A Statistical Investigation of Species Distribution Models and Communication of Statistics Across Disciplines

by

Sara A. Stoudt

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor William Fithian, Co-chair
Professor Perry de Valpine, Co-chair
Professor Steven Beissinger
Assistant Professor Peng Ding

Summer 2020

A Statistical Investigation of Species Distribution Models and Communication of Statistics Across Disciplines

Abstract

A Statistical Investigation of Species Distribution Models and Communication of Statistics Across Disciplines

by

Sara A. Stoudt

Doctor of Philosophy in Statistics

University of California, Berkeley

Assistant Professor William Fithian, Co-chair

Professor Perry de Valpine, Co-chair

Ecologists commonly make strong parametric assumptions when formulating statistical models. Such assumptions have sparked repeated debates in the literature about statistical identifiability of species distribution and abundance models, among others. Part I of this dissertation draws upon the econometrics literature to introduce a broader view of the identifiability problem than has been taken in ecological debates. In particular we use a simulation approach to illustrate the concepts of non-parametric and parametric identifiability and their implications for ecologists. The fact that all models are approximations has very different implications for these two cases of identifiability. When non-parametric identifiability holds, even a mis-specified parametric model provides a useful approximation to the truth, and the fit of alternative models can be compared. When non-parametric identifiability does not hold, parametric assumptions create artificial identifiability, and alternative models cannot be distinguished empirically.

Joint species distribution models (JSDMs) have become a popular tool for helping ecologists understand properties of a community while accounting for relationships between species. Part II of this dissertation stress tests a foundational JSDM to understand how well properties of the community are estimated in the presence of model mis-specification. Community diversity metrics summarize community characteristics that ecologists have historically been interested in, so it is of interest to ask whether estimation of these more complicated metrics is robust to inevitable model mis-specification.

Being a statistician is a “hands-on” job that requires communicating with stakeholders and researchers in a variety of fields. Part III of this dissertation leverages the communication skills I have built while working at the intersection of ecology and statistics to teach statistics students how to write about statistical analyses in an accessible way that is still faithful to the

data. A pedagogical approach is described that builds upon that of traditional writing and science communication. This approach adds to the solid foundation with concrete examples in the context of statistics, particular focus on the nuances of statistical language, and a focus on narrative that carries throughout the data analysis process itself.

To my dad who encouraged me to “never change, never stop”

Contents

Acknowledgments

This dissertation was made possible through the support (and funding) of many. I would like to thank my advisors Perry de Valpine and Will Fithian for their patience, guidance, and help navigating the balance between statistical scholarship and applied impact. I would also like to thank my committee members Steve Beissinger and Peng Ding for providing helpful suggestions and food for thought throughout the process.

I appreciate funding from the National Physical Sciences Consortium fellowship, the Gordon and Betty Moore Foundation (through Grant GBMF3834), the Alfred P. Sloan Foundation (through Grant 2013-10-27), and the Art of Writing program for allowing me to pursue my studies.

To Deb Nolan, thank you for your mentorship, support, and advocacy. I would never have started identifying as a writer, or even a teacher, without our work together, and I'm grateful that I get to learn from you.

I am grateful to the interdisciplinary communities I was lucky to be a part of at BIDS and through the Environment and Society: Data Science for the 21st Century program. You all helped cultivate my curiosity and kept my eyes open to data problems in a broader range of topics.

I am also grateful that my mentors who helped me get to Berkeley continued to support me throughout my time here. Thank you to Ben Baumer, Nick Horton, and Antonio Possolo for your constant encouragement and for helping me remember my strengths when I was caught up in my weaknesses.

To Rebecca Barter, Kellie Ottoboni, and Ciera Martinez, thank you for being such strong and kind role models. I really would have been lost without your support.

To the staff of the statistics department and BIDS, thank you for making sure everything runs smoothly and for helping me navigate Berkeley as an institution and its resources.

To the statistics graduate students and the colleagues turned friends, I am glad that we are able to support one another in pursuing our dreams (and avoiding our nightmares), and I hope this continues beyond our time at Berkeley.

Finally, to my family and friends, thank you for always being there, enthusiastically celebrating my wins and tirelessly listening to recounts of my losses.

Chapter 1

Introduction

Ecology has traditionally been a “small-data” field in comparison to other disciplines, and relatedly as Borgman et al. put it, a “little science” [18]. Collecting data on plants, animals, and the broader environment often requires expensive and labor intensive field work, and it can be hard to gather data across a wide enough spatial, temporal, and taxonomic range to make conclusions beyond a specific study site or species. Large-scale historical data-collection efforts did exist though. Examples include expeditions by Joseph Grinnell and natural history museum collections [63, 101]; they provide specimen data dating back to the nineteenth and twentieth centuries. However, there was a historical barrier to specifying and fitting complicated models that account for the intricacies of this ecological data.

New data collection protocols and technologies have increased the speed and decreased the cost of data collection, helping ecology to transition towards a “medium-data” field and even a “big-data” field in some cases [51]. Examples of these technologies include camera traps [92], remote sensing [5], marine and terrestrial microphones [15, 84], genetic data [36], and phone applications that allow anyone to collect and contribute data [178, 123]. More data requires even more computational sophistication.

As the field transitions between regimes of data size and model complexity, researchers have to navigate a new set of problems including modeling data that spans multiple scales, contains interactions between species, and has spatio-temporal patterns [138, 62]. Therefore, methods that are computationally feasible and able to incorporate data from many sources and even of different types, like data fusion approaches [137, 132, 54], are needed.

Along with the methodological challenges that more data brings, ecologists have had to face data sharing, formatting, and management challenges as well [152, 115, 67]. Ecological data has been compiled in central databases such as the Global Biodiversity Information Facility (GBIF) [57], the Neotoma Paleoecology Database [121], the Paeleobiology Database [133], and the Forest Inventory and Analysis Database [55] and across large scale monitoring networks such as the Long Term Ecological Research (LTER) Network [102] and the National Ecological Observatory Network (NEON) [120].

This is the context in which my dissertation came together. My role became identifying common problems in the ecology literature, recognizing potential solutions that could be borrowed from the statistics literature, and acting as the go-between for the two fields.

The specific goals of this dissertation are to:

1. address identifiability problems that ecologists face when using statistical models for estimating species distributions and abundances,

2. investigate how statistical models perform under model mis-specification in the context of community ecology, and

3. teach others how to communicate statistics in an accessible yet accurate way to broader audiences, including collaborators in other disciplines.

In the rest of this introduction section, I will provide a broad overview of the motivation and more specific context for each chapter of this dissertation.

1.1 Clarifying Identifiability Controversies in Species Distribution and Abundance Modeling

Much of the increase in data availability has been due to citizen science efforts such as iNaturalist and eBird [81, 30] that allow amateurs and professionals alike to contribute data on wildlife sightings. For a sense of scale, there are over 16 million occurrences in the iNaturalist data on GBIF and over 550 million occurrences in the eBird data on GBIF. These occurrences span the globe and have been collected since 2008 and 2002 respectively.

However, this citizen science data comes with its own challenges including inexperienced observers, sampling bias (e.g. in space), and variation in sampling effort [37, 82]. These new sources of data, and their challenges, have inspired new methods for estimating ecological quantities of interest including correcting for opportunistic data [86], correcting for observer bias [190], and correcting for overall bias using data from multiple species [53].

The differing collection protocols of some citizen science efforts inspired the first part of this dissertation. For example, iNaturalist participants collect opportunistic (or presence-only) data while eBird participants collect data in the form of check-lists (yielding presence-absence data). Since ecologists often want to interpret parameters of a model in context (e.g estimate prevalence, average occurrence, or average abundance) rather than just make predictions about where a particular species might be spotted, identifiability of these parameters is an important property.

Informally, identifiability means that properties of a statistical model can be estimated from the data that is available, i.e. alternative models can be distinguished empirically. It is crucial to have identifiability if our goal is statistical inference, while it is possible to have strong predictive performance without identifiability. We wanted to know what is and is not identifiable using data collected under different protocols and when common modeling assumptions do not hold. This question led us to a variety of identifiability controversies in species distribution and abundance models. Recognizing these disparate back-and-forth controversies as different flavors of one problem and providing a unified clarification to them became the subject of the first chapter of the dissertation. Our work shows that there are different forms of identifiability (originally discussed by econometricians), some stronger than others, that can help us reason about a variety of identifiability controversies that have been debated in the literature.

1.2 Stress Testing Latent Factor Approaches to Joint Species Distribution Models

When data or appropriate computational approaches were hard to come by, ecologists often focused on single species distribution models for a species of interest, knowing that ignoring the distribution of other species will plausibly impact inference. For example, "stacked" species distribution models that combine information from single species distribution models without accounting for relationship between species often overestimate species richness [166]. The creation of joint species distribution models is a consequence of having more data available for many species within the same spatio-temporal range as well as progress in the ability to statistically model associations between species.

However, through working on the first chapter I gained an appreciation for the havoc that model mis-specification can cause for inference in species distribution and abundance models. The next step was to move to more complicated joint species distribution models, in the case where data quality was not an issue, to assess when they break down in the presence of inevitable model mis-specification.

Joint species distribution models are used to both get more realistic estimates of single species distributions and give ecologists insight into the relationships between species. Hierarchical Bayesian methods have become a popular way to account for the many sources of uncertainty and variation in ecological models [28, 32], although they have their practical challenges. These include the need to worry about technical details of the MCMC algorithms, philosophical differences between Frequentist and Bayesian methods, and a mismatch in ecologists' computational training and the tools available for fitting these types of models [16, 95, 186]. As hierarchies get complicated, these models can become black-box-like and hard to parse, making the goal of interpretation harder to achieve.

Joint species distribution models are fairly new, first occurring in the literature in the late 2000s and early 2010s [93, 129, 146, 60]. Performance of these models in service of some goals have been assessed (e.g. for prediction accuracy and global goodness of fit [201, 126]). We wanted to add to these comparisons to see if any misfit of joint species distribution models were hidden by a focus on traditional summaries of performance, especially in the presence of model mis-specification. Crucially, could characteristics of a community, such as measures of diversity, still be estimated well if the model is mis-specified? Stress testing joint species distribution models for the purpose of inference became the focus of the second chapter of the dissertation.

1.3 Communicating with Data

Learning how to read and write about data is increasingly important for everyone as a new form of literacy emerges, digital literacy [44]. The American Library Association's Digital Literacy Taskforce defines "digital literacy" as "the ability to use information and communication technologies to find, evaluate, create, and communicate information, requiring both cognitive and technical skills" [2].

Weighty issues such as climate change, systemic racism and sexism, and public health crises are often discussed using data and statistics. Being able to both parse that information as consumers and produce accessible material as data-intensive researchers are both important skills. Additionally, becoming critical consumers of numbers used in the media is especially necessary in an era where the term "fake news" is often used (and abused) [193].

It is clear that the next generation of statisticians and data scientists will increasingly need communication skills. However, students are rarely formally trained to write effectively while properly accounting for the subtleties of statistical language. As a way to remedy this, I became involved in the development of a course to teach undergraduate students how to write about statistics for their peers and the broader public. This effort expanded into a book to share that experience with more students and instructors eager to teach similar courses at their own institutions. Deb Nolan and my approach to teaching students about statistical writing, and the pedagogical framework behind it, became the focus of the third chapter of the dissertation.

This focus on communication was also important for my research in quantitative ecology. I was able to use the communication techniques I was teaching to explain my own work to a broader audience. Inspired by my work on identifiability in species distribution and abundance models, I wrote an article for Logic, a magazine about technology and society, about the benefits and limitations of relying on data collected by citizen scientists [175]. My identifiability project was also featured in the Berkeley Science Review [197], and my experience with citizen science data was relevant to the work I did at the Los Angeles Times during a data journalism internship.

Beyond making my own work more accessible, I also got involved in making the statistics field as a whole more approachable. While working at the interface of ecology and statistics, I found that there was plenty of interest in having core concepts from statistics demystified for ecologists. This demand was how I got involved in writing the Stats Corner section of the Ecology for the Masses blog, whose goal is to "make good ecological science accessible to people outside of science" [42]. In this work it was important to be able to explain methods in contexts that were relevant to ecologists' work. My experience navigating the boundaries between statistics and ecology made my involvement in this project a natural fit.

Part I

Identifiability in Species Distribution and Abundance Modeling

Chapter 2

Clarifying Identifiability Controversies

Co-authored with Perry de Valpine and Will Fithian

2.1 Introduction

There has been considerable debate in the species distribution and abundance literature about the inherent information content of data collected under different protocols [198, 66, 7]. Examples include estimating overall prevalence with presence-only v. presence-absence data, estimating occupancy or abundance with single-visit v. multiple-visit data and heterogeneous detection probabilities across sites, and estimating abundance with capture-recapture data and heterogeneous detection probabilities of individuals.

In all of these scenarios, some authors present a model where the quantities of interest are identifiable while others present a model where they are not. The debate then becomes about whose model is more realistic: the model that seems to be identifiable despite our intuition that we are not collecting enough information v. the counter-example that seems contrived but exhibits a lack of identifiability.

With single-visit occupancy data, Lele et al. proposed a method to estimate occurrence even with imperfect detection [97]. By modeling detection and occurrence probabilities using logistic regression and sets of covariates that differ by at least one covariate, parameters for occurrence and detection probabilities can be estimated separately despite each site being visited only once. Solymos et al. proposed a similar approach for disentangling abundance from detection with single-visit data [171]. They modeled detection probabilities with logistic regression and abundance with Poisson regression, using sets of covariates that again differ by at least one covariate. However, these approaches were critiqued by Knape and Korner-Nievergelt because the ability to disentangle detection from occurrence or abundance depends on the choice of link function [88, 89]. They give a counter-example where the true occurrence and detection probabilities come from scaled logistic functions of covariates, one scaled by α and the other by $1/\alpha$. The range of observed values of covariates do not produce probabilities greater than one. Solymos and Lele rebutted, provided a variety of commonly used link functions for which identifiability holds, and stated that the modeling choices made in the counter-examples were unrealistic [170].

A similar controversy occurs in the debate over estimating prevalence using presence-only, rather than presence-absence, data. Royle et al. proposed a model that appears to estimate prevalence with presence-only data [161]. Although this violates a natural intuition that the number of sightings alone, without data on absences, fails to inform the overall proportion of occupied sites, Ward et al. proved identifiability of prevalence for the model [188]. However, they showed that identifiability is not guaranteed in a broader model. Hastie and Fithian provided a counter-example that reveals this lack of broader identifiability [69].

Competing models are also present in the capture-recapture literature. When there is individual heterogeneity in detection probabilities beyond what is generated by covariates, identifiability of the overall abundance is controversial. Link refined the conclusions of Huggins and showed that abundance cannot be identified without restrictions on the detection probability distribution [99, 75]. Holzmann et al. showed that if we assume a distribution family for the detection probabilities, we can identify its parameters and hence abundance [73]. However, Link pointed out that choosing between families with similar model fit can lead to different abundance estimates [100].

In this paper we change the focus of the conversation from a debate about the realism of particular parametric models to one about the consequences of parametric assumptions. To do this we present a unifying framework, rooted in ideas from the econometrics literature, in which these individual controversies are special cases. The econometrics literature differentiates between identifiability in parametric and non-parametric models [91, 160, 156]. We introduce the concept of non-parametric identifiability, a strong form of identifiability that exists when a model could be well approximated without parametric assumptions – the data are informative on their own. Note that a non-parametric data-generating process, e.g. any continuous function, cannot technically be perfectly estimated by a finite sample from the data-generating process, but as Koopmans and Reiersol state: "identification problems are not problems of statistical inference in a strict sense, since the study of identifiability proceeds from a hypothetical exact knowledge of the probability distribution of observed variables rather than from a finite sample of observations. However, it is clear that the study of identifiability is undertaken in order to explore the limitations of statistical inference" [91].

We include discussion of an intermediate form of identifiability, partial identifiability, where at least a range of plausible results can be determined even in the absence of non-parametric identifiability [109]. Finally, we give a procedure to explore non-parametric identifiability by approximating a non-parametric model as a flexible unpenalized spline to which simulated data can be fit. This procedure complements use of specific counter-examples and thus offers a new avenue for discovering identifiability problems and their consequences.

As the adage says, "all models are wrong, but some are useful" [19]. In this paper we illustrate a framework to help determine when models are useful. We argue that one must establish identifiability within a super-model, a model that is sufficiently general to encompass any plausible analysis model, even if such a general model will not actually be estimated from data in practice. For example, if an analysis model (or sub-model) of interest uses a logistic curve to relate x and y, as in logistic regression, a relevant super-model would include all continuous curves of any shape relating x and y.

When identifiability holds in a more general model, then an approximating sub-model—what is commonly used in practice—enjoys a kind of robustness to model mis-specification. The approximating sub-model is "wrong but useful," and its parameters carry reasonable scientific interpretations. When identifiability does not hold in a more general model, then what appears to be an approximating sub-model can have parameter estimates that are artifacts of the choice of sub-model and are unrelated to the underlying data-generating process. Existence of a sub-model that lacks identifiability (i.e. a counter-example) proves that finding a suitable super-model is impossible.

The rest of this chapter is laid out as follows: the "Materials and Methods" section provides our working definitions of different forms of identifiability and outlines our simulation method and the ecological examples for which we assess identifiability. The "Results" section describes the implications of model mis-specification in different regimes of identifiability that are revealed by the simulation study. The "Discussion" section connects the ideas in this paper to other identifiability debates more broadly and provides some guidance for future work.

2.2 Materials and Methods

Identifiability and Its Different Forms

Informally, *properties* of a statistical model are called *identifiable* when they can be estimated from the available data. These properties answer a question of interest and can be any summary of the *data-generating process*, a statistical concept for the ecological and measurement processes that produced the data. Ecologically relevant properties include the average occurrence probability of a species and particular model parameters relating occurrence to covariates describing a landscape.

A *model* is a set of candidate data-generating processes, each of which defines a joint distribution of observed data D and unobserved, or latent, data E (following the notation of Cressie et al. [32]). As an example, with detection/non-detection data, E could denote the true occurrence or non-occurrence of a species at each site, while D denotes detection/non-detection data from all visits to all sites. In many ecological applications, the unobserved data represents an underlying natural process (the "ecology"), so properties of interest typically concern the distribution of the unobserved data. In contrast, if prediction of the observed data is our main goal, we may care only about the *observable distribution*, the distribution of observed data D.

We call two data-generating processes within a model *observationally equivalent* if they result in identical observable distributions, though they may differ in their distributions for the latent data E. A simple example is single-visit detection/non-detection data, without covariates, where either 50% of sites are occupied and a species is detected with 100%

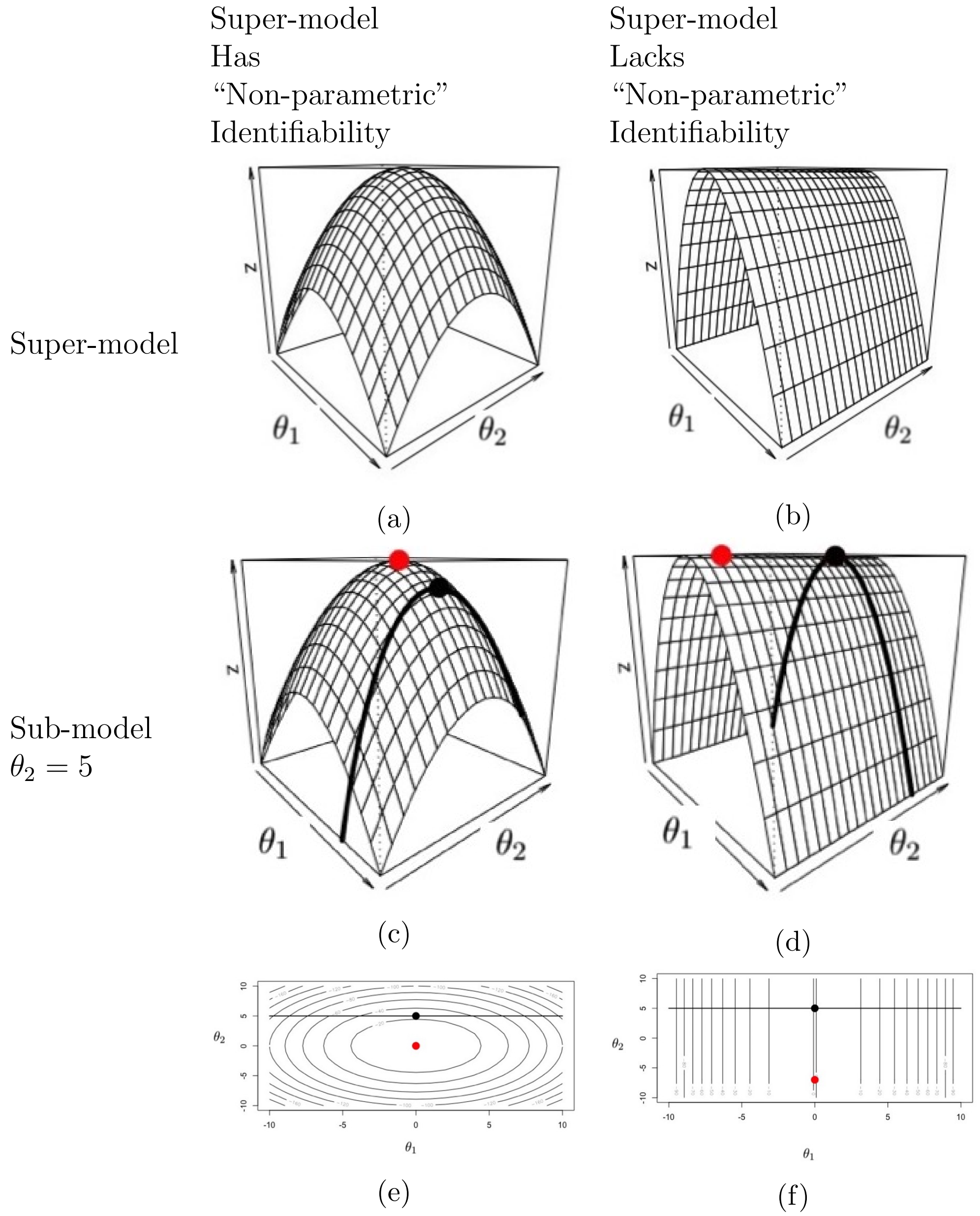
probability or 100% of sites are occupied and a species is detected with 50% probability. These two data-generating processes lead to the same observable distribution for one visit per site, no matter how many sites we visit.

A property of a model is formally *identifiable* if no two data-generating processes within the model are observationally equivalent but imply different values of the property. For example, in simple linear regression where the properties of interest are the parameters (intercept, slopes, and residual variance), there are no sets of different parameters that define the same distribution of data, so the parameters are identifiable. In a parametric model like linear regression, the type of identifiability is *parametric identifiability.*

If we broaden the model to a non-parametric model by relaxing assumptions, then the type of identifiability is *non-parametric identifiability.* A non-parametric extension of linear regression would allow the relationship between a response y and covariate x to be any continuous function. The properties of interest would be the shape of the function and the residual variance. A property is *partially identifiable* if some, but not all, observationally-equivalent data-generating processes can be eliminated from consideration.

We will refer to a non-parametric model as a *super-model* and a parametric model as a *sub-model.* For example, linear regression is a sub-model of non-parametric regression. Imposing parametric assumptions corresponds to choosing a sub-model from within a super-model. Identifiability in a super-model implies identifiability in sub-models that it contains. The converse is not true; sub-model identifiability does not imply super-model identifiability. Therefore, non-parametric identifiability is stronger than parametric identifiability. These identifiability concepts are illustrated next.

Illustration of Parametric and Non-parametric Identifiability The top row of Fig. 2.1 illustrates super-models with (Fig. 2.1a) and without identifiability (Fig. 2.1b). The second row depicts the corresponding sub-models along the black curves. The θ_1 and θ_2 axes represent parameters, or more generally, properties of the model. The z axis shows the log-likelihood of parameters given a very large data set. Data generating-processes make up the points on the surface.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2.1: Illustration of Identifiability Scenarios: Each sub-plot shows a hypothetical log-likelihood surface for a very large sample size as a function of parameters θ_1 and θ_2. For illustration, estimating parameters in two-dimensions (top row) is analogous to estimating a "non-parametric" super-model. The red points are the true data-generating processes for each scenario. The black points in the middle row are the estimates made by each sub-model (black curves) given very large sample sizes. The bottom row shows the contour plots of the middle row with the horizontal lines representing the sub-models.

Our choice of model defines how many parameter dimensions (axes) can be estimated. In real problems, a super-model might be infinite-dimensional, such as the space of all continuous curves of covariates, while a sub-model might be finite-dimensional, such as the space of all logistic curves from linear predictors. More generally, a super-model is sufficiently general to encompass any plausible data-generating process and is free of restrictive assumptions. For simple visualization purposes, estimating parameters in two-dimensions (top row) is analogous to estimating a super-model. Estimating one parameter in a single dimension is analogous to estimating a sub-model (bottom row).

In Fig. 2.1a there is a single maximum, corresponding to the super-model maximum likelihood estimate. Therefore the super-model parameters are identifiable and there is an unambiguous "best" data-generating process for explaining the observable distribution. In Fig. 2.1b there is a ridge in the log-likelihood surface, so the super-model parameters are not identifiable. Different combinations of the parameters are observationally equivalent, yielding the same likelihood of the data no matter how large a sample we collect.

Within the chosen sub-models, the black dots in Figs 2.1c and 2.1d show the best approximation to the true data-generating processes (the red dots). Since the super-model in Fig. 2.1a is identifiable, the sub-model in Fig. 2.1c is also identifiable, and the best sub-model parameters are as close as possible to the best super-model parameters. Since the super-model is not identifiable in Fig. 2.1b, the identifiability of the sub-model in Fig. 2.1d is artificial. The notion of "best" approximation is not what we expect in this case. The "distance" between the estimate and the truth depends arbitrarily on the choice of sub-model. Here the model is wrong and not useful.

In the Results section we further illustrate, within ecological scenarios, the concepts of super-models and sub-models in cases with and without non-parametric identifiability. Typical identifiability debates involve mathematical insight to show a ridge in a likelihood surface such as in Fig. 2.1b. Below, we instead approximate the whole non-parametric surface using flexible models and large simulated data sets.

Simulation Method for Diagnosing Lack of Non-parametric Identifiability

The concept of non-parametric identifiability helps resolve identifiability debates that have arisen among ecologists. To see this we need to illustrate parametric v. non-parametric identifiability in real ecological scenarios and discover the consequences of estimating a property of interest in each, especially when models are mis-specified. Next we give a method for doing so with simulation studies, covering examples from the presence-only v. presence-absence and single-visit v. double-visit controversies. Similar discussion of single-visit v. double-visit abundance and capture-recapture examples are provided in the Appendix. Our simulation method's key step is to create super-models as very flexible models that can be interpreted as "almost nonparametric."

Crafting Sub- and Super-Models

For each example, we use simulations to compare scenarios (model and sampling protocol) with and without non-parametric identifiability (bottom and top rows of each result, respectively). For each scenario we craft a sub-model and super-model, fitting them to many simulated data sets.

Since "all models are wrong" the sub-models are designed to exclude the true data-generating process, i.e. the simulation model. To fit the sub-models to the simulated data, we use established analysis methods for each kind of data, described below in more detail. In all cases, the sub-models are mis-specified yet parametrically identifiable.

For super-models, we use unregularized cubic splines with seven knots per function of a covariate. Although splines are really parametric models, this choice represents a flexible parametric model (with eleven parameters per function of a covariate) and hence approximates a fully non-parametric model. In principle, one could use an arbitrarily flexible spline model to more closely approximate a fully non-parametric model containing any continuous function of the covariate. However, in our examples, even a seven knot spline is sufficient to illustrate lack of non-parametric identifiability. We include the correct covariate for each response distribution so that issues of identifiability focus on the continuous relationships between variables instead of the variables themselves.

The simulation results in each example address several specific questions. First, is the sub-model a useful approximation to the truth? Second, is the super-model identifiable? Identifiability of the super-model is approximately the same as non-parametric identifiability. Third, how is lack of super-model identifiability reflected in the sub-model results?

Column 1 of each results figure shows whether the mis-specified sub-model is a useful approximation to the truth and how the lack of super-model identifiability impacts the sub-model results. To determine super-model identifiability, we increase sample size from moderate (n=100) to large (n=1,000) to enormous (n=100,000) (columns 2-4 in each results figure). If an enormous (approximately infinite) amount of data does not uniquely identify a single set of best-fit parameters, the super-model is not identifiable, and the scenario lacks non-parametric identifiability. In such a case, the data fundamentally lack information about the question(s) of interest. Even when one will not have very large sample sizes, and even when only a parametric sub-model will be estimated, this exercise illustrates when the data can inform the question(s) of interest.

Ecological Scenarios

Next we illustrate the presence-only v. presence-absence and single-visit v. double-visit occurrence controversies. Corresponding results on the single-visit v. double-visit abundance and capture-recapture controversies are in the Appendix.

Presence-Only v. Presence-Absence Data In this example, y_i indicates presence (1) or absence (0) of the species of interest at site i, and x_i is a covariate from site i. We assume

S sites were visited and the species was present at S_1 of them. We also assume that we have perfect detection, that we know the distribution $\pi(x)$ of the covariate (e.g., we have access to the values of explanatory variables across all sites), and we do not have geographic sampling bias. In the presence-only scenario, data are the S_1 pairs $(x_i, y_i = 1)$ representing the sites where we observe a species. In the presence-absence scenario, data are all S pairs (x_i, y_i). The property of interest is the overall prevalence, $P(y_i = 1)$, which is intuitively difficult to estimate in the presence-only scenario.

In the analysis models for both scenarios, we assume probability of occurrence is a function of the covariate x_i and some parameters β: $P(y_i = 1|x_i; \beta) = \psi(x_i, \beta)$. In the sub-model, this takes a logistic form with two parameters: $\text{logit}(\psi(x_i, \beta)) = \beta_0 + \beta_1 x_i$. In the super-model, $\text{logit}(\psi(x_i, \beta))$ is the flexible unpenalized spline. In both scenarios, we assume $y_i \sim \text{Bernoulli}(\psi(x_i, \beta))$, independently.

In the presence-absence scenario, the sub-model and super-model represent logistic regression with few or many parameters, respectively. In the presence-only scenario, we apply the method of Royle et al. [161]. The probability that x_i appears in the data is given by Bayes Law as $P(x_i|y_i = 1; \beta) = \frac{P(y_i=1|x_i;\beta)\pi(x_i)}{\sum_{i=1}^{S} P(Y=1|x_i;\beta)\pi(x_i)}$. The likelihood is then $\prod_{i=1}^{S_1} P(x_i|y_i = 1; \beta)$ and we estimate β by maximum likelihood. One can think of x_i as "pixel identity," as in Royle et al., or as the covariate value(s) at pixel (site) i, which is the perspective taken here [161].

Intuitively, presence-only data does not yield information about the sampling effort, so we do not know if the number of sites with detections is small or large with respect to the total effort [140]. Thus there is a debate about whether identifiability of prevalence in Royle et al.'s method is an arbitrary outcome of parametric assumptions. Our example places this debate in the more general context of non-parametric identifiability, illustrating that any counter-example – any case where parametric identifiability fails – reveals lack of non-parametric identifiability and therefore lack of information in the data.

The simulation model used for this example is as follows. The number of sites visited is random, $S \sim \text{Pois}(n)$ where n is the sample size level, varying across columns 2-4 in the results figures. The presence-absence scenario naturally has *more* data than the presence-only case. Therefore we double the sample size for the latter case to ensure that conclusions are made about the *quality* of the data rather than the *quantity*. Covariate values are independent and uniformly distributed, $x_i \sim \text{Unif}(-2.5, 8)$. Observations are drawn from Hastie and Fithian's scaled logistic, $\text{logit}(\psi(x_i, \beta)/\alpha) = \beta_0 + \beta_1 x_i$, with $\alpha = 0.5$, $\beta_0 = -1$, and $\beta_1 = 1$ [69]. The scaled logistic has the familiar sigmoidal shape of a logistic curve but asymptotes at a maximum for $\psi(x_i, \beta)$ equal to α. This means that the sub-model for both the presence-only and presence-absence scenarios is mis-specified. The super-model is also formally mis-specified, but it is so flexible that it should be able to closely approximate the scaled logistic.

Single-visit v. Double-visit Occupancy Data In this example, y_{ij} is detection (1) or non-detection (0) of the study species on visit j to site i. There are S sites and either one (single-visit) or two (double-visit) visits to each site. Again we give the single-visit case twice as many sites to make the quantity of data equal between the two protocols. Each site has a covariate x_i related to occurrence and z_i related to detection. E_i is the unobserved indicator for true occurrence (1) or non-occurrence (0) at site i. The properties of interest are the relationships between occurrence probability and x and between detection probability and z.

In the analysis models for both scenarios, we assume the occurrence probabilities are a function of x_i and some parameters β, $P(E_i = 1|x_i;\beta) = \psi(x_i, \beta)$ and the detection probabilities are a function of z_i and some parameters θ, $P(y_i = 1|E_i = 1, z_i; \theta) = p(z_i, \theta)$. In the sub-model, each takes a logistic form with two parameters: $\text{logit}(\psi(x_i, \beta)) = \beta_0 + \beta_1 x_i$ and $\text{logit}(p(z_i, \theta)) = \theta_0 + \theta_1 x_i$. In the super-model, each of $\text{logit}(\psi(x_i, \beta))$ and $\text{logit}(p(z_i, \theta))$ is a flexible unpenalized spline. In both examples, we assume $E_i \sim \text{Bernoulli}(\psi(x_i, \beta))$ and $y_{ij} \sim \text{Bernoulli}(E_i p(z_i, \beta))$, with all outcomes independent.

In the double-visit scenario, the sub-model and super-model represent standard occupancy models. In the single-visit scenario, we use the un-penalized version of Lele et al.'s method, which relies on the assumption that there is at least one distinct covariate for each of occurrence and detection and a logistic form for the sub-model [97]. The identifiability of this approach has been debated. Knape and Korner-Nievergelt's counterexample to the scenario's identifiability involves a scaling parameter α such that different values of α yield the same distribution of all y_{ij} values (the *observable distribution* in this case, with $j = 1$) but different occurrence and detection probabilities [88]. Our example uses a different counter-example to illustrate that the scaled logistic is not just a uniquely troublesome "corner case."

The simulation model used for this example assumes linear relationships for both occurrence and detection probabilities, with parameters and data ranges such that probabilities fall between zero and one. Values for each of x_i and z_i are independent and uniformly distributed as above, $x_i \sim \text{Unif}(-2.5, 8)$ and $z_i \sim \text{Unif}(-2.5, 8)$. Occurrence and detection probabilities are $\psi(x_i, \beta) = 0.071x_i + 0.18$ and $p(z_i, \beta) = 0.048z_i + 0.12$, respectively. These values of β and θ imply that the occurrence and detection probabilities do not reach one within the range of the observed covariates. As intended, this means that the sub-model for both the single-visit and double-visit data is mis-specified. Again, the super-model is also formally mis-specified but is so flexible that it should be able to approximate the truth closely.

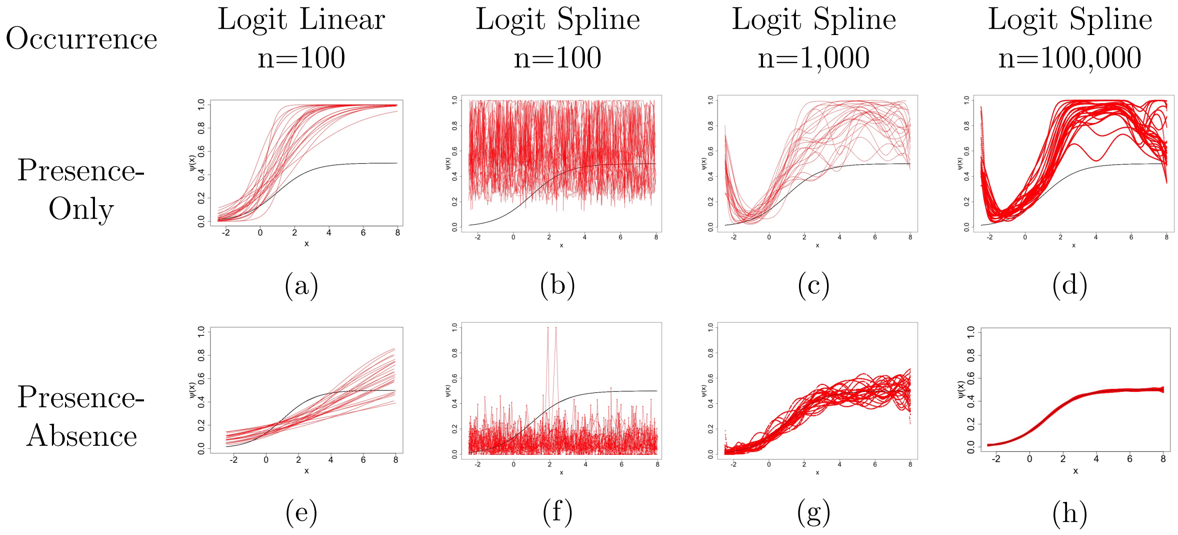
Single-visit cases were estimated with the R package detect [172]. Double-visit cases were estimated with the R package unmarked [52].

2.3 Results

Presence-Only v. Presence-Absence Data

In the first column of Fig. 2.2, we see that the (mis-specified) sub-model provides a useful approximation to the truth when estimated from presence-absence data (Fig. 2.2e) but not from presence-only data (Fig. 2.2a). In each sub-figure, the true relationship between x and $\psi(x)$ is in black and the estimated relationships from many simulated data sets are in red. The use of multiple data sets can be thought of like a bootstrap procedure where the width of the region spanned by the red curves is analogous to a confidence interval for any red curve. With presence-absence data, even though the approximation is rough because the red curves do not match the black curve, the salient point is that they do approximate the black curve as well as possible. The implied estimated prevalence (driven by the height of the estimated curve) is in a reasonable range. With presence-only data, the model is identifiable but the red curves do not even approximate the black curve. The implied prevalence is much larger than the true prevalence. Typical debates about identifiability have focused on this type of comparison between special cases, resulting in debate about whether the black curve is a reasonable special case to worry about.

The remaining columns (2-4) illustrate that the presence-absence scenario enjoys non-parametric identifiability, while the presence-only scenario does not. For the small sample size, neither scenario does well; there are too many parameters to fit. However, in the presence-absence scenario, the data are fundamentally informative because, as sample size increases (Figs 2.2f-2.2h), estimates of the flexible super-model converge to the true model. Because the estimates using different simulated data sets converge narrowly around the truth, we can feel confident that estimates are precise as well as accurate. In the presence-only scenario (Figs 2.2b-2.2d), we see the opposite: as sample size increases, the *shape* of the curve is similar to the truth but the *height* of the curve remains undetermined by the data. Even with enormous sample size, the estimated super-model implies a prevalence often twice as large as the true prevalence. There is an appearance that increasing sample size does improve estimates albeit very slowly; this is likely in part due to identifiability of the shape and in part because the flexible spline model is not fully non-parametric. An inability for a flexible model to approximate the truth, even when given more and more data, reflects lack of nonparametric identifiability.

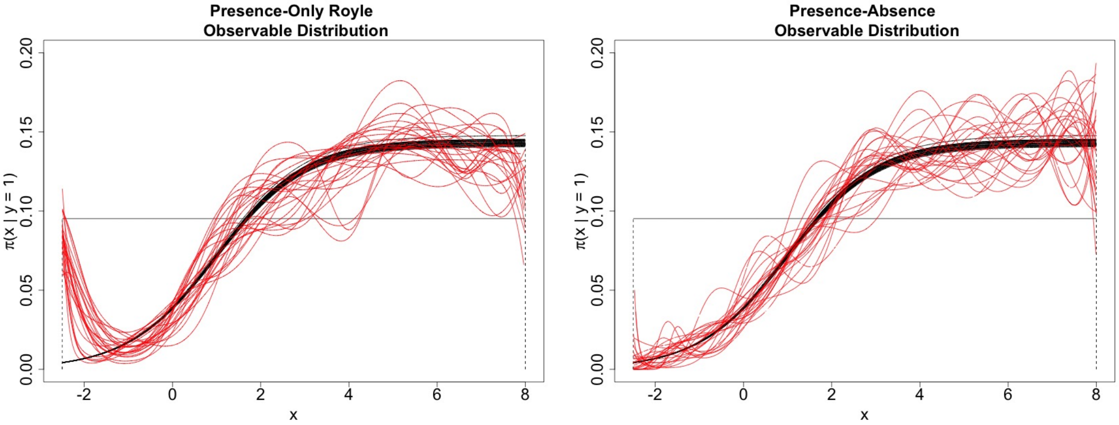


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2.2: The x-axis displays the value of the covariate that predicts occurrence. The y-axis displays the occurrence probability. Black curves show the truth while red curves show estimates from various simulations. The first column shows the fit using a parametric sub-model while the remaining columns show the fit using a more flexible super-model. The first row illustrates the implications of model mis-specification when prevalence is parametrically identifiable but not non-parametrically identifiable. The bottom row shows that when prevalence is also non-parametrically identifiable, the mis-specified parametric sub-model now gives a useful approximation; the flexible super-model reveals that the data can inform the parameter of interest.

Insufficiency of Approximating the Observable Distribution Even though they can't identify prevalence, presence-only data can nevertheless approximate $P(x|y_i = 1)$, the observable distribution in this case, well. Fig. 2.3 shows estimates of $P(x|y_i = 1)$ for the simulations with $n = 1,000$, i.e. from the third column of Fig. 2.2. The estimates fit well for both presence-only (Fig. 2.3a) and presence-absence (Fig. 2.3b) scenarios, indicating that the identifiability problems occur in this example only when the property of interest is prevalence. There is some mis-behavior in the presence-only case on the left boundary, but this could be mis-attributed to spline edge effects rather than an identifiability problem.



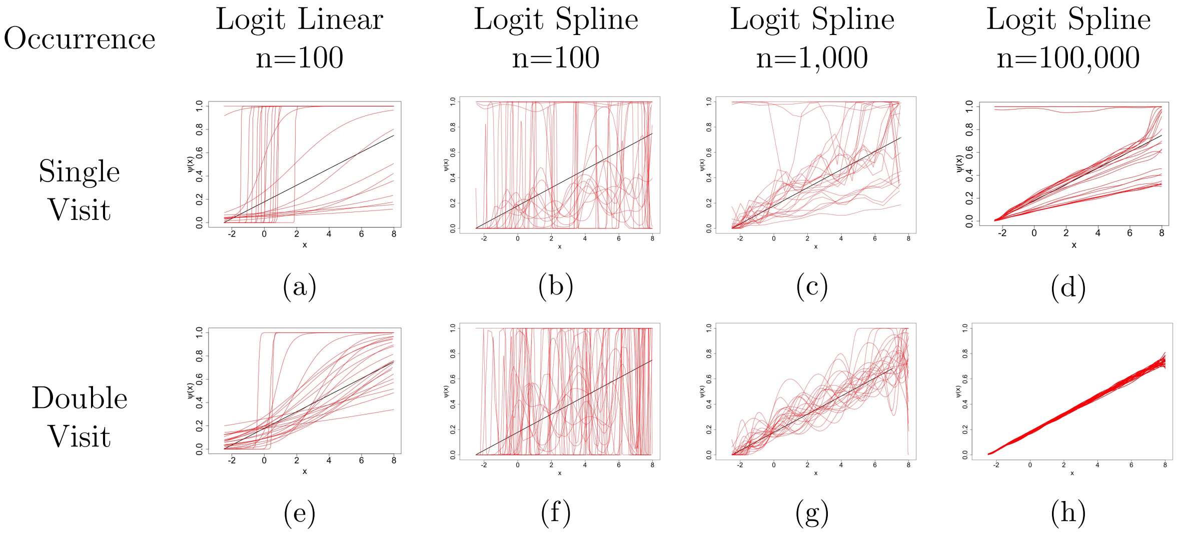
(a) (b)

Figure 2.3: The x-axis displays the value of the covariate that predicts occurrence. The y-axis displays the distribution of the covariate given that the species occurs. Black curves show the truth while red curves show estimates from various simulations. These plots compare the observable distribution using presence-only v. presence-abasence data in the non-parametric super-model case where $n = 1,000$. With either presence-only or presence-absence data, the observable distribution of the covariate given a presence is identifiable.

