A real data-based simulation procedure to select an imputation strategy for mixed-type trait data

Jacqueline A. May[1*], Zeny Feng[2], Sarah J. Adamowicz[1]

[1] Department of Integrative Biology & Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada.

[2] Department of Mathematics & Statistics, University of Guelph, Guelph, Ontario, Canada.

* Corresponding author

Email: mayj@uoguelph.ca

Abstract

Missing observations in trait datasets pose an obstacle for analyses in myriad biological disciplines. Imputation offers an alternative to removing cases with missing values from datasets. Imputation techniques that incorporate phylogenetic information into their estimations have demonstrated improved accuracy over standard techniques. However, previous studies of phylogenetic imputation tools are largely limited to simulations of numerical trait data, with categorical data not evaluated. It also remains to be explored whether the type of genetic data used affects imputation accuracy. We conducted a real data-based simulation study to compare the performance of imputation methods using a mixed-type trait dataset (lizards and amphisbaenians; order: Squamata). Selected methods included mean/mode imputation, *k*-nearest neighbour, random forests, and multivariate imputation by chained equations (MICE). Known values were removed from a complete-case dataset to simulate different missingness scenarios: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Each method (with and without phylogenetic information derived from mitochondrial and nuclear gene trees) was used to impute the removed values. The performances of the methods were evaluated for each trait and in each missingness scenario. A random forest method supplemented with a nuclear-derived phylogeny performed best overall, and this method was used to impute missing values in the original squamate dataset. Data with imputed values better reflected the characteristics and distributions of the original data compared to the complete-case data. However, phylogeny did not always improve performance for every trait and in every missingness scenario, and caution should be taken when imputing trait data, particularly in cases of extreme bias. Ultimately, these results support the use of a real data-based simulation procedure to select a suitable imputation strategy for a given mixed-type trait dataset. Moreover, they highlight the potential biases that complete-case usage may introduce into analyses.

Author summary

The issue of missing data is problematic in trait datasets as observations for rare or threatened species are often missing disproportionately. When only complete cases are used in an analysis, derived results may be biased. Imputation is an alternative to complete-case analysis and entails filling in the missing values using known observations. It has been demonstrated that including phylogenetic information in the imputation process improves accuracy of predicted values. However, most previous evaluations of imputation methods for trait datasets are limited to numerical, simulated data, with categorical traits not considered. Using a reptile dataset comprised of both numerical and categorical trait data, we employed a real data-based simulation strategy to select an optimal imputation method for the dataset. We evaluated the performance of four different imputation methods across different missingness scenarios (e.g. missing completely at random, values missing disproportionately for smaller species. Results indicate that imputed data better reflected the original dataset characteristics compared to complete-case data; however, the optimal imputation strategy for a given scenario was contingent on missingness scenario and trait type. As imputation performance varies depending on the properties of a given dataset, a real data-based simulation strategy can be used to provide guidance on best imputation practices.

Introduction

Trait data are used in a wide variety of biological disciplines, including evolutionary biology, community ecology, and biodiversity conservation. For instance, trait data pertaining to the life history of a species, such as longevity, metabolic rate, and generation time, are integral in studies of biological aging (1,2). Environmental trait data, such as latitude, temperature, and habitat type, may be used to identify those species most at risk of extinction (3,4). However, an extensive proportion of these trait data are often missing. Missingness may stem from a taxonomic bias: data are available in copious amounts for well-researched or charismatic species and are lacking for endangered species or those that inhabit remote environments (e.g. deep sea) (5–7). Mammal and bird taxa tend to be well sampled, and data for a large and diverse array of traits are available for many groups (8,9). However, regional and phylogenetic biases are common in trait data for groups such as reptiles and amphibians, and observations are largely limited to body size and habitat traits (9). Species traits are often tied to evolutionary history, a concept referred to as phylogenetic signal (10). Closely related species can share the characteristics that render them elusive or difficult to study (e.g. small body size), resulting in sparse or unreliable data for entire taxonomic clades (5,6,8). Certain types of trait data may also be easier to quantify (e.g. morphometric data) as opposed to traits that require arduous or invasive data collection techniques (e.g. age or reproductive data) (11; see Fig 1 for a visualization of missingness in reptiles). When trait datasets are used in studies, these biases can lead researchers to make erroneous conclusions about the data. Consequently, the development of approaches for handling missing data is an important area of research that spans across multiple biological disciplines.

Fig 1. Visualization of missingness. Visualization of missingness (proportion of present vs. missing observations) in Squamata trait data obtained from the primary literature. Superscripts indicate the original sources of the trait data: 1) amniote life history database (12,13), 2) vertebrate home range sizes dataset (14,15), 3) traits of lizards of the world (16,17) and 4) AnAge (18,19). See S1 File for further detail on trait sources.

The use of complete-case datasets can result in a large proportion of information being discarded (7,20). If data are "missing completely at random" (MCAR), the removal of cases leads to a reduction in the size of the dataset, and in turn, a reduction in statistical power (7,21). Trait data, however, are often "missing at random" (MAR): observations that are missing for a particular trait are related to known values for some other traits. Simply removing incomplete cases when data are MAR can result in biased estimations of model parameters (7,11,22). In more extreme cases, trait data may be "missing not at random" (MNAR): the reason data are missing is related to the unobserved data themselves. In such scenarios, the reason for missingness may be unclear to the researcher and thus difficult to verify empirically (23).

Imputing missing observations is a common alternative to the complete-case analysis. Imputation techniques use known observations to estimate the missing and unobserved values of a variable (or variables) of interest. Single imputation techniques such as hot deck imputation or *k*-nearest neighbour (KNN; 16) offer an efficient means for estimating missing values; however, these methods provide only a single estimate of the missing value. Random forest methods such as missForest (25) are also growing in popularity as they make no prior assumptions about the distributions of variables. Multiple imputation techniques have been developed that perform single imputation several times and are therefore capable of providing a measure of uncertainty of the imputed values (7,26). An example of a multiple imputation method is multivariate imputation by chained equations (MICE; 19), which offers numerous models for imputing data of different types. Incorporating phylogenetic information into the imputation process has also been shown to increase the accuracy of imputed values (11,28). This increase in accuracy is a result of the phylogenetic signal that is often inherent in trait data. A commonly used method for incorporating phylogenetic information into the imputation process is the use of phylogenetic eigenvectors. More specifically, methods such as phylogenetic eigenvector regression (PVR) (29) and phylogenetic eigenvector mapping (PEM) (30) employ a principal coordinates analysis (PCoA) to derive eigenvectors from a phylogenetic tree. PEM expands on the PVR method by applying an additional branch length transformation based on the Ornstein-Uhlenbeck evolutionary model (30,31). Phylogenetic eigenvectors may then be used as additional predictor variables in the imputation process (see 11,24,25).

As missing data are a major concern in trait datasets, we are motivated to consider imputing these missing values. The correlative nature of trait data makes them suitable candidates for imputation, particularly when phylogenetic signal is also present (34). In an evaluation of imputation methods using mammalian trait data, Penone *et al.* (11) found that supplementing the imputation process with phylogenetic information improved the accuracy of KNN, missForest, and MICE for several life history traits. Kim *et al.* (24) similarly found that adding phylogenetic information to MICE improved accuracy rates of estimated functional diversity metrics. However, when imputing bird demographic traits with moderate phylogenetic signal (Pagel's λ < 0.8), Johnson *et al.* (27) found that use of phylogenetic information improved error rates by a margin of less than 1%. Moreover, they suggest that the use of auxiliary traits (traits that are present in the dataset but not the target of imputation) were often sufficient for accurate imputations. In sum, these findings indicate that improvements conferred by phylogenetic imputation methods are context-dependent, contingent upon the presence of phylogenetic signal and relationships among traits in the dataset.

Trait data exist in several forms, ranging from the discrete categories of foraging behaviour to the countable number of eggs in a nest. Available trait datasets are often comprised of mixed types that contain categorical, count, and numerical data. Many contemporary imputation methods are able to estimate both categorical and numerical values. However, most previous studies have only evaluated their performances using simulated trait data, and the few studies that have utilized real data are limited to numerical traits. Additionally, phylogenetic information is usually included in the form of a multigene tree; it remains to be explored whether the type of genetic data used to construct the phylogeny affects imputation accuracy. Phylogenetic resolution varies among gene trees (36,37), and certain genes may be more or less suited for imputation in a given taxon and taxonomic rank. To determine the best-suited imputation method for a given mixed-type dataset, we propose a method-selection strategy that employs real data-based simulations. Results from the real-data simulations will address: 1) whether there is an optimal imputation strategy for a specific data type (continuous, count, categorical) and missingness scenario (MCAR, MAR, and MNAR); 2) which imputation method performs the best for a given dataset containing mixed data types; 3) whether phylogenetic information improves the imputation performance; and 4) which type of phylogenetic information is influential (mitochondrial, nuclear). The strategy proposed here may be considered for future trait-based analyses to reduce biases that may occur if researchers analyze only complete cases, bolster sample size and improve statistical power, and mitigate error rates when imputing missing values. In turn, this will facilitate the pursuit of new research directions, particularly in those fields impeded by sparsely available trait data.

Results

Performance comparison without phylogeny

In general, when missing data were generated under MCAR, error rate increased with missingness proportion; this trend was observed for all trait and method combinations (Fig 2). Under the same simulation setting, *k*-nearest neighbour (*KNN;* 24,38), random forests (*RF*; "missForest" R package 25,39) and multivariate imputation by chained equations (*MICE;* 27) outperformed mode and mean imputation for the majority of traits. However, there were exceptions to this pattern. For the categorical trait activity time, mode imputation resulted in a lower error rate than *RF* and *KNN* at 30-40% missingness (Fig 2a). Additionally, for smallest clutch, the mean imputation method outperformed *KNN* (10-40%) and *RF* (10%, 30-40%) (Fig 2d). *MICE* resulted in lower error rates than *KNN* and *RF* for most traits across all missingness proportions. However, *KNN* resulted in the lowest error rate for activity time at 10%, and *RF* resulted in the lowest error rate across all missingness proportions settings for the insular endemic trait (Fig 2b) and at 10% missingness for largest clutch (Fig 2g). In both MAR and MNAR scenarios without phylogenetic information added, *MICE* generally outperformed both *RF* and *KNN* (see Fig 3).

Fig 2. MCAR performance without phylogeny. Performance of the methods mean imputation, *KNN*, missForest (*RF*), and multivariate imputation by chained equations (*MICE*) across different proportions of missingness when data were MCAR. *MICE_LR* and *MICE_PMM* signify the use of logistic regression and predictive mean matching for imputing categorical and numerical traits, respectively. Error rate was measured as PFC for the categorical traits a) activity time and b) insular endemic and as MSE for the numerical traits c) largest clutch, d) smallest clutch, e) female snout-vent length (SVL), f) maximum SVL, and g) latitude. In both cases, error rates closer to 0 are indicative of better performance.

Fig 3. Imputation performance across all missingness scenarios. Comparison of error rates for the methods mode imputation, *KNN*, *RF*, and *MICE* for different missingness scenarios with and without the addition of phylogenetic information. Phylogenetic information was added in the form of trees built from sequence data of mitochondrial cytochrome *c* oxidase subunit I (COI) and nuclear oocyte maturation factor (c-mos) and recombination activating gene 1 (RAG1). Performance was quantified using PFC for the categorical traits a) activity time and b) insular endemic and using MSE for the numerical traits c) largest clutch, d) smallest clutch, e) female SVL, f) maximum SVL, and g) latitude. MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random.

Phylogenetic imputation performance

All traits exhibited significant phylogenetic signal in all gene trees (S1 Fig; see S1 File for more details on phylogenetic signal measures). However, improvements to imputation performance through the addition of phylogeny were contingent on method, data type, and missingness scenario (Fig 3). For instance, when considering the categorical trait activity time, supplementing phylogenetic information from any of the three genes generally improved performance for each method and in each missingness scenario (Fig 3a). On the contrary, in the case of the binary trait insular endemic, adding phylogenetic information to *MICE* at low missingness levels (10%) resulted in an increased error rate (Fig 3b). For most traits, MAR results reflected those in the MCAR scenarios; however, deviations from the general pattern occurred in some MNAR cases. For example, in the MNAR scenario for insular endemic, phylogeny was only beneficial when nuclear information was added to *KNN*.

For the traits largest clutch, smallest clutch, and latitude, *KNN* and *RF* performances were improved by the addition of any type of phylogenetic information in the MCAR and MAR scenarios; this was particularly evident in the case of nuclear oocyte maturation factor (c-mos) (Fig 3c-d, g). However, phylogeny did not improve *MICE* performance consistently for these traits. In the MAR scenarios, phylogenetic information improved *MICE* performance for smallest clutch and latitude; conversely, for largest clutch, any type of phylogenetic information increased error rate for *MICE*. In the MNAR scenarios, the addition of any type of phylogenetic information increased error rate for *MICE* imputation for all of these traits drastically in several situations (e.g. more than doubling the error rate for largest clutch and latitude). The traits female snout-vent length (SVL) and maximum SVL displayed somewhat dissimilar patterns from the other traits (Figs 4e-f) as phylogenetic information tended to decrease imputation performance for most methods and in most scenarios.

