	2	5	5	4
	71–80%	0	2	3	3	7	3	3
	81–90%	0	1	3	1	2	2	2
	91–100%	0	1	1	1	0	2	2
Missingness type	MCAR	5	6	6	8	7	8	16
	MAR	0	2	1	3	5	3	3
	MNAR	0	1	0	3	3	2	2
Data type	Numerical	0	2	3	2	5	0	2
	Categorical	0	1	0	1	0	0	0
	Longitudinal	5	3	8	1	8	6	17
	Image	0	1	1	2	1	2	2
	Text	0	0	0	0	0	1	2
	Mixed	0	2	0	7	8	3	4

Note: GNN, graph neural network (4 articles); *FFNN*, feed-forward neural network (9 articles); *RNN*, recurrent neural network (9 articles); *AE*, autoencoder (9 articles); *GAN*, generative adversarial network (16 articles); Hybrid models (7 articles); Other models (16 articles). *MCAR*, missing completely at random; *MAR*, missing at random; *MNAR*, missing not at random. Shading refers to number of counts; green is higher, yellow is lower.

of their data. Further details regarding the missingness scenarios to which each DL method was applied are available in Supporting information C.

3.4. Deep learning methods for imputing different data types

3.4.1. Generative adversarial networks

A total of 13 studies used GAN-based methods. In eight of these, GANs were outperformed by conventional trait data methods (Lalande and Doya, 2022; Santos et al., 2022; Wang et al., 2022; Fuchs et al., 2023; Kim et al., 2024). This group includes the only studies to evaluate GANs on MAR and MNAR data (He, Yu et al., 2024; Zhou et al., 2024), and the only application of a GAN to species trait data (Gendre et al., 2024).

The remaining five studies reported that GAN-based methods, including a GAN-VAE hybrid architecture, outperformed conventional methods (Vi et al., 2021; Ali et al., 2024; Wang et al., 2024b; Bannany and Qahtan, 2025; Sarmad Ali et al., 2025). However, these studies either used MCAR data or did not specify the missingness mechanism. Overall, the performance of GAN imputation methods, relative to conventional methods, was dictated primarily by missingness mechanism.

3.4.2. Autoencoders

AE-based models outperformed baseline alternatives including kNN, RF, and MICE in five of the eight studies retained (Cappelletti et al., 2020; Ali et al., 2024; He, Yu et al., 2024; Wang et al., 2024b; Sarmad Ali et al., 2025). AEs were applied to several data types including image, categorical, longitudinal, and mixed data, with missingness proportions, when defined, ranging from 5% up to 80%. He, Yu et al. (2024) and Zhou et al. (2024) evaluated their models on biased MAR and MNAR datasets. He, Yu et al. found that their novel AE-Neural network classifier hybrid method consistently achieved the best performance. Their baseline AE model, MIDA (Multiple imputation using overcomplete denoising autoencoders), also performed comparably to conventional alternatives. Zhou et al. (2024) reported that while AE-based methods performed well in terms of RMSE and correlation coefficient imputation, the imputed datasets performed poorly in downstream tasks.

In a study by Antonenko et al. (2024), auto-replicative RF models outperformed vanilla AEs when imputing mixed data at low missingness rates (<= 30%). When applied to image data in a study by Santos et al. (2022), VAEs were found to add too much noise into imputed images compared to baseline alternatives including kNN and MICE. AEs were predominantly evaluated using MCAR data, or in studies that did not define their missingness scenario, and as such further analysis is required to better understand their capacity to handle MAR and MNAR data.

3.4.3. Recurrent neural networks

RNNs were used and compared to conventional methods four times. Bidirectional Recurrent Imputation for Time Series (BRITS) was used twice; in studies by both Amerise et al. (2024) and Wang et al. (2024b), it was shown to outperform kNN on longitudinal data and under MCAR conditions, although in the latter study BRITS was also found to be outperformed by RF. An LSTM model using the ReLU activation function was found to outperform all other models by Sowmya and Kayarvizhy (2021).

3.4.4. Feed-forward neural networks

Lin et al. (2022) utilized both a multilayered perceptron (MLP) based method, and a deep belief network to impute numerical data under MCAR conditions, with missing rates between 10% and 50%. Here, both DL methods significantly outperformed baseline alternatives including kNN, decision trees, and support vector machines. In a study by Santos et al. (2022), a deep image prior (DIP) network was used to impute image data across a range of missingness rates between 5% and 80%. Though DIP was only comparable to baseline alternatives including kNN and MICE when missingness rates were low, it consistently outperformed these baseline methods as missingness rates increased past 30%. As with their implementation of AEs, Cappelletti et al. (2020) found convolutional neural networks to outperform baseline kNN models.

3.4.5. Transformers

Four transformer-based methods were utilized across two studies. In a study by Amerise et al. (2024), Self-Attention-based Imputation for Time Series (SAITS), Multivariate time series imputation with transformers (MTSIT), and a transformer encoder model were used to impute longitudinal data under MCAR conditions. All three transformer methods outperformed the baseline kNN model, with SAITS showing the strongest performance, followed by the transformer encoder model. SAITS was also used by Wang et al. (2024b) to again impute longitudinal MCAR data. Here, SAITS outperformed kNN whilst showing comparable performances to RF.

3.4.6. Other

Four studies used architectures categorized as "Other." A mixed deep Gaussian mixture model was used by Fuchs et al. (2023) and was shown to have lower variation in performance across missingness scenarios (MCAR and MAR) and missingness rates compared to RF and kNN when imputing mixed type data. Zhou et al. (2024) found that conditional score–based diffusion models performed poorly when imputing continuous data across all three missingness scenarios, and at missingness rates of 30%, 50%, and 70%. So et al. (2025) found that self-organizing maps were consistently outperformed by baseline alternatives including RF and MICE when imputing mixed data under MNAR conditions at low missingness rates (5%, 10%, and 15%). A novel GNN method was introduced by Amerise et al. (2024) and was found to outperform all alternative methods including DL and kNN methods when imputing longitudinal data under MCAR conditions.