Single-Visit v. Double-Visit Occupancy Data

The first columns of Figs 2.4 and 2.5 show that the sub-model provides a more useful approximation to the truth when estimated from double-visit data (Figs 2.4e and 2.5e) than from single-visit data (Figs 2.4a and 2.5a). With double-visit data, estimated curves cluster more around and typically cross the true linear relationships. With single-visit data, estimated curves are more variable and often nearly completely miss the truth.

Results from the super-model with increasing sample sizes show that non-parametric identifiability holds for double-visit data but not for single-visit data (columns 2-4 of Figs 2.4 and 2.5). With moderate sample sizes (column 2) neither scenario provides good estimates for the super-model. However, for large and enormous sample sizes, double-visit data closely approximate the truth while single-visit data approximate only the shape but not the height of truth. Using double-visit data, estimates for different simulated data sets converge more and more narrowly around the truth as sample sizes increase. In contrast, even with large sample sizes single-visit data give estimates of occurrence probabilities reaching 100% even though the truth only reaches about 80%, and estimates of detection probabilities reach 100% even though the true maximum true detection probability is about 50%. Just as we saw in Section 2.3, the observable distribution of the product of the occurrence and detection probabilities can be matched well without identifying the individual distributions of occurrence and detection.



Occurrence

Logit Linear n=100

Logit Spline n=100

Logit Spline n=1,000

Logit Spline n=100,000

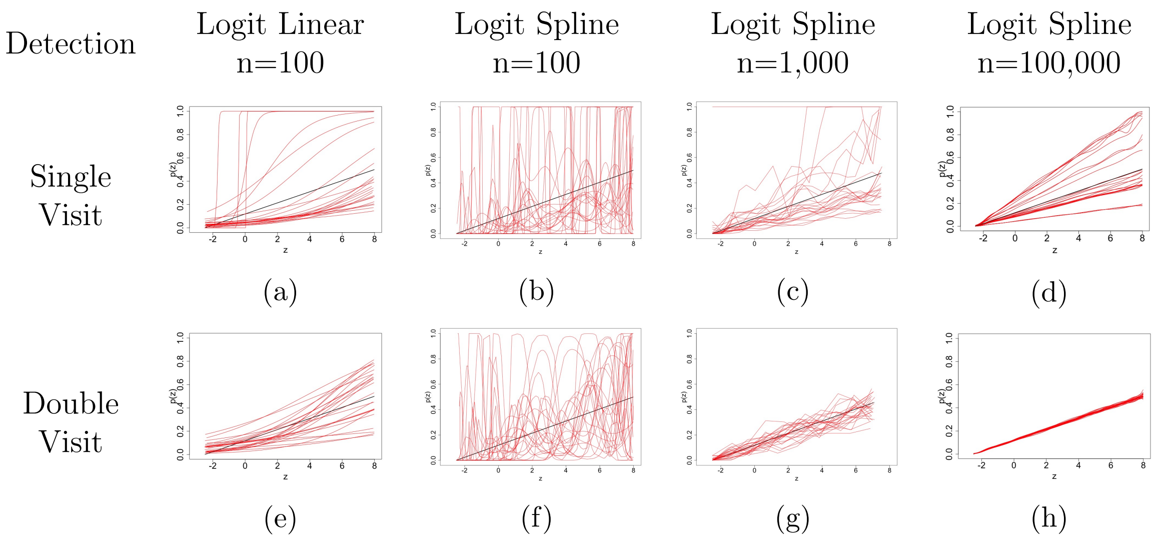
Single Visit

(a) (b) (c) (d)

Double Visit

(e) (f) (g) (h)

Figure 2.4: The x-axis displays the value of the covariate that predicts occurrence. The y-axis displays the probability of occurrence. Black curves show the truth while red curves show estimates from various simulations. The top row illustrates the implications of model mis-specification when average occurrence is only parametrically identifiable by single-visit data. The bottom row shows that in the non-parametrically identified double-visit case more data improves the estimates of occurrence.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2.5: The x-axis displays the value of the covariate that predicts detection. The y-axis displays the probability of detection. Black curves show the truth while red curves show estimates from various simulations. The first row illustrates the implications of model mis-specification when average detection is only parametrically identifiable by single-visit data. The bottom row shows that detection probabilities are non-parametrically identified using double-visit data, and estimation is robust to model mis-specification.

2.4 Discussion

In this chapter we aim to clarify controversies about identifiability in the species distribution and abundance literature. We do this by broadening the discussion to be about multiple senses of identifiability and the consequences of parametric assumptions rather than the realism of individual model choices. By clarifying that existing identifiability counter-examples serve to diagnose a lack of non-parametric identifiability, we change the focus of debates over controversial data-collection protocols.

We have shown that parametric identifiability alone does not suffice for reliable inference about properties of an unobserved distribution. Ecologists should be most confident in their results when a model satisfies identifiability in a sufficiently general super-model, even when they have no intention of estimating this model. Non-parametric identifiability ensures that the choice of parametric model does not create artificial identifiability and provides some robustness to model mis-specification.

In this section we connect our reasoning to other conversations in the literature about identifiability, discuss some practicalities, and provide ideas about how to further leverage our spline simulation approach.

Identifiability by Fiat in Other Contexts

When using species distribution and abundance models, the goal is often to do inference on properties of the unobserved distribution, such as prevalence, the average occurrence probability, or total abundance. We have shown that some assumptions that allow us to technically estimate these quantities lack robustness to model mis-specification. Because our properties of interest (e.g. prevalence) are often functions of the unobserved distribution, standard diagnostics fail to reveal model mis-specification when a property lacks non-parametric identifiability. What we need to check is that we chose the correct data-generating process from all of the ones that explain the observed data equally well. This is inherently uncheckable because we do not observe the components that we need to check.

For example, without non-parametric identifiability in a super-model, two data-generating processes will have the same likelihood of the observed data even though they differ in the joint likelihood of the observable and unobservable parameters. Since we cannot calculate the full likelihood, model selection metrics such as the Akaike information criterion (AIC) will not be able to distinguish between the two data-generating processes. Similarly, the unobserved data are unknown, so we cannot look for patterns in the residuals as an indicator of poor fit. Although there is a diagnostic proposed by Lele, Nadeem, and Schmuland and recommended by Lele et al. to warn of identifiability problems, the approach assumes that the true data-generating process lies within the proposed sub-model [98, 97]. Therefore this method will also not be able to diagnose the lack of identifiability under model mis-specification

Our examples show why having a narrow sense of parametric identifiability can be dangerous. Using an identifiable sub-model that restricts the possible space of data-generating process too stringently gives us identifiability by fiat, choosing from a wide range of feasible data-generating processes in a way that is completely determined by the model's parametric form. Cautions based on similar intuition are expressed in the literature beyond species distribution and abundance models. For example, Lele et al. warns those working on resource selection against making the jump from estimating use distributions with relative methods to estimating occupancy distributions with absolute methods [96]. The use distribution (the distribution of covariates among used sites) is identifiable using presence-only data, but the use distribution alone cannot help us estimate the probability of selection (the probability that an individual will select a resource given that it is encountered) and occupancy (the probability that a specific resource unit will be used during a specified period at least once). Assumptions used to get the needed available distribution (analogous to using background data to supplement presence-only data) are inherently uncheckable and may impact inference.

What identifiability means in the context of Bayesian analysis is debated [164]. Concerns about weak parametric identifiability and wariness towards model mis-specification arise when hierarchical Bayesian models are used to compensate for sparse data. Poirier explains that technically, parameters are always identifiable if an appropriate prior is chosen [144]. However, there may be parameters that the data do not inform. Eberly and Carlin warn against fitting complicated models without enough data since the posterior can depend heavily on the prior distributions [41]. In the case of sparse data, Lele advocated for performing sensitivity analysis on the parameters of interest [94]. We agree that sensitivity analysis is an important check. This paper shows that we not only need to be wary of the impact of prior choices, but we also need to carefully consider the impact of our likelihood choices. Relying on parametric identifiability without non-parametric identifiability can lead to undetectable sensitivity of inference under model mis-specification.

Paths Forward

Presence-only and single-visit methods have become popular due to more readily available opportunistically collected data and constraints on researchers that make multiple visits too resource-intensive. Although it can be tempting to do more with less, a lack of non-parametric identifiability makes inference perilous under inevitable model mis-specification. However, ecologists can focus on answering different questions, rely on survey design, and combine data from different sources to still make progress towards understanding species distributions and abundances.

Some ecological questions can be answered by only relying on prediction or relative inference. Presence-only data can be, in principle, used for these purposes. For conservation and management scenarios, the occurrences at the particular sites in a study area are of interest. Therefore, prediction as a metric of success makes sense. Methods that yield relative suitability of sites often suffice and are used to assess habitat suitability [135]. Critically, there are diagnostics for assessing the performance of relative probability of presence predictions [20, 128, 72, 141].

Some management decisions can still be made with a range of estimates for prevalence and occurrence probabilities. Although the single-visit occurrence scenario lacks the stronger form of non-parametric identifiability, we can partially identify the average occurrence and average detection. Some of the observationally equivalent data-generating processes revealed by our counter-example may not be plausible. For example, if we identify the product of average occurrence and average detection as 0.75, we know that the case where $\bar{\psi} = 0.625$ and $\bar{p} = 1.20$ is not plausible because detection probabilities cannot be above one. With partial identifiability we can bound $\bar{\psi}$ and $\bar{p}$ to both be in the interval $[0.75, 1]$ (where either, but not both could be equal to one), and depending on the scenario, this might be enough to make a management decision. Results on partial-identifiability with presence-only data can be found in the Appendix.

MacKenzie and Royle and Guillera-Arroita et al. make recommendations for designing a study such that the parameter estimates are the most precise for a given budget [105, 65]. Further work could recommend study designs that minimize sensitivity to model mis-specification. For example, within a single-visit, multiple surveys could be conducted, multiple observers could conduct independent surveys, or multiple subplot surveys per site could be conducted [105]. Alternatively, if using a multiple-visit protocol, the closure assumption is testable in certain scenarios where richer data is collected. Rota et al. proposed a test of whether the closure assumption is violated that works in conjunction with Pollock's robust design where an observer surveys the site multiple times per visit [157, 145]. We then assume closure between the multiple surveys within a visit but allow for changes between visits.

Data fusion and integrated population models provide another way forward. In the case of presence-only data, Phillips and Elith showed that we need additional information to anchor the observed data to the unobserved data [140]. If there is an independently and systematically collected presence-absence data set in the same region and for the same species, it can be used to evaluate our inference [43]. Having a site and species specific match may be unlikely, but multiple researchers provide a way to leverage a mixture of presence-only and presence-absence data [39, 53, 153].

Similarly, in the case of single-visit data, methods could be adapted to leverage a mixture of single-visit and multiple-visit data, i.e. a subset of sites could be visited more than once within a season [106, 104]. Certain aspects of the model could then be checked. For example, with some multiple-visit data the models for detection and occurrence probabilities could be checked at the cost of accepting the uncheckable assumption that the sites visited multiple times have the same detection and occurrence probabilities as (or can be modeled like) the ones visited once.

By supplementing lower quality data with higher quality data, some unobserved data becomes observed data. This makes some model checking feasible. Peel et al. show these mixture approaches fairing well under correct specification, but it would be interesting for future study to assess the accuracy and precision of estimates under model mis-specification as the proportion of presence-absence data or sites visited multiple times increases in these mixture approaches [136]. Another open question is how to diagnose a lack of generalizability when combining data from different sources.

Further Leveraging the Spline Simulation Method

As computational tools become more readily available, ecologists can fit more complex models of species distributions and abundance, attempting to account for more and more aspects of the underlying data-generating process. It would be prudent to probe identifiability to make sure the data is not pushed beyond the limitations of their information content.

Remembering that identifiability in one sub-model does not imply identifiability in a more flexible super-model and that a lack of non-parametric identifiability impacts our sense of having a useful approximation to the truth under model mis-specification, we must consider the potential impact of each parametric assumption. Approximating a non-parametric super-model with an unpenalized, flexible spline, as we did in our simulation studies, can be done in a variety of ecological scenarios beyond the ones discussed in this paper. Any time we make a distributional choice such as using the logit v. the probit link or a Poisson distribution v. a Negative Binomial distribution, fits of increasing spline complexity may reveal instabilities that hint at identifiability concerns. If we see stable behavior of estimates under one parametric model but not in an alternative one, we may have less confidence in our results, as they depend heavily on our model choice. When using unpenalized splines to allow for increased flexibility in this diagnostic scenario, we should not be confused with the more typically used spline formulation that regularizes to avoid too much “wiggliness” and overfitting.

We hope that by bringing these multiple forms of identifiability to the attention of the ecology literature, explaining their nuances, and designing an approach to reveal which identifiability regime a scenario is in, proper data-collection protocols will be adopted. We suggest researchers avoid estimating prevalence with presence-only data, estimating occurrence probabilities or abundance in the presence of imperfect detection with single-visit data, and estimating abundance in the presence of heterogeneous detection probabilities with capture-recapture data as we have illustrated the dangers. Richer data, rather than more data, will ensure robust inference as we continue to try to understand species distributions and abundance.

The first chapter of this dissertation clears up the identifiability controversy in the species distribution and abundance literature by explaining different levels of identifiability that relate to assumptions made about the data-generating process. As part of this effort we see the consequences for inference of the weaker form of parametric identifiability in the presence of model mis-specification. If our estimation model is wrong when using presence-only or single-visit data, our estimates of important properties such as species prevalence, species occurrence and detection probabilities, and species abundance can be very different from the true values. The second chapter of this dissertation continues with a focus on inference in the presence of model mis-specification, but this time the focus is on the case where we have the stronger form of non-parametric identifiability. Next is an investigation of identifiable joint species distribution models that account for multiple species at once to see how estimation of properties of the community fares under model mis-specification.

Part II

Stress Testing Latent Factor Approaches to Joint Species Distribution Models

Chapter 3

Stress Testing a Latent Factor JSDM

Co-authored with Perry de Valpine and Will Fithian

3.1 Introduction

Ecologists intuitively know that the occurrence and abundance of one species is likely related to that of another. As more data has become available on multiple species in overlapping regions and as more sophisticated species distribution and abundance models have been developed, it has become possible to explicitly account for known or suspected relationships between species in community ecology analyses.

A variety of approaches exist for incorporating community structure into ecological analyses including using another species as a covariate [87], using joint species distribution models [93, 129, 146, 60], applying multi-response approaches [124], analyzing networks [68, 147, 143], and using copulae [3]. However, the dimension of the problem quickly escalates when accounting for all possible covariances between pairs of species.

A latent factor approach to joint species distribution models (JSDMs) has been popular as it reduces the complexity to estimation of a number of latent factors that is often much smaller than the number of species [192]. These latent factors induce correlations between species and can be thought of as representing unobserved covariates that impact species occurrence. Many approaches to estimating these latent factor models have appeared in the literature [130, 76, 131, 177].

There have been a variety of comparisons of these multiple approaches to JSDMs. Zhang et al. investigate how characteristics of the sampling data, such as number of sampling sites and species, affect the predictive capability of JSDMs [201]. Wilkinson et al. compare the computational performance and the parameter estimates of a variety of JSDMs [195]. Norberg et al. evaluate the predictive performance of species distribution models (together with some JSDMs), including prediction of held out data to assess both interpolation and extrapolation capabilities [126]. Zurell et al. and Thurman et al. test how JSDMs detect species interactions (the former using point-processes to simulate interactions, the latter using real data with known species interactions) [203, 179]. However, interpreting co-occurrence as evidence of an interaction remains controversial [13].

These comparisons tend to either focus on the case where the model is correctly specified or assess performance by predictive metrics or global measures of goodness-of-fit. Since ecologists often use joint species distribution models to understand relationships between species [48, 148], assess the species composition across regions of interest [8, 202], and understand species' responses to environmental variables [29, 113, 155], there is still room for further investigation into model performance in terms of community properties.

Our work here adds insight into the performance of JSDMs in two ways. First, we stress test the latent factor approach under mis-specification of the species covariance structure. As we have no a priori reason to believe that a joint species distribution model correctly reflects the underlying data-generating process, it is important to test that the approach is fit for its purpose and sufficiently robust to realistic levels of model mis-specification. Second, we evaluate performance based on estimation of community metrics, like Shannon's Entropy, species richness, Pielou's evenness, and Jaccard similarity, rather than on global prediction metrics. Although there have been critiques of reliance on community diversity metrics that fail to fully capture the ecology (e.g. [80]), defenders have argued that each metric gives insight into a certain aspect of the ecology (e.g. [150]), and practical guides help ecologists navigate the many metrics available for analysis [4, 119]. Community diversity metrics have a long history in community ecology and represent the community characteristics that researchers have been interested in over time, so it is of interest to ask whether JSDMs estimate these characteristics well.

3.2 Materials and Methods

Wilkinson et al. show that many of the different approaches to latent factor modeling for joint species distributions are variations on Hui's Bayesian Ordination and Regression Analysis (BORAL) model by standardizing notation across a variety of JSDM methods [195, 76]. In this Bayesian model (described in Table 3.1) the observed data for J species come from I sites and are the true occurrences Y_{ij} (where $Y_{ij} = 1$ denotes a presence and $Y_{ij} = 0$ denotes an absence) for species j at site i. To simplify the investigation we make the perfect detection assumption. In practice this is a tenuous assumption [9, 191, 199], but imperfect detection has been tackled in multispecies N-mixture models [40], in a joint species distribution model for two species [159, 158], and in a joint species distribution model for more species [180].

The probability that $Y_{ij} = 1$ is assumed to be drawn from a Bernoulli distribution with the probability of the event being $\Phi(u_{ij} + v_{ij})$ (where Φ is the cumulative distribution function of the standard normal). The u_{ij} are fixed effects related to covariates X_{ij} that account for variation in the observed occurrences. In this simulation study we keep the fixed effect simple by only estimating species-specific intercepts β_j (using a zero-mean normal distribution with variance 10 as a prior distribution). The v_{ij} contain information about the correlations between species. We assume the v_{ij} are determined by the inner product of $K << J$ random effects $\eta_{i.}$ (a $1 \times K$ vector) and parameters $\lambda_{.j}$ (a $K \times 1$ vector). The random effects are assumed to come from independent standard normal distributions and the λ_{kj} are parameters to be estimated (using a zero-mean normal distribution with variance 10 as a prior distribution).

Latent factors have been used in a variety of other contexts to reduce the dimension of the estimation problem, from an approach similar to Hui et al. that is more focused on ecological prediction [78, 187] and ecological structural equation models [108, 49] to many approaches in psychology and the social sciences [17]. In this context, the λ_{kj} parameters can be thought of as species-specific coefficients for the unobserved covariates $\eta_{i.}$ at the site level. The $K \times J$ matrix Λ where each row k is the vector $\lambda_{k.}$ becomes part of the $J \times J$ species covariance matrix $\Lambda'\Lambda$. This species covariance matrix represents the associations between species that are accounted for by the latent factors.

Throughout the simulation study we use the R package, *boral* to do the model fitting [77]. This implementation uses MCMC to fit the model described in Table 3.1. We use the default values of number of iterations (40000), burn-in (1000) and thinning (30).

Description	**Notation**	**Distribution / Property**
Site Index	i	
Total Number of Sites	I	
Species Index	j	
Total Number of Species	J	
Latent Factor Index	k	
Total Number of Latent Factors	K	$K << J$
Total Number of Covariates	M	
Probability of Occurrence	$P(Y_{ij} = 1)$	$\sim$ Bern$(\Phi(\mu_{ij} + v_{ij}))$
Fixed Effects	μ_{ij}	$= X_{i.}\beta_{.j}$
Matrix of Covariates	X	$I \times M$
Species-specific Coefficients (parameters to be estimated)	β_{mj}	prior $\sim N(0, 10)$ $m = 1, ..., M$ $j = 1, ..., J$
Linear Predictor of Unmeasured Covariates	v_{ij}	$= \eta_{i.}\lambda_{.j}$
Site-specific Random Effects	η_{ik}	prior $\sim N(0, 1)$
Species-specific Factor Loadings (parameters to be estimated)	λ_{kj}	prior $\sim N(0, 10)$
Factor Loadings Matrix	Λ	$K \times J$ each row k is vector $\lambda_{k.}$
Species Covariance Matrix	$\Lambda'\Lambda$	$J \times J$

Table 3.1: For a choice of K, denoted K_{est}, this is the BORAL estimation model that is fitted to the observed Y_{ij} [76]. The R package, *boral*, is used to implement and fit this model via MCMC [77].

In this simulation study the primary mis-specification focus is on the relationship between species, ignoring differing species prevalences. We assume the species-specific intercepts are all zero, yielding occurrence probabilities of 0.5 for every species.

We also investigate a subset of the simulation scenarios using more realistic species prevalences. To do this we get a distribution of realistic species prevalences from the bryophyte data used in Ovaskainen et al. and use it to define species-specific intercepts in the simulation model [131]. Due to the computational cost of each simulation in the R implementation we use, we investigate only one site-to-species case. The distribution of 25 species prevalences used is shown in Figure 3.1. There are many species with small prevalences, and a few species with prevalences greater than 0.5.

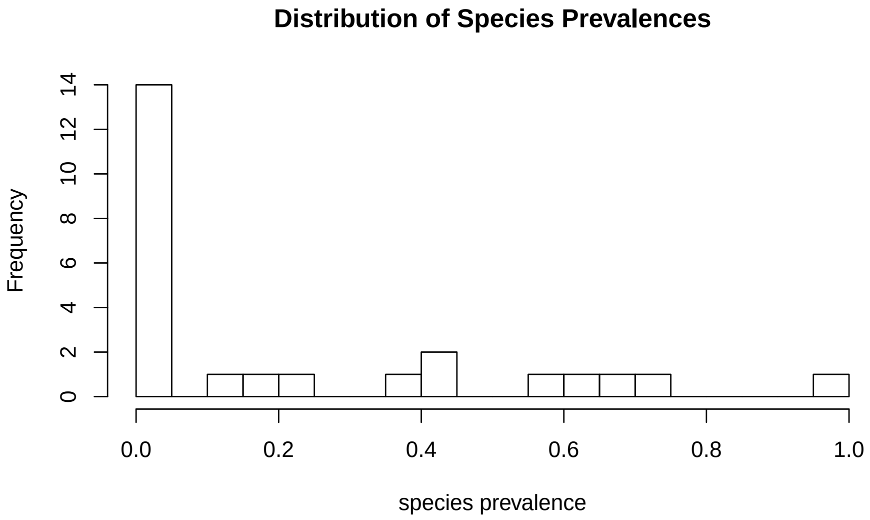


Figure 3.1: Distribution of 25 species' prevalences subsampled from bryophyte data in Ovaskainen et al. [131]. Contrasting the scenario with equal prevalences of 0.5, there are many species with small prevalences, and a few species with larger prevalences.

Approaches to Mis-Specification Simulation

With the BORAL **estimation model** in mind, there are several ways that a **simulation model** can be chosen such that the species covariance matrix $\Lambda'\Lambda$ is mis-specified. These choices include breaking modeling assumptions typical of latent factor models [45, 46].

Three types of mis-specification are considered. The structure of the data-generating process could be correct but we choose the wrong number of latent factors to include in the model (conceptually similar to omitting an important covariate or including a superfluous covariate in the fixed effect). We'll refer to this scenario in the results as the "wrong-K" scenario (in contrast to the "correct-K" scenario where the number of latent factors in both the estimation and simulation model are the same).

The assumption that a small number of latent factors captures the associations between species may not be appropriate (the true relationship is not well approximated by a low-rank matrix). We'll refer to this scenario as the "block-covariance" scenario in the results (an explanation of the name follows).

We could choose the wrong distribution for the underlying latent factors (typically assumed to be normally distributed). We'll refer to this scenario as the "heavy-tail" scenario in the results. The heavy-tail scenario can be partitioned into cases where the model is fitted with the correct or wrong number of latent factors. We will refer to these sub-scenarios as "correct-K, heavy-tail" and "wrong-K, heavy-tail" scenarios in the results.

In this simulation study we generate occurrence data from a variety of scenarios that approximate these different types of mis-specifications to assess whether the latent factor model approach is robust to each in turn.

Correct-K: $K_{sim} = K_{est}$

To get a baseline for performance we generate occurrence data from a simulation model chosen to match the estimation model (see Table 3.1). We also choose the number of latent factors to fit to the data, K_{est}, to be equivalent to the number of simulated latent factors, K_{sim}, making this the "correct-K" special case. Since Zhang et al. show that the ratio of sites to species can be influential in prediction performance, we generate scenarios using $J =$ 10, 25, and 45 species across $I = 50$ sites [201]. We can also think of the site-to-species ratio as a site-per-parameter ratio as the number of parameters scales linearly with the number of species. Having data collected at more sites provides additional information to help estimate the species-specific intercepts and factor loadings.

Wrong-K: $K_{sim} \neq K_{est}$

In practice, we do not know what the appropriate number of latent factors is, so we rely on model selection strategies such as cross-validation and information criteria [78] or, in a Bayesian framework, using shrinkage to choose the number of latent factors [11].

If we think about latent factors as substitutes for covariates we were unable to measure, the case when $K_{sim} < K_{est}$ means that we have included covariates in the model that do not explain the covariances between species. If $K_{sim} > K_{est}$, we may have omitted covariates that would help explain covariances between species.

For each of the datasets simulated from the "correct-K" scenario described above, we fit "wrong-K" models with $K_{est} \in \{\{1, 2, 3, 4, 5\} | K_{est} \neq K_{sim}\}$.

Block-Covariance: $K_{sim} >> K_{est}$

A special case of using the wrong number of latent factors in a model is when the underlying relationship between species is based on a large number of latent factors. More formally, this means that the covariance matrix between species has a large rank. It seems feasible that the relationships between a large number of species may be described by many latent factors. For example, in an ecological context, a full rank block diagonal covariance would represent a scenario where species would cluster, i.e. a species would be closely related to a small number of other species but unrelated to a majority of others. We want to know if a small number of latent factors can approximate a full rank covariance matrix, i.e. will a low-rank approximation suffice.

To assess robustness to the assumption that a small number of latent factors can capture the covariances between species, we generate occurrence data using a full rank block diagonal covariance matrix. We define a block to be a $B \times B$ portion of the $J \times J$ species covariance matrix (where $J = 30$ and $I = 100$ in these simulations) with ones on the diagonal and 0.9 on the off-diagonals. As we increase the block size B from 3 to 5 to 10, the total number of blocks in the covariance matrix decreases from 10 to 6 to 3.

Despite data being generated from a distribution governed by a full rank block diagonal covariance matrix, the covariance matrix could in principle be described well by a low rank approximation of an appropriate size. For example consider the species covariance matrix below that is made up of two blocks of three species. Within a group, species have covariance a with one another, but they are unrelated to species outside of the group.

$$\Sigma = \begin{bmatrix} 1 & a & a & 0 & 0 & 0 \\ a & 1 & a & 0 & 0 & 0 \\ a & a & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & a & a \\ 0 & 0 & 0 & a & 1 & a \\ 0 & 0 & 0 & a & a & 1 \end{bmatrix}$$

This covariance matrix Σ would be well approximated in a latent factor model by $\Lambda'\Lambda$ formed by two latent factors where $\Lambda = \begin{bmatrix} \sqrt{a} & \sqrt{a} & \sqrt{a} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{a} & \sqrt{a} & \sqrt{a} \end{bmatrix}$, especially if $\sqrt{a}$ is close to one.

However, when using a latent factor model in practice, it is rare to fit beyond a handful of latent factors. For example, for ordination, a popular approach for visualizing multivariate data that inspired the BORAL model, only one or two dimensions are used, and the *boral* implementation warns users against fitting more than five latent factors [78, 77]. In the case where the number of blocks is large, the number of latent factors needed for a good approximation might be too large to be considered in the fitting process. In the block-covariance scenario we expect the three block case to be well approximated by a model with three latent factors, but we do not expect the six and ten block cases to be well approximated by five or fewer latent factors.

Heavy-Tail: $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0, 1)$

In the estimated latent factor model we assume that the latent factors η_{ik}^{est} come from independent normal distributions, but what if the simulation model has latent factors that come from a heavier-tailed distribution? Others have studied model mis-specification of this type in linear and generalized linear models to assess impact on prediction of both fixed and random effects (see [114] and citations within). We further investigate its impact on prediction of functions of model parameters, i.e. community metrics.

In these simulation models we generate latent factors η_{ik}^{sim} from independent t distributions (with 3, 5, and 20 degrees of freedom) to assess robustness to mis-specification of the random effects themselves. Note we standardize the simulated $v_{ij} = \eta_{i.}\lambda_{.j}$ to make a fair comparison to other results.

Metrics of Performance

In this simulation study we evaluate the estimation of Shannon's Entropy H, species richness S, Pielou's evenness E, and Jaccard similarity J [107]. The chosen metrics are calculated at the site-level or at the site-pair level (e.g. Jaccard similarity), so we need to further aggregate, typically by taking an average across sites or pairs. See Table 3.2 for definitions of these metrics and the aggregations used. However, there are some dissimilarity metrics that do not require an additional aggregation that could also be used [38, 154].

Name	**Notation**	**Description**
Shannon's Entropy (site level)	H_i	$= -\sum_{j=1}^{J} \frac{Y_{ij}}{J} \log \frac{Y_{ij}}{J}$ $= S_i \log J$
Shannon's Entropy (average)	H	$\frac{1}{I} \sum_{i=1}^{I} H_i$
Species Richness (site level)	S_i	$= \sum_{j=1}^{J} Y_{ij}$
Species Richness (average)	S	$= \frac{1}{I} \sum_{i=1}^{I} S_i$
Pielou's Evenness (site level)	E_i	$= H_i / \log(S_i)$ $= \frac{S_i \log J}{\log S_i}$
Pielou's Evenness (average)	E	$= \frac{1}{I} \sum_{i=1}^{I} E_i$
Jaccard Similarity (between two sites)	$J_{ii'}$	$= a/(b + c - a)$
Number of Species in Both Site i and i'	a	
Number of Species in Site i'	b	
Number of Species in Site i	c	
Jaccard Similarity (average)	J	$= \frac{1}{\binom{I}{2}} \sum_{i=1}^{I} \sum_{i' \neq i} J_{ii'}$

Table 3.2: Community Metric Definitions for Occupancy Data: Note that H_i, S_i, and E_i are closely related to one another in the special case where we use occupancy data as opposed to abundance data. We display results for all three as researchers may be used to working with one over another.

Determining the "true" value of community metrics corresponding to model assumptions is not straightforward as they are not simple functions of the model parameters. Consider observed data Y that come from a data-generating process D (the simulation model). Also consider a parametric Bayesian estimation model $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ with a prior on θ, denoted $\pi(\theta)$. In this simulation study, the model parameters θ are the species-specific intercepts, denoted together as a vector $\boldsymbol{\beta}$, and the species-specific factor loadings $\lambda_{k.}$ that make up the rows of Λ. We can simplify notation by referring to these parameters as $\theta = \{\boldsymbol{\beta}, \Lambda\}$.

The true data-generating process D may or may not be included in the estimation model $\mathcal{P}$ (defined for these scenarios in Table 3.1).

For any set of observations Y, there is an observed value of a community metric $T(Y)$. This value will vary between replicates within a simulation model. We are interested in estimating the expected value of $T(Y)$ over the distribution of communities $\mathbf{Y}$ generated by D, denoted by the following.

$$\gamma(D) = \mathbb{E}_{Y \sim D}[T(Y)] \tag{3.1}$$

The parameter of interest $\gamma(D)$ exits for any data-generating process D. In the BORAL estimation model, data-generating processes take the form of a parametric model p_θ, so it can be helpful to denote $g(\theta) = \gamma(p_\theta)$. We can approximate $\gamma(D)$ via Monte Carlo simulation. For $b = 1, ..., B$ we draw a new dataset $Y^{(b)}$ from the data-generating process D. Then we approximate $\gamma(D)$ by the following expression.

$$\tilde{\gamma}(D) = \frac{1}{B}\sum_{b=1}^{B} T(Y^{(b)}) \tag{3.2}$$

The values of $\tilde{\gamma}(D)$ for each of the simulation scenarios (using $B = 1000$) can be found in the Appendix.

To assess model fit, we need to compare $\tilde{\gamma}(D)$ with corresponding predictions from estimated models. This simulation study uses Bayesian estimation to fit a parametric model, so the mean posterior value of $g(\theta)$ is of interest. By Bayes Rule, the posterior distribution of θ, denoted $\pi_{post}(\theta|Y)$, is proportional to the product of the prior on θ, $\pi(\theta)$, and the likelihood of the data given θ, $p_\theta(Y)$.

$$\pi_{post}(\theta|Y) = \frac{\pi(\theta)p_\theta(Y)}{\int \pi(\theta)p_\theta(Y)d\theta} \tag{3.3}$$

Remember that the choice of estimation model $\mathcal{P}$ may or may not include the true data-generating process D. We choose to estimate the predicted value of $\gamma(D)$ using the posterior mean of $g(\theta)$ under the Bayesian model as follows.

$$\hat{\gamma}_{Bayes}(Y) = \mathbb{E}_{\theta \sim \pi_{post}(\theta|Y)}[g(\theta)] \tag{3.4}$$

This mean is actually a double expectation: an expectation over the posterior distribution of model parameters and an expectation over the distribution of communities predicted by given values of the model parameters. The predicted value of $\gamma(D)$ can be approximated by first using posterior samples $\theta^{(a)} \sim \pi_{post}(\theta|Y)$ for each $a = 1, ..., A$ to generate new observations $Y^{(a,b)}$ for $b = 1, ..., B$ and then computing each $\tilde{\gamma}(p_{\theta^{(a)}})$ as described above. We use $A = 250$ and $B = 500$. Then we approximate $\hat{\gamma}_{Bayes}(Y)$ by the following.

$$\tilde{\gamma}_{Bayes}(Y) = \frac{1}{A}\sum_{a=1}^{A} \tilde{\gamma}(p_{\theta^{(a)}}) \tag{3.5}$$

In the next section we compute and compare $\tilde{\gamma}(D)$ and $\tilde{\gamma}_{Bayes}(Y)$ for each set of simulation models and estimated parameters. We focus on the relative bias, the expected relative difference between the true expected value of a community metric and the posterior expectation of the community metric (over communities $\mathbf{Y}$ generated by D), as follows.

$$\mathbb{E}_{Y\sim D}\left[\frac{\gamma_{Bayes}(Y)-\gamma(D)}{\gamma(D)}\right]\approx \mathbb{E}_{Y\sim D}\left[\frac{\tilde{\gamma}_{Bayes}(Y)-\tilde{\gamma}(D)}{\tilde{\gamma}(D)}\right] \tag{3.6}$$

To approximate the sampling distribution of relative error (over possible communities $\mathbf{Y}$ generated by a fixed D), we replicate the generation of observed data and model fitting for every simulation model. In the R implementation that we use, the combination of model fit and computation of the predicted expected community metric for each scenario can take on the order of minutes to complete. Some combinations even take more than ten minutes to run for each replicate. Since we aim to explore a variety of scenarios, computation time was a limiting factor. Therefore, in the results we limit the number of simulation replicates for each scenario to 50.

Simulation Evaluation

The results for each simulation scenario answer specific questions about estimation of community metrics under model mis-specification. The primary question is: is the estimation unbiased under the particular data-generating process and estimation model? Second, within a particular scenario, is there robustness to "wrong-K" cases?

When we evaluate estimation of community metrics for the "correct-K" cases of each scenario, each panel of a results plot shows the outcome for a different community metric. The x-axis shows the number of latent factors used in the model fit, K_{est}, the y-axis shows the relative error, and the color represents the number of species (where the number of sites is constant at 50). The error bars represent plus-or-minus two standard deviations of the relative error across 50 replicates within a particular scenario. If these cover zero, estimation of the particular community metric is unbiased in the scenario.

There are a few exceptions to these plots. The block-covariance scenario does not have a "true" number of latent factors to display in a plot. The heavy-tail scenario has panels that correspond to the combination of community metric and the degrees of freedom for the random effects.

When we evaluate the estimation of community metrics for "wrong-K" cases, each row in the results plots represents a community metric, each panel in a row denotes the true number of latent factors, K_{sim}, and the x-axis shows how many latent factors were used in the fitting procedure, K_{est}. If the magnitude of the relative error does not increase between the "correct-K" and "wrong-K" cases within a scenario, there is some robustness in estimation of the community metric. Again, the block-covariance and heavy-tail scenarios are the exceptions as noted above.

In the following results plots we expect the magnitude of the relative bias to be larger for the mis-specified scenarios than for the correctly specified scenario if the estimation of community metrics is not robust to the mis-specification. We also expect the magnitude of the relative bias to increase if $K_{sim} > K_{est}$ and to be fairly unchanged if $K_{sim} < K_{est}$ (although we expect more variability) if the estimation of community metrics is not robust to the "wrong-K" form of model mis-specification.

3.3 Results

In this section we show the results for the four scenarios described in Section 3.2. For each specification scenario, we first look at the special "correct-K" case where the true number of latent factors, K_{sim}, matches the number of latent factors used in the model fit, K_{est}, (even if other aspects of the estimation and simulation models don't match). Then we investigate the "wrong-K" version of these scenarios, estimating community metrics when $K_{est} \neq K_{sim}$. This reflects a more realistic situation; in practice we do not know the true number of latent factors, K_{sim}.

Correct-K

In the correct-K scenario Figure 3.2 shows that the relative biases of the Shannon Entropy H, species richness S, and Pielou's evenness E are not significantly different from zero across all of the site-to-species parameter ratios, although estimation variance is higher when the ratio of sites to the number of species parameters to be estimated is larger. The same is mostly true for the Jaccard similarity J, but the predicted values are often smaller than the truth. In practice, this means that the true species community is more similar across sites than the predicted species community. We provide some insight into this downward bias in Section 3.4.

When we include more realistic species prevalences, the Jaccard similarity estimation is unbiased, but the Shannon Entropy and species richness tend to be overestimated. We provide some insight into this upward bias in Section 3.4. However, it should be noted that particularly for the species richness, the bias is not large in context. In this example, 25% of the species are present on each site. A 10% bias for species richness out of a possible 25 species would only decrease the species richness by a fraction of a species. See the Appendix for true values of the community metrics to get a further sense of scale.

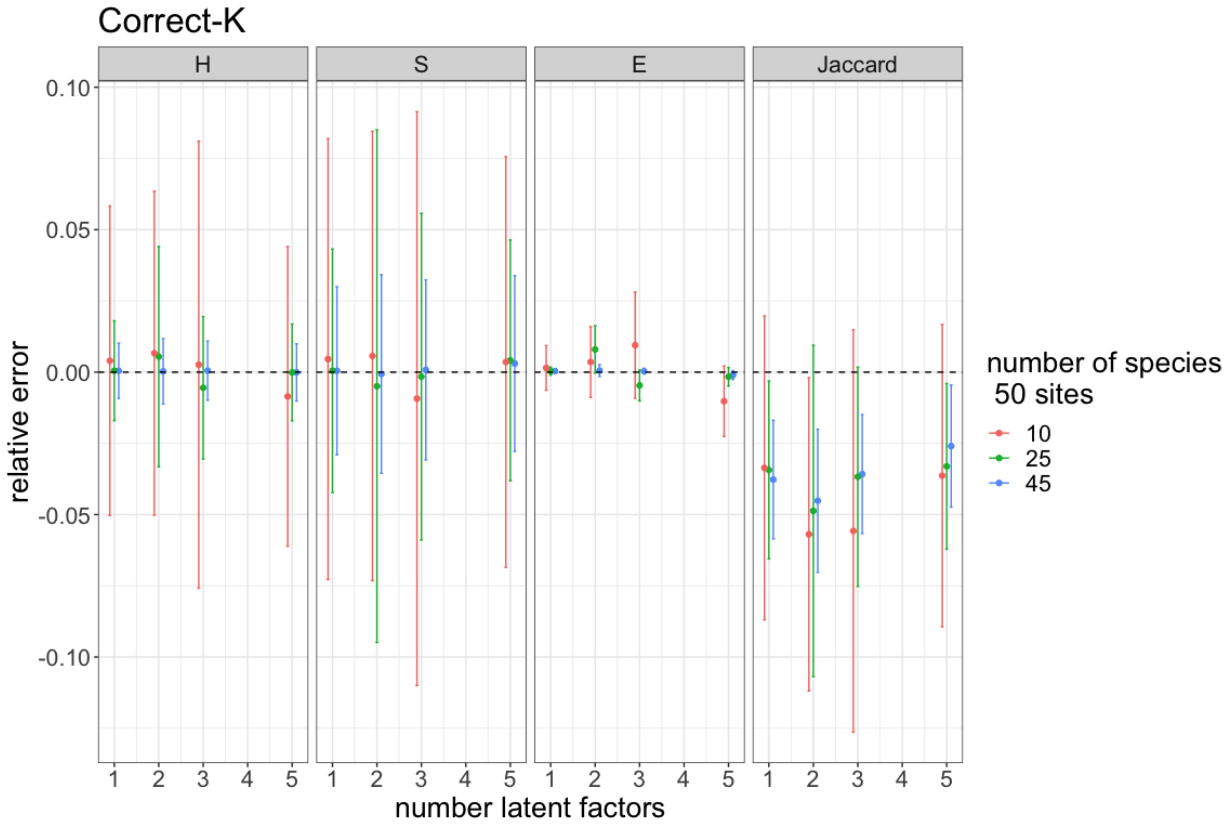


Figure 3.2: Correct-K ($K_{sim} = K_{est}$): The x-axis represents $K_{est} = K_{sim}$, and the y-axis represents the relative error. Shannon Entropy, species richness, and evenness are unbiased across a variety of site-to-species parameter ratios and number of latent factors. The Jaccard similarity is downward biased for most of the scenarios, implying that the predicted species community is less similar across sites than the true one.

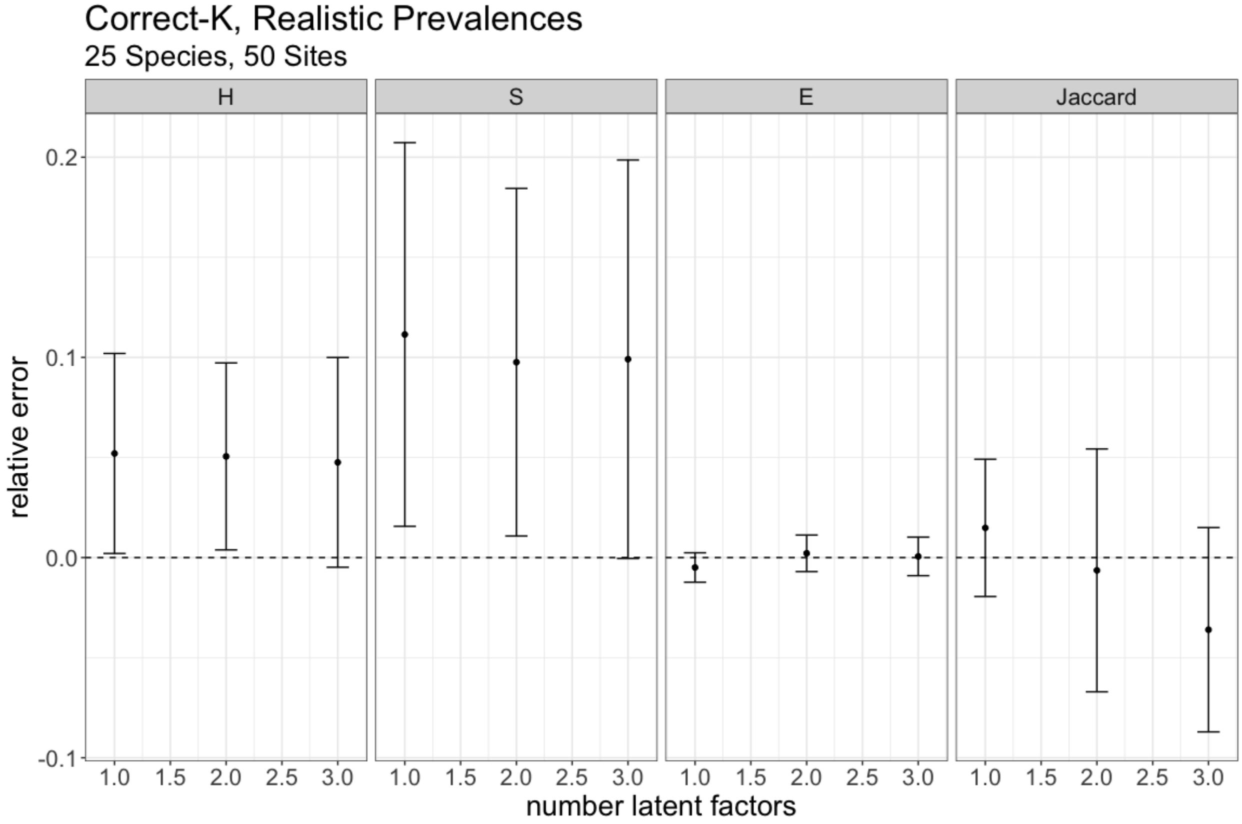


Figure 3.3: Correct-K with Realistic Species Prevalences ($K_{sim} = K_{est}$): The x-axis represents $K_{est} = K_{sim}$, and the y-axis represents the relative error. Only one site-to-species parameter ratio is pictured (25 species for 50 sites). Evenness and the Jaccard similarity are unbiased while the Shannon Entropy and species richness are slightly upward biased.

