Missing data in ecology: Some synthesis, clarification, and recommendations

Michael Dumelle[a,*], Rob Trangucci[b], Amanda M. Nahlik[a], Anthony R. Olsen[a], Kathryn M. Irvine[c], Karen Blocksom[a], Jay M. Ver Hoef[d], Claudio Fuentes[b]

[a]*United States Environmental Protection Agency, Office of Research and Development, 200 SW 35th St, Corvallis, OR, USA.*
[b]*Department of Statistics, Oregon State University, Corvallis, OR, USA*
[c]*United States Geological Survey, 2327 University Way, Suite 2, Bozeman, MT, USA.*
[d]*Marine Mammal Laboratory, Alaska Fisheries Science Center, NOAA Fisheries, Seattle, WA, USA*

Abstract

In ecology and related sciences, missing data are common and easily mishandled. When mishandled, missing data obfuscate ecological understanding. We review and synthesize several approaches for handling missing data. Generally, missing data can be grouped into one of three categories: missing completely at random (MCAR); missing at random (MAR); or missing not at random (MNAR). We review each category and pay special attention to MAR, which is quite flexible and useful but often misunderstood. We compare the benefits and drawbacks of several modern missing data methods, including complete case analysis, imputation, and data augmentation. We clarify the important distinction between imputation and prediction and argue that using predictive metrics to evaluate imputation methods is bad statistical practice and should be avoided. We describe a novel framework called a contingency filter, which clarifies whether missing data have a basis for measurement, and highlight its utility several contexts. We also study the impact of missing data on spatially explicit statistical models. Throughout, we illustrate missing data concepts using wetland data from the United States Environmental Protection Agency's 2016 National Wetland Condition Assessment (NWCA). We end by providing ten explicit recommendations for ecologists to consider while handling missing data.

*Corresponding Author; An author biosketch is provided in Appendix S1
Email address: Dumelle.Michael@epa.gov; michael.dumelle@oregonstate.edu (Michael Dumelle)

Keywords:

Bayesian modeling; Complete case analysis; Contingency filter; Data augmentation; Imputation, Prediction; Spatially explicit modeling

Disclaimer:

This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and pre-decisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Open Research Statement

All data and `R` code used in the creation of this manuscript are available on GitHub at https://github.com/USEPA/missing.data.in.ecology.

Upon (hopeful) manuscript acceptance, we will archive our data in a permanent repository (e.g., Dryad, Zeonodo) per ESA guidelines.

1. Introduction

A data analysis aims to answer important scientific questions by studying relationships among variables, estimating population parameters, or making future predictions. Ecological data often have some amount missingness that makes these important scientific questions harder to answer. Throughout this work, we synthesize methods for analyzing missing data, clarify some common misunderstandings regarding missing data, and details some recommendations. Our intent is to provide ecologists and scientists in other disciplines (e.g., environmental science, geology, geochemistry, soil science, etc.) with tools to more effectively handle missing data.

Missing data methods have matured tremendously over the past few decades (Rubin, 1996). Specific missing data methods (e.g., imputation) have been popularized in disciplines such as medicine (Bennett, 2001; Wood et al., 2004; Little et al., 2012; Austin et al., 2021; Carpenter and Smuk, 2021), psychology (Roth, 1994; Graham, 2009; Schlomer et al., 2010; Schoemann et al., 2024), social science (Little and Rubin, 1989; Newman, 2014; Saunders et al., 2006), epidemiology (Donders et al., 2006; Pedersen et al., 2017; Perkins et al., 2018), and survey research (Brick and Kalton, 1996; Andridge and Little, 2010; Lumley, 2011; Laaksonen, 2018), among many others (Carpenter et al., 2023). In ecology specifically, missing data methods have been used to study biodiversity (Taugourdeau et al., 2014; Kim et al., 2018; Bowler et al., 2024), species trait indices (Johnson et al., 2021; Penone et al., 2014), species distribution models (Ter Braak et al., 1994; Łopucki et al., 2022), meta-analyses (Ellington et al., 2015), behavioral ecology (Nakagawa and Freckleton, 2011), plant ecology (Dray and Josse, 2015), water quality (Srebotnjak et al., 2012), forestry (Van Deusen, 1997), and animal movement (Scharf et al., 2017). However, modern missing data methods still lack widespread adoption in the ecological and related sciences (Nakagawa and Freckleton, 2008; Nakagawa, 2015).

We begin our review by comparing and contrasting three types of missing data: missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR). We then synthesize modern missing data methods like complete case analysis, imputation, and data augmentation. We illustrate how diagnostic tools can be used to evaluate missingness patterns,

clarify the crucial distinction between imputation and prediction, and introduce a novel framework called a contingency filter, which clarifies whether there is a basis for certain types of missing data to exist. Using both simulated data and real data from the United States Environmental Protection Agency's (USEPA) National Wetland Condition Assessment (NWCA, Kentula and Paulsen, 2019), we show how missing data methods can be applied to various aspects of ecological analyses like data exploration, parameter estimation, prediction, and more. We conclude by outlining ten explicit recommendations for ecologists to consider while handling missing data.

2. The NWCA Data

The NWCA measures US wetland condition on repeating, five-year cycles (2011, 2016, 2021, and planned for 2026). For each cycle, approximately 1000 sample locations (i.e., 0.5 ha wetland sites) are selected throughout the conterminous US using a spatially balanced probability design (Stevens Jr and Olsen, 2004; Olsen et al., 2019). Each site is sampled during a one-day field visit during an Index Period (typically, April to October) for dozens of variables, each characterizing a distinct aspect of wetland condition. Chemical, physical, and biological samples are collected and analyzed using consistent field-based and lab-based protocols (USEPA, 2015, 2016; Herlihy et al., 2019; McCauley et al., 2019). Throughout this paper, we motivate the missing data problem in ecology using several variables from the NWCA 2016: two soil variables (soil hardening and soil modification); three vegetation variables (vegetation multi-metric index, vegetation removal, and vegetation type); and two water chemistry variables (surface water presence and total nitrogen). For each variable, Table 1 provides a brief summary, while Magee et al. (2019) and USEPA (2023) provide technical details. In total 967 wetlands were sampled for these variables as part of the NWCA 2016.

To collect soil hardening, soil modification, and vegetation removal data, field crews used a set list of eight human-mediated physical alterations associated with these variables within the 0.5-ha wetland site Assessment Area (AA) and the 100-m buffer around the AA (USEPA, 2016). Physical alterations that indicated soil hardening included, e.g., soil compaction from cattle,

non-paved trails, vehicle ruts, paved roads, and impervious surfaces. Observations that indicated soil modification included, e.g., trash and dumping, landfill activity, excavation and dredging, soil tilling, sedimentation, and soil erosion. Physical alterations that indicated vegetation removal included, e.g., forest clear cutting and selective cutting, herbicide use, grazing, and mowing.

Each observed physical alteration was scored based on the proximity to the AA, ranging from 25 points (inside the AA) to 1 point (at the furthest edge of the buffer). The eight physical alteration scores associated with each of the three variables were summed to calculate the Soil Hardening Index, the Soil Modification Index, and the Vegetation Removal Index. The maximum score for each index was 424 points (USEPA, 2023). For the examples in this paper, the Soil Hardening Index is reported as the numeric score. The Soil Modification Index score was translated into a binary result (i.e., "Yes", at least one soil modification was observed or "No", a soil modification was not observed). The Vegetation Removal Index score was translated into three stressor level categories: Low, zero points of vegetation removal; Medium, between 0 and 50 points of vegetation removal; and High, greater than 50 points of vegetation removal.

Wetland type of the sampled sites was determined from US Fish and Wildlife Survey National Wetland Inventory spatial layers (Wilen and Bates, 1995) and confirmed in the field. Seven Cowardin wetland types (Cowardin, 1979) were aggregated into wetland classes describing the dominant vegetation type at the site: woody or herbaceous.

To collect vegetation data, five 100-m^2 plots within the sampled wetland sites were surveyed by a field botanist for the presence, cover, and height class of all vascular plant species, tree species, nonvascular groups, and for ground surface attributes (USEPA, 2016). The Index Period coincides with peak vegetation to maximize observations and identification of plant species on site. Observed plant nomenclatures were standardized by name-site pairs to the US Department of Agriculture Natural Resources Conservation Service PLANTS database (USDA-NRCS, 2020). Metrics describing species traits (life history, wetland indicator status, native status, and coefficients of conservatism) were described for each taxon, taxon-state, or taxon-region (depending on the metric). Candidate metrics for use in Vegetation Multimetric Indices (VMMIs)

were screened for range, repeatability, responsiveness, and redundancy. The strongest performing metrics (e.g., relative cover of native monocots, percent richness of native species, relative cover of native species, among others) were used to develop and calculate four VMMIs: one for each combination of a) site type (i.e., tidal or inland) and b) dominant vegetation type (i.e., woody or herbaceous). Different metrics were used in each VMMI, with tidal VMMIs comprised of six metrics and inland VMMIs comprised of four metrics. Metric scores were combined for each VMMI, which was ultimately scored on a continuous scale from 0 to 100, with higher values reflecting healthier vegetiation (i.e., less disturbed conditions); see USEPA (2023) for further details.

The Index Period typically coincides with the warmest, driest part of the year in the US. Because of the diverse wetland types and hydrologic regimes across the conterminous US (Mitsch et al., 2023), not all wetlands have surface water presence during the field visit (i.e., a proportion of the sampled wetlands are dry). Field crews recorded whether surface water is present at the site as part of the observational data. Using these data, a binary indicator ("Yes", water was present or "No", water was not present) was created to reflect surface water presence. When surface water was present at a site, field crews collected a water sample for analysis of total nitrogen, among other analytes. Total nitrogen was analyzed using a persulfate digestion followed by an automated colorimetric analysis via Flow Injection Analysis (FIA, USEPA, 2015). When measured, concentrations of total nitrogen in the surface water are reported in mg/L.

3. Three Types of Missing Data

Causes of missing data are highly variable but can be generally classified into three broad groups: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Little and Rubin, 2019). Before making these definitions explicit, we describe some general notation that will be used henceforth. Following Rubin (1976) and Little and Rubin (2019), let $\mathbf{Y}$ be an $n \times p$ matrix whose rows index each of the n units (i.e., sites) and whose columns index each of the p variables potentially observable for each unit. Each unit-variable

combination is called an item, and hence, there are np items in $\mathbf{Y}$ (p items for each of n units). The matrix $\mathbf{Y}$ represents all outcomes that could possibly be observed for each item. Next, let $\mathbf{R}$ be an $n \times p$ matrix that identifies whether the items in $\mathbf{Y}$ are observed or missing. That is, an item in $\mathbf{R}$ has a value of one if the corresponding item in $\mathbf{Y}$ is observed (i.e., not missing) and a value of zero if the corresponding item in $\mathbf{Y}$ is missing (i.e., not observed). Just as $\mathbf{Y}$ represents all possible outcomes a researcher might observe, $\mathbf{R}$ represents the collection of all possible missingness patterns. Based on $\mathbf{R}$, we may partition $\mathbf{Y}$ into two distinct groups: $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$. The group $\mathbf{Y}_{obs}$ contains the items in $\mathbf{Y}$ that are observed, while the group $\mathbf{Y}_{mis}$ contains the items in $\mathbf{Y}$ that are missing (i.e., not observed). Providing these definitions is important because notation varies throughout the literature – e.g., Rubin (2004) places variables whose items are fully observed into a separate matrix called $\mathbf{X}$. We simply assume $\mathbf{Y}$ contains all possible data.

The matrices $\mathbf{Y}$, $\mathbf{R}$, $\mathbf{Y}_{obs}$, and $\mathbf{Y}_{mis}$ are very general and represent all possible observed items and missingness patterns. In practice, however, we only see one data set with a single set of observed items and a single missingness pattern. We indicate the values $\mathbf{Y}$, $\mathbf{R}$, $\mathbf{Y}_{obs}$, and $\mathbf{Y}_{mis}$ which are realized (i.e., seen) for a single data set as $\mathbf{y}$, $\mathbf{r}$, $\mathbf{y}_{obs}$, and $\mathbf{y}_{mis}$, respectively (Table 2). Shortly we provide an example of these quantities using wetland data, and in the Supporting Information, we provide an additional example along with `R` code to further elucidate these relationships. While we directly observe the values in $\mathbf{y}_{obs}$, the values in $\mathbf{y}_{mis}$ are hidden from us. However, they generally still exist in practice and hence, could have been directly observed under other missingness patterns (i.e., other random outcomes in $\mathbf{R}$). We explore specific scenarios where $\mathbf{y}_{mis}$ is not directly observable in Section 6.

Consider the following matrix $\mathbf{y}$, which, for six units, contains the observable values of three NWCA 2016 variables (USEPA, 2023): wetland type (WT), soil hardening (SH) scores, and

vegetation multi-metric index (VMMI) scores (Table 1).

$$\mathbf{y} = \left[\begin{array}{c|c|cc} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Woody} & 97.5 & 44.0 \\ 2 & \text{Herbaceous} & 29.5 & 31.2 \\ 3 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \\ 6 & \text{Woody} & 65.9 & 38.7 \end{array}\right].$$

Suppose that WT and SH are observed for each unit but VMMI is missing in the second and sixth unit. This missingness pattern is represented by $\mathbf{r}$, which is given by

$$\mathbf{r} = \left[\begin{array}{c|ccc} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 0 \\ 3 & 1 & 1 & 1 \\ 4 & 1 & 1 & 1 \\ 5 & 1 & 1 & 1 \\ 6 & 1 & 1 & 0 \end{array}\right].$$

Combining $\mathbf{y}$ and $\mathbf{r}$ yields the matrices

$$\mathbf{y}_{obs} = \left[\begin{array}{c|c|cc} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Woody} & 97.5 & 44.0 \\ 2 & \text{Herbaceous} & 29.5 & * \\ 3 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \\ 6 & \text{Woody} & 65.9 & * \end{array}\right] \text{ and } \mathbf{y}_{mis} = \left[\begin{array}{c|ccc} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & * & * & * \\ 2 & * & * & 31.2 \\ 3 & * & * & * \\ 4 & * & * & * \\ 5 & * & * & * \\ 6 & * & * & 38.7 \end{array}\right]. \quad (1)$$

This example highlights that $\mathbf{y}_{mis}$ represents a single outcome, namely the outcome that the second element of VMMI is 31.2 and the sixth element is 38.7. These outcomes are not observed by the analyst, however. Instead, to the analyst, the second and sixth elements of VMMI could be anywhere in the interval $[0, 100]$. We will return to these $\mathbf{y}$, $\mathbf{r}$, $\mathbf{y}_{obs}$, and $\mathbf{y}_{mis}$ values shortly, after describing the MCAR, MAR, and MNAR paradigms.

Data are missing completely at random (MCAR) if the probability that $\mathbf{R} = \mathbf{r}$ does not depend on any $\mathbf{y}$. Put another way, the probability of missingness *for the realized data set* is unrelated to the data values themselves. More formally, MCAR data satisfy the following condition (Little, 2021):

$$Pr(\mathbf{R} = \mathbf{r}|\mathbf{Y} = \tilde{\mathbf{y}}) = Pr(\mathbf{R} = \mathbf{r}|\mathbf{Y} = \mathbf{y}^{\star})$$

for any two realizations $\tilde{\mathbf{y}}$ and $\mathbf{y}^{\star}$. The term $Pr(\mathbf{R} = \mathbf{r}|\mathbf{Y} = \tilde{\mathbf{y}})$ is read as "the probability that $\mathbf{R} = \mathbf{r}$ given $\mathbf{Y}$ is set to a particular value, namely $\tilde{\mathbf{y}}$." When data are MCAR, the underlying distributions of the observed and missing data are the same (Example 1.13, Little and Rubin, 2019). A more stringent condition than MCAR is that all possible outcomes of $\mathbf{R}$ (the missingness patterns) are independent of all possible outcomes of $\mathbf{Y}$ (the data sets). When this condition is satisfied, the data are missing-always-completely-at-random (MACAR); this nuance is subtle (see Heitjan and Basu, 1996; Seaman et al., 2013; Mealli and Rubin, 2015, for more details).

A more general data assumption than MCAR is missing at random (MAR). Data are MAR if the probability that $\mathbf{R} = \mathbf{r}$ depends on $\mathbf{y}$ only through $\mathbf{y}_{obs}$, the observed data. Put another way, the probability of the observed missingness pattern is related to the observed data values but unrelated to the missing data values. More formally, MAR data satisfy the following condition:

$$Pr(\mathbf{R} = \mathbf{r} \mid \mathbf{Y}_{obs} = \mathbf{y}_{obs}, \mathbf{Y}_{mis} = \tilde{\mathbf{y}}_{mis}) = Pr(\mathbf{R} = \mathbf{r} \mid \mathbf{Y}_{obs} = \mathbf{y}_{obs}, \mathbf{Y}_{mis} = \mathbf{y}^{\star}_{mis})$$

for any two realizations $\tilde{\mathbf{y}}_{mis}$ and $\mathbf{y}^{\star}_{mis}$. When data are MAR, the underlying distributions of the observed and missing data are the same given the values of the observed data (Example 1.13, Little and Rubin, 2019). The MAR assumption is powerful but often misunderstood (perhaps in part because of the term "random") . As with MCAR and MACAR, there is a stronger assumption whereby we enforce independence between $\mathbf{Y}$ and $\mathbf{R}$ given $\mathbf{Y}_{obs}$, which is called missing-always-at-random, or MAAR (Seaman et al., 2013).

The most general missing data assumption is missing not at random (MNAR). Data are MNAR if $\mathbf{R}$ depends on $\mathbf{Y}$ through $\mathbf{Y}_{mis}$, the missing data. Put another way, the probability of missingness is related to the missing data values themselves. More formally, MNAR data satisfy the following condition:

$$Pr(\mathbf{R} = \mathbf{r} \mid \mathbf{Y}_{obs} = \mathbf{y}_{obs}, \mathbf{Y}_{mis} = \tilde{\mathbf{y}}_{mis}) \neq Pr(\mathbf{R} = \mathbf{r} \mid \mathbf{Y}_{obs} = \mathbf{y}_{obs}, \mathbf{Y}_{mis} = \mathbf{y}^{\star}_{mis})$$

for some values of $\tilde{\mathbf{y}}_{mis}$ and $\mathbf{y}^{\star}_{mis}$ such that $\tilde{\mathbf{y}}_{mis} \neq \mathbf{y}^{\star}_{mis}$. When data are MNAR, the underlying distributions of the observed and missing data are not the same.

These missing data definitions are at first technical and esoteric, so making them more concrete is helpful to aid in their understanding. Distributions of hypothetical MCAR, MAR, and MNAR VMMI scores are shown in Figure 1. When VMMI scores are MCAR, the distributions of the observed and missing VMMI scores are the same (Figure 1). When VMMI scores are MAR given WT, the distributions of the observed and missing VMMI scores are the same within all woody wetlands, the same within all herbaceous wetlands, but different between woody and

herbaceous wetlands (Figure 1). This pattern emerges because the probability of missingness is higher for woody wetlands than for herbaceous wetlands. Finally, when VMMI scores are MNAR, the distributions of the observed and missing VMMI scores are different for all wetlands, even after accounting for WT (Figure 1).