The relationship between phylogenetic signal and error ratio varied depending on data type. For categorical traits, higher error ratio, indicative of better performance due to phylogeny, was associated with higher phylogenetic signal strength (Fig 4a). This same pattern was not observed for numerical traits (Fig 4b). Moreover, in MNAR scenarios for numerical traits, many error ratio values fell below 1 at higher levels of phylogenetic signal, indicative of a reduction in performance due to phylogeny. Generally, the improvement in imputation performance resulting from phylogeny was most apparent for *KNN* and *RF*, as these methods account for the majority of error ratio values greater than 1; error ratio values for *MICE*, however, often fell below 1, particularly in the case of numerical traits.

Fig 4. Association between error ratio and phylogenetic signal. Association between error ratio (error rate without phylogeny/error rate with phylogeny) and phylogenetic signal for the c-mos gene (Fritz and Purvis' *D* (40) for categorical traits and Pagel's λ (41) for numerical traits) at different proportions of missingness. Error ratio values above 1 (indicated by the gray line)

signify an improvement in performance when phylogeny is added. In the case of *D,* lower values are indicative of higher levels of phylogenetic conservation for the trait; conversely, higher values of λ suggest stronger phylogenetic signal. Results are not shown for MAR in a) as only one trait (activity time) was simulated for this scenario. To improve visualization, values were jittered (random noise introduced to data) using the package "ggplot2" (42). Additionally, results are only shown for the c-mos gene as results for cytochrome *c* oxidase subunit I (COI) and recombination activating gene 1 (RAG1) follow similar patterns.

Imputation of original dataset using best strategy

Although results varied considerably, particularly in MNAR scenarios, the method that resulted in the lowest error rates overall was *RF* with c-mos. Consequently, this method was chosen to impute the original dataset. Out of the total species in the original dataset (*n* = 6657), those with available c-mos sequence records were included in the imputed subset (*n* = 921). The proportion of missingness varied for each trait in this subset as 0.16 for activity time, 0 for insular endemic, 0.21 for largest clutch, 0.21 for smallest clutch, 0.23 for female SVL, 0 for maximum SVL, and 0 for latitude. As insular endemic, maximum SVL, and latitude had complete observations in this subset, these traits were not imputed.

Distributions and categorical frequencies of the complete-case, original, and imputed data can be observed in Fig 5. For the trait activity time, when compared to the original data, discrepancies in the categorical frequencies were more apparent in the complete-case data than in the imputed data (Fig 5a). The complete-case data displayed a greater overrepresentation of the rarest category (cathemeral: 11% vs. 8.9%) and underrepresentation of the most common category (diurnal: 57.9% vs. 64.4%). Conversely, the imputed data displayed a greater representation of observations in the most common category compared to the original data (diurnal: 67.5% vs. 64.4%). For all numerical traits, the imputed data distributions followed the distributions of the original more closely than did the complete-case distributions (Figs 6b-d; Table 1). Perhaps most apparent are the discrepancies in the maximum values in the complete-case data compared to those in the original and imputed data (e.g. for largest clutch, 68 vs. 88; for smallest clutch, 8 vs. 30). Although the discrepancies in the complete-case data were greater, both complete-case and imputed data displayed reduced variance relative to the original data for the traits largest clutch, smallest clutch, and female SVL.

Fig 5. Comparison of quantitative characteristics across datasets. Comparison of a) categorical frequencies for the trait activity time and distributions for the traits b) largest clutch, c) smallest clutch, and d) female SVL of the complete-case, original, and imputed data. The natural logarithm (ln) of the numerical data were taken to improve visualization.

Discussion

In agreement with previous evaluations of imputation methods using trait data (11,34,43), there was no “optimal” method for imputing values in all scenarios. In the absence of phylogeny, the best overall method for imputing mixed-type trait data was *MICE*. This trend was apparent even in cases of MNAR, as *MICE* resulted in the lowest error rates for five out of seven traits in these scenarios when phylogeny was not included. *MICE* demonstrated strong performances in previous evaluations of imputation techniques in mammalian (11) and plant (43) trait datasets. Furthermore, the robustness of predictive mean matching is appealing for the non-linear relationships and non-normal distributions commonly observed in numerical trait data (44,45). This may explain the superior performance of *MICE* in the case of smallest clutch, a count trait with a right-skewed distribution (many species with smallest clutch size = 1).

Table 1. Summary statistics for the complete-case, original, and imputed datasets.

	Largest clutch (# eggs/neonates)			Smallest clutch (# eggs/neonates)			Female SVL (mm)			Maximum SVL (mm)		Latitude (°)	
	CC	*O*	*I*	*CC*	*O*	*I*	*CC*	*O*	*I*	*CC*	*O*	*CC*	*O*
N	141	731	921	141	731	921	137	705	921	152	921	152	921
Min	1	1	1	1	1	1	18.7	18.7	18.7	21.7	21.7	-40.36	-47.89
Max	68	88	88	8	30	30	499.5	534.3	534.3	1170	1170	56.6	56.6
Range	67	87	87	7	29	29	480.8	515.6	515.6	1148.3	1148.3	96.96	104.49
Median	2	3	3	1	2	2	60.1	62.7	65.2	77	80	-11.36	-9.48
Mean	5.79	6.08	6.06	1.63	2.06	2.14	75.82	83.4	84.36	103.44	110.35	1.65	-3.8
SE (mean)	0.67	0.33	0.27	0.09	0.07	0.06	4.76	2.43	2.07	8.58	3.30	2.03	0.75
0.95 CI (mean)	1.33	0.65	0.53	0.18	0.14	0.12	9.41	4.76	4.06	16.96	6.48	4.01	1.48
Variance	63.85	79.46	67.41	1.22	3.90	3.46	3103.49	4151.01	3950.16	11197.05	10048.39	627.38	521.53
Standard deviation	7.99	8.91	8.21	1.10	1.98	1.86	55.71	64.43	62.85	105.82	100.24	25.05	22.84

Summary statistics of the complete-case (*CC*), original (*O*), and imputed (*I*) datasets for the numerical traits largest clutch, smallest clutch, female snout-vent length (SVL), and latitude. As the proportion of missingness was 0 for the traits maximum SVL and latitude in the original data subset, these traits were not imputed. Original trait data obtained from Meiri (16).

Predictive mean matching has also been shown to perform well on smaller sample sizes (45), as seen in the current study (*n* = 152). Its use in trait imputation is therefore an appealing option when phylogenetic information is scarce.

As reported in previous studies (11), imputation error rates tended to increase with missingness proportion and varied amongst different traits. Adding phylogenetic information, however, did not always improve imputation performance; on the contrary, in some instances its inclusion led to increased error rates. The effect of phylogeny therefore appears to be situational and linked to the method used, the underlying mechanism of the missingness in the data, and quantitative attributes and evolutionary history of the target trait. The performances of *KNN* and *RF* were often improved when any type of phylogenetic information was provided, even in some cases of MNAR. This pattern was more prominent at higher missingness proportions, as phylogeny can offset the loss of the trait data. Conversely, phylogeny often increased the error rate for *MICE*. This increase in error rate was also found in Johnson *et al.* (34) when phylogenetic information was added to *MICE*, particularly in MNAR scenarios (e.g. larger values more likely to be missing). The authors suggest this may stem from an issue relating to the large number of eigenvectors used in the imputation process (e.g. more than 20 eigenvectors were included in biased missingness scenarios). Penone *et al.* (11) restricted their maximum number of eigenvectors to 10 and suggest that the use of too many eigenvectors can mask the information provided by other traits in the imputation process. Indeed, in the current study, *MICE* performed well when the number of predictors were low, as in the case of trait-only imputation. As phylogenetic resolution varies between nuclear and mitochondrial gene trees, the number of eigenvectors used for imputation varied in accordance. In this study, the 65% variation method was used to determine the number of eigenvectors to be included; however, it is possible that the use of too many eigenvectors (i.e. more than 40; 62), with less information provided by each eigenvector, would introduce more noise or lead to overfitting by the regression-based models. Thus, analyses using phylogenetic eigenvectors for imputation may consider the use of tree-based methods such as *RF* (or recursive partitioning; see Kim *et al.* (32)) that are more robust to high-dimensional data. Future studies may also consider exploring whether the optimal number of phylogenetic eigenvectors to use for imputation changes under varying degrees of missingness bias.

RF with phylogeny demonstrated the strongest performance overall as it resulted in the lowest error rates across all missingness scenarios. This result supports previous evaluations of the effectiveness of *RF* for mixed-type data (25). For both *KNN* and *RF*, adding phylogenetic information reduced imputation error rate for traits of all types (categorical, count, continuous). Nuclear-derived phylogenetic information (i.e. c-mos or RAG1) generally conferred a greater improvement in imputation performance relative to mitochondrial COI. Due to their faster rates of nucleotide substitution, mitochondrial genes are less adept at resolving deeper phylogenetic relationships relative to nuclear genes (47). Consequently, the relationships resolved by nuclear gene trees may more closely follow the evolutionary trajectory of the traits used in this study. However, COI often still conferred a reduction in error rate, in some cases more so than the nuclear genes (e.g. smallest clutch); mitochondrial sequences therefore should be used when nuclear data are unavailable and may be more advantageous when studying more closely related species. Strength of phylogenetic signal also appeared to correlate with error ratio (i.e. the magnitude of performance enhancement) for categorical traits. The same pattern was not apparent for numerical traits, however. This may stem from the limited range of phylogenetic signal observed for these other types: all genes displayed significant levels of phylogenetic signal for all traits, many of which verged toward λ = 1 (higher trait conservation). This may suggest that the boost in performance due to phylogeny is negligible beyond a certain level of phylogenetic signal. However, imputation of a greater number and variety of traits that do not display any evidence of phylogenetic signal would need to be included to test this assertion.

The comparison between the distributions and categorical frequencies of the complete-case, original, and imputed trait data support the efficacy of imputation for mixed-type data. A greater than 6-fold increase in sample size when using imputed data (n = 151 for complete-case vs. n = 921 for imputed data) is striking and illustrates the information loss that can occur when using a complete-case approach. Moreover, complete-case data often do not capture the true variability of the data; instead, they comprise a biased subset and, in turn, the potential for erroneous inferences. Previous studies using clinical data (64) and mammalian trait data (11) found that inferences derived from imputed datasets are less biased when compared to those obtained using complete-case datasets. However, the missing values in these studies were introduced either completely at random (MCAR) or at random (MAR). Although imputation performs well under MCAR and MAR, the mechanism of missingness is often difficult to determine in practice (23,49). Imputation has been shown to perform poorly in scenarios with biased missingness, such as when extreme values or values in the tails of the distribution of the population are disproportionately missing (34). The results from our study provide reason for further discretion in these instances as the most extreme error rates were observed in MNAR scenarios. If data are truly MNAR and the imputation method is not carefully chosen, imputed values and the inferences derived therein may be inaccurate. A recent study completed by Jardim et al. (50) suggests that accurate estimation of phylogenetic signal from imputed datasets is contingent on several variables, including the amount of missing data, missing mechanism, and the evolutionary trajectory of the trait itself. For example, as values closer to the equator were missing in the latitude MNAR scenario, mean imputation outperformed most other imputation methods. Due to the prevalence of allopatric speciation modes in diversification (51,52), closely related species can inhabit different latitudes or distributions; traits with such evolutionary histories may be less suitable for imputation. Therefore, we agree with Johnson *et al.* (34) and Jardim *et al.* (50) that caution should be taken when imputing data and the properties of the dataset of interest be inspected beforehand. Testing imputation methods using a real data-based simulation strategy as we demonstrate here would provide useful insight as to whether imputation is a suitable alternative to complete-case analysis.

As is often the case when constructing a complete-case dataset, several traits were excluded from this study. These included many categorical traits that were invariant in the complete-case dataset, such as those containing information about geography or habitat. In turn, the range of phylogenetic signal for traits was also limited. It was therefore not feasible to truly gauge the relationship between error ratio and phylogenetic signal strength in traits as they all exhibited significant levels. The continued collection of high-quality trait data for both known and novel species is necessary to further probe these types of relationships. For instance, in the case of Squamata, snake species are disproportionally undersampled (9) and were thus not included in the current study. An increase in data availability would also facilitate additional research on the use of imputation methods in real datasets. Simulated trait data do not fully capture the nuances of real datasets, and comparative evaluations using real data and different taxonomic groups are needed to test whether imputing values is practical, particularly in cases of severe biases.

Missingness in datasets is a pervasive issue in the realm of biological research. It is particularly problematic for those taxonomic groups threatened by extinction, or that are small or reside in understudied areas of the globe. As trait data can take on many forms, methods that can accurately predict missing values for diverse data types are invaluable for the study of these obscure groups. Previous research has focused largely on numerical data, and consideration of imputation performance for categorical traits is imperative in driving this field forward. The results presented here provide support for the use of imputation methods in real mixed-type datasets. Supplementing these methods with phylogenetic information is often beneficial, even if sequence data are available for only one or a limited number of markers. However, researchers should take care to understand the properties of their dataset and consider the ramifications of using imputation. In such situations, a real data-based simulation strategy can provide guidance on best imputation practices for a given biological or ecological dataset. Simulating missingness using real data more accurately reflects the characteristics and the nature of the unobserved values. The imputation method that is robust in these scenarios and across diverse trait types can be used to bolster sample size while simultaneously preserving the original properties of a dataset. Derived inferences may then more accurately represent the biological phenomena under investigation.