3.4.7. Additional DL architectures

A total of 18 studies used DL methods for imputation without comparison to conventional trait imputation methods. These studies evaluated a range of DL architectures, including VAEs, GANs, transformer-based models, deep Markov models, and novel GNN-based methods. DL-based methods outperformed non-DL alternatives in 9 of the 18 studies (Zawistowski and Grzenda, 2009; Fraccaro et al., 2017; Lai et al., 2018; Nguyen et al., 2020; Utomo et al., 2024; Wang et al., 2024a; Gan and Yan, 2025; Gonzalez-Calabuig et al., 2025; Wang et al., 2025); 2 studies found that non-DL methods outperformed DL methods (Elezaj et al., 2018; Xie et al., 2024); and a further two studies found DL methods could outperform non-DL methods under certain conditions. Pingi et al. (2024) found that their novel FaIC-GAN model showed varied performance across datasets, with poorer performance attributed to data with significant class imbalances. Skudnik and Jevšenak (2022) imputed species trait data, comparing the performance of an MLP with a mixed-effect model (MEM) to impute tree height for forest mensuration. While the MEM was more accurate at the plot level, when plots were clustered or additional variables were included, the MLP proved more accurate in terms of RMSE.

In the remaining 5 of the 18 studies, three developed novel methods, including a spatial-temporal aware data recovery network (Liang et al., 2023), a diffusion probabilistic model (DBiker) (Huang et al., 2024), and a novel use of large language models for information extraction from metadata (Cannizzaro, et al., 2021). One method used DL in an array of available imputation methods (Olier and Vellido, 2008), and Karlaš et al. (2020) used generative topographic mapping to fill gaps in time series data.

4. Discussion

4.1. Differences between domains

Ecological and evolutionary studies differed markedly from data science in terms of preferred imputation methods. Within ecological and evolutionary studies, ML and phylogenetic imputation methods dominate all four time periods, with kNN, phylogenetic imputation, and RF being particularly prominent. Data science studies, on the other hand, make use of a wider variety of methods across the range of data contexts found within the domain.

DL has been prominent since the earliest instances of imputation in data science studies examined here (2000–2009). Of the three data science studies within this period, two used DL methods: Zawistowski and Grzenda (2009) included self-organizing maps in a repertoire of imputation methods, and Olier and

Vellido (2008) used generative topographic mapping. Following these early studies, although not the dominant method, DL has been consistently used across various data contexts, and its applications are becoming increasingly widespread. Conversely, DL methods have only recently garnered attention within ecological and evolutionary studies, with only two studies (Skudnik and Jevšenak, 2022; Gendre et al., 2024) using DL for trait data imputation.

This difference in methodological preference is exemplified by the variations in the use of different kNN-based methods between the two domains. Within ecological and evolutionary studies, nine kNN methods are utilized across 26 studies. Considering that multiple methods are employed within individual studies, kNN for species trait data imputation is primarily dominated by gradient kNN and vanilla kNN, which are used in 43.8% and 28.1% of studies, respectively. However, in data science, 11 different kNN methodologies are applied over only 15 studies, including fuzzy case clustering kNN, a Gaussian mixture model kNN, and weighted kNN. While over time there has been a shift within ecological and evolutionary studies away from kNN to RF methods (likely due to the improved performance of RF over kNN for trait data [Penone et al., 2014]), there is a general preference by ecologists and evolutionary biologists for a subset of established and accessible methods.

4.2. Why do they differ?

Data science applies imputation across a wide range of domains, such as financial forecasting (John et al., 2019), traffic analysis (Duan et al., 2016; Chen et al., 2021), and web traffic prediction (Liu et al., 2015), while the ecological and evolutionary studies selected for this review focus solely on species trait data. Across the articles we reviewed, species trait data are often presented in a static tabular format, with predominantly discrete nominal or numerical data types. In contrast, the type of data used within data science studies is more variable, with imputation of longitudinal, image, text, or mixed type data. This difference in data structure encourages the development of innovative multimodal methods to imputation in data science compared to ecology and evolutionary work. Conventional methods are also often user-friendly, including readily available R packages for species trait data imputation (Buuren and Groothuis-Oudshoorn, 2011; Stekhoven and Bühlmann, 2012). Conversely, DL packages (such as Haghish, 2022) for imputation are less well established or require expert knowledge, which adds an additional barrier to their adoption in ecology and evolution.

Additionally, the intention of imputation has historically differed between the two domains. In ecological and evolutionary contexts, trait data are imputed primarily to facilitate downstream inference and enable the study of relationships and patterns in data. Consequently, imputed trait data isn't focused so much on accuracy as it is on propagating uncertainty to the point of inference, with multiple imputation being more appropriate (Nakagawa and Freckleton, 2008). Whereas in data science contexts, the aim of imputation is often to yield complete datasets for prediction (Shmueli, 2010; Daoud and Dubhashi, 2020), and therefore, focuses far more on point accuracy and single imputation methods. This difference explains to some extent the difference in methods used, as ML and DL methods have often followed this path of being primarily focused on accuracy, and so may not have been appropriate for use in trait data contexts. However, many modern DL methods do lend themselves to error propagation and multiple imputation, making them more applicable to ecological and evolutionary studies.

4.3. Suitability of DL for eco-evolutionary data imputation

GANs were the most commonly used DL architectures for imputation, although they were often outperformed by baseline alternatives, including when applied to categorical trait data (Gendre et al., 2024). Despite this, GANs still show the potential for outperforming conventional trait data imputation methods, with several studies finding that they performed particularly well when incorporated into hybrid models (Wang et al., 2024b). Additionally, GAN-based methods have yet to be used to impute continuous or mixed type trait data.

AEs were often found to outperform baseline alternatives, showing strong performance even under MAR and MNAR conditions, and hybrid VAE architectures have been found to be particularly promising

(He et al., 2024). Despite this promise, too few studies have yet investigated their performance with regards to tabular data under MAR and MNAR conditions to come to a robust conclusion with regards to their suitability for outperforming conventional trait data imputation methods.

Transformer-based methods consistently performed well, with no transformer models performing worse than baseline methods. These models utilize attention mechanisms (Vaswani et al., 2017), which may allow such networks the capacity to better account for inter-trait dependencies when imputing likely trait values. However, no studies in our review investigated the capacity of transformer models to impute missing data under MAR or MNAR conditions. RNNs and FFNNs also performed well across data types and missingness proportions, but as for transformers no studies utilized them for imputing MAR or MNAR data. Due to the low number of studies investigating the performance of transformers, RNNs, and FFNNs for data imputation, their potential for imputing species trait data is questionable, particularly when data are MAR or MNAR.

Multiple imputation is an essential capability for any DL method to be suitable for imputing species trait data. It allows us to propagate uncertainty in our imputed data through to our downstream analyses, yielding more realistic confidence in parameter estimates (Blomberg and Todorov, 2025). Although many DL approaches are trained explicitly for deterministic, single outputs, such as FFNNs trained to minimize mean squared error, there are several ways to incorporate predictive uncertainty and enable multiple imputation.