Unfortunately the MCAR, MAR, and MNAR assumptions are challenging, or even impossible, to evaluate directly. Little (1988b) proposed a test to evaluate the plausibility of the MCAR assumption. Enders (2022) review some approaches to compare MCAR and MAR assumptions, but Van Buuren (2018) note that many of these these approaches are not widely used and their practical value is lacking. The MNAR assumption is not generally testable because such a test would require the missing data values themselves, which we do not observe.

The terms "ignorable" and "nonignorable" are often used in the missing data literature. When a missingness process is ignorable, the distribution $Pr(\mathbf{R} = \mathbf{r} \mid \mathbf{Y} = \mathbf{y})$ need not be included in our statistical model for $Pr(\mathbf{Y} = \mathbf{y})$. MCAR and MAR data can be "ignorable" because the mechanism yielding $\mathbf{R}$ does not depend on $\mathbf{Y}_{mis}$, as long as the parameters that govern the missingness distribution are distinct from the parameters that govern the data distribution (Little and Rubin, 2019). The term ignorable does not imply, however, that we can ignore the process of thoughtfully handling the missing data (Van Buuren, 2018). MNAR data are synonymously called nonignorable because the underlying mechanism yielding $\mathbf{R}$ must be accounted for while estimating population characteristics about $\mathbf{Y}$ (i.e., these must be modeled together). Sometimes it is possible to relax the MNAR assumption after adding variables to $\mathbf{Y}_{obs}$, making a MAR assumption reasonable (Collins et al., 2001). Molenberghs et al. (2014) and Van Buuren (2018) provide thorough review of MNAR data and associated techniques for handling them; specific techniques include selection models (Heckman, 1976), pattern-mixture models (Little, 2008; Glynn et al., 2013), and parameter selection models (Creemers et al., 2011).

Approaches for handling missing data that are MNAR (i.e., nonignorable) are often problem-specific and hence, lack generality. MNAR approaches are problem-specific because they explicitly model the missing data mechanism, which tends to require specific information

about the process being studied. When the missing data mechanism is not fully understood, it is possible to choose a poor model for the mechanism, and then, subsequent analyses may communicate misleading results. Given these MNAR data constraints, we focus henceforth on general methods for MCAR and MAR data, which cover an expansive range of possible assumptions and data structures. We highlight the flexibility and utility of the MAR assumption, arguing that it is a practical default assumption for ecological applications.

4. An Overview of Missing Data Methods for MCAR and MAR data

Throughout this section, we describe several methods for handling MCAR and MAR missing data. We also conduct a simulation study that empirically compares the performance of 13 different missing data methods; performance was evaluated using various metrics measuring inferential and predictive capacity.

4.1. Preventing and Recording Missing Data

While not a missing data method *per se*, it is helpful to consider ways to prevent and record missing data when designing a study (Little et al., 2012). Suggestions include simplifying data collection protocols (De Leeuw, 2001; Lin et al., 2012; O'neill and Temple, 2012) and conducting pilot studies to identify sources of missing data (Kang, 2013), among others. Missing data should also be recorded in a clear consistent manner that cannot be confused with a valid value (Pearson, 2006; Fraser et al., 2009). For example, it may be confusing to assign a value of zero to a missing item when this scenario cannot be distinguished from an observed item with a value of zero. In addition to recording whether an item is missing, it can be beneficial to also record the reason for missingness (if available), as this may inform subsequent handling of the missing data. So-called planned missing data designs explicitly build missing data into studies, often to balance variables that are easy (or cheap) to observe with others that are difficult (or expensive) to observe (for some commentary, see Gelman et al., 1998; Graham et al., 2006; Graham, 2012; Hossie et al., 2021; Noble and Nakagawa, 2021).

4.2. Complete Case Analysis

The first missing data method we detail is complete case analysis (CCA; i.e., listwise deletion). CCA is an intuitive approach that discards units with at least one missing item prior to analysis. Units two and six from $\mathbf{y}_{obs}$ in Equation (1) have missing VMMI items, so the relevant CCA subset is:

$$\mathbf{y}_{CCA} = \left[\begin{array}{c|cccc} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Herbaceous} & 90.4 & 39.4 \\ 3 & \text{Woody} & 97.5 & 44.0 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \end{array}\right].$$

CCA is simple and straightforward but completely omits partially observed (and potentially useful) data. When data are MCAR, CCA yields unbiased (i.e., correct on average) parameter estimates and standard errors appropriate for the reduced sample size (Van Buuren, 2018). When data are not MCAR, CCA can yield biased parameter estimates and standard errors, especially when variables related to missingness are omitted from the analysis (Schafer and Graham, 2002; Little and Rubin, 2019). CCA is attractive especially when the proportion of missing data is small, but Little and Rubin (2019) and Vach (2012) discourage the use of heuristic rules that determine whether CCA is appropriate, arguing that decisions be made uniquely for each data set based on relevant context.

4.3. Imputation

Imputation methods replace (i.e., fill in, impute) missing items with plausible values called imputations that are informed by the observed data, yielding a complete data set consisting of observed and imputed items. The imputed items themselves can be deterministic (i.e., nonrandom) or random (i.e., stochastic). Deterministic imputation methods always return the same set of imputations, while random imputation methods build in some component of randomness (D'Agostino McGowan et al., 2024). The process of completing the data set with random imputed items is called "single imputation." Single imputation methods sometimes

incorporate additional randomness in the parameters of underlying models used to generate the imputations themselves; we discuss these nuances in more detail later. After imputation (deterministic or single), the complete data set can analyzed (e.g., using a linear regression model) as if there were no missing items.

Shortly we review several imputation methods, detailing the benefits and drawbacks of each and illustrating them while revisiting $\mathbf{y}_{obs}$ in Equation (1). Throughout, we first assume items are missing only for a single variable, and then we generalize to settings where several variables have missing items. We discuss "multiple imputation," whereby complete data sets from separate single imputation draws are pooled together, sharing information (Rubin, 1996). Multiple imputation is visually compared and contrasted against complete case analysis, deterministic imputation, single imputation in Figure 2.

4.3.1. Mean Imputation

The simplest imputation method is a deterministic method called mean imputation. In mean imputation, a variable's missing items are imputed by the mean of its observed items (Wilks, 1932). Imputing VMMI items two and six from $\mathbf{y}_{obs}$ yields

$\mathbf{y}_{Mean} =$

Unit	WT	SH	VMMI
1	Woody	97.5	44.0
2	Herbaceous	29.5	**37.4**
3	Herbaceous	90.4	39.4
4	Woody	79.6	25.0
5	Herbaceous	33.6	41.2
6	Woody	65.9	**37.4**

,

where the VMMI imputations (in bold) equal 37.4, the mean of the four observed VMMI items: 44.0, 39.4, 25.0 and 41.2. A benefit of mean imputation is that the mean of the observed items equals the mean of the imputed items. Mean imputation is also quite computationally efficient.

Unfortunately, mean imputation tends to artificially inflate the strength of the relationships in the data, leading to biased parameter estimates and notably underestimated variances (Gleason and Staelin, 1975; Kromrey and Hines, 1994; Olinsky et al., 2003). Enders (2022) claim that "in no situation is mean imputation defensible, and you should absolutely avoid this approach."

4.3.2. Regression Imputation

The next deterministic imputation method is regression imputation (i.e., conditional mean imputation). In regression imputation, a linear regression model is first fit using the completely observed units. Then, the fitted model is used to generate an imputation for each missing item (Buck, 1960). Suppose we generate a linear regression model that captures the effect of wetland type (WT) and soil hardening (SH) on VMMI. This model is written as:

$$\text{VMMI}_i = \beta_0 + \beta_1 \mathcal{I}(\text{WT}_i) + \beta_2 \text{SH}_i + \epsilon_i, \tag{2}$$

where i indexes each of the four completely observed units ($i = 1, 3, 4, 5$), β_0 represents the overall intercept, β_1 represents the shift in average VMMI for woody wetlands,

$$\mathcal{I}(\text{WT}_i) = \begin{cases} 1 & \text{if WT}_i = \text{Woody} \\ 0 & \text{if WT}_i = \text{Herbaceous} \end{cases},$$

β_2 represents the change in average VMMI for a unit change in SH score, and each ϵ_i are independent and identically distributed (i.e., i.i.d.) random errors with a mean of zero and a (positive) variance of σ^2. Fitting the model in Equation (2) using ordinary least squares yields the following parameter estimates: $\hat{\beta}_0$ = 36.142; $\hat{\beta}_1$ = -7.581; and $\hat{\beta}_2$ = 0.067. These fitted values from the model are combined with the WT and SH measurements for units two and six to generate the

VMMI imputations:

$$\text{VMMI-IMP}_2 = 36.142 - 7.581(0) + 0.067(29.5) = 38.12$$

$$\text{VMMI-IMP}_6 = 36.142 - 7.581(1) + 0.067(65.9) = 32.98$$

Then the complete data set is

$$\mathbf{y}_{Reg} = \left[\begin{array}{c|cc|c} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Woody} & 97.5 & 44.0 \\ 2 & \text{Herbaceous} & 29.5 & \mathbf{38.1} \\ 3 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \\ 6 & \text{Woody} & 65.9 & \mathbf{33.0} \end{array}\right],$$

While intuitive, regression imputation, like mean imputation, tends to artificially inflate the strength of the relationships in the data, biasing parameter estimates and notably underestimating variances (Gleason and Staelin, 1975; Kromrey and Hines, 1994; Olinsky et al., 2003; Little and Rubin, 2019). Because its intuitive appeal is met with lackluster performance, Van Buuren (2018) claims that "Regression imputation, as well as its modern incarnations in machine learning, is probably the most dangerous of all [missing data] methods described here [in this book]."

4.3.3. Stochastic Regression Imputation

Stochastic regression imputation is a single imputation method that builds upon regression imputation by adding a random (i.e., stochastic) error to each regression imputation. The random error is simulated from an appropriate error distribution (often, a Gaussian distribution) with mean zero and variance σ^2. Suppose that we simulate two i.i.d. random errors from the Gaussian distribution with mean zero and variance 174.14 (the value 174.14 is the estimated residual variance from the fitted regression model). Further suppose that the random error realized for unit

two is -11.0 while the random error realized for unit six is 21.1. The stochastic regression imputations are then:

$$\text{VMMI-IMP}_2 = 36.142 - 7.581(0) + 0.067(29.5) - 11.83 = 26.29 \text{ and}$$
$$\text{VMMI-IMP}_6 = 36.142 - 7.581(1) + 0.067(65.9) + 21.14 = 54.12.$$

The complete data set is

$$\mathbf{y}_{StReg} = \left[\begin{array}{c|c|cc} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Woody} & 97.5 & 44.0 \\ 2 & \text{Herbaceous} & 29.5 & \mathbf{26.3} \\ 3 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \\ 6 & \text{Woody} & 65.9 & \mathbf{54.1} \end{array}\right].$$

While not intuitive at first, stochastic regression imputation tends to perform much better than regression imputation because stochastic regression imputation incorporates uncertainty in the imputations themselves (Little and Rubin, 2019). Unfortunately stochastic regression imputation is not perfect and ignores uncertainty in estimated model parameters, often yielding standard errors that are too small and confidence intervals that are too narrow.

4.3.4. Bootstrap Stochastic Regression Imputation

Bootstrap stochastic regression imputation (henceforth called bootstrap imputation) is a single imputation method that builds upon stochastic regression imputation by incorporating an additional source of randomness that is associated with the parameter estimates (Efron, 1994). Borrowing bootstrap resampling ideas (Efron and Tibshirani, 1994), bootstrap imputation applies stochastic regression imputation to a random, resampled set of observed data (i.e., a bootstrap sample) instead of the observed data. Suppose the bootstrap sample contained two copies of item

one, zero copies of item three, one copy of item four, and one copy of item five:

$$\mathbf{y}_{BSample} = \left[\begin{array}{c|lll} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Herbaceous} & 90.4 & 39.4 \\ 1 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \end{array}\right].$$

The ordinary least squares parameter estimates for this model are $\hat{\beta}_0$ = 42.265, $\hat{\beta}_1$ = -14.742, and $\hat{\beta}_2$ = -0.032 and the random errors realized are 5.14 and -2.76, respectively. Together, they yield the bootstrap imputations:

$$\text{VMMI-IMP}_2 = 42.265 - 14.742(0) - 0.032(29.5) + 5.14 = 46.47 \text{ and}$$

$$\text{VMMI-IMP}_6 = 42.265 - 14.742(1) - 0.032(65.9) - 2.76 = 22.67.$$

Then the complete data set is

$$\mathbf{y}_{Boot} = \left[\begin{array}{c|lll} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Woody} & 97.5 & 44.0 \\ 2 & \text{Herbaceous} & 29.5 & \mathbf{46.5} \\ 3 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \\ 6 & \text{Woody} & 65.9 & \mathbf{22.8} \end{array}\right],$$

Because bootstrap imputation also incorporates parameter uncertainty, it can yields estimates with appropriate standard errors and confidence interval coverage (Schomaker and Heumann, 2018; Little and Rubin, 2019). Other imputation methods that borrow resampling ideas are Jackknife imputation (Miller, 1974) and Bayesian imputation (Rubin, 2004; Box and Tiao, 2011).

4.3.5. Hot Deck Imputation

Several of the aforementioned imputation methods use a statistical model to generate imputations. A fundamentally different approach is to use hot deck imputation (Kalton and Kasprzyk, 1986; Scheuren, 2005; Myers, 2011). Hot deck imputation imputes missing items by matching (i.e., copying) the value of an observed item from an appropriate donor unit. Donors are typically selected from the observed data, though it is possible to select from alternative data sources. Selecting donors from alternative data sources is called cold deck imputation. Cold deck imputation is far less common than hot deck imputation, in large part because finding suitable alternative data sources can be challenging. Because hot deck imputations match observed items, they are constrained within the range of the observed data. They avoid strong parametric assumptions and tend to be robust to data transformations like logarithms or exponentiation. However, they do not always perform well with small sample sizes or highly skewed data (Kleinke, 2017).

Hot deck imputation involves two steps. The first step is to define a closeness measure between the item requiring imputation and all other items (Andridge and Little, 2010). Suppose that for $\mathbf{y}$, we define a closeness measure that 1) treat units with the same wetland type as closer than units with different wetland types and 2) after accounting for wetland type, treats units with more similar soil hardening scores (measured by distance) as closer than units with more different soil hardening scores. In this example, unit two is closest to unit five because it has the same wetland type and a more similar SH score than unit three; and unit six is closest to unit four because it has the same wetland type and a more similar SH score than unit one.

The second hot deck imputation step is to choose a close donor to generate the imputation. A hot deck deterministic imputation method called nearest neighbor always selects the closest donor to generate the imputation (Rancourt et al., 1994; Chen and Shao, 2000). Using the

aforementioned closeness measure for $\mathbf{y}_{obs}$ and nearest neighbor, the complete data set is

$$\mathbf{y}_{NN} = \left[\begin{array}{c|ccc} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Woody} & 97.5 & 44.0 \\ 2 & \text{Herbaceous} & 29.5 & \mathbf{41.2} \\ 3 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \\ 6 & \text{Woody} & 65.9 & \mathbf{25.0} \end{array}\right],$$

The imputed VMMI value for unit two matches the VMMI value for unit five (41.2), and the imputed VMMI value for unit six matches imputed value for unit four (25.0). While intuitive and straightforward, nearest neighbor fails to incorporate randomness in the closeness measure or donor selection and hence, performance suffers.

A widely used hot deck single imputation method is predictive mean matching (Rubin, 1986; Little, 1988a). Predictive mean matching finds the d donors closest donors and randomly selects one to generate the imputation. Using the aforementioned closeness measure for $\mathbf{y}_{obs}$ and predictive mean matching with $d = 2$, a potential complete data set is

$$\mathbf{y}_{PMM} = \left[\begin{array}{c|ccc} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Woody} & 97.5 & 44.0 \\ 2 & \text{Herbaceous} & 29.5 & \mathbf{39.4} \\ 3 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & 33.6 & 41.2 \\ 6 & \text{Woody} & 65.9 & \mathbf{25.0} \end{array}\right],$$

where the VMMI value of unit three was chosen randomly over the VMMI value of unit five (for the VMMI imputation of unit two) and the VMMI value of unit four was chosen randomly over

the VMMI value of unit one (for the VMMI imputation of unit six).

Predictive mean matching closeness can be determined using various rules. Two such rules are Type-0 matching, which uses a deterministic approach to determine closeness, and Type-1 matching, which uses an approach to determine closeness that has randomness (White et al., 2011; Van Buuren, 2018). Generally, rules that incorporate randomness while determining closeness are preferred (Van Buuren, 2018). Van Buuren (2018) recommends choosing d to be 3, 5, or 10, while Schenker and Taylor (1996) use an adaptive method to select d. Siddique and Belin (2008) discuss weighting among the d donors so that closer donors are more likely to be selected. Predictive mean matching has been shown to perform well for a variety of data types and missing data scenarios (Marshall et al., 2010; Morris et al., 2014; Vink et al., 2014; Kleinke, 2017).

4.3.6. Multiple Imputation

Singly imputed data (i.e., a complete data set from a single imputation draw) represent one of many possible imputed data sets and hence, fail to capture all the variability in the imputations themselves. Using singly imputed data in a subsequent analysis tends to yield parameter estimates whose corresponding standard errors are too small and confidence errors too narrow, no matter the imputation procedure used (Little and Rubin, 2019). To address this drawback, Rubin (1977) proposed a groundbreaking approach called multiple imputation. In multiple imputation, several singly imputed data sets are generated, an analysis is performed on each, and the results are appropriately pooled (Rubin, 1996). While the general concept behind multiple imputation is simplistically elegant, there are several important yet subtle questions lurking below the surface. Some examples: "How many imputed data sets should be used?"; "How should the results from different imputed data sets be pooled?"; and "Can the impact of missingness on an analysis be measured?". We explore answers to these types of questions next.

Rubin's Rules (Rubin, 2004) provide a framework for using multiple imputation to pool results from m singly imputed data sets. Let Q represent a parameter requiring estimation, which could be a simple mean, a slope parameter in a regression model, a population total, or some other quantity of interest. Suppose that an appropriate analysis of each of the m singly imputed data

sets yields an estimate of Q, denoted by $\hat{Q}_i$ ($i = 1, 2, \ldots, m$). Rubin's Rules state that the multiple imputation estimator of Q, denoted $\hat{Q}$, is given by

$$\hat{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i. \tag{3}$$

Equation (3) plainly states that $\hat{Q}$ is the average of each $\hat{Q}_i$. While the estimator $\hat{Q}$ is simple and intuitive, calculating its variance is more challenging. Rubin's Rules state that the variance of $\hat{Q}$, denoted $\hat{T}$, is given by

$$\begin{aligned} \hat{T} &= \frac{1}{m} \sum_{i=1}^{m} \hat{U}_i + \left(1 + \frac{1}{m}\right) \sum_{i=1}^{m} \frac{(\hat{Q}_i - \hat{Q})^2}{m-1} \\ &= \text{Var(Within)} + \text{Var(Between)} \end{aligned}$$

where $\hat{U}_i$ is an estimate of the variance of $\hat{Q}_i$. The term $\frac{1}{m} \sum_{i=1}^{m} \hat{U}_i$ represents the variability within each $\hat{Q}_i$ while the term $\left(1 + \frac{1}{m}\right) \sum_{i=1}^{m} \frac{(\hat{Q}_i - \hat{Q})^2}{m-1}$ represents the variability between each $\hat{Q}_i$. Henceforth we call these terms Var(Within) and Var(Between), respectively.