Materials and methods

Complete-case dataset creation

Traits are defined here as characteristics that are typical of a species. These may refer to characteristics relating to the biology of a species or the environment in which it resides. Data for squamates (lizards and amphisbaenians; order: Squamata) were selected for analysis as complete-case observations were available for at least 100 species as well as both categorical and numerical traits. In addition, these species had DNA sequence records publicly available for both mitochondrial and nuclear markers. Squamata represent an incredibly diverse group of vertebrates (~10,000 species; 30), inhabiting disparate environments and boasting a broad range of morphological features. However, trait data for Squamata are undersampled relative to mammal and bird groups, particularly in tropical regions that are home to diverse species at risk (9). As of 2022, 19.6% of squamate species are estimated to be under threat of extinction (54). Imputation may offer additional avenues to identify those traits correlated with risk status in squamates (e.g. 32,33) and in doing so, contribute to biodiversity conservation efforts in vulnerable areas. Trait data were obtained from a dataset published by Meiri (16) (other datasets were also considered, see S1 File). This dataset contains information about the habitat, life history, morphology, behaviour, and conservation threat level of 6,657 squamate species (lizards and amphisbaenians, not including snakes) (34,35). The raw trait data were downloaded into R v. 4.0.3 (57).

The Barcode of Life Data System (BOLD) (58) was used as the source for mitochondrial sequence data as it contains thousands of published cytochrome *c* oxidase subunit I (COI) partial gene sequence records (16,676 sequences for over 2000 Squamata species as of July 16th, 2021). COI sequence data were downloaded into R on March 12th, 2020 (59). Data were filtered for records that have been identified to the species level, as this information was necessary for trait matching purposes. Additional quality control checks on the sequence data included trimming N and gap content from sequence ends and removing sequences with greater than 1% of internal N and/or gap content across their entire sequence length. Sequences between 650 and 1000 bp were retained to facilitate downstream multiple sequence alignment. As multiple COI sequence records are available for many species, a centroid sequence selection process was employed to find a typical representative sequence for each species (Orton et al., 2019; see S1 File for details on this process). The *AlignTranslation* function from the R package "DECIPHER" v. 2.18.1 (61,62) was used to perform a multiple sequence alignment on the centroid sequences. *AlignTranslation* was used as it performs a multiple sequence alignment guided by the translated amino acid sequence, which is more reliable than an alignment based on nucleotide data alone (61). The translated final alignment was visualized using the *ggmsa* function from the R package "ggmsa" v. 0.06 (42) to verify the nucleotides were in the correct reading frame and to check for the presence of stop codons. Nuclear sequence data were obtained from a multigene alignment published in Pyron *et al.* (64,65). This alignment is comprised of sequence data for 12 genes (seven nuclear, five mitochondrial) and 4161 species of Squamata (64). The alignment was partitioned into its constituent gene alignments using RAxML v. 8 (66).

Species names from the COI alignment were matched against the species names in the trait dataset. Those species that had available data for at least five traits (both categorical and numerical) and a corresponding COI sequence record were then matched against the species names in the nuclear multigene alignment. The nuclear markers oocyte maturation factor (c-mos) and recombination activating gene 1 (RAG1) had the largest number of available records for the species in the complete-case dataset and were selected for analyses (see S1 Table for sequence identifiers of those records selected). Final checks were performed on the trait data in the complete-case subset. Categorical traits with severe class imbalances and very low variability (e.g. more than 90% of observations in one of the categories and/or the remaining observations sparsely dispersed across other categories), such as reproductive mode, geographic range, and substrate, were excluded from the study. The distributions of numerical trait data were visualized to check for the presence of severe outliers. For each numerical trait, an upper threshold was calculated as follows: quartile 3 + (3 × the interquartile range of the data). Severe outliers are defined here as those values that exceed the upper threshold. If identified, these values were verified in the primary literature to ensure they were real datapoints and not the result of data entry error. The final dataset contains information for the seven most complete traits, including the categorical traits: activity time and insular endemic, the count traits: largest clutch and smallest clutch, and the continuous traits: female snout-vent length (SVL), maximum snout-vent length (SVL), and latitude (geographic centroid for the species; Roll et al. 2017). The final dataset is referred to as the "complete-case dataset" including, 152 species, representing 25 Squamata families (S2 Table). To maintain a sufficient sample size, we permitted some missing values (no more than 10% for each trait) present in the so-called "complete-case dataset"; otherwise, the sample size will drop to 121 if only species without missing values in their traits are included. For further details on these traits, see S3 Table.

Phylogenetic information

The alignments for the COI, c-mos, and RAG1 sequences were used to build maximum likelihood gene trees in RAxML v. 8 (66). The model GTRGAMMAI was specified (option -m), and the alignment was partitioned based on codon position (option -q). The gene trees were then read into R and made ultrametric using the *chronos* function in the R package "ape" v. 5.4.1 (68). Phylogenetic eigenvectors were extracted from each gene tree and for each trait using the "MPSEM" package v. 0.3.6 in R (47). To prevent overfitting, the number of eigenvectors that explained greater than or equal to 65% of the phylogenetic structure variance was used (see S1 File for further details on this process). Following the method of Penone *et al.* (11), the phylogenetic eigenvectors were appended to the complete-case dataset and treated as predictors in the model to impute the missing value of a given trait.

Previous studies have suggested that phylogenetic signal strength in simulated trait data is positively correlated with imputation accuracy (32,70). To assess this association using real data, we measured phylogenetic signal for each trait using Pagel's λ (41) for numerical traits and the *D* metric (40) for categorical traits. Pagel's λ is estimated using maximum likelihood and represents the value that optimally transforms a phylogenetic variance-covariance matrix to fit the observed trait data structure. A λ value of 0 indicates no phylogenetic signal (star-shaped phylogeny), whereas a λ value of 1 suggests that the trait data adhere to a Brownian motion (BM) model of evolution (41). The *D* metric represents whether the number of transitions of a binary trait varies from the expected number under a BM model (40). A *D* value of 0 indicates that the trait data adhere to a BM model, and a *D* value of 1 indicates that there is no phylogenetic signal in the trait data. A *D* value greater than 1 signifies phylogenetic overdispersion. Alternatively, a *D* value less than 0 suggests the trait is phylogenetically conserved (40). These metrics were calculated separately for each trait using each gene tree (S1 File). The *phylosig* function in the R package "phytools" v. 0.7.70 (51) and *phylo.d* function in the R package "caper" v. 1.0.1 (52) were used to measure λ and *D*, respectively.

Imputation process

Four imputation methods were considered: mean/mode imputation, *k*-nearest neighbour (*KNN*) ("VIM" package v. 6.1.0; 16), random forests (*RF*) ("missForest" package v. 1.4; 53,54), and multivariate imputations by chained equations (*MICE*) ("mice" package v. 3.13.0; 19). Mean (for numerical traits) / mode (for categorical traits) imputation, the simplest method, was used as a baseline for comparison. The remaining methods were chosen due to their popularity in trait-based studies (e.g. 27,55) and capacity to impute both continuous and categorical traits. These methods have also been evaluated in previous studies of trait data imputation (11,34,43). *KNN* and *RF* are single imputation methods as they provide a single estimation of the missing value. *MICE* is a multiple imputation method that performs imputation *m* times on the dataset with missing values, resulting in *m* imputed datasets. The *MICE* algorithm utilizes chained equations to estimate missing values and offers several different models for imputing data. In this study, the predictive mean matching model was used to estimate missing continuous data. Predictive mean matching is the default model for continuous data in *MICE* and performed well in previous evaluations using trait data (43,75). Predictive mean matching fills the missing observation with a random value selected from a "donor" pool for the missing observations. This pool is created by fitting a regression model on the observed data and selecting *k* fitted values that are closest to the predicted value for the missing observation (44,45). Logistic regression is a common approach for predicting missing categorical data and is the default method for imputing categorical data in *MICE*. Logistic regression and polytomous logistic regression models were used to impute values for the binary trait insular endemic and the nominal multi-categorical trait activity time, respectively. To obtain a final imputed value for *MICE*, the mean and mode values were taken across the *m* datasets for numerical traits and categorical traits, respectively. See S1 File for further details on imputation algorithms.

When imputing the missing values of each trait ("target trait") using the observed values of the other traits ("auxiliary traits"), not all of the auxiliary traits are useful for imputing the missing values of the target trait. Association tests between each pair of traits were used to filter out irrelevant auxiliary traits and build a more parsimonious imputation model for the target trait. Regression models were used in the association tests in which the target trait was specified as the response variable and each one of the auxiliary traits was specified as the covariate. Linear regression, Poisson regression, and logistic regression models were used for continuous, count, and categorical target traits, respectively. Only auxiliary traits with a coefficient not significantly equal to zero were retained in the imputation model for a particular target trait. Finally, as methods such as *KNN* are sensitive to the range of the data, numerical traits were natural log-transformed prior to imputation.

Simulation study

To simulate missing data, three different missingness scenarios were considered: 1) missing completely at random ("MCAR"); 2) missing at random ("MAR"); and 3) missing not at random ("MNAR"). Within the MCAR scenario, missing values were randomly introduced into the complete-case dataset at different proportions (0.10, 0.20, 0.30, and 0.40). In cases where traits had values that were already missing (up to 10%), missing values were introduced on top of these (i.e. up to 50% missingness). To reduce stochasticity and maintain a fair comparison of imputation performance across different missing proportions, and not introducing variability relating to species identity, missing data for each increase in proportion (e.g. from 0.10 to 0.20 missingness) were added upon the missing values of the previous proportion. To simulate MAR scenarios using real data, logistic regression models were fitted to the original Meiri (16) dataset (*n* = 6657) to identify which auxiliary traits were significantly associated with the missingness for each target trait. In the fitted model, the indicator of whether an observation is missing or not was treated as the response variable and auxiliary traits specified as predictors. The fitted models were then used to introduce missing values into the complete-case datasets (for further details see S1 File). To test how the imputation methods perform in cases of extreme bias, MNAR scenarios were simulated for each trait. Values were removed from the 10th percentile of the tail of data distribution for numerical biological traits, e.g., the 10th percentile of the lower latitudes (between 10° and -10°); and from a single category for categorical traits, e.g., "nocturnal" category for activity time and "yes" category for insular endemic. These values were removed to emulate realistic MNAR scenarios for Squamata (see S1 File for further information).

A range of parameters and their values were considered for the different imputation methods (see S1 File for details on this process). The parameters that resulted in the lowest error rate were used in the imputation model. Imputations using only trait data were first performed on the simulated missing dataset. Imputations were again performed using trait data and phylogenetic eigenvectors derived from either COI, RAG1, or c-mos gene trees. This amounted to 78 different combination settings with respect to method and missingness scenario. The entire process was repeated 100 times for each combination of settings, resulting in 7,800 runs of the simulation and imputation pipeline procedure (see Fig 6 for a visualization of the process).

Fig 6. Workflow of the pipeline for a particular combination of variables. 1) 20% of the trait observations are removed missing completely at random (MCAR) from the complete-case dataset; 2) missing values are imputed using *k*-nearest neighbour (*KNN*). Phylogenetic information in the form of a cytochrome *c* oxidase subunit I (COI) gene tree and known trait data are used to estimate the missing trait data; and 3) the imputed values are compared to those in the complete-case dataset. Mean squared error (MSE) or proportion falsely classified (PFC) are calculated for numerical and categorical traits, respectively, and averaged across 100 replicates.

Evaluation of methods

To assess imputation accuracy, imputed values were compared against the known values in the complete-case dataset. Mean squared error (MSE) rates and proportion falsely classified (PFC) rates were computed for numerical and categorical traits, respectively. These rates were averaged across the 100 replicates for each combination of methods for each trait. For both metrics, values closer to 0 are indicative of better performance. The packages “ggplot2” v. 3.3.5 (42) and “plotly” v. 4.10.0 (76) were used to visualize results in R.

Real data imputation application and comparison

To select the most suitable method for imputing missing values in the original trait dataset, the results of the MAR simulations were first considered as these mimic realistic biological scenarios. In case of more than one method performing equally well, the method that was most robust across different missingness scenarios and that resulted in the lowest average error rate for the majority of traits was selected. To investigate whether imputed values alter the quantitative distributional characteristics of the data, summary statistics for each trait were calculated using the dataset that includes imputed values and compared with the corresponding summary statistics of both the original and complete-case datasets. To investigate whether the phylogenetic information improves the imputation accuracy for a given trait and imputation method, the following error ratio was calculated for each trait and each method:

$$Error\ ratio = \frac{Error\ rate\ (MSE\ or\ PFC)\ without\ phylogeny}{Error\ rate\ (MSE\ or\ PFC)\ with\ phylogeny}$$

An error ratio value greater than 1 indicates an improvement in imputation performance resulting from the addition of phylogenetic information. To observe the trend of the effect of phylogenetic signal strength on the imputation of different traits, the error ratio values were plotted against the λ and *D* metrics for numerical and categorical traits, respectively.