One approach is to modify the training objective so that the models predict parameters that approximate the probability distribution of the missing data conditional on the observed values. For example, Mixture Density Networks (Bishop, 1994) combine neural networks with mixture models to predict probability distributions rather than single values, allowing for uncertainty to be explicitly represented. Model uncertainty can also be approximated using ensembles of DL models (Lakshminarayanan et al., 2016). In this framework, multiple models are trained with different random initializations and/or on bootstrap resamples of the training data to provide an ensemble of predictions. This ensemble of predictions can then be interpreted as an approximation of the predictive distribution. Bayesian neural networks also offer a powerful, albeit computationally expensive (Lakshminarayanan et al., 2016), alternative. With Bayesian neural networks, parameters are sampled from prior distributions, which are subsequently updated through back propagation (Goan and Fookes, 2020). Repeated forward passes from these models generate different samples reflective of the predictive distribution of the missing values, naturally reflecting multiple imputation. Some DL models are inherently well suited to multiple imputation, particularly generative models. For example, VAEs, through their latent-variable framework, can be prompted multiple times to yield a distribution of likely values for a given missing value. Similarly, some models, such as TabPFN, are explicitly designed to approximate Bayesian inference, yielding posterior predictive distributions for their targets from which imputations can be sampled (Hollmann et al., 2022). Overall, through either adapted training procedures, or native capability, many DL models demonstrate strong suitability for multiple imputation.

Closely related species often have similar trait values owing to their shared evolutionary history. As such, representations of how species relate to one another can be useful in determining likely values in imputation processes. Several studies have shown that the inclusion of such phylogenetic information can improve imputation performances (Penone et al., 2014; Johnson et al., 2021). The capacity for DL models to also incorporate phylogenetic information is likely to be an important requirement if they are to serve as viable alternatives to conventional imputation methods. Phylogenetic information is often represented using phylogenetic trees, and some imputation approaches are explicitly designed to account for this (Goolsby et al., 2017). However, phylogenetic eigenvector mapping can be used to represent phylogenies in a tabular format that can be readily incorporated into ML workflows. Given their capacity to model complex relationships of input features, DL architectures have strong potential to effectively integrate tabular phylogenetic information for trait data imputation.

Within the studies found in this review, many evaluate their respective imputation methods using imputation accuracy metrics, such as RMSE or MAPE. Based on these measures we identify several approaches that are suitable for imputing species trait data, However, predictive accuracy is often limited

as an imputation metric and may be insufficient for evaluating the utility of imputed datasets in ecological and evolutionary contexts. By focusing on the recovery of missing values, these metrics implicitly treat imputed values as equivalent to observed data, potentially overlooking the extent to which important biological patterns are preserved (Blomberg and Todorov, 2025). Instead, performance metrics should focus on quantifying the preservation of insightful biological patterns and data structure, such as trait–trait relationships or phylogenetic signal. Consequently, the true performance of the methods we identify may differ when evaluated using such criteria rather than conventional accuracy metrics alone. Considering this, we highlight that future evaluations of DL imputation method performances in trait data contexts should place greater emphasis on performance measures that assess the suitability of imputed datasets in downstream analyses. Developing and adopting such metrics will be essential for determining whether improvements in predictive accuracy translate into meaningful gains for ecological and evolutionary research. Ultimately, we want to avoid judging imputation methods by predictive accuracy alone. Rather we should ascertain whether they support valid inference, an aim which requires multiple imputation rather than single imputation.

A major consideration when using DL methods for imputation is the volume of empirical data available for training. Often, DL models are trained on vast amounts of training data, a requirement that may be difficult to satisfy in species trait databases, where data are often sparse, incomplete, or taxonomically biased. Recent advances in DL may offer alternatives for overcoming this constraint. Foundational neural networks are models that are pre-trained on large, generic datasets, and then fine-tuned using comparatively small amounts of domain-specific data for a specific task. The foundational transformer model TabPFN has recently shown state-of-the-art performance in both classification and regression tasks with small- and medium-sized datasets (Hollmann et al., 2025). Additionally, models such as LEKA:LLM have made advances in harmonizing data from different domains for use in transfer learning, improving target domain performances (Zhang et al., 2025). In-context learning (ICL) likewise offers an approach for overcoming data shortfalls. ICL refers to the ability of neural networks to adapt predictions to a new task using importation provided in the input context without updating their parameters, that is, without training for the specific task (Dong et al., 2023). Specifically, the models use examples, instructions, or contextual information at inference to adapt their behaviors. This phenomenon has been used in large language models, where the use of a series of prompts can adapt a network to a previously unseen task. Powerful novel methods such as TabPFN have also started using ICL for tabular data to improve performance on small datasets (Hollmann et al., 2022).

Alternatively, several studies have shown that synthetic data can be used in place of empirical data for training DL models when data availability is limited (Gutierrez et al., 2021). Although some studies have highlighted the lack of a consensus with regards to the best approach for generating synthetic data (Petrakos et al., 2025), and others have warned of a lack of evidence for its utility (Manousakas and Aydöre, 2023), many studies found strong performance when trained on synthetic data (Nikolenko, 2019; Hollmann et al., 2025). As such, the use of virtual species in ecological and evolutionary contexts could provide suitable training data for DL imputation models. By generating realistic trait datasets under controlled ecological and evolutionary processes, virtual species could enable the creation of large training corpora that would otherwise be impossible to obtain empirically. However, further research is necessary to identify the best approach for generating them. These state-of-the-art DL techniques and models offer powerful avenues for maximizing model performance in imputing missing species trait data when available empirical data is otherwise prohibitively sparse.

Taken together, synthetic data generation, alongside advances in foundation models, transfer learning, and ICL, offers a promising pathway for overcoming the data limitations that have traditionally constrained the application of DL methods. These developments may prove particularly valuable for species trait imputation, where empirical datasets are often sparse, incomplete, and unevenly distributed across taxa.

Overall, the most promising architectures to trial at an early stage are AEs, especially VAEs and hybrid models such as WGAIN-VAE (Wang et al., 2024b); FFNNs, such as MLPs may also perform well (Lin et al., 2022), although for these models to be appropriately used in ecological and evolutionary contexts,

measures must be taken to adapt them to a multiple imputation framework. Following these, despite mixed performance, and poor performance on categorical species trait data, GANs show some promise, particularly advanced implementations such as cue Wasserstein models (Wang et al., 2024a) or when used in hybrid architectures (Ali et al., 2024). Additional architectures, not typically designed for data types common to trait data contexts, also show strong performance, such as transformers and GNNs, and have been adjusted for use on tabular continuous and categorical type data (Huang et al., 2020; You et al., 2020; Guo et al., 2021).