Wrong-K

Estimation of Shannon's Entropy H and species prevalence S are robust to the "wrong-K" case that might occur in practice (see Figures 3.4 and 3.5). Estimation variance across replicates depends on the ratio of sites to number of estimated parameters. Evenness E appears slightly less robust to the choice of K_{est} used in the fit, and the Jaccard similarity maintains its downward bias across all choices of K_{est} (see Figures 3.6 and 3.7).

When we allow the distribution of species prevalences to be more realistic, estimates of evenness and the Jaccard similarity tend to be unbiased while Shannon's Entropy and species richness tend to be overestimated (see Figures 3.8 and 3.9).

When $K_{sim} = 1$, we see worse and worse performance as K_{est} increases for H, S, and the Jaccard similarity. For other community metrics and K_{sim} pairings we do not see a strong decline in performance when $K_{sim} < K_{est}$ or $K_{sim} > K_{est}$.

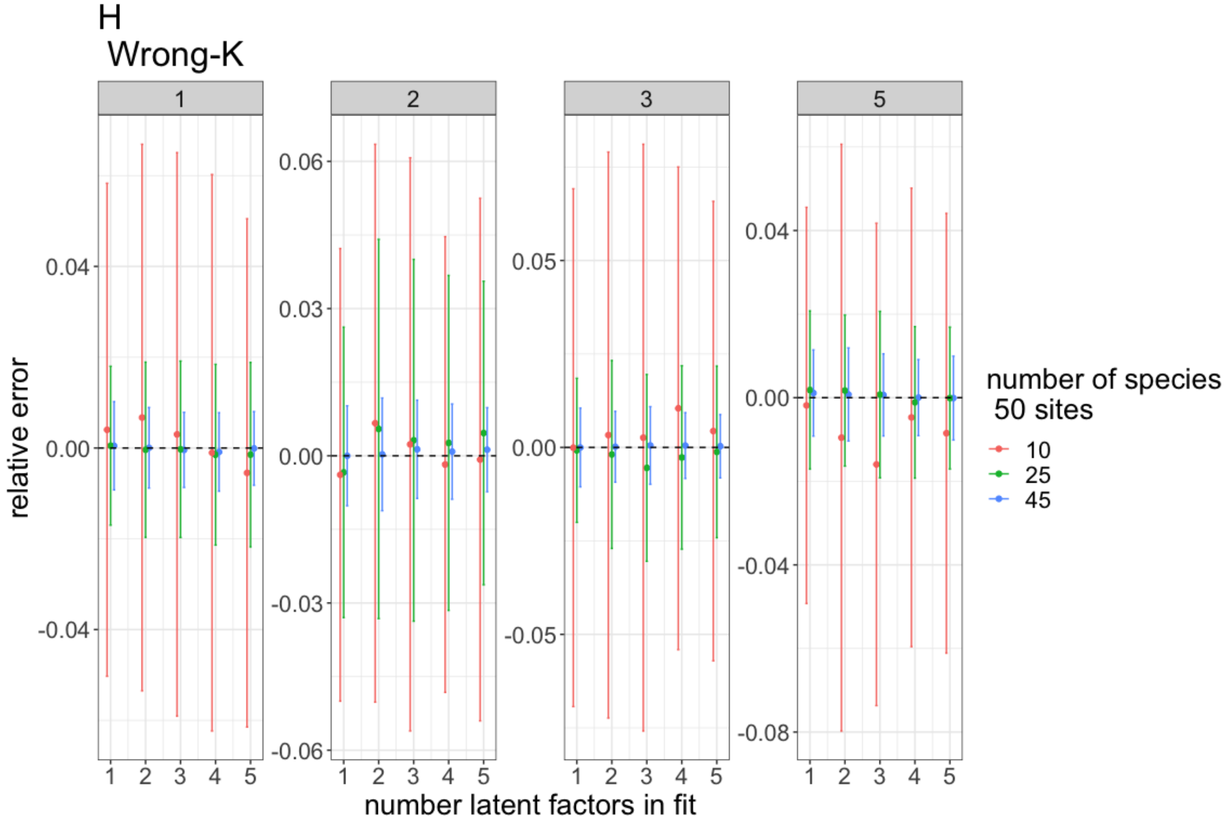


Figure 3.4: Wrong-K ($K_{est} \neq K_{sim}$): Shannon's Entropy has consistently unbiased estimates across scenarios.

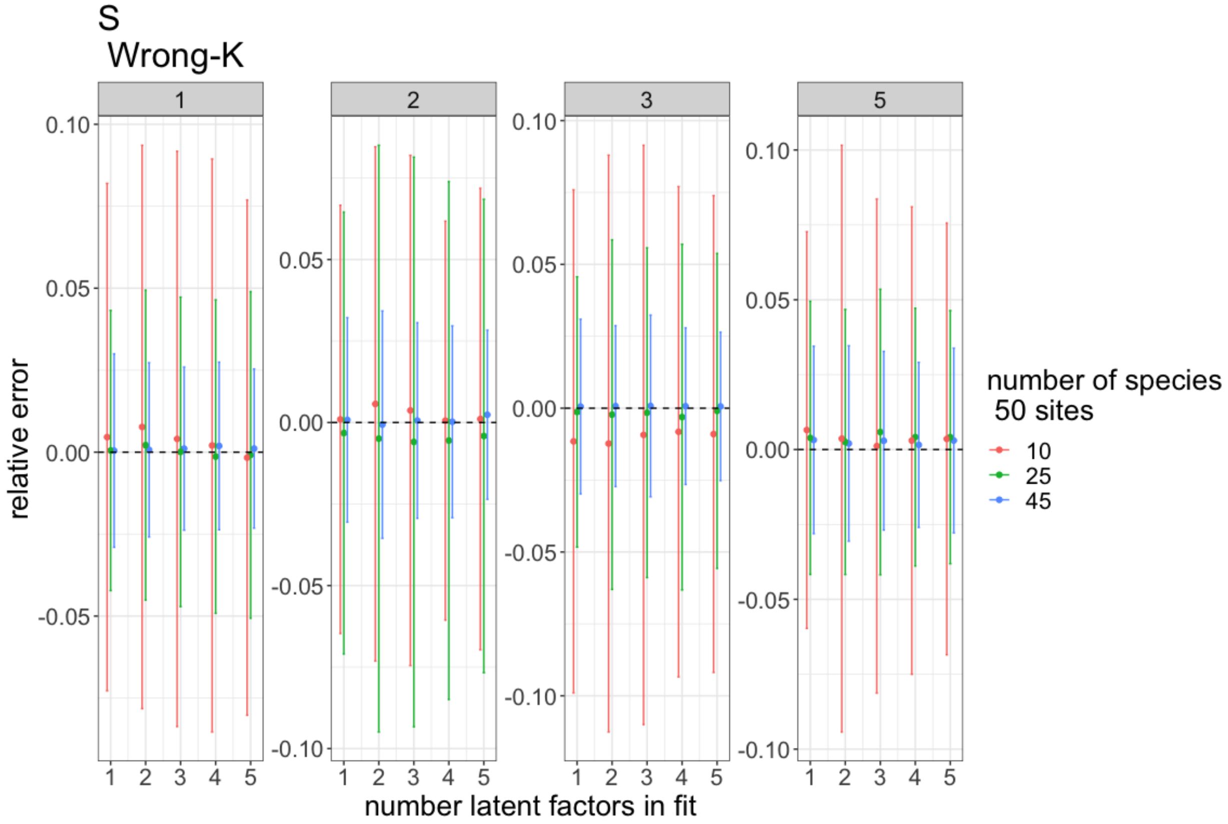


Figure 3.5: Wrong-K ($K_{est} \neq K_{sim}$): Species richness also has consistently unbiased estimates across scenarios.

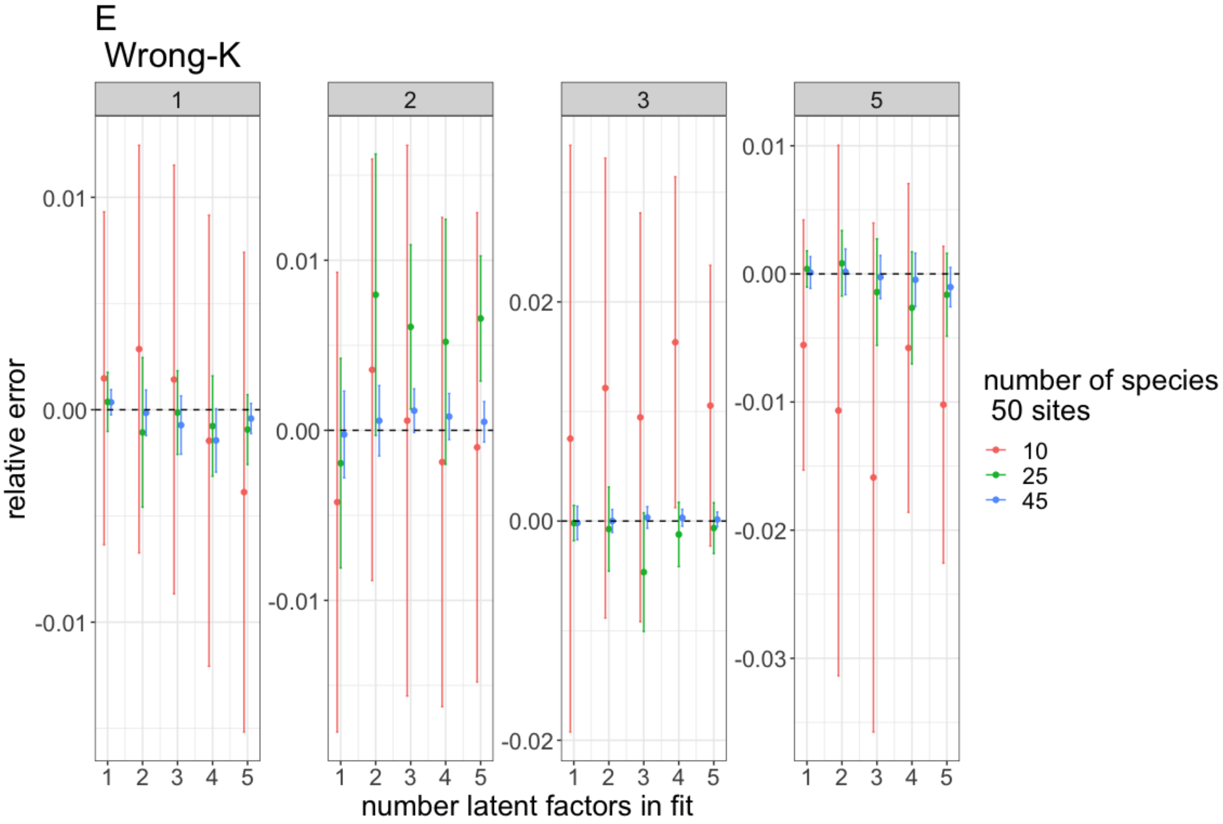


Figure 3.6: Wrong-K ($K_{est} \neq K_{sim}$): Evenness is also unbiased fairly consistently.

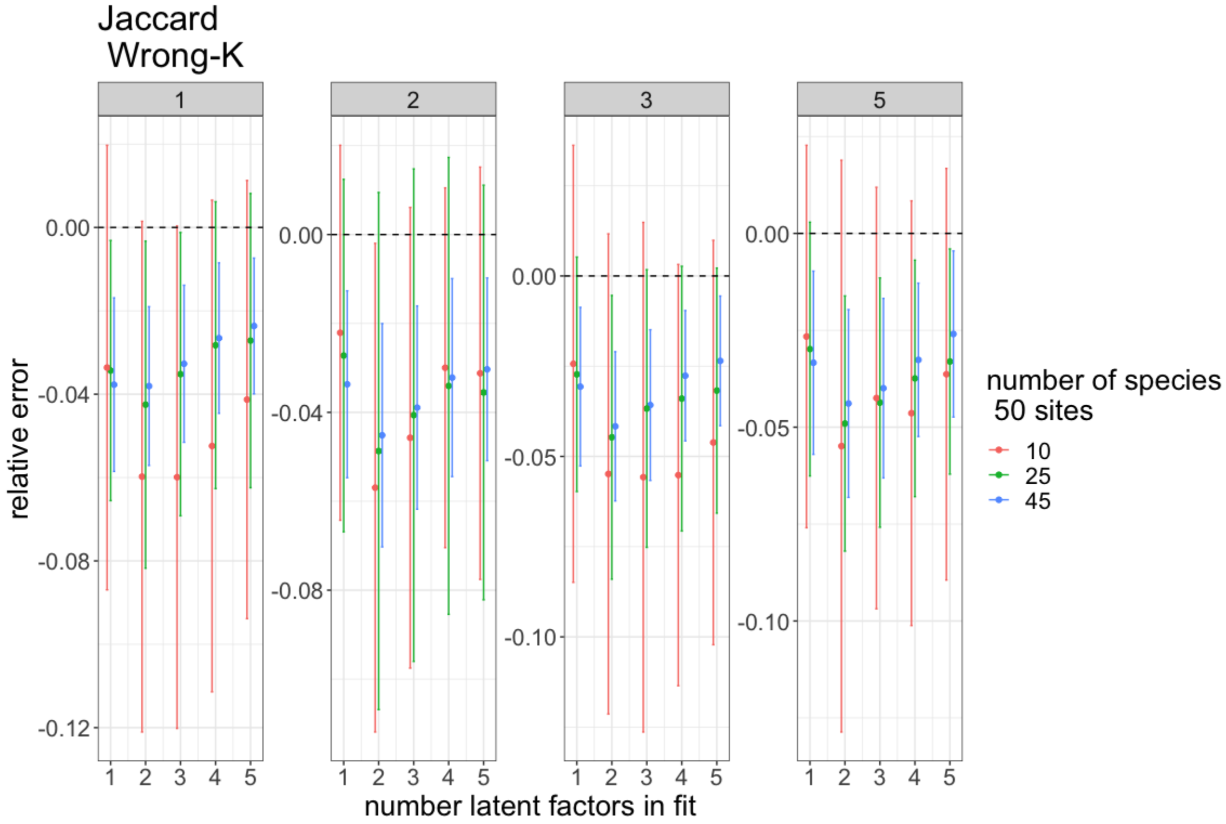


Figure 3.7: Wrong-K ($K_{est} \neq K_{sim}$): The Jaccard similarity remains downward biased across most scenarios.

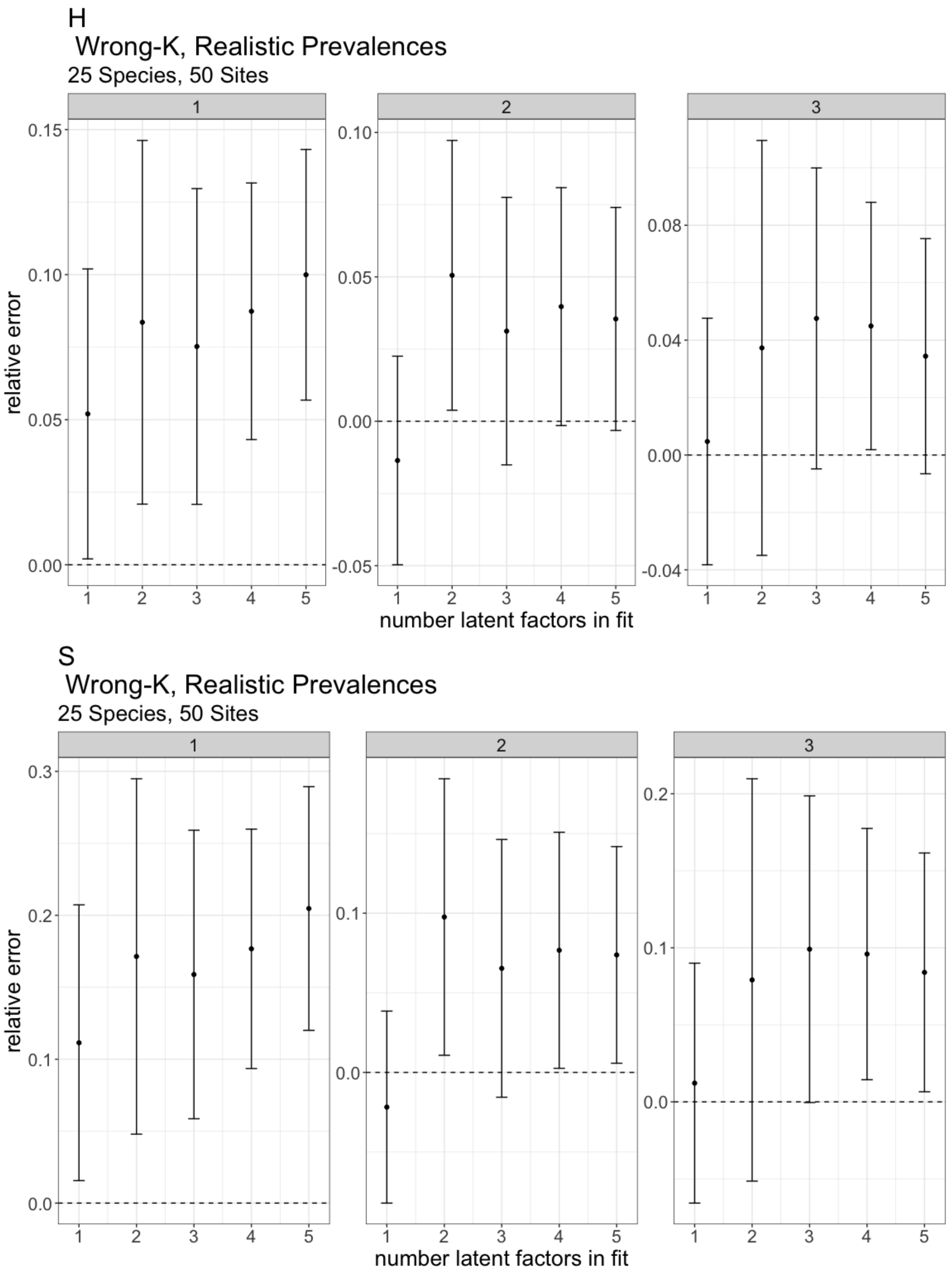


Figure 3.8: Wrong-K with Realistic Species Prevalences ($K_{est} \neq K_{sim}$): As in the correctly specified case with realistic prevalences, Shannon's Entropy and species richness are slightly upward biased.

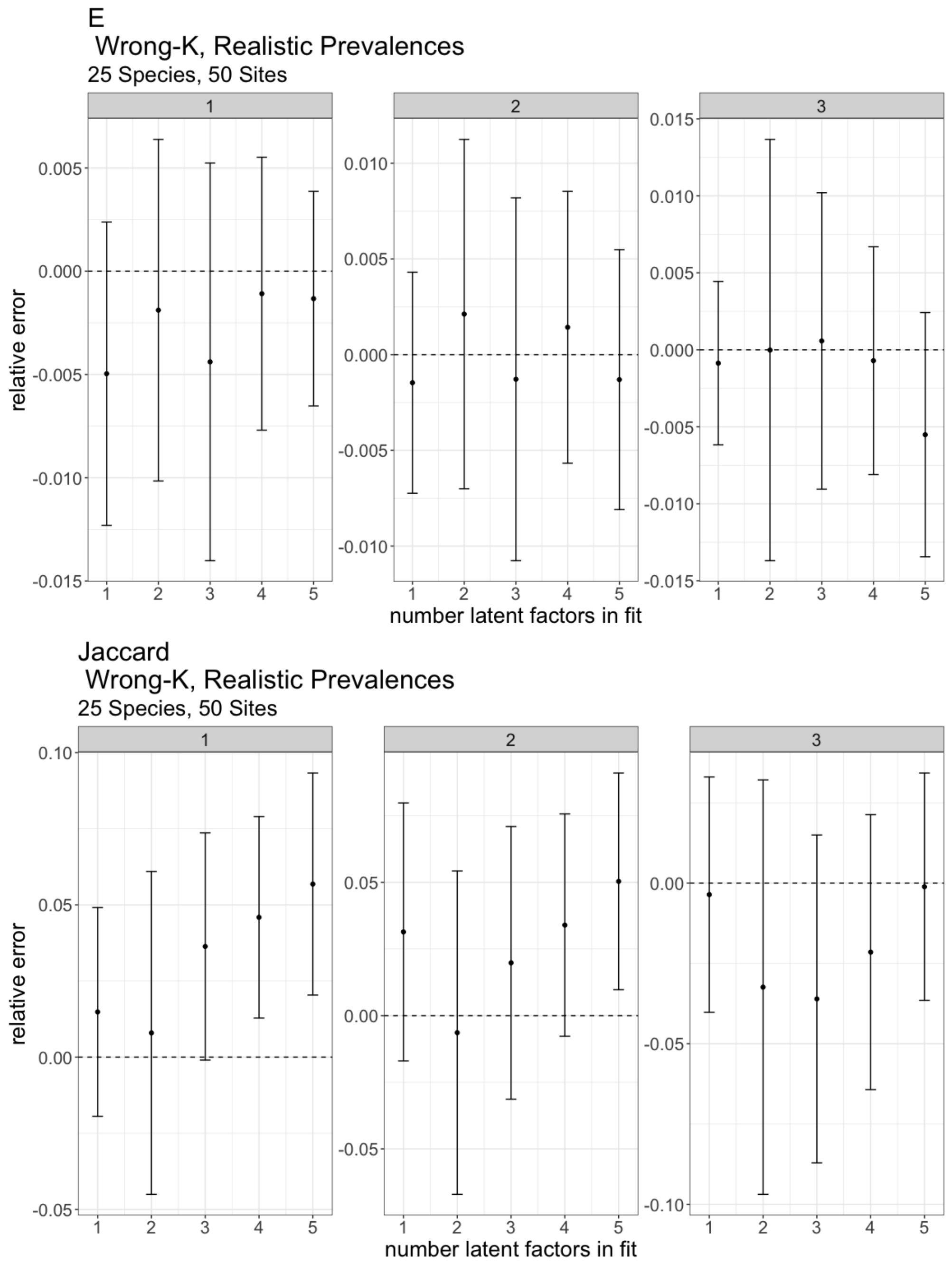


Figure 3.9: Wrong-K with Realistic Species Prevalences ($K_{est} \neq K_{sim}$): Evenness and the Jaccard similarity appear to be fairly robust to the number of latent factors.

Block-Covariance

The estimation of H, S, and the Jaccard similarity is fairly robust to the block-covariance scenario (see Figure 3.10). In the three block scenario, E is not well estimated by more than three latent factors. This is appropriate as those models are mis-specified in the sense that they have too many latent factors.

For realistic species prevalences estimation performance breaks down (see Figure 3.11). The relative bias for estimation of H, S, and the Jaccard similarity get increasingly worse as K_{est} increases. The performance of E here is similar to its performance in the case of equal prevalences.

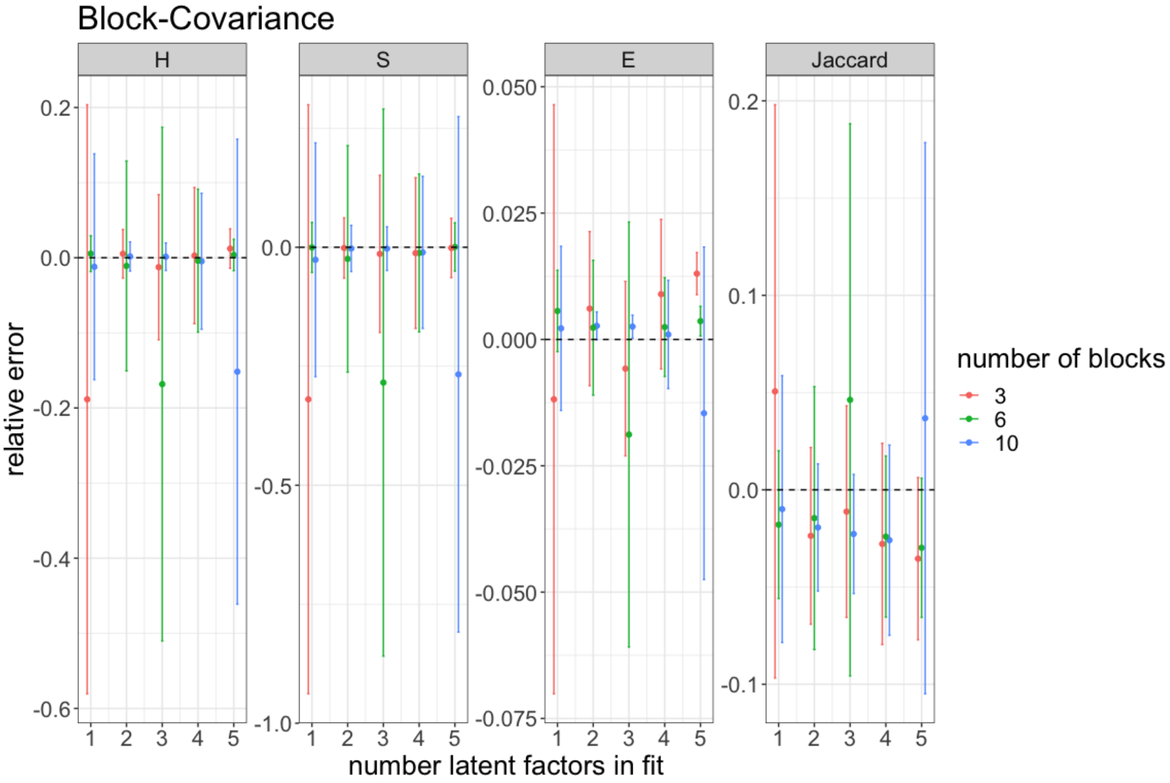


Figure 3.10: Block-Covariance ($K_{sim} >> K_{est}$): When all of the species prevalences are the same, community metrics are mostly unbiased except for evenness in some scenarios.

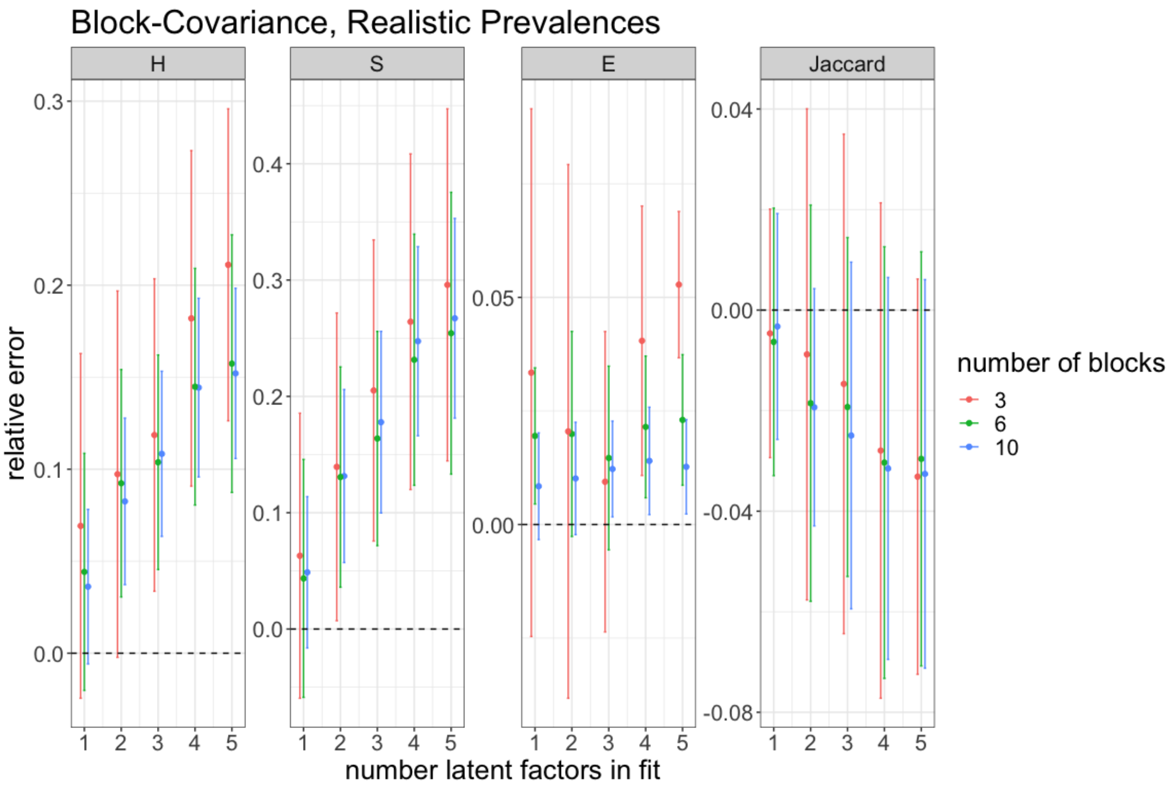


Figure 3.11: Block-Covariance with Realistic Species Prevalences ($K_{sim} >> K_{est}$): When there are realistic species prevalences, performance breaks down, especially for Shannon's entropy and species richness.

Heavy-Tail

In the "correct-K, heavy-tail" scenario, even with very few degrees of freedom for the t-distributed random effects, estimates of H and S are robust to this type of model misspecification (see Figure 3.12) although estimation variance across replicates depends on the ratio of sites to number of parameters estimated. Evenness E and the Jaccard similarity estimates don't always contain zero relative bias across replicates with larger K_{sim}. The negative biases for these metrics indicate that the the truth is more even across sites than predicted and is more similar across sites than predicted.

In the "correct-K, heavy-tail" scenario with realistic species prevalences results are slightly more mixed (see Figure 3.13). Evenness and Jaccard similarity often have estimates that cover zero relative bias across replicates while H and S are sometimes overestimated, especially for larger K_{sim}.

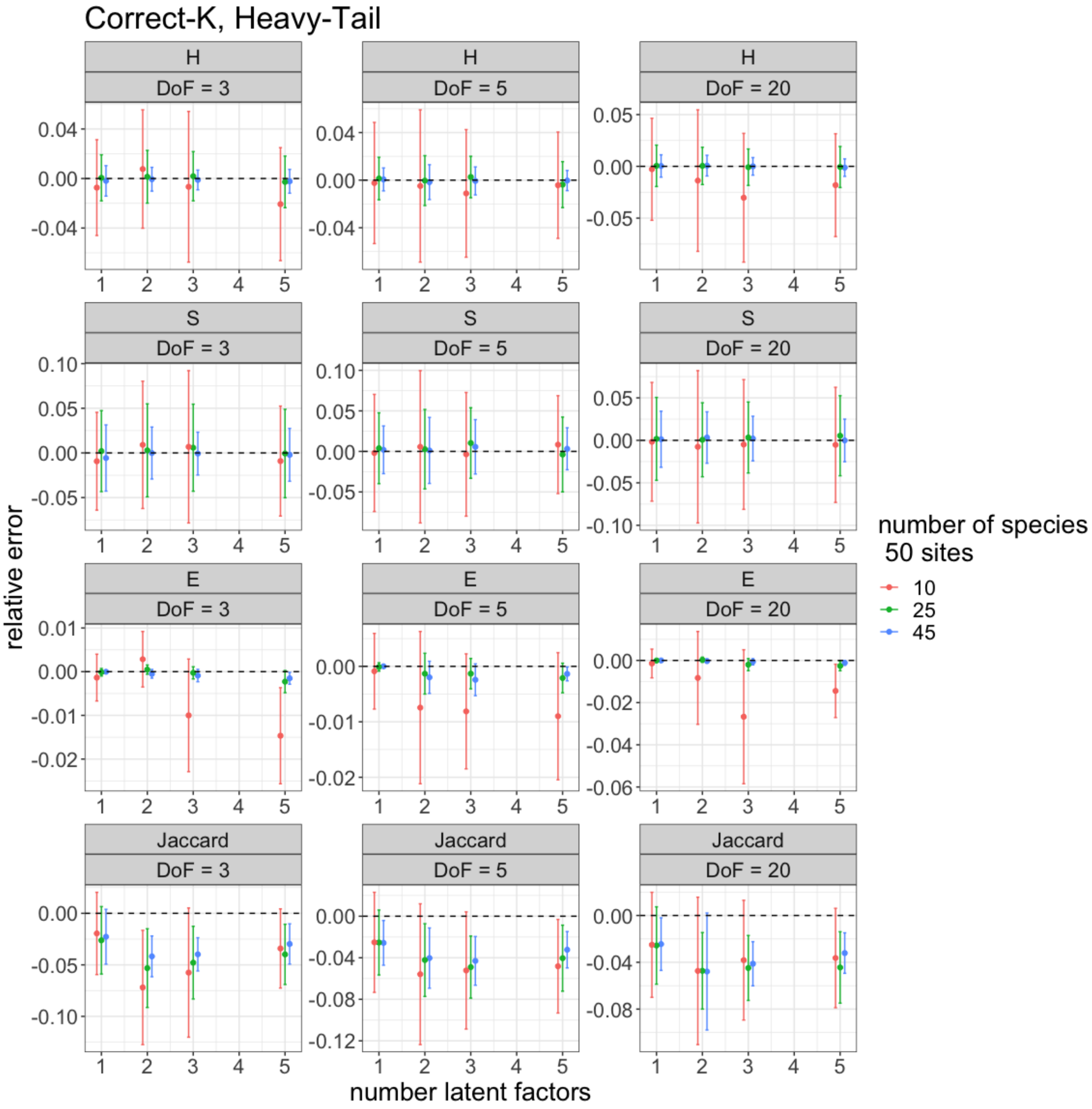


Figure 3.12: Correct-K, Heavy-Tail ($K_{est} = K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): Most scenarios have estimates of community metrics that are robust to heavy tailed random effects. Variability across replicates notably decreases as the number of species coefficients to fit increases for a given number of sites.

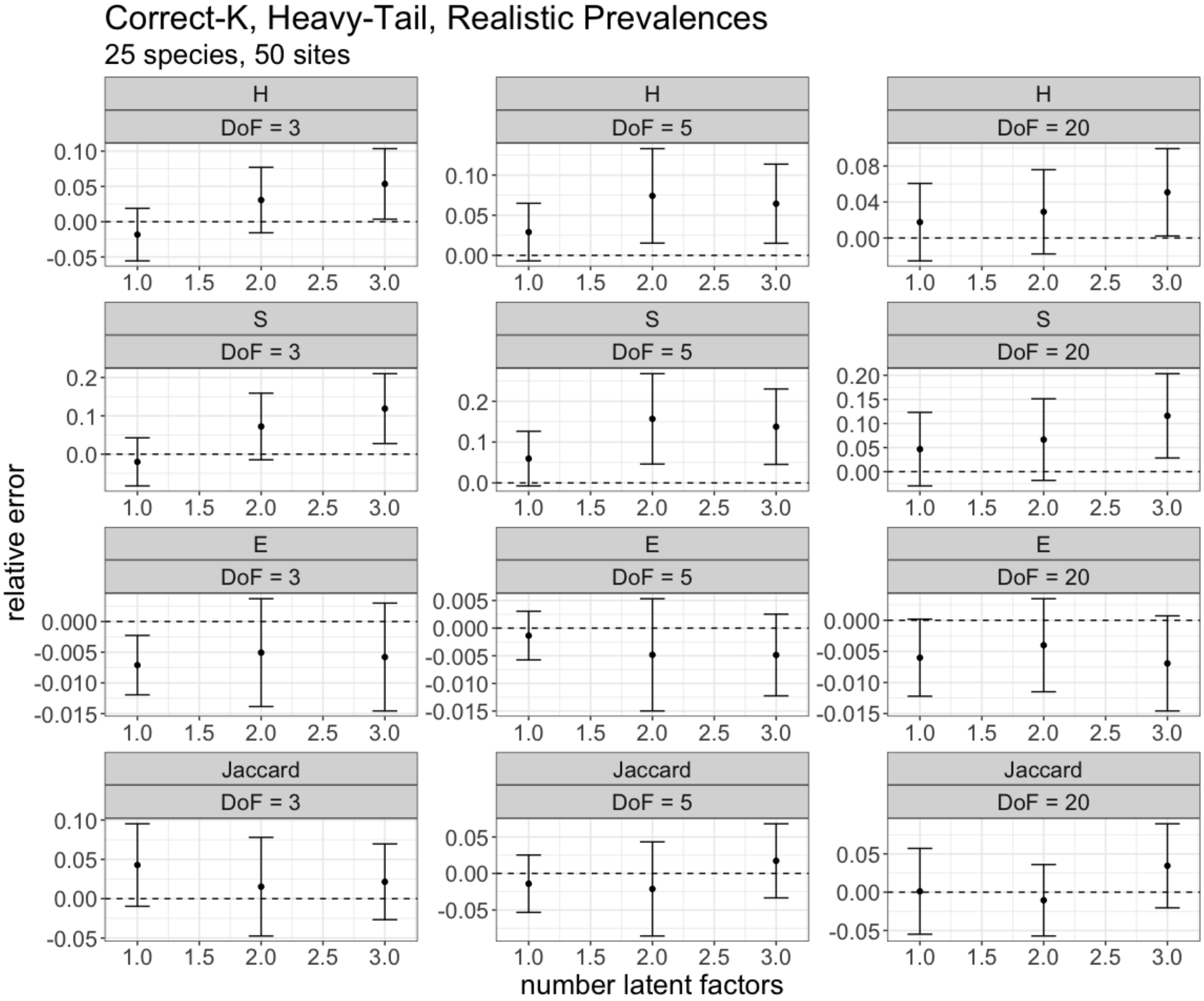


Figure 3.13: Correct-K, Heavy-Tail with Realistic Species Prevalences ($K_{est} = K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): Evenness and Jaccard similarity are robust to heavy tailed random effects. Shannon's Entropy and species richness are sometimes overestimated, especially for larger K_{sim}.

In the "wrong-K, heavy-tail" scenario those metrics that were robust in the "correct-K, heavy-tail" scenario remain robust to differing K_{est} while metrics that tended to be biased in the "correct-K, heavy-tail" scenario remain biased for differing number of latent factors (see Figures 3.14- 3.17). The same is true when realistic prevalences occur. All but the Jaccard similarity result in biased estimates (see Figures 3.18- 3.21).

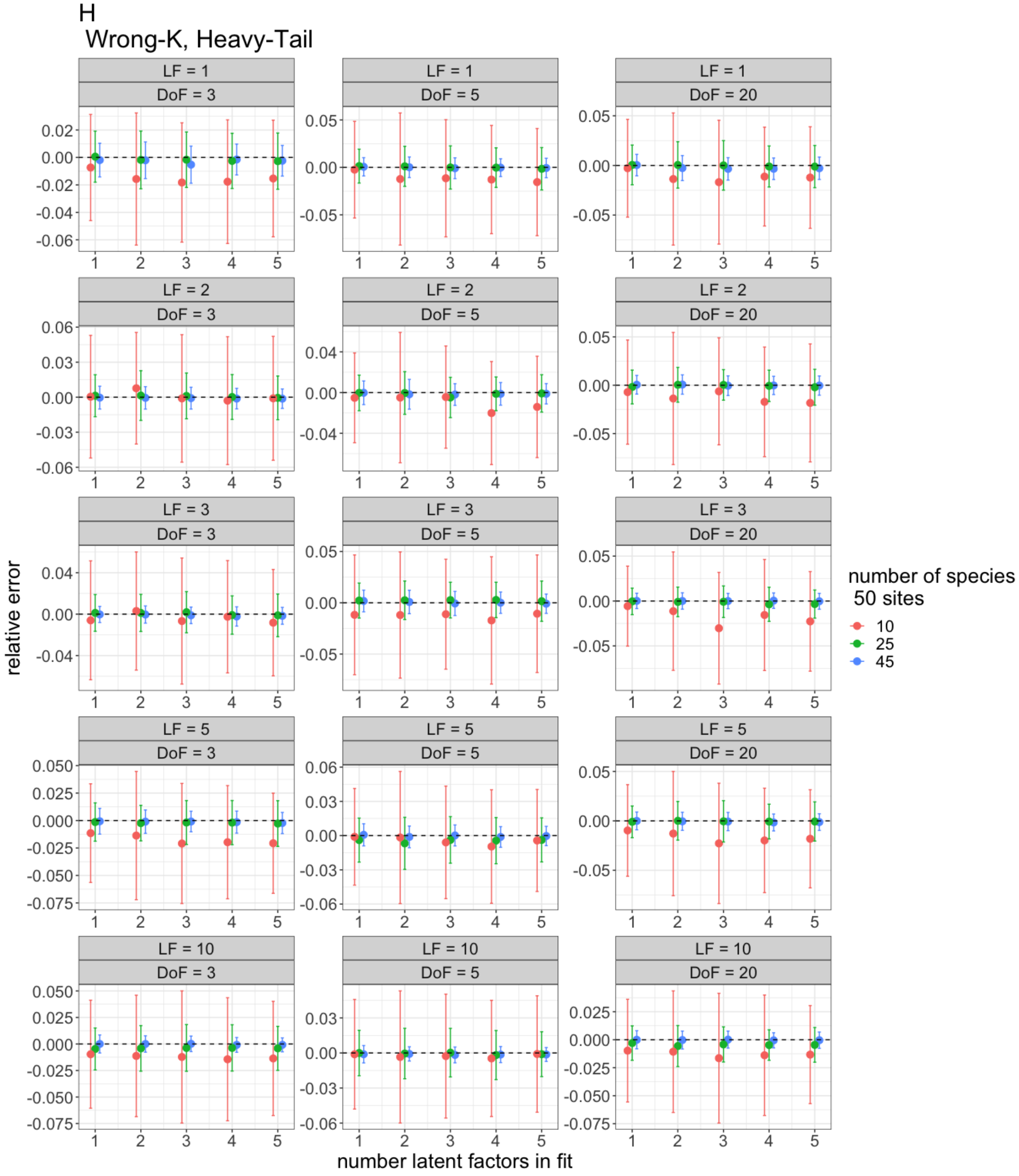


Figure 3.14: Wrong-K, Heavy-Tail ($K_{est} \neq K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): Shannon's Entropy remains robust to the wrong number of latent factors even in the presence of heavy tailed random effects.

