Supplementary text for "Placing probabilities on a taxon's true absence: applications to the hominin genus *Paranthropus*"

Andrew Du, Eric Friedlander, John Rowan, Zeresenay Alemseged

Contents

1 Detailed overview of the model

1.1 Motivation

While the methodology presented in this work is broadly applicable to any taxon—regarding the problem of inferring its true absence across a collection of sites—we present the model using the example of the hominin genus *Paranthropus* from eastern Africa, as was done in the main text. Inferring whether *Paranthropus* was truly absent at a site is difficult because observed absence is consistent with two mutually exclusive outcomes: (A) *Paranthropus* never occupied the site (i.e., true absence), or (B) it did occupy the site but has not been sampled yet. We are ultimately interested in estimating the probability of each of these two possibilities, conditional on not finding a *Paranthropus* specimen after sampling n mammalian specimens at a given site. To accomplish this, we first fit a zero-inflated beta-binomial model to the data using maximum likelihood estimation. This model can be viewed as a mixture model wherein each of the two previously mentioned outcomes is modeled with a different probability distribution called a *mixture component*. As we will see, our model, along with Bayes' theorem, will allow us to compute the probability that a given site was generated from possibility (A) or (B), conditional on sampling n mammalian specimens without finding *Paranthropus*, and thus estimate the probability that *Paranthropus* was truly absent from a given site or has not been sampled yet.

1.2 Derivation

We now present the proposed model. In order to instill intuition, we outline the model in a generative fashion (i.e., how data would be generated from the model). Let ψ be the probability that a given site belongs to component (B) (i.e., *Paranthropus* is truly present). In particular, if one were to select a site uniformly at random with no additional information about the recovered specimens, ψ represents the probability that *Paranthropus* was truly present. Intuitively, one can think of ψ as the expected proportion of all sites that contains *Paranthropus*, regardless of whether it has been observed yet. Thus, there is a single ψ that needs to be specified for the model, which applies across all sites. It follows that $1 - \psi$ is the probability that a site belongs to component (A) (i.e., *Paranthropus* was truly absent from the site).

For site i, let Z_i be a latent (unobserved) variable that is equal to 1 if *Paranthropus* was truly present and 0 otherwise. It follows that $P(Z_i = 1) = \psi$ and $P(Z_i = 0) = 1 - \psi$. In the language of statistics, we say that Z_i is a $Bernoulli(\psi)$ random variable and write:

$$Z_i \sim Bernoulli(\psi). \tag{1}$$

For each value of Z_i (i.e., 0 or 1), we must specify the distribution on the number of *Paranthropus* specimens that will be sampled if n_i total mammalian specimens are collected. In general, if we were to know the sampling probability of *Paranthropus*, denoted p_i, then the number of sampled *Paranthropus* specimens can be modeled using a $Binomial(n_i, p_i)$ distribution. Using an example, a $Binomial(n, p)$ random variable represents the number of heads one may observe if flipping a weighted coin n times, given that the probability of obtaining heads from a single flip is p. In our setting, if we know that p_i is the probability that a randomly chosen specimen from site i will be *Paranthropus* (i.e., its sampling probability), then the number of *Paranthropus* specimens out of n_i mammalian specimens has a $Binomial(n_i, p_i)$ distribution.

We now use the value of Z_i to generate a value for p_i. Recall that if $Z_i = 0$, then *Paranthropus* never occupied the ith site, so the probability that a sampled mammalian specimen is *Paranthropus* is zero (i.e., $p_i = 0$). This is the first mixture component in our model, and we say that there is a *point mass* at 0 because all of the mass of the binomial distribution is concentrated at 0 (i.e., if $p_i = 0$, then the number of sampled *Paranthropus* specimens must also equal 0). If $Z_i = 1$, then *Paranthropus* did occupy the site. However, the sampling probability of *Paranthropus* varies from site to site. Therefore, we model p_i as a random effect using a one-parameter beta distribution, which has the probability density function $(1 + \lambda)(1 - p_i)^\lambda$ if $\lambda \geq 0$. This beta distribution is simply a reparametrization of the standard beta distribution with parameters $\alpha = 1$ and $\beta = 1 - \lambda$, i.e., $Beta(1, 1 + \lambda)$ (cf. the "Standard Reflected Beta Distribution" from Wang et al. (2016)). The parameter λ is the same across all sites, and it determines the shape of the distribution, which in our case ($\lambda \geq 0$) can only be monotonically decreasing. This is an appropriate assumption here because such a distribution has the majority of its probability density concentrated towards lower values of p_i, reflecting the general ecological pattern that a species is rare, and thus harder to sample, at most sites within its geographic range (Brown et al., 1995; Murray et al., 1999). As with ψ, we estimate λ from the data. Figure 1 shows how different values of λ affect the shape of the beta distribution.