Acknowledgements

We would like to thank Dr. Cameron Nugent and Dr. Karl Cottenie for their helpful insights regarding the design and structure of the simulation pipeline. We also thank Dr. Tyler Elliott for his valuable comments on the manuscript and code. Finally, we thank many researchers who have collected trait and sequence data and made them publicly available. This work would not have been possible without you.

References

1. Voituron Y, de Fraipont M, Issartel J, Guillaume O, Clobert J. Extreme lifespan of the human fish (*Proteus anguinus*): a challenge for ageing mechanisms. Biol Lett. 2011;7(1):105–7.

2. Valcu M, Dale J, Griesser M, Nakagawa S, Kempenaers B. Global gradients of avian longevity support the classic evolutionary theory of ageing. Ecography. 2014 Oct 1;37(10):930–8.

3. Howard SD, Bickford DP. Amphibians over the edge: silent extinction risk of Data Deficient species. Divers Distrib. 2014 Jul 1;20(7):837–46.

4. Pacifici M, Visconti P, Butchart SHM, Watson JEM, Cassola FM, Rondinini C. Species' traits influenced their response to recent climate change. Nat Clim Change. 2017 Mar 1;7(3):205–8.

5. Garamszegi LZ, Møller AP. Nonrandom variation in within-species sample size and missing data in phylogenetic comparative studies. Syst Biol. 2011;60(6):876–80.

6. González-Suárez M, Lucas PM, Revilla E. Biases in comparative analyses of extinction risk: mind the gap. J Anim Ecol. 2012 Nov 1;81(6):1211–22.

7. Nakagawa S, Freckleton RP. Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol. 2008;23:592–6.

8. Titley MA, Snaddon JL, Turner EC. Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. PLOS ONE. 2017 Dec 14;12(12):e0189577.

9. Etard A, Morrill S, Newbold T. Global gaps in trait data for terrestrial vertebrates. Glob Ecol Biogeogr. 2020;29(12):2143–58.

10. Blomberg SP, Garland TJr, Ives AR. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution. 2003;57:717–45.

11. Penone C, Davidson AD, Shoemaker KT, Di Marco M, Rondinini C, Brooks TM, et al. Imputation of missing data in life-history trait datasets: which approach performs the best? Methods Ecol Evol. 2014;5:961–70.

12. Myhrvold NP, Baldridge E, Chan B, Sivam D, Freeman DL, Ernest SKM. An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles. Ecology. 2015;96(11):3109.

13. Nathan P. Myhrvold, Elita Baldridge, Benjamin Chan, Dhileep Sivam, Daniel L. Freeman, S. K. Morgan Ernest. Data from: An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles [Internet]. Wiley. Collection.; 2016. Available from: https://wiley.figshare.com/articles/dataset/Full_Archive/3563457

14. Tamburello N, Côté IM, Dulvy NK. Energy and the Scaling of Animal Space Use. Am Nat. 2015 Aug 1;186(2):196–211.

15. Tamburello N, Côté IM, Dulvy NK. Data from: Energy and the Scaling of Animal Space Use. Dryad Dataset. 2015;

16. Meiri S. Traits of lizards of the world: Variation around a successful evolutionary design. Glob Ecol Biogeogr. 2018;27(10):1168–72.

17. Meiri S. Data from: Traits of lizards of the world: Variation around a successful evolutionary design. Dryad Dataset [Internet]. 2019; Available from: https://doi.org/10.5061/dryad.f6t39kj

18. de Magalhães JP, Costa J. A database of vertebrate longevity records and their relation to other life-history traits. J Evol Biol. 2009;22:1770–4.

19. Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, et al. Human Ageing Genomic Resources: new and updated databases. Nucleic Acids Res. 2018 Jan 4;46(D1):D1083–90.

20. Zhang Z. Missing data imputation: focusing on single imputation. Ann Transl Med. 2015;4(1):9.

21. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.

22. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087–91.

23. van Buuren S. Flexible Imputation of Missing Data. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2012.

24. Kowarik A, Templ M. Imputation with the R Package VIM. J Stat Softw. 2016;74(7):1–16.

25. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012 Jan 1;28(1):112–8.

26. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8:3–15.

27. Van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67.

28. Swenson NG. Phylogenetic imputation of plant functional trait databases. Ecography. 2014;37:105–10.

29. Diniz-Filho JAF, Ramos de Sant'ana CE, Bini LM. An eigenvector method for estimating phylogenetic inertia. Evolution. 1998;52:1247–62.

30. Guénard G, Legendre P, Peres-Neto P. Phylogenetic eigenvector maps: a framework to model and predict species traits. Methods Ecol Evol. 2013 Dec 1;4(12):1120–31.

31. Guénard G. A phylogenetic modelling tutorial using Phylogenetic Eigenvector Maps (PEM) as implemented in R package MPSEM (0.3-6). 2019.

32. Kim SW, Blomberg SP, Pandolfi JM. Transcending data gaps: a framework to reduce inferential errors in ecological analyses. Ecol Lett. 2018;21(8):1200–10.

33. Fournier A, Penone C, Pennino MG, Courchamp F. Predicting future invaders and future invasions. Proc Natl Acad Sci U S A. 2019/03/29 ed. 2019 Apr 16;116(16):7905–10.

34. Johnson TF, Isaac NJB, Paviolo A, González-Suárez M. Handling missing values in trait data. Glob Ecol Biogeogr. 2021 Jan 1;30(1):51–62.

35. James TD, Salguero-Gómez R, Jones OR, Childs DZ, Beckerman AP. Bridging gaps in demographic analysis with phylogenetic imputation. Conserv Biol. 2021;35(4):1210–21.

36. Keck BP, Near TJ. Assessing phylogenetic resolution among mitochondrial, nuclear, and morphological datasets in Nothonotus darters (Teleostei: Percidae). Mol Phylogenet Evol. 2008 Feb 1;46(2):708–20.

37. Blom MPK, Bragg JG, Potter S, Moritz C. Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards. Syst Biol. 2017 May 1;66(3):352–66.

38. Templ M, Kowarik A, Alfons A, de Cillia G, Prantner B, Rannetbauer W. R package "VIM": Visualization and Imputation of Missing Values [Internet]. 2021. Available from: https://cran.r-project.org/web/packages/VIM/VIM.pdf

39. Stekhoven DJ. missForest: Nonparametric Missing Value Imputation using Random Forest. 2013.

40. Fritz SA, Purvis A. Selectivity in Mammalian Extinction Risk and Threat Types: a New Measure of Phylogenetic Signal Strength in Binary Traits. Conserv Biol. 2010 Aug 1;24(4):1042–51.

41. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999 Oct 1;401(6756):877–84.

42. Wickham H. ggplot2: Elegant Graphics for Data Analysis. [Internet]. New York: Springer-Verlag; 2016. Available from: https://ggplot2.tidyverse.org

43. Poyatos R, Sus O, Badiella L, Mencuccini M, Martínez-Vilalta J. Gap-filling a spatially explicit plant trait database: comparing imputation methods and different levels of environmental information. Biogeosciences. 2018;15:2601–17.

44. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Med Res Methodol. 2014 Jun 5;14(1):75.

45. Kleinke K. Multiple Imputation by Predictive Mean Matching When Sample Size Is Small. Methodology. 2018 Jan 1;14(1):3–15.

46. Diniz-Filho JAF, Bini LM, Rangel TF, Morales-Castilla I, Olalla-Tárraga MÁ, Rodríguez MÁ, et al. On the selection of phylogenetic eigenvectors for ecological analyses. Ecography. 2012 Mar 1;35(3):239–49.

47. Springer MS, DeBry RW, Douady C, Amrine HM, Madsen O, de Jong WW, et al. Mitochondrial Versus Nuclear Gene Sequences in Deep-Level Mammalian Phylogeny Reconstruction. Mol Biol Evol. 2001;18(2):132–43.

48. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol. 2019;110:63–73.

49. Enders CK. Applied Missing Data Analysis. New York: The Guilford Press; 2010. (Methology in the Social Sciences).

50. Jardim L, Bini LM, Diniz-Filho JAF, Villalobos F. A Cautionary Note on Phylogenetic Signal Estimation from Imputed Databases. Evol Biol. 2021 Jun 1;48(2):246–58.

51. Esquerré D, Brennan IG, Catullo RA, Torres-Pérez F, Keogh JS. How mountains shape biodiversity: The role of the Andes in biogeography, diversification, and reproductive biology in South America’s most species-rich lizard radiation (Squamata: Liolaemidae). Evolution. 2019;73(2):214–30.

52. Skeels A, Cardillo M. Reconstructing the Geography of Speciation from Contemporary Biodiversity Data. Am Nat. 2019 Feb 1;193(2):240–55.

53. Uetz P, Aguilar P, Hošek J, editors. The Reptile Database. 2021; Available from: http://www.reptile-database.org

54. Cox N, Young BE, Bowles P, Fernandez M, Marin J, Rapacciuolo G, et al. A global reptile assessment highlights shared conservation needs of tetrapods. Nature [Internet]. 2022 Apr 27; Available from: https://doi.org/10.1038/s41586-022-04664-7

55. Böhm M, Williams R, Bramhall HR, McMillan KM, Davidson AD, Garcia A, et al. Correlates of extinction risk in squamate reptiles: the relative importance of biology, geography, threat and range size. Glob Ecol Biogeogr. 2016;25(4):391–405.

56. Munstermann MJ, Heim NA, McCauley DJ, Payne JL, Upham NS, Wang SC, et al. A global ecological signal of extinction risk in terrestrial vertebrates. Conserv Biol [Internet]. n/a(n/a). Available from: https://doi.org/10.1111/cobi.13852

57. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/

58. Ratnasingham S, Hebert PDN. bold: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes. 2007 May 1;7(3):355–64.

59. Data from: Barcode of Life Data System: DS-IMPMIX2: Squamata cytochrome c oxidase subunit I (COI) dataset. [Internet]. 2020. Available from: dx.doi.org/10.5883/DS-IMPMIX2

60. Orton MG, May JA, Ly W, Lee DJ, Adamowicz SJ. Is molecular evolution faster in the tropics? Heredity. 2019 May 1;122(5):513–24.

61. Wright ES. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics. 2015 Oct 6;16(1):322.

62. Wright ES. RNAconTest: comparing tools for noncoding RNA multiple sequence alignment based on structural consistency. RNA. 2020 May 1;26(5):531–40.

63. Yu G, Zhou L, Huang H. Package "ggmsa". Plot Multiple Sequence Alignment using 'ggplot2. 2021.

64. Pyron RA, Burbrink FT, Wiens JJ. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evol Biol. 2013 Apr 29;13(1):93.

65. Pyron RA, Burbrink FT, Wiens JJ. Data from: A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. Dryad Dataset [Internet]. 2013; Available from: https://doi.org/10.5061/dryad.82h0m

66. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

67. Roll U, Feldman A, Novosolov M, Allison A, Bauer AM, Bernard R, et al. The global distribution of tetrapods reveals a need for targeted reptile conservation. Nat Ecol Evol. 2017 Nov 1;1(11):1677–82.

68. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35:526–8.

69. Guénard G, Legendre P. Modeling Phylogenetic Signals using Eigenvector Maps. 2019.

70. Molina-Venegas R, Moreno-Saiz JC, Castro Parga I, Davies TJ, Peres-Neto PR, Rodríguez MÁ. Assessing among-lineage variability in phylogenetic imputation of functional trait datasets. Ecography. 2018 Oct 1;41(10):1740–9.

71. Revell L. phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3:217–23.

72. Orme D, Freckleton RP, Thomas G, Petzoldt T, Fritz S, Isaac N, et al. Package "caper": Comparative Analyses of Phylogenetics and Evolution in R. 2018.

73. Stekhoven DJ, Buehlmann P. MissForest - non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8.

74. Richards C, Cooke RSC, Bates AE. Biological traits of seabirds predict extinction risk and vulnerability to anthropogenic threats. Glob Ecol Biogeogr. 2021 May 1;30(5):973–86.

75. Taugourdeau S, Villerd J, Plantureux S, Huguenin-Elie O, Amiaud B. Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data. Ecol Evol. 2014;4(7):944–58.

76. Sievert C. Interactive Web-Based Data Visualization with R, plotly, and shiny. [Internet]. Florida: Chapman and Hall/CRC; 2020. Available from: https://plotly-r.com

Supporting information

S1 File. Supplementary Information.

S1 Fig. Phylogenetic signal measurements. Measures of phylogenetic signal for a) categorical and b) numerical traits in gene trees constructed for mitochondrial COI and nuclear c-mos and RAG1. Asterisks indicate significance at the 0.05 level, according to results from hypothesis tests comparing the results to a null model (no phylogenetic signal). Fritz and Purvis' *D* metric (40) and Pagel's λ (41) were used to measure phylogenetic signal for categorical and numerical traits, respectively. In the case of *D,* lower values are indicative of higher levels of phylogenetic conservation for the trait; conversely, higher values of λ suggest stronger phylogenetic signal. As the *D* metric only measures the phylogenetic signal of binary traits, the three-level categorical trait AT was broken down into the binary traits "AT: Diurnal" and "AT: Nocturnal".

S1 Table. Sequence identifiers.

S2 Table. Taxonomic composition of complete-case trait dataset (n = 152).