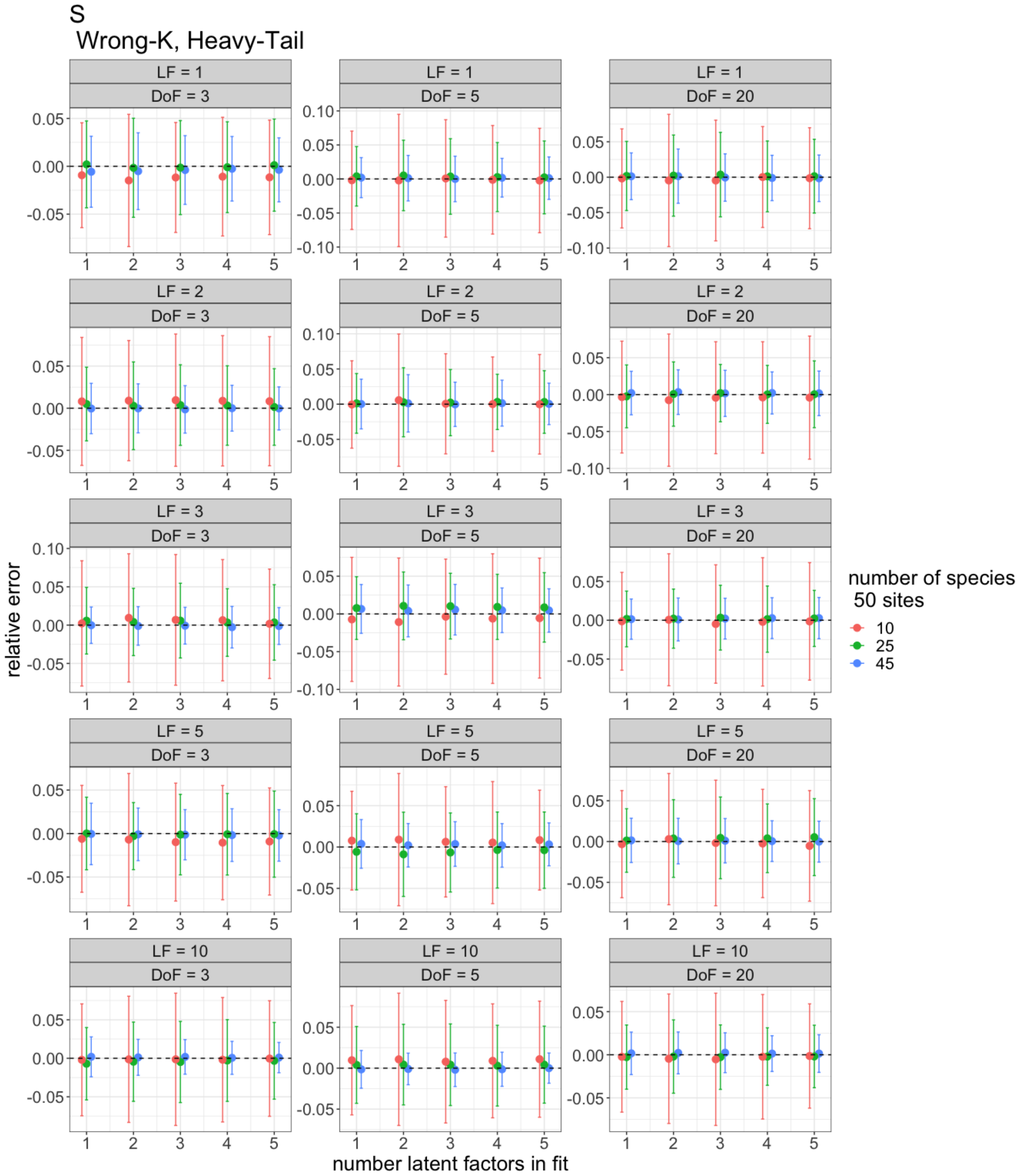


Figure 3.15: Wrong-K, Heavy-Tail ($K_{est} \neq K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0, 1)$): Species richness remains robust to the wrong number of latent factors even in the presence of heavy tailed random effects.

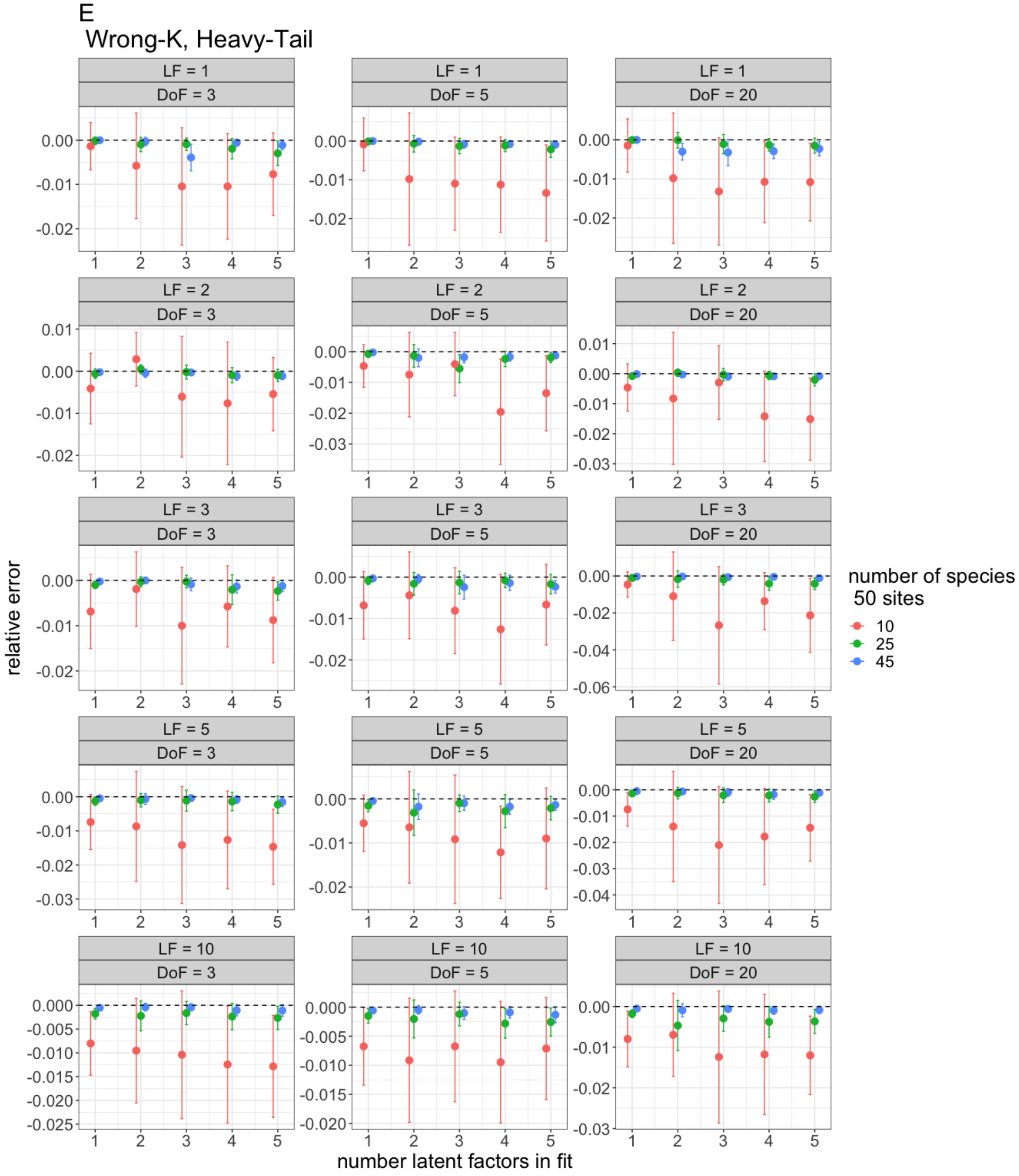


Figure 3.16: Wrong-K, Heavy-Tail ($K_{est} \neq K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): In the presence of heavy tailed random effects evenness maintains its slight biases irregardless of the number of latent factors used in the model fit.

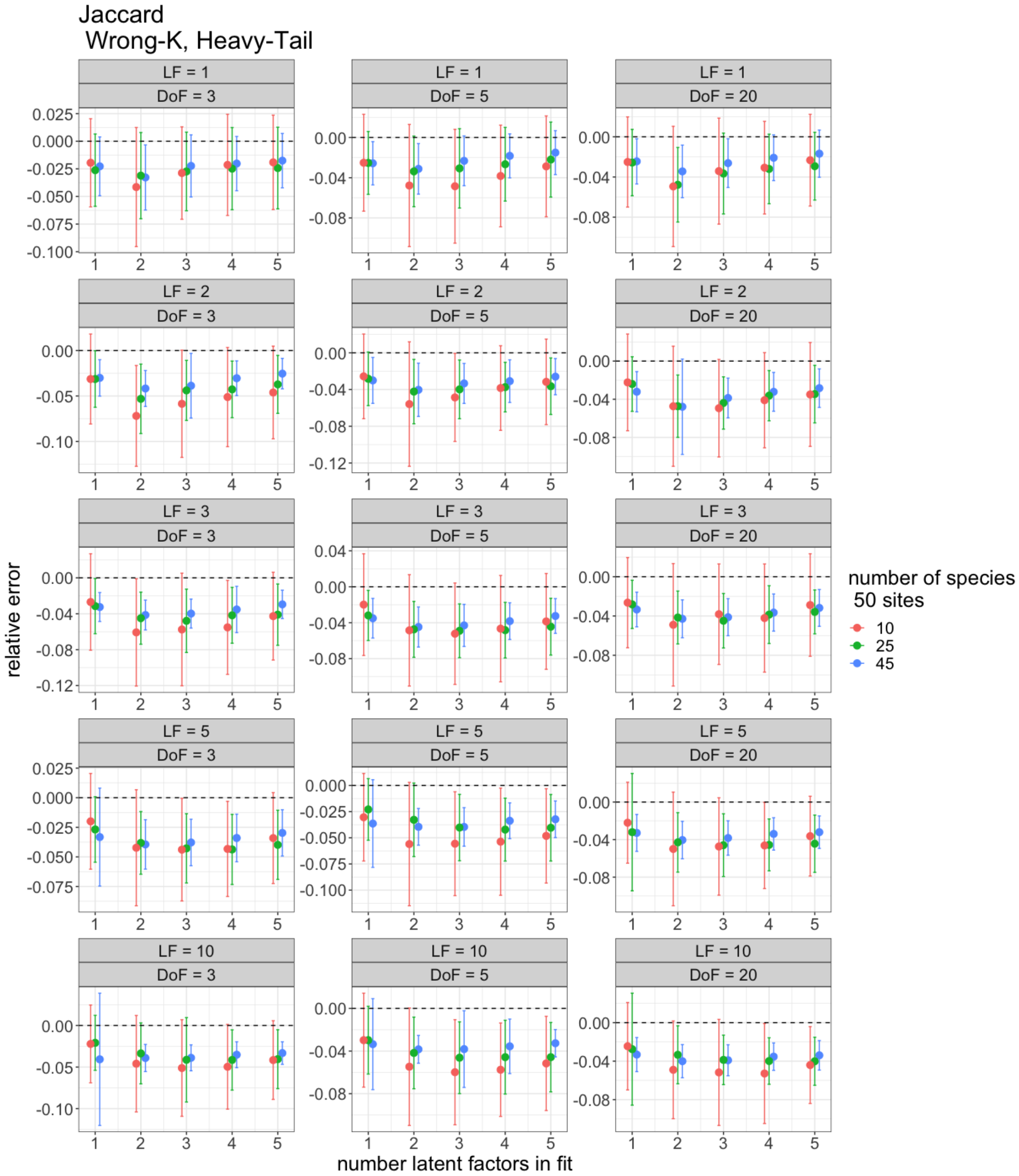


Figure 3.17: Wrong-K, Heavy-Tail ($K_{est} \neq K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): In the presence of heavy tailed random effects the Jaccard similarity maintains its slight biases irregardless of the number of latent factors used in the model fit.

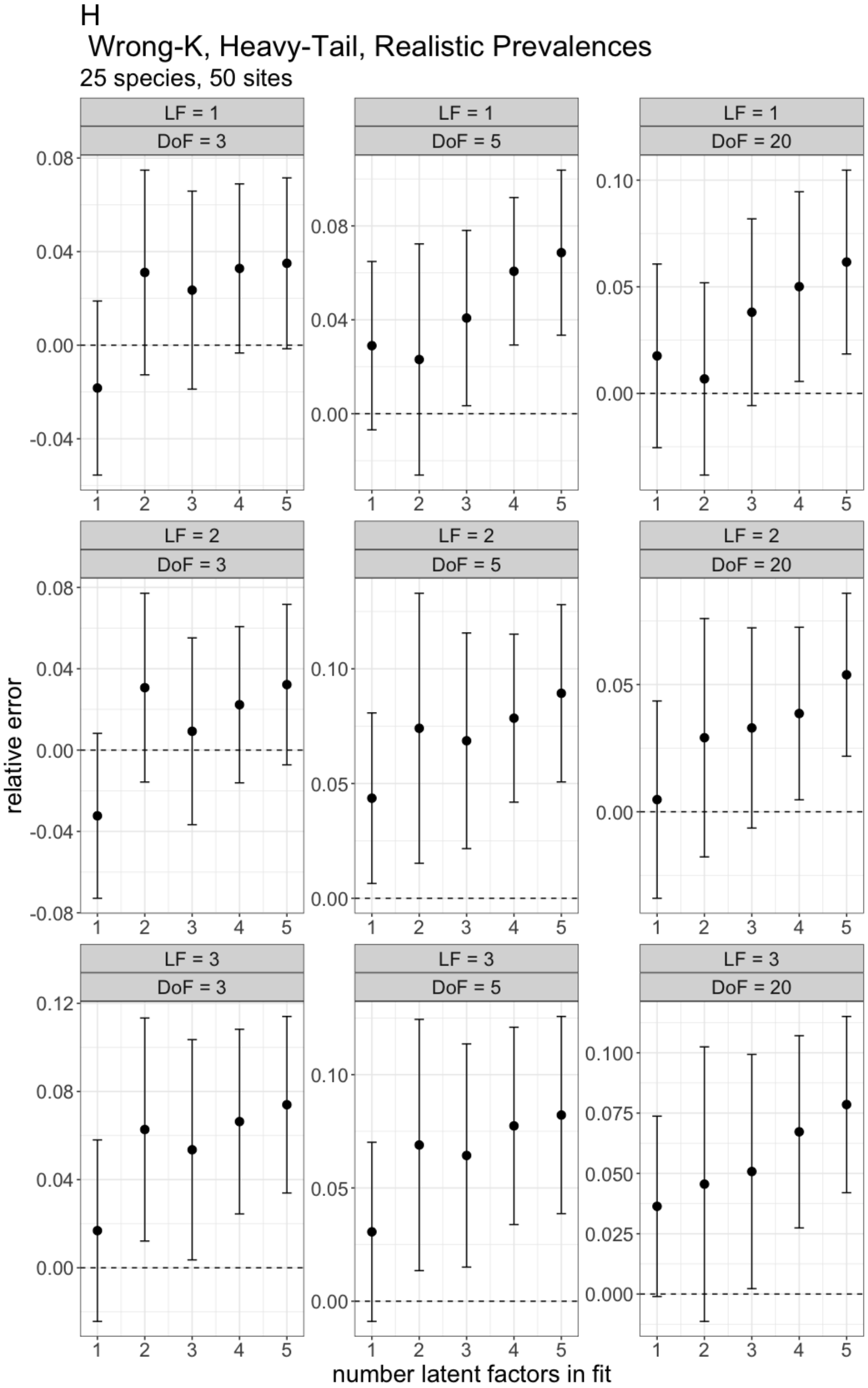


Figure 3.18: Wrong-K, Heavy-Tail with Realistic Species Prevalences ($K_{est} \neq K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): With realistic species prevalences, Shannon's Entropy tends to be overestimated, especially when larger number of latent factors are used.

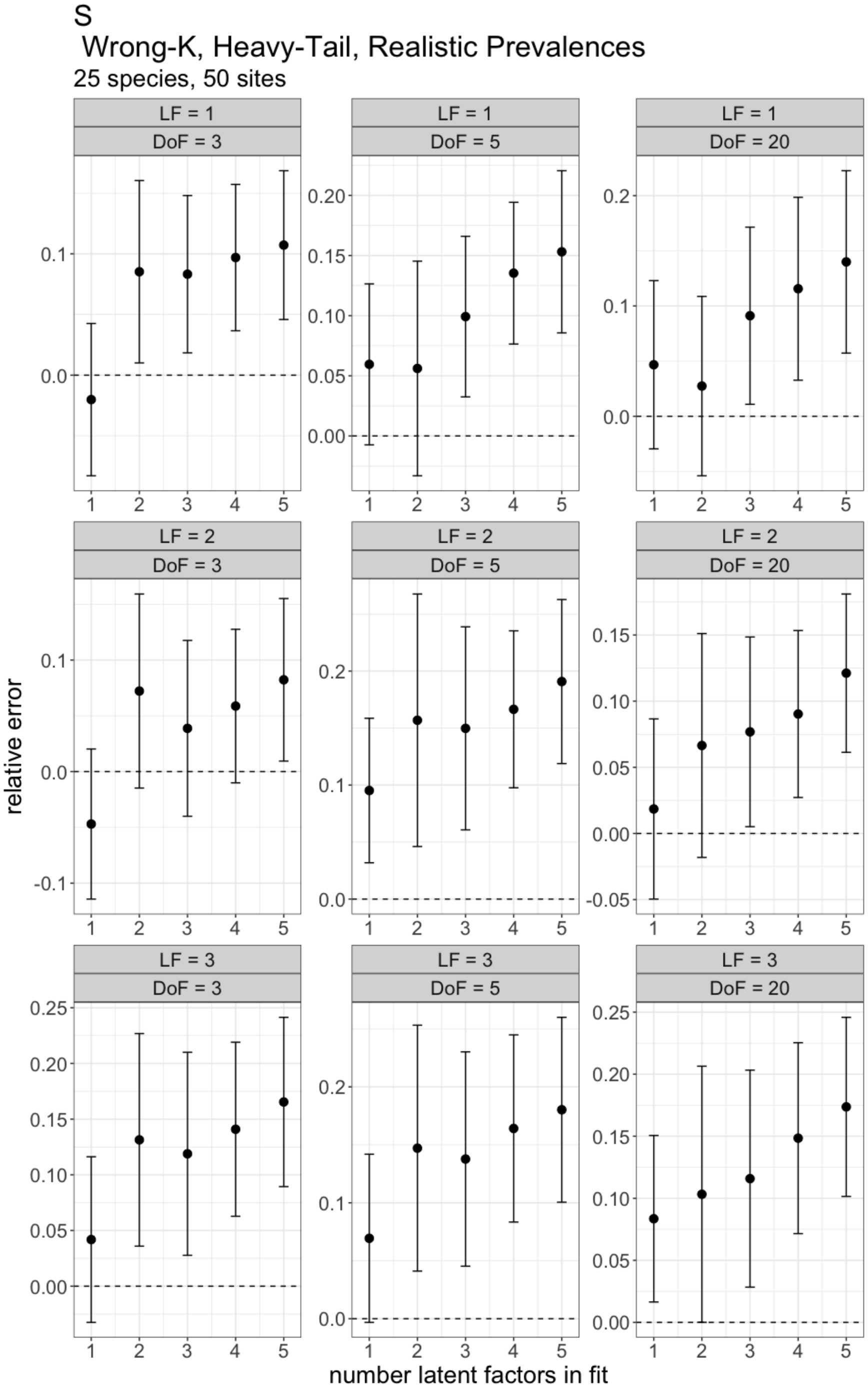


Figure 3.19: Wrong-K, Heavy-Tail with Realistic Species Prevalences ($K_{est} \neq K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): With realistic species prevalences, species richness tends to be overestimated, especially when larger number of latent factors are used.

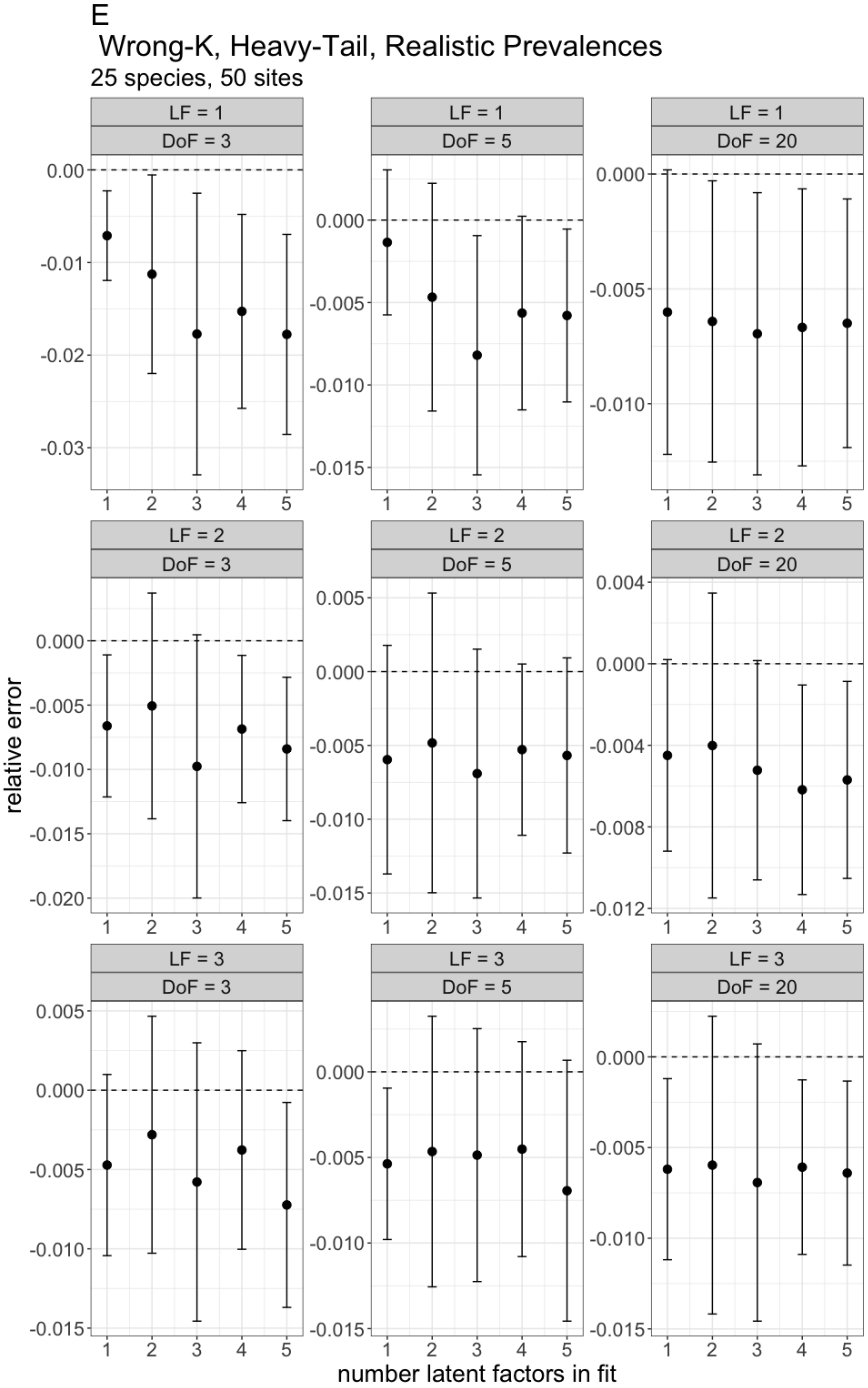


Figure 3.20: Wrong-K, Heavy-Tail with Realistic Species Prevalences ($K_{est} \neq K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): In the presence of heavy-tailed random effects evenness is downward biased.

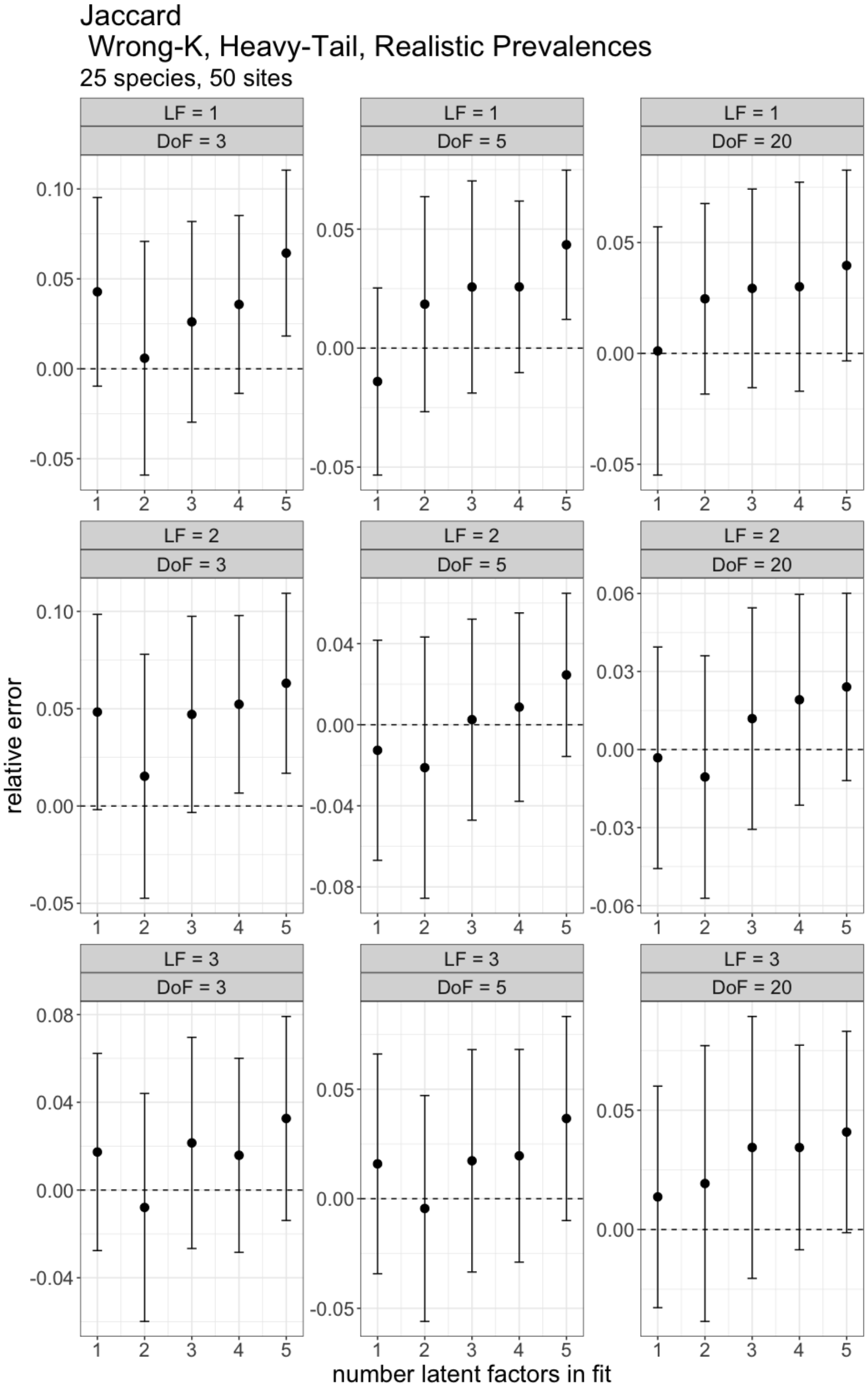


Figure 3.21: Wrong-K, Heavy-Tail with Realistic Species Prevalences ($K_{est} \neq K_{sim}$, $\eta_{ik}^{sim} \sim t$, $\eta_{ik}^{est} \sim N(0,1)$): Jaccard similarity appears robust to differing number of latent factors and heavy-tailed random effects. However, estimates increase as K_{est} increases.

3.4 Discussion

Unpacking Trends in the Results

We found that the Jaccard similarity is often downward biased, even in the "correct-K" case, but this bias does not appear when more realistic species prevalences are involved. To investigate further, we can examine the distribution of the Jaccard similarity metric across site pairs in a subset of scenarios instead of relying on the average.

Figure 3.22 shows the empirical cumulative distribution functions of the Jaccard similarity for a correctly specified scenario (where $K_{sim} = K_{est} = 1$) with and without realistic species prevalences. The black line shows the observed distribution of the Jaccard similarity in one replicate. The grey lines show the Jaccard similarity from other replicates within the same simulated model scenario. The red lines show predicted Jaccard similarity distributions generated from the model fitted to the data (whose Jaccard similarity distribution was plotted in black).

The left plot shows a scenario where the Jaccard similarity is underestimated. We can see that the entire predicted distribution is shifted to the left; the result is not an artifact of the choice of summarizing the Jaccard similarity distribution by its mean. The right plot shows a scenario where the estimate of Jaccard similarity is unbiased on average. When species prevalences are realistic, the entire predicted distribution maps more closely onto the true one.

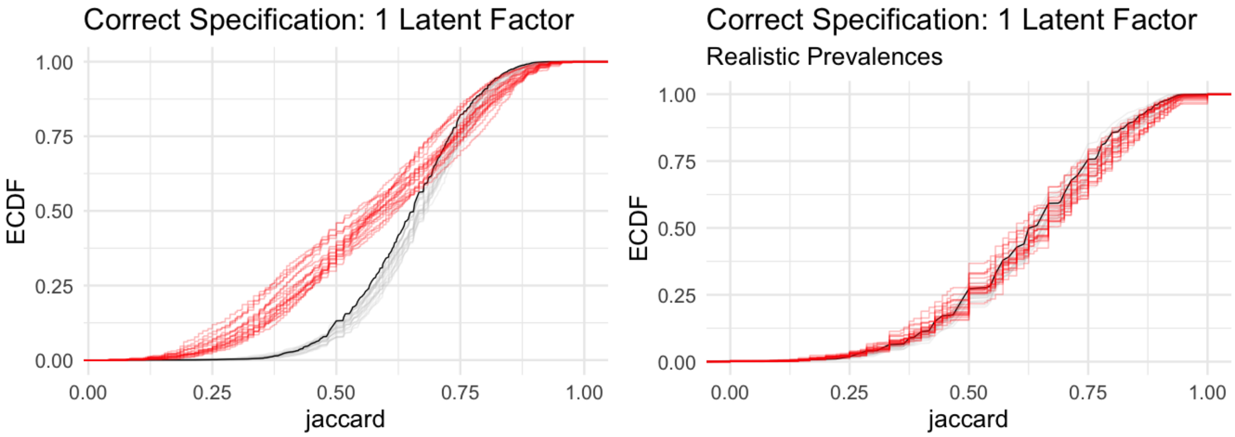


Figure 3.22: These plots show the empirical cumulative distribution function of the Jaccard similarity across pairs of sites. The black line shows the observed distribution of the Jaccard similarity in one replicate. The grey lines show the Jaccard similarity from other replicates within the same scenario. The red lines show predicted Jaccard similarity distributions generated from the model fitted to the data. The left plot shows a scenario where the Jaccard similarity is underestimated. We can see that the entire predicted distribution is shifted to the left. The right plot shows a scenario where the estimate of Jaccard similarity is unbiased on average.

Understanding why this bias occurs is beyond the scope of this work, but we can offer some speculation. In the left case, estimated prevalences vary more than the true prevalences (which are all the same). This creates more dissimilarity, shifting the whole distribution of the Jaccard to the left. In the right case, since true prevalences themselves vary already, the extra dissimilarity induced by prediction error is less noticeable. To test this theory and further diagnose the estimation error, future work could include a calculation of the expected Jaccard similarity as a function of species prevalences.

We also found that Shannon's Entropy and species richness were often unbiased across a variety of mis-specification scenarios, but when realistic species prevalences were included, estimates of these metrics tended to be too large. Figure 3.23 shows a similar phenomenon to the Jaccard similarity shown above. We can see that the entire predicted distributions are shifted to the right; the result is not an artifact of the choice of summarizing the Shannon's Entropy and species richness distributions by their means. It is known that species richness is sensitive to rare species and Shannon's Entropy is sensitive to rare and abundant species, so this is not entirely surprising [71, 181]. The ability to estimate evenness well was more variable across mis-specification scenarios. This could just be due to the nature of the metric. Since evenness is a ratio of Shannon's Entropy and species richness, there are a variety of situations that can lead to its poor estimation. However, it is also possible for biases in the estimation of the numerator and the denominator to cancel out, leaving the ratio well estimated.

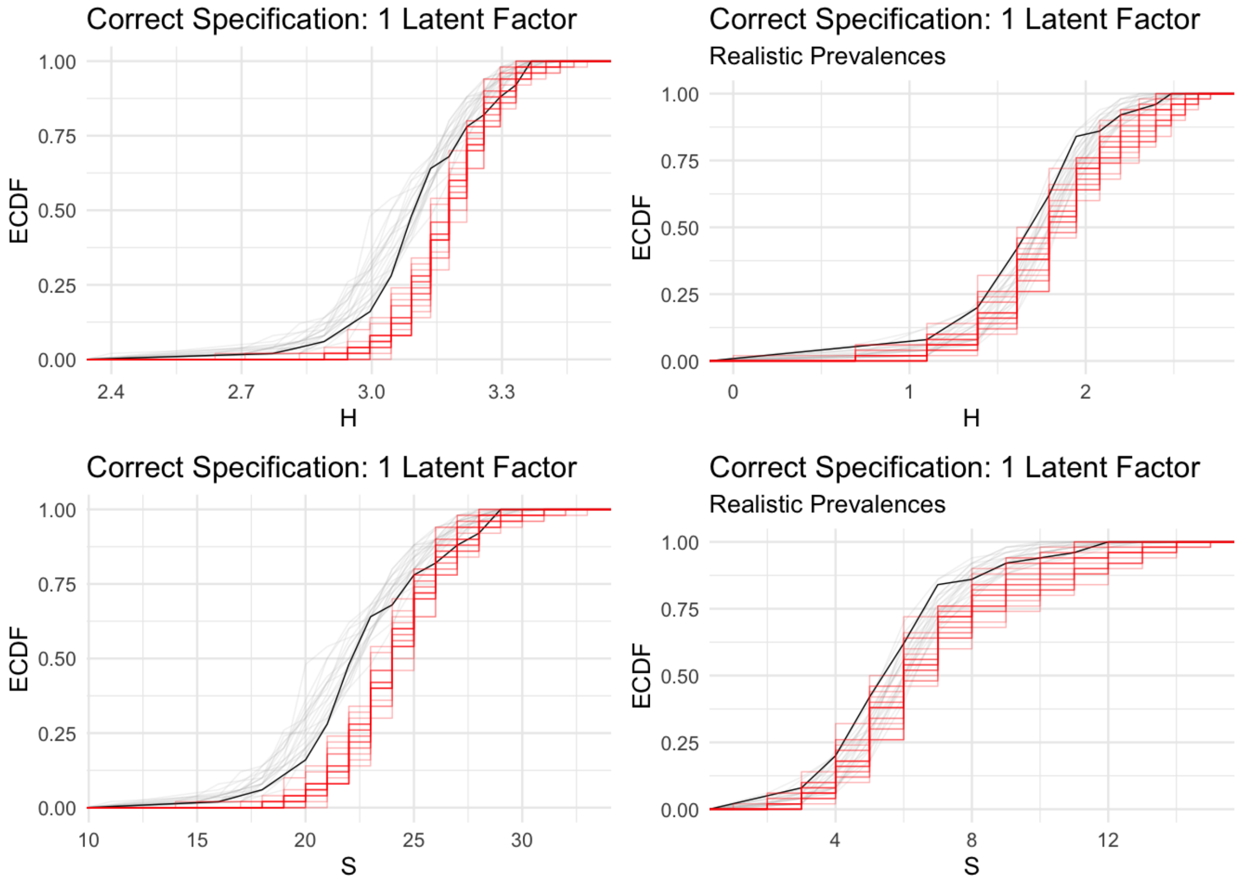


Figure 3.23: These plots show the empirical cumulative distribution function of the Shannon's Entropy and species richness across sites. The black line shows the observed distribution in one replicate. The grey lines show the distribtuions from other replicates within the same scenario. The red lines show predicted distributions generated from the model fitted to the data. The left plots show a scenario where H and S are overestimated. We can see that the entire predicted distribution is shifted to the right. The right plots show a scenario where the estimates are unbiased on average.

A final investigation we might be interested in is deciding whether these patterns are driven by systematic bias in estimation of the species prevalences. For example, are estimates of the prevalences for rare species pulled upward while those of common species are pulled downward? This would be especially problematic if it was happening in the correctly specified cases. Figure 3.24 shows the true v. predicted prevalences for the same scenarios as shown in Figures 3.22 and 3.23. On the left when the true prevalences are all the same (0.5) we do not see any prominent tendency to over- or under-estimate although we see more extreme over-estimates than under-estimates for individual species. On the right when the true prevalences follow a more realistic distribution, we do not see any systematic over- or under-estimation either.

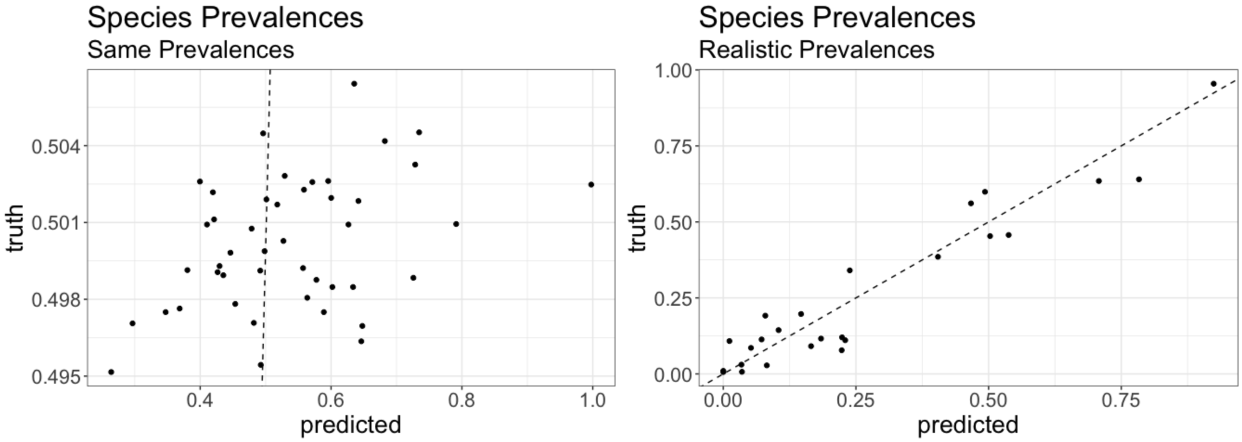


Figure 3.24: For the same situations as displayed in Figures 3.22 and 3.23 we compare the true prevalences with the predicted prevalences to see if systematic over- or under-estimating is occurring. We do not see an obvious pattern in either case.

Understanding why we do not see a more prominent decline in performance when $K_{sim} > K_{est}$ and diminishing returns in performance when $K_{sim} < K_{est}$ is beyond the scope of this work, but this is an interesting avenue for future work.

Mis-specifications Involving Nonlinearities

More complicated mis-specifications could also occur. For example higher order or indirect interactions may not be fully captured by estimating pairwise interactions, interactions between species may be nonlinear, or the interaction of two species could change depending on a third species [12, 116]. Future work could include more complete stress testing of JSDMs against higher order interaction mis-specification, but we can start to get insight about the last case using a toy example here.

Suppose there are three species and three latent factors, η_1, η_2, and $\eta_3 = \eta_1 * \eta_2$ (all are 1×3 vectors). Because the third latent factor is an interaction between the other two, this violates the assumption of the latent factor model that latent factors are unrelated [47]. We craft a scenario (see Table 3.3) where when Species 1 has small prevalence, Species 2 and 3 are positively correlated, but when Species 1 has high prevalence Species 2 and 3 are negatively correlated.

Description	**Notation**	**Value**
Number of Sites	I	200
Number of Species	J	3
First Latent Factor	η_{i1}	$\sim N(0,1)$
Second Latent Factor	η_{i2}	$\sim N(0,1)$
Third Latent Factor	η_{i3}	$= \eta_{i1} * \eta_{i2}$
First Latent Factor Loadings	λ_1	$= \begin{bmatrix} 0 & 1 & 0.5 \end{bmatrix}$
Second Latent Factor Loadings	λ_2	$= \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$
Third Latent Factor Loadings	λ_3	$= \begin{bmatrix} 0 & 0.5 & -1 \end{bmatrix}$

Table 3.3: Interaction Simulation Model: Note we use a large number of sites (200) to better see the signal of this interaction in the observed data.

Figure 3.25 shows the results for this interaction scenario; these results are similar to those of the other mis-specification scenarios. However, this time the worst case scenarios from each metric are combined in one scenario. The Shannon's Entropy and species richness are overestimated (although a relative of error of zero is mostly covered across replicates), the Jaccard similarity is underestimated, and the evenness has mixed results depending on the number of latent factors used in the fit.

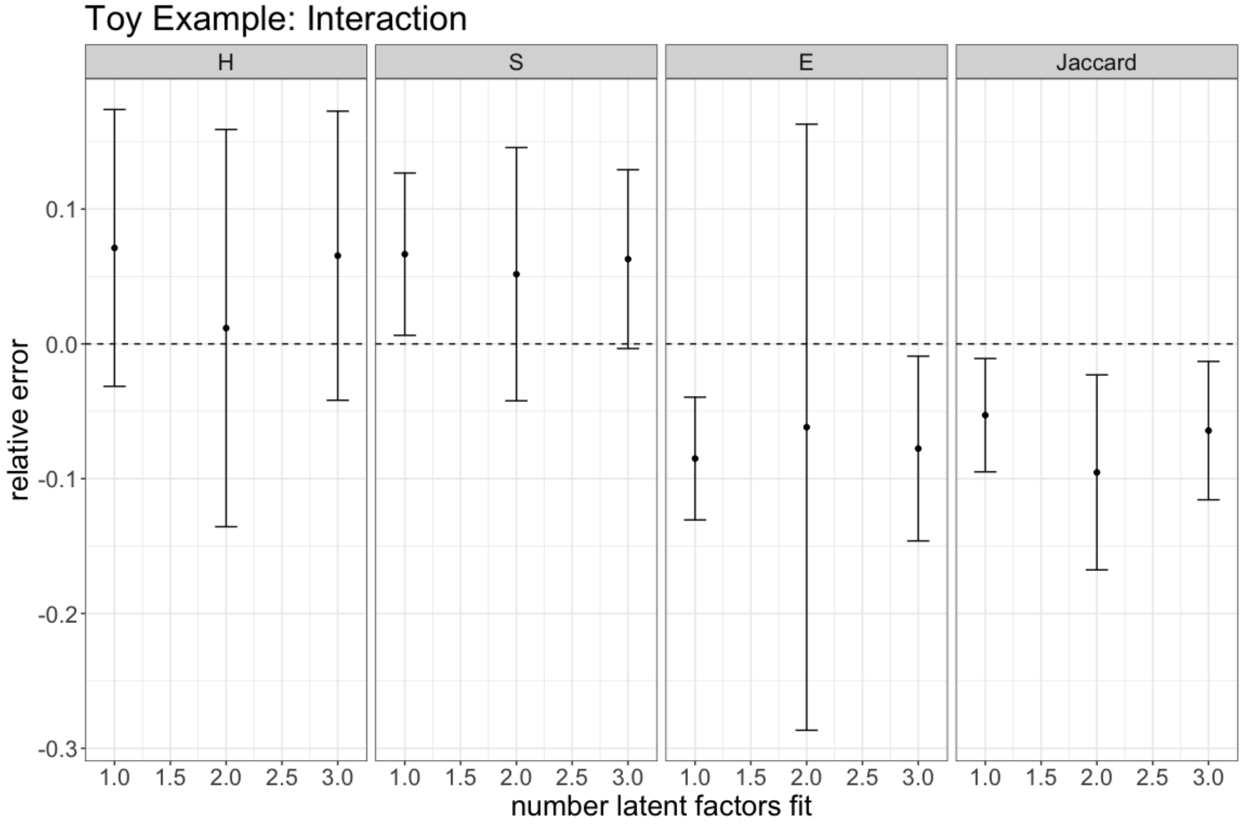


Figure 3.25: Toy Interaction Example: The Shannon's Entropy and species richness are often overestimated, the Jaccard similarity is underestimated, and the evenness has mixed results.

Recall that in this scenario, Species 1 drives the relationship between Species 2 and 3. If Species 1 has small prevalence, Species 2 and 3 are positively correlated, but when Species 1 has high prevalence, Species 2 and 3 are negatively correlated. The left plot of Figure 3.26 shows the number of sites where the occurrence or absence of Species 2 and Species 3 are the same in the presence of Species 1 (x-axis) and in the absence of Species 1 (y-axis) across datasets simulated from this interaction scenario. We expect points to lie mostly above the $y = x$ dashed line. When we make the same plot for predicted occurrences, we see the opposite. Species 2 and 3 are predicted to be more likely to have the same occurrence or absence in the presence of Species 1. This inability to predict the appropriate pairwise relationships between species sheds light on the bias in the community metrics. Future study of more complicated species interactions could build upon this toy example.