Combining scenario (A), where *Paranthropus* is truly absent at the ith site and $Z_i = 0$, with scenario (B), where *Paranthropus* is truly present at the ith site and $Z_i = 1$, we can write the distribution of p_i as:

$$P(p_i = w | Z_i = 0) = \begin{cases} 1 & \text{if } w = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2a}$$

$$f(p_i | Z_i = 1) = (1 + \lambda)(1 - p_i)^\lambda, \quad \text{if } \lambda \geq 0, \quad 0 \leq p_i \leq 1. \tag{2b}$$

Equation 2a can also be expressed more conveniently using an *indicator function*: $P(p_i | Z_i = 0) = I(p_i = 0)$, where $I(p_i = 0)$ equals 1 if $p_i = 0$ and 0 otherwise.
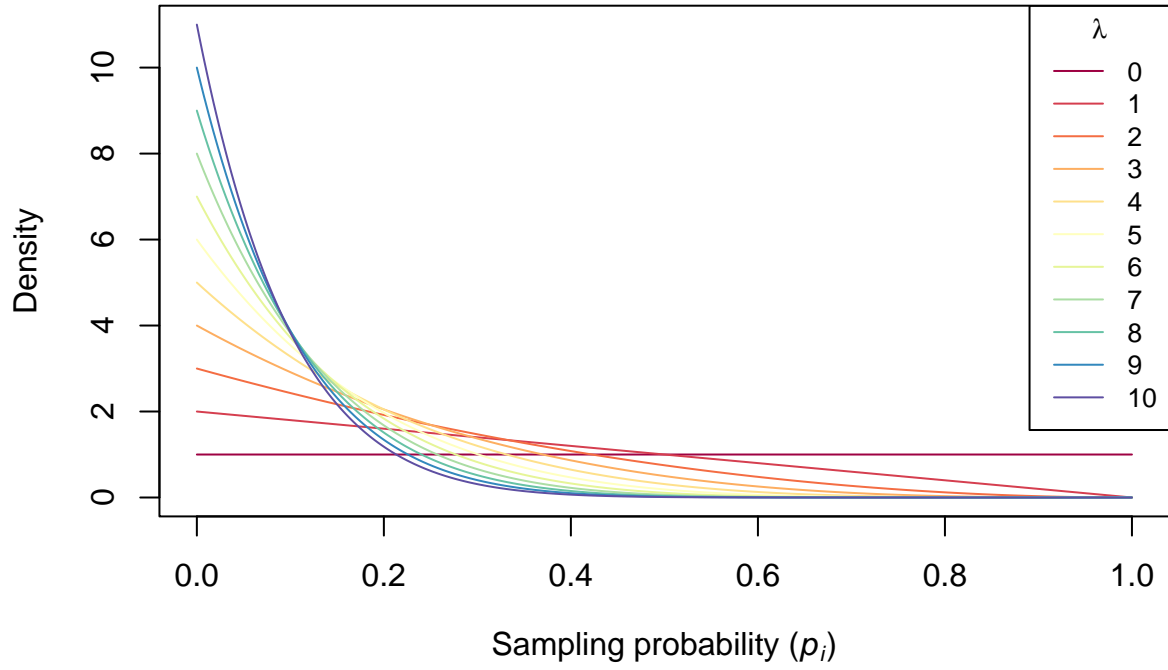
Figure 1: Shape of the monotonically decreasing, one-parameter beta distribution under different values of the parameter, λ. Larger values of λ generate more sharply decreasing functions.

Finally, with p_i defined for each site, we can now model the number of *Paranthropus* specimens recovered at site i, denoted X_i, as:

$$X_i \sim Binomial(n_i, p_i) \tag{3}$$

where n_i is the total number of mammalian specimens recovered at site i and is given by the observed data.

Since we have now specified the entire generative model, we can use this model to compute the probability of recovering, for any i, X_i specimens of *Paranthropus* from the ith site. First, focus on the case of $Z_i = 0$. Recall that if $Z_i = 0$, then $p_i = 0$, and our sampling probability becomes a point mass at 0. Therefore, the number of sampled *Paranthropus* specimens must also be zero (i.e., $X_i = 0$), so:

$$P(X_i = x | Z_i = 0) = I(x_i = 0). \tag{4}$$

Alternatively, if $Z_i = 1$, we must account for both the randomness in selecting p_i and the randomness in sampling X_i given p_i, making the sampling probability hierarchical. Recalling that p_i is distributed according to $Beta(1, 1 + \lambda)$, we have:

$$P(X_i = x | Z_i = 1) \sim Binomial(n_i, p_i), \text{ where } p_i \sim Beta(1, 1 + \lambda). \tag{5}$$

This commonly used distribution is referred to as the *beta-binomial distribution* and has the following probability mass function (PMF):

$$P(X_i = x | Z_i = 1) = f_B(x_i; n_i, \lambda) = \binom{n_i}{x_i} \frac{B(x_i + 1, n_i - x_i + 1 + \lambda)}{B(1, 1 + \lambda)}, \tag{6}$$

where B denotes the beta function.