Often $\hat{Q}$ is a sum of random quantities (e.g., a mean, a regression slope, a total). In this context, it is reasonable, based on the Central Limit Theorem, to assume that $\hat{Q}$ is approximately normally distributed with mean Q and variance T. Under general conditions, this assumption implies that the test statistic

$$t_v^* = \frac{\hat{Q} - Q}{\sqrt{\hat{T}}}$$

is approximately t distributed with v degrees of freedom. Thus, p-values for appropriate hypothesis tests can be obtained by comparing t_v^* to the t-reference distribution. Moreover, a (1 - α)% confidence interval is given by

$$\left(\hat{Q} - t_{v,1-\alpha/2}\sqrt{\hat{T}}, \hat{Q} + t_{v,1-\alpha/2}\sqrt{\hat{T}}\right),$$

where $t_{v,1-\alpha/2}$ is the critical value from the t-reference distribution. The degrees of freedom for

these hypothesis tests and confidence intervals depends on the number of singly imputed data sets (m) and the sample size within each imputed data set. Various approaches to determining degrees of freedom are provided by Barnard and Rubin (1999); Lipsitz et al. (2002); Reiter (2007); Wagstaff and Harel (2011).

Rubin (2004) derives two statistics that elucidate the impact missing data has on an analysis. The first statistic is the proportion of variability in $\hat{Q}$ that results from the missing data, denoted by λ. An estimate of λ, denoted by $\hat{\lambda}$, is given by

$$\hat{\lambda} = \frac{\text{Var(Between)}}{\hat{T}},$$

the ratio of the between-imputation variability to the total variability. The statistic λ varies from zero to one; larger values indicate more variability in $\hat{Q}$ is attributable to the missing data. For example, a value of 0.5 implies that half of the variability in $\hat{Q}$ is attributable to the missing data.

The second statistic is the relative increase in variability in $\hat{Q}$ due to the missing data, denoted by r. An estimate of r, denoted by $\hat{r}$, is given by

$$\hat{r} = \frac{\text{Var(Between)}}{\text{Var(Within)}},$$

the ratio of the between-imputation variability to the within-imputation variability. For example, a value of one implies that the between-imputation is the same as the within-imputation variability. If half the variability is between-imputation variability, then roughly half the variability in $\hat{Q}$ is attributable to the missing data. This perspective reveals a connection between $\hat{\lambda}$ and $\hat{r}$, and it can be shown that $\hat{r} = \hat{\lambda}/(1 - \hat{\lambda})$. Rubin (2004) discusses other statistics that capture different aspects of imputation-related variability.

One reason the magnitude of between-imputation variability is important is because it can inform the choice of m, the number of singly imputed data sets. Choosing an appropriate m is a balance between performance and computational feasibility. As m increases, multiple imputation becomes more reliable, but the benefits rapidly decrease with large m (Rubin, 2004). Moreover,

storing and analyzing m singly imputed data sets can be computationally cumbersome. Van Buuren (2018) mention that classical advice is to set m to three, four, or five when there are low to moderate amounts of missing data. They also mention that substantive conclusions, especially when the primary interest is point estimates, are unlikely to change significantly as a result of raising m beyond five. Rubin (2004), Schafer (1997), and Schafer and Olsen (1998) argue that using m greater than 10 is rarely necessary. Graham et al. (2007) and Bodner (2008), however, recommend larger m; they advise setting $m = 50$ when $\lambda \leq 0.5$ and $m \geq 100$ when $\lambda \geq 0.7$. Building from Von Hippel (2009), White et al. (2011) provide a simple rule to set $m \geq 100\lambda$. For example, if approximately 34% of the data are missing, set $m \geq 34$. One way to balance all these perspectives is use $m = 100$ as a default, adjusting as necessary depending on context.

4.3.7. Missing Data in More Than One Variable

So far we have only discussed missing items in a single variable. But in many practical applications, items are missing in several variables. Sometimes, missing items have a specific pattern that facilitates their handling (Little and Rubin, 2019), but this is not true in general. Suppose $\mathbf{y}_{obs}$ in Equation (1) now also has missing items in the WT and SH variables and looks like

$$\mathbf{y}_{obs} = \left[\begin{array}{c|c|c|c} \text{Unit} & \text{WT} & \text{SH} & \text{VMMI} \\ \hline 1 & \text{Woody} & 97.5 & 44.0 \\ 2 & * & 29.5 & * \\ 3 & \text{Herbaceous} & 90.4 & 39.4 \\ 4 & \text{Woody} & 79.6 & 25.0 \\ 5 & \text{Herbaceous} & * & 41.2 \\ 6 & \text{Woody} & 65.9 & * \end{array}\right]. \tag{4}$$

Items are missing for all three variables: WT, SH, and VMMI. Fortunately, imputation methods naturally extend to accommodate missing items in several variables. For example, mean imputation applies a straightforward extension – simply take the mean of each variable. For

categorical variables, one approach is to impute using the most common category (i.e., level).

Regression and hot deck imputations are more challenging to extend to the multivariable case because they must account for relationships that exist among several variables. There are two main approaches to this extension: the joint approach and the conditional approach. In the joint approach, imputations for all missing items are generated simultaneously (Little and Rubin, 2019). Schafer (1997) uses a joint approach for data that are multivariate Gaussian, and Demirtas et al. (2008) show that even when this Gaussian assumption is violated, the method tends to work well. Allison (2005) describes a joint approach for data that are categorical.

The conditional approach leverages conditioning to capture relationships among variables instead of specifying a multivariate distribution. The conditional approach has many names (see Van Buuren, 2018, for a review) but is commonly called the fully conditional specification (FCS, Van Buuren et al., 2006) or chained equations (Azur et al., 2011; White et al., 2011). We adopt the chained equations nomenclature henceforth. In chained equations, univariate distributions are assumed for each variable that condition on the remaining variables. After determining suitable initial values for each missing item, missing items are updated with new imputations one variable at a time. Starting with $\mathbf{y}_{obs}$ from Equation (4), we can use chained equations to impute a complete data set. First, we use mean imputation to determine initial values:

Unit	WT	SH	VMMI
1	Woody	97.5	44.0
2	**Woody**	29.5	**37.4**
3	Herbaceous	90.4	39.4
4	Woody	79.6	25.0
5	Herbaceous	**72.6**	41.2
6	Woody	65.9	**37.4**

.

Second, we assume following relationship for VMMI given WT and SH,

$$
\begin{aligned}
[\text{VMMI}|\text{SH}, \text{WT}] &\sim \text{Gaussian}(\mu_1, \sigma_1^2) \\
\mu_1 &\equiv \beta_{0,1} + \beta_{1,1}\text{SH} + \beta_{2,1}\mathcal{I}(\text{WT}) \\
\mathcal{I}(\text{WT}) &= \begin{cases} 1 & \text{if WT} = \text{Woody} \\ 0 & \text{if WT} = \text{Herbaceous} \end{cases},
\end{aligned}
$$

and use stochastic regression imputation to update VMMI imputations:

Unit	WT	SH	VMMI
1	Woody	97.5	44.0
2	**Woody**	29.5	**42.1**
3	Herbaceous	90.4	39.4
4	Woody	79.6	25.0
5	Herbaceous	**72.6**	41.2
6	Woody	65.9	**24.7**

.

The distributional notation $[\text{VMMI}|\text{SH}, \text{WT}]$ simply means the distribution of VMMI given SH and WT. Third, we use nearest neighbor hot deck (determining closeness using WT and breaking ties with VMMI) to update SH imputations:

Unit	WT	SH	VMMI
1	Woody	97.5	44.0
2	**Woody**	29.5	**42.1**
3	Herbaceous	90.4	39.4
4	Woody	79.6	25.0
5	Herbaceous	**90.4**	41.2
6	Woody	65.9	**24.7**

.

Fourth, we assume following relationship for VMMI given WT and SH,

$$
\begin{aligned}
[\text{WT}|\text{VMMI}, \text{SH}] &\sim \text{Bernoulli}(\mu_2) \\
\mu_2 &\equiv \text{Probability that WT} = \text{Woody} \\
\log\left(\frac{\mu_2}{1-\mu_2}\right) &= \beta_{0,2} + \beta_{1,2}\text{VMMI} + \beta_{2,2}\text{SH},
\end{aligned}
$$

and use logistic regression (with a cutoff probability of 0.5) to update WT imputations:

Unit	WT	SH	VMMI
1	Woody	97.5	44.0
2	**Woody**	29.5	**42.1**
3	Herbaceous	90.4	39.4
4	Woody	79.6	25.0
5	Herbaceous	**90.4**	41.2
6	Woody	65.9	**24.7**

Here, the WT imputation stayed Woody. Fifth, we use stochastic regression imputation to update VMMI imputations again:

Unit	WT	SH	VMMI
1	Woody	97.5	44.0
2	**Woody**	29.5	**51.4**
3	Herbaceous	90.4	39.4
4	Woody	79.6	25.0
5	Herbaceous	**90.4**	41.2
6	Woody	65.9	**33.8**

This process continues iteratively until a suitable number of imputations have been updated, at

which point the complete data set may look like

Unit	WT	SH	VMMI
1	Woody	97.5	44.0
2	**Herbaceous**	29.5	**47.2**
3	Herbaceous	90.4	39.4
4	Woody	79.6	25.0
5	Herbaceous	**29.5**	41.2
6	Woody	65.9	**17.6**

.

For multiple imputation, this process is repeated m times and the analysis results are appropriately pooled.

4.3.8. Inclusive vs Restrictive Imputation Approaches

It is important to clarify that the imputation approach used to generate a complete data set is **is separate from** the analysis approach applied to the complete data set. For example, an analyst can use nearest neighbor imputation to impute a complete data set and then linear regression to analyze the complete data set. When all variables used in the analysis approach are also used in the imputation approach, the imputation approach is said to be "inclusive" for the analysis approach (Collins et al., 2001). For example, if the analysis approach uses the SH, WT, and VMMI variables, an imputation approach is inclusive if it also uses the SH, WT, and VMMI variables. If the analysis approach only uses the WT and VMMI variables, an imputation approach that uses the SH, WT, and VMMI variables is also inclusive, as it contains all the variables used in the analysis approach. When at least one variable used in the analysis approach is not also used in the imputation approach, the imputation approach is said to be "restrictive" for the analysis approach (Collins et al., 2001). For example, if the analysis approach uses the SH, WT, and VMMI variables, an imputation approach is restrictive if it omits at least one of the SH, WT, or VMMI variables (e.g., only uses the SH and WT variables in the imputation approach).

Generally, imputation approaches should be inclusive. This helps ensure plausibility of MAR and validity of the imputed values later used in the analysis approach. There is one exception to this guideline, however. When using deterministic imputation, the variable that is eventually the "response" variable in an analysis approach should be omitted from the imputation approach, making it restrictive (D'Agostino McGowan et al., 2024). For example, if the analysis approach is a linear regression of SH (response variable) on WT and VMMI (explanatory variables), SH should be omitted from the deterministic imputation approach. Several authors note that when the imputation approach is single or multiple imputation (i.e., has a random component), the eventual response variable in the analysis approach should be included in the imputation approach (Graham, 2009; Moons et al., 2006; Kenward and Carpenter, 2007; Tilling et al., 2016; Hughes et al., 2019). For example, if the analysis approach is a linear regression of SH (response variable) on WT and VMMI (explanatory variables), SH should be included in the single or multiple imputation approach.

4.3.9. Imputation Diagnostics

Diagnostic tools are useful to evaluate the adequacy of a statistical procedure. In imputation, there are several scenarios in which diagnostics can be useful. The first scenario is examining diagnostics in the underlying imputation approach. For example, regression imputation uses the observed data to create a regression model, which is later leveraged to generate imputations. This regression model can be checked using standard regression diagnostics like residual plots (Fox, 2019; Montgomery et al., 2021; Gelman et al., 2005). The second scenario is examining diagnostics of the analysis approach applied to the complete data set. Diagnostics in these scenarios are important but specific to the assumed models in the imputation and analysis approaches. A third scenario, which we focus on, is examining distributions of the imputations themselves as compared to the observed data.

Several authors have provided overviews of helpful diagnostic tools that can be used to assess the adequacy of the imputations (Abayomi et al., 2008; Su et al., 2011; Eddings and Marchenko, 2012; Nguyen et al., 2017). Van Buuren (2018) note that visualizations are particularly useful for

identifying discrepancies between observed and imputed data that can occur in the means, spread, scales, or relationships among variables. These visualizations can include histograms or densities of individual variables or scatterplots between variables that are separated by observed and imputed classification; these visualizations can also be conditioned on other variables (Bondarenko and Raghunathan, 2016). One such visualization is shown in Figure 3, which compares observed and imputed data for mean imputation, regression imputation, and bootstrap imputation of a hypothetical variable $\mathbf{z}_2$ using a hypothetical variable $\mathbf{z}_1$. The bootstrap imputations appear the most reasonable because they are indistinguishable from the observed data; namely, they recreate the general trend and spread in the observed data and lack outliers. Because the MAR assumption is difficult to verify empirically, these diagnostic visualizations tend to be especially helpful.

4.4. Data Augmentation

Unlike imputation, which separates the handling of missing data from the subsequent analysis, data augmentation approaches unite these processes (Tanner and Wong, 1987). Two commonly used data augmentation approaches are maximum likelihood and Bayesian modeling. Maximum likelihood approaches work by maximizing a likelihood function built from the observed and missing data. Bayesian approaches work by sampling from a posterior distribution of missing data and model parameters. Next we broadly clarify the intuition behind each approach and provide references for some of the more technical details.

The first data augmentation approach is maximum likelihood. Little and Rubin (2019) describes a likelihood function that incorporates the observed data, missing data, and model parameters of scientific interest. When the missing data mechanism is ignorable (i.e., the data are MCAR or MAR), this likelihood is marginalized (i.e., averaged, integrated) over all possible values of the missing items. This is useful because then the likelihood can be maximized with respect to only the observed data and model parameters. Dempster et al. (1977) proposed a novel method to evaluate this likelihood function: The EM algorithm. The EM algorithm consists of two steps: an expectation (E) step, and a maximization (M) step. The algorithm begins with a set

of initial values chosen for the model parameters. Then the E step augments the observed data with plausible values of specific functions of the missing items created using the current model parameters. After the E step, the M step maximizes the likelihood that includes both the observed data and missing data to updated parameter estimates. The E step is then repeated, followed by another M step, yielding new parameter estimates. This process continues until convergence, which means that the parameters are not meaningfully changing from iteration to iteration of the algorithm. The EM algorithm is has seen widespread use in a variety of applications across many scientific disciplines (McLachlan and Krishnan, 2007).

The second data augmentation approach is Bayesian modeling (Gelman et al., 1995; McElreath, 2018; Johnson et al., 2022), which has seen widespread use in ecology (Cressie et al., 2009; Hobbs and Hooten, 2015; Hooten and Hobbs, 2015). Bayesian models differ from frequentist models by assuming underlying model parameters are random variables and using a prior distribution to formally incorporate prior information into the model. Combining the prior information and the observed data yields the posterior distribution of the model parameters. Typically these posterior distributions are sampled directly using a Markov Chain Monte Carlo (MCMC) algorithm like Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970; Chib and Greenberg, 1995; Tierney, 1994), Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990; Casella and George, 1992), or Hamiltonian Monte Carlo (Duane et al., 1987; Carpenter et al., 2017). For a review of various MCMC approaches, see Robert et al. (1999). In a Bayesian model, missing data can be treated as model parameters, combined with the observed data and prior information to yield posterior distributions of each missing item. Samples from the various posterior distributions of each missing item provide valuable insights into their marginal and joint behaviors. These samples can be loosely viewed as "imputations" in that they "fill-in" plausible values of the missing data. They can be also used as part of a formal imputation approach like Bayesian regression imputation (Rubin, 2004).

603 *4.5. Inverse Probability Weighting*

Statistical survey designs leverage probability (i.e., random) sampling to estimate population parameters like means and totals; this is called the design-based approach to statistics (Särndal et al., 2003). The design-based approach, which is based on random sampling, is different from a model-based approach, which is based on assumptions of an underlying model (e.g., a linear regression model) generating the data (Särndal et al., 1978; Brus and De Gruijter, 1997; Dumelle et al., 2022). The design-based approach assigns weights to each unit in the sample that quantify how representative the unit is of the broader population of interest (i.e., target population). These weights generally equal the inverse of the probability the unit was selected in the sample and are called inverse probability weights (IPW, Seaman and White, 2013; Seaman and Vansteelandt, 2018). Usually, the sum of the IPW weights equal the population (or sampling frame) size. For example, consider a simple random sample (SRS) of size n from a known population of size N in which each unit is selected with probability $\frac{n}{N}$. Each IPW equals $\frac{N}{n}$, the sum of which (over the n units selected in the sample) equals

$$\sum_{i=1}^{n} \frac{N}{n} = \frac{nN}{n} = N,$$

604 the known population size.

Units selected in a sample may not be available for data collection, which leads to missing data. Fortunately, IPW can be adjusted, or rescaled, to account for the missing units (Little, 1988a; Little and Rubin, 2019; Seaman and White, 2014; Little et al., 2024). Often the adjusted weights are rescaled to sum to the original population size. When the data are MCAR, this involves multiplying each weight by the inverse of the response probability (Kott, 2012). For example, let the SRS IPW for each sampled unit equal $\frac{N}{np}$, where p (the probability that unit i is sampled) is the ratio $p = \frac{n^*}{n}$, where n^* is the number of sampled units (out of a possible n units).

The sum of the adjusted weights for the n^* sampled units is

$$\sum_{i=1}^{n^*} \frac{N}{np} = \frac{n^* N}{np} = \frac{npN}{np} = N,$$

the known population size. When the data are MAR, the weights can be adjusted within MAR categories (i.e., cells) using a similar rescaling approach (Heeringa et al., 2017). Rescaling over multiple variables is often achieved using approaches like post-stratification, raking, and calibration (Lumley, 2011). Doubly-robust estimators incorporate missingness mechanisms into nonresponse weight adjustments (Robins et al., 1994; Robins, 2000; Seaman and Vansteelandt, 2018). When units are partially sampled (i.e., some items are sampled), weights can be adjusted separately for each variable based its nonresponse rate.

Carpenter and Smuk (2021) note some drawbacks of IPW, one being that weight adjustments and subsequent analyses are sensitive to large weights. Little (1986) makes connections between IPW and imputation, while Reiter et al. (2006) highlight the importance of including variables that influence sampling probabilities (i.e., survey design variables) in an imputation model, and Seaman et al. (2012) combines elements of IPW and imputation. There is debate about whether weighted data should be used in an imputation model when performing design-based analysis on the complete data (Lohr, 2021). Generally, imputation and data augmentation are much more flexible than IPW and hence, receive our primary focus henceforth.