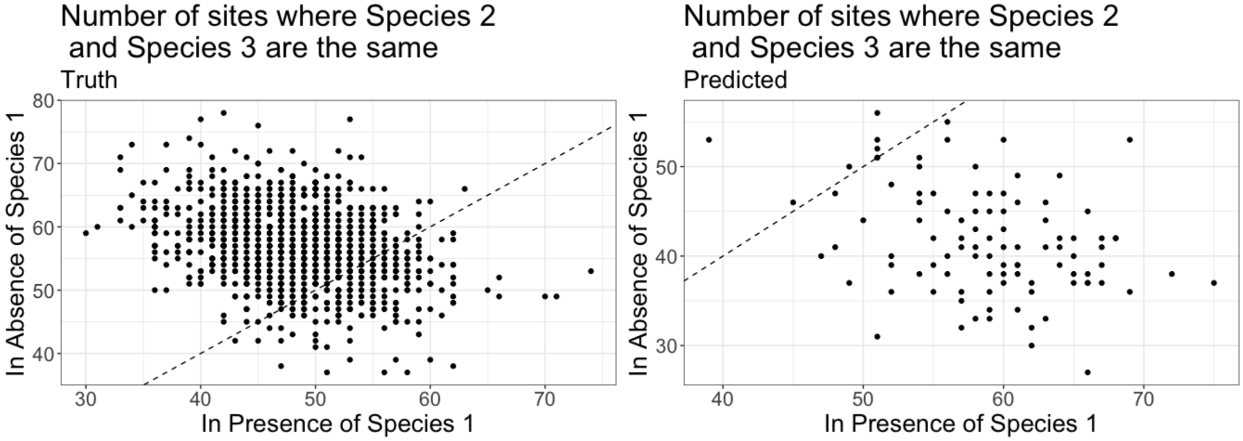


Figure 3.26: Relationship between Species 2 and 3 Depends on Species 1: In the simulation model Species 2 and 3 are positively correlated when Species 1 has low prevalence but are negatively correlated when Species 1 has high prevalence (left plot). The opposite relationship is predicted by the estimated model (right plot).

Although the discussion in this work has focused on occupancy data, future work could similarly stress test abundance models. For abundance, there is an added potential model mis-specification, the distribution of the abundances themselves (Poisson, Negative Binomial, etc.).

Performance of JSDMs and Related Findings

Joint species distribution models have been evaluated in terms of estimation of model parameters in general and parameters related to the covariances between species in particular. This simulation study contributes additional insight into the estimation of community metrics and into how model mis-specification affects estimation.

Throughout the simulation study we see some fairly large error bars for estimation of community metrics. These may be narrowed with more replicates within a scenario, but it should be noted that Wilkinson et al. found that the BORAL approach often had parameter estimates characterized by more uncertainty than other methods they considered [195]. It would be interesting to assess whether an alternative JSDM that gives more precise estimates in general would continue to cover a zero relative bias under model mis-specification.

We also see that cases with a smaller number of species for a given number of sites often have more uncertainty than cases with a large number. Relatedly, Zhang et al. found that a decrease in number of species included in the model decreased the prediction performance although they excluded species based on occurrence probability rather than focusing on the explicit site-to-species parameter ratio [201]. Both of these results may suggest that the information borrowed across species may be more valuable than the extra burden incurred by having to estimate more parameters for an additional species.

When assessing JSDMs for their ability to find interactions between species, Zurell et al. found that the residual correlation estimated by the Pollock et al. model depended on the species prevalences rather than the interaction strength [203, 146]. The Pollock et al. model is similar to the BORAL model (it estimates the full species covariance matrix rather than using latent factors), so this may be connected to our finding that whether or not an estimate of a community metric was unbiased under model mis-specification differed depending on the distribution of species prevalences considered [146]. The relationship between prevalence and species covariances is worthy of further study. Two simple cases to augment the results found in this work and further investigate whether robustness to model mis-specification depends on species prevalence would be: all species have a prevalence smaller than 0.5 and all species have a prevalence larger than 0.5.

Under a different data-generating process Thurman et al. found that the Golding et al. JSDM (just a different implementation of the Pollock et al. model) was better at predicting positive interactions than negative interactions with data generated from an Erdős-Rényi random network (where an interaction between any pair of species in equally likely) [179, 60, 146]. This finding may be tied to poor performance in the toy interaction case which has competing signs of interaction between species. Although the resulting set of covariances between species from the block-covariance scenario would not be likely under the Erdős-Rényi network data-generating process used by Thurman et al., it could still be interesting to try the block-covariance scenario with negative covariances within a block to see if the signs of species covariances affect robustness [179].

Wilkinson et al. also found that the parameter estimates of the BORAL approach were strongly correlated with the results from other JSDMs for both fixed effect coefficients and covariance coefficients [195]. Because of this, the conclusions about estimation of community metrics that follow may also be relevant to other JSDMs whose parameter estimates are similar.

3.5 Conclusion

In this paper we stress tested a commonly used joint species distribution model that is the foundation of a variety of more complicated methods. Our goal was to see how well a latent factor JSDM was able to estimate community metrics in the face of model mis-specification of the assumed covariance relationship between species. These mis-specifications include choosing the wrong number of latent factors to include in the model, wrongly assuming that a small number of latent factors captures the covariances between species, wrongly assuming the latent factor distribution is normal, or wrongly assuming latent factors are unrelated.

When species prevalences are similar to one another, estimates are robust to a variety of model mis-specifications, although the Jaccard similarity is often slightly downward biased even in correctly specified cases. When faced with a realistic distribution of species prevalences, estimates remain fairly robust to choosing the wrong number of latent factors and heavy tailed random effects. In these cases, when biases do occur (as in the overestimation of Shannon's Entropy and species richness), they generally are small in magnitude. These findings may give ecologists more confidence in their estimates if the expected bias would result in a negligible difference in a community metric in practice.

Ecologists who are using JSDMs to help them understand properties of species communities can use the results in this paper to better understand what biases they may be facing in their community metric estimates when they suspect a particular form of model mis-specification. They can also use this information to decide when they feel comfortable using estimated community metrics to make decisions about community conservation and management.

A key contribution of the first two chapters of this dissertation has been bridging the gap between statisticians and ecologists which requires a conscious navigation of the differing styles and background knowledge of researchers in both fields. In both chapters we reveal insight into challenging problems in ecology through the use of statistical concepts and tools. This effort requires clear communication with ecologists. My dissertation culminates in a discussion of how to train statisticians to talk about their work amongst peers, with interdisciplinary collaborators, and to a wider public.

Part III

Communicating with Data: The Art of Writing for Data Science

Chapter 4

Teaching the Art of Writing for Data Science

Book excerpts co-authored with Deb Nolan

4.1 Pedagogical Context for "Communicating with Data"

The American Statistical Association updated their curriculum guidelines for undergraduate statistics classes in 2014 [26]. One of the key points referenced in the executive summary is the "ability to communicate" to facilitate "thinking with data." This recommendation includes the goal of teaching students to be able to "communicate complex statistical methods in basic terms to managers and other audiences and to visualize results in an accessible manner" and a mandate that "programs should provide multiple opportunities to practice and refine these statistical practice skills." The guidelines go on to define communication goals to include "effective technical writing, presentation skills, visualizations" and an "ability to interact with and communicate with a variety of clients and collaborators." Pedagogical advice includes offering regular opportunities to practice communication skills that are tied to the more technical aspects of statistics instruction.

Similarly, in 2016 the American Statistical Association released the updated Guidelines for Assessment and Instruction in Statistics Education (GAISE) [25]. In these guidelines, communication is particularly emphasized as a core competency for presentations and projects including visual presentations and verbal communication of findings in addition to traditional written communication.

Typical undergraduate statistics courses involve a final project that is accompanied by a written report and/or an oral presentation. Increasingly, courses dedicated specifically to communication of statistics, interpreted broadly, are appearing in undergraduate curricula. The inspiration for this dissertation chapter indeed occurred through my involvement in co-developing and co-teaching such a course in the Statistics Department of UC Berkeley with Professor Deborah Nolan through the Art of Writing Program. This course was an undergraduate seminar capped at 15 students with only one introductory statistics class required as a prerequisite. Through our experience developing and teaching the course, we recognized the gap in resources for instructors who want to teach such a statistical communication course.

There are books that focus on science writing such as “The Craft of Scientific Writing” by Alley [1], “A Field Guide for Science Writers” by Blum, Knudson, and Marantz Henig [14], “Writing Science: How to write papers that get cited and proposals that get funded” by Schimel [167] and “The Scientist’s Guide to Writing: How to write more easily and effectively throughout your scientific career” by Heard [70] that we referenced as we prepared the course. However, the communication of statistics differs from the communication of more general science (e.g. words like “significance” carry more weight), and many of these science writing resources are focused at a more advanced level for graduate students learning how to write formal research articles which can be inaccessible for an undergraduate audience. Our manuscript aims to be practical and accessible with its variety of examples and exercises and useful to those presenting results from data with its in-depth treatment of graphs, coding, data descriptions, and statistical terminology.

In framing our approach to teaching communication of statistics and research involving data, we took inspiration from both the science communication literature and the general teaching pedagogy literature.

Burns et al. give an overview of the different levels of science communication [22]. The public can respond to science by being aware of science, enjoying science, being interested in science, having an opinion about science, and finally understanding science. They mention Jesse Shore’s approach to designing exhibits in science museums. Shore aims to both “attract and involve visitors who are uninformed or disinterested in the overall subject” and “maintain interest of those who are informed (interested public) or even specialists (attentive public).” Relatedly, as part of our Communicating with Data class we visited the Phoebe A. Hearst Museum of Anthropology to learn about how museum curators write the museum placards. We connected the writing of placards that describe each part of the exhibit to writing captions that describe each visualization in a statistical report.

Bubela et al. distinguish the more passive deficit model of communicating with the public from the more interactive public engagement model [21]. In the former, miscommunication of science is blamed on the public’s lack of knowledge while the latter puts the onus on scientists to create a dialogue with the public. They advocate for teaching graduate students how to communicate with the media and a variety of audiences since they are the future of science. We agree and expand this focus to undergraduate students. We incorporate instruction and activities for writing press releases throughout the book to introduce students to how journalists often consume science. We also emphasize the need to make our writing accessible and stress narrative as a way to engage readers (Chapter 6).

In a review of the writing literature, Raimes notes a variety of instructional activities that we leveraged in our own classroom and advocate for in our textbook [151]. These include using journals to brainstorm ideas (Activity 10.8.3), making connections between writing and other language skills such as reading (Chapters 1 and 2) and speaking (Sections 11.1.1-11.1.3), writing collaboratively (classroom activities accompanying Chapters 6 and 7), responding to peer and instructor writing (Activity 10.8.2), using revision strategies (Section 10.2), and separating high-level revision from low-level editing (Chapters 8 and 9).

Beyond writing formally, informal styles of writing are also important for students to learn. In an overview of new tools for teaching writing, Warschauer highlights blogs and Wikis as ways to help students ease into writing and developing a voice beyond a stilted academic style [189]. Ben-Zvi also highlights Wikis as a tool for teaching collaborative writing [10]. We incorporate instruction and activities for writing blogs throughout the book, and Wikipedia activities appear in Chapters 8, 9, and 11.

The following section includes the Preface of our book which further outlines our objectives, intended audience, and structure of the book. Then, to introduce each part of the book, I add more specific pedagogical context, referencing relevant chapters, sections, and activities in the full version of our book.

4.2 Book Preface

Communication is a critical yet often overlooked stage in the data science pipeline. Nolan and my book aims to help students and researchers write about their data insights in a way that is both compelling and faithful to the data. We address writing challenges specific to scientific investigations of data, such as how to describe data succinctly, create effective visualizations, write clean code, and accurately summarize statistical findings. We also provide more general advice on science writing, including how to distill findings into a story, how to organize and revise the story, and how to write clearly, concisely, and precisely.

Our Objective

In our experience, university training in writing rarely addresses the challenges associated with technical writing. Our students lack opportunities to practice writing about their data-analytic processes and to learn from examples of good, domain-specific writing. To compound this problem, instructors of science courses typically have little experience in teaching technical writing. Many of us find it difficult to give students advice when we have only our personal experience with writing to draw from. In this book we attempt to address both the teaching and learning challenges relevant to communicating the story behind a data analysis.

Intended Audience

We aim for this book to be a resource for students who want to learn how to write about scientific findings where the focus is on presenting the results of a data analysis. Instructors teaching a course in science communication can use it as a textbook, and others teaching a science course that has a writing component can use the book as a supplement. In addition, a researcher who is looking for help writing can use this book to self-train. Practicing statisticians, data scientists, or scientists who need assistance with writing about their data analysis findings will hopefully find guidance they can use to practice their communication skills in the context of their own work.

The only prerequisite is a knowledge of statistics at the introductory level. While we expect the reader to have at least a rudimentary understanding of statistics, the principles of communication found on these pages carry over to writing about more complex data analyses.

Examples and Activities

Each chapter includes many examples and concludes with a collection of activities for practice writing. These examples and activities come from several scientific fields and a broad variety of publications. The main sources are scientific journals, but the advice is equally relevant to writing a report for a supervisor, a paper for an instructor, or an article for a popular magazine. To this point, we use the terms *article*, *paper*, and *report* interchangeably throughout. Additionally, many of the activities at the end of chapters give practice for those who want to write for a broader audience, such as a blog post or press release. We have honed the examples and activities in this book to focus on the essentials of writing about data and, at the same time, we have attempted to create scenarios that allow for individual creativity. Also, we use samples from our own writing and anonymized student work for examples of what not to do.

Book Organization

Our book consists of five parts. Part I aims to help the novice learn how to write by reading the work of others. We identify the main components of a data analysis, examine the argument, and point out how components of an analysis are organized into a story and written for a technical article. In addition, we read and examine material written for broader audiences, e.g., press releases and blog posts. Part II delves into the specifics of how to describe data at a level appropriate for publication, create informative and effective visualizations that support the main findings, and communicate an analysis pipeline through well-written, reproducible code. Part III demonstrates how to distill a data analysis into a compelling story and organize and write the first draft of a technical paper. Part IV addresses revision; this includes advice on writing about statistical findings in a clear and accurate way, general writing advice, and strategies for proof reading and revising. Part V gives advice about communication strategies beyond the page, which includes giving talks, building a professional network, and participating in online communities. This part also contains over twenty portfolio assignments that are aimed at building upon the guidance and examples in the earlier parts of the book and providing continued writing practice.

In addition to the book, we plan to provide additional materials online to use in a course. These will include a detailed week-by-week syllabus that describes the topics covered, in-class activities, assignments, and additional reading used in the course that inspired this book. Along with the syllabus, we will provide pointers to avoid potential problems with some classroom activities and ideas for grading written work. We also will give ideas for how to use this book as a supplemental text in a science course or as the main text for a large course in technical writing where the student work is more limited in scope.

In this dissertation chapter I provide excerpts from each part of the book that illustrate the pedagogical novelty while providing some narration to fill in the context of each.

Acknowledgements
This book grew out of a course on technical writing that we developed and first taught in the fall of 2017. We are grateful to the *Art of Writing Program* at UC Berkeley for providing a grant that enabled us to design and co-teach the course. We also owe a debt of gratitude to our students who willingly worked with us as we class-tested and developed the activities, assignments, and course materials. Their respect for each other, good humor, and diligence made the course a rewarding experience for all of us. Thanks also go to those who read earlier versions of this manuscript and tried it out with their students. In particular, we thank Ben Baumer, Beth Chance, Kelly McConville, Amelia McNamara, Pamela Matheson, and Susan Wang.

The *Art of Writing Program* also supported our creation of this book by organizing a workshop where participants read and commented on all aspects of the book. We are much indebted to the workshop participants: Andrew Bray, Kathleen Donegan, Nicholas Horton, Ramona Naddaff, and Evan Variano, for their thoughtful and constructive comments that led to a much improved manuscript.

4.3 Part I: Reading to Write

Pedagogical Context for Reading to Write

We start our book by teaching students how to "read as writers" so they can learn from other writers. Examples of how others structure their writing and choose content in both formal (Chapter 1) and informal (Chapter 2) contexts can act as concrete models of writing to aspire to.

This "reading to write" approach is often used at the graduate level. For example, Matarese outlines a graduate level course where students write a research paper along with reading and discussing papers in their discipline [112]. This process helped students "develop a framework of knowledge that helps them assess the effectiveness of their own writing." Matarese also makes the point that researchers read the literature to understand what is at the cutting edge of their field, so it seems natural that they should read the literature to see what is at the cutting edge of communication in their field.

Similarly, Parke describes a statistics communication course for graduate students in a different discipline (education) who are learning statistics [134]. In the course, students read journal articles to see how results are reported in text, tables, and plots. Since the students are relatively new to the statistical content, reading articles where statistics methods and terms are used helps them recognize conventions and common usage of statistical concepts.

A "reading to write" approach can work at the undergraduate level, but it can be harder to find examples that are accessible yet still representative of what kind of writing they are aiming for. The strength of our book is that we have curated a series of examples that can be digested by undergraduate students, even if they do not have a lot of experience with statistics. Beyond the content, we also echo Matarese about teaching students that researchers rarely read (or write) linearly. This can be surprising to students used to reading and writing from the beginning to the end.

Reading Science Articles

In this chapter we walk through a variety of scientific articles and provide guidance for how to read from the perspective of a science writer.

We use the following articles from a variety of scientific fields and written at a variety of levels of formality as re-occurring examples to identify the components of a scientific article.

- *Do the Golden State Warriors Have Hot Hands?* by Daks, Desai, and Goldberg [34], published in *Mathematical Intelligencer* which publishes articles for a general audience, written in an informal style

- *Evaluation of the accuracy, consistency, and stability of measurements of the Planck constant used in the redefinition of the international system of units* by Possolo et al. [149], published in *Metrologia* which publishes papers on measurement problems in physics

- *The longest period transiting planet candidate from* $K2$ by Giles et al. [59], published in *Astronomy & Astrophysics* which is an open access journal that publishes on a variety of topics in astronomy and astrophysics

Excerpt: *Main Components of a Scientific Article*
Effectively learning how to write about data involves a strategy for reading an article and examining how the author organizes and writes about their findings. When we read to write, we identify the main components of an analysis and notice how the author brings these components together to form a logical and compelling story. During this process, we discover examples and templates that we can use to organize our own work and write about our findings. To get started, the following three steps can be helpful.

- *Identify the elements of the data analysis.* We begin by looking throughout the article for various building blocks of the data analysis. We often find that some of these elements are included in the article, while others are not. This investigation helps us understand the choices that an author makes in writing about their data analysis, and as a reader, we assess whether particular omissions impact the credibility of the conclusions or whether any included details are superfluous to the main story.

- *Examine the argument.* When we read an article, we expect to be convinced of the importance and validity of the findings. We look for context that explains how the findings fit with others' work, and we try to discern whether the new insights support, counter, or extend current views in the field. We consider the appropriateness of the analysis, the generalizability of the conclusions drawn, and whether others' work have been adequately and convincingly presented. To help make this assessment, we pay close attention to the words the writer uses and consider how an alternative, similar word choice could impact the strength of a claim. We also pay attention to whether the article's tone adds credibility and whether the connections between paragraphs and between sections convey a cohesive story.

- *Map the organization of the document.* At a basic level, science articles have three main parts – a beginning, which serves to define and motivate the problem; a middle that presents the findings and explains what they mean; and an end, which summarizes the conclusions and their importance. Mapping out these parts of the document helps us see how the author chooses to organize their analysis, gives a sense of what a reader might expect when reading an article, and provides templates that we might follow.

■

The chapter concludes with an extended example, going through a template for identifying the elements of the data analysis, examining the argument, and mapping the organization of *How to weigh a donkey in the Kenyan countryside* by Milner and Rougier [117], published in *Significance* magazine, a venue for accessible data analyses.

Reading Materials for Broader Publics

This chapter extends the reading guidelines presented in the previous chapter to written work for broader audiences. We focus on press releases and blog posts as examples. We can still identify the elements, examine the argument, and map the organization for both of these less formal mediums.

Successful press releases often follow a common template of what information to include, how to include it, and where to present each piece.

Excerpt: *Identify the Elements – The Five Ws and H*
News stories in the US tend to follow a template to identify the basic elements of the story; only some of these story elements match the statistical elements of the data analysis from the previous chapter. A press release typically begins with answers to five questions, referred to as the five 'Ws': Who? When? Where? What? and Why? However, when we read a news story about a data finding, it can be confusing to answer these questions because some pertain to the data and others to the researcher. To help, we have expanded these five questions in Table 4.1 to specify whether the question refers to the investigator or the study. There we see that the 'what' and 'why' questions refer to the findings from the data analysis, but the other three are about the investigators. When a scientific study is the topic of the press release, then the 'who', 'when', and 'where' of the data are also relevant. These are described in Table 4.2. All together we look for the answers to the combined set of eight Ws.

The eight Ws for the investigators and the data often appear at the beginning of the press release in the introductory paragraphs. The answers to these questions can help us figure out how we might summarize our technical report for a general audience.

W	**Question**
Who	Who are the researchers investigating the problem and conducting the analysis?
When	When was the analysis carried out (e.g., "today", "this week", "recently", etc. are acceptable if the press release carries a date)?
Where	Where was the analysis carried out?
What	What were the findings from the analysis?
Why	Why are the findings important?

Table 4.1: The Five Ws. The traditional set of questions addressed in early paragraphs of a press release.

W	**Question**
Who	Who are the subjects of study?
When	When were the data collected, e.g.,when did the subjects participate in the study?
Where	Where were the data collected, e.g., where were subjects under study located?

Table 4.2: The Three Ws. The three additional questions about the data that are answered in press releases about scientific findings.

■

Excerpt: *Examine the Argument – The Role of Quotes*

A press release usually contains quotes from the investigator and quotes from others who are knowledgable in the field and understand the potential impact of the findings. Quotes are typically interspersed with more detailed information about the statistical elements, and used to expand further on themes and introduce new information. Quotes also offer human-interest perspectives about the researchers and others involved.

When we read a press release or news story, we examine the quotes and how they help make the argument. Quotes from experts can convey a broader or different perspective than the investigator's, and they can serve to assess the impact of the findings. For example, others can say things about the importance of the findings that might be awkward for an investigator to say about their own work. In addition to adding credibility to the story, quotes can make the story more personable. When we read news stories, we generally like to hear directly from investigators through a quote about why they work in their area and what they find exciting about their discovery.

■

Excerpt: *Map the Organization - The Inverted Pyramid*
The W-questions are often addressed in the first few paragraphs of a press release. However, the beginning of a press release also has the job of capturing our interest so that we continue reading. After we read about the essential elements about the researchers and their findings, we look for additional details about the research. These details are organized from most to least important in what is referred to as an *inverted pyramid*. Additionally, the details are interspersed with quotes that help make the argument for the findings and add a human interest component to the story. Finally, the press release finishes with a brief conclusion. In this section, we provide examples of this organizational structure of a press release.

The Hook A catchy opening is referred to as a hook. The first sentence of a press release should catch our attention and entice us to read more. While the first sentence needs to hook us, we are wary of exaggerated findings. Press releases sometimes leave it to the journalist writing a follow-up story to provide the hook and simply begin with a straightforward presentation of the Ws.

After the essential facts about the findings have been conveyed (i.e., the W-questions in Tables 4.1 and 4.2), the press release provides additional details, organized from most to least important. This inverted pyramid allows us to stop reading before the end of the article without missing the main point(s), and it allows the journalist to easily cut the press release and create a shorter story without losing the most important details. It can be a useful exercise to read a press release with an eye toward this kind of truncation. That is, we identify places in the story where a reader might stop, and we consider what would be their understanding of the scientific discovery at that point. In print journalism there are space constraints, so this cut often ensures that the essential elements are placed “above the fold” in the physical paper. In digital journalism, the “above the fold” analogy becomes the scroll (i.e. the essential elements must be readable before the first scroll).

■

The audience for a blog differs from a formal scientific article or a press release because the author has more control over who they want to reach. The writer defines the audience that they are aiming for in a blog, so they can take a more varied approach to writing blog posts. Bloggers may want to make their work more accessible, teach a concept, help others avoid having to reinvent the wheel when solving a technical challenge, foster discussion, synthesize an experience, or represent themselves in their field. Each of these motivations lead to different strategies that an author can use to overcome particular writing challenges that come from different flavors of blog posts.

4.4 Part II: Preparing to Write

Pedagogical Context for Preparing to Write

Describing data itself, creating informative data visualizations, and writing readable code are typically addressed in statistics and computer science curricula already. Our contribution in this part of the book is to emphasize that these steps are not only part of a broader sense of communication but are also essentially connected to reading, writing, and narrative.

Visualization has already made its way into the statistics classroom. Nolan and Perrett provide assignments for data visualization units in undergraduate statistics courses; these are informed by Friel, Curcio, and Bright's investigation of what affects graph comprehension [125, 56]. Friel, Curcio, and Bright identify different levels of comprehension from the most basic ability to extract information from the data, to the intermediate ability of being able to find relationships in the data, to the most advanced ability to move beyond the data that is presented [56]. Building on this hierarchy, we explicitly show how visualization is part of a statistical argument. We emphasize that crafting a visualization helps reveal a part of our narrative (Section 4.4) and revising a visualization both aesthetically and in terms of overall effectiveness in communication helps hone our message (Section 4.5).

Wolfe makes the point that to make a data visualization we make "rhetorical choices that underlie our decisions on how to summarize, aggregate, and synthesize the data we visualize" [196]. Although we talk about how to make appropriate visualizations that can stand alone in Chapter 4, we also consider the integration of our visualizations into our narrative. By choosing and ordering plots during the storyboard process in Chapter 6, we make rhetorical choices for our overall data analysis report.

Effective coding strategies are typically taught in a computer science course, but they are increasingly appearing in statistical courses that contain computing elements. We approach the topic from a higher level, focusing our approach on showing students that code is also a form of communication that must adhere to style guidelines (Section 5.3 makes the analogy to grammar in typical writing) and can be revised for both correctness and elegance (Sections 5.4 and 5.5). The "reading to write" paradigm (Section 5.1) also has precedent in the coding literature; Simon et al. discuss the relationship between reading code and writing code [168]. They find that their students' ability to read code is strongly related to their ability to write code. Lopez et al. find that the ability for students to perform "code tracing" (i.e. stepping through the code by hand to assess expected output) is also positively correlated with their ability to write code [103]. These findings support the importance of viewing code as something to be read as well as written.

Describing Data

Before jumping into the mechanics of code or design principles for making visualizations, describing the data itself is an accessible place for students to start writing. Even before we have solidified our analysis, we can be describing the data that we have access to. Since the data is an input to both visualizations and code, it is important to understand it on its own first. This chapter introduces two re-occurring examples, e-Cigarettes and infant health, to illustrate how to write about data provenance and preparation.

The tension between being precise and concise is a common theme that arises throughout the book. When talking about data, it can be challenging to determine what details to include in the main report v. in the supplementary materials.

For example, we often summarize the data cleaning phase of the analysis in a few sentences or paragraphs. We do not write the description as a chronological sequence of our discoveries and data modifications. Instead, we explain and justify the actions taken to prepare the data for analysis.

Excerpt: *Excluded E-Cigarette Users*
In their article on e-cigarettes [182], the authors justify dropping one of the four groups of e-cigarette users from their analysis due to small sample size and discuss their treatment of missing values. The article also explains, without going into detail, that student responses are weighted according to the survey design and nonresponse. This statement assures the reader that the analysis was done with careful consideration of the complex design and of missing values in key variables. A more technical article such as the survey's associated Methodology Report might describe the design in greater detail [127].

■

Another theme that occurs throughout is separating discussion of process from results. When describing data, we need to write specific descriptions of data fields without getting into implementation details. For example, the reader does not need to know how information is coded in the source file. All they require is a simple name in plain English for a feature and information about the measurements, such as the units or category descriptors. However, tracking our analysis and its implementation is important for reproducibility, so this chapter also discusses reproducible data wrangling and accessibility of raw data. These concepts are reiterated in the coding chapter.

This chapter also provides guidance about avoiding potential pitfalls in describing simple summaries of data including overly-precise reporting of numbers, leaving out information about sample size, and confusion in units between proportions, rates, and ratios.

Communicating Through Statistical Graphs

Just as we read a statistical report, we also "read" plots. A statistical graph provides support to a written argument and can provide a more digestible presentation of numerical summaries. In this chapter, we consider how to create plots that are effective in communicating statistical findings. We address how to select an appropriate type of plot to reveal underlying structure in the data, facilitate important comparisons, and create context for interpreting the distributions and relationships observed.

Each plot has a take-away message. Attention to detail in a plot can make a big difference in making this message stand out. In this chapter, we collect advice on making readable visualizations and organize them into five categories that address how to: incorporate the study design, choose scales, handle large amounts of data with smoothing and aggregation, facilitate meaningful comparisons, and add contextual information. Excerpts illustrating how to strengthen a plot's argument follow.

Excerpt: *Incorporating the Data Design: Race Times in the Cherry Blossom 10-mile Run* The boxplots on the left in Figure 4.1 show the race time (minutes) by age for male runners in the annual Cherry Blossom 10-mile run from 1999 to 2012 (scraped from [31]). It is tempting to interpret the curvature in the lower quartile, median, and upper quartile of the boxplots as the trend in how an individual's performance changes with age. However, this is a cohort study, not a longitudinal study. The 25-year olds running in, say, 2001, are a different group of people than the 50-year olds running in the race that year, and these two groups can be different in ways that would affect the relationship between race time and age. For example, the 50-year olds in the run are likely to be fitter for their age than the 25-year olds.

Furthermore, these data have a time component, the year of the run. The plot on the right conditions on year, where each curve is a local average of race time for runners of the same age for that year. The plot reveals interesting structure: average race times have slowed over the years. This is likely due to the increased popularity of the race with greater participation by novice runners in recent years.

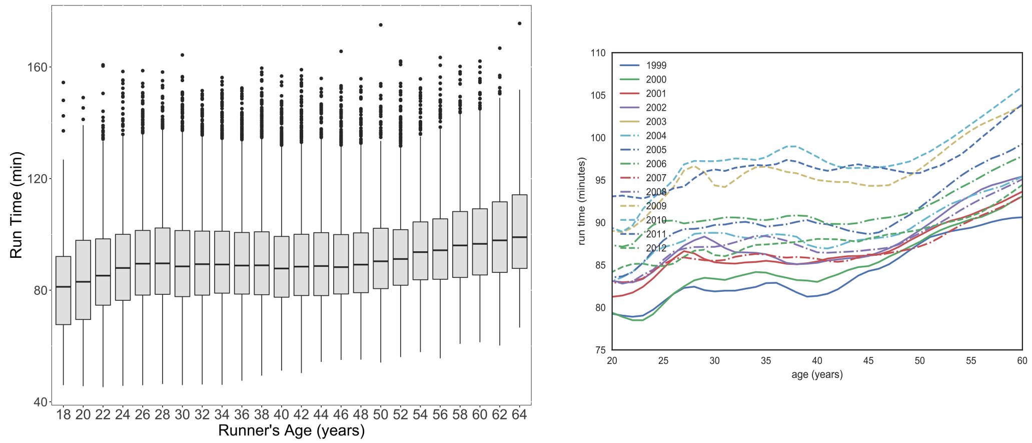


Figure 4.1: Race Time by Age. The boxplots on the left show the race time (minutes) by age of the male runners in the Cherry Blossom 10-mile run from 1999 to 2012. The plot on the right contains a curve for each year of the race. Each curve is a local average of the race time for runners of roughly the same age. One feature apparent from this plot is that average race times have slowed over the years.

■

Excerpt: *Choosing the Scale to Reveal Structure: San Francisco Housing Price Distribution*
The density plot in the lefthand plot of Figure 4.2 contains the sale price of all houses sold in 2004 in San Francisco (scraped from the San Francisco Chronicle site of weekly sales that is no longer available). It is difficult to closely examine the distribution because a few unusually expensive houses force the bulk of the distribution into a small portion at the left of the plot. In contrast, the density curve in the righthand plot excludes these high priced houses. The shape of the distribution for the bulk of the houses is much clearer in this plot. There we can more easily observe the skewness of the main mode. If we do not include all of the data in the graph, then we must mention this exclusion in the caption or on the plot itself.

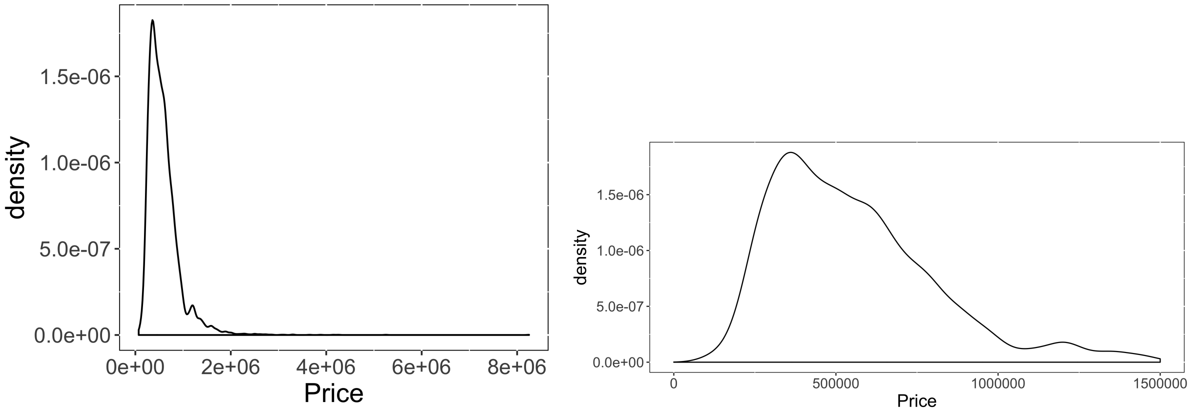


Figure 4.2: Distribution of House Price. Both plots show the distribution of sale price of houses in 10 cities in the San Francisco Bay Area in 2004. The one difference between them is that sales over $1.5m are excluded from the plot on the right. This exclusion makes it easier to see the shape of the bulk of the data. The distribution has a short left tail with a market entry point of about $250k and a mode around $350k. The distribution is right skew with a large right shoulder indicating that many houses are in the $500-$750k range. Even clipping houses that sold for over $1.5m, we see that the distribution has a long right tail with many houses selling for $1m and more.

■

Excerpt: *Smooth to Uncover Trends: Cherry Blossom 10-mile Run, continued*
The scatter plot on the left in Figure 4.3 shows the race time (minutes) and age for male runners in the Cherry Blossom 10-mile run from 1999 to 2012. There are more than 50,000 points in the scatter plot and over plotting makes it impossible to see any relationship. In contrast, the plot on the right smooths the data by taking local averages of the race times for runners of the same age in each year. As noted earlier, this plot reveals that average race times have slowed over the years.

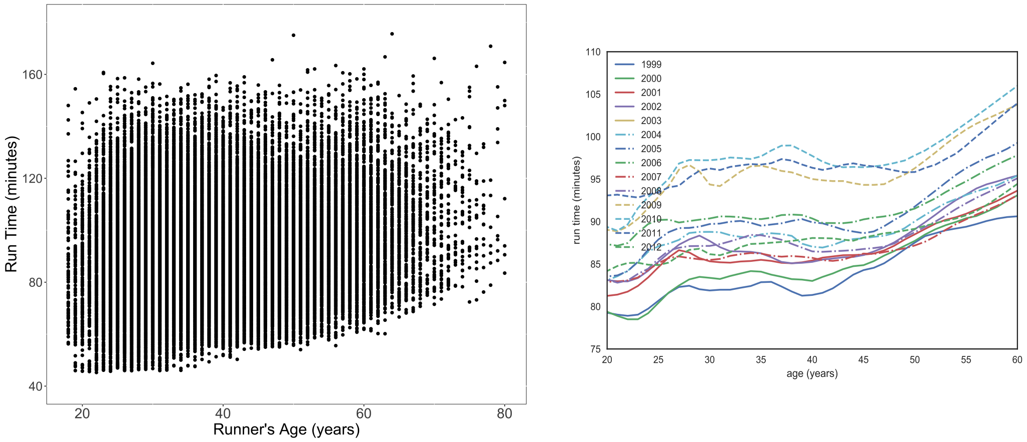


Figure 4.3: Race Time by Age. The scatter plot on the left shows the race time (minutes) and age for male runners in the Cherry Blossom 10-mile run from 1999 to 2012. There are over 50,000 points in the scatter plot. Due to over-plotting we cannot see any patterns in the data. The plot on the right contains a curve for each year of the run, where one curve is a running average of the race times by age. One feature apparent from this plot is that race times have slowed over the years.

■

Making good statistical graphs is hard and usually is an iterative process. As we refine our written argument, we refine the accompanying visuals to best support our findings.

Excerpt: *California Voter Registration Trends.*

A typical sequence to create a plot goes as follows:

- Select a plot type according to the kind of data.
- Make the plot taking the software defaults and using cryptic labels.
- Consider transformations to symmetrize and straighten relationships and the choice of scale (this may involve several iterations).
- Address issues with over-plotting.
- Consider variable(s) to condition on that would inform whether the shape or relationship observed is maintained across subgroups.
- Try different approaches to visualizing this comparison, such as a grid of plots, use of color and plotting symbols, smoothing, etc.
- Determine whether there is any additional information that would help put the findings in context, such as reference markers or highlighting particularly interesting observations or features.

- Add informative labels, legends, and titles.
- Write the caption.

This process is similar to the writing process. The default statistical graph produced by the software is like a first draft, and we put a large effort into revising to improve the story we are telling. On occasion we even decide to discard our current draft and begin again. In the moment, we might think some of the steps listed above are unnecessary for a particular plot, but it is usually a good idea to work through each step. As we go from one step to the next, we might uncover something that we didn't expect to find. This may signal a need to redo an earlier step. We provide an example of a cautionary tale about not getting caught up in the editing process before finding the best plot for telling our story.

We are interested in voter trends in California and visit the online voter registration site of the California Secretary of State [24]. There we find county summaries of voter registration for seven presidential election years (1992, 1996, ..., 2016). We scrape these data and make the plot in Figure 4.4.

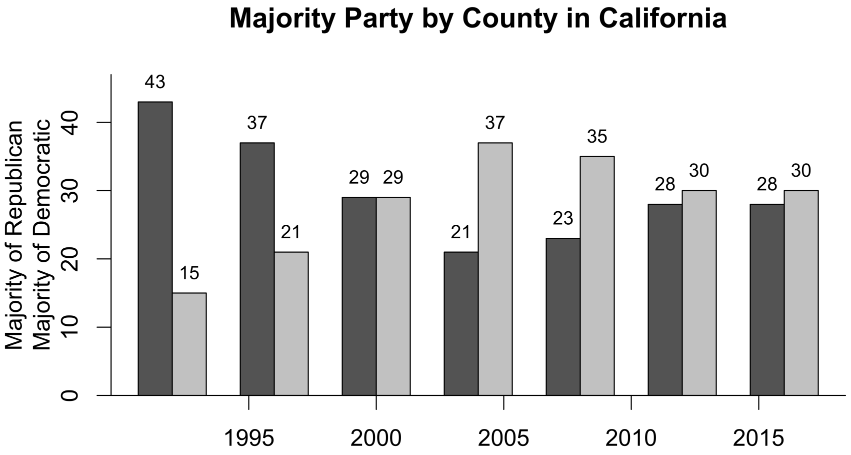


Figure 4.4: Bar Chart of Majority Party in California Counties. The side-by-side bars show the number of counties in California that have a majority of voters registered Republican (pale gray) or Democrat (dark gray). Each pair of columns adds to 58, the number of counties in California. This plot was made on `swivel.com` (no longer a live site) from data available at the state of California's voter registration site.

This plot has many problems, as listed below.

- Tick marks: x-axis tick marks are at 5-year intervals and do not line up with the locations of the bars so the reader has to work too hard to figure out that the bars correspond to measurements made at 4-year intervals.
- Color: atypical use of light and dark green for the Democratic and Republican parties, which are traditionally represented by blue and red.
- Legend: lack of a legend means that we cannot discern which color represents which party.
- Axis Label: y-axis label does not indicate the units of measurement.
- Title: confusing title that does not illuminate the content of the plot.

It is tempting to quickly try to fix these problems. We can "improve" the plot (see Figure 4.5), but we should first think some more about the story and whether this plot makes the best argument. Our aim is to show change in voter registration over the past seven presidential elections. However, it is people who register to vote, not counties. County size is a lurking variable–small counties tend to be rural and conservative–so using counties overstates the Republican presence. Rather than record counties, we should tally voter registration counts. To do this, we revisit the registration website to obtain new data. There we find a page of voter registration counts by party, including counts for other parties and for unaffiliated voters. Since the number of unaffiliated voters is sizable, it could be informative to include these registration numbers in the plot.

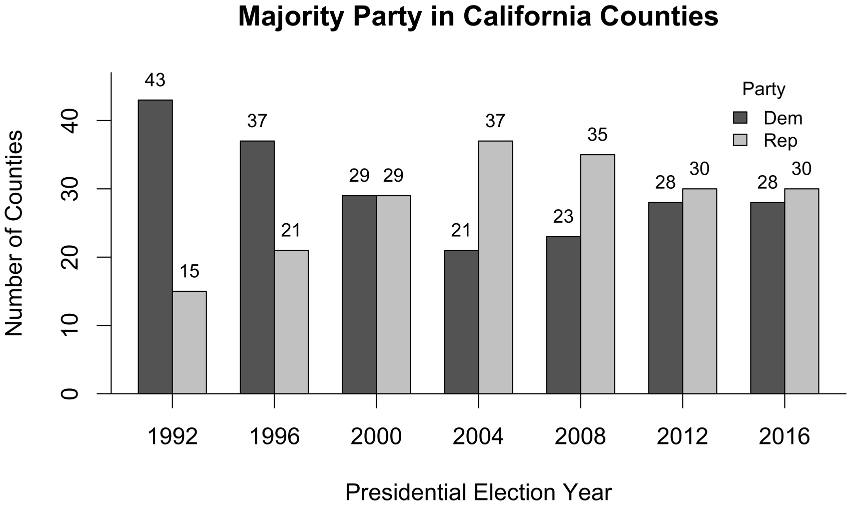


Figure 4.5: Bar Chart of Majority Party in California Counties – Revised. This bar chart addresses many of the problems found with the bar chart in Figure 4.4, including the inaccurate y-axis label, ill-positioned tick marks on the x-axis, poor choice of colors, and lack of legend.

What kind of plot should we make? We have registration figures over time so a line plot seems appropriate. Also, given that the California population has grown dramatically in the past 25 years, rather than compare raw registration numbers, we scale them by each year's total registration and compare percentages.

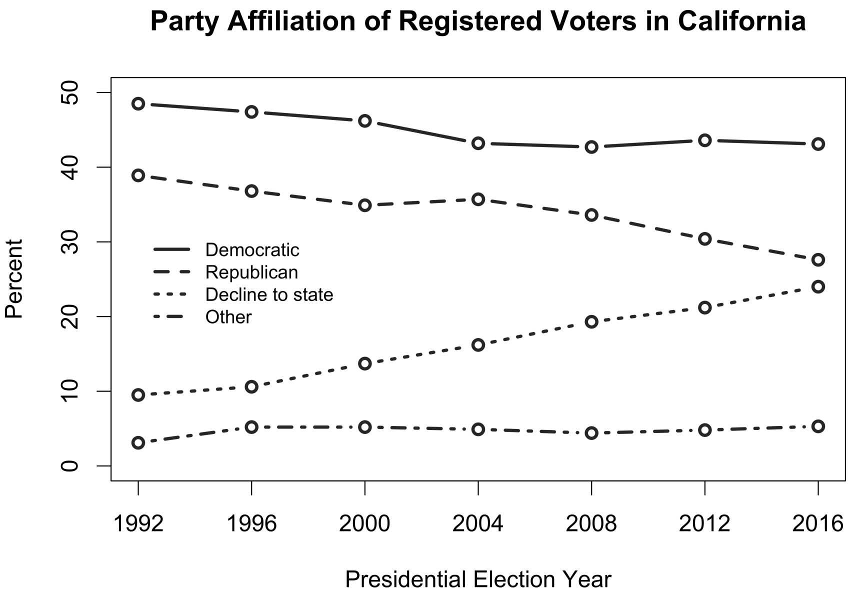


Figure 4.6: Distribution of California Voters by Party. The line chart addresses the essential problem with the bar chart in Figure 4.5, i.e., we are interested in the change in voter registration over the years, not in the number of counties that are majority Republican or Democratic. Here we see that the percentage of registered Democrats and Republicans have declined in this period, the percentage of unaffiliated voters has dramatically increased, and the gap between Democrats and Republicans has grown from about 10% to 15%.

We discard the first plot and make an entirely new one (see Figure 4.6). This new graph makes a more interesting and accurate depiction of voter registration trends in California. We see that: the percentages of registered Democrats and Republicans have declined over this period; the percentage of Democrats was about 10% higher than the Republicans in the earlier years but the spread has grown to about 15%; and the percentage of unaffiliated voters has dramatically increased from about 10% to about 25%.

The time that we have spent editing, revising, and entirely remaking our graph was well spent. We now have a plot that makes a compelling visual presentation of the change in California voter behavior over the last 25 years. ■

Communicating through Code

Code both carries out a data analysis and provides a way to communicate our ideas about the analysis. It is important to write clear code to help others understand our process, check our results, or repurpose our approach to their own problem. Clear code helps us remember our own thought process and helps us avoid logic errors. Just as we explained how to read articles to learn how others organize and write about their findings, we can learn how to write code from reading others' code.