Combining Equations 1, 4, and 6 by adding over the two possible values of the latent variable Z_i, the law of total probability implies that:

$$\begin{aligned}
P(X_i = x) &= P(X_i = x | Z_i = 0)P(Z_i = 0) + P(X_i = x | Z_i = 1)P(Z_i = 1) \\
&= I(x = 0)(1 - \psi) + f_B(x; n_i, \lambda)\psi.
\end{aligned} \tag{7}$$

Equation 7 is the probability of recovering x number of *Paranthropus* specimens for the ith site if the sample size is n_i.

1.3   Estimating model parameters

In the previous section, we discussed how the data are generated by our model, but ultimately, we are interested in going in the opposite direction: that is, given the data, $X_1, X_2, ...X_{56}$ (where X_1 is the number of *Paranthropus* specimens in the first site, X_2 is the number of *Paranthropus* specimens in the second site, and we have a total of 56 sites), what are the parameters (ψ, λ)? Furthermore, we are interested in using these parameter estimates to infer the latent variables $Z_1, Z_2, ..., Z_{56}$, using data from both within and outside the ith site. In other words, we would like to assess the probability that a given site has *Paranthropus* truly present or not, conditional on observations from the site and parameter estimates in our model (which are informed by data from all sites). To estimate the parameters (ψ, λ), we use a method called *maximum likelihood estimation*. Intuitively, this method selects the parameter values that will make the data most probable, with larger probabilities indicating more likely parameter estimates (Wang, 2010). Using the

estimated parameter values, data, and Bayes' theorem (see next section), we can estimate the probability that $Z_i = 1$, conditional on the parameter estimates and data, which is the conditional probability that *Paranthropus* was truly present at the site; the complement of this probability is the conditional probability that *Paranthropus* was truly absent from the ith site. We note that the case that is most interesting is when no *Paranthropus* specimens have yet been observed at the given site (i.e., $X_i = 0$). To do all this, we operate in the framework of likelihood and define a *likelihood function.*

A likelihood function quantifies information about model parameters that is contained in the data (Wang et al., 2016). It is numerically equivalent to the *joint probability* of the data given the parameters, so to get the likelihood function, we multiply the probability of the data by itself N times, where N is sample size (Wang, 2010): e.g., $L(\psi, \lambda | X) = \prod_{i=1}^{N} P(X_i | \psi, \lambda)$. For our study, the probability of the data is given by Equation 7, so our likelihood function is defined as:

$$L(\psi, \lambda | X) = P(X | \psi, \lambda) = \prod_{i=1}^{N} I(x_i = 0)(1 - \psi) + f_B(x_i; n_i, \lambda)\psi. \tag{8}$$

We then use the L-BFGS-B optimization algorithm (Byrd et al., 1995) to find the parameter values (ψ, λ) that maximize the likelihood function (Equation 8). In practice, we work with the log transformation of Equation 8 for analytical tractability.

1.4 Bayes' theorem

For this subsection, we first provide a brief introduction to Bayes' theorem and what the equation aims to accomplish, with respect to our research question of interest. We then show how we use Bayes' theorem to derive our main probability of interest: the probability that *Paranthropus* was truly absent at a given site, conditional on the estimated parameters and data.

1.4.1 Introduction to Bayes' theorem

In scientific inference, we understand that a process(es) generates a pattern(s) (i.e., data), and we ultimately want to learn something about the generative process from the resulting data. In Bayesian inference, processes are described with models and their unobserved quantities (e.g., latent variables, parameters), and we learn something about these unobserved quantities of interest from observed data (Hobbs and Hooten, 2015). For example, whether *Paranthropus* was truly absent or present at a given site is an unobserved quantity (i.e., latent variable), and we want to learn something about this latent variable at each site, given the observed data (i.e., sampling X_i number of *Paranthropus* specimens after sampling n_i number of mammalian specimens). Equations 4 and 6 tell us how the data are generated, given that we know (or assume) the true state of the latent variable at a given site, i.e., $P(X_i | Z_i)$. However, we want our direction of inference to go in the opposite direction: we want to transpose $P(X_i | Z_i)$ to get $P(Z_i | X_i)$, i.e., we want to infer the latent variable from the observed data. Bayes' theorem tells us how to transpose conditional probabilities, using the basic algebraic rules of conditional probability:

$$P(Z_i | X_i) = \frac{P(X_i | Z_i) P(Z_i)}{P(X_i)}. \tag{9}$$

Thus, based on this transposition of conditional probabilities (Equation 9), Bayes' theorem enables us to infer something about unobserved processes of interest (e.g., latent variables, parameters) from the observed data.