5. Conclusions and future directions

Species trait data are an essential resource within ecological and evolutionary studies. Their reliability is therefore paramount for improving our understanding of various ecological and evolutionary patterns and processes, as well as facilitating decisions relevant to conservation management. Missing data pose a particular issue within these datasets, and while comprehensive data collection is an ideal solution, this is infeasible at the scale necessary to fill gaps in global species trait databases. Consequently, alternative methods for handling missing data are required. In this study, we show how the methods deployed for one such technique, imputation, differ within ecological and evolutionary studies compared to the domain of data science, which has a far greater focus on the evaluation of state-of-the-art data handling techniques.

We find that ecological and evolutionary studies use a narrower set of methods compared to data science and show a lack of uptake of advanced DL methods. DL imputation methods are rapidly advancing and are gaining popularity in data science, but have been underexplored in ecology and evolution. Though current applications of DL to trait data contexts show mixed results (Skudnik and Jevšenak, 2022; Gendre et al., 2024), most DL architectures have yet to be evaluated in trait data contexts. We highlight AEs (specifically variational AEs), transformer-based architectures such as TabPFN, Bayesian neural networks, and FFNNs explicitly trained for multiple imputation as architectures that could be particularly promising for imputing missing species trait data. Trait data are most often available in tabular format; hence, architectures such as transformers and GNNs also hold promise as they have recently been used with tabular data. Alongside these general architectures, we also recommend investigations into the use of frontier DL models and techniques, including foundational models and transfer learning to help account for limited data availability for training. In particular, TabPFN is promising due to its state-of-the-art performance in both classification and regression tasks with tabular data. We highlight the need for further quantitative investigations into the applicability of these DL methods for imputing missing species trait data in ecological and evolutionary contexts and studies to compare them to conventional methods, particularly when applied to biased MAR and MNAR data. Furthermore, for these investigations to be robust, DL imputation performance must be carried out with multiple imputation in mind, and evaluated not just in terms of prediction accuracy, but also in terms of downstream task performance.

The improved uptake of state-of-the-art imputation methods could also be facilitated within ecological and evolutionary studies through interdisciplinary collaborations with data scientists. Such collaborative efforts would not only allow for rapid uptake of state-of-the-art imputation methodologies but also the development of bespoke techniques with specific data contexts in mind. Comparative studies looking into the efficacy of such novel methods would also benefit from established, well-curated datasets for use in benchmarking. In the computer vision community for instance, datasets such as ImageNet, CIFAR-10/100, PASCAL VOC, and others are well-regarded benchmark datasets, allowing the DL community to use a common ground to compare methods (Mukhoti et al., 2021). The development of similar datasets in ecological and evolutionary studies would allow for standardized analyses of novel methods, streamlining their further development and uptake by ecologists and evolutionary biologists.

Abbreviations

AE(s)	autoencoder(s)
AI	artificial intelligence
BRITS	bidirectional recurrent imputation for time series
DIP	deep image prior

DL	deep learning
FFNN	feed-forward neural network
GAIN	generative adversarial imputation network
GAN	generative adversarial network
GNN	graph neural network
kNN	*k*-nearest neighbors
LEKA	large language model–enhanced knowledge augmentation
LLM	large language model
LSTM	long short-term memory
MAPE	mean absolute percentage error
MAR	missing at random
MCAR	missing completely at random
MEM	mixed effect model
MICE	multiple imputation by chained equations
ML	machine learning
MLP	multilayer perceptron
MNAR	missing not at random
MTSIT	multivariate time series imputation with transformers
PICOC	population, intervention, comparison, outcome(s), and context
PSALSAR	protocol and reporting result with search, appraisal, synthesis, and analysis framework
ReLU	rectified linear unit
RF	random forest
RMSE	root mean squared error
RNN	recurrent neural network
SAITS	self-attention-based imputation for time series
TabPFN	tabular prior fitted network
VAE	variational autoencoder

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2026.10067.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2026.10067.

Acknowledgments. This work was supported by the Natural Environment Research Council (grant number NE/S007377/1) via the Quadrat DTP.

Author contribution. Conceptualization: M.T.T., G.L., D.FRP.B., J.MJ.T., and R.C.H.; Data curation: M.T.T.; Formal analysis: M.T.T.; Funding acquisition: G.L., D.FRP.B., J.MJ.T., and R.C.H.; Investigation: M.T.T.; Methodology: M.T.T., G.L., D.FRP.B., J.MJ.T., and R.C.H.; Project Administration: M.T.T.; Resources: M.T.T.; Software: M.T.T.; Supervision: G.L., D.FRP.B., J.MJ.T., and R.C.H.; Validation: M.T.T., G.L., D.FRP.B., J.MJ.T., and R.H.; Visualization: M.T.T.; Writing – original draft: M.T.T.; Writing – review and editing: M.T.T., G.L., D.FRP.B., J.MJ.T., and R.C.H.

Competing interests. The authors declare none.

Data availability statement. The data that support this study are available on GitHub: https://github.com/MTurner-ABDN/Imputation-Review-Data.

Funding statement. This work was supported by the Natural Environment Research Council (grant number NE/S007377/1).

References

Adhikari D, **Jiang W**, **Zhan J**, **He Z**, **Rawat DB**, **Aickelin U and Khorshidi HA** (2023) A comprehensive survey on imputation of missing data in Internet of Things. *ACM Computing Surveys 55*(7), 1–38. https://doi.org/10.1145/3533381

Ali SS, **Ren J and Wu J** (2024) Framework to improve software effort estimation accuracy using novel ensemble rule. *Journal of King Saud University – Computer and Information Sciences 36*(9), 102189. https://doi.org/10.1016/j.jksuci.2024.102189

Amerise IL, **Fionda V and Pirro G** (2024) Bipartite time series network for data imputation. In Endriss U, Melo FS, Bach K, Bugarin-Diz A, Alonso-Moral JM, Barro S and Heintz F (eds.), *Frontiers in Artificial Intelligence and Applications. IOS Press*. https://doi.org/10.3233/FAIA240756

Antonenko E, **Carreno A and Read J** (2024) Autoreplicative random forests with applications to missing value imputation. *Machine Learning 113*(10), 7617–7643. https://doi.org/10.1007/s10994-024-06584-1

Bannany O and Qahtan AA (2025) TGAIN: Missing data imputation for mixed-type relational datasets using generative adversarial networks. In Stahlbock R and Arabnia HR (eds.), *Data Science*, Vol. *2253*. Cham: Springer Nature Switzerland, pp. 364–380. https://doi.org/10.1007/978-3-031-85856-7_27

Bishop CM (1994) *Mixture Density Networks*. Birmingham: Aston University.