S3 Table. Descriptions and additional details for traits in the complete-case dataset.

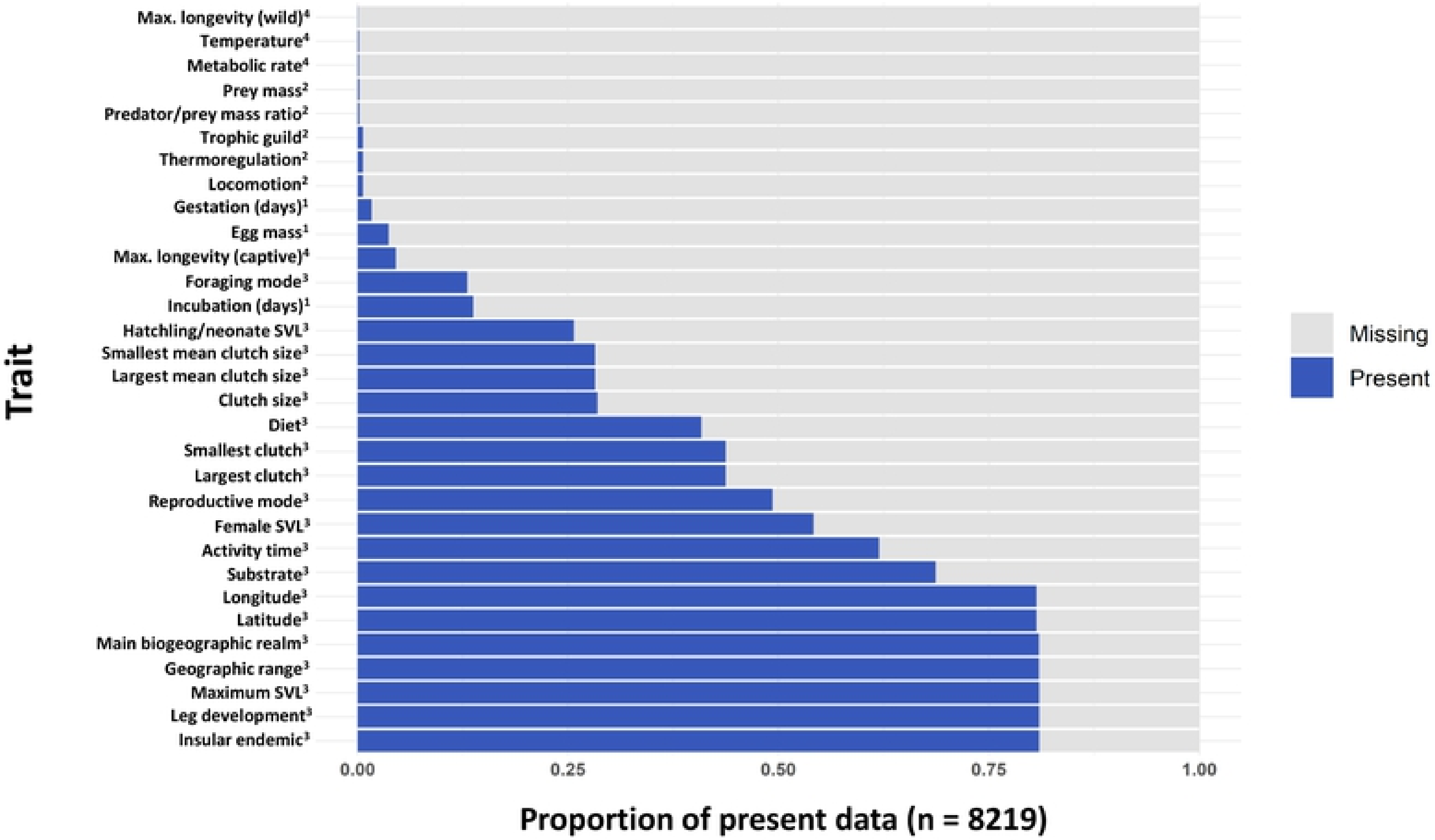
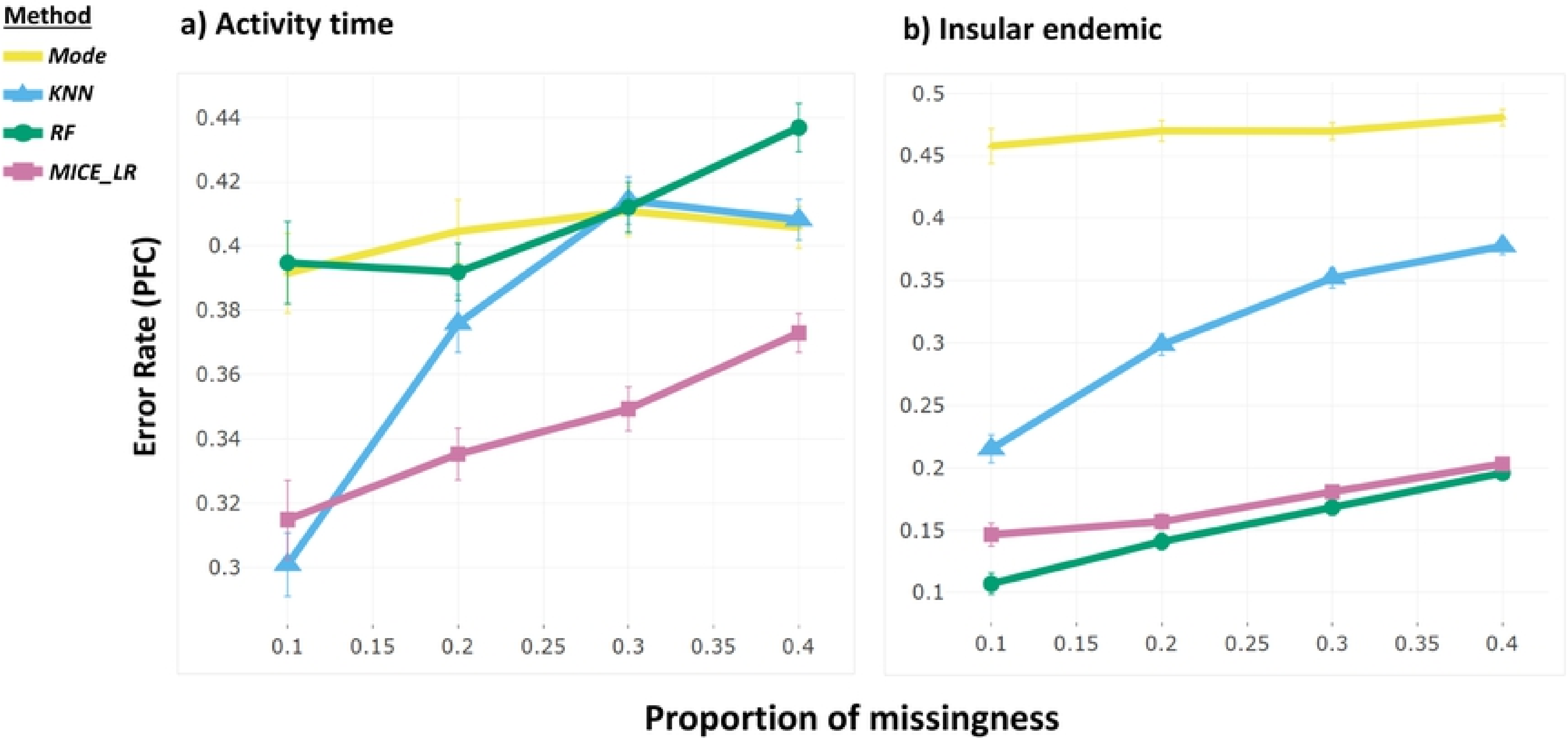
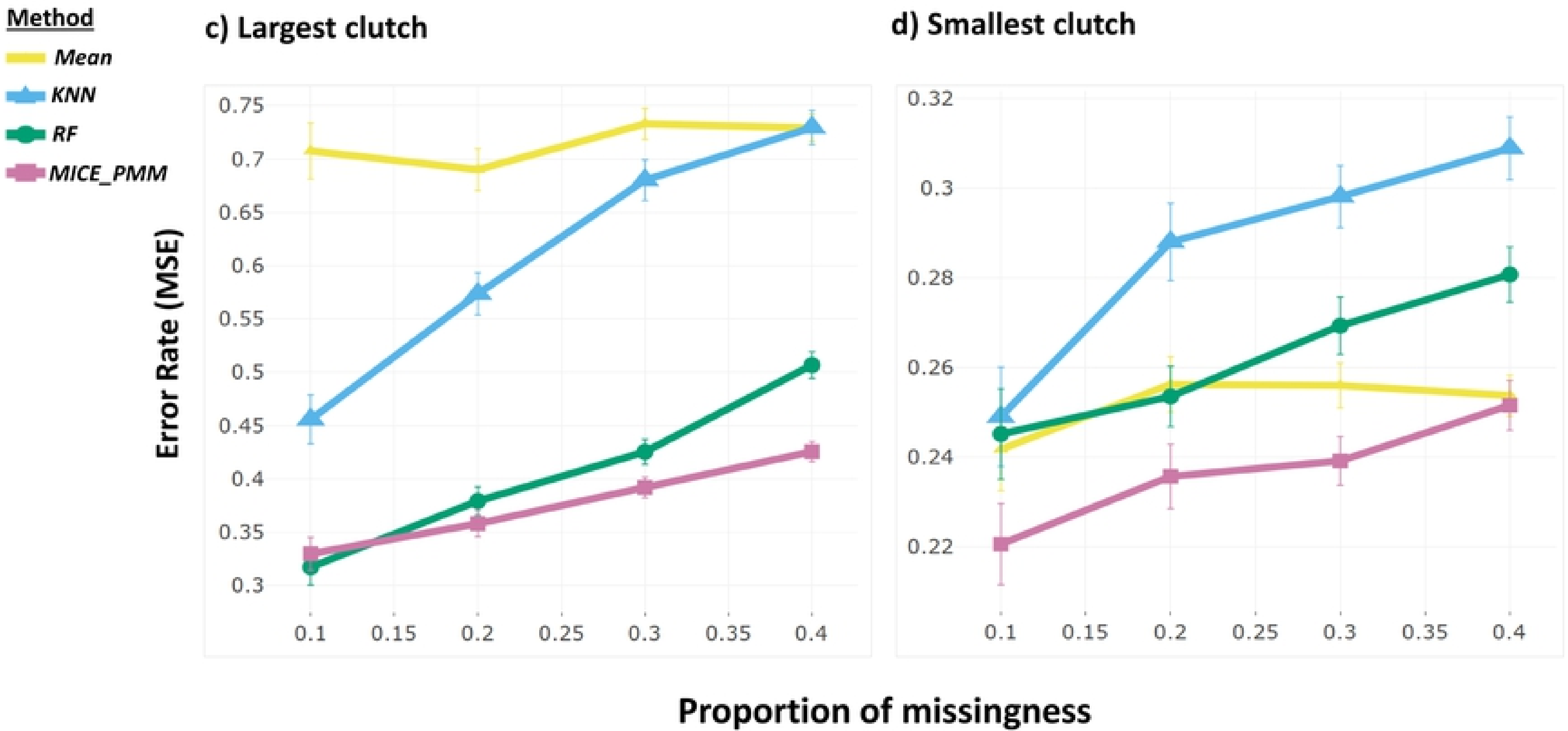
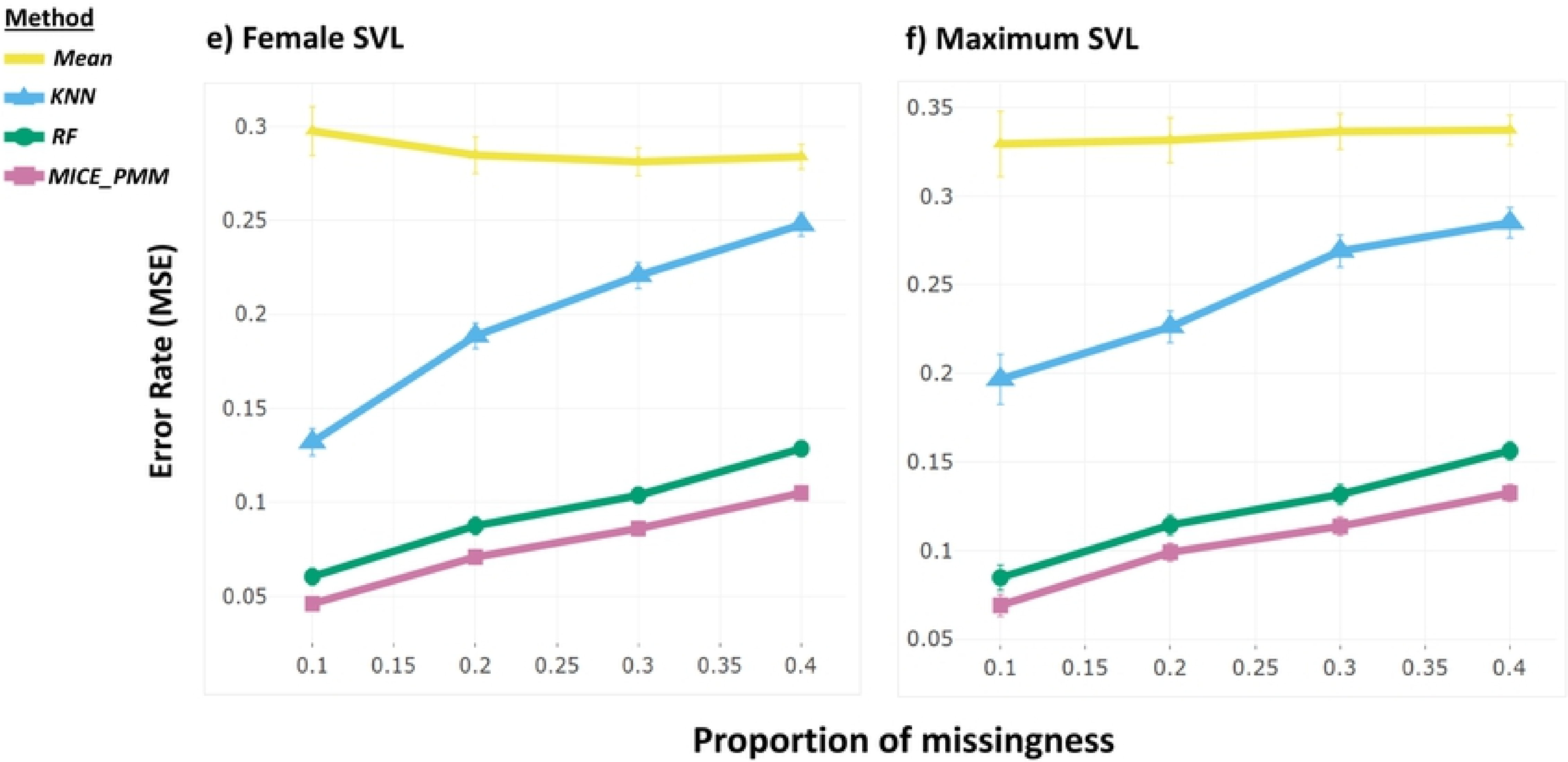
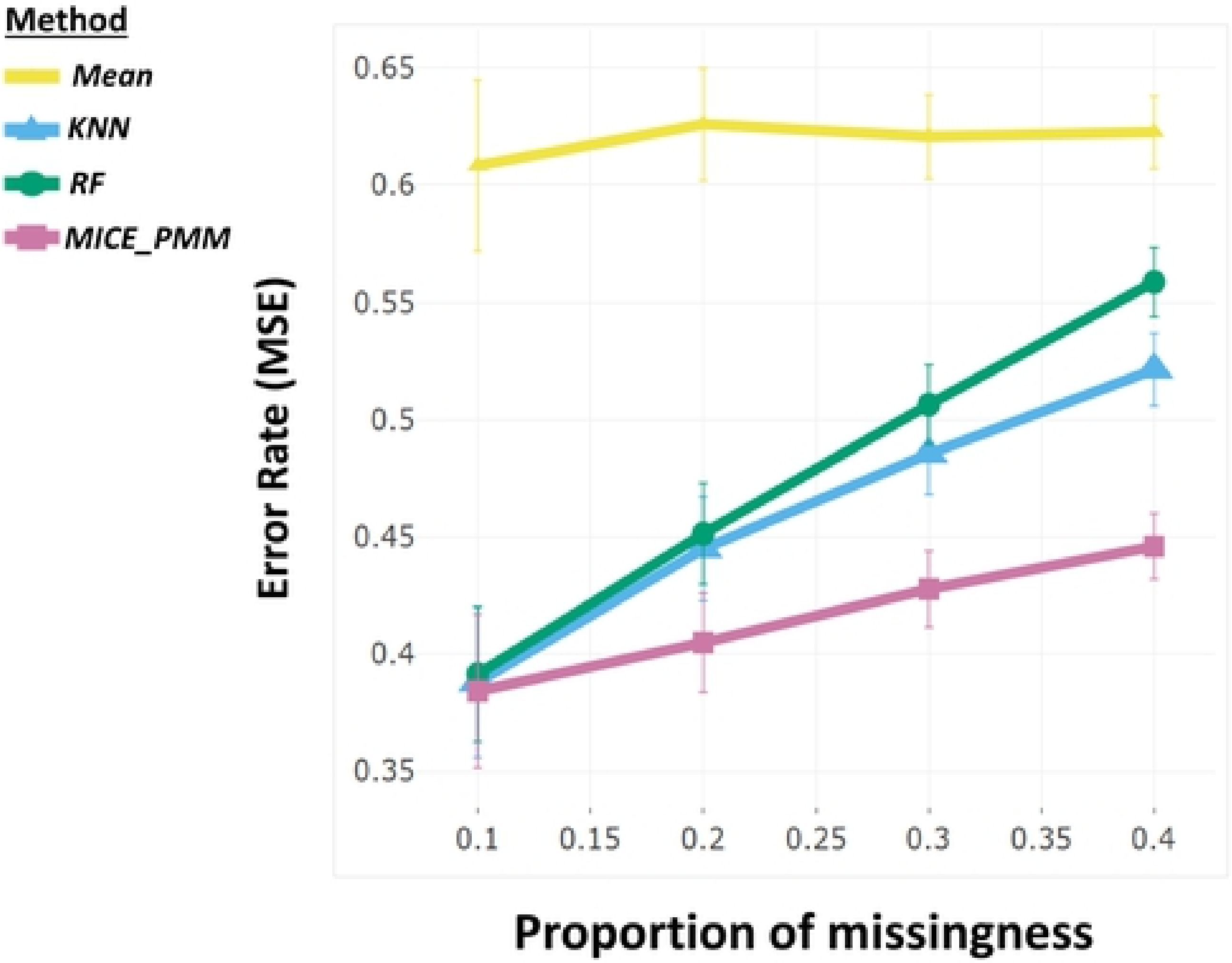