Coding has a variety of analogues to writing. We have style guidelines and writing conventions for both traditional writing and code. Refactoring code to make it more clear or efficient is similar to revision of a typical manuscript. We write comments and documentation which have to narrate the purpose of the code. Pseudocode provides a skeleton of code to come and signals our intent for a particular analysis or functionality just as an outline provides a starting point for writing a draft of a paper.

Excerpt: *Pseudocode*
Pseudocode provides an overview of a computational task without the precise implementation details and without adhering to a particular programming language syntax. Pseudocode can be included in a publication to explain the implementation of an algorithm, an idea behind a statistical method, or the data processing pipeline for an analysis; it keeps the focus at a high-level and does not take up too much space. The full code used to produce the results is often provided online as supplementary material, rather than in the article. Ideally, the code is provided via a notebook that integrates explanatory text with code.

Pseudocode focuses on logic, not syntax, and can help us organize our thoughts and write better code. The fundamental elements in pseudocode are inputs, outputs, and key verbs that describe operations. We want to use concrete, specific, and active verbs to concisely represent the core actions. Although we do not need to specify the format of our inputs and outputs (e.g., data frame, list, dictionary, array, etc.), it can be helpful to consider what type they are (e.g., string, integer, double, factor, etc.). As we describe how to take the inputs and create the outputs, we need to consider the order of operations, including conditional and repeated evaluation of expressions, and ensure that a piece of information is presented ahead of a step that relies on it.

Just as in writing text, there is a balance between being precise and concise when writing pseudocode. We need to include enough detail so that a reader can understand and possibly reproduce our approach, but not so much detail that the reader loses sight of the key computational features. If we use pseudocode as an outline before writing the actual code, it can help us write clearer more focused code. The pseudocode can also act as documentation for our code. ■

This chapter also touches on computational reproducibility and tools used to work towards this goal including version control and computational notebooks. Version control is just as important for our writing projects, and computational notebooks are often the first place where the narrative of our research starts to emerge.

4.5 Part III: Composing the Story

Pedagogical Context for Composing the Story

The statistical report has been the most traditional way to practice statistics communication in undergraduate courses. For example, Spurrier describes a capstone course for undergraduate statistics majors [173]. This course includes "modules on important nonstatistical skills" such as written and oral reports. Students learn the "role of each major section of a technical report" and "presentation ground rules." We cover this material in Chapter 7 and Section 11.1. Our additional contribution is to include many examples of each part of a technical report in a data context but with content that is accessible to an undergraduate. These examples give students concrete writing to refer to as they write their own reports.

Technical reports can be stilted and formulaic, especially if written by novices. Pfannkuch et al. note a lack of opportunity for students to go beyond the description of their analysis process and tell data stories: "Not one of the (introductory statistics) books clearly demonstrated reasoning comparatively all of the way from looking at the plots, unlocking the story and the underlying concepts, and synthesising the whole data story through a transparent reasoning process" [139].

Our contribution is to focus on crafting a narrative by developing a storyboard (Chapter 6) before writing a first draft. This deviates from the formal outline that is often used before writing the first draft; an outline assumes you already know what you are going to say. In contrast, the storyboard process requires students to grapple with their *findings*, decide what the interesting and defensible storyline is, and choose which material is crucial to telling that story. Going through this process before writing a first draft can help avoid a more formulaic report that details only the statistical analysis *process*.

Organizing the Story

Rather than providing excerpts, this full chapter is included below. The idea for storyboarding our research was inspired by an activity led by Sara ElShafie in the Data Science for the 21st Century NSF Training Program Science Communication Short Course at UC Berkeley taken by Sara Stoudt. The activities described in this chapter of the book are also inspired by exercises in this short course.

Before we start writing a full draft, it can be helpful to organize both the structure and the story. In this chapter, we discuss how to build a narrative, select relevant details, and order them in a compelling way.

Creating a Storyboard

In the film industry, a storyboard is a series of sketches that depict the important changes of scene and action. These are arranged in a sequence to visually layout the progression of the story. Although storyboards are traditionally used as a means to brainstorm and plan out movies, TV shows, and other visual arts, the storyboard is fundamentally a tool to organize one's thoughts and streamline them into a story that is accessible and compelling. Data scientists too can benefit by mapping out their results into a storyboard before beginning to write.

We often start a data analysis with summary statistics and simple plots to understand the data. We then dig deeper to build and test models. Throughout the exploratory and formal data analysis, we make more plots and do more analyses than we want or need to display in a paper. Choosing what is most relevant to our argument is a valuable skill, and being economical is key. Determining the order to present our findings can be challenging because this order is often different from the sequence in which we performed the analysis. Creating a storyboard helps bridge the gap between our knowledge and the knowledge of our reader. Remember that you have spent much more time with the material than your reader so it can be difficult to pare down your findings and insights to those most relevant to your story. On the other hand, it can be easy to omit details that you are overly familiar with or think obvious.

We have organized storyboarding into six basic steps.

1. *Collect tables and plots.* In your work, you have most likely made many preliminary tables and plots to help you understand your data and models. Gather these tables and plots all in one place and in a format where you can easily move them around.

2. *Group related findings.* Group tables and plots that contain similar messages, and summarize each group by one or two sentences or bullet points. Some groups will naturally form, but also consider regrouping and rearranging the plots and materials to uncover groups that are not immediately obvious. You may want to duplicate some plots and tables to put them into more than one group.

3. *Make an argument (find the story).* Consider the groups that you constructed from your findings and ask what they are telling you. Look at connections between groups and see if a story emerges. Sequence the groups so that they tell your story.

   The groups and the tables and figures within them may serve different purposes. A group of plots might build off of one another. Identify the train of thought: what is the starting point, what is the end, and what intermediate steps are needed to connect the two?

   Alternatively, plots in a group might supplement one another or supplement another group of plots. Determine how these plots connect to the main story. Do they branch off from an initial line of inquiry? Do they support a core argument?

4. *Choose the tables and plots needed to tell your story.* Be sparing with your plots and tables; choose the relevant details that explain the story while removing redundancies and unnecessary lines of inquiry. However, make sure that your core argument can be seen through your plots and tables.

   - If your plots present similar ideas, then decide which single plot or table suffices to explain or clarify the point. Remove the rest; they are redundant.
   - When you have material that builds off of one another, decide which intermediate steps, if any, are necessary for a reader to understand how to get from the start to the end of your story.

- Identify holes in your argument that need a plot to support them, and make and add them to your storyboard.
- In the case where materials supplement each other, decide if the supplementary information is necessary for your main storyline or if these details can go in an appendix or be summarized briefly in writing, and not shown with plots.

5. *Sequence the chosen tables and plots.* This requires a sense of fluidity between ideas. Examine all of the groups together and consider how each group relates to the others. Is there a temporal component between them? Does one group motivate another? Decide which details you need to see first to understand the details that follow (e.g., would A make sense without first seeing B?). Think about how the supplemental information that survived the parsimony step ties back to the next major detail in the storyline. Can it be placed after the main storyline as a discussion point or before to set the stage?

   Remember that the order that best expresses your *ideas* in a story does not necessarily match the order that best expresses your *research process.* A fluid order most likely does not match the chronology of your analysis.

6. *Add captions and transitions.* Write a brief caption for each plot that conveys the message you want the reader to glean. An important consideration in this step is whether, for the material being summarized, there is a mismatch in experience between you and your eventual audience. It can be easy to forget that the reader only knows what is being presented, and details were necessarily left out as part of the parsimony goal above. Is there any information that you could add between plots and tables that would smooth the transitions between them or provide context to ease interpretation?

Think of storyboarding as a visual outline that informs a formal written outline. We cannot write this formal outline without having an idea of what we are trying to say. The main goal of a visual outline is to identify the narrative. What problem exists? What did we do to solve that problem? Why does it matter that we have solved this problem? After we have identified the story and experimented with the ordering of details by rearranging panels of plots and text summaries, we can build a formal outline to tell the story we have identified. An intermediate step is to take notes on your storyboard.

Taking Notes on the Storyboard

The storyboarding process helps you separate the journey you have made to arrive at your findings from the findings themselves and prepare to write. Note taking is another step in this direction. Before jumping into writing a first draft, take the opportunity to look over your storyboard and make notes. Reflect on the storyboard and jot down information needed to support your argument. These notes should consider the following topics.

1. *Foundations.* Make a list of the concepts that are most important to your findings. For each, determine whether it is: a concept that needs to be explained before presenting

your findings; an idea that your work extends or that your work fundamentally relies on; or a motivating or competing notion that needs to be mentioned.

2. *Assumptions.* Augment your notes with any assumptions you made in your analysis. Annotate them with a brief indication of why they are reasonable, how they were confirmed, and what went unchecked.

3. *Eliminations.* Examine the plots and summaries that were dropped from your storyboard. Why were they dropped? Do they warrant mention in the report? If so, then add them to your notes. Write down whether the point should be included before, concurrent with, or after the findings. Also, if appropriate, mention where supplementary material on the topic can be found, e.g., an online resource or an appendix that you plan to write.

4. *References.* Undoubtedly, you have read many articles that relate to your work. Look over your findings and identify a few of those articles that are most relevant to your story. These can include an article that your work extends, one that justifies the method you have used, or one that supplements your findings.

Taking notes on your storyboard will help you settle on an outline for your narrative and solidify your argument. These notes (together with your storyboard) form the basis of the introduction, background, motivation, and discussion sections of your report. They are not meant to be exhaustive. Instead, their purpose is to identify the supporting material needed to tell how you arrived at the findings in your story. With the storyboard fresh in your mind, the core ideas, terms, and references are readily apparent. Noting them down now will save time when preparing your first draft.

Iterating

Storyboarding is part of the cycle of discovery. As we annotate plots, write brief transitions between them, and take notes on the storyboard, new ideas often occur to us, and we find gaps in our argument. With the help of the storyboard, our core arguments surface from our plots and tables, but as the essential pieces come together, we often discover that the justification of a point is missing or that a new line of inquiry appears. When this happens, we find ourselves updating our analysis, adding to our storyboard, and taking more notes.

Neither a storyboard nor its initial summary are fixed, and we don't always respond in the same way when we identify something missing. There are three basic approaches that we take to address newly found holes in our argument.

1. *Acknowledgement.* A hole does not always need to be filled, but at a minimum it needs to be acknowledged. In the report, this acknowledgement could be an identified assumption, a point of discussion, or a topic of future work. We update our notes to this effect.

2. *Patchwork.* The hole may have a quick fix, e.g., update a plot, run an additional simulation, skim articles on a related subject, or run-through a variant of the analysis.

This new work may be included in the storyboard, or it may simply result in an additional note that summarizes the new information.

3. *Rework.* When we think the new issue needs to be more fully investigated, then we take a more thorough approach in revisiting the analysis. For example, we may need to carry out a parallel avenue of investigation or check out a special case. If we believe it will strengthen our argument to flesh out this new idea, then it's best to tackle it at this stage.

Iteration happens throughout the writing process. After we have a storyboard and notes, we prepare a formal outline. Even this part of the process is iterative and can send us back to the data and our analysis. As we write up the essential pieces of our report, we may again find a hole and need to iterate. Our continued willingness to iterate will further strengthen our argument.

It's common for holes in your argument to crop up and lead you to retrace your steps. That's why we storyboard. However, when we are thorough in the early stages, the iterations in report writing are likely to uncover issues of the acknowledgement and patchwork types that won't require major rework and rewriting.

Creating a Storyboard for Drug-Related ED Visits

We demonstrate the storyboarding process with an exploratory analysis of data from the Drug Abuse and Warning Network (DAWN) [184]. Each observation in this dataset represents a drug-related visit to the emergency department (ED) of a hospital in the United States. Other research has investigated the drugs accidentally ingested by children and those drugs responsible for over-medication among the elderly. We have chosen to focus our analysis on these two age groups to learn more about their drug-related ED visits.

We walk through the steps of creating a storyboard from a collection of crude tables and plots that represent the kind of output that we typically work with in the exploratory phase of an analysis. The collection includes tables of percentages and rates, barplots of univariate statistics on the general characteristics of the visits, and bivariate plots that reveal relationships between age and other factors.

Age	≤ 5	6 - 11	12 - 17	18 - 20	21 - 24	25 - 29
Percent	5.8	1.4	5.6	6.6	8.1	9.4

Age	30 - 34	35 - 44	45 - 54	55 - 64	65+
Percent	8.6	14.3	15.3	10.2	14.8

Table 4.3: Age Distribution of Drug-related Admits to an Emergency Department. Age Distribution of Drug-related Admits to an Emergency Department.

1. Collect tables and plots.

Our brief, simple analysis begins with one table and several plots (Table 4.3 and Figures 4.7 through 4.12). The captions for these figures and table are brief descriptions of the information plotted. The process of ferreting out the message in plots and writing informative captions is the purpose of storyboarding. Over the next steps, we cull the plots, revise them, and possibly add more as we piece the story together.

Table 4.3 gives an overview of the distribution of drug-related emergency department visits by age. Figure 4.7 shows the distribution of the type of drug-related visits across age groups. The left plot in Figure 4.8 compares the age distribution by sex, and the collection of four plots on the right side of this figure examine the types of visits by sex for a subset of ages. Figure 4.9 shows where patients go after their visit, by age. The bar plot in Figure 4.10 examines the time of day of the visit, by age. The set of dot plots in Figure 4.11 show where patients go when released from their ED visit by time of day across a subset of ages. The four bar plots in Figure 4.12 explore the types of cases per quarter of the year across four of the age groups.

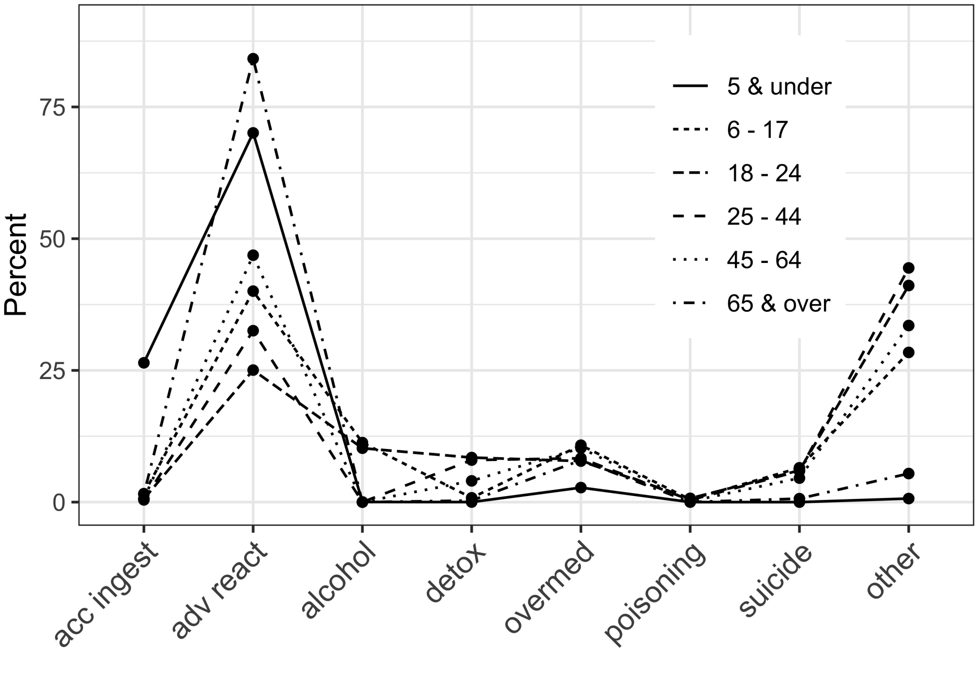


Figure 4.7: Type of Case by Age. This line plot shows the proportion of types of cases within each age group; that is, values across types within an age group sum to one.

2. Group related findings

We group together the set of dotplots in Figure 4.9 on where the patients go after their visit with those in Figure 4.11 that further break down disposition (where the patients go) by time-of-day because disposition appears most relevant to the elderly patients. We also group together Figure 4.10 and Figure 4.12 because they are most relevant to the youngest patients; that is, the time-of-day and seasonal differences are most striking for the five and under group.

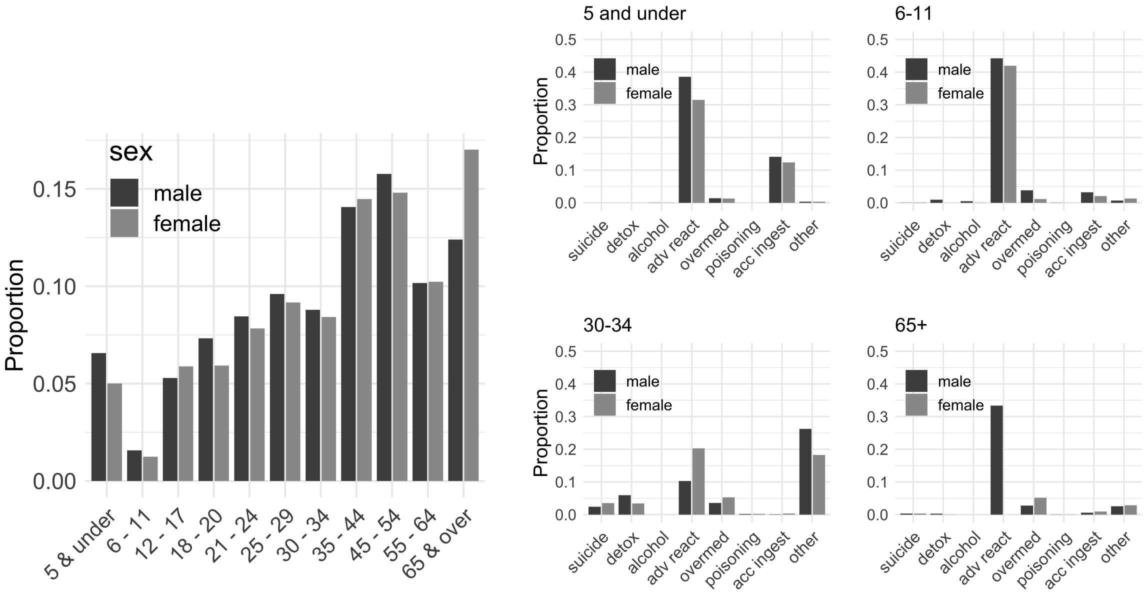


Figure 4.8: Barplot of Age by Gender. The plot on the left shows the breakdown of age for each sex; e.g., the sum of all bars for males is one. Each plot in the group of four on the right shows the proportion of case type by sex. Values within an age panel sum to one.

We have grouped the plots into four sets. We have kept Figure 4.7 on the type of visits by age and Figure 4.8 on the age distribution by sex as separate “set”s. Figures 4.9 and 4.11 form a third set that pertains to the disposition of the elderly, and the last set consists of Figures 4.10 and 4.12 on time patterns in the visits of the youngest patients.

Other arrangements are possible. For example, we may want to duplicate the plots in Figure 4.11 and put them with Figure 4.10, as well as keep them in their original group. In this case, we don’t bother to duplicate Figure 4.11 because these plots add little to the story.

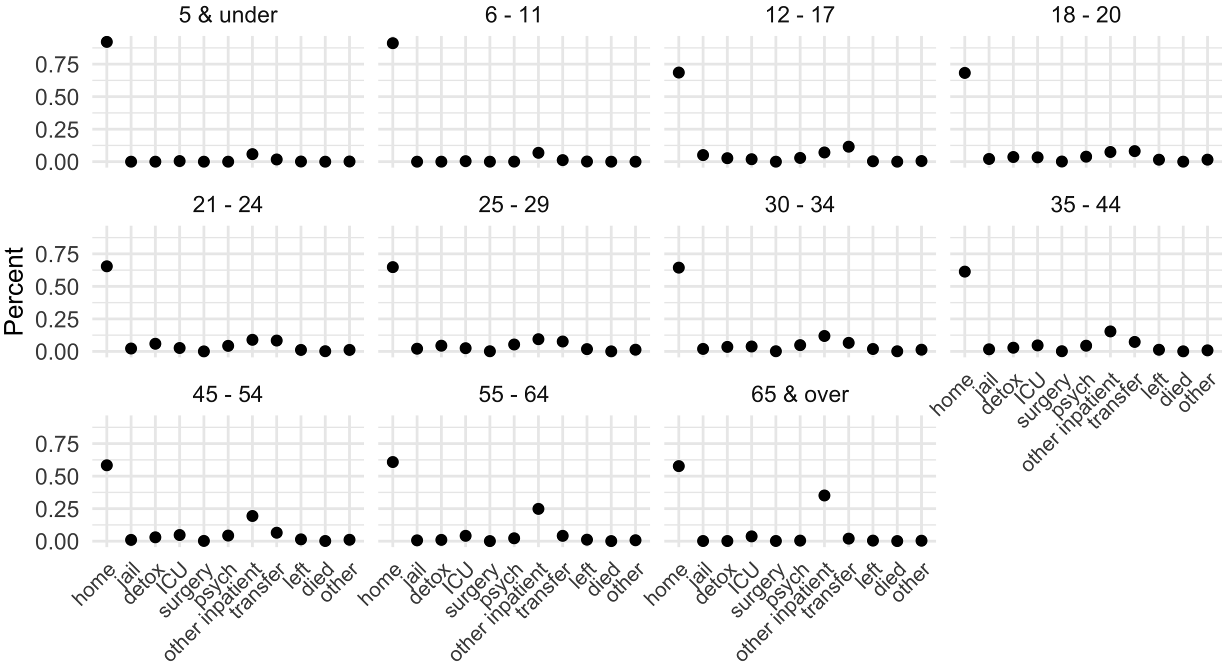


Figure 4.9: Dotplot of Disposition by Age. Each panel of this plot corresponds to an age group. Within a panel, the proportions for teach type of outcome, i.e., where the patient goes after the ED visit, sum to one.

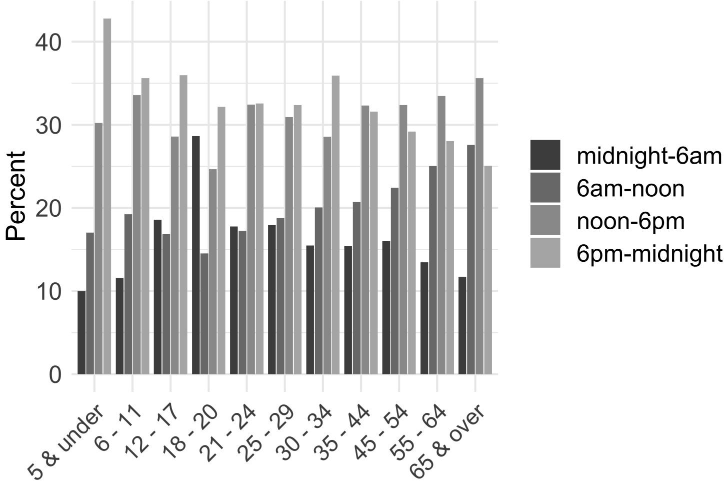


Figure 4.10: Barplot of Time of Visit by Age. This plot shows the proportion of each age group that arrives to the ED during different times of the day. The proportions within an age group sum to one.

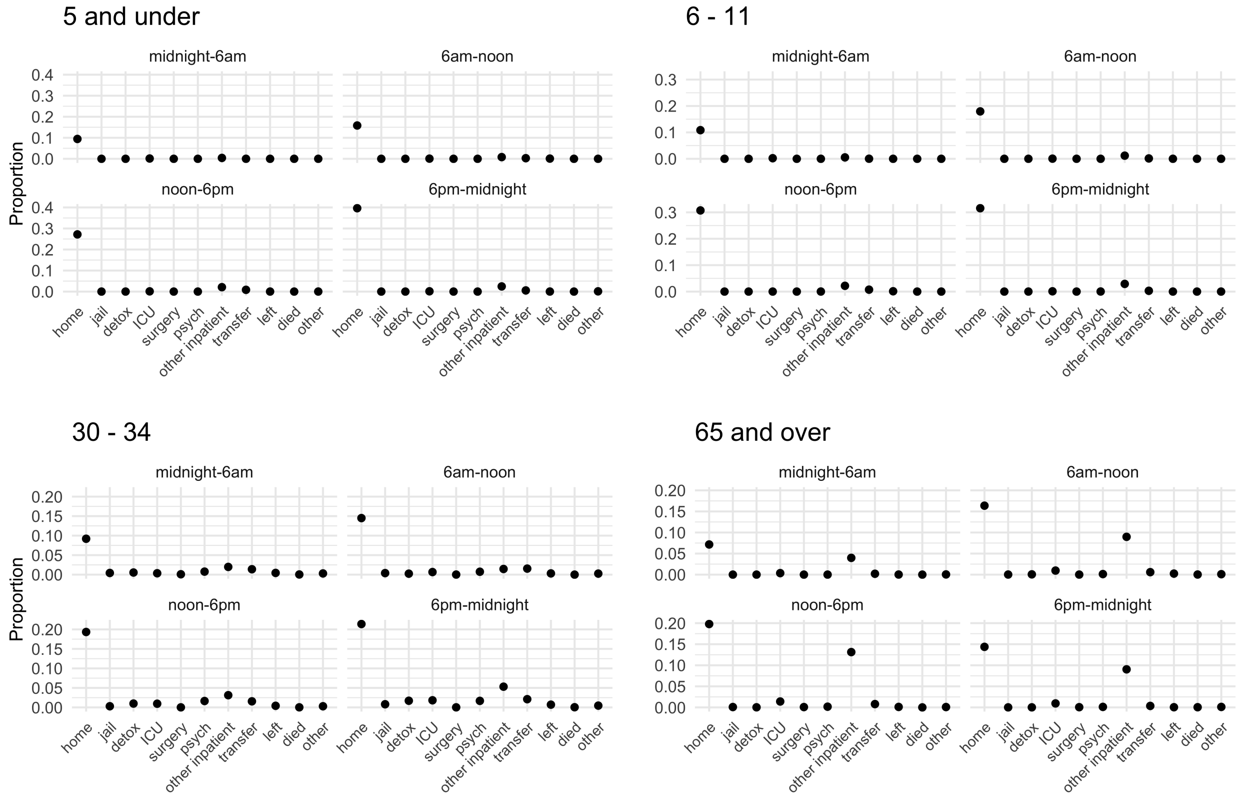


Figure 4.11: Dotplot of Time of Visit by Disposition and Age. These plots show where the patient goes after the ED visit by time of day. Each set of four plots represents a different age group: 5 and under, 6 to 11, 30 to 34, and 65 and older. Points across all four panels within an age group sum to one.

3. Make the argument (find the story)

We have found some important features of the ED related visits for the youth (5 and under) and the elderly (65 and over). We chose to further focus on these two groups because our analysis may offer insights for informational campaigns that aim to reduce the frequency of ED visits for these two vulnerable groups. Specifically, we have found:

- These two groups make up a large share of the ED visits (over 20%) and they share a large proportion of visits due to adverse reactions (at least 80% for each age group).
- For those five and under, the second most common reason for a drug-related ED visit is accidental ingestion.
- For the elderly, the second most prevalent reason is over-medication.
- Males outnumber females in the youth category, while the opposite is true for the elderly category.

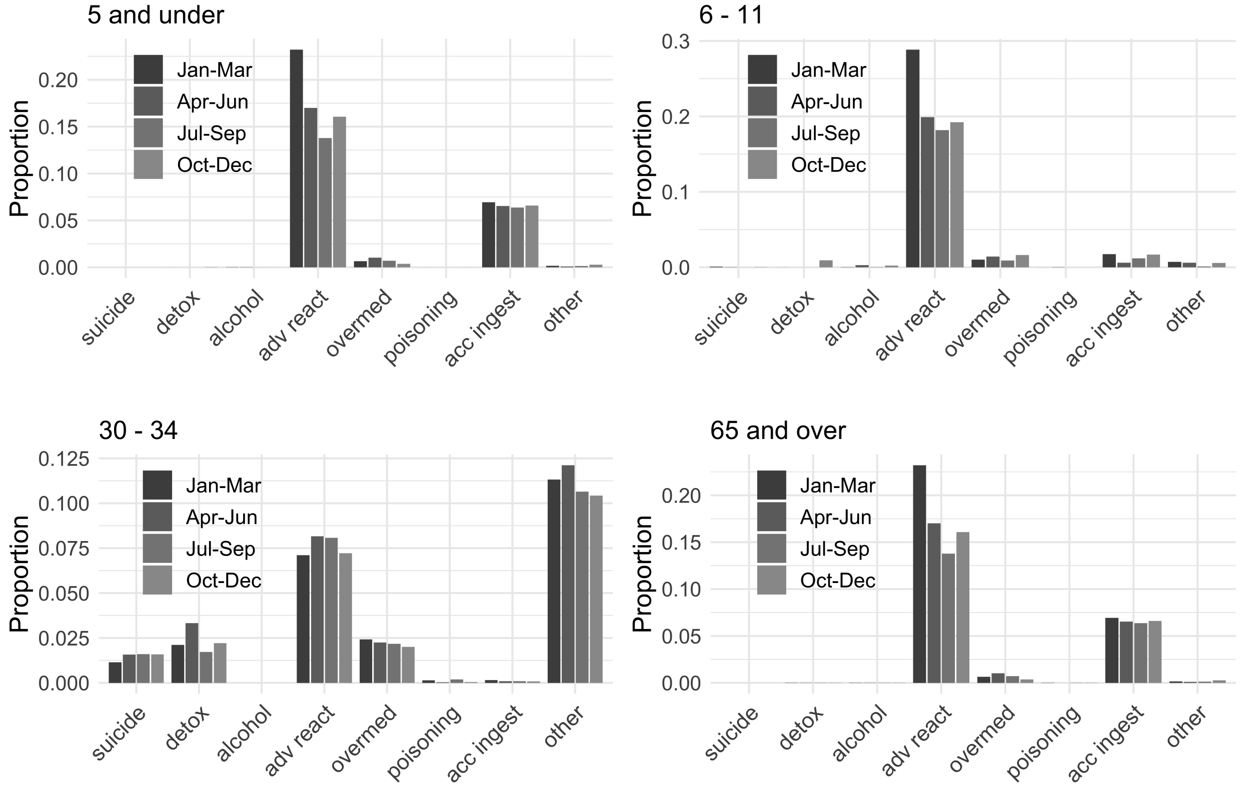


Figure 4.12: Barplot of Season by Type of Case and Age. This group of four plots show type of case (x-axis) by the season for four age groups. The heights of all of the bars within a panel sum to one.

- Youth visits occur most often at night (6 pm to midnight), and we see a larger percentage of youth visits during the winter months.
- After the ED visit, a noticeable proportion of elderly go to other inpatient care.

4. Choose the tables and plots needed to tell your story

We can streamline Table 4.3 because it contains more age categories than needed, if our focus is only on the 5-and-under and the 65+ groups. A problem with this evidence is the lack of any comparison figures. For example, it could be helpful to compare the age distribution for ED visits to the age distribution of the US population. (Note that the Census data was provided for cruder categories so we collapsed categories in the DAWN data in order to make like comparisons between age groups.) According to the 2010 US Census (Table 4.4) the proportion of youngest patients nearly matches the prevalence in the population, but the fraction of 6 to 11 year olds is less than one fifth the Census figure. This comparison may offer insight.

	Age Group								
Source	5 & under	6 to 11	12 to 17	18 to 24	25 to 34	35 to 44	45 to 54	55 to 64	65-plus
DAWN	5.8	1.4	5.6	14.7	18.0	14.3	15.3	10.2	14.8
Census	7.9	8.0	8.3	9.8	13.3	13.3	14.6	11.8	13.0

Table 4.4: Age Distribution. The percentage of drug-related admits to an emergency department (top row) and US 2010 Census (bottom row) in each age group.

Next we consider the line plot in Figure 4.7 on the type of visit. This plot motivates the study and displays key features for both age groups of interest.

On the other hand, the plots in Figure 4.8 compare the age distribution (on the left) and the type of visit (on the right) by sex. Without a more in-depth analysis that compares our findings with age-sex breakdown of the US population, we can't determine whether the differences in sex are noteworthy or simply reflective of general age-sex patterns. The advantage of comparing the sexes for the purpose of our analysis is not evident. We have identified an avenue of analysis that we don't wish to pursue.

We move on to the third group of plots, i.e., those in Figures 4.9 and 4.11. These plots contain useful information about where elderly patients go after their visit to the ED. For now, we keep only the bottom three in Figure 4.9 as they contain the relevant information.

The fourth group of plots (Figures 4.10 and 4.12) concern the time of day and season when people visit the ED. The barplot in Figure 4.10 shows the most striking pattern for those five and under so we prioritize it and drop the others.

In sum, we have selected Table 4.4, the line plot in Figure 4.7, the three panels of dot plots along the bottom of Figure 4.9, and the bar plot in Figure 4.10 to tell our story. Later, we will remake the plots to improve them.

5. Sequence the chosen tables and plots

The following sequence appears to be a natural progression of ideas.

- Start with the numbers from Table 4.4 to give context and a sense of scale.

- Then, describe the reason for the drug-related visits by age to motivate studying the youngest and oldest patients; use Figure 4.7.

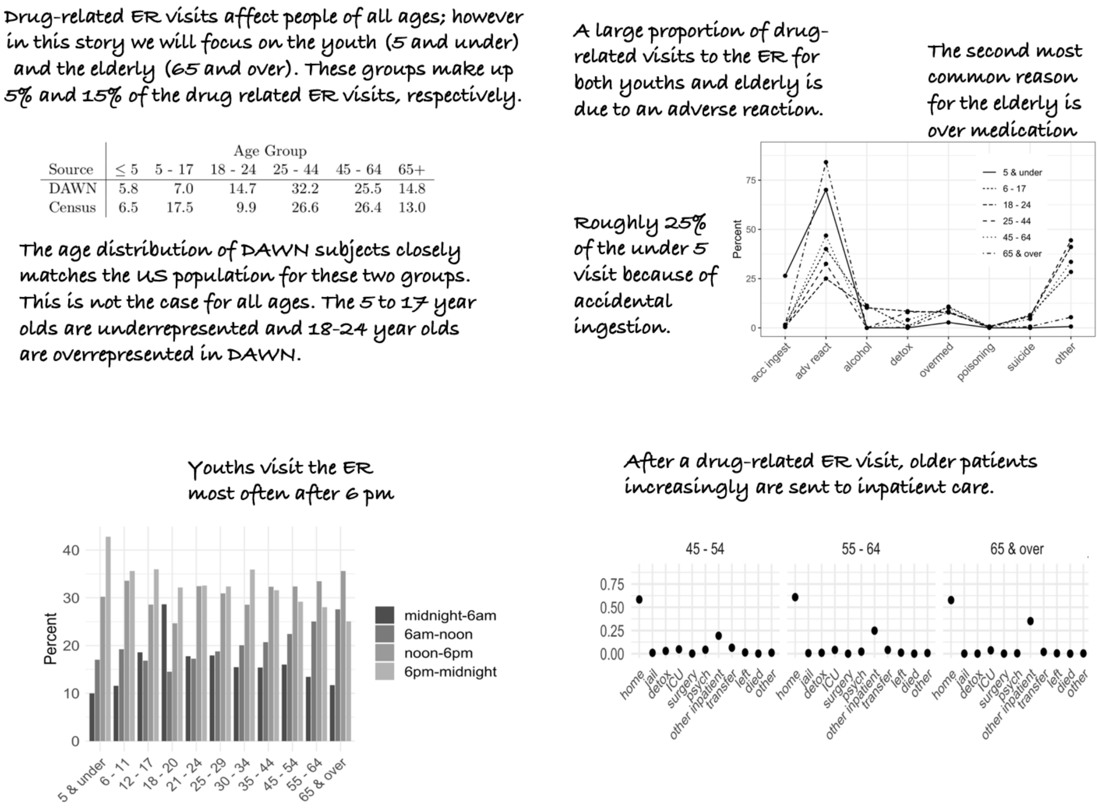
- After we have established the vulnerability of the youngest, discuss *when*. Support from Figure 4.10.

- For the elderly, focus on the increase with age in the proportion of patients being admitted to the hospital. Base this on the bottom three panels in Figure 4.9.

6. Add captions and transitions

Figure 4.13 displays the flow of the tables and plots that we established in step 5. That is, we put the figures and text together to form our storyboard. The captions and transitions shown in the figure reflect the thought process from the earlier steps in storyboarding.

Taking Notes on the DAWN Storyboard



Drug-related ER visits affect people of all ages; however in this story we will focus on the youth (5 and under) and the elderly (65 and over). These groups make up 5% and 15% of the drug related ER visits, respectively.

Source	Age Group					
	≤ 5	5 - 17	18 - 24	25 - 44	45 - 64	65+
DAWN	5.8	7.0	14.7	32.2	25.5	14.8
Census	6.5	17.5	9.9	26.6	26.4	13.0

The age distribution of DAWN subjects closely matches the US population for these two groups. This is not the case for all ages. The 5 to 17 year olds are underrepresented and 18-24 year olds are overrepresented in DAWN.

A large proportion of drug-related visits to the ER for both youths and elderly is due to an adverse reaction.

The second most common reason for the elderly is over medication

Roughly 25% of the under 5 visit because of accidental ingestion.

Youths visit the ER most often after 6 pm

After a drug-related ER visit, older patients increasingly are sent to inpatient care.

Figure 4.13: Example Storyboard. Here is an example of a storyboard coming together. Plots with a similar theme are gathered and summaries are added to them. From the original two tables and seven plots (with 38 panels in total) we winnowed down to one table and three plots to tell the story.

With the fresh storyboard, we make notes of the supplemental information we expect to need for our first draft.

Our analysis relies on the the soundness of the DAWN survey. Basic information about this survey is important for the reader to understand and believe our results. We want to identify records and how they were selected for the survey. Our analysis consists primarily of comparing percentages, and we will want to provide a measure of the accuracy of these percentage.

We eliminated the portions of the analysis that examined sex and season. We note these deletions and that we do not plan to include them as discussion points. This note serves as a record of our decision.

For references, we list the source of the data and description of the DAWN survey methodology [184], the source for Census figures [74], and references most relevant to the two age groups under study, such as [163, 23, 118].

Iterate Over the Storyboard

	Age Group				
Source	5 & under	6 to 11	12 to 17	18 to 24	25 to 34
DAWN %	5.8	1.4	5.6	14.7	18.0
2010 Census %	7.9	8.0	8.3	9.8	13.3
Visits / 100,000 people	1201	288	1107	2455	2217

	35 to 44	45 to 54	55 to 64	65-plus
DAWN %	14.3	15.3	10.2	14.8
2010 Census %	13.3	14.6	11.8	13.0
Visits/ 100,000 people	1762	1718	1416	1864

Table 4.5: Age Distribution Table for Drug-Related Emergency Department Visits. The top row of the table displays the percentage of drug-related visits to the emergency department of a hospital by age group, and second row gives US Census figures. The third row provides rates of the number of visits per 100,000 population. Together the youngest and oldest age groups account for 20% of all emergency department visits, which roughly matches their presence in the population (21%). Most noticeable is the rate of visits for the youngest is four times the rate for 6 to 11 year olds.

We continue to review the storyboard and ask ourselves whether our argument holds together or needs additional work. For example, when we created the storyboard, we identified a hole in the lack of comparison figures for the age distribution and so added comparative statistics from the 2010 US Census (Table 4.4). As we iterate again, we consider whether the revised table adequately addresses the problem. What is the best way to illustrate the vulnerability of the youngest and oldest populations? Rates that normalize the number of visits by the size of the population are likely to offer a more informative comparison. We have identified a new hole in the hole that we patched earlier. This time, our patch includes an additional row to the table that contains the number of visits per 100,000 people in each age group (see Table 4.5).

We revisit our notes and add an item about the definition of the rates. Also when we computed these rates, we found that the DAWN age categories don't match exactly the Census categories. Since we made approximations, we add a note about this.

After further contemplation, we deem the plot of time of day unnecessary because it doesn't focus adequately on our age groups. Instead, we decide to include some of the statistics from that plot rather than the plot itself. We make a note of this decision, but for the time being, we leave the plot in the storyboard as a placeholder.

In these iterations we continue to push our argument forward. We trim away unnecessary information, fill holes, and carefully prepare a defensible and convincing argument. Note taking helps us keep track of the decisions that we make and why.

Writing the First Draft

This chapter describes the typical structure of a formal article and provides guidance about what should be included and emphasized in each section of a paper. Throughout, we reference the They Say/I Say framework of placing our own work in the context of other's work while distinguishing what sets our work apart [61].

We give examples of different structures that a formal report could take depending on the venue. We also discuss differing audiences and introduce the idea of the secondary audience to have an eye towards broader impact.

Then we go on to talk about the different sections of a typical report and how to tackle writing the first draft of each. We advocate for starting in the middle and writing about the data, methods, and results as well as choosing figures and writing accompanying captions. It can be easiest to gain traction in the middle of the report as we have already organized our thoughts via a storyboard and taken note of particular details that we need to communicate to our audience. Then we recommend writing the end of the report, typically made up of a discussion and conclusion. The beginning of a paper, including the background, introduction, abstract, title, and keywords, can be the hardest to write, so we recommend saving it for last after we've had a chance to synthesize our own work.

Here we explain key features of the discussion and introduction sections, comparing and contrasting their function in a formal report.

Excerpt: *Discussion Section*
The discussion section takes a step back and gives our specific work a more general perspective. We want to broaden from the particular details of our findings to the field we are working in. After reading our discussion, a reader should know:

- What are the key features of our analysis and results?
- Do our results confirm or contradict our initial expectations or earlier work?
- What problems did we run into, and if they still remain unsolved, what are our suggestions for starting to address them?
- Knowing what we know now, after doing our analysis, what recommendations do we have?
- What should someone work on next to build off of our work?

Be careful of overstating your work. Be honest about the obstacles you faced and the generalizability of your findings. This does not weaken your report, but rather helps to ensure that your work is appropriately understood and used by others.

As an example, we write a discussion section based on our DAWN analysis from the previous chapter. Our discussion begins with the limitations of the study so that the reader understands any problems with generalizability of the study in other contexts.

> The DAWN data give a complete picture of drug-related ED visits across the United States by age, sex, race, and drug type that was previously inaccessible due to the lack of a comprehensive survey. However, there are some limitations that affect the results in this report. Information on race and ethnicity is often sparse, and some hospitals do not report this information at all since they consider this to be private information. Without race being consistently reported, we are not able to investigate how this factor interacts with ED visits. Another data limitation pertains to the variety of pharmaceuticals involved in the visits. Variety may be overstated because it can be challenging to determine whether or not a patient's current medications are unrelated to the visit. This potential bias could be a greater problem for older patients since they are often on multiple medications.

The next paragraph summarizes the findings for the youngest age groups, and makes suggestions for how to act on the findings with the goal of decreasing the number of drug related ED visits..

> We found adverse reactions and accidental ingestion to be the major types of drug-related ED visits for the youngest patients. Additionally, we found that visits for this group often occur during the evening when they are most likely home and under their parent's control. Emphasis on better storage of medication in households with small children has the potential to reduce drug-related ED visits by youths. In addition, better education for new parents about the dangers of medications in the household could possibly help decrease the number of incidents.

The end of the discussion summarizes the main features of drug-related visits for the elderly, provides a hypothesis of why they may occur, and offers suggestions for how to act on the findings to decrease the number of visits in this vulnerable age group.

> Our investigation also found that visits from the elderly were commonly due to adverse reactions, and together with over-medication these cases make up nearly 95% of drug-related visits to the ED in this age group. The increase in inpatient services after the ED visit for the elderly may be attributable to the opportunity for a closer evaluation of the person's ability to care for themselves or to deterioration of other health conditions brought on by the drug-related incident. For the elderly, clearer instructions and more monitoring for prescribed drugs (including interactions between multiple drugs) could help reduce drug-related ED visits and keep them out of in-patient care.

■

Excerpt: *Introduction Section*

The introduction usually goes in the opposite direction of the discussion section. Rather than move from the details of our study to the broader field, an introduction first gives an overview of the broader field and then identifies the specific question we are trying to answer within that space. When describing a statistical analysis, we must identify both the scientific and statistical question of interest and explain how our data and results provide answers to those questions.

After reading the introduction, a reader should know:

- What is the problem?
- What is the motivation?
- Why is it important?
- What was found?

You may think that you want to save your main conclusions for the end of the report, but you want to state them in the introduction as well. Telling the reader ahead of time what you found peaks their curiosity in how you obtained the results and may convince them to continue to read.