4.6. Other Missing Data Approaches

Here we briefly review several other missing data methods. Van Buuren (2018) provides an overview of imputation approaches specific to certain data types like binary, count, categorical, semi-continuous, hierarchical, longitudinal, multivariate, and censored data. Helsel (2011) discusses imputation in the context of censored ecological data. Imperfect detection, whereby the presence or absence of a species cannot be detected with certainty, is another type of missing data problem (MacKenzie, 2005; Conn et al., 2012; Kellner and Swihart, 2014). Propensity score approaches (Rosenbaum and Rubin, 1983; Rubin, 2001; Austin, 2011; Guo and Fraser, 2014) aim

to match units that have similar items for several variables (i.e., covariates) to control for confounding factors while studying the effect of a treatment. Several authors have accommodated various types of missing data while modeling propensity scores (D'Agostino Jr and Rubin, 2000; Little and Vartivarian, 2005; Cham and West, 2016; Choi et al., 2019). Finally, the missing data indicator method (MDIM) aims to control for the effect of missing data by building it directly into a regression model. More formally, the MDIM adds to a regression model 1) an indicator variable that represents missingness and 2) a variable that is the interaction between the indicator variable in 1) and another variable. The MDIM has mixed results; some highlight its utility (White and Thompson, 2005; Groenwold et al., 2012; Sullivan et al., 2018), while some highlight its drawbacks (Jones, 1996; Donders et al., 2006; Knol et al., 2010; Groenwold et al., 2012).

4.7. An Inferential Comparison of Various Missing Data Methods

Here we compare the performance of 13 unique missing data methods: complete case analysis, three deterministic imputation methods, four single imputation methods, four multiple imputation methods, and a data augmentation method (Table 3). The three deterministic imputation methods are mean imputation, regression imputation, and nearest neighbor imputation. The four single imputation methods are stochastic regression imputation, predictive mean matching (Type-0 matching, three donors), bootstrap regression imputation, and predictive mean matching (Type-1 matching, three donors). Recall that stochastic regression imputation and predictive mean matching (Type-0 matching) incorporate randomness in the imputations but not the underlying imputation approach parameters, while bootstrap regression imputation and predictive mean matching (Type-1 matching) incorporate randomness in both the imputations and underlying imputation approach parameters. The four multiple imputation methods are the multiple imputation version of aforementioned single imputation methods. The data augmentation method is a fully Bayesian model. Henceforth we refer to each method using the abbreviations from Table 3.

We measured the inferential performance of each of the 13 missing data methods by tracking various metrics across 2,000 simulation trials. Each trial was conducted independent of other

trials. Within each trial, 100 observations were independently generated from the model:

$$\mathrm{y}_i = \beta_0 + \beta_1 \mathcal{I}(\mathrm{x}_{1,i}) + \beta_2 \mathrm{x}_{2,i} + \epsilon_i, \tag{5}$$

where y_i is the ith value of a general response variable, $\mathrm{x}_{1,i}$ is the ith value of a general categorical variable that belongs to one of two equally-likely levels ("A" and "B"), $\mathcal{I}(\mathrm{x}_{1,i})$ is an indicator variable defined as

$$\mathcal{I}(\mathrm{x}_{1,i}) = \begin{cases} 1 & \text{if } \mathrm{x}_{1,i} = \mathrm{B} \\ 0 & \text{if } \mathrm{x}_{1,i} = \mathrm{A} \end{cases},$$

$\mathrm{x}_{2,i}$ is ith value of a general continuous variable from a Gaussian distribution with mean zero and variance one, ϵ_i is the ith random error from a Gaussian distribution with mean zero and variance one, β_0 is an intercept parameter, and β_1 and β_2 are slope parameters that control the impact of $\mathcal{I}(\mathrm{x}_{1,i})$ and $\mathrm{x}_{2,i}$ on y_i, respectively. In each simulation trial, the true values of the β_0, β_1, and β_2 were fixed at one, implying Equation (5) can be synonymously rewritten as

$$\mathrm{y}_i = 1 + \mathcal{I}(\mathrm{x}_{1,i}) + \mathrm{x}_{2,i} + \epsilon_i.$$

In each trial, we randomly assigned the 100 simulated observations simulated to distinct training and test sets (Gareth et al., 2013). The training set contained 99 observations and was used to estimate the $\boldsymbol{\beta}$ parameters of a linear regression model with the same form as in Equation (5). The test set contained one observation which was later used to evaluate the fitted linear regression model's predictive capacity. We revisit prediction in Section 5.2.

After simulating the training data, each item in $\mathbf{y}$ and $\mathbf{x}_2$ was randomly and independently assigned to be observed or missing. The probability that each item was missing depended on $\mathbf{x}_1$ (Table 4) and hence, the data were assumed MAR (given $\mathbf{x}_1$). For CCA and the imputation methods, the $\boldsymbol{\beta}$ parameters were estimated using ordinary least squares (or synonymously, restricted maximum likelihood). For FBDA, default priors were assumed for the $\boldsymbol{\beta}$ parameters and each missing item. We used the `mice` package (Van Buuren and Groothuis-Oudshoorn, 2011) for

the imputation methods and the `brms` package (Bürkner, 2017) for FBDA. Both packages are part of the **R** programming language (R Core Team, 2025).

Inferential performance for the 13 missing data methods was measured using three metrics: mean bias, root-mean-squared error, and mean 95% confidence interval coverage. These metrics were calculated for both β_1 and β_2 slope parameters (we ignored the intercept, β_0, as this tends to be of little scientific interest). Mean bias (MBias) is the average deviation of $\hat{\beta}$ (an estimate) from β (the true value) across trials:

$$\text{MBias}(\hat{\beta}) = \frac{1}{2000}\sum_{i=1}^{2{,}000}(\hat{\beta}_i - \beta)$$

Root-Mean-Squared error (RMSE) is the square root of the average squared deviation of $\hat{\beta}$ from β across trials.

$$\text{RMSE}(\hat{\beta}) = \sqrt{\frac{1}{2000}\sum_{i=1}^{2{,}000}(\hat{\beta}_i - \beta)^2}$$

Mean 95% confidence interval coverage (Cover95) is the proportion of 95% Gaussian confidence intervals that contain β:

$$\text{Cover95}(\hat{\beta}) = \frac{1}{2000}\sum_{i=1}^{2{,}000}\mathcal{I}(\text{Cover}_i)$$

$$\mathcal{I}(\text{Cover}_i) = \begin{cases} 1 & \text{if 95LB}(\hat{\beta}_i) \leq \beta \leq \text{95UB}(\hat{\beta}_i) \\ 0 & \text{Otherwise} \end{cases},$$

where $\text{95LB}(\hat{\beta}_i)$ ($\text{95UB}(\hat{\beta}_i)$) is the 95% confidence interval lower (upper) bound for β based on $\hat{\beta}_i$. For FBDA, $\text{95LB}(\hat{\beta}_i)$ and $\text{95UB}(\hat{\beta}_i)$ are actually bounds of 95% credible intervals (using quantiles 0.025 and 0.975), but we call them confidence intervals for consistency with the other methods. Well-fitting models should have mean bias (MBias) close to zero (i.e., parameter estimates are correct on average) and "proper" coverage (Cover95). We define proper coverage as being between 0.94 and 0.96, a range of plausible empirical coverages given natural simulation

variability. For unbiased models with proper coverage, lower RMSE is preferred, as this indicates more precise estimation of β.

Table 5 provides MBias, RMSE, and Cover95 for the missing data methods. All 13 methods except mean imputation had little to no MBias (relative to RMSE). Deterministic imputation methods (Mean, Reg, NN) performed worse (higher RMSE; lower Cover95) than single imputation methods (StReg-S, PMMT0-S, Boot-S, PMMT1-S). Single imputation methods performed worse (higher RMSE; lower Cover95) than their multiple imputation counterparts (StReg-M, PMMT0-M, Boot-M, PMMT1-M). The only methods with proper Cover95 were CCA, Boot-M, PMMT1-M, and FBDA. Among these, Boot-M, PMMT1-M, and FBDA had better (lower) RMSE than CCA. There was little difference in performance among Boot-M, PMMT1-M, and FBDA.

5. Imputation vs Prediction

The terms imputation and prediction are sometimes used synonymously (Eskelson et al., 2009; Ver Hoef and Temesgen, 2013), and, confusingly, this can vary across disciplines. We have been explicit thus far that imputation refers to the process of imputing missing items from units that are part of the complete data set used in a subsequent analysis approach. We define prediction as the process of constructing best guesses (i.e., predictions) of new (i.e., future) items belonging to units **that are distinct** from the units in the complete data set. A well-performing imputation method should yield imputations that reflect each variable's natural, underlying variability and hence, may not be close to the true value of the missing item being imputed. A well-performing predictive model, however, should yield predictions that are as close as possible to the true value of the missing item being predicted. Imputation and prediction are solutions to completely different questions, and as Van Buuren (2018) firmly clarifies: "imputation is not prediction." Evaluating the effectiveness of an imputation method using a predictive metric (i.e., closeness to the true values of the missing items) is bad statistical practice and should be avoided.

5.1. Revisiting the Inferential Comparison

We revisit the inferential comparison from Section 4.7 to show, empirically, that the missing data methods best at recreating missing items are worst at the study's primary goal of estimating the population slope parameters β_1 and β_2. Recall that in each simulation trial, we simulated true values of the explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_2$ and then assigned missingness indicators to $\mathbf{x}_2$. This implies we can compare the imputed values of each missing element in $\mathbf{x}_2$ to the true value. Then we can compute statistics like mean bias (Mbias) and root-mean-squared error (RMSE) of the imputations themselves. More formally, imputation Mbias is

$$\text{MBias}(\text{Imp}(\mathbf{x}_2)) = \frac{1}{2000 \times J} \sum_{i=1}^{2{,}000} \sum_{j} (\tilde{\mathrm{x}}_{2,j} - \mathrm{x}_{2,j}),$$

where i indexes the simulation trials, j indexes the J missing items for $\mathbf{x}_2$ within a simulation trial (which vary from trial to trial), $\tilde{\mathrm{x}}_{2,j}$ is the imputed value of x_2, and $\mathrm{x}_{2,j}$ is the true value of x_2. Similarly, imputation RMSE is

$$\text{RMSE}(\text{Imp}(\mathbf{x}_2)) = \sqrt{\frac{1}{2000 \times J} \sum_{i=1}^{2{,}000} \sum_{j} (\tilde{\mathrm{x}}_{2,j} - \mathrm{x}_{2,j})^2}.$$

All missing data methods produced unbiased imputations. Figure 4 shows that mean and regression imputation had by far the lowest imputation RMSE. Recall, however, that mean and regression imputation had the worst performance regarding the primary goal of efficiently estimating β_1 and β_2. This result seems counterintuitive at first, but Figure 3 helps clarify the intuition. In Figure 3, the mean and regression imputations clearly deviate from the remainder of the observed data while the bootstrap imputations look natural. Full imputation Mbias and RMSE results are provided in the Supporting Information.

5.2. A Predictive Comparison of Various Missing Data Methods

We also evaluated predictive performance of the 13 missing data methods. We used each fitted linear regression model (from Section 4.7) to predict y_{test}, the true value of y from the test data.

Predictions were derived for CCA, the deterministic imputation methods (Mean, Reg, NN), and the single imputation methods (StReg-S, PMMT0-S, Boot-S, PMMT1-S) using the single fitted model:

$$\text{Pred}(\text{y}_{test}) = \tilde{\text{y}}_{test} = \hat{\beta}_0 + \hat{\beta}_1 \mathcal{I}(\text{x}_{1,test}) + \hat{\beta}_2 \text{x}_{2,test},$$

where $\text{x}_{1,test}$ and $\text{x}_{2,test}$ are the x_1 and x_2 values for the test unit and the $\hat{\boldsymbol{\beta}}$ are estimated parameters from the respective linear regression model. We derive the variance of $\text{Pred}(\text{y}_{test})$ in the Supporting Information. For the multiple imputation methods (StReg-M, PMMT0-M, Boot-M, PMMT1-M), predictions of y_{test} were made separately for each complete data set and then pooled using Rubin's Rules to create a final prediction; this multiple imputation approach to prediction is called "Predict-Combine (PC)" (Miles, 2016). Another approach is "Combine-Predict (CP)," where the parameter estimates from each complete data set are pooled and then used to create a final prediction (Figure 5). PC follows Rubin's Rules, but CP can produce similar results in a more computationally efficient manner. We prefer PC, however, because PC adheres to Rubin's Rules. Finally, FBDA makes predictions by sampling from the model's posterior predictive distribution of y_{test}.

Prediction performance metrics were mean prediction bias, root-mean-squared-prediction error, and mean 95% prediction interval coverage of y_{test}. Mean prediction bias (MPBias) is the average deviation of $\tilde{\text{y}}_{test}$ (a prediction) from y_{test} (the true value) across trials:

$$\text{MPBias}(\tilde{\text{y}}_{test}) = \frac{1}{2000} \sum_{i=1}^{2{,}000} (\tilde{\text{y}}_{test,i} - \text{y}_{test,i}). \tag{6}$$

Root-Mean-Squared-Prediction error (RMPSE) is the square root of the average squared deviation $\tilde{\text{y}}_{test}$ from y_{test} across trials:

$$\text{RMPSE}(\tilde{\text{y}}) = \sqrt{\frac{1}{2000} \sum_{i=1}^{2{,}000} (\tilde{\text{y}}_{test,i} - \text{y}_{test,i})^2}. \tag{7}$$

Mean 95% prediction interval coverage (PCover95) is the proportion of 95% Gaussian prediction

intervals that contain y_{test}:

$$\text{PCover95}(\tilde{\mathrm{y}}_{test}) = \frac{1}{2000}\sum_{i=1}^{2{,}000}\mathcal{I}(\text{PCover}_i) \tag{8}$$

$$\mathcal{I}(\text{PCover}_i) = \begin{cases} 1 & 95\text{LB}(\tilde{\mathrm{y}}_{test,i}) \leq \mathrm{y}_{test,i} \leq 95\text{UB}(\tilde{\mathrm{y}}_{test,i}) \\ 0 & \text{Otherwise} \end{cases}.$$

As with the confidence intervals from the inferential comparison, the FBDA prediction intervals ar actually credible intervals, but we call them prediction intervals for consistency with other methods. We call prediction interval coverage proper when it is between 0.94 and 0.96 (Section 4.7).

Table 6 provides MPBias, RMPSE, and PCover95 for all 13 missing data methods. All methods except mean imputation had little to no MPBias (relative to RMPSE). Deterministic imputation methods (Mean, Reg, NN) performed worse (higher RMSPE; lower PCover95) than the single imputation methods (StReg-S, PMMT0-S, Boot-S, PMMT1-S), although NN did have proper coverage. Single imputation methods performed worse (higher RMSPE; lower PCover95) than their multiple imputation counterparts (StReg-M, PMMT0-M, Boot-M, PMMT1-M). The multiple imputation methods and FBDA tended to have the best (lowest) RMSPE, followed closely by CCA and the single imputation methods. Overall, this simulation study suggests that the performance gap between the worst and best missing data methods may be much narrower for prediction (Table 6) than for inference (Table 5).

As with $\mathbf{x}_2$ in Section 5.2, we also compared missing data methods for their capacity to impute $\mathbf{y}$. We found that again, mean and regression imputation most closely recreated the missing elements of $\mathbf{y}$ in the complete (training) data set but performed among the worst at prediction of y_{test} in the test data set. Further details are provided in the Supporting Information.

6. Contingency filters

Sometimes items are missing because there is no basis by which the item can be measured. To formalize this notion, we shortly define a "contingency filter [variable]" and highlight its utility in ecology and other sciences. Contingency filters are closely related to the idea of follow-up (i.e., filter) questions in surveys (Eckman et al., 2014), whereby participants are asked separate sets of questions based on their responses to previous answers.

A contingency filter is a variable with two levels that "filters" an auxiliary variable into two groups, one for each contingency filter level. The levels of the contingency filter determine whether each item of the auxiliary variable is "measurable" or not. An item is measurable if there is a basis by which a measurement can exist. An item is not measurable if there is no basis by which a measurement can exist. In the NWCA 2016 data, surface water presence is a contingency filter for total nitrogen (TN) in surface water. If the wetland has surface water, TN in the surface water can be measured. But if the wetland does not have surface water, TN in the surface water cannot be measured. Contingency filters may also be binary variables of interest themselves. For example, the surface water presence contingency filter can be used characterize the proportion of wetlands with and without surface water.

If an item is not measurable, it is technically "missing", but it is not missing in the same way as MCAR or MAR (or MNAR) missing data. If a wetland does not have surface water, we would not want to assume TN is MCAR or MAR and perform imputation. However, if a wetland does have surface water and TN is missing (e.g., the sample was lost), a MCAR or MAR assumption may be plausible depending on context. The distinction between observable but meaningful missingness and nonmeasurable missingness exists in causal inference literature but (to our knowledge) has not been codified. For example, in studies aimed at measuring a survivor average treatment effect (SATE), whereby one wishes to measure the impact of a treatment on, say, blood pressure, this measurement is only meaningful in patients who survive until the end of the study. See Frangakis and Rubin (2002), Tchetgen Tchetgen (2014), Wang et al. (2017), and Hudgens and Halloran (2006) for more details in the causal inference literature.

The NWCA 2016 characterized units (i.e., sites) as being in Good, Fair, Poor, or Missing (i.e., Not Assessed) condition for TN based on measured data at the site and a comparison to the reference (i.e., undisturbed) benchmark (Stoddard et al., 2006; USEPA, 2023) at the site (Figure 6). All missing TN data resulted from a lack of surface water at the wetland. Leveraging the NWCA survey design, we could use design-based inference to estimate the proportion of wetlands throughout the conterminous United States (CONUS) in either Good, Fair, Poor condition for TN, but these proportion estimates would not sum to one because of the wetlands without surface water. Instead, we can use the surface water presence contingency filter to ask and answer two separate questions: 1) for the CONUS wetlands with surface water, what is the distribution of Good, Fair, and Poor sites for TN?; and 2) what is the proportion of CONUS wetlands with surface water? From Figure 7A, approximately 34%, 26%, and 40% of CONUS wetlands with surface water are in Good, Fair, and Poor condition for TN. And from Figure 7B, approximately 60% of wetlands have surface water and 40% do not. We computed the estimates and uncertainties in Figure 7A and Figure 7B using the IPW weights from the NWCA 2016 survey design, the Horvitz-Thompson estimator (Horvitz and Thompson, 1952; Cordy, 1993), the local neighborhood variance estimator (Stevens Jr and Olsen, 2003), and the `spsurvey` **R** package (Dumelle et al., 2023b).

In the previous example, we showed how contingency filters can be applied to an auxiliary variable or studied as a binary variable of interest. Contingency filters can also be used as explanatory variables in a statistical regression model. Suppose the goal is to study the effect of TN and soil modification (SM, a binary variable) on VMMI using a linear regression model. VMMI and SM are measurable whether or not there is surface water, but TN is only measurable when there is surface water. We could subset the data to include only wetlands with surface water, but this omits valuable data elucidating the impact of SM on VMMI. A more nuanced approach is to incorporate the surface water presence contingency filter by creating a new variable that

captures the interaction between surface water presence and TN:

$$\mathrm{x}_i = \begin{cases} \mathrm{TN}_i & \text{if surface water is present} \\ 0 & \text{if surface water is not present,} \end{cases} \tag{9}$$

where i indexes each site (i.e., unit). Equation (9) clarifies that when there is surface water, x_i equals TN_i, and when there is not surface water, x_i equals zero.