Max. longevity (wild)4
Temperature4
Metabolic rate4
Prey mass2
Predator/prey mass ratio2
Trophic guild2
Thermoregulation2
Locomotion2
Gestation (days)1
Egg mass1
Max. longevity (captive)4
Foraging mode3
Incubation (days)1
Hatchling/neonate SVL3
Smallest mean clutch size3
Largest mean clutch size3
Clutch size3
Diet3
Smallest clutch3
Largest clutch3
Reproductive mode3
Female SVL3
Activity time3
Substrate3
Longitude3
Latitude3
Main biogeographic realm3
Geographic range3
Maximum SVL3
Leg development3
Insular endemic3
Trait
0.00
0.25
0.50
0.75
1.00
Proportion of present data (n = 8219)
Missing
Present

Fig 1


Method
Mode
KNN
RF
MICE_LR
a) Activity time
b) Insular endemic
Error Rate (PFC)
0.44
0.42
0.4
0.38
0.36
0.34
0.32
0.3
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Proportion of missingness

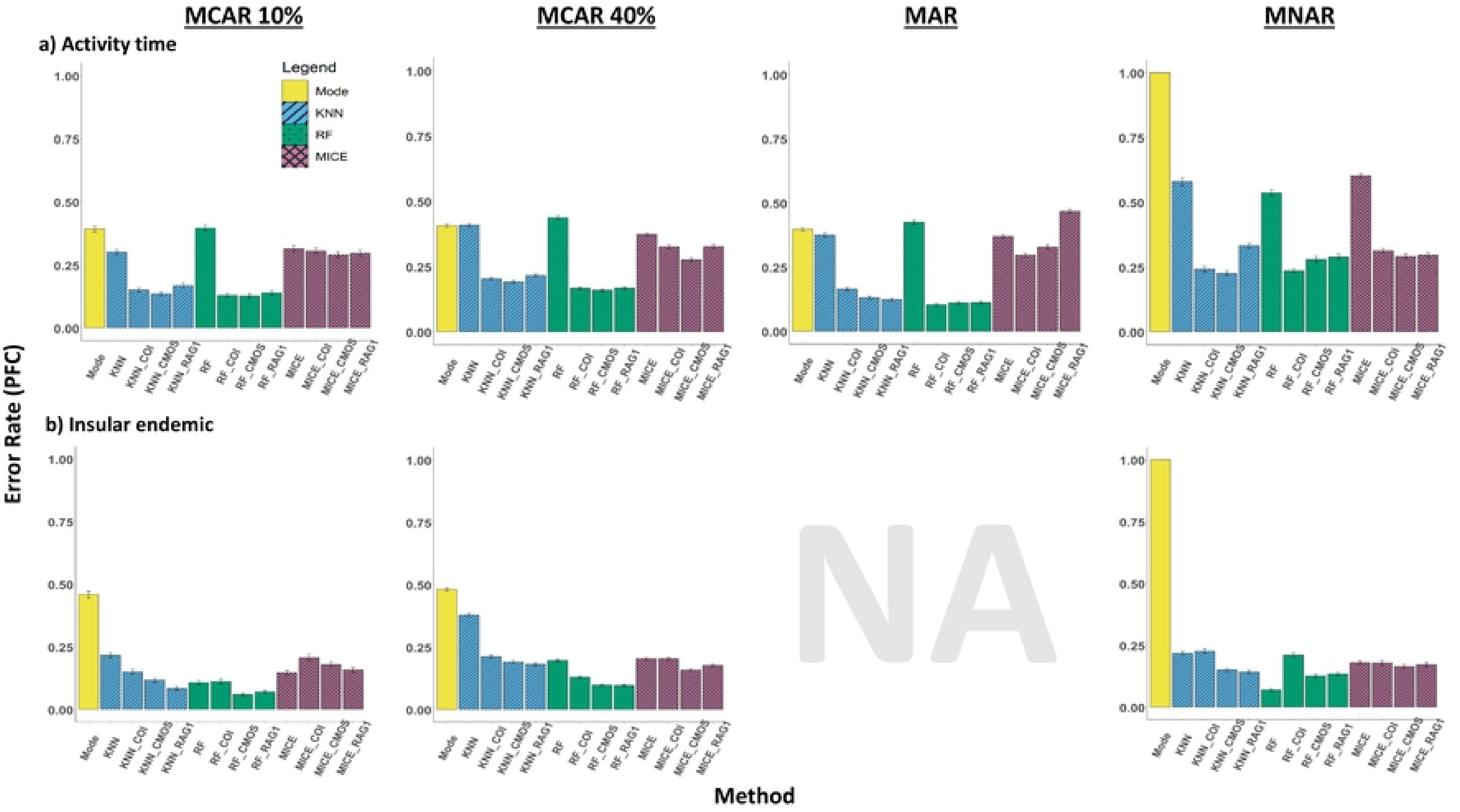

Method
Mean
KNN
RF
MICE_PMM
c) Largest clutch
d) Smallest clutch
Error Rate (MSE)
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.32
0.3
0.28
0.26
0.24
0.22
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Proportion of missingness

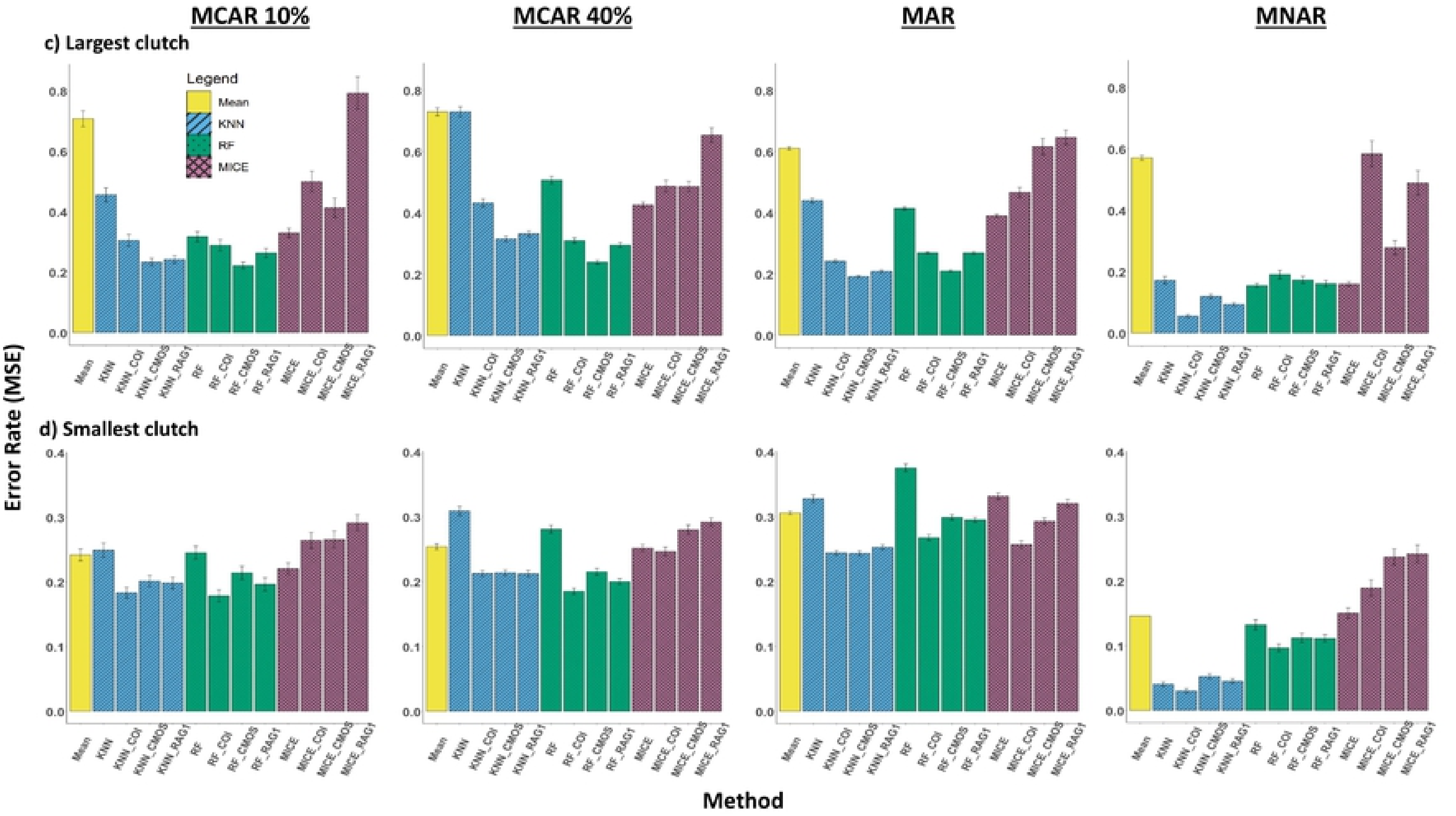

Method
Mean
KNN
RF
MICE_PMM
e) Female SVL
f) Maximum SVL
Error Rate (MSE)
0.3
0.25
0.2
0.15
0.1
0.05
0.35
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Proportion of missingness