Blomberg SP and Todorov OS (2025) The fallacy of single imputation for trait databases: Use multiple imputation instead. *Methods in Ecology and Evolution 16*(4), 658–667. https://doi.org/10.1111/2041-210X.14494

Buuren SV and Groothuis-Oudshoorn K (2011) Mice: Multivariate imputation by chained equations in *R*. *Journal of Statistical Software 45*(3). https://doi.org/10.18637/jss.v045.i03

Cannizzaro G, **Leone M**, **Bernasconi A**, **Canakoglu A and Carman MJ** (2021) Automated integration of genomic metadata with sequence-to-sequence models. In Dong Y, Ifrim G, Mladenić D, Saunders C and Van Hoecke S (eds.), *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track 12461*. Springer International Publishing, pp. 187–203. https://doi.org/10.1007/978-3-030-67670-4_12

Cappelletti L, **Fontana T**, **Di Donato GW**, **Di Tucci L**, **Casiraghi E and Valentini G** (2020) Complex data imputation by auto-encoders and convolutional neural networks—A case study on genome gap-filling. *Computers 9*(2), 37. https://doi.org/10.3390/computers9020037

Cardoso Pereira R, **Seoane Santos M**, **Pereira Rodrigues P and Henriques Abreu P** (2020) Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. *Journal of Artificial Intelligence Research 69*, 1255–1285. https://doi.org/10.1613/jair.1.12312

Chen J, **Zhang S**, **Chen X**, **Jiang Q**, **Huang H and Gu C** (2021) Learning traffic as videos: A spatio-temporal VAE approach for traffic data imputation. In Farkaš I, Masulli P, Otte S and Wermter S (eds.), *Artificial Neural Networks and Machine Learning – ICANN 2021*, Vol. *12895*. Cham: Springer International Publishing, pp. 615–627. https://doi.org/10.1007/978-3-030-86383-8_49

Culumber ZW, **Anaya-Rojas JM**, **Booker WW**, **Hooks AP**, **Lange EC**, **Pluer B**, **Ramírez-Bullón N and Travis J** (2019) Widespread biases in ecological and evolutionary studies. *Bioscience 69*(8), 631–640. https://doi.org/10.1093/biosci/biz063

Daoud A and Dubhashi D (2020) *Statistical Modeling: The Three Cultures (Version 1)*. https://doi.org/10.48550/ARXIV.2012.04570

Dhar V (2013) Data science and prediction. *Communications of the ACM 56*(12), 64–73. https://doi.org/10.1145/2500499

Dong Q, **Li L**, **Dai D**, **Zheng C**, **Ma J**, **Li R**, **Xia H**, **Xu J**, **Wu Z**, **Liu T**, **Chang B**, **Sun X and Sui, Z.** (2023). *A Survey on In-context Learning (Version 6)*. https://doi.org/10.48550/ARXIV.2301.00234

Duan Y, **Edwards JS and Dwivedi YK** (2019) Artificial intelligence for decision making in the era of big data – Evolution, challenges and research agenda. *International Journal of Information Management 48*, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Duan Y, **Lv Y**, **Liu Y-L and Wang F-Y** (2016) An efficient realization of deep learning for traffic data imputation. *Transportation Research Part C: Emerging Technologies 72*, 168–181. https://doi.org/10.1016/j.trc.2016.09.015

Elezaj O, **Yildirim S and Kalemi E** (2018) Data-driven machine learning approach for predicting missing values in large data sets: A comparison study. In Nicosia G, Pardalos P, Giuffrida G and Umeton R (eds.), *Machine Learning, Optimization, and Big Data*, Vol. *10710*. Cham: Springer International Publishing, pp. 268–285. https://doi.org/10.1007/978-3-319-72926-8_23

Etard A, **Morrill S and Newbold T** (2020) Global gaps in trait data for terrestrial vertebrates. *Global Ecology and Biogeography 29*(12), 2143–2158. https://doi.org/10.1111/geb.13184

Fang C and Wang C (2020). *Time Series Data Imputation: A Survey on Deep Learning Approaches (Version 1)*. https://doi.org/10.48550/ARXIV.2011.11347

Fielding S, **Fayers PM**, **McDonald A**, **McPherson G and Campbell MK** (2008) Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes 6*(1), 57. https://doi.org/10.1186/1477-7525-6-57

Fraccaro M, **Kamronn S**, **Paquet U and Winther O** (2017). *A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning*. Available at http://arxiv.org/abs/1710.05741

Fuchs R, **Pommeret D and Stocksieker S** (2023) MI2AMI: Missing data imputation using mixed deep gaussian mixture models. In Nicosia G, Ojha V, La Malfa E, La Malfa G, Pardalos P, Di Fatta G, Giuffrida G and Umeton R (eds.), *Machine Learning, Optimization, and Data Science*, Vol. *13810*. Cham: Springer Nature Switzerland, pp. 211–222. https://doi.org/10.1007/978-3-031-25599-1_16

Galimard J-E, **Chevret S**, **Curis E and Resche-Rigon M** (2018) Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Medical Research Methodology 18*(1), 90. https://doi.org/10.1186/s12874-018-0547-1

Gan G and Yan Y (2025) Using neural networks for imputing missing values in insurance data. *Japanese Journal of Statistics and Data Science 8*(2), 1113–1151. https://doi.org/10.1007/s42081-025-00304-2

Gendre M, **Hauffe T**, **Pimiento C and Silvestro D** (2024) Benchmarking imputation methods for categorical biological data. *Methods in Ecology and Evolution 15*(9), 1624–1638. https://doi.org/10.1111/2041-210X.14339

Goan E and Fookes C (2020) Bayesian neural networks: An introduction and survey. In Mengersen KL, Pudlo P and Robert CP (eds.), *Case Studies in Applied Bayesian Data Science*, Vol. *2259*. Cham: Springer International Publishing, pp. 45–87. https://doi.org/10.1007/978-3-030-42553-1_3

Gomer B and Yuan K-H (2023) A realistic evaluation of methods for handling missing data when there is a mixture of MCAR, MAR, and MNAR mechanisms in the same dataset. *Multivariate Behavioral Research 58*(5), 988–1013. https://doi.org/10.1080/00273171.2022.2158776