Explaining each component of Equation 9, $P(Z_i | X_i)$ is the *posterior probability*, $P(X_i | Z_i)$ is the *likelihood*, $P(Z_i)$ is the *prior probability*, and $P(X_i)$ is the *marginal distribution of the data.* The likelihood connects observed data with unobserved quantities in the model, while the prior probability summarizes what we know about the unobserved quantities before our study is conducted. For our model, the parameter determining the prior distribution, ψ (Equation 1), is estimated from the data, so our model is an empirical Bayes one (Carlin

and Louis, 2000). The marginal distribution of the data serves as a normalizing factor, ensuring that the area under the posterior probability curve equals one. This means that the posterior probability distribution is a true probability density/mass function, and we can apply to it the rules and intuitions associated with probability theory.

1.4.2 Applying Bayes' theorem to our research question

Once our model's ψ and λ parameters are estimated (Section 1.3), we use Bayes' theorem to compute our main probability of interest: the posterior probability that *Paranthropus* was truly absent at the ith site (i.e., $Z_i = 0$), given the estimated parameters and data. This probability equals zero for those sites where *Paranthropus* is observed to be present, so we focus on deriving the posterior probability equation for those sites where *Paranthropus* is currently observed to be absent. Fleshing out Equation 9 by including additional data (n_i) and the estimated parameters ($\hat{\psi}, \hat{\lambda}$):

$$
\begin{aligned}
P(Z_i = 0 | X_i = 0, n_i, \hat{\psi}, \hat{\lambda}) &= \frac{P(X_i = 0 | Z_i = 0, n_i, \hat{\psi}, \hat{\lambda}) P(Z_i = 0 | n_i, \hat{\psi}, \hat{\lambda})}{P(X_i = 0 | n_i, \hat{\psi}, \hat{\lambda})} \\
&= \frac{1 \times (1 - \hat{\psi})}{I(X_i = 0)(1 - \hat{\psi}) + f_B(0; n_i, \hat{\lambda}) \hat{\psi}} \\
&= \frac{(1 - \hat{\psi})}{1 \times (1 - \hat{\psi}) + \frac{\Gamma(1)\Gamma(n_i + 1 + \hat{\lambda})}{\Gamma(n_i + 2 + \hat{\lambda})} \frac{\Gamma(2 + \hat{\lambda})}{\Gamma(1)\Gamma(1 + \hat{\lambda})} \hat{\psi}} \\
&= \frac{1 - \hat{\psi}}{1 - \hat{\psi} + \frac{(n_i + \hat{\lambda})!}{(n_i + 1 + \hat{\lambda})!} \frac{(1 + \hat{\lambda})!}{\lambda!} \hat{\psi}} \\
&= \frac{1 - \hat{\psi}}{1 - \hat{\psi} + \frac{1 + \hat{\lambda}}{n_i + 1 + \hat{\lambda}} \hat{\psi}}.
\end{aligned}
\tag{10}
$$

In Equation 10 going from Lines 1 to 2, the likelihood, $P(X_i = 0 | Z_i = 0, n_i, \hat{\psi}, \hat{\lambda})$, equals one because the probability that we sample zero *Paranthropus* specimens at a given site where it is truly absent ($Z_i = 0$) equals one. The equation for the prior, $P(Z_i = 0 | n_i, \hat{\psi}, \hat{\lambda})$, is given by Equation 1, and the equation for the marginal distribution of the data, $P(X_i = 0 | n_i, \hat{\psi}, \hat{\lambda})$, is given by Equation 7.

With Equation 10, we plug in our estimated parameters ($\hat{\psi}, \hat{\lambda}$) and data for the ith site (i.e., the number of sampled mammalian specimens, n_i) into Equation 10 to get our posterior probability of interest. The complement of this probability—the posterior probability that *Paranthropus* was truly present at a site given the estimated parameters and data—can be calculated as $P(Z_i = 1 | X_i = 0, n_i, \hat{\psi}, \hat{\lambda}) = 1 - P(Z_i = 0 | X_i = 0, n_i, \hat{\psi}, \hat{\lambda})$ (Equation 10).

1.5 Quantifying uncertainty in the parameter estimates and posterior probabilities

As discussed in the main text, we quantify uncertainty by bootstrapping the data (1,000 iterations). For each bootstrapped iteration, we first resampled our 56 sites with replacement. Next for each resampled site, we resampled with replacement the observed number of sampled *Paranthropus* specimens. We then estimate the ψ and λ parameters, following the protocol outlined in Section 1.3. Finally, we