The linear regression model studying the effect of SM and TN on VMMI can be written as

$$\mathrm{VMMI}_i = \beta_0 + \beta_1 \mathcal{I}(\mathrm{SM}_i) + \beta_2 \mathrm{x}_i + \epsilon_i, \tag{10}$$

where

$$\mathcal{I}(\mathrm{SM}_i) = \begin{cases} 1 & \text{if soil modification is present} \\ 0 & \text{if soil modification is not present.} \end{cases}$$

Equation (10) restricts wetlands with surface water and no TN to have the same average VMMI as wetlands without surface water, which may not be realistic. We expand this model by adding an explanatory variable for surface water presence, which allows wetlands with surface water and no TN to have different average VMMI as wetlands without surface water:

$$\mathrm{VMMI}_i = \beta_0 + \beta_1 \mathcal{I}(\mathrm{SM}_i) + \beta_2 \mathcal{I}(\mathrm{SWP}_i) + \beta_3 \mathrm{x}_i + \epsilon_i, \tag{11}$$

where

$$\mathcal{I}(\mathrm{SWP}_i) = \begin{cases} 1 & \text{if surface water is present} \\ 0 & \text{if surface water is not present.} \end{cases}$$

We fit the model in Equation (11) using ordinary least squares (after $\log_e$ transforming TN, a common transformation for chemical concentrations). The linear regression model implied (Table 7) that SM (β_1) was associated with a significant decrease in VMMI (and thus healthier vegetation; p-value < 0.001), SWP was associated with a significant increase in VMMI (p-value

< 0.001), and log TN (given SWP) was not associated with a significant increase or decrease in VMMI (p-value ≈ 0.35). We clarify that this example was chosen to elucidate the utility of contingency filters, not to necessarily find the best possible model for VMMI.

Contingency filters provide a formal structure for handling missing data that are not measurable. They can filter auxiliary variables (e.g., TN condition), be the object of a statistical analysis themselves (e.g., proportion of surface water presence), be used in a statistical model (e.g., a linear regression model) as part of the explanatory variable structure, or for some other application not mentioned here.

7. The Effect of Missing Data on a Spatially Explicit Model

Applying modern missing data methods to spatially explicit statistical models deserves further study, as many data sets in ecology are spatially structured. We call a statistical model "spatially explicit" if it implies a form of spatial dependence among units (i.e., sites). Spatial dependence embodies Tobler's First Law of Geography, which states that nearby units in space tend to be more similar than distant units (Tobler, 1970). Spatially explicit models formally measure spatial dependence using autocovariance (or similarly, autocorrelation) and incorporate it into modeling. The benefit of using a spatially explicit model for spatial data is that spatially explicit models tend to notably outperform their nonspatial counterparts, yielding more reliable inference and more accurate predictions (Cressie, 1993; Schabenberger and Gotway, 2017). Zimmerman and Ver Hoef (2024) thoroughly review spatially explicit models for for environmental and ecological data.

The linear regression model for VMMI in Equation (11) (Section 6) was not spatially explicit. That is, the model assumed that proximity of nearby spatial locations was independent of (i.e., not important for characterizing) VMMI. However, there is evidence of spatial patterning in VMMI (Figure 8), as healthier vegetation (i.e., larger VMMI) tends to be clustered in the Upper Midwest and along the Gulf and Atlantic coasts and less healthy vegetation (i.e., smaller VMMI) tends to be clustered in the Southwest and along the Pacific coast. Thus there is some visual evidence that

VMMI has spatial dependence and hence, our understanding of VMMI can be improved by spatially explicit models.

We build upon the VMMI linear regression model in Equation (11) by adding additional explanatory variables: wetland type (WT), soil hardening (SH), and vegetative removal stress (VRMV). Recall that these variables are defined in Section 2 and Table 1. Again, we don't claim this enhanced linear regression model is the best possible VMMI model, but rather we use it to clarify the effect of missing data on spatially explicit models. The new enhanced model's response structure is given by

$$\begin{aligned}\text{VMMI}_i =& \beta_0 + \beta_1 \mathcal{I}(\text{SM}_i) + \beta_2 \mathcal{I}(\text{SWP}_i) + \beta_3 \text{x}_i + \\ & \beta_4 \text{WT}_i + \beta_5 \text{SH}_i + \beta_6 \mathcal{I}(\text{VRMV-M}_i) + \beta_7 \mathcal{I}(\text{VRMV-H}_i) + \\ & \epsilon_i,\end{aligned}$$

where

$$\mathcal{I}(\text{VRMV-M}_i) = \begin{cases} 1 & \text{if there is medium vegetation removal stress} \\ 0 & \text{if there is low or high vegetation removal stress, and} \end{cases}$$

$$\mathcal{I}(\text{VRMV-H}_i) = \begin{cases} 1 & \text{if there is high vegetation removal stress} \\ 0 & \text{if there is low or medium vegetation removal stress.} \end{cases}$$

Spatially dependence is formally incorporated into the model in Equation 7 through the error term, $\boldsymbol{\epsilon}$. The covariance of the error term is assumed to follow an exponential form, though many other forms exist (for a list, see Zimmerman and Ver Hoef, 2024). The exponential spatial covariance is given by

$$\text{Cov}(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2_{sp} \exp(-\mathbf{h}_{ij}/\phi) & i \neq j \\ \sigma^2_{sp} + \sigma^2_{ind} & i = j \end{cases}, \tag{12}$$

where σ^2_{sp} is a variance parameter that represents spatially explicit variability, $\mathbf{h}_{ij}$ is the Euclidean

(i..e, straight-line) distance between the spatial locations (i.e., coordinates) of VMMI_i and VMMI_j, ϕ is a range parameter that controls the distance-decay rate of the spatial dependence, and σ^2_{ind} is a variance parameter that describes nonspatial (i.e., independent) variability. Equation (12) reflects intuition, as nearby locations (as determined by $\mathbf{h}_{ij}$) share more spatial dependence than distant locations.

To study the impact of missing data on spatially explicit models, we simulated missingness indicators for the SH, SM, and VRMV variables at varying rates (Table 8) that depended on WT and surface water presence (SWP). Thus the data are MAR (given WT and SWP). Then we fit three separate spatially explicit models. The first model used all the (true) data (i.e., there was no simulated missigness). The second model used multiple imputation (MI) with $m = 100$ for the data with missingness. The imputation method was predictive mean matching (Type-1 matching) with three donors. The third model used CCA, removing the units with at least one simulated missing item entirely. Because we simulated the missigness indicators, we could compare the all data (AD) model to the MI and CCA models to quantify the information lost from the missingness. We also compared the fit of the spatially explicit models to their nonspatial counterparts to quantify the information gained from spatial dependence. We fit the spatially explicit models using the using the `spmodel` **R** package (Dumelle et al., 2023a).

We evaluated the inferential performance of the six models (three spatial, three nonspatial) via the $\boldsymbol{\beta}$ slope parameters (Table 9). The estimates and standard errors from the MI models were much closer to those from the AD models than from the CCA models. This suggests that MI recovered more missing information than CCA. Within model type (spatial vs nonspatial), slope parameter p-values from the AD and MI models were much more similar than for the CCA models (Table 9). Slope parameter standard errors for CCA were sometimes twice as large (or more) as their MI counterparts, regardless of whether the models were spatial or nonspatial. These superiority of MI to CCA here is intuitive; there are several variables potentially missing and when a single item is missing, CCA omits the entire unit. Comparing the nonspatial and spatial models, the ratio of the MI slope parameter standard errors to the AD slope parameter

standard errors was smaller for the spatial models. This suggests that some of the information loss from the missing data was mitigated by incorporating spatial dependence.

We also evaluated the predictive capacity of the six models using leave-one-out cross validation (LOOCV, Gareth et al., 2013). LOOCV measures out-of-sample performance by removing a unit from the training data, fitting a model, and predicting the value of the response variable item from the removed unit. Then the removed unit is placed back in the training data the process is repeated for a different unit. After all n units have an associated LOOCV prediction, we can compute predictive metrics by comparing the leave-one-out predictions to their true values. The predictive metrics we computed were mean bias (MPBias), root-mean-squared-prediction error (RMSPE), 95% prediction interval coverage (PCover95), and predictive R-squared (R2). MPBias, RMSPE, and PCover95 are calculated using Equations (6), (7), and (8), respectively. Predictive R-squared is given by

$$\text{Cor}(\mathbf{y}, \tilde{\mathbf{y}}_{LOOCV})^2,$$

the squared correlation between VMMI ($\mathbf{y}$) and its LOOCV predictions ($\tilde{\mathbf{y}}_{LOOCV}$). This definition borrows from the idea that R-squared can be represented as the squared correlation between data and their corresponding modeled values (Rencher and Schaalje, 2008).

All nonspatial and spatial models had MPBias near zero and proper PCover95 (Table 10). The nonspatial models had similar RMSPE for all three methods (AD, MI, CCA), while the spatial AD and MI methods had 11% lower (better) RMSPE than the CCA method. Compared to their nonspatial counterparts, the spatial methods had 18% lower RMSPE for the AD and MI approaches and 6.5% lower RMSPE for the CCA approach. For the nonspatial and spatial models, predictive R2 was larger (better) for the AD and MI methods than the CCA method. Predictive R2 was much larger for the spatial models than for the nonspatial models. While generally there was little difference in predictive performance among the nonspatial models, the AD and MI methods noticeably outperformed CCA among the spatial models.

The results in Tables 9 and 10 suggest that MI was able to recapture much of the information loss from the (sometimes drastic) missing data implied by Table 8, while CCA struggled to keep

up. For both MI and CCA, however, the information loss from the missing data was more noticeable for the slope estimates (Table 9) than for prediction (Tables 10), replicating similar findings from Sections 4.7 and 5.2. Together, these results suggest that 1) MI may be particularly useful in spatial contexts, as MI had a more beneficial impact on the spatial models than the nonspatial models; and 2), incorporating spatial dependence yields more useful models in general, as the spatial models outperformed the nonspatial models for all methods (AD, MI, CCA).

8. Conclusion

Missing data in ecology, and science more broadly, are common and easily mishandled. There are many approaches for handling missing data, each ranging in utility and complexity. Our intent is that this work provides researchers with tools to adequately address their missing data. Concise and explicit recommendations for handling data missing data in the ecological literature are lacking, so we end by providing ten recommendations for ecologists and other scientists to consider.

Recommendation 1: Prevent missing data when possible

The simplest and most effective way to handle missing data is to avoid it in the first place. Prior to sampling, identify potential sources of missing data and plan appropriately. Consider streamlining data collection protocols (e.g., field-based or lab-based protocols) to make accurately recording data easier.

Recommendation 2: Record missing data clearly and unambiguously

Options for handling missing data that are not recorded clearly and unambiguously are limited. Details regarding this cause for missingness are important because they may prove useful while evaluating the plausibility of assumptions like MAR or determining contingency filters. Missing data should be coded with a unique value that cannot be confused with another measured item. For example, missing items should not be coded as a zero if it is possible the measured item can have a true value of zero. Measure variables that are related to perceived causes of missingness,

even if these variables are not of primary scientific interest, as they can be used to inform imputation approaches.

Recommendation 3: Determine whether missing data are MCAR, MAR, MNAR, or not measurable

The missing completely at random (MCAR) assumption is simple yet generally impractical. The MNAR assumption is generally practical but rarely simple. The missing at random (MAR) assumption is a balance of simplicity and practicality, being far more general than MCAR and far more simple than MNAR. The MAR assumption is often misunderstood but is flexible and useful. The MAR assumption is also quite plausible when variables related to missingness are measured and subsequently incorporated (e.g., conditioned upon) into imputation and analysis approaches. Because of this, we believe MAR is a reasonable default for ecologists to assume while handling missing data (that are measurable).

Sometimes missing data are not measurable (e.g., there cannot be total nitrogen measured in the surface water of a wetland when there is no surface water in the wetland). Contingency filters filter an auxiliary variable into two groups based on whether items are measurable. Contingency filters may also be binary variables of interest themselves.

Recommendation 4: Determine whether complete case analysis or inverse probability weighting are appropriate

Complete case analysis (CCA) is a simple, well-performing missing data method when the proportion of units with at least one missing item is relatively low and the missing data are MCAR or MAR. If the missing data are MAR, include variables related to the missingness in the analysis approach (e.g., as explanatory variables in a linear regression model). As the proportion of units with at least one missing item increases, the practicality and performance of CCA relative to other missing data methods diminishes, sometimes rapidly.

Inverse probability weighting (IPW) is a design-based approach (i.e., based on random sampling) for estimating population quantities like means and totals that, like CCA, omits missing data. IPW rescales the weights of the observed data in a way that retains broader population

characteristics. IPW is straightforward, computationally efficient, and useful but does lack the flexibility of other missing data methods like imputation and data augmentation.

Recommendation 5: Use multiple imputation or data augmentation instead of deterministic or single imputation

Multiple imputation and data augmentation are more effective methods for missing data than single or deterministic imputation, a conclusion reached empirically here (Table 5 and Table 6) and repeatedly reinforced throughout decades of research (e.g., Rubin, 1996; Van der Heijden et al., 2006; Van Buuren, 2018; Little and Rubin, 2019). Multiple imputation and data augmentation incorporate uncertainty in the missing items and provide a framework for pooling model results. An advantage of multiple imputation over data augmentation is that the imputation and analysis approaches can be separated. Multiple imputation can naturally incorporate variables related to missingness in the imputation approach but omit these variables in the analysis approach, while data augmentation cannot easily separate these approaches.

Recommendation 6: Set the number of multiple imputations, m, to be at least 100 (when computationally feasible)

As m increases, multiple imputation performance improves. However, this improvement rate diminishes rapidly with increasing m, while the computational cost can increase dramatically. So, if the computational cost associated with multiple imputation is not a concern, set m to be at least 100. If multiple imputation with large m is not computationally feasible, set m to be at least three.

Recommendation 7: Use single imputation methods instead of deterministic imputation methods when imputing a single complete data set

Sometimes multiple imputation and data augmentation are not feasible. Single imputation methods (which have a random component) dramatically outperform deterministic imputation methods (which do not have a random component); see Tables 5 and 6.

Recommendation 8: Favor an inclusive imputation approach over a restrictive one

An inclusive imputation approach requires that all variables used in the analysis approach are also used in the imputation approach. A restrictive imputation approach requires that at least one variable used in the analysis approach is not used in the imputation approach. Favor an inclusive approach, as it helps ensure that imputations are reasonable and informative for complete data set used by the analysis approach. For single or multiple imputation, include the eventual response variable(s) for the analysis approach (e.g., the response variable in a linear regression model) in the imputation approach. For deterministic imputation, omit the eventual response variable(s) for the analysis approach in the imputation approach.

Recommendation 9: Recognize the crucial distinction between imputation and prediction

Imputation is not prediction (Van Buuren, 2018). Imputation is the process of imputing missing items that are part of the complete data set intended for analysis. Prediction is the process of constructing best guesses of new (i.e., future) items that belong to units separate from those in the complete data set (e.g., in time, space, or out-of-sample). A well-performing imputation method should yield imputations that reflect each variable's natural, underlying variability and hence, may not be close to the true item being imputed. A well-performing predictive model, however, should yield predictions that are as close as possible to the item being predicted. Evaluating the effectiveness of an imputation method using a predictive metric (i.e., closeness to the true values of the missing items) is bad statistical practice and should be avoided.

Recommendation 10: Be explicit and transparent about the amount of missing data and which methods were used to handle them

Strive to be explicit and transparent about the sources of missingness, the and amounts of missingness, and the methods used to handle missingness in data. Simply ignoring the missing data without further mention, while convenient, is neither explicit nor transparent. Following this recommendation helps ensure scientific conclusions and decision-making processes are reproducible, replicable, and valid (National Academies, 2019). Some scientific disciplines (e.g.,

clinical trials) have encouraged this practice and provided guidance (e.g., via CONSORT, Moher et al., 2010).

Acknowledgments

The views expressed in this manuscript are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency, the U.S. Geological Survey, or the National Oceanic and Atmospheric Administration. Any mention of trade names, products, or services does not imply an endorsement by the U.S. government, the U.S. Environmental Protection Agency, the U.S. Geological Survey, or the National Oceanic and Atmospheric Administration. The U.S. Environmental Protection Agency, the U.S. Geological Survey, and the National Oceanic and Atmospheric Administration do not endorse any commercial products, services or enterprises.

Author Contributions

- M.D.: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft, Writing – Review and Editing
- R.T.: Conceptualization, Methodology, Software, Visualization, Validation, Writing – Original Draft, Writing – Review and Editing
- A.M.N: Conceptualization, Visualization, Writing – Original Draft, Writing – Review and Editing
- A.R.O.: Conceptualization, Methodology, Writing – Original Draft, Writing – Review and Editing
- K.M.I: Conceptualization, Writing – Original Draft, Writing – Review and Editing
- K.B. Conceptualization, Data Curation, Writing – Original Draft, Writing – Review and Editing

- J.M.VH. Conceptualization, Writing – Original Draft, Writing – Review and Editing
- C.F. Conceptualization, Writing – Original Draft, Writing – Review and Editing

Conflicts of Interest Statement

The authors do not have any conflicts of interest.

References

Abayomi, K., Gelman, A., Levy, M., 2008. Diagnostics for multivariate imputations. Journal of the Royal Statistical Society Series C: Applied Statistics 57, 273–291.

Allison, P.D., 2005. Imputation of categorical variables with proc mi. SUGI 30 proceedings 113, 1–14.

Andridge, R.R., Little, R.J., 2010. A review of hot deck imputation for survey non-response. International Statistical Review 78, 40–64.

Austin, P.C., 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research 46, 399–424.

Austin, P.C., White, I.R., Lee, D.S., van Buuren, S., 2021. Missing data in clinical research: a tutorial on multiple imputation. Canadian Journal of Cardiology 37, 1322–1331.

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple imputation by chained equations: What is it and how does it work? International Journal of Methods in Psychiatric Research 20, 40–49.

Barnard, J., Rubin, D.B., 1999. Small-sample degrees of freedom with multiple imputation. Biometrika 86, 948–955.

Bennett, D.A., 2001. How can i deal with missing data in my study? Australian and New Zealand Journal of Public Health 25, 464–469.

Bodner, T.E., 2008. What improves with increased missing data imputations? Structural Equation Modeling: A Multidisciplinary Journal 15, 651–675.

Bondarenko, I., Raghunathan, T., 2016. Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. Statistics in Medicine 35, 3007–3020.

Bowler, D.E., Boyd, R.J., Callaghan, C.T., Robinson, R.A., Isaac, N.J., Pocock, M.J., 2024. Treating gaps and biases in biodiversity data as a missing data problem. Biological Reviews .

Box, G.E., Tiao, G.C., 2011. Bayesian inference in statistical analysis. John Wiley & Sons.

Brick, J.M., Kalton, G., 1996. Handling missing data in survey research. Statistical Methods in Medical Research 5, 215–238.

Brus, D., De Gruijter, J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–44.

Buck, S.F., 1960. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society: Series B (Methodological) 22, 302–306.

Bürkner, P.C., 2017. brms: An R package for bayesian multilevel models using stan. Journal of Statistical Software 80, 1–28.

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

Carpenter, J.R., Bartlett, J.W., Morris, T.P., Wood, A.M., Quartagno, M., Kenward, M.G., 2023. Multiple imputation and its application. John Wiley & Sons.

Carpenter, J.R., Smuk, M., 2021. Missing data: A statistical framework for practice. Biometrical Journal 63, 915–947.