Typically, at the end of the introduction, the rest of the paper is outlined. A roadmap for the paper helps the reader know what to expect and helps them skip over sections to read what interests them.

The beginning of an introduction for our DAWN analysis identifies the subgroups of interest and explains why we were previously unable to explore their drug-related emergency visit behavior.

> We often think about the rise in drug use in the United States in terms of illegal drugs and the teenage to middle-aged populations, but understanding the types of drug use that impacts vulnerable age groups such as young children and the elderly is also an important consideration. However, assessing heterogeneity in drug-related emergencies across sex, race, age, etc. has been challenging due to a lack of comprehensive data.

The introduction continues by explaining what has changed so that we can now study drug-related visits more easily and completely.

> To address this gap in knowledge, the United States Department of Health and Human Services consolidated data from the Drug Abuse and Warning Network (DAWN). These data contain information about drug-related visits to hospital emergency departments (EDs) in over 250 hospitals across the county.

The introduction ends with a more specific description of the subpopulation of interest and an overview of the major findings.

> We focus on the youngest (five and under) and oldest (65 and over) age groups because to some extent, these drug-related ED visits could be prevented by increased parental or medical supervision. Combined, these two groups make up more than one in five drug-related visits to emergency departments. For the youngest children, visits are due to an adverse reaction or accidental ingestion, and visits for children five and under are over-represented when compared to the 6 to 11 year olds. The elderly population's drug-related ED visits are primarily due to an adverse reaction or over medication, and the proportion of these patients returning home directly after the ED steadily declines with age.

■

4.6 Part IV: Editing and Revising

Pedagogical Context for Editing and Revising

Many students write reports for their course projects, but because these projects typically occur at the end of a course, they rarely get the opportunity to revise their work based on instructor feedback. This part of the book gives students guidance on how to revise their work both through strategies they can use on their own and through eliciting feedback from a peer.

Chapter 8 addresses the specific challenges of writing about statistics. These differ from challenges faced by those writing about science more generally. Words in their statistical context have different interpretations than their everyday usage which can cause confusion for non-statistical audiences. This phenomenon is called "lexical ambiguity." Kaplan et al. study how undergraduate students defined common statistical terms that also have a common usage (e.g. "confidence" and "random") at the start of an introductory statistics source [85]. They found, for example, that students associate "confidence" to mean "a high degree of assurance" while the statistical meaning has a specific "level of surety based on probability." These double meanings may cause confusion when introductory students write about their statistical findings or when a non-statistician reads a more statistical report. Chapter 9 contains more classical writing advice (like in [1]) but provides it in the context of data and statistics examples.

Chapter 10 covers revision more broadly including the use of peer review to get feedback. Peer review is not just important for building writing skills; it is also part of the scientific process. Guilford provides a teaching method for walking undergraduates through the publishing process [64]. Students had to write a letter of inquiry, write a research paper, submit a draft for peer review, revise the paper based on feedback, and resubmit as well as provide feedback to another peer given guidelines for review. Our course implemented this strategy, replacing the letter of inquiry with storyboarding, and followed the material in Chapters 6, 7, and 10.

Cho and Schunn give more guidance on the peer review process itself and describe a "scaffolded writing and rewriting the discipline" approach [27]. Their peer review occurs online for a large course, but we can easily transfer their approach to smaller, in-person courses. Peer reviewers evaluate work along three dimensions: flow, logic, and insight. Those who receive feedback also provide "back-review" about how helpful the feedback was when they were revising their work. To evaluate "flow" the overall prose is examined for whether or not the main points and transitions are clear. To evaluate "logic" the structure of the argument is examined for whether or not it is convincing. To evaluate "insight" the content itself is evaluated for whether it contributes new knowledge to the reader. Our rubrics for peer and instructor review resemble these criteria and include overall grammar, structural organization, and statistical accuracy (Activity 10.8.2).

Taking Care with Statistical Terms

It is challenging to craft clear sentences, choose appropriate words, and convey findings in a compelling manner that is faithful to the data and avoids overstating implications. One of the first things we want to check when re-reading a draft of our work is that the statistical content is accurately portrayed. This chapter provides advice on how to differentiate statistical terminology from everyday language, represent numbers in text, incorporate mathematical expressions, and choose the correct nouns and adjectives (e.g., fewer or less, percent or percentage). Examples follow.

Excerpt: *Statistical Terms and Everyday Usage*
Many statistical terms give unique and specific meaning to words from everyday language. Examples of these include *confidence*, *error*, *sensitivity*, and *significant*. To reduce confusion, we recommend that you avoid these words in their common usage and exercise care when writing them in their technical setting. Depending on the audience, it can help to clarify the statistical meaning of a term.

An *error* in plain English typically describes a mistake or something wrong. Statistical error is quite different. It refers to a difference, such as the difference between an individual measurement and the average of several measurements. The terms *margin of error*, *measurement error*, *sampling error*, and *standard error* are a few examples of statistical error, each of which has a precise statistical definition.

Given so many related, but distinct, terms, it is important to clearly distinguish between them in our writing. For example, The New York Times' 2006 Polling Standards [122], states that articles containing polling results should give the margin of error and have an explanation of what that margin of error means. In a report by the Pew Research Center on smartphone usage [169], the margin of error is provided for their survey without any explanation. Alternatively, the report might include an additional sentence that briefly explains margin of error, e.g. as suggested by the Times Polling Standards [122],

> This means that in 95 cases out of 100, overall results based on such samples will differ by no more than 2.5 percentage points in either direction from what would

> have been obtained by seeking out all American adults.

■

Excerpt: *Help the Reader with Statistical Terms*
Even though a data analysis is inherently technical, we want our writing to be understandable. Using technical terms and acronyms without first defining them alienates the reader.

For example, the acronym "GLM" is a blackbox term in the next sentence.

> We use a GLM to predict whether an e-mail is spam or not.

Below, we define the abbreviation and explain the term by juxtaposing something with which the reader is already familiar.

> Since the prediction of whether an e-mail is spam or not is a binary decision, we use a generalized linear model (GLM) that allows for a different distribution of errors than the traditional normal distribution.

Notice that we have also justified the use of a GLM. ■

Excerpt: *Similar Words with Distinct Meanings*
When we write about data, we discuss percentages, mean effects, above and below average, approximate values, etc. Many times we are confronted with a choice between similar words, such as percent or percentage, affect or effect, above or over, and about or approximate. The precise definitions for each of these terms matters, and to write well, you will need to select the correct word.

For instance, a *percent* refers to a specific number, such as 20 percent, and a *percentage* refers to an unspecified portion, such as a large percentage. That is, *percentage* is used without numbers. When we take the difference between two percentages, that difference is measured in *percentage* points, e.g., 20 percent is 3 percentage points more than 17 percent.

The following examples demonstrate the usage of *percent*, *percentage*, and *percentage point.*

> Twenty *percent* of the eligible voters did not vote; this is a large *percentage.*

> A greater *percentage* of eighth-grade students met the state standards than third-graders.

> About *40 percent* of mothers smoked when pregnant in the '60s compared to *8 percent* today; that's a decrease of *32 percentage points*, or an *80 percent* decrease.

Although incorrect, the following use of *percent* has become common.

> What *percent of your time* do you spend watching TV?

The more grammatically correct way uses *percentage of your time.*

■

This chapter also cautions against absolute statements and slipping into causal statements when they aren't warranted. These statements are especially problematic if they arise in writing for broader audiences. We provide an example of a press release that avoids causal language and one that is potentially misleading.

Excerpt: *Avoiding Causal Exaggerations in Press Releases*

The first sentence of the press release about the correctional officer survey [183] makes a causal claim, using the phrase "make them more likely" to connect exposure to traumatic events with mental health outcomes.

The study is a survey so it is difficult to assert causality. We intuitively believe this to be the case and presumably other studies support the link between traumatic events and depression. However, a more accurate description of the findings would point out the link without implying causation.

In contrast, the first paragraph of a press release about a study of children with different religious upbringings and their health in their adolescence [176] begins with a statement about how spiritual practice may be a protective factor for health in adolescence and early adulthood. The choice of *may be* is not a strong statement, but the press release is careful to not overstate the findings.

■

Crafting Words and Sentences

This chapter focuses on strengthening the details in our writing. We give examples of how to trim unnecessarily complicated phrases, write straightforwardly, use active verbs and concrete nouns, balance specific and general statements, and smoothly transition between ideas. We also provide some general guidance on commonly made grammar mistakes. Examples follow.

Excerpt: *Eliminate Empty Phrases*
Phrases that contain no information, such as *it is interesting to note that, the fact that*, and *it should be pointed out that*, should be eliminated. If something wasn't interesting then we would not be writing about it. If something is more interesting than other things, then we can find a more compelling way to draw attention to it.

Particularly offensive empty phrases are pompous ones, such as *as is well known*, *of course*, *clearly demonstrate*, and *it is obvious.* These are unnecessary and annoying to the reader.

The following sentence explains the reasons for analyzing a subset of the available data. It takes a cumbersome approach that includes unnecessary phrases, redundant information, and clunky descriptors.

> In this part of our analysis, we assume that flight delays that last shorter than 15 minutes have minimal effects on passengers, and so we reduce our large dataset into a smaller subset in which all departure delays are at least fifteen minutes long.

A more streamlined version is

> Since short departure delays have minimal impact on travelers, we analyzed only those flights where the delay was longer than 15 minutes.

The opening phrase *In this part of our analysis* does not contain information so we dropped it. The adjectives *large* and *smaller* don't add information, and the process of reducing the dataset doesn't need to be described. The essential information is that a subset of the data was analyzed. Notice too that we replaced the clunky description *last shorter than 15 minutes* with *longer than 15 minutes* and mentioned the criteria for subsetting only once.

Note: The term "fat phrase" from comes from Alley's *The Craft of Scientific Writing* [1] while "empty phrase" is also discussed in Andrew Gelman's blog post [58]. ■

Excerpt: *Tell What You Found, Not the Path You Traveled*
For beginning scientists, one of the hardest skills to learn about writing is to avoid presenting findings in the order in which the analysis was carried out. There are a few circumstances where we do want to write about the process in chronological order, for example, in a blog post that demonstrates the thought process behind an analysis, but most often we want to present a summary of our findings not a description of the analysis process. For example, if we fitted several alternative models to the data, then we typically mention these models in a few sentences in a discussion section and do not dedicate space to describing these alternatives. As another example, exploratory data analysis is an important and often lengthy stage of the analysis that can uncover problems with the data, the need for transformations, and unexpected relationships to consider when modeling. It is important for replicability that we describe these findings and make our code available so that the reader knows exactly what we did to the data and whether there was data snooping, but this description typically constitutes a brief summary in our written report.

The following paragraph justifies the researcher's transformation of the data.

> During my exploratory data analysis, I found that the distribution of house prices is heavily right tailed. Hence, I applied a log transform to housing price and achieved a relatively normal distribution.

The relevant point is that the log transformed data have an approximate normal distribution.

> The log transformation of house price follows an approximate normal distribution, so we analyze the transformed data.

We trimmed the phrase *during my exploratory data analysis* and the description of the untransformed data. If the distribution of the log-transformed data looks normal, then the distribution of the untransformed data must be right skewed. ■

In this chapter we also discuss differences in word and sentence level choices between blogs and formal writing. Because a blog is written informally, this form of communication gives the writer more flexibility. Blogs “break the rules” in a variety of ways including using empty phrases, including information about the data analysis process to add insight into the writer’s thought process, and using conversational language to connect with the reader and increase accessibility.

Excerpt: *Ourselves as a Character in Blogs*
The use of conversational language helps us transition away from our more guarded writing. Although in our formal writing we try to remove ourselves from the content we are discussing, a blog allows us to blur the line between our professional and personal ties to the content. In a blog post, the content may be interjected with personal anecdotes or information. Similarly, unlike in a formal report where the emphasis is placed on telling the reader what you found, a blog post may be the venue where you want to insert yourself into the story and talk about the path you traveled. Many blog posts document the process behind work that is more formally explained elsewhere.

For example, three sentences (adapted from content in [162]) can capture an analysis of music data grouped by different workouts. These sentences describe the findings without any process information.

> Tracks associated with certain muscle groups have more missing values than others, so the representativeness of the sample may be a problem. The shoulders track is associated with less pop or dance, and the cool down track is quieter with no electronic music. Future work could study beats per minute data.

The blog version is noticeably longer because intermediate steps and conjectures are added [162]. The reader not only gets to read about the results but also how the author found those results (e.g. what plots were made). The steps are outlined in a stream of consciousness style, interjected with mini-brainstorms the author has as they go through the analysis process.

■

Revising: Drafts #2 Through...

Although writing a first draft can be intimidating, the revision process is often the hardest part of writing. It can be time consuming, and progress can be hard to see. This chapter provides a plan of attack for revising our work so that we can attain a polished draft efficiently.

Typically, we edit considerably before showing others our writing because we want a reviewer to focus on our ideas, not our writing weaknesses. However, enlisting a peer to give feedback before the final draft can give a helpful, new perspective. If we make sure to clean up the small details such as spelling, grammar, etc. a reviewer can focus on the higher level aspects of our writing such as content and structure. The reader’s role grows in the revision and editing process. Not only are we reading to take inspiration from writers we admire, now we are reading with a critical eye, looking for deviations from good practices.

Excerpt: *Preparing to Rewrite*

Once we have our ideas written into a first draft, we typically refine the draft many times before it becomes a polished product. In this process, we often need to get some distance from our draft. Distance can be created in both time and format.

Time permitting, it can be helpful to work on something other than our paper for a few days. After we have spent some time away from our paper, we can revisit our writing with fresh eyes. Before revisiting the draft, we suggest that you write down what you expect your reader to think about the topic before and after reading your report. Then, you can reread the paper and see if your intention matches the reality.

Changing the spacing and printing the draft so that the draft looks physically different can be helpful to build distance in formatting. It can be easy to overlook mistakes on a computer screen, especially when reading the same thing over and over again. Reading out loud can be useful in this situation. When we read in our head, we often subconsciously fix mistakes without recognizing that they are there. When we read out loud slowly, we notice small grammatical details as well as higher level problems such as ideas that do not flow or abrupt transitions. As you read aloud, we recommend that you don't stop to fix anything. Instead, mark what you want to fix with a small comment on what you have in mind and then come back to it later. You will want to examine the draft as a whole without getting sidetracked with the specifics.

Once we have caught the more obvious flaws in our writing, it can be helpful to revise with more specific goals in mind. By working with different parts of the paper in different revision stages, we also ensure that we don't fatigue of rereading the whole document over and over again.

Note: Tips about revising your writing can come from a variety of sources including Alley's *The Craft of Scientific Writing* (e.g., needing "distance" from a draft) [1]. ■

After a round of general proofreading, it can be helpful to target problem areas in different rounds of revision. Focused reading can help us catch flaws that we may not notice when we read with a more global emphasis. In this chapter we explain how thinking about our writing weaknesses can help focus our revisions.

Excerpt: *Targeted Revision: Lack of Focus*

If you suspect that your writing lacks focus, read through your article looking for backtracking. Note all the places where the subject changes. Do you ever start one idea, jump to another one, and then come back to the original idea? A lack of focus can easily happen in early drafts when our main goal is to get all of our ideas written down. In the revision process we can see the whole and reorganize to avoid this stream of consciousness style.

The following paragraph (adapted from student work [142]) bounces back and forth between ideas rather than organizing the ideas and presenting them in a logical, compelling order.

> A mistake I notice is that my students are not good with word problems. It's pretty hard to visualize and understand information from word problems. For example, if a question used the phrase "How much more," you may not know that you will compare two things, probably using subtraction. Since some kids don't know what the phrase means in context, I would help them draw and ask guiding questions to take it step by step. There are other phrases in word problems too, but teaching them how to visualize by drawing out the pictures helps a lot. Another tricky type of word problem is converting between units, like 12 inches is equal to 1 foot. Students do not know how to visualize length and do not understand how units work. It's good to have examples to show them the difference between the units, like a ruler or measuring cup.

Here we group ideas and order them in a way that keeps the narrative moving in one direction.

> I notice that word problems are tricky for my students. I think this is because they have a hard time visualizing the problem, interpreting certain phrases in the problems, and avoiding pitfalls such as changes in units. For example, if a question used the phrase "How much more," students may not understand that this is a comparison problem involving subtraction. Similarly if the problem involves a conversion between units, like inches to feet, students may not be able to identify the main objective of the problem. Both problems can be made more concrete with examples. Drawing a picture of the groups being compared can help show the difference in size in terms of "how much more." Physical objects, such as a ruler or measuring cup, can help show the difference between units. My strategy to help students is to have them draw the problem, ask guiding questions, and then go step by step.

This example shows how complex a revision can be. ■

Giving Feedback

Before reading a draft and giving feedback to a peer, it is helpful to make a plan of what to read for. Knowing where they aim to submit their work can help assess whether the structure is appropriate and if the content is presented at the right level for the intended audience. Having a sense for what they are most concerned about in the paper can also help us target the review and focus our energy on the weaker aspects of the draft. If the review is for a journal instead of a friend, reviewing the journal guidelines will help give structure to the revision process.

Excerpt: *Reviewer's Template*

As you read a peer's work, consider the following prompts to guide your review.

- First, skim through the paper to get a sense for the organizational structure. Are all the major components of the report accounted for?

- Next, read through the paper without making any edits. Summarize the paper in a few sentences.
- What are you confused about and why? (i.e., what remains unclear?)
- What are you curious about and why? (i.e., what details would you like to see added?)
- Read through the paper again, and now make comments and edits. Underline sections that are particularly clear/well written, and try to articulate why.
- Try to avoid copyediting. Focus on organization and flow rather than comma usage and spelling.
- Look through your comments and edits. Organize your feedback into themes. Try to avoid just giving an unstructured list of comments. What is one thing you would have done differently and why?
- Don't be afraid to give serious constructive feedback. You aren't being mean; you are being helpful.
- End your feedback by identifying the biggest strength of the paper.

The questions to answer for peer review in this section were partially inspired by a response template for peer review shared with the authors by Kathleen Donegan (UC Berkeley English Department), teaching resources from University of Michigan's Sweetland Center for Writing [185], and sample peer review materials (no longer accessible online) from Brandeis University's University Writing Program. ■

Receiving Feedback

Receiving feedback can be difficult, but feedback is meant to help us strengthen our writing, not act as a judgement of our writing ability. Since we are naturally invested in our work, we can be resistant to making the major changes that a reviewer advocates for. It is reasonable to first read through the comments and take a minute to be defensive. What are the reasons for ignoring the advice of the reviewer? Then we can take a higher level approach and consider where the reviewer is coming from, realizing that the feedback is not personal. Finally we identify the common themes within the feedback. Are the main criticisms about content, big picture writing aspects such as structure and clarity, or detailed aspects of the writing such as word choice or sentence flow? Once we've synthesized the feedback we can create a revision to-do list that is specific and contains all of the comments made by our reviewer to help us stay organized.

Excerpt: *Feedback about Narrative*

The comments below pertain to the emphasis and narrative of our article. The first concerns the visualizations.

> Given that a key aspect of the paper is data visualization, one way to improve the paper would be to suggest a few more graphical displays. More specifically, images that are more exciting and tell a stronger story about the data would be useful.

The second comment indicates that the reviewer is not satisfied with the data description.

> It seems to me that the main uses of this dataset are to illustrate the many difficulties involved in collecting good data. A few of these difficulties are discussed briefly, but I would advocate an expansion of these sections.

We could strengthen our argument by focusing on the challenges inherent in the type of data we are working with and making our graphics more compelling. To do this, we might revisit the figures in the paper and determine whether the story is clear in the visualization and whether the captions are as informative as possible. We should also consider whether a different type of plot might better emphasize the story. After editing the figures, we can give them to a peer to see if they understand what the graphs are trying to say and if they find them compelling.

As for the data description, we can start by brainstorming difficulties in the collection process. We might draft a new section that focuses on these ideas. If we add a new section, then we need to consider where this section best fits in the paper and adjust the transitions as needed. ■

In the revision process, we should consider whether we have made a convincing and compelling argument to the reader about our findings. We also want to keep in mind the venues that we are interested in submitting our work to. While in the revision phase we should make sure that we have written at the appropriate level for the intended audience and that the intended audience matches the readership of the prospective venue.

4.7 Part V: Science Writing and You

Pedagogical Context for Science Writing and You

Beyond teaching students the skills they need to join an academic discipline, we as teachers also need to introduce students to the academic community and help them form their identity as part of that community. This community includes the science discourse community discussed in Yore et al. and the community of practice discussed in Hunger et al. [200, 79].

Yore et al. assess scientists' writing practices and how they see writing fitting into their science [200]. Scientists reported that "writing, reviewing, and revising helped improve the clarity and understanding of the embedded science ideas," identified their discourse communities as "social," and considered "writing as knowledge building instead of just knowledge telling." Even though they note that novice scientists typically enter their field's discourse community as part of their graduate research, we aim to make this introduction earlier, at the undergraduate level. Chapter 11 of our book focuses on community and identity. This closing chapter emphasizes that being a scientist involves being a writer and that science has a social component (e.g. conference networks (Section 11.1), social media (Sections 11.2 and 11.4), and online coding communities (Section 11.3)).

To allow students the freedom to form their own identity and shape their own experience in the class, we used an apprentice-style approach in our Communicating with Data course. We modeled examples of effective and ineffective writing, wrote together as a class and in small groups, and used peer review in addition to instructor feedback. Throughout the course, students created a portfolio of written work (Chapter 12), choosing from a variety of low stakes prompts (evaluated on check, check-plus, check-minus system) each week. Student motivation was high, since exercises could be repurposed outside the classroom as evidence of their writing and data analysis skills. Students chose their favorite portfolio pieces to refine based on peer and instructor feedback into final "products" (analogous to a final project report).

Students learned to write by frequently writing while the choose-your-own-adventure style allowed students to have flexibility and autonomy. Others have advocated for this apprentice approach. For example, Hunter et al. explain how an apprentice-style of teaching is used to introduce undergraduates to research [79]. The authors found that students saw their experience with undergraduate research as helpful in seeing how science works in practice and helpful for personal growth.

Embracing Your Role as a Scientist

Rather than providing excerpts, this full chapter is included below.

Polished draft in hand, you may ask yourself, now what? It is now time to (finally) share your work with the world. So far this book has aimed to teach you how to read as a writer, prepare to write as a writer, write as a writer, and edit as a writer. We hope you now feel comfortable identifying as a writer (because you definitely are one). Now we want to give some advice about wielding the identity of a writer as you advocate for your science.

Despite being both a scientist and a person existing beyond our work we often must separate our personal feelings from the work we do to remain objective. It can be helpful to have a network (a source of professional support), a research focus (professional interests), a community (a source of personal support), and an identity (personal interests) to help balance between the professional and personal. Your community helps support your communication within your own circle of influence while your network can help you reach across community boundaries. Your research focus and identity help you navigate the social aspect of research by signaling what interests you.

This chapter discusses opportunities to embrace the social aspects of communicating your work and gives advice about venues for sharing your work beyond the page.

Expanding Our Professional Network

To expand our professional network we often must physically leave our place of work or study to meet others and showcase our work. These meetings may occur in annual conferences for professional societies or more informally at “meet up” groups (i.e., in-person events organized online by people with similar interests).

When we seek others to add to our network we can’t expect someone we have just met to read our paper while we stand there waiting for them to get caught up on what we are working on. Instead, to take advantage of the meeting we rely on oral communication to give them an overview of what we are working on. This type of communication can take the form of a formal talk, a lightning or speed talk, or a poster. It can also take the form of a one-on-one conversation.

Up until this point we have talked about written communication, but many of the strategies and advice we gave for preparing written material applies to preparing formal talks and being ready to chat informally with new people. Each of the formal oral communication venues mentioned above has an analogous written form. A formal talk is like a formal paper, a lightning or speed talk is like a press release, and a poster is like a blog post. More informal networking opportunities happen in unstructured social settings and don’t have an obvious writing analogue.

Formal Talk

An effective talk is like an effective paper; it has an organized structure, a well-defined audience, a compelling narrative, and is delivered clearly. A talk has many of the same elements as a paper. We start with an introduction of ourselves and the background of the problem we are working on, we give some grounding in the literature (e.g., what has already been done, what holes are we trying to fill), we launch into the heart of our solution/approach, and we conclude with some discussion of the impact of our approach and an announcement of any future work we have planned. Just as we would storyboard before drafting a paper, we can also storyboard before a talk. In some ways the storyboard is even more natural in the case of a talk that involves slides. Each note card represents a slide, and we can physically arrange them to design the flow of the talk.

The level of detail provided in each part of the talk, and the relative length of each part depends on the venue. If we are at a conference for our particular sub-field or in a section dedicated to our sub-field, we may assume our audience will have a lot of the background needed to follow along with our work. We can then focus more of our attention on our approach. At a conference with a broad theme or in a more interdisciplinary venue, we may want to heavily emphasize the background and context while only providing the high level version of our approach.

Just as we follow the journal length requirements for a formal piece of writing, in a formal talk we conform to the venue time requirements. Having the right talk length is essential. We do not want to be rushing through slides because we have too much material to cover, and we don’t want to run out of things to say. A good pacing guideline is to have no more than one slide per minute of talk time allotted. You may spend an extra minute or two on a few key slides. Once our slides have solidified in order and content, we recommend that you practice the talk and time yourself. Make sure you have time left for questions (a good rule of thumb is about five or ten minutes of questions in an hour long talk). Also consider that you might talk a little faster during the actual talk if you are nervous; remember to talk slowly, even if it feels too slow.

One important distinction between a paper and a talk is that we do not want us or our audience to be doing a lot of reading during our talk. Keeping our slides uncluttered with minimal words and one or two figures will ensure the focus is on what we are saying. Including visualizations, rather than blocks of text, can help guide a listener through our narrative and keep their attention. Just as we don't want to give our listener blocks of formal text, we do not recommend that you write out the talk and read it word for word. Having an outline and perhaps a few notecards with phrases to jog our memory if we get stuck are fine, but relying on rigid text in a talk format makes it more obvious if we slip up. We want to aim for a natural delivery, as if we were just talking with a friend. This ease comes from practice and preparation, not memorization.

Remember that like our paper, the talk is a showcase for our work. The important thing is that after the talk, listeners can follow up with us to ask questions and find out more about our work. We recommend that you make followup easy for the audience. Provide contact information on your slides, and if possible, remain at the venue location after your talk to mingle and give people the opportunity to ask questions one-on-one rather than in front of a crowd.

Lightning/Speed Talk

A formal talk showcases our work in a venue where there is ample time to address at least some details. In contrast, a lightning or speed talk (often only about five minutes long) is an abbreviated advertisement for our work. We cannot possibly explain our entire project in 5 or 10 minutes, so we need to give the audience enough detail to understand the big picture and be invested in the answer to our question. We recommend that you think of this style of talk as an elevator pitch for your work; make the listener want to come talk to you afterward to learn more.

We have a short amount of time to make people interested in talking with us further. Like a press release, this format of talk requires that we put the important information first, avoid getting into technical details, and have an enticing ending that will make listeners want to follow up with us.

Preparation for a short talk is much like preparation for a long talk, although we do not spend any time on details. We can storyboard and practice the talk with a timer. However it is important to note that just because the talk is short, this doesn't mean the talk is easy to give. It can actually be much harder to give a short talk without preparation than an unpracticed long talk. In a short presentation, your pace must be precise; there is not a lot of leeway for stumbling due to nerves. We recommend that you practice, practice, practice.

Poster

A poster is a low-stakes way to advertise our work. Sometimes a lightning talk is paired with a poster so that listeners who were intrigued by our talk can follow up with us. Other times posters are there for conference attendees to browse on their own time.

Think of a poster like a blog post. A viewer stumbles upon it and must decide whether to commit to engaging or rather move on to one of the many other posts out there. To capture and hold the attention of a viewer our poster should have a succinct story and a striking visual. Bullet points and white space are powerful tools. Just as we avoided slides crammed with text, our poster should also be free of dense blocks of text.

Storyboarding can also be used to prepare our poster. In this scenario each note card in our storyboard is a "block" of our poster. Using these blocks of content we can explore potential roadmaps of how a viewer will step through our poster. Left to right? Top to bottom? A mix, separated by columns? You can decide what visually makes sense.

A poster still requires talking to people. It can be helpful to also have a few minutes of commentary planned for a viewer who prefers to be guided through the poster. We suggest walking them through your main points, answering any specific questions they have, and taking note of any suggestions that they give you.

Networking

When we hear "networking" we often think about social interaction with strangers in artificial settings. However, networking does not have to have a superficial connotation. Any time we talk to a person, we are networking. Writing for broader audiences should strengthen our ability to talk casually about what we do.

Conferences are not just about giving and listening to talks; networking is also a major part. Take advantage of having many people with similar interests all together in the same area. Before you go to a conference, do some preparation. Reach out to people whose work you admire and see if they would like to meet during the conference. Professional relationships can also be formed serendipitously in between sessions and during meals and coffee breaks, so push yourself to start conversations with other conference attendees.

As we chat with others, we may want to formalize a connection by referencing a product of our work. However, our work is not always in a polished state when we find ourselves in these networking situations. Instead, we can post a draft of our formal paper on the arXiv, a freely accessed archive for scholarly articles, to document and provide access to our works in progress [6]. If we are not even at the formal drafting phase of research, we may consider writing occasional blog posts to keep track of our research process publicly.

Building a Research Focus

We have talked about expanding our professional network in the context of showcasing our work at different, in-person, venues. Connections can be made by attending conferences, visiting different campuses or departments, and taking advantage of the connections of those we already have a connection with within our network. By connecting with others with similar interests, we can find out about papers to read, job or internship opportunities to apply for, conferences to attend, and potential collaborators to work with. However, connections can also be made remotely. Academics have become active on social media (e.g., Twitter), taking advantage of its networking and information sharing power. Online connections can be made and leveraged at a faster pace. All it takes is one "influencer" to promote us to make other people aware of our work.

In order to fully take advantage of online community connections, we must make others aware of what we are interested in and what our area of expertise is. This way they can point us to relevant materials and know who to come to for questions on a particular topic. We build a professional identity by repeatedly showcasing our work, skills, and strengths. To help carve out our niche within the broader research community, we consider our expertise. Is our research focus deep; have we invented a particular method and know it inside and out? Is our focus more broad; do we solve problems in a wide range of scenarios? Is our focus domain specific; do we feel comfortable being the go-between for quantitative and non-quantitative people? Is our focus on tools; do we build software to simplify analyses or data wrangling? Is our focus on teaching; do we explain concepts well and advocate for our students?

Whereas at individual speaking events, we focus on exhibiting one piece of work, it is also important to have a complete body of work in an easily accessed place can help people explore at their leisure. Building a portfolio of writings and code and having it in an easy to link to place (e.g., GitHub, your personal website) so that we can provide links to resources when asked, is a great way to display our professional interests. A full portfolio can also display how our interests and expertise have changed over time.

We cannot always travel to talk about our work with others. Using social media to alert others when we have added something to our portfolio, whether it's a talk, a blog post, or a new draft, is a less resource-intensive way to get our work noticed. We can also use social media to solicit feedback on a draft of our work or advice on a question we are pondering. By crowdsourcing the feedback process, we can get faster feedback more frequently.

Promoting our work may feel like bragging, but exposure is an important part of being a researcher. We want to be part of scientific discourse and tell others what we have been working on so that others can learn from it, give feedback, and extend it. Small announcements via social media maintain a living curriculum vitae and also gives us a sense of milestones reached. The data analysis and writing processes can be intense, so it is nice to celebrate the little victories in a less formal venue. However, if we are going to take advantage of the speed of social media to promote work and get feedback, we need to be sure to return the favor. We should support others' work and give our feedback when requested.

Fostering a Personal Community

A professional network can help us build a career, but a personal community is also a key to our success. We cannot always be in work-mode, so we also need a source of support that is separate from our job.

Our personal communities can be related to our work (e.g., coding communities, question and answer communities) or they could be a bookclub, a knitting circle, a basketball team, or an improv group. We will primarily talk about communities that can be connected to your work in this section, but alternative communities are also an important source of support.

Coding communities ranging from open source projects to Stack Overflow can both provide help for work related efforts and support for our more creative endeavors. They can also be a way for us to give back or pay it forward by providing an opportunity for us to use our expertise to help others. "Side projects" where we use our work related skills in a non-work related setting can be a rewarding way to re-charge after a challenging day of data analysis or writing. The following examples provide insight into both contributing to and benefiting from these communities.

Excerpt: *Community Standards: Codes of Conduct*
Contributions to online communities do not have to be code. We can contribute content to Wikipedia. However, we must recognize that Wikipedia is an online community that has community standards and protocols that we must conform to. In this excerpt from Wikipedia's guidelines for contributing we are told about how we should conduct ourselves when we disagree with another editor [194].

> While discussing matters, it is very important that you conduct yourself with civility and assume good faith on the part of others. Edit warring (repeatedly overriding or reimplementing contributions) is highly discouraged. There is a bright-line rule called the three-revert rule, the violation of which may lead someone to be blocked from editing to prevent further disruption. Disruptive editing is not always intentional, as new editors may simply not understand the ins and outs of Wikipedia.

■

Excerpt: *Community Standards: Minimal Reproducible Example*
On Stack Overflow (and similar sites) there are guidelines for how to ask a question such that you are most likely to get a useful answer. The idea of a minimal, reproducible example (reprex) is outlined in this excerpt of the guidelines [174].

> Your code examples should be...
>
> - ...Minimal – Use as little code as possible that still produces the same problem
> - ...Complete – Provide all parts someone else needs to reproduce your problem in the question itself
> - ...Reproducible – Test the code you're about to provide to make sure it reproduces the problem

■

Other forms of personal community that can relate to our work are writer and accountability groups. Writer groups often are formalized peer review sessions. Writers each bring an excerpt of something they are working on and perhaps a specific aspect of the writing that they are most concerned about to the meeting and then ask for feedback from the group. You can start your own writer's group with people from your professional network. You may even want to meet virtually rather than in person.

Accountability groups provide a venue to set and report on goals to peers rather than supervisors, creating a judgment-free accountability mechanism. These goals could be a mix of work and personal. In these groups members take turns reporting on what they accomplished since the previous meeting, reflecting on why they did or did not meet their goals, and making revised goals for the upcoming time period. Group members can also provide advice for how to overcome obstacles. It can help if the membership of this group is not made up of individuals in our inner circle since maximum benefit comes from being able to be honest about our struggles and to fail without worry of judgement.

Welcoming Who We Are

We carry our identity with us throughout all of our professional endeavors. Sometimes our identity and our profession specifically intertwine. There are identity based conferences (e.g., Women in Data Science, Society for Advancement of Chicanos/Hispanics and Native Americans in Science (SACNAS), Grace Hopper Celebration) and identity based coding groups (e.g Girls Who Code, Black in AI) that provide an opportunity for sharing work, advice, and support among people with a shared identity experience. Other times we must carve out our own place, perhaps by leaning on our personal community.

Having a visible presence, on- or off-line, regardless of the main focus of the content, can increase the visibility of traditionally underrepresented groups in a particular field. Sharing our work and experience balancing personal and professional identities can bolster others who may be interested in a similar field but do not see many like them in it.

Excerpt: *Blogging to Bring Perspective*

In this blog post excerpt, Daly tells her personal story of her first time at a conference [35]. She does not specifically reference identity, but she is honest about feeling anxious and overwhelmed, helping to normalize those feelings.

This blog post acknowledges networking nerves. Readers can benefit from this vulnerability and moderate their own feelings of nervousness around new people. ■

Excerpt: *Blogging to Share Background*

In this blog post excerpt, D'az shares his experience as a first generation tech worker with the explicit goal of helping others facing a similar situation [33]. This blog post teaches us about the complexities of navigating between past and current identities. ■

Curating our professional persona is useful for our career, but it is also necessary to reflect on our personal values and goals. What drives us to do the work that we do and what matters to us beyond our work? As we communicate our work, we should take note of what energizes us and follow that energy as we move forward in our careers.

Although we can give you tools and advice for *how* to communicate your work effectively to a variety of audiences, we cannot guide you in *what* you ultimately say. We hope that by embracing your dual role as scientist and writer and using the skills you have developed by working through this book, you will go on to advocate for what you are passionate about in both your professional and personal spheres.

Building a Portfolio

This chapter of the book includes over 20 more extensive prompts, meant to supplement each chapter's set of activities, to help readers build a portfolio of written work. These prompts are all forward referenced according to which book chapter's skills they focus on. We provide general guidance for what to write and how much to write, but these prompts are meant to be freely adapted based on the reader/writer's interest.

Appendix A

Appendix

A.1 Identifiability Controversy Supporting Material

Partially Identifying Prevalence Using Presence-Only Data (with perfect detection)

For this discussion it will be useful to denote the occurrence probability ψ_i at site i as $\psi(x_i)$ to emphasize the dependency on the covariate x.

Each $\psi(x_i) \in (0, 1)$ so each $\alpha\psi(x_i) \in (0, 1)$. Therefore, $\alpha \in \left(0, \frac{1}{\sup_{x \in X} \psi(x)}\right)$ for the set X of possible realizations of the feature x. This bound on α translates into a bound for the overall prevalence ρ^*, $\rho^* \in \left(0, \frac{\rho}{\sup_{x \in X} \psi(x)}\right)$, where ρ is the prevalence before scaling. Note that $\frac{\sup_{x \in X} \psi(x)}{\rho} = \sup_{x \in X} \frac{\psi(x)}{\rho} = \sup_{x \in X} \frac{\pi(x|Y=1)}{\pi(x)}$. Then the bound can be estimated: $\rho^* \in \left(0, \inf_{x \in X} \frac{\pi(x)}{\pi(x|Y=1)}\right)$. A lower bound exists for $\sup_{x \in X} \frac{\pi(x|Y=1)}{\pi(x)}$, so an upper bound exists for $\inf_{x \in X} \frac{\pi(x)}{\pi(x|Y=1)}$ because $\pi(x)$ is assumed to be known and $\pi(x|Y = 1)$ is the observable distribution of the covariate x given $Y = 1$. In context, to get a narrower interval a big differential between $\pi(x|Y = 1)$ and $\pi(x)$ would need to exist. This would correspond to x being a strong predictor of occurrence. In practice, one could search for a region within the data that had a big difference. There is potential that this partial identifiability would suffice in practice depending on the research goals.

Single Visit Occurrence Example with Penalization

Figs A.1 and A.2 show that the penalization in the *detect* implementation does not mitigate problems due to a lack of nonparametric identifiability [172].

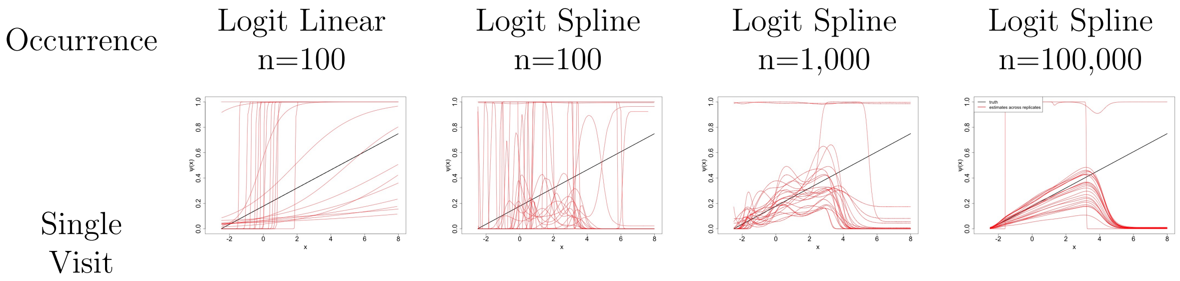


Figure A.1: The x-axis displays the value of the covariate that helps us predict occurrence. The y-axis displays the occurrence probabilities. Penalization in the implementation of the single-visit analysis does not make up for the lack of nonparametric identifiability for the occurrence probabilities.

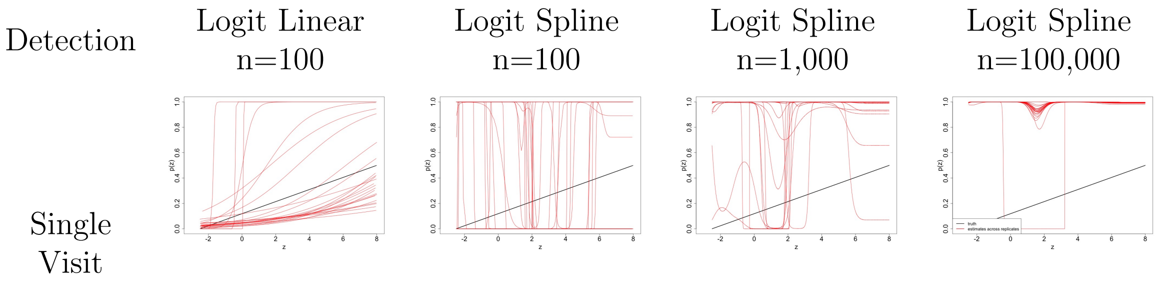


Figure A.2: The x-axis displays the value of the covariate that helps us predict detection. The y-axis displays the detection probabilities. Penalization in the implementation of the single-visit analysis does not make up for the lack of nonparametric identifiability for the detection probabilities.

Single-Visit Abundance Scenario

The single-visit abundance scenario follows analogously to the single-visit occurrence scenario. Parametric identifiability for the site-specific abundance and detection probabilities comes from particular choices of link functions, but these properties lack the stronger nonparametric identifiability. In the single-visit abundance scenario there is an underlying data-generating process that produces an abundance N_i at each site i. A random sample of S sites are visited and how many individuals seen are recorded. The N_i are assumed to be independent Poisson random variables with parameters λ_i. Given the abundance N_i at a site, and the site-specific detection probability p_i, the number y_i of individuals observed is assumed to follow a Binomial distribution with N_i trials and probability p_i. Then the y_i are marginally independent Poisson distributions with parameter $\lambda_i p_i$. The abundances are typically of interest, and the average parameter $\frac{1}{S}\sum_{i=1}^{S}\lambda_i$ of the abundance distribution may be of particular interest.

An approach using covariates x to model abundance and z to model detection probabilities is proposed by Solymos et al. that estimates the parameters λ_i of the site-specific abundance distribution separately from the site-specific detection probabilities p_i [171].

Their approach is able to parametrically identify the site-specific abundance and detection probabilities, but Knape and Korner-Nievergelt proposed a counter-example model that reveals a lack of nonparametric identifiability for these properties [88]. This counter-example model is similar to that of the single-visit occurrence example: $\lambda_i = \alpha\frac{e^{\beta_0+\beta' x_i}}{1+e^{\beta_0+\beta' x_i}}$; $p_i = \frac{1}{\alpha}\frac{e^{\theta_0+\theta' z_i}}{1+e^{\theta_0+\theta' z_i}}$. The same observable distribution of the y_i can arise from different components λ_i and p_i.

A lack of nonparametric identifiability can be shown in even these idealized conditions, but it should be noted that the assumptions of the Poisson distribution, the independence between sites, and the Binomial distribution may also be tenuous. Barker et al. and Knape et al. provide a discussion of identifiability and robustness with respect to the Binomial and Poisson assumptions in the scenario where abundance data is recorded from multiple visits [7, 90].

Again the scaling counter-example was a convenient way to show that the average abundance lacks nonparametric identifiability, but the breakdown under model mis-specification can be shown using yet another data-generating process that differs from the assumptions of the single-visit model. Figs A.3 and A.4 present the same identifiability scenarios as in the main manuscript for estimating abundance and detection respectively. The top rows use counts with imperfect detection from a single-visit, while the bottom rows use counts with imperfect detection from two visits. The first columns show estimation via Poisson and logistic regression respectively when the true data-generating processes come from quadratic functions of covariates.

In the single-visit case, the "best approximation" within the parametric family of the Poisson underestimates the abundance and overestimates detection probabilities. However, with two visits, the abundance and detection probabilities are locally approximated within the Poisson and logistic families respectively. In the second column, the added flexibility of the spine terms allow both the single-visit and double-visit cases to estimate abundances and detection probabilities closer to the truth, albeit with fairly high variability across simulations. With extra data the nonparametrically identified double-visit case has decreased variability across simulations and is starting to converge to the truth (although even more data seems to be needed) while the estimates across simulations in the single-visit case still cover too wide a range of potential abundances and detection probabilities to be useful in practice.

In the single-visit abundance example, an upper bound for abundance cannot be found since detection probabilities could be arbitrarily small, but the number of detected individuals (assuming no double-counts) can provide a lower bound.