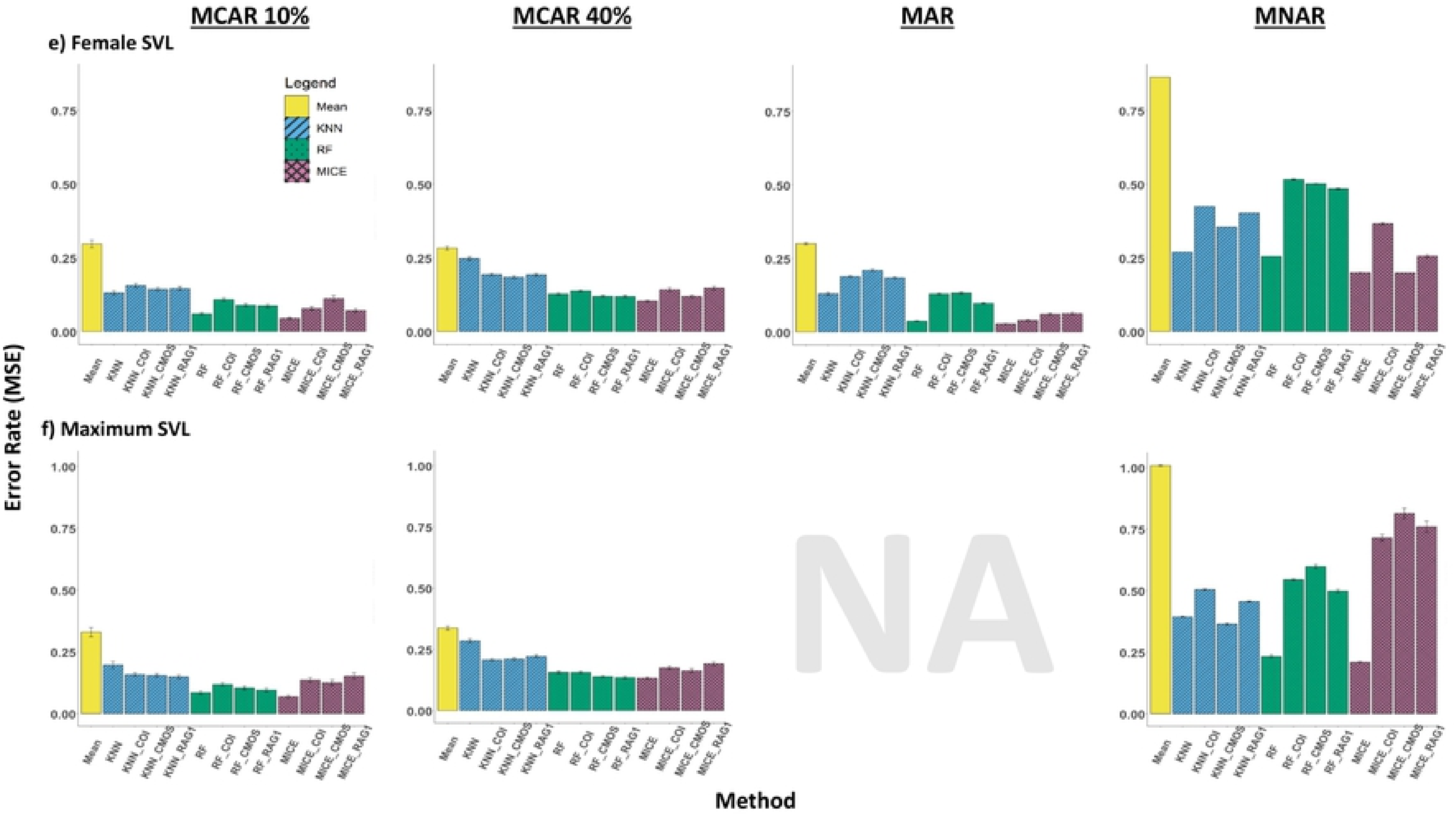


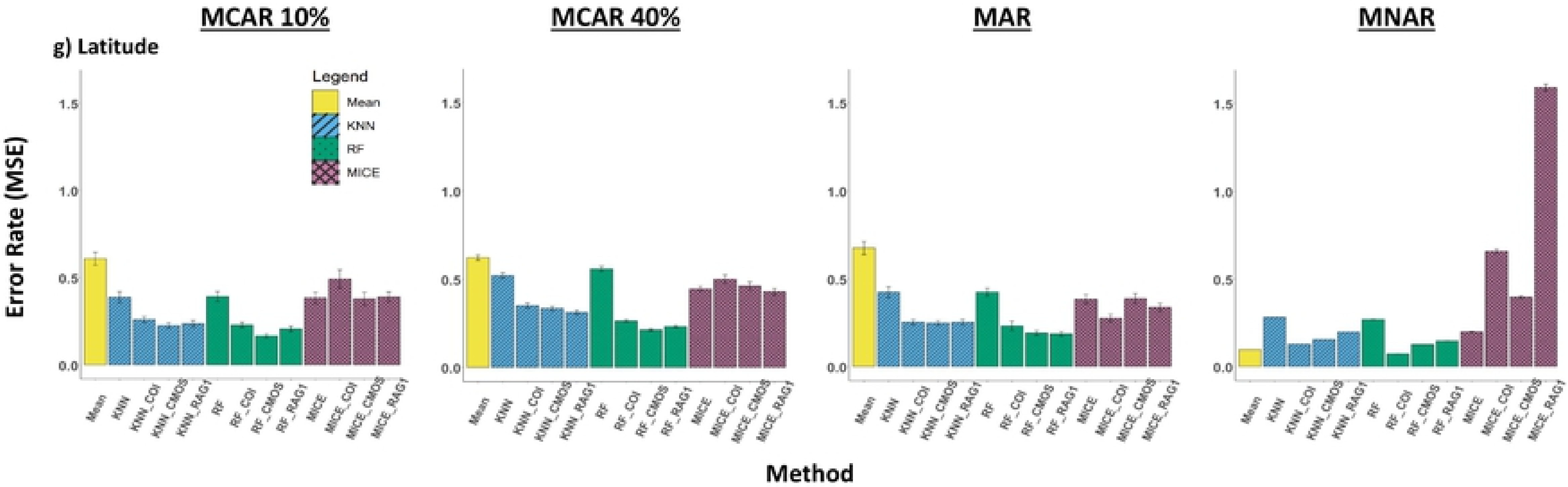
g) Latitude

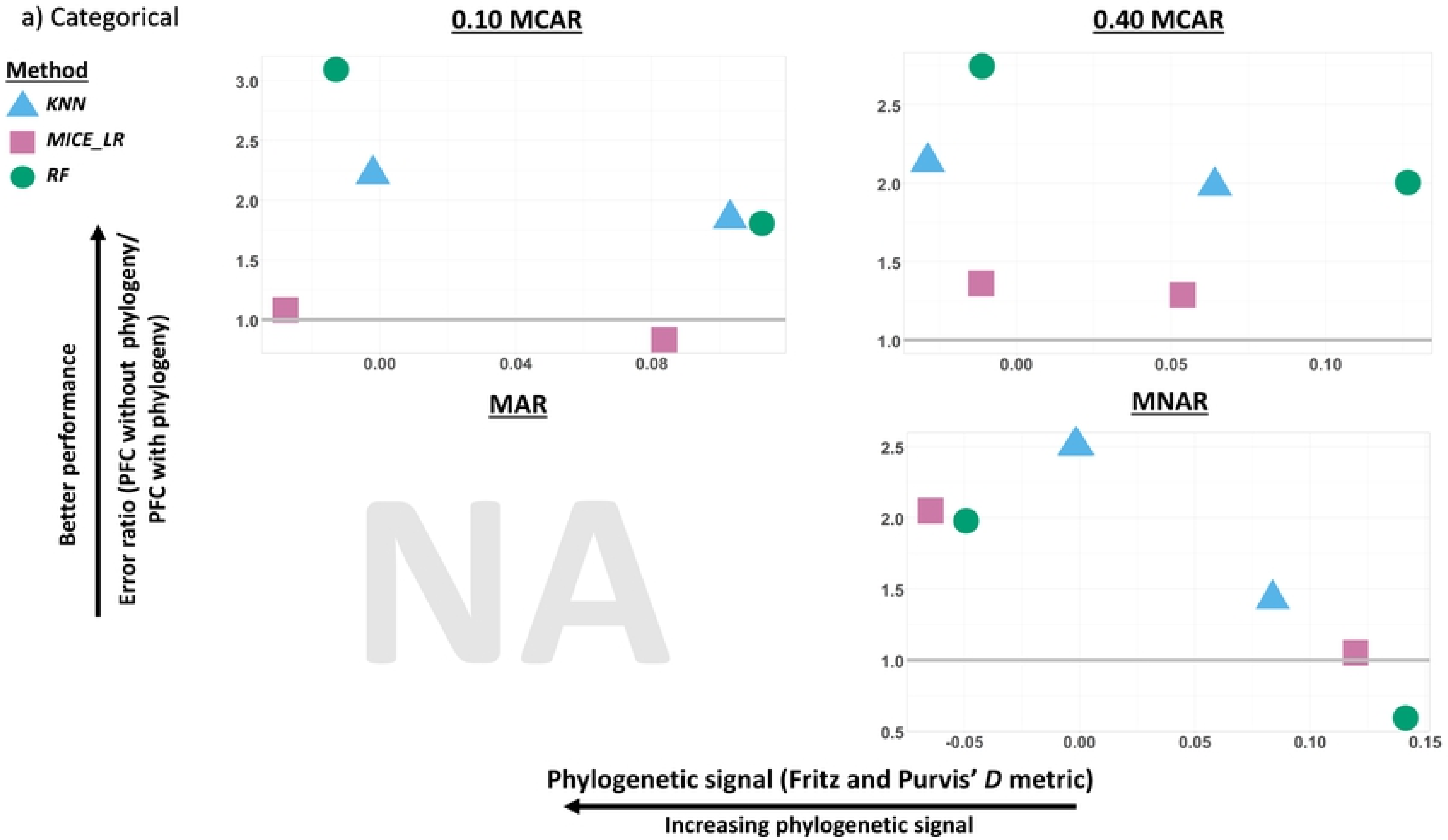
Method
Mean
KNN
RF
MICE_PMM
Error Rate (MSE)
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Proportion of missingness