Casella, G., George, E.I., 1992. Explaining the Gibbs sampler. The American Statistician 46, 167–174.

Cham, H., West, S.G., 2016. Propensity score analysis with missing data. Psychological Methods 21, 427.

Chen, J., Shao, J., 2000. Nearest neighbor imputation for survey data. Journal of Official Statistics 16, 113.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician 49, 327–335.

Choi, J., Dekkers, O.M., le Cessie, S., 2019. A comparison of different methods to handle missing data in the context of propensity score analysis. European Journal of Epidemiology 34, 23–36.

Collins, L.M., Schafer, J.L., Kam, C.M., 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods 6, 330.

Conn, P.B., Johnson, D.S., London, J.M., Boveng, P.L., 2012. Accounting for missing data when assessing availability in animal population surveys: An application to ice-associated seals in the Bering Sea. Methods in Ecology and Evolution 3, 1039–1046.

Cordy, C.B., 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Statistics & Probability Letters 18, 353–362.

Cowardin, L., 1979. Classification of wetlands and deepwater habitats of the united states. US Department of the Interior Fish and Wildlife Service Office of Biological Services .

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., Kenward, M.G., 2011. Generalized shared-parameter models and missingness at random. Statistical Modelling 11, 279–310.

Cressie, N., 1993. Statistics for Spatial Data. Wiley, Hoboken, NJ.

Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M.V., Wikle, C.K., 2009. Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. Ecological Applications 19, 553–570.

D'Agostino Jr, R.B., Rubin, D.B., 2000. Estimating and using propensity scores with partially missing data. Journal of the American Statistical Association 95, 749–759.

De Leeuw, E.D., 2001. Reducing missing data in surveys: An overview of methods. Quality and Quantity 35, 147–160.

Demirtas, H., Freels, S.A., Yucel, R.M., 2008. Plausibility of multivariate normality assumption

when multiply imputing non-gaussian continuous outcomes: A simulation assessment. Journal of Statistical Computation and Simulation 78, 69–84.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39, 1–22.

Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T., Moons, K.G., 2006. A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology 59, 1087–1091.

Dray, S., Josse, J., 2015. Principal component analysis with missing values: A comparative survey of methods. Plant Ecology 216, 657–667.

Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D., 1987. Hybrid monte carlo. Physics Letters B 195, 216–222.

Dumelle, M., Higham, M., Ver Hoef, J.M., 2023a. spmodel: Spatial statistical modeling and prediction in r. PLOS ONE 18, e0282524.

Dumelle, M., Higham, M., Ver Hoef, J.M., Olsen, A.R., Madsen, L., 2022. A comparison of design-based and model-based approaches for finite population spatial sampling and inference. Methods in Ecology and Evolution 13, 2018–2029.

Dumelle, M., Kincaid, T., Olsen, A.R., Weber, M., 2023b. spsurvey: Spatial sampling design and analysis in r. Journal of Statistical Software 105, 1–29.

D'Agostino McGowan, L., Lotspeich, S.C., Hepler, S.A., 2024. The "why" behind including "y" in your imputation model. Statistical Methods in Medical Research 33, 996–1020.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., Presser, S., 2014. Assessing the mechanisms of misreporting to filter questions in surveys. Public Opinion Quarterly 78, 721–733.

Eddings, W., Marchenko, Y., 2012. Diagnostics for multiple imputation in stata. The Stata Journal 12, 353–367.

Efron, B., 1994. Missing data, imputation, and the bootstrap. Journal of the American Statistical Association 89, 463–475.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. CRC press.

Ellington, E.H., Bastille-Rousseau, G., Austin, C., Landolt, K.N., Pond, B.A., Rees, E.E., Robar, N., Murray, D.L., 2015. Using multiple imputation to estimate missing data in meta-regression. Methods in Ecology and Evolution 6, 153–163.

Enders, C.K., 2022. Applied missing data analysis. Guilford Publications.

Eskelson, B.N., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T., 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research 24, 235–246.

Fox, J., 2019. Regression diagnostics: An introduction. Sage publications.

Frangakis, C.E., Rubin, D.B., 2002. Principal Stratification in Causal Inference. Biometrics 58, 21–29.

Fraser, G.E., Yan, R., Butler, T.L., Jaceldo-Siegl, K., Beeson, W.L., Chan, J., 2009. Missing data in a long food frequency questionnaire: Are imputed zeroes correct? Epidemiology 20, 289–294.

Gareth, J., Daniela, W., Trevor, H., Robert, T., 2013. An introduction to statistical learning: With applications in R. Spinger.

Gelfand, A.E., Smith, A.F., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian data analysis. Chapman and Hall/CRC.

Gelman, A., King, G., Liu, C., 1998. Not asked and not answered: Multiple imputation for multiple surveys. Journal of the American Statistical Association 93, 846–857.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D.F., Meulders, M., 2005. Multiple imputation for model checking: Completed-data plots with missing and latent data. Biometrics 61, 74–85.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence , 721–741.

Gleason, T.C., Staelin, R., 1975. A proposal for handling missing data. Psychometrika 40, 229–252.

Glynn, R.J., Laird, N.M., Rubin, D.B., 2013. Selection modeling versus mixture modeling with nonignorable nonresponse, in: Drawing inferences from self-selected samples. Routledge, pp. 115–142.

Graham, J.W., 2009. Missing data analysis: Making it work in the real world. Annual Review of Psychology 60, 549–576.

Graham, J.W., 2012. Missing data: Analysis and design. Springer Science & Business Media.

Graham, J.W., Olchowski, A.E., Gilreath, T.D., 2007. How many imputations are really needed? some practical clarifications of multiple imputation theory. Prevention Science 8, 206–213.

Graham, J.W., Taylor, B.J., Olchowski, A.E., Cumsille, P.E., 2006. Planned missing data designs in psychological research. Psychological Methods 11, 323.

Groenwold, R.H., White, I.R., Donders, A.R.T., Carpenter, J.R., Altman, D.G., Moons, K.G., 2012. Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. Cmaj 184, 1265–1269.

Guo, S., Fraser, M.W., 2014. Propensity score analysis: Statistical methods and applications. SAGE publications.

Hastings, W.K., 1970. Monte carlo sampling methods using markov chains and their applications. Biometrika 57.

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, in: Annals of economic and social measurement, volume 5, number 4. NBER, pp. 475–492.

Heeringa, S.G., West, B.T., Berglund, P.A., 2017. Applied survey data analysis. Chapman and Hall/CRC.

Van der Heijden, G.J., Donders, A.R.T., Stijnen, T., Moons, K.G., 2006. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. Journal of Clinical Epidemiology 59, 1102–1109.

Heitjan, D.F., Basu, S., 1996. Distinguishing "missing at random" and "missing completely at random". The American Statistician 50, 207–213.

Helsel, D.R., 2011. Statistics for censored environmental data using Minitab and R. John Wiley & Sons.

Herlihy, A.T., Kentula, M.E., Magee, T.K., Lomnicky, G.A., Nahlik, A.M., Serenbetz, G., 2019. Striving for consistency in the national wetland condition assessment: Developing a reference condition approach for assessing wetlands at a continental scale. Environmental Monitoring and Assessment 191, 1–20.

Hobbs, N.T., Hooten, M.B., 2015. Bayesian models: A statistical primer for ecologists. Princeton University Press.

Hooten, M.B., Hobbs, N.T., 2015. A guide to bayesian model selection for ecologists. Ecological Monographs 85, 3–28.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association 47, 663–685.

Hossie, T.J., Gobin, J., Murray, D.L., 2021. Confronting missing ecological data in the age of pandemic lockdown. Frontiers in Ecology and Evolution 9, 669477.

Hudgens, M.G., Halloran, M.E., 2006. Causal Vaccine Effects on Binary Postinfection Outcomes. Journal of the American Statistical Association 101, 51–64.

Hughes, R.A., Heron, J., Sterne, J.A., Tilling, K., 2019. Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. International Journal of Epidemiology 48, 1294–1304.

Johnson, A.A., Ott, M.Q., Dogucu, M., 2022. Bayes rules!: An introduction to applied Bayesian modeling. Chapman and Hall/CRC.

Johnson, T.F., Isaac, N.J., Paviolo, A., González-Suárez, M., 2021. Handling missing values in trait data. Global Ecology and Biogeography 30, 51–62.

Jones, M.P., 1996. Indicator and stratification methods for missing explanatory variables in multiple linear regression. Journal of the American Statistical Association 91, 222–230.

1241 Kalton, G., Kasprzyk, D., 1986. The treatment of missing survey data. Survey Methodology 12,
1242 1–16.
1243 Kang, H., 2013. The prevention and handling of the missing data. Korean Journal of
1244 Anesthesiology 64, 402–406.
1245 Kellner, K.F., Swihart, R.K., 2014. Accounting for imperfect detection in ecology: A quantitative
1246 review. PLOS ONE 9, e111436.
1247 Kentula, M.E., Paulsen, S.G., 2019. The 2011 national wetland condition assessment: Overview
1248 and an invitation. Environmental Monitoring and Assessment 191, 325.
1249 Kenward, M.G., Carpenter, J., 2007. Multiple imputation: Current perspectives. Statistical
1250 Methods in Medical Research 16, 199–218.
1251 Kim, S.W., Blomberg, S.P., Pandolfi, J.M., 2018. Transcending data gaps: A framework to reduce
1252 inferential errors in ecological analyses. Ecology Letters 21, 1200–1210.
1253 Kleinke, K., 2017. Multiple imputation under violated distributional assumptions: A systematic
1254 evaluation of the assumed robustness of predictive mean matching. Journal of Educational and
1255 Behavioral Statistics 42, 371–404.
1256 Knol, M.J., Janssen, K.J., Donders, A.R.T., Egberts, A.C., Heerdink, E.R., Grobbee, D.E., Moons,
1257 K.G., Geerlings, M.I., 2010. Unpredictable bias when using the missing indicator method or
1258 complete case analysis for missing confounder values: an empirical example. Journal of
1259 Clinical Epidemiology 63, 728–736.
1260 Kott, P.S., 2012. Why one should incorporate the design weights when adjusting for unit
1261 nonresponse using response homogeneity groups. Survey Methodology 38, 95–99.
1262 Kromrey, J.D., Hines, C.V., 1994. Nonrandomly missing data in multiple regression: An
1263 empirical comparison of common missing-data treatments. Educational and Psychological
1264 Measurement 54, 573–593.
1265 Laaksonen, S., 2018. Survey methodology and missing data. Survey Methodology and Missing
1266 Data: Tools and Techniques for Practitioners. Berlin, Germany: Springer International
1267 Publishing .

Lin, J.Y., Lu, Y., Tu, X., 2012. How to avoid missing data and the problems they pose: Design considerations. Shanghai Archives of Psychiatry 24, 181.

Lipsitz, S., Parzen, M., Zhao, L.P., 2002. A degrees-of-freedom approximation in multiple imputation. Journal of Statistical Computation and Simulation 72, 309–318.

Little, R., 2008. Selection and pattern-mixture models, in: Longitudinal Data Analysis. Chapman and Hall/CRC, pp. 423–446.

Little, R.J., 1986. Survey nonresponse adjustments for estimates of means. International Statistical Review/Revue Internationale de Statistique , 139–157.

Little, R.J., 1988a. Missing data adjustments in large surveys. Journal of Business & Economic Statistics 6, 287–296.

Little, R.J., 1988b. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association 83, 1198–1202.

Little, R.J., 2021. Missing Data Assumptions. Annual Review of Statistics and Its Application 8, 89–107. doi:10.1146/annurev-statistics-040720-031104.

Little, R.J., Carpenter, J.R., Lee, K.J., 2024. A comparison of three popular methods for handling missing data: Complete-case analysis, inverse probability weighting, and multiple imputation. Sociological Methods & Research 53, 1105–1135.

Little, R.J., D'agostino, R., Cohen, M.L., Dickersin, K., Emerson, S.S., Farrar, J.T., Frangakis, C., Hogan, J.W., Molenberghs, G., Murphy, S.A., et al., 2012. The prevention and treatment of missing data in clinical trials. New England Journal of Medicine 367, 1355–1360.

Little, R.J., Rubin, D.B., 1989. The analysis of social science data with missing values. Sociological Methods & Research 18, 292–326.

Little, R.J., Rubin, D.B., 2019. Statistical analysis with missing data. 3rd ed., John Wiley & Sons.

Little, R.J., Vartivarian, S., 2005. Does weighting for nonresponse increase the variance of survey means? Survey Methodology 31, 161.

Lohr, S.L., 2021. Sampling: Design and analysis. CRC press.

Łopucki, R., Kiersztyn, A., Pitucha, G., Kitowski, I., 2022. Handling missing data in ecological

studies: Ignoring gaps in the dataset can distort the inference. Ecological Modelling 468, 109964.

Lumley, T., 2011. Complex surveys: A guide to analysis using R. John Wiley & Sons.

MacKenzie, D.I., 2005. What are the issues with presence-absence data for wildlife managers? The Journal of Wildlife Management 69, 849–860.

Magee, T.K., Blocksom, K.A., Fennessy, M.S., 2019. A national-scale vegetation multimetric index (VMMI) as an indicator of wetland condition across the conterminous united states. Environmental Monitoring and Assessment 191, 322.

Marshall, A., Altman, D.G., Royston, P., Holder, R.L., 2010. Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. BMC Medical Research Methodology 10, 1–16.

McCauley, D., Arnold, W., Saxton, J., Turner, C., 2019. Applying adaptive management and lessons learned from national assessments to address logistical challenges in the national wetland condition assessment. Environmental Monitoring and Assessment 191, 329.

McElreath, R., 2018. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC.

McLachlan, G.J., Krishnan, T., 2007. The EM algorithm and extensions. John Wiley & Sons.

Mealli, F., Rubin, D.B., 2015. Clarifying missing at random and related definitions, and implications when coupled with exchangeability. Biometrika 102, 995–1000. doi:10.1093/biomet/asv035.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, 1087–1092.

Miles, A., 2016. Obtaining predictions from models fit to multiply imputed data. Sociological Methods & Research 45, 175–185.

Miller, R.G., 1974. The jackknife-a review. Biometrika 61, 1–15.

Mitsch, W.J., Gosselink, J.G., Anderson, C.J., Fennessy, M.S., 2023. Wetlands, 6th Ed. John Wiley & Sons.

Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P.J., Elbourne, D., Egger, M., Altman, D.G., 2010. Consort 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. BMJ 340.

Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A., Verbeke, G., 2014. Handbook of missing data methodology. CRC Press.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2021. Introduction to linear regression analysis. John Wiley & Sons.

Moons, K.G., Donders, R.A., Stijnen, T., Harrell Jr, F.E., 2006. Using the outcome for imputation of missing predictor values was preferred. Journal of Clinical Epidemiology 59, 1092–1101.

Morris, T.P., White, I.R., Royston, P., 2014. Tuning multiple imputation by predictive mean matching and local residual draws. BMC Medical Research Methodology 14, 1–13.

Myers, T.A., 2011. Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. Communication Methods and Measures 5, 297–310.

Nakagawa, S., 2015. Missing data: Mechanisms, methods and messages. Ecological Statistics: Contemporary Theory and Application , 81–105.

Nakagawa, S., Freckleton, R.P., 2008. Missing inaction: The dangers of ignoring missing data. Trends in Ecology & Evolution 23, 592–596.

Nakagawa, S., Freckleton, R.P., 2011. Model averaging, missing data and multiple imputation: A case study for behavioural ecology. Behavioral Ecology and Sociobiology 65, 103–116.

National Academies, 2019. Reproducibility and Replicability in Science. The National Academies Press, Washington, DC. URL: https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science, doi:10.17226/25303.

Newman, D.A., 2014. Missing data: Five practical guidelines. Organizational Research Methods 17, 372–411.

Nguyen, C.D., Carlin, J.B., Lee, K.J., 2017. Model checking in multiple imputation: An overview and case study. Emerging Themes in Epidemiology 14, 1–12.

Noble, D.W., Nakagawa, S., 2021. Planned missing data designs and methods: Options for strengthening inference, increasing research efficiency and improving animal welfare in ecological and evolutionary research. Evolutionary Applications 14, 1958–1968.

Olinsky, A., Chen, S., Harlow, L., 2003. The comparative efficacy of imputation methods for missing data in structural equation modeling. European Journal of Operational Research 151, 53–79.

Olsen, A.R., Kincaid, T.M., Kentula, M.E., Weber, M.H., 2019. Survey design to assess condition of wetlands in the United States. Environmental Monitoring and Assessment 191, 268.

O'neill, R., Temple, R., 2012. The prevention and treatment of missing data in clinical trials: An FDA perspective on the importance of dealing with it. Clinical Pharmacology & Therapeutics 91, 550–554.

Pearson, R.K., 2006. The problem of disguised missing data. Acm Sigkdd Explorations Newsletter 8, 83–92.

Pedersen, A.B., Mikkelsen, E.M., Cronin-Fenton, D., Kristensen, N.R., Pham, T.M., Pedersen, L., Petersen, I., 2017. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology , 157–166.

Penone, C., Davidson, A.D., Shoemaker, K.T., Di Marco, M., Rondinini, C., Brooks, T.M., Young, B.E., Graham, C.H., Costa, G.C., 2014. Imputation of missing data in life-history trait datasets: Which approach performs the best? Methods in Ecology and Evolution 5, 961–970.

Perkins, N.J., Cole, S.R., Harel, O., Tchetgen Tchetgen, E.J., Sun, B., Mitchell, E.M., Schisterman, E.F., 2018. Principled approaches to missing data in epidemiologic studies. American Journal of Epidemiology 187, 568–575.

R Core Team, 2025. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rancourt, E., Särndal, C., Lee, H., 1994. Estimation of the variance in the presence of nearest

neighbor imputation, in: Proceedings of the section on survey research methods, American Statistical Association. pp. 888–893.

Reiter, J.P., 2007. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. Biometrika 94, 502–508.

Reiter, J.P., Raghunathan, T.E., Kinney, S.K., 2006. The importance of modeling the sampling design in multiple imputation for missing data. Survey Methodology 32, 143.

Rencher, A.C., Schaalje, G.B., 2008. Linear models in statistics. John Wiley & Sons.

Robert, C.P., Casella, G., Casella, G., 1999. Monte Carlo statistical methods. Springer.

Robins, J.M., 2000. Robust estimation in sequentially ignorable missing data and causal inference models, in: Proceedings of the American Statistical Association, Indianapolis, IN. pp. 6–10.

Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846–866.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Roth, P.L., 1994. Missing data: A conceptual review for applied psychologists. Personnel Psychology 47, 537–560.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Rubin, D.B., 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. Journal of the American Statistical Association 72, 538–543.

Rubin, D.B., 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. Journal of Business & Economic Statistics 4, 87–94.

Rubin, D.B., 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91, 473–489.

Rubin, D.B., 2001. Using propensity scores to help design observational studies: Application to the tobacco litigation. Health Services and Outcomes Research Methodology 2, 169–188.

Rubin, D.B., 2004. Multiple imputation for nonresponse in surveys. John Wiley & Sons.

Särndal, C.E., Swensson, B., Wretman, J., 2003. Model assisted survey sampling. Springer Science & Business Media.