Gonzalez-Calabuig M, **Fernández-Torres M-Á and Camps-Valls G** (2025) Generative networks for spatio-temporal gap filling of Sentinel-2 reflectances. *ISPRS Journal of Photogrammetry and Remote Sensing 220*, 637–648. https://doi.org/10.1016/j.isprsjprs.2025.01.016

González-Suárez M, **Lucas PM and Revilla E** (2012) Biases in comparative analyses of extinction risk: Mind the gap. *Journal of Animal Ecology 81*(6), 1211–1222. https://doi.org/10.1111/j.1365-2656.2012.01999.x

Goolsby EW, **Bruggeman J and Ané C** (2017) Rphylopars: Fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution 8*(1), 22–27. https://doi.org/10.1111/2041-210X.12612

Greenberg DA, **Palen WJ and Mooers AØ** (2017) Amphibian species traits, evolutionary history and environment *Batrachochytrium dendrobatidis* infection patterns, but not extinction risk. *Evolutionary Applications 10*(10), 1130–1145. https://doi.org/10.1111/eva.12520

Gromski PS, **Xu Y**, **Kotze HL**, **Correa E**, **Ellis DI**, **Armitage EG**, **Turner ML and Goodacre R** (2014) Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites 4*(2), 433–452. https://doi.org/10.3390/metabo4020433

Guo X, **Quan Y**, **Zhao H**, **Yao Q**, **Li Y and Tu W** (2021) *TabGNN: Multiplex Graph Neural Network for Tabular Data Prediction (Version 1)*. https://doi.org/10.48550/ARXIV.2108.09127

Gutierrez P, **Luschkova M**, **Cordier A**, **Shukor M**, **Schappert M and Dahmen T** (2021) Synthetic training data generation for deep learning based quality inspection. In Komuro T and Shimizu T (eds.), *Fifteenth International Conference on Quality Control by Artificial Vision*. Bellingham, WA: SPIE, p. 13. https://doi.org/10.1117/12.2586824

Haghish EF (2022) *mlim: Single and Multiple Imputation with Automated Machine Learning* (p. 0.3.0) [Dataset]. https://doi.org/10.32614/CRAN.package.mlim

He N, **Yan P**, **Liu C**, **Xu L**, **Li M**, **Van Meerbeek K**, **Zhou G**, **Zhou G**, **Liu S**, **Zhou X**, **Li S**, **Niu S**, **Han X**, **Buckley TN**, **Sack L and Yu G** (2023) Predicting ecosystem productivity based on plant community traits. *Trends in Plant Science 28*(1), 43–53. https://doi.org/10.1016/j.tplants.2022.08.015

He Y-L, **Yu J-Y**, **Li X**, **Fournier-Viger P and Huang JZ** (2024) A novel and efficient risk minimisation-based missing value imputation algorithm. *Knowledge-Based Systems 304*, 112435. https://doi.org/10.1016/j.knosys.2024.112435

Hollmann N, **Müller S**, **Eggensperger K and Hutter F** (2022). *TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second (Version 6)*. https://doi.org/10.48550/ARXIV.2207.01848

Hollmann N, **Müller S**, **Purucker L**, **Krishnakumar A**, **Körfer M**, **Hoo SB**, **Schirrmeister RT and Hutter F** (2025) Accurate predictions on small data with a tabular foundation model. *Nature 637*(8045), 319–326. https://doi.org/10.1038/s41586-024-08328-6

Huang X, **Khetan A**, **Cvitkovic M and Karnin Z** (2020). *TabTransformer: Tabular Data Modeling Using Contextual Embeddings (Version 1)*. https://doi.org/10.48550/ARXIV.2012.06678

Huang L, **Li P**, **Gao Q**, **Liu G**, **Luo Z and Li T** (2024) Diffusion probabilistic model for bike-sharing demand recovery with factual knowledge fusion. *Neural Networks 179*, 106538. https://doi.org/10.1016/j.neunet.2024.106538

Hunt LA (2017) Missing data imputation and its effect on the accuracy of classification. In Palumbo F, Montanari A and Vichi M (eds), *Data Science*. Cham: Springer International Publishing, pp. 3–14. https://doi.org/10.1007/978-3-319-55723-6_1

Irwin BWJ, **Whitehead TM**, **Rowland S**, **Mahmoud SY**, **Conduit GJ and Segall MD** (2021) Deep imputation on large-scale drug discovery data. *Applied AI Letters 2*(3), e31. https://doi.org/10.1002/ail2.31

John C, **Ekpenyong EJ and Nworu CC** (2019) Imputation of missing values in economic and financial time series data using five principal component analysis (PCA) approaches. *Central Bank of Nigeria Journal of Applied Statistics 10*(1), 51–73. https://doi.org/10.33429/Cjas.10119.3/6

Johnson TF, **Isaac NJB**, **Paviolo A and González-Suárez M** (2021) Handling missing values in trait data. *Global Ecology and Biogeography 30*(1), 51–62. https://doi.org/10.1111/geb.13185

Karlaš B, **Li P**, **Wu R**, **Gurel NM**, **Chu X**, **Wu W and Zhang C** (2020) Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. *Proceedings of the VLDB Endowment 14*(3), 255–267. https://doi.org/10.14778/3430915.3430917

Kim M, **Benabderrahmane S and Rahwan T** (2024) Interpretable machine learning model for predicting activist investment targets. *The Journal of Finance and Data Science 10*, 100146. https://doi.org/10.1016/j.jfds.2024.100146

Kitchin R and Lauriault TP (2015) Small data in the era of big data. *GeoJournal 80*(4), 463–475. https://doi.org/10.1007/s10708-014-9601-7

Lai S, **Zheng W-S**, **Hu J-F and Zhang J** (2018) Global-local temporal saliency action prediction. *IEEE Transactions on Image Processing 27*(5), 2272–2285. https://doi.org/10.1109/TIP.2017.2751145

Lakshminarayanan B, **Pritzel A and Blundell C** (2016). *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles (Version 3)*. https://doi.org/10.48550/ARXIV.1612.01474

Lalande F and Doya K (2022) Numerical data imputation: Choose kNN over deep learning. In Skopal T, Falchi F, Lokoč J, Sapino ML, Bartolini I and Patella M (eds.), *Similarity Search and Applications*, Vol. *13590*. Cham: Springer International Publishing, pp. 3–10. https://doi.org/10.1007/978-3-031-17849-8_1

LeCun Y, **Bengio Y and Hinton G** (2015) Deep learning. *Nature 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Liang W, **Li Y**, **Xie K**, **Zhang D**, **Li K-C**, **Souri A and Li K** (2023) Spatial-temporal aware inductive graph neural network for C-ITS data recovery. *IEEE Transactions on Intelligent Transportation Systems 24*(8), 8431–8442. https://doi.org/10.1109/TITS.2022.3156266

Lin W-C, **Tsai C-F and Zhong JR** (2022) Deep learning for missing value imputation of continuous data and the effect of data discretization. *Knowledge-Based Systems 239*, 108079. https://doi.org/10.1016/j.knosys.2021.108079

Liu M, **Li S**, **Yuan H**, **Ong MEH**, **Ning Y**, **Xie F**, **Saffari SE**, **Shang Y**, **Volovici V**, **Chakraborty B and Liu N** (2023) Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial Intelligence in Medicine 142*, 102587. https://doi.org/10.1016/j.artmed.2023.102587.