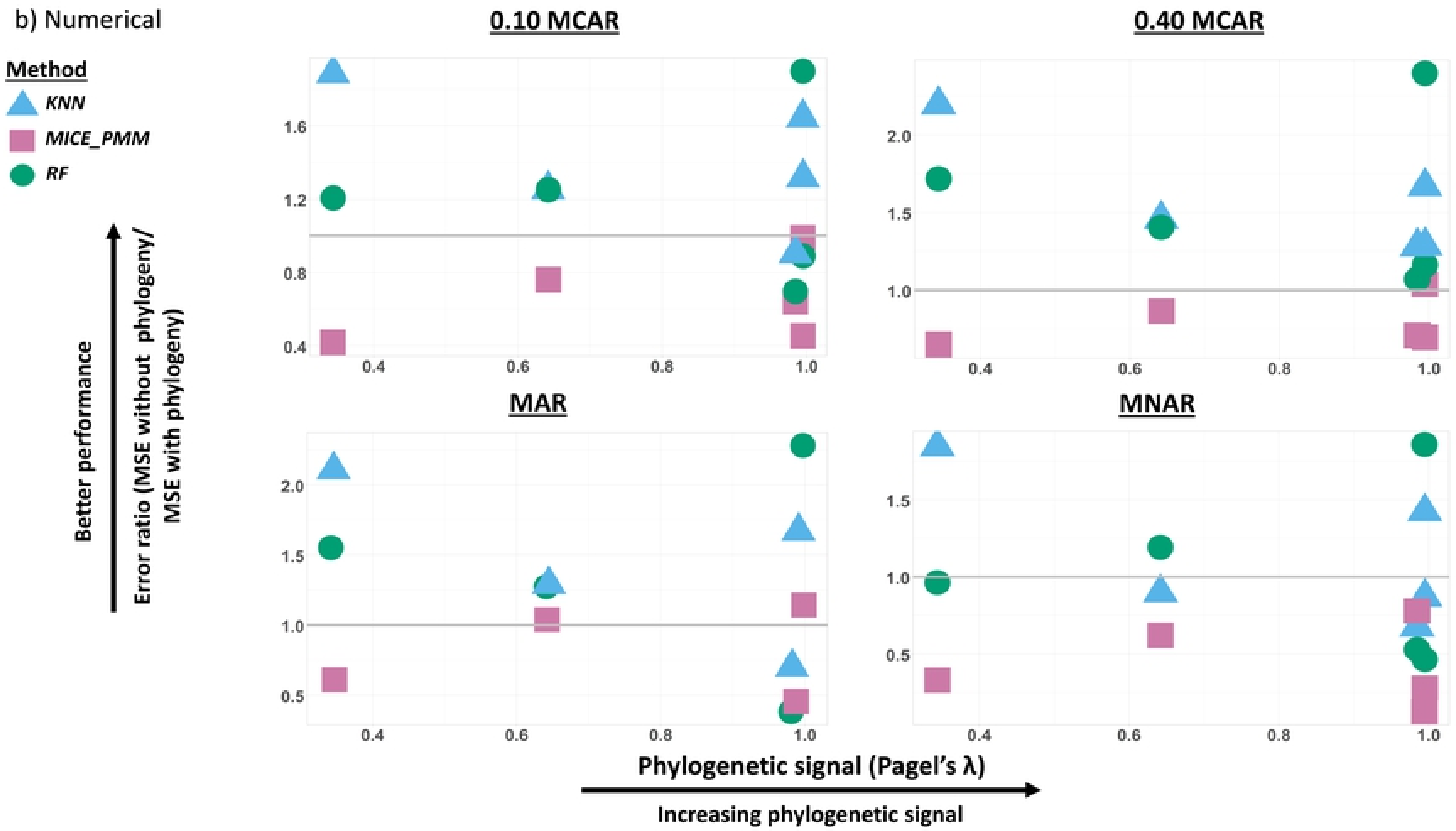

MCAR 10%
MCAR 40%
MAR
MNAR
a) Activity time
b) Insular endemic
Legend
Mode
KNN
RF
MICE
1.00
0.75
0.50
0.25
0.00
Mode
KNN
KNN_COI
KNN_CMOS
KNN_RAG1
RF
RF_COI
RF_CMOS
RF_RAG1
MICE
MICE_COI
MICE_CMOS
MICE_RAG1
NA
Error Rate (PFC)
Method


MCAR 10%
MCAR 40%
MAR
MNAR
c) Largest clutch
d) Smallest clutch
Legend
Mean
KNN
RF
MICE
Error Rate (MSE)
Method
0.0
0.2
0.4
0.6
0.8
0.1
0.3
Mean
KNN
KNN_COI
KNN_CMOS
KNN_RAG1
RF
RF_COI
RF_CMOS
RF_RAG1
MICE
MICE_COI
MICE_CMOS
MICE_RAG1


MCAR 10%
MCAR 40%
MAR
MNAR
e) Female SVL
Legend
Mean
KNN
RF
MICE
0.00
0.25
0.50
0.75
Mean
KNN
KNN_COI
KNN_CMOS
KNN_RAG1
RF
RF_COI
RF_CMOS
RF_RAG1
MICE
MICE_COI
MICE_CMOS
MICE_RAG1
f) Maximum SVL
0.00
0.25
0.50
0.75
1.00
NA
Error Rate (MSE)
Method


MCAR 10%
MCAR 40%
MAR
MNAR
g) Latitude
Legend
Mean
KNN
RF
MICE
Error Rate (MSE)
1.5
1.0
0.5
0.0
Mean
KNN
KNN_COI
KNN_CMOS
KNN_RAG1
RF
RF_COI
RF_CMOS
RF_RAG1
MICE
MICE_COI
MICE_CMOS
MICE_RAG1
Method


a) Categorical
Method
KNN
MICE_LR
RF
Better performance
Error ratio (PFC without phylogeny/ PFC with phylogeny)
0.10 MCAR
3.0
2.5
2.0
1.5
1.0
0.00
0.04
0.08
0.40 MCAR
2.5
2.0
1.5
1.0
0.00
0.05
0.10
MAR
NA
MNAR
2.5
2.0
1.5
1.0
0.5
-0.05
0.00
0.05
0.10
0.15
Phylogenetic signal (Fritz and Purvis' D metric)
Increasing phylogenetic signal


b) Numerical
Method
KNN
MICE_PMM
RF
0.10 MCAR
0.40 MCAR
MAR
MNAR
Better performance
Error ratio (MSE without phylogeny/ MSE with phylogeny)
1.6
1.2
0.8
0.4
2.0
1.5
1.0
0.5
0.4
0.6
0.8
1.0
Phylogenetic signal (Pagel's λ)
Increasing phylogenetic signal

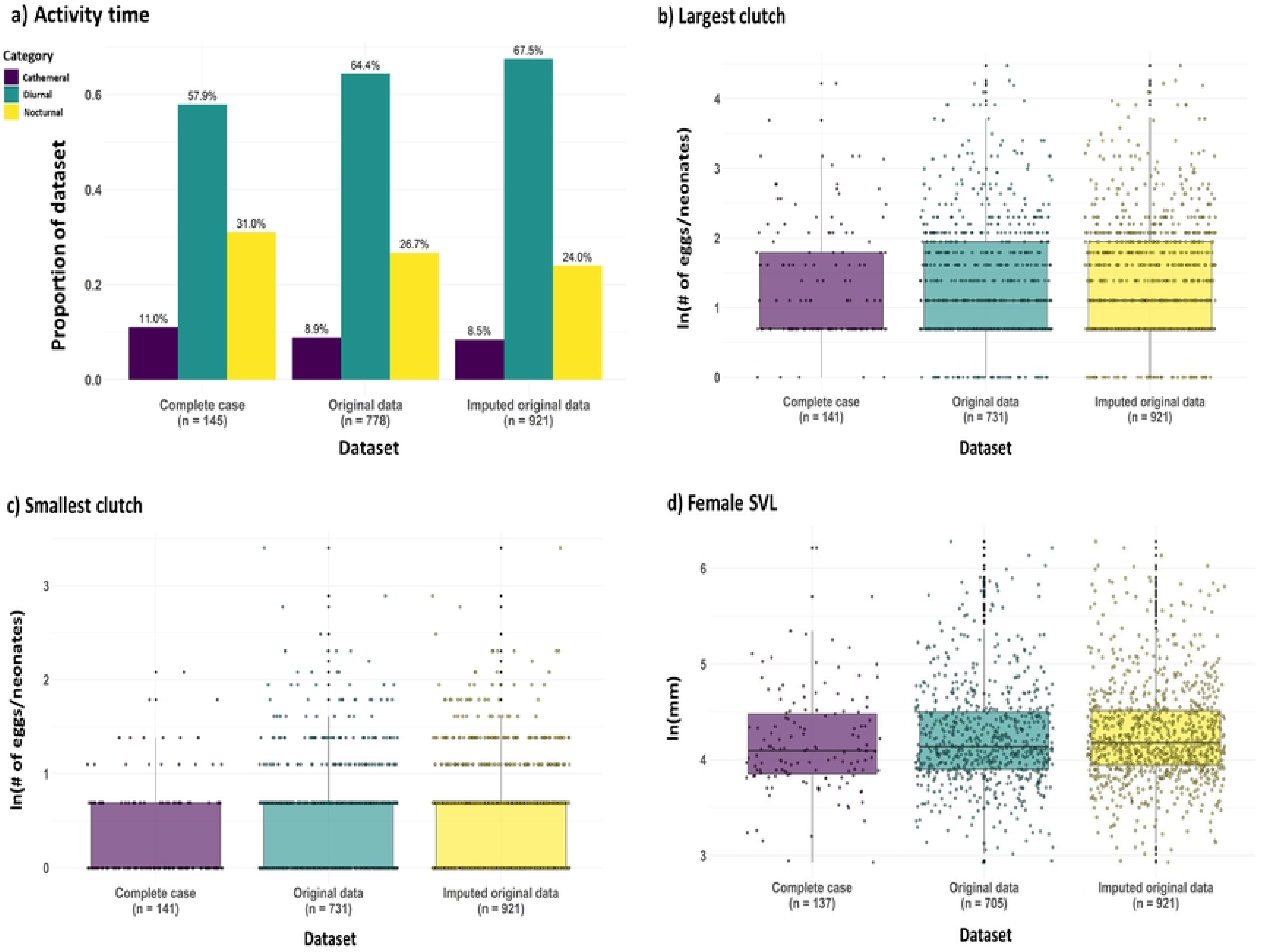
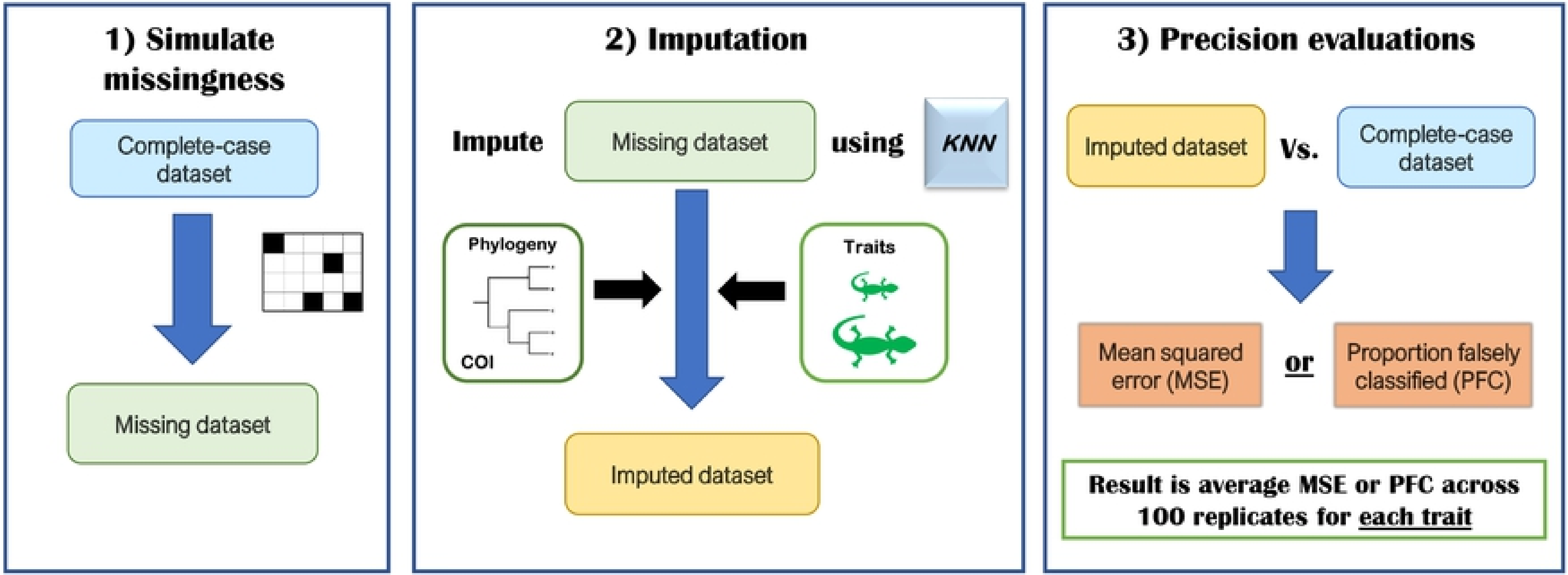

a) Activity time
Category
Cathemeral
Diurnal
Nocturnal
Proportion of dataset
0.0
0.2
0.4
0.6
11.0%
57.9%
31.0%
8.9%
64.4%
26.7%
8.5%
67.5%
24.0%
Complete case (n = 145)
Original data (n = 778)
Imputed original data (n = 921)
Dataset
b) Largest clutch
ln(# of eggs/neonates)
0
1
2
3
4
Complete case (n = 141)
Original data (n = 731)
Imputed original data (n = 921)
Dataset
c) Smallest clutch
ln(# of eggs/neonates)
0
1
2
3
Complete case (n = 141)
Original data (n = 731)
Imputed original data (n = 921)
Dataset
d) Female SVL
ln(mm)
3
4
5
6
Complete case (n = 137)
Original data (n = 705)
Imputed original data (n = 921)
Dataset

Fig 5



1) Simulate missingness

Complete-case dataset

Missing dataset

2) Imputation

Impute Missing dataset using *KNN*

Imputed dataset

3) Precision evaluations

Imputed dataset Vs. Complete-case dataset

Mean squared error (MSE) or Proportion falsely classified (PFC)

Result is average MSE or PFC across 100 replicates for each trait

Fig 6