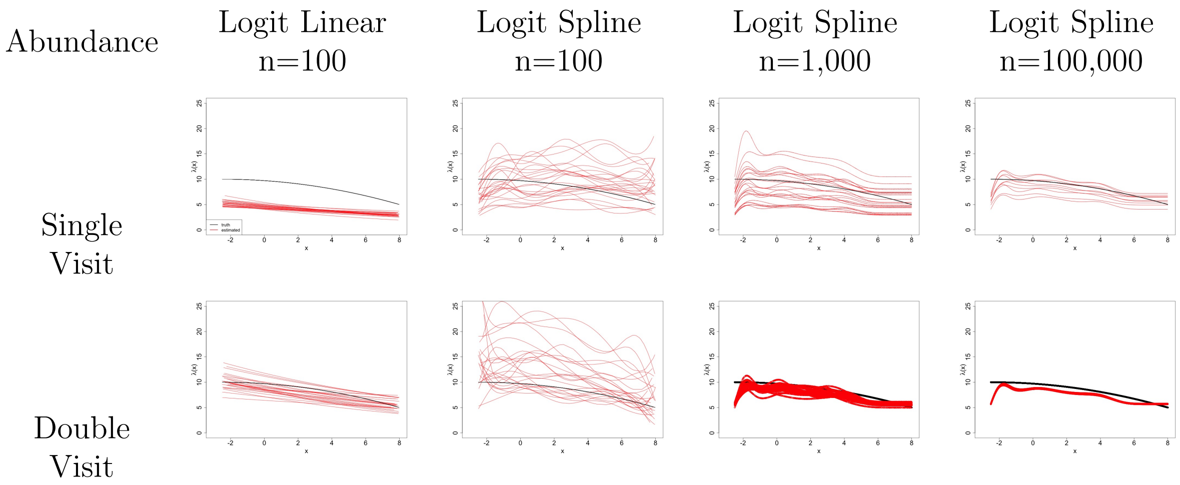


Figure A.3: The x-axis displays the value of the covariate that helps us predict abundance. The y-axis displays the abundance. The first row shows that single-visit data only parametrically identifies the average abundance. With model mis-specification, the abundances are underestimated. Added flexibility increases the variability in the estimation of the abundance. In the nonparametrically identified double-visit case depicted in the bottom row, more data improves the estimates.

Capture-Recapture Abundance Scenario

Capture-recapture is a data-collection strategy that requires repeated visits to the same locations over time and the ability to uniquely identify individuals. This way a repeated sighting of a particular individual is recorded. Closure is assumed and replication given by revisiting locations is used to allow for imperfect detection probabilities. The controversy in this scenario is how to estimate the abundance when nothing is known about the distribution of the individual detection probabilities.

To estimate an unknown total abundance N across a region of interest, S sites are visited T times. When an individual is spotted it is marked such that it can be distinguished from others if it is seen again during another visit. The X_i are the number of times individual i is observed, and n is the number of individuals that are seen at least once. The X_i given individual detection probabilities p_i are independent Binomial random variables with T trials and probability p_i of success. The detection probabilities p_i are identically distributed from some unknown distribution $g(p)$. The sighting frequencies f_x are the number of X_i where $X_i = x$. However, sighting frequencies are observed only given that the individual is spotted at all, f_x^c. Similarly the only probability observed is that of seeing an individual x times given that it is seen at least once.

With the language of nonparametric and parametric identifiability, previous results in

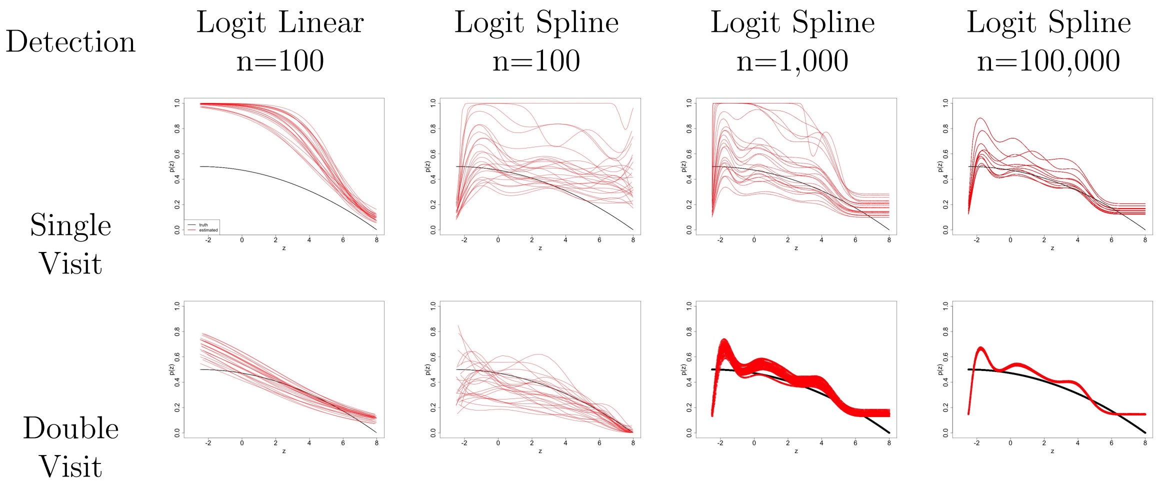


Figure A.4: The x-axis displays the value of the covariate that helps us predict detection. The y-axis displays the probability of detection. The first row shows that single-visit data only parametrically identifies the average detection. With model mis-specification, the detection probabilities are overestimated. Even with more data, the single-visit case yields estimates that cover the majority of the range of plausible values, not providing much insight into the true average detection probability. In contrast, the bottom row shows that detection probabilities are nonparametrically identified using double-visit data, and estimation is robust to model mis-specification (although even more data would be needed to get even closer to the truth).

the literature are more easily interpretable. Huggins showed that there are different unconditional distributions of the total abundance (unobserved) that have the same distributions when conditioned on the captured individuals (observed) [75]. Link refined the conclusions of Huggins and showed that if two distributions of detection probabilities conditioned on the captured individuals are close (in function space), their unconditional distributions of the total abundance are not necessarily close [99, 75]. This result shows a lack of nonparametric identifiability; the abundance cannot be identified without restrictions on the detection probability distribution $g(p)$.

Holzmann et al. showed that if $g(p)$ is assumed to belong to certain probability distribution families, abundance is identifiable [73]. Abundance is identifiable if it is assumed that the detection probabilities follow a Uniform distribution (with more than one visit per site), follow a Beta distribution (with more than two visits per site) or follow a finite mixture model (with at least twice as many visits as mixture components). These are parametric identifiability results. Link responded that even if it is assumed that there are no individuals who are undetectable, there is still no identifiability across different families of assumptions for $g(p)$ e.g. a Beta distribution can be found that gives an identical observable distribution to a two-point mixture but implies a different overall abundance [100]. This illustrates the gap between a parametrically and nonparametrically identifiable property.

Mao determined a lower bound on the odds of an individual animal not being captured and used it to lower bound the abundance [110, 111]. However, they also showed that an upper bound on abundance cannot be found without placing restrictions on $g(p)$. These results partially identify abundance.

Lack of Nonparametric Identifiability for Capture-Recapture Data (with heterogeneous detection)

Link gave examples that have the same observable distribution but imply different values of total abundance, showing lack of nonparametric identifiability when working with the distribution of the abundance conditional on the sighted individuals [99].

However, Farcomeni and Tardella showed that abundance is technically identifiable when analyzing the unconditional likelihood rather than the conditional likelihood [50]. The conditional likelihood was focused on by Huggins, Link, and Holzman et al. because Sanathanan stated that the conditional likelihood of the abundance (conditional on the captured individuals) is asymptotically equivalent to its unconditional likelihood [75, 99, 73, 165]. However, Farcomeni and Tardella pointed out that the conditions for this statement are not met in the capture-recapture scenario [50]. Despite technical identifiability, Farcomeni and Tardella showed that there is no consistent estimator for the total abundance. Here, some intuition about why the technical identifiability is so tenuous is provided [50].

Recall that if N is considered fixed, $n \sim \text{Binom}(N, 1 - \pi_g(0))$. Technically N and $\pi_g(0)$ are identifiable because it is not possible to find a $N \neq N^*$ and $\pi_g(0) \neq \pi_g^*(0)$ such that for all x:

$$\binom{N}{x} \pi_g(0)^x (1 - \pi_g(0))^{N-x} = \binom{N^*}{x} \pi_g^*(0)^x (1 - \pi_g^*(0))^{N^*-x}$$

However, with a single realization of the data generating process, these parameters are not practically identifiable. Importantly, more realizations of the data generating function would be needed, not a larger sample size.

If the model is broadened to allow N to be random this tenuous identifiability goes away. Suppose $N \sim \text{Poisson}(\lambda)$. Note that N could be chosen to follow a nonparametric distribution, but since N and $\pi_g(0)$ can be proved to be not nonparametrically identifiable in this case, they certainly are not nonparametrically identifiable in a more general case. Then the number of individuals seen at least once follows a Poisson distribution with parameter $\lambda(1 - \pi_g(0))$.

Multiple data generating processes can have the same observable distribution by scaling up λ and scaling down $(1-\pi_g(0))$ by equal amounts (or vice versa). Therefore, the abundance is not nonparametrically identifiable.

Johndrow et al. proposed estimating a different quantity, the abundance of individuals who have detection probabilities above a particular threshold [83]. They provide a risk analysis of their estimator and some guidance on how to choose the required threshold.

A.2 JSDM Supporting Material

	number factors	number species	H	S	E	Jaccard
1	1	10	1.56	5.01	0.97	0.66
2	2	10	1.56	5.00	0.97	0.65
3	3	10	1.53	5.00	0.95	0.67
4	5	10	1.57	5.00	0.98	0.65
5	10	10	1.56	5.00	0.97	0.66
6	1	25	2.51	12.49	0.99	0.65
7	2	25	2.49	12.51	0.98	0.66
8	3	25	2.51	12.50	0.99	0.66
9	5	25	2.51	12.49	0.99	0.66
10	10	25	2.51	12.49	0.99	0.66
11	1	45	3.10	22.50	1.00	0.66
12	2	45	3.10	22.52	1.00	0.66
13	3	45	3.10	22.51	1.00	0.65
14	5	45	3.10	22.51	1.00	0.66
15	10	45	3.10	22.49	1.00	0.66

Table A.1: Correct-K True Values

	number factors	number species	H	S	E	Jaccard
1	1	25	1.78	6.16	0.98	0.63
2	2	25	1.85	6.65	0.97	0.66
3	3	25	1.93	7.21	0.98	0.69

Table A.2: Correct-K with Realistic Species Prevalences True Values

	block size	number species	H	S	E	Jaccard
1	3	30	2.68	15.02	0.99	0.67
2	5	30	2.67	14.99	0.98	0.67
3	10	30	2.63	15.00	0.97	0.68

Table A.3: Block-Correlation True Values

	block size	number species	H	S	E	Jaccard
1	3	30	1.62	5.55	0.95	0.76
2	5	30	1.60	5.56	0.93	0.76
3	10	30	1.54	5.54	0.90	0.76

Table A.4: Block-Correlation with Realistic Prevalences True Values

	number factors	number species	degrees of freedom	H	S	E	Jaccard
1	1	10	3	1.57	5.00	0.98	0.64
2	2	10	3	1.57	5.00	0.98	0.66
3	3	10	3	1.58	5.01	0.98	0.66
4	5	10	3	1.57	4.98	0.98	0.66
5	10	10	3	1.57	5.00	0.98	0.66
6	1	25	3	2.51	12.47	0.99	0.65
7	2	25	3	2.51	12.50	0.99	0.66
8	3	25	3	2.51	12.50	0.99	0.66
9	5	25	3	2.51	12.50	0.99	0.66
10	10	25	3	2.51	12.49	0.99	0.66
11	1	45	3	3.11	22.48	1.00	0.65
12	2	45	3	3.11	22.51	1.00	0.66
13	3	45	3	3.11	22.50	1.00	0.66
14	5	45	3	3.11	22.50	1.00	0.66
15	10	45	3	3.11	22.49	1.00	0.66
16	1	10	5	1.57	5.01	0.98	0.65
17	2	10	5	1.57	5.00	0.98	0.65
18	3	10	5	1.57	5.01	0.98	0.66
19	5	10	5	1.57	5.00	0.98	0.66
20	10	10	5	1.57	5.00	0.98	0.66
21	1	25	5	2.51	12.51	0.99	0.64
22	2	25	5	2.51	12.49	0.99	0.66
23	3	25	5	2.51	12.47	0.99	0.66
24	5	25	5	2.51	12.51	0.99	0.66
25	10	25	5	2.51	12.48	0.99	0.66
26	1	45	5	3.11	22.50	1.00	0.65
27	2	45	5	3.11	22.48	1.00	0.66
28	3	45	5	3.11	22.50	1.00	0.66
29	5	45	5	3.11	22.48	1.00	0.66
30	10	45	5	3.11	22.52	1.00	0.66
31	1	10	20	1.57	4.99	0.98	0.65
32	2	10	20	1.57	5.00	0.98	0.66
33	3	10	20	1.57	5.00	0.98	0.65
34	5	10	20	1.57	5.00	0.98	0.66
35	10	10	20	1.57	4.98	0.98	0.66
36	1	25	20	2.51	12.50	0.99	0.65
37	2	25	20	2.51	12.51	0.99	0.66
38	3	25	20	2.51	12.50	0.99	0.66
39	5	25	20	2.51	12.49	0.99	0.66
40	10	25	20	2.51	12.52	0.99	0.66
41	1	45	20	3.11	22.50	1.00	0.65
42	2	45	20	3.11	22.49	1.00	0.66
43	3	45	20	3.11	22.50	1.00	0.66
44	5	45	20	3.11	22.51	1.00	0.66
45	10	45	20	3.11	22.48	1.00	0.66

Table A.5: Heavy Tail True Values

	number factors	number species	degrees of freedom	H	S	E	Jaccard
1	1	25	3	1.84	6.48	0.98	0.64
2	2	25	3	1.86	6.65	0.98	0.64
3	3	25	3	1.82	6.39	0.98	0.65
4	1	25	5	1.85	6.60	0.98	0.63
5	2	25	5	1.78	6.14	0.98	0.67
6	3	25	5	1.80	6.24	0.98	0.65
7	1	25	20	1.85	6.59	0.98	0.64
8	2	25	20	1.86	6.61	0.98	0.65
9	3	25	20	1.83	6.40	0.98	0.64

Table A.6: Heavy Tail with Realistic Prevalences True Values

	H	S	E	Jaccard
1	0.41	1.47	1.08	0.69

Table A.7: Toy Interaction True Values

Bibliography

1. Alley, M. *The Craft of Scientific Writing* 3rd. Chap. 17 (Springer-Verlag New York, 1996).
2. American Library Association. *Digital Literacy* `https://literacy.ala.org/digital-literacy/`. 2020.
3. Anderson, M. J., de Valpine, P., Punnett, A. & Miller, A. E. A pathway for multivariate analysis of ecological communities using copulas. *Ecology and Evolution* **9,** 3276–3294 (2019).
4. Anderson, M. J. *et al.* Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecology Letters* **14,** 19–28 (2010).
5. Aplin, P. Remote sensing: ecology. *Progress in Physical Geography* **29,** 104–113 (2005).
6. arXiv. *About arXiv* `https://arxiv.org/`.
7. Barker, R. J., Schofield, M. R., Link, W. A. & Sauer, J. R. On the Reliability of N-Mixture Models for Count Data. *Biometrics* **74,** 369–377 (2018).
8. Baselga, A. & Araujo, M. B. Do community-level models describe community variation effectively? *Journal of Biogeography* **37,** 1842–1850 (2010).
9. Beissinger, S. R. *et al.* Incorporating Imperfect Detection into Joint Models of Communities: A response to Warton et al. *Trends in Ecology and Evolution* **31,** 736–737 (2016).
10. Ben-Zvi, D. Using Wiki to Promote Collaborative Learning in Statistics Education. *Technology Innovations in Statistics Education* **1,** 1–18 (2007).
11. Bhattacharya, A. & Dunson, D. B. Sparse Bayesian infinite factor models. *Biometrika* **98,** 291–306 (2011).
12. Billick, I. & Case, T. J. Higher Order Interactions in Ecological Communities: What Are They and How Can They be Detected? *Ecology* **75,** 1529–1543 (1994).
13. Blanchet, F. G., Cazelles, K. & Gravel, D. Co-occurrence is not evidence of ecological interactions. *Ecology Letters* **23,** 1050–1063 (2020).
14. Blum, D., Knudson, M. & Henig, R. M. *A Field Guide for Science Writers: The Offical Gude of the National Association of Science Writers* 2nd ed. (Oxford University Press, 2005).

15. Blumstein, D. T. *et al.* Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology* **48,** 758–767 (2011).

16. Bolker, B. Learning hierarchical models: advice for the rest of us. *Ecological Applications* **19,** 588–592 (2009).

17. Bollen, K. A. Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology* **53,** 605–634 (2002).

18. Borgman, C. L., Wallis, J. C. & Enyedy, N. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries* **7,** 17–30 (2007).

19. Box, G. E. P. in. Chap. Robustness in the strategy of scientific model building (Academic Press, 1979).

20. Boyce, M. S., Vernier, P. R., Nielsen, S. E. & Schmiegelow, F. K. A. Evaluating resource selection functions. *Ecological Modelling* **157,** 281–300 (2002).

21. Bubela, T. *et al.* Science communication reconsidered. *Nature Biotechnology* **27,** 514–518 (2009).

22. Burns, T. W., O'Connor, D. J. & Stocklmayer, S. M. Science Communication: A Contemporary Definition. *Public Understanding of Science* **12,** 183–202 (2003).

23. Bush, D. M. Emergency Department Visits Attributed to Overmedication That Involved the Insomnia Medication Zolpidem. *The Center for Behavioral Health Statistics and Quality Report* (2014).

24. California Secretary of State. *Voter Registration Statistics* `https://www.sos.ca.gov/elections/voter-registration/voter-registration-statistics/`. 2020.

25. Carver, R. *et al. Guidelines for Assessment and Instruction in Statistics Education (GAISE)* tech. rep. (American Statistical Association, 2016).

26. Chance, B. *et al. Curriculum Guidelines for Undergraduate Programs in Statistical Science* tech. rep. (American Statistical Association Undergraduate Guidelines Workgroup, 2014).

27. Cho, K. & Schunn, C. D. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education* **48,** 409–426 (2007).

28. Clark, J. S. Why environmental scientists are becoming Bayesians. *Ecology Letters* **8,** 2–14 (2004).

29. Clark, J. S., Gelfand, A. E., Woodall, C. W. & Zhu, K. More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications* **24,** 990–999 (2014).

30. Cornell Lab of Ornithology, Ithaca, New York. *eBird: An online database of bird distribution and abundance* `http://www.ebird.org`. 2017.

31. Credit Union Cherry Blossom. *Credit Union Cherry Blossom Ten Mile Run & 5K Run-Walk* `http://cherryblossom.org/`. 2019.

32. Cressie, N., Calder, C. A., Clark, J. S., Ver Hoef, J. M. & Wikle, C. K. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications* **19,** 553–570 (2009).

33. D'az, A. *Inside Martyr Complex for First-Gen* `https://angelddaz.substack.com/p/inside-martyr-complex-for-first-gen`. Dec. 2019.

34. Daks, A., Desai, N. & Goldberg, L. R. Do the Golden State Warriors have hot hands? *The Mathematical Intelligencer* **40,** 1–5 (2018).

35. Daly, J. *My Experience at the 2017 rOpenSci Unconference* `https://jasminedaly.com//2017-05-28-runconf17-experience/`. 2017.

36. DeYoung, R. W. & Honeycutt, R. L. The Molecular Toolbox: Genetic Techniques in Wildlife Ecology and Management. *Wildlife Management* **69,** 1362–1384 (2005).

37. Dickinson, J. L., Zuckerberg, B. & Bonter, D. N. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics* **41,** 149–172 (2010).

38. Diserud, O. H. & Odegaard, F. A multiple-site similarity measure. *Biology Letters* **3,** 20–22 (2006).

39. Dorazio, R. M. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography* **23,** 1472–1484 (2014).

40. Dorazio, R. M., Connor, E. F. & Askins, R. A. Estimating the Effects of Habitat and Biological Interactions in an Avian Community. *PLOS One* **10,** e0135987 (2015).

41. Eberly, L. E. & Carlin, B. P. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine* **19,** 2279–2294 (2000).

42. *Ecology for the Masses Blog* `https://ecologyforthemasses.com/`. 2020.

43. Elith, J. *et al.* Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29,** 129–151 (2006).

44. Eshet, Y. Digital Literacy: A Conceptual Framework for Survival Skills in the Digital era. *Journal of Educational Multimedia and Hypermedia* **13,** 93–106 (2004).

45. Fabrigar, L. R. & Wegener, D. T. in. Chap. Factor Analysis Assumptions (Oxford Scholarship, 2011).

46. Fabrigar, L. R. & Wegener, D. T. in. Chap. Requirements for and Decisions in Choosing Exploratory Common Factor Analysis (Oxford Scholarship, 2011).

47. Fabrigar, L. R. & Wegener, D. T. in. Chap. Factor Analysis Assumptions (Oxford Scholarship, 2015).

48. Faisal, A., Dondelinger, F., Husmeier, D. & Beale, C. M. Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. *Ecological Informatics* **5,** 451–464 (2010).

49. Fan, Y. *et al.* Applications of structural equation modeling (SEM) in ecological studies: an updated review. *Ecological Processes* **5,** 19 (2016).

50. Farcomeni, A. & Tardella, L. Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electronic Journal of Statistics* **6,** 2602–2626 (2012).

51. Farley, S. S., Dawson, A., Goring, S. J. & Williams, J. W. Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience* **68,** 563–576 (2018).

52. Fiske, I. & Chandler, R. unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. *Journal of Statistical Software* **43,** 1–23 (2011).

53. Fithian, W., Elith, J., Hastie, T. & Keith, D. A. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution* **6,** 424–438 (2015).

54. Fletcher Jr., R. J. *et al.* A practical guide for combining data to model species distributions. *Ecology* **100,** e02710 (2019).

55. *Forest Inventory and Analysis Database* `https://www.fia.fs.fed.us/`. 2020.

56. Friel, S. N., Curcio, F. R. & Bright, G. W. Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education* **32,** 124–158 (2001).

57. GBIF. *What is GBIF?* `https://www.gbif.org/what-is-gbif`. 2020.

58. Gelman, A. *Advice on writing research articles* `https://statmodeling.stat.columbia.edu/2009/07/30/advice_on_writi/`. 2018.

59. Giles, H. A. C. *et al.* The longest period transiting planet candidate from K2. *Astronomy & Astrophysics* **615,** 1–5 (2018).

60. Golding, N., Nunn, M. A. & Purse, B. V. Identifying biotic interactions which drive the spatial distribution of a mosquito community. *Parsites and Vectors* **8,** 367 (2015).

61. Graff, G. & Birkenstein, C. *They Say / I Say: The Moves That Matter in Academic Writing* (W. W. Norton & Company, 2009).

62. Green, J. L. *et al.* Complexity in Ecology and Conservation: Mathematical, Statistical, and Computational Challenges. *BioScience* **55,** 501–510 (2005).

63. Grinnell, J. & Storer, T. I. *Animal life in the Yosemite: an account of the mammals, birds, reptiles, and amphibians in a cross-section of the Sierra Nevada* (University of California Press, 1924).

64. Guilford, W. H. Teaching Peer Review and the Process of Scientific Writing. *Advances in Physiology Education* **25,** 167–175 (2001).

65. Guillera-Arroita, G., Ridout, M. S. & Morgan, B. J. T. Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution* **1.** `https://doi.org/10.1111/j.2041-210X.2010.00017.x`, 131–139 (2010).

66. Guillera-Arroita, G. *et al.* Is my species distribuiton model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography* **24.** `https://doi.org/10.1111/geb.12268`, 276–292 (2015).

67. Hampton, S. E. *et al.* Big data and the future of ecology. *Frontiers in Ecology and the Environment* **11,** 156–162 (2013).

68. Harris, D. J. Inferring species interactions from co-occurrence data with Markov networks. *Ecology* **97,** 3308–3314 (2016).

69. Hastie, T. & Fithian, W. Inference from presence-only data; the ongoing controversy. *Ecography* **36,** 864–867 (2013).

70. Heard, S. B. *The Scientist's Guide to Writing: How to write more easily and effectively throughout your scientific career* (Princeton University Press, 2016).

71. Hill, M. O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54,** 427–432 (1973).

72. Hirzel, A. H., Lay, G. L., Helfer, V., Randin, C. & Guisan, A. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling* **199,** 142–152 (2006).

73. Holzmann, H., Munk, A. & Zucchini, W. On Identifiability in Capture-Recapture Models. *Biometrics* **62,** 934–939 (2006).

74. Howden, L. M. and Meyer, J. A. Age and Sex Composition: 2010. *2010 Census Briefs* (2011).

75. Huggins, R. A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneous capture probabilities. *Statistics and Probability Letters* **54,** 147–152 (2001).

76. Hui, F. K. C. BORAL - Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution* **7,** 744–750 (2016).

77. Hui, F. K. C. *boral: Bayesian Ordination and Regression AnaLysis* R package version 1.8. 2020.

78. Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D. & Warton, D. I. Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution* **6,** 399–411 (2015).

79. Hunter, A., Laursen, S. L. & Seymour, E. Becoming a Scientist: The Role of Undergraduate Research in Students' Cognitive, Personal,and Professional Development. *Science Education* **91,** 36–74 (2006).

80. Hurlbert, S. H. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology* **52,** 577–586 (1971).

81. *iNaturalist* `https://www.inaturalist.org`. 2020.

82. Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P. & Roy, D. B. Statistics for citizen science: extracting signals ofchange from noisy ecological data. *Methods in Ecology and Evolution* **5,** 1052–1060 (2014).

83. Johndrow, J. E., Lum, K. & Manrique-Vallier, D. Low-risk population size estimates in the presence of capture heterogeneity. *Biometrika* **106,** 197–210 (2019).

84. Johnson, M., Aguilar de Soto, N. & Madsen, P. T. Studying the behaviour and sensory ecology of marine mammals using acoustic recording tags: a review. *Marine Ecology Progress Series* **395,** 55–73 (2009).

85. Kaplan, J. J., Fisher, D. G. & Rogness, N. T. Lexical Ambiguity in Statistics: What do Students Know about the Words Association, Average, Confidence, Random and Spread? *Journal of Statistics Education* **17,** 1–20 (2009).

86. Kery, M. *et al.* Site-Occupancy Distribution Modeling to Correct Population-Trend Estimates Derived from Opportunistic Observations. *Conservation Biology* **24,** 1388–1397 (2010).

87. Kissling, W. D. *et al.* Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography* **39,** 2163–2178 (2012).

88. Knape, J. & Korner-Nievergelt, F. Estimates from non-replicated population surveys rely on critical assumptions. *Methods in Ecology and Evolution* **6,** 298–306 (2015).

89. Knape, J. & Korner-Nievergelt, F. On assumptions behind estimates of abundance from counts at multiple sites. *Methods in Ecology and Evolution* **7,** 206–209 (2016).

90. Knape, J. *et al.* Sensitivity of binomial N-mixture models to overdispersion: The importance of assessing model fit. *Methods in Ecology and Evolution* **9,** 2102–2114 (2018).

91. Koopmans, T. C. & Reiersol, O. The Identification of Structural Characteristics. *The Annals of Mathematical Statistics* **21,** 165–181 (1950).

92. Kucera, T. E. & Barrett, R. H. in (eds O'Connell, A. F. & Nichols, J. D.) chap. A History of Camera Trapping (Springer, 2011).

93. Latimer, A. M., Banerjee, S., Sang, H., Mosher, E. S. & Silander Jr., J. A. Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecology Letters* **12,** 144–154 (2009).

94. Lele, S. R. Model complexity and information in the data: Could it be a house built on sand? *Ecology* **91,** 3493–3496 (2010).

95. Lele, S. R. & Dennis, B. Bayesian methods for hierarchical models: Are ecologists making a Faustian bargain. *Ecological Applications* **19,** 581–584 (2009).

96. Lele, S. R., Merrill, E. H., Keim, J. L. & Boyce, M. S. Selection, use, choice and occupancy: clarifying concepts in resource selection studies. *Journal of Animal Ecology* **82,** 1183–1191 (2013).

97. Lele, S. R., Moreno, M. & Bayne, E. Dealing with detection error in site occupancy surveys: what can we do with a single survey? *Journal of Plant Ecology* **5,** 22–31 (2012).

98. Lele, S. R., Nadeem, K. & Schmuland, B. Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning. *Journal of the American Statistical Association* **105,** 1617–1625 (2012).

99. Link, W. A. Nonidentifiability of Population Size from Capture-Recapture Data with Heterogeneous Detection Probabilities. *Biometrics* **59,** 1123–1130 (2003).

100. Link, W. A. Rejoinder to "On Identifiability in Capture-Recapture Models". *Biometrics* **62,** 936–939 (2006).

101. Lister, A. R. & Climate Change Research Group. Natural history collections as sources of long-term datasets. *Trends in Ecology and Evolution* **26,** 153–154 (2011).

102. *Long Term Ecological Research* `https://lternet.edu/`. 2020.

103. Lopez, M., Whalley, J., Robbins, P. & Lister, R. Relationships between reading, tracing and writing skills in introductory programming. *The Fourth International Computing Education Research Workshop* (2008).

104. MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G. & Franklin, A. B. Estimating Site Occupancy, Colonization, and Local Extinction When a Species Is Detected Imperfectly. *Ecology* **84,** 2200–2207 (2003).

105. MacKenzie, D. I. & Royle, J. A. Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology* **42,** 1105–1114 (2005).

106. MacKenzie, D. I. *et al.* Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83,** 2248–2255 (2002).

107. Magurran, A. E. *Ecological Diversity and Its Measurement* (Springer, 1988).

108. Malaeb, Z. A., Summers, J. K. & Pugesek, B. H. Using structural equation modeling to investigate relationships among ecological variables. *Environmental and Ecological Statistics* **7,** 93–111 (2000).

109. Manski, C. F. *Partial Identification of Probability Distributions* (Springer, 2003).

110. Mao, C. X. Estimating population sizes for capture-recapture sampling with binomial mixtures. *Computational Statistics and Data Analysis* **51,** 5211–5219 (2007).

111. Mao, C. X. On the Nonidentifiability of Population Sizes. *Biometrics* **64,** 977–981 (2008).

112. Matarese, V. in. Chap. Using strategic, critical reading of research papers to teach scientific writing: the reading–research–writing continuum (Chandos Publishing, 2013).

113. Maynard, D. S. *et al.* Species associations overwhelm abiotic conditions to dictate the structure and function of wood-decay fungal communities. *Ecology* **99,** 801–811 (2018).

114. McCulloch, C. E. & Neuhaus, J. M. Prediction of Random Effects in Linear and Generalized Linear Models under Model Misspecification. *Biometrics* **67,** 270–279s (2011).

115. Michener, W. K. & Jones, M. B. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology and Evolution* **27,** 85–93 (2012).

116. Miller, T. E. Direct and Indirect Species Interactions in an Early Old-Field Plant Community. *The American Naturalist* **143,** 1007–1025 (1994).

117. Milner, K. & Rougier, J. How to weigh a donkey in the Kenyan countryside. *Significance* **11,** 40–43 (2014).

118. Mohanty, M. & Slattum, P. Alcohol, Medications, and Older Adults. *Age in Action* **26,** 1–5 (2011).

119. Morris, E. K. *et al.* Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution* **4,** 3514–3524 (2014).

120. *National Ecological Observatory Network* `https://www.neonscience.org/`. 2020.

121. *Neotoma Paleoecology Database* `https://www.neotomadb.org/`. 2020.

122. New York Times. *Polling Standards* `http://www.nytimes.com/packages/pdf/politics/pollingstandards.pdf`. 2006.

123. Newman, G. *et al.* The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment* **10,** 298–304 (2012).

124. Nieto-Lugilde, D., Maguire, K. C., Blois, J. L., Williams, J. W. & Fitzpatrick, M. C. Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Methods in Ecology and Evolution* **9,** 834–848 (2017).

125. Nolan, D. & Perrett, J. Teaching and Learning Data Visualization: Ideas and Assignments. *The American Statistician* **70,** 260–269 (2016).

126. Norberg, A. *et al.* A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs* **89,** e01370 (2019).

127. Office on Smoking and Health. *2018 National Youth Tobacco Survey: Methodology Report* tech. rep. (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2018).

128. Ottaviani, D., Lasinio, G. J. & Boitani, L. Two statistical methods to validate habitat suitability models using presence-only data. *Ecological Modelling* **179,** 417–443 (2004).

129. Ovaskainen, O., Hottola, J. & Siitonen, J. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology* **91,** 2514–2521 (2010).

130. Ovaskainen, O., Roy, D. B., Fox, R. & Anderson, B. J. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution* **7,** 428–436 (2015).

131. Ovaskainen, O. *et al.* How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters* **20,** 561–576 (2017).

132. Pacifici, K. *et al.* Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology* **98,** 840–850 (2017).

133. *Paleobiology Database* `https://paleobiodb.org/`. 2020.

134. Parke, C. S. Reasoning and Communicating in the Language of Statistics. *Journal of Statistics Education* **16,** 1–25 (2008).

135. Pearce, J. L. & Boyce, M. S. Modelling distribution and abundance wiht presence-only data. *Journal of Applied Ecology* **43,** 405–412 (2006).

136. Peel, S. L. *et al.* Reliable species distributions are obtainable with sparse, patchy and biased data by leveraging over species and data types. *Methods in Ecology and Evolution* **10,** 1002–1014 (2019).

137. Peng, C., Guiot, J., Wu, H., Jiang, H. & Luo, Y. Integrating models with data in ecology and palaeoecology: advances towards a model-data fusion approach. *Ecology Letters* **14,** 522–536 (2011).

138. Petrovskii, S. & Petrovskaya, N. Computational ecology as an emerging science. *Interface Focus* **2,** 241–254 (2012).

139. Pfannkuch, M., Regan, M., Wild, C. & Horton, N. J. Telling Data Stories: Essential Dialogues for Comparative Reasoning. *Journal of Statistics Education* **18,** 1–39 (2010).

140. Phillips, S. J. & Elith, J. On estimating probability of presence from use-availability or presence-background data. *Ecology* **94,** 1409–1419 (2013).

141. Phillips, S. J. & Elith, J. POC plots: calibrating species distribution models with presence-only data. *Ecology* **91,** 2476–2484 (2010).

142. Phu, B. *Mistakes that Kids Make When They Do Math* `https://stat198-spring18.github.io/blog/2018/04/17/math-mistakes`. 2018.

143. Pichler, M., Boreaux, V., Klein, A., Schleuning, M. & Hartig. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution* **11,** 281–293 (2019).

144. Poirier, D. J. Revising Beliefs in Nonidentified Models. *Econometric Theory* **14,** 483–509 (1998).

145. Pollock, K. H. A Capture-Recapture Design Robust to Unequal Probability of Capture. *The Journal of Wildlife Management* **46.** 10.2307/3808568, 752–757 (1982).

146. Pollock, L. J. *et al.* Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution* **5,** 397–406 (2014).

147. Pomeranz, J. P. F., Thompson, R. M., Poisot, T. & Harding, J. S. Inferring predatory-prey interactions in food webs. *Methods in Ecology and Evolution* **10,** 356–367 (2018).

148. Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K. C. & Moles, A. T. Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution* **10,** 1571–1583 (2019).

149. Possolo, A., Schlamminger, S., Stoudt, S., Pratt, J. R. & Williams, C. J. Evaluation of the accuracy, consistency, and stability of measurements of the Planck constant used in the redefinition of the international system of units. *Metrologia* **55,** 29–37 (2018).

150. Purvis, A. & Hector, A. Getting the measure of biodiversity. *Nature* **405,** 212–219 (2000).

151. Raimes, A. Teaching Writing. *Annual Review of Applied Linguistics* **18,** 142–167 (1998).

152. Reichman, O. J., Jones, M. B. & Schildhauer, M. P. Challenges and Opportunities of Open Data in Ecology. *Science* **331,** 703–705 (2011).

153. Renner, I. W., Louvrier, J. & Gimenez, O. Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization. *Methods in Ecology and Evolution* **10,** 2118–2128 (2019).

154. Ricotta, C. & Pavoine, S. A multiple-site dissimilarity measure for species presence/absence data and its relationship with nestedness and turnover. *Ecological Indicators* **54,** 203–206 (2015).

155. Rocha, R. *et al.* Secondary forest regeneration benefits old-growth specialist bats in a fragmented tropical landscape. *Scientific Reports* **8,** 1–9 (2018).

156. Roehrig, C. S. Conditions for Identification in Nonparametric and Parametric Models. *Econometrica* **56,** 433–447 (1988).

157. Rota, C. T., Fletcher Jr., R. J., Dorazio, R. M. & Betts, M. G. Occupancy estimation and the closure assumption. *Journal of Applied Ecology* **46,** 1173–1181 (2009).

158. Rota, C. T. *et al.* A multispecies occupancy model for two or more interacting species. *Methods in Ecology and Evolution* **7,** 1164–1173 (2016).

159. Rota, C. T. *et al.* A two-species occupancy model accommodating simultaneous spatial and interspecific dependence. *Ecology* **97,** 48–53 (2016).

160. Rothenberg, T. J. Identification in Parametric Models. *Econometrica* **39,** 577–591 (1971).

161. Royle, J. A., Chandler, R. B., Yackulic, C. & Nichols, J. D. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution* **3,** 545–554 (2012).

162. Salmon, M. *The music of Les Mills Body Pump, with Spotify data* `http://www.masalmon.eu/2017/10/01/bodypump`. 2017.

163. SAMHSA, Office of Applied Studies. *Emergency Department Visits Involving Accidental Ingestion of Drugs by Children Aged 5 or Younger* tech. rep. (National Criminal Justice Reference Service, 2010).

164. San Martin, E. & Gonzalez, J. Bayesian identifiability: Contributions to an inconclusive debate. *Chilean Journal of Statistics* **1,** 69–91 (2010).

165. Sanathanan, L. Estimating the size of a multinomial population. *The Annals of Mathematical Statistics* **43,** 142–152 (1972).

166. Scherrer, D., Mod, H. K. & Guisan, A. How to evaluate community predictions without thresholding. *Methods in Ecology and Evolution* **11,** 51–63 (2019).

167. Schimel, J. *Writing Science: How to write papers that get cited and proposals that get funded* (Oxford University Press, 2011).

168. Simon, Sutton, K., Lopez, M. & Clear, T. Surely We Must Learn to Read before We Learn to Write! *Proceedings of 11th Australasian Computing Education Conference* (2009).

169. Smith, A. *U.S. Smartphone Use in 2015* `https://www.pewresearch.org/internet/2015/04/01/us-smartphone-use-in-2015/`. 2015.

170. Solymos, P. & Lele, S. R. Revisiting resource selection probability functions and single-visit methods: clarification and extensions. *Methods in Ecology and Evolution* **7,** 196–205 (2016).

171. Solymos, P., Lele, S. R. & Bayne, E. Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. *Environmetrics* **23,** 197–205 (2012).

172. Solymos, P., Moreno, M. & Lele, S. R. *detect: Analyzing Wildlife Data with Detection Error* (2018).

173. Spurrier, J. D. A Capstone Course for Undergraduate Statistics Majors. *Journal of Statistics Education* **9,** 1–11 (2001).

174. StackOverflow. *How to create a Minimal, Reproducible Example* `https://stackoverflow.com/help/minimal-reproducible-example`. 2020.

175. Stoudt, S. *Tag Yourself* `https://logicmag.io/nature/tag-yourself/`. 2019.

176. Sweeny, C. *Religious upbringing linked to better health and well-being during early adulthood.* Harvard T.H. Chan School of Public Health. `https://www.hsph.harvard.edu/news/press-releases/religious-upbringing-adult-health/`. 2018.

177. Taylor-Rodriguez, D., Kaufeld, K., Schliep, E. M., Clark, J. S. & Gelfand, A. E. Joint Species Distribution Modeling: Dimension Reduction Using Dirichlet Processes. *Bayesian Analysis* **12,** 939–967 (2017).

178. Teacher, A. G. F., Griffiths, D. J., Hodgson, D. J. & Inger, R. Smartphones in ecology and evolution: a guide for theapp-rehensive. *Ecology and Evolution* **3,** 5268–5278 (2013).

179. Thurman, L. L., Barner, A. K., Garcia, T. S. & Chestnut, T. Testing the link between species interactions and species co-occurrence in a trophic network. *Ecography* **42,** 1658–1670 (2019).

180. Tobler, M. W. *et al.* Joint species distribution models with species correlations and imperfect detection. *Ecology* **100,** e02754 (2019).

181. Tomasovych, A. & Kidwell, S. M. Predicting the effects of increasing temporal scale on species composition, diversity, and rank-abundance distributions. *Paleobiology* **36,** 672–695 (2010).

182. Tsai, J. *et al.* Reasons for Electronic Cigarette Use Among Middle and High School Students - National Youth Tobacco Survey, United States, 2016. *Morbidity and Mortality Weekly Report* **67,** 196–200 (2018).

183. UCB Public Affairs. *Correctional officers at high risk for depression, PTSD, suicide, survey finds.* Media Relations, UC Berkeley. `http://news.berkeley.edu/2018/08/23/california-correctional-officers-at-high-risk-for-depression-ptsd-and-suicide-new-survey-finds/`. 2018.

184. United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality. *Drug Abuse Warning Network (DAWN)* `https://doi.org/10.3886/ICPSR34565.v3`. 2011.

185. University of Michigan Sweetland Center for Writing. *Using Peer Review to Improve Student Writing* `https://lsa.umich.edu/sweetland/instructors/teaching-resources/using-peer-review-to-improve-student-writing.html`. 2020.

186. Uriarte, M. & Yackulic, C. B. Preaching to the unconverted. *Ecological Applications* **19,** 592–596 (2009).

187. Walker, S. C. & Jackson, D. A. Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs* **81,** 635–663 (2011).

188. Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J. R. Presence-Only Data and the EM Algorithm. *Biometrics* **65,** 554–563 (2009).

189. Warschauer, M. Invited Commentary: New Tools for Teaching Writing. *Language Learning & Technology* **14,** 3–8 (2010).

190. Warton, D. I., Renner, I. W. & Ramp, D. Model-Based Control of Observer Bias for the Analysis of Presence-Only Data in Ecology. *PLoS ONE* **8,** e79168 (2013).

191. Warton, D. I. *et al.* Extending joint models in community ecology: a response to Beissinger et al. *Trends in Ecology and Evolution* **31,** 737–738 (2016).

192. Warton, D. I. *et al.* So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution* **30,** 766–779 (2015).

193. Watson, J. M. in (eds Gal, I. & Garfield, J. B.) chap. Assessing Statistical Thinking Using the Media (IOS Press, 1997).

194. Wikipedia. *Wikipedia:Contributing to Wikipedia* `https://en.wikipedia.org/wiki/Wikipedia:Contributing_to_Wikipedia`. 2020.

195. Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R. & McCarthy, M. A. A comparison of joint species distribution models for presence-absence data. *Methods in Ecology and Evolution* **10,** 198–211 (2019).

196. Wolfe, J. Teaching Students to Focus on the Data in Data Visualization. *Journal of Business and Technical Communication* **29,** 344–359 (2015).

197. Wong-Fannjiang, C. in *BSR: Fall 2019* `https://berkeleysciencereview.com/article/filling-species-gap/` (Berkeley Science Review, 2019).

198. Yackulic, C. B. *et al.* Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution* **4,** 236–243 (2013).

199. Yamaura, Y., Blanchet, F. G. & Higa, M. Analyzing community structure subject to incomplete sampling: hierarchical community model vs. canonical ordinations. *Ecology* **100,** e02759 (2019).

200. Yore, L. D., Hand, B. M. & Florence, M. K. Scientists' Views of Science, Models of Writing, and Science Writing Practices. *Journal of Research in Science Teaching* **41,** 338–369 (2004).

201. Zhang, C., Chen, Y., Xu, B., Xue, Y. & Ren, Y. Comparing the prediction of joint species distribution models with respect to characteristics of sampling data. *Ecography* **41,** 1876–1887 (2018).

202. Zipkin, E. F., DeWan, A. & Royle, J. A. Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. *Journal of Applied Ecology* **46,** 815–822 (2009).

203. Zurell, D., Pollock, L. J. & Thuiller, W. Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography* **41,** 1812–1819 (2018).