Liu Z, **Yan Y**, **Yang J and Hauskrecht M** (2015) Missing value estimation for hierarchical time series: A study of hierarchical web traffic. *2015 IEEE International Conference on Data Mining*, 895–900. https://doi.org/10.1109/ICDM.2015.58.

Loehle C, **Irwin L**, **Manly BFJ and Merrill A** (2015) Range-wide analysis of northern spotted owl nesting habitat relations. *Forest Ecology and Management 342*, 8–20. https://doi.org/10.1016/j.foreco.2015.01.010.

Luo Y (2022) Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics 23*(1), bbab489. https://doi.org/10.1093/bib/bbab489.

Manousakas D and Aydöre S (2023) *On the Usefulness of Synthetic Tabular Data Generation (Version 1)*. https://doi.org/10.48550/ARXIV.2306.15636.

May JA, **Feng Z and Adamowicz SJ** (2023) A real data-driven simulation strategy to select an imputation method for mixed-type trait data. *PLoS Computational Biology 19*(3), e1010154. https://doi.org/10.1371/journal.pcbi.1010154.

Mengist W, **Soromessa T and Legese G** (2020) Method for conducting systematic literature review and meta-analysis for environmental science research. *MethodsX 7*, 100777. https://doi.org/10.1016/j.mex.2019.100777.

Mukhoti J, **Kirsch A**, **van Amersfoort J**, **Torr PHS and Gal Y** (2021) *Deep Deterministic Uncertainty: A Simple Baseline (Version 3)*. https://doi.org/10.48550/ARXIV.2102.11582.

Myers TA (2011) Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures 5*(4), 297–310. https://doi.org/10.1080/19312458.2011.624490

Nakagawa S and Freckleton RP (2008) Missing inaction: The dangers of ignoring missing data. *Trends in Ecology & Evolution 23*(11), 592–596. https://doi.org/10.1016/j.tree.2008.06.014.

Nathan R (2001) The challenges of studying dispersal. *Trends in Ecology & Evolution 16*(9), 481–483. https://doi.org/10.1016/S0169-5347(01)02272-8.

Newbold T, **Scharlemann JPW**, **Butchart SHM**, **Şekercioğlu ÇH**, **Alkemade R**, **Booth H and Purves DW** (2013) Ecological traits affect the response of tropical forest bird species to land-use intensity. *Proceedings of the Royal Society B: Biological Sciences 280*(1750), 20122131. https://doi.org/10.1098/rspb.2012.2131.

Nguyen M, **He T**, **An L**, **Alexander DC**, **Feng J and Yeo BTT** (2020) Predicting Alzheimer's disease progression using deep recurrent neural networks. *NeuroImage 222*, 117203. https://doi.org/10.1016/j.neuroimage.2020.117203.

Nikolenko SI (2019) *Synthetic Data for Deep Learning (Version 1)*. https://doi.org/10.48550/ARXIV.1909.11512.

Olier I and Vellido A (2008) Advances in clustering and visualization of time series using GTM through time. *Neural Networks 21*(7), 904–913. https://doi.org/10.1016/j.neunet.2008.05.013

Penone C, **Davidson AD**, **Shoemaker KT**, **Di Marco M**, **Rondinini C**, **Brooks TM**, **Young BE**, **Graham CH and Costa GC** (2014) Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution 5*(9), 961–970. https://doi.org/10.1111/2041-210X.12232.

Petrakos NZ, **Moodie EEM and Savy N** (2025) A framework for generating realistic synthetic tabular data in a randomized controlled trial setting. *Statistics in Medicine 44*(18–19), e70227. https://doi.org/10.1002/sim.70227.

Pingi ST, **Zhang D**, **Bashar MA and Nayak R** (2024) Joint representation learning with generative adversarial imputation network for improved classification of longitudinal data. *Data Science and Engineering 9*(1), 5–25. https://doi.org/10.1007/s41019-023-00232-9.

R Core Team (2025) *_R: A Language and Environment for Statistical Computing_* [Computer software]. Available at https://www.R-project.org

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons. https://doi.org/10.1002/9780470316696.

Santos JC, **Henriques Abreu P and Santos MS** (2022) The identification of cancer lesions in mammography images with missing pixels: Analysis of morphology. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–8. https://doi.org/10.1109/DSAA54385.2022.10032420.

Sarmad Ali S, **Ren J**, **Wu J**, **Zhang K and Chao L** (2025) Advancing software project effort estimation: Leveraging a NIVIM for enhanced preprocessing. *Journal of Software: Evolution and Process 37*(1), e2745. https://doi.org/10.1002/smr.2745.

Shmueli G (2010) To explain or to predict? *Statistical Science 25*(3). https://doi.org/10.1214/10-STS330.

Skudnik M and Jevšenak J (2022) Artificial neural networks as an alternative method to nonlinear mixed-effects models for tree height predictions. *Forest Ecology and Management 507*, 120017. https://doi.org/10.1016/j.foreco.2022.120017.

So HY, **Ma J**, **Griffith LE and Balakrishnan N** (2025) Application of machine learning methods in the imputation of heterogeneous co-missing data. *Japanese Journal of Statistics and Data Science 8*(1), 691–720. https://doi.org/10.1007/s42081-025-00298-x.

Sowmya V and Kayarvizhy N (2021) An efficient missing data imputation model on numerical data. *2021 2nd Global Conference for Advancement in Technology (GCAT)*, 1–8. https://doi.org/10.1109/GCAT52182.2021.9587886

Stekhoven DJ and Bühlmann P (2012) MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597.

Stork NE, **Coddington JA**, **Colwell RK**, **Chazdon RL**, **Dick CW**, **Peres CA**, **Sloan S and Willis K** (2009) Vulnerability and resilience of tropical forest species to land-use change. *Conservation Biology 23*(6), 1438–1447. https://doi.org/10.1111/j.1523-1739.2009.01335.x.