Särndal, C.E., Thomsen, I., Hoem, J.M., Lindley, D., Barndorff-Nielsen, O., Dalenius, T., 1978. Design-based and model-based inference in survey sampling [with discussion and reply]. Scandinavian Journal of Statistics , 27–52.

Saunders, J.A., Morrow-Howell, N., Spitznagel, E., Doré, P., Proctor, E.K., Pescarino, R., 2006. Imputing missing data: A comparison of methods for social work researchers. Social Work Research 30, 19–31.

Schabenberger, O., Gotway, C.A., 2017. Statistical methods for spatial data analysis. CRC press, New York.

Schafer, J.L., 1997. Analysis of incomplete multivariate data. CRC press.

Schafer, J.L., Graham, J.W., 2002. Missing data: Our view of the state of the art. Psychological Methods 7, 147.

Schafer, J.L., Olsen, M.K., 1998. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research 33, 545–571.

Scharf, H., Hooten, M.B., Johnson, D.S., 2017. Imputation approaches for animal movement modeling. Journal of Agricultural, Biological and Environmental Statistics 22, 335–352.

Schenker, N., Taylor, J.M., 1996. Partially parametric techniques for multiple imputation. Computational Statistics & Data Analysis 22, 425–446.

Scheuren, F., 2005. Multiple imputation: How it began and continues. The American Statistician 59, 315–319.

Schlomer, G.L., Bauman, S., Card, N.A., 2010. Best practices for missing data management in counseling psychology. Journal of Counseling Psychology 57, 1.

Schoemann, A.M., Moore, E.W.G., Yagiz, G., 2024. How and why to follow best practices for testing mediation models with missing data. International Journal of Psychology .

Schomaker, M., Heumann, C., 2018. Bootstrap inference when using multiple imputation. Statistics in Medicine 37, 2252–2266.

Seaman, S., Galati, J., Jackson, D., Carlin, J., 2013. What is meant by "missing at random"? Statistical Science 28, 257–268.

Seaman, S., White, I., 2014. Inverse probability weighting with missing predictors of treatment assignment or missingness. Communications in Statistics-Theory and Methods 43, 3499–3515.

Seaman, S.R., Vansteelandt, S., 2018. Introduction to double robust methods for incomplete data. Statistical Science: A Review Journal of the Institute of Mathematical Statistics 33, 184.

Seaman, S.R., White, I.R., 2013. Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research 22, 278–295.

Seaman, S.R., White, I.R., Copas, A.J., Li, L., 2012. Combining multiple imputation and inverse-probability weighting. Biometrics 68, 129–137.

Siddique, J., Belin, T.R., 2008. Multiple imputation using an iterative hot-deck with distance-based donor selection. Statistics in Medicine 27, 83–102.

Srebotnjak, T., Carr, G., de Sherbinin, A., Rickwood, C., 2012. A global water quality index and hot-deck imputation of missing data. Ecological Indicators 17, 108–119.

Stevens Jr, D.L., Olsen, A.R., 2003. Variance estimation for spatially balanced samples of environmental resources. Environmetrics 14, 593–610.

Stevens Jr, D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99, 262–278.

Stoddard, J.L., Larsen, D.P., Hawkins, C.P., Johnson, R.K., Norris, R.H., 2006. Setting expectations for the ecological condition of streams: The concept of reference condition. Ecological Applications 16, 1267–1276.

Su, Y.S., Gelman, A., Hill, J., Yajima, M., 2011. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. Journal of Statistical Software 45, 1–31.

Sullivan, T.R., White, I.R., Salter, A.B., Ryan, P., Lee, K.J., 2018. Should multiple imputation be the method of choice for handling missing data in randomized trials? Statistical Methods in Medical Research 27, 2610–2626.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–540.

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O., Amiaud, B., 2014. Filling the gap in functional trait databases: Use of ecological hypotheses to replace missing data. Ecology and Evolution 4, 944–958.

Tchetgen Tchetgen, E.J., 2014. Identification and estimation of survivor average causal effects. Statistics in Medicine 33, 3601–3628.

Ter Braak, C., Van Strien, A., Meijer, R., Verstrael, T., 1994. Analysis of monitoring data with many missing values: which method?, in: Bird Numbers 1992. Distribution, monitoring and ecological aspects: Proceedings 12th International Conference of IBCC and EOAC, Noordwijkerhout, The Netherlands, pp. 663–673.

Tierney, L., 1994. Markov chains for exploring posterior distributions. The Annals of Statistics , 1701–1728.

Tilling, K., Williamson, E.J., Spratt, M., Sterne, J.A., Carpenter, J.R., 2016. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. Journal of Clinical Epidemiology 80, 107–115.

Tobler, W.R., 1970. A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240.

USDA-NRCS, 2020. The plants database. URL: (`http://plants.usda.gov`).

USEPA, 2015. National Wetland Condition Assessment 2016: Laboratory Operations Manual. U.S. Environmental Protection Agency. Washington, DC, USA.

USEPA, 2016. National Wetland Condition Assessment 2016: Field Operations Manual. U.S. Environmental Protection Agency. Washington, DC, USA. URL: `https://www.epa.gov/sites/default/files/2017-08/documents/nwca2016_fom_v1_1a_full_0.pdf`.

USEPA, 2023. National Wetland Condition Assessment: 2016 Technical Support Document. U.S. Environmental Protection Agency. Washington, DC, USA. URL:

https://www.epa.gov/system/files/documents/2023-04/NWCA%202016%20Technical%20Support%20Document_20230216.pdf.

Vach, W., 2012. Logistic regression with missing values in the covariates. Springer Science & Business Media.

Van Buuren, S., 2018. Flexible imputation of missing data. CRC press.

Van Buuren, S., Brand, J.P., Groothuis-Oudshoorn, C.G., Rubin, D.B., 2006. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 76, 1049–1064.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45, 1–67.

Van Deusen, P.C., 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. Canadian Journal of Forest Research 27, 379–384.

Ver Hoef, J.M., Temesgen, H., 2013. A comparison of the spatial linear model to nearest neighbor (k-nn) methods for forestry applications. PLOS ONE 8, e59129.

Vink, G., Frank, L.E., Pannekoek, J., Van Buuren, S., 2014. Predictive mean matching imputation of semicontinuous variables. Statistica Neerlandica 68, 61–90.

Von Hippel, P.T., 2009. How to impute interactions, squares, and other transformed variables. Sociological Methodology 39, 265–291.

Wagstaff, D.A., Harel, O., 2011. A closer examination of three small-sample approximations to the multiple-imputation degrees of freedom. The Stata Journal 11, 403–419.

Wang, L., Zhou, X.H., Richardson, T.S., 2017. Identification and estimation of causal effects with outcomes truncated by death. Biometrika 104, 597–612.

White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine 30, 377–399.

White, I.R., Thompson, S.G., 2005. Adjusting for partially missing baseline measurements in randomized trials. Statistics in Medicine 24, 993–1007.

Wilen, B.O., Bates, M., 1995. The US Fish and Wildlife Service's National Wetlands Inventory Project. Classification and Inventory of the World's Wetlands , 153–169.

Wilks, S.S., 1932. Moments and distributions of estimates of population parameters from fragmentary samples. The Annals of Mathematical Statistics 3, 163–195.

Wood, A.M., White, I.R., Thompson, S.G., 2004. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical Trials 1, 368–376.

Zimmerman, D.L., Ver Hoef, J.M., 2024. Spatial Linear Models for Environmental Data. CRC Press.

Tables

Class	Name	Description	Abbreviation	Valid Values
Soil	Hardening	Soil hardening index	SH	[0, 424]
Soil	Modification	Whether soil was modified	SM	Yes; No
Vegetation	Type	Dominant wetland vegetation type	WT	Woody; Herbaceous
Vegetation	Removal	Vegetation removal stressor	VRMV	Low; Medium; High
Vegetation	VMMI	Vegetation Multimetric Index	VMMI	[0, 100]
Surface Water	Presence	Whether surface water is present	SWP	Yes; No
Surface Water	Nitrogen	Total nitrogen in surface water	TN	NA (no water); [0, Inf) mg/L (water)

Table 1: NWCA 2016 variables, descriptions, and valid values. Brackets indicate a continuous range of values from the first to last number.

Notation	Description
$\mathbf{Y}$	All possible values of each item
$\mathbf{y}$	A single value of each item
$\mathbf{R}$	All possible missingness patterns
$\mathbf{r}$	A single missingness pattern
$\mathbf{Y}_{obs}$	All possible observed subsets of $\mathbf{Y}$
$\mathbf{y}_{obs}$	A single observed subset of $\mathbf{Y}$
$\mathbf{Y}_{mis}$	All possible missing subsets of $\mathbf{Y}$
$\mathbf{y}_{mis}$	A single missing subset of $\mathbf{Y}$

Table 2: Descriptions of missing data notation.

Method	Abbreviation	Imputations	Parameters
Complete Case Analysis	CCA	-	-
Mean Imputation	Mean	Deterministic	Deterministic
Regression Imputation	Reg	Deterministic	Deterministic
Nearest Neighbor Imputation	NN	Deterministic	Deterministic
Stochastic Regression Imputation (Single)	StReg-S	Random	Deterministic
Predictive Mean Matching Type-0 Imputation (Single)	PMMT0-S	Random	Deterministic
Bootstrap Regression Imputation (Single)	Boot-S	Random	Random
Predictive Mean Matching Type-1 Imputation (Single)	PMMT1-S	Random	Random
Random Regression Imputation (Multiple)	StReg-M	Random	Deterministic
Predictive Mean Matching Type-0 Imputation (Multiple)	PMMT0-M	Random	Deterministic
Bootstrap Regression Imputation (Multiple)	Boot-M	Random	Random
Predictive Mean Matching Type-1 Imputation (Multiple)	PMMT1-M	Random	Random
Fully Bayesian Data Augmentation	FBDA	Random	Random

Table 3: Missing data methods and abbreviations used in the simulation study. The "Imputations" column clarifies whether the imputation process is deterministic (i.e., nonrandom) or has a random component. The "Parameters" column clarifies whether the parameters used to fill in missing data items are deterministic or have a random component.

Missingness Probabilities		
	$\mathbf{y}$	$\mathbf{x}_2$
$\mathbf{x}_1$ = A	0.35	0.7
$\mathbf{x}_1$ = B	0.7	0.35

Table 4: Missingness probabilities for $\mathbf{y}$ and $\mathbf{x}_2$ given $\mathbf{x}_1$.

	β_1			β_2		
Method	MBias	RMSE	Cover95	MBias	RMSE	Cover95
CCA	0.00	0.49	**0.95**	-0.00	0.26	**0.95**
Mean	0.59	0.62	0.16	0.59	0.61	0.06
Reg	0.00	0.47	0.31	-0.31	0.40	0.12
NN	0.04	0.51	0.65	0.32	0.45	0.38
StReg-S	0.00	0.46	0.60	0.02	0.26	0.59
PMMT0-S	0.08	0.46	0.62	0.10	0.31	0.55
Boot-S	-0.00	0.54	0.52	0.00	0.28	0.55
PMMT1-S	0.07	0.52	0.56	0.12	0.33	0.53
StReg-M	-0.00	0.40	0.87	0.02	0.22	0.91
PMMT0-M	0.08	0.40	0.88	0.11	0.25	0.90
Boot-M	0.00	0.40	**0.95**	0.00	0.22	**0.95**
PMMT1-M	0.07	0.39	**0.96**	0.12	0.26	**0.94**
FBDA	0.00	0.38	**0.95**	0.04	0.22	**0.95**

Table 5: Missing data method inferential performance across 2,000 independent simulation trials. Mean bias (MBias), root-mean-squared error (RMSE) and 95% confidence interval coverage (Cover95) are reported separately for the β_1 and β_2 slope parameters. The Method and Cover95 columns are bolded when the method has proper Cover95 (i.e., between 0.94 and 0.96).

	Predictions		
Method	MPBias	RMSPE	PCover95
CCA	-0.01	1.08	**0.96**
Mean	-0.21	1.24	0.88
Reg	0.00	1.12	0.57
NN	-0.02	1.17	**0.96**
StReg-S	-0.00	1.09	0.92
PMMT0-S	-0.02	1.10	0.93
Boot-S	-0.02	1.11	0.90
PMMT1-S	-0.02	1.12	0.92
StReg-M	-0.01	1.06	**0.94**
PMMT0-M	-0.02	1.07	**0.94**
Boot-M	-0.01	1.06	**0.94**
PMMT1-M	-0.02	1.07	**0.95**
FBDA	-0.01	1.06	**0.95**

Table 6: Missing data method prediction performance across 2,000 independent simulation trials. Mean prediction bias (MPBias), root-mean-squared-prediction error (RMPSE) and 95% prediction interval coverage (PCover95) are reported for predictions. The Method and PCover95 columns are bolded when the method has proper PCover95 (i.e., between 0.94 and 0.96).

Parameter	Estimate	SE	p-value
β_0 (Intercept)	53.88	1.16	< 0.001
β_1 (SM)	-10.49	1.49	< 0.001
β_2 (SWP)	8.94	1.40	< 0.001
β_3 (Log TN, SWP)	0.77	0.83	0.35

Table 7: Parameter estimates, standard errors, and p-values for the model in Equation (10).

Missingness Probabilities			
	SH	SM	VRMV
WT = Woody, SWP = Present	0.2	0.2	0.2
WT = Herbaceous, SWP = Present	0.4	0.4	0.4
WT = Woody, SWP = Not Present	0.6	0.6	0.6
WT = Herbaceous, SWP = Not Present	0.8	0.8	0.8

Table 8: Missingness probabilities for the SH, SM, and VRMV variables given WT and SWP.

Slope Parameter	Nonspatial Models AD		MI		CCA	
β_1 (Soil Modified)	-7.99	$(1.49)^+$	-8.02	$(2.15)^+$	-5.79	$(2.70)^+$
β_2 (Surface Water)	7.65	$(1.37)^+$	8.22	$(1.45)^+$	-2.67	(5.62)
β_3 (Log Nitrogen, Surface Water)	0.18	(0.81)	0.38	(0.83)	3.28	$(1.31)^+$
β_4 (Woody Wetland)	-6.14	$(1.33)^+$	-6.06	$(1.40)^+$	-10.30	$(2.80)^+$
β_5 (Soil Hardening)	-0.18	$(0.03)^+$	-0.15	$(0.06)^+$	-0.23	$(0.08)^+$
β_6 (Moderate Veg Removal Stress)	-2.46	(1.48)	-0.34	(2.18)	7.00	$(2.85)^+$
β_7 (High Veg Removal Stress)	-1.22	(3.81)	-0.43	(4.67)	-2.81	(7.38)
Slope Parameter	Spatial Models AD		MI		CCA	
β_1 (Soil Modified)	-5.73	$(1.29)^+$	-5.09	$(1.65)^+$	-5.08	(2.61)
β_2 (Surface Water)	2.98	$(1.23)^+$	3.29	$(1.28)^+$	-4.45	(5.30)
β_3 (Log Nitrogen, Surface Water)	-0.60	(0.70)	-0.70	(0.72)	2.48	(1.27)
β_4 (Woody Wetland)	-10.13	$(1.25)^+$	-10.16	$(1.28)^+$	-10.64	$(2.77)^+$
β_5 (Soil Hardening)	-0.08	$(0.03)^+$	-0.06	(0.04)	-0.15	(0.09)
β_6 (Moderate Veg Removal Stress)	-1.44	(1.25)	0.69	(1.69)	7.21	$(2.66)^+$
β_7 (High Veg Removal Stress)	1.22	(3.38)	1.43	(3.97)	-3.88	(7.23)

Table 9: Linear regression slope parameter estimates and standard errors $(\cdot)$ for the nonspatial and spatial models fit using all data (AD), multiple imputation (MI), and complete case analysis (CCA). The $+$ indicates the p-value for the corresponding parameter estimate is less than 0.05.

Model	Method	MPBias	RMSPE	PCover95	R2
Nonspatial	AD	0.00	20.21	0.95	0.13
Nonspatial	MI	0.00	20.16	0.96	0.13
Nonspatial	CCA	0.00	19.77	0.95	0.08
Spatial	AD	-0.01	16.51	0.94	0.42
Spatial	MI	-0.02	16.48	0.95	0.42
Spatial	CCA	-0.10	18.48	0.95	0.19

Table 10: LOOCV prediction performance for the nonspatial and spatial models fit using all data (AD), multiple imputation (MI), and complete case analysis (CCA). Metrics evaluated were mean bias (MBias), root-mean-squared-prediction error (RMSPE), 95% prediction interval coverage (PCover95), and predictive R-squared (R2).

Figure Captions

- Figure 1 Caption: MCAR (a), MAR (b), and MNAR (c) distributions for hypothetical VMMI data.

- Figure 2 Caption: A flowchart for specific missing data methods. Complete case analysis removes units with missing items and then implements an analysis approach. Deterministic imputation fills in missing values in a nonrandom way and then implements an analysis approach. Single imputation fills in missing values in a random way and then implements an analysis approach. Multiple imputation pools the results from m separate single imputation steps.

- Figure 3 Caption: Observed and imputed data comparison. The hypothetical variable $\mathbf{z}_1$ is fully observed and used to impute the hypothetical variable $\mathbf{z}_2$ using mean (Mean), regression (Reg), and bootstrap (Boot) imputation. Observed items are represented by blue circles and imputed items are represented by orange triangles. The green line is a linear regression slope characterizing the average effect of $\mathbf{z}_1$ on $\mathbf{z}_2$ using the observed data.

- Figure 4 Caption: The $\mathbf{x}_2$ imputation RMSE versus the β_2 parameter RMSE. Mean and regression imputation have the lowest $\mathbf{x}_2$ imputation RMSE but among the highest β_2 parameter RMSE. The third deterministic imputation method, NN, had high $\mathbf{x}_2$ imputation RMSE and high β_2 parameter RMSE. Remaining missing data methods were grouped into "Other" except CCA, which was omitted because CCA does not fill in missing values.

- Figure 5 Caption: The Predict-Combine and Combine-Predict approaches to prediction using multiple imputation. In Predict-Combine, models are fit and predictions made separately for each complete data set and then pooled according to Rubin's Rules to create a final prediction. In Combine-Predict, models are fit and pooled across complete data sets and the pooled model is used to create a final prediction.

- Figure 6 Caption: NWCA 2016 surface water presence and total nitrogen (TN) condition

categories. Wetlands with surface water presence are grouped into Good, Fair, or Poor condition for TN.

- Figure 7 Caption: In (A), the proportion of CONUS wetlands with surface water that are in Good, Fair, or Poor TN condition. In (B), proportion of CONUS wetlands with surface water. 95% confidence intervals for the estimates in (A) and (B) are represented by black bars.

- Figure 8 Caption: Spatial distribution of NWCA 2016 VMMI. VMMI scores range from 0 to 100. The larger the VMMI, the healthier the vegetation.

1549 **Figures**

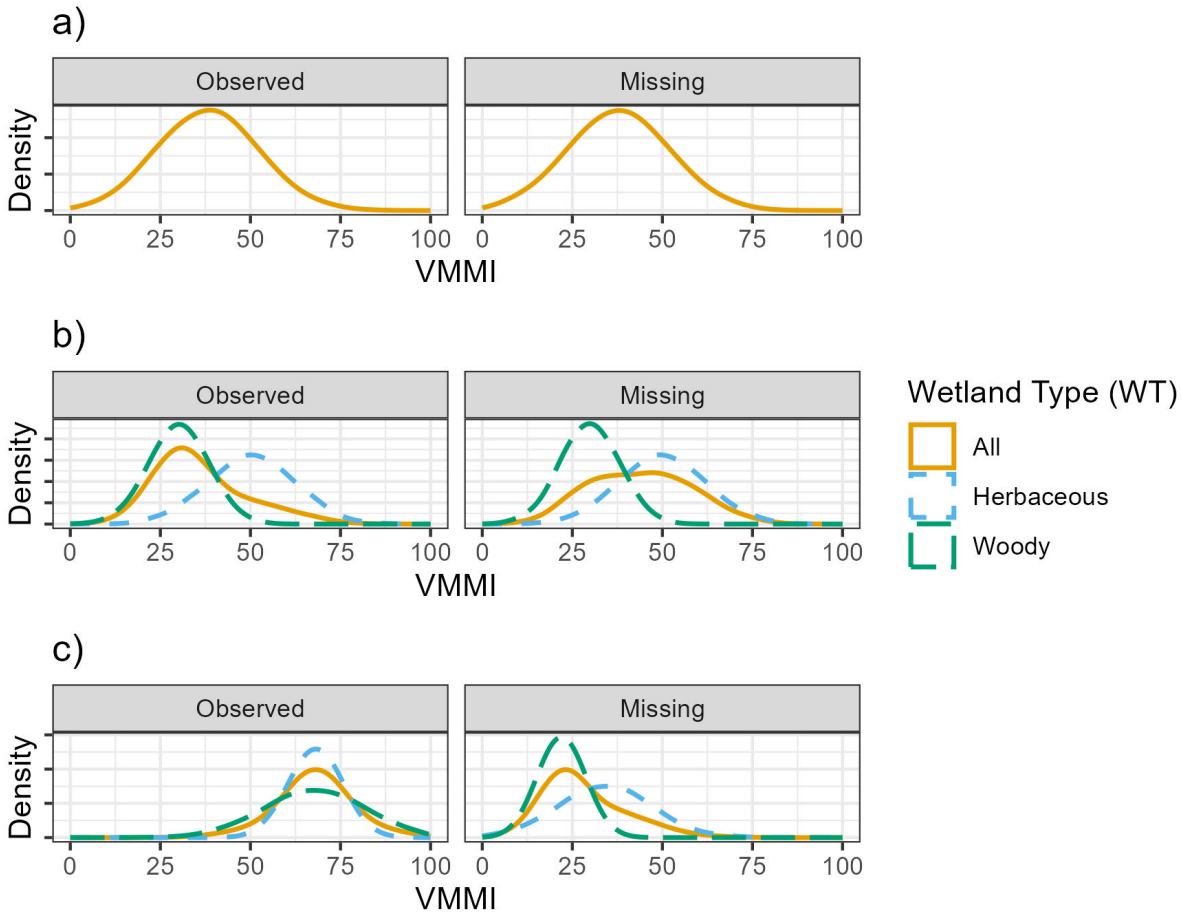


Figure 1

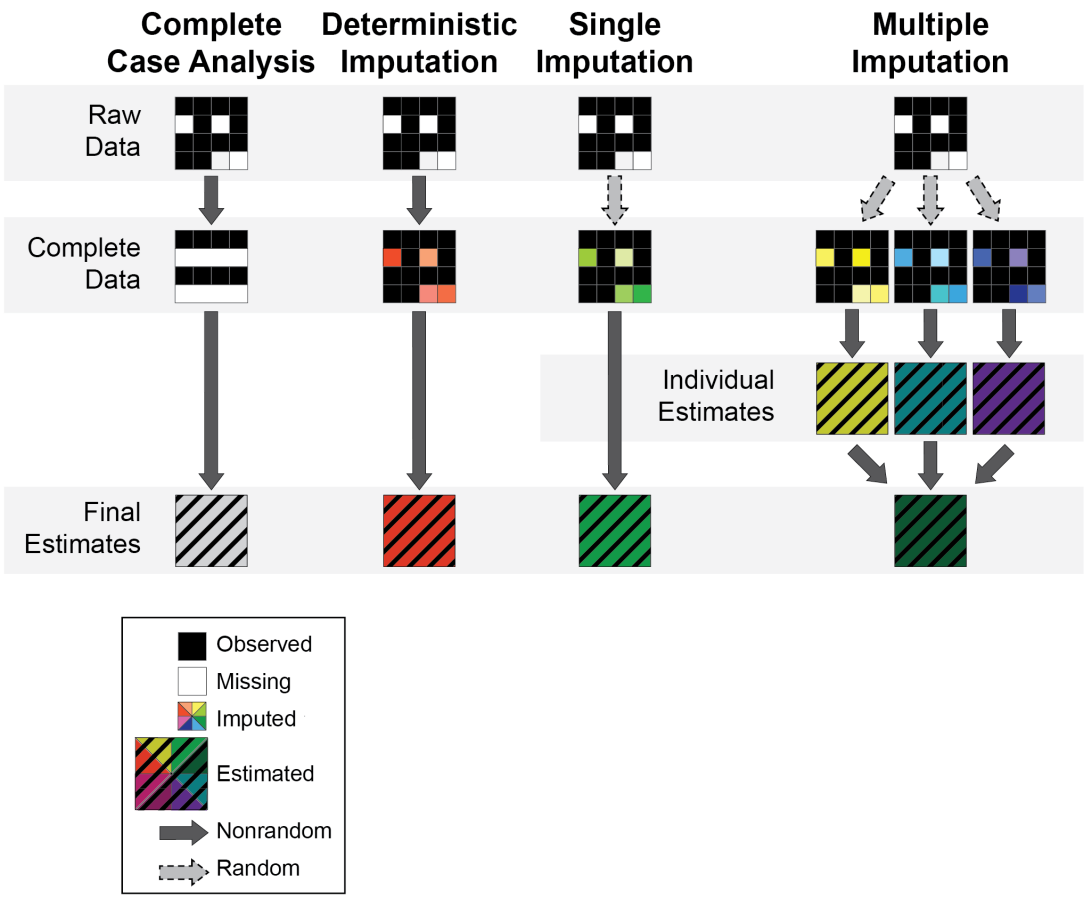

Complete Case Analysis
Deterministic Imputation
Single Imputation
Multiple Imputation
Raw Data
Complete Data
Individual Estimates
Final Estimates
Observed
Missing
Imputed
Estimated
Nonrandom
Random

Figure 2

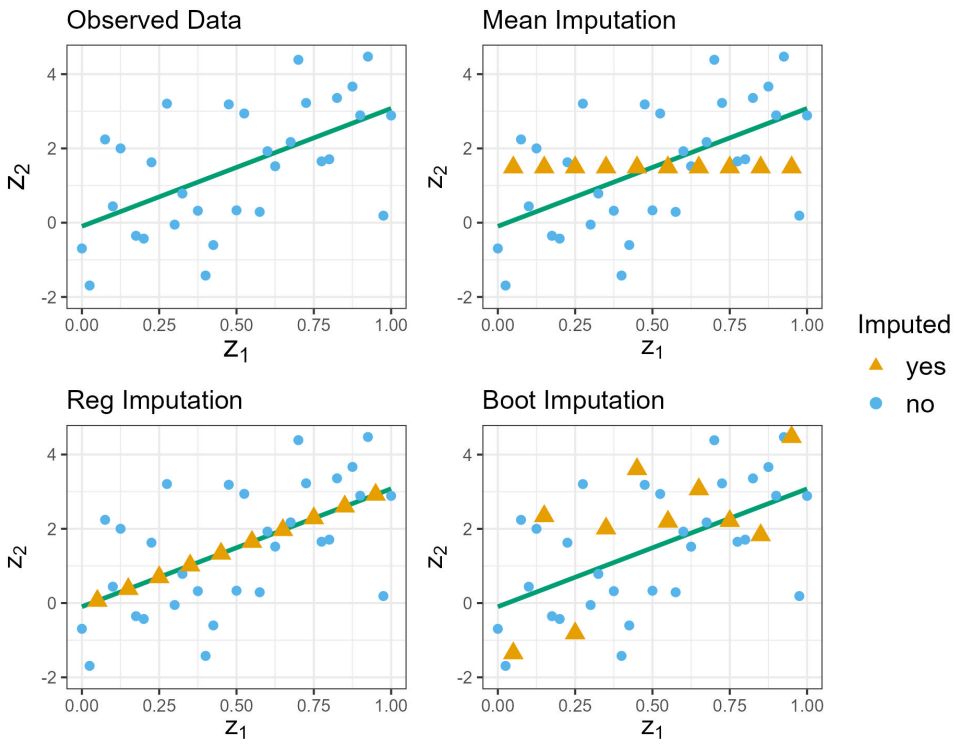

Observed Data
Mean Imputation
Reg Imputation
Boot Imputation
z_1
z_2
0.00
0.25
0.50
0.75
1.00
4
2
0
-2
Imputed
yes
no

Figure 3

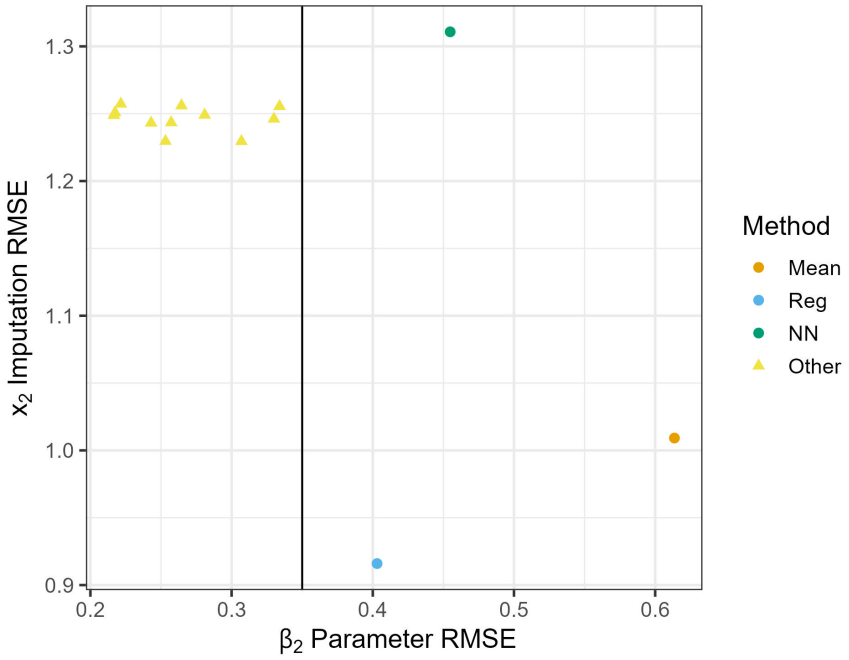

1.3
1.2
1.1
1.0
0.9
x_2 Imputation RMSE
0.2
0.3
0.4
0.5
0.6
β_2 Parameter RMSE
Method
Mean
Reg
NN
Other

Figure 4

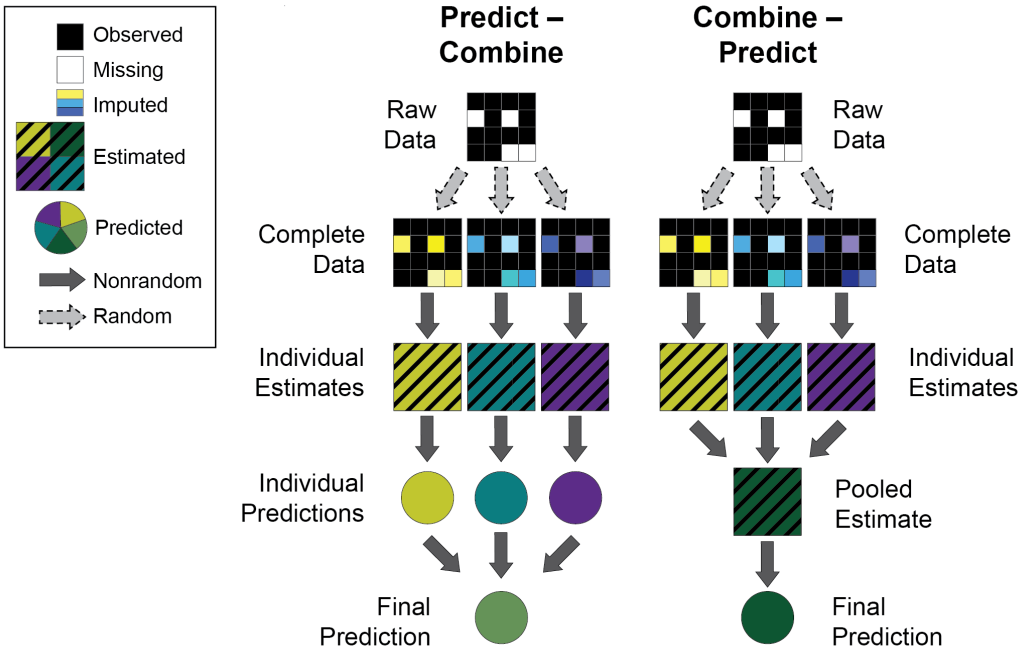

Observed
Missing
Imputed
Estimated
Predicted
Nonrandom
Random
Predict –
Combine
Combine –
Predict
Raw
Data
Raw
Data
Complete
Data
Complete
Data
Individual
Estimates
Individual
Estimates
Individual
Predictions
Pooled
Estimate
Final
Prediction
Final
Prediction

Figure 5

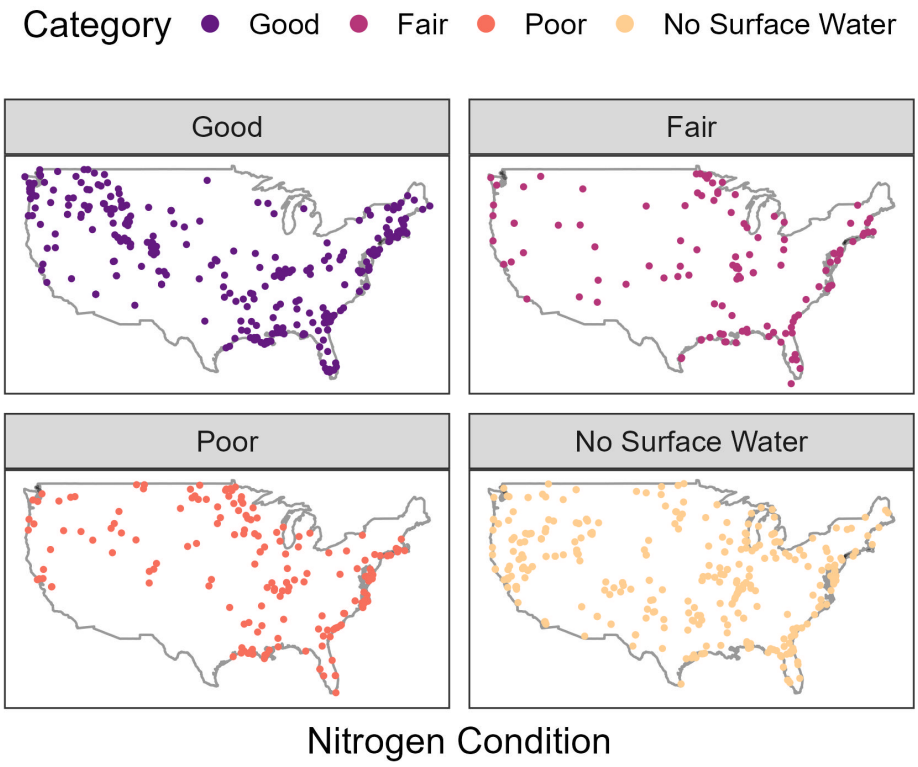
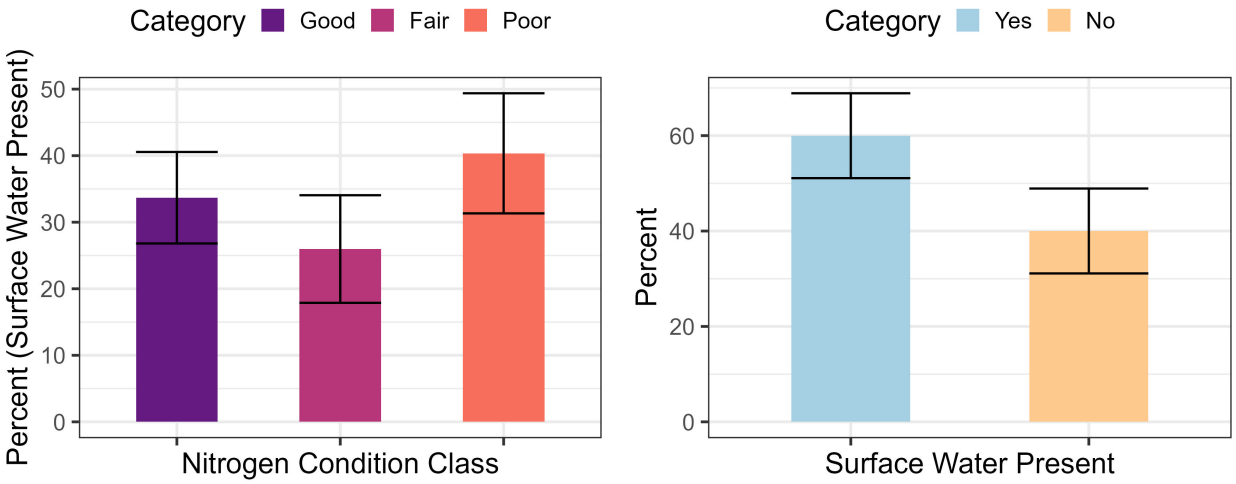

Category
Good
Fair
Poor
No Surface Water
Good
Fair
Poor
No Surface Water
Nitrogen Condition

Figure 6


Category
Good
Fair
Poor
Percent (Surface Water Present)
50
40
30
20
10
0
Nitrogen Condition Class
Category
Yes
No
Percent
60
40
20
0
Surface Water Present

(A) (B)

Figure 7

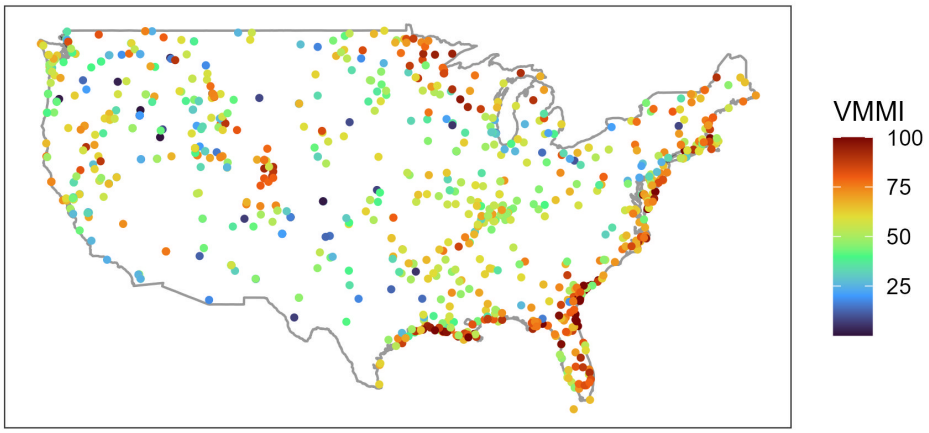


Figure 8