Sun Y, **Li J**, **Xu Y**, **Zhang T and Wang X** (2023) Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications 227*, 120201. https://doi.org/10.1016/j.eswa.2023.120201.

Taugourdeau S, **Villerd J**, **Plantureux S**, **Huguenin-Elie O and Amiaud B** (2014) Filling the gap in functional trait databases: Use of ecological hypotheses to replace missing data. *Ecology and Evolution 4*(7), 944–958. https://doi.org/10.1002/ece3.989.

Urban MC, **Bocedi G**, **Hendry AP**, **Mihoub J-B**, **Pe'er G**, **Singer A**, **Bridle JR**, **Crozier LG**, **De Meester L**, **Godsoe W**, **Gonzalez A**, **Hellmann JJ**, **Holt RD**, **Huth A**, **Johst K**, **Krug CB**, **Leadley PW**, **Palmer SCF**, **Pantel JH and Travis JMJ** (2016) Improving the forecast for biodiversity under climate change. *Science 353*(6304), aad8466. https://doi.org/10.1126/science.aad8466.

Urban MC, **Travis JMJ**, **Zurell D**, **Thompson PL**, **Synes NW**, **Scarpa A**, **Peres-Neto PR**, **Malchow A-K**, **James PMA**, **Gravel D**, **De Meester L**, **Brown C**, **Bocedi G**, **Albert CH**, **Gonzalez A and Hendry AP** (2022) Coding for life: Designing a platform for projecting and protecting global biodiversity. *Bioscience 72*(1), 91–104. https://doi.org/10.1093/biosci/biab099.

Utomo D, **Hu L-C and Hsiung P-A** (2024) Deep neural network-based spatiotemporal heterogeneous data reconstruction for landslide detection. *International Journal of Data Science and Analytics 17*(1), 93–109. https://doi.org/10.1007/s41060-022-00358-5.

Vaswani A, **Shazeer N**, **Parmar N**, **Uszkoreit J**, **Jones L**, **Gomez AN**, **Kaiser L and Polosukhin I** (2017). *Attention Is All You Need (Version 7)*. https://doi.org/10.48550/ARXIV.1706.03762.

Vi BN, **Tan Nguyen D**, **Tran CT**, **Phuc Ngo H**, **Nguyen CC and Phan H-H** (2021). Multiple imputation by generative adversarial networks for classification with incomplete data. *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 1–6. https://doi.org/10.1109/RIVF51545.2021.9642138.

Violle C, **Navas M**, **Vile D**, **Kazakou E**, **Fortunel C**, **Hummel I and Garnier E** (2007) Let the concept of trait be functional! *Oikos 116*(5), 882–892. https://doi.org/10.1111/j.0030-1299.2007.15559.x.

Wang K, **Luo M**, **Deng M and Chen H** (2022). Nested Random Forest: A personalized imputation method for missing data. *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, pp. 119–126. https://doi.org/10.1109/BigDIA56350.2022.9874127.

Wang Y, **Xu X**, **Hu L**, **Fan J and Han M** (2024a) A time series continuous missing values imputation method based on generative adversarial networks. *Knowledge-Based Systems 283*, 111215. https://doi.org/10.1016/j.knosys.2023.111215.

Wang Y, **Xu X**, **Hu L**, **Liu J**, **Yan X and Ren W** (2024b) Continuous imputation of missing values in time series via Wasserstein generative adversarial imputation networks and variational auto-encoders model. *Physica A: Statistical Mechanics and its Applications 647*, 129914. https://doi.org/10.1016/j.physa.2024.129914.

Wang X, **Chen S**, **Zhou C**, **Duan S-B and Shi Z** (2025) Cross-sensor data reconstruction for optical remote sensing gap-filling with attention-enhanced multi-scale fusion network. *International Journal of Applied Earth Observation and Geoinformation 140*, 104571. https://doi.org/10.1016/j.jag.2025.104571.

Weber MG, **Wagner CE**, **Best RJ**, **Harmon LJ and Matthews B** (2017) Evolution in a community context: On integrating ecological interactions and macroevolution. *Trends in Ecology & Evolution 32*(4), 291–304. https://doi.org/10.1016/j.tree.2017.01.003.

Whitmee S and Orme CDL (2013) Predicting dispersal distance in mammals: A trait-based approach. *Journal of Animal Ecology 82*(1), 211–221. https://doi.org/10.1111/j.1365-2656.2012.02030.x.

Xie H, **Gong Y and Dong X** (2024) Spatial–temporal regularized tensor decomposition method for traffic speed data imputation. *International Journal of Data Science and Analytics 17*(2), 203–223. https://doi.org/10.1007/s41060-023-00412-w.

You J, **Ma X**, **Ding DY**, **Kochenderfer M and Leskovec J** (2020). *Handling Missing Data with Graph Representation Learning (Version 1)*. https://doi.org/10.48550/ARXIV.2010.16418.

Zawistowski P and Grzenda M (2009) Handling incomplete data using evolution of imputation methods. In Kolehmainen M, Toivanen P and Beliczynski B (eds.), *Adaptive and Natural Computing Algorithms*, Vol. *5495*. Berlin, Heidelberg: Springer, pp. 22–31. https://doi.org/10.1007/978-3-642-04921-7_3.

Zhang S (2012) Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software 85*(11), 2541–2552. https://doi.org/10.1016/j.jss.2012.05.073.

Zhang Z (2016) Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine 4*(1), 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.

Zhang X, **Zhang J**, **Mo F**, **Wang D**, **Fu Y and Liu K** (2025). *LEKA:LLM-Enhanced Knowledge Augmentation (Version 2)*. https://doi.org/10.48550/ARXIV.2501.17802.

Zhou Y, **Bouadjenek MR and Aryal S** (2024) Missing data imputation: Do advanced ML/DL techniques outperform traditional approaches?In Bifet A, Krilavičius T, Miliou I and Nowaczyk S (eds.), *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track 14950*. Springer Nature Switzerland, pp. 100–115. https://doi.org/10.1007/978-3-031-70381-2_7

Zhuang Y, **Wu F**, **Chen C and Pan Y** (2017) Challenges and opportunities: From big data to knowledge in AI 2.0. *Frontiers of Information Technology & Electronic Engineering 18*(1), 3–14. https://doi.org/10.1631/FITEE.1601883.

Cite this article: Turner MT, Leontidis G, Burslem DFRP, Travis JMJ and Henry RC (2026). Methods for imputing trait and community data in ecological and evolutionary research. *Environmental Data Science*, 5: e41. doi:10.1017/eds.2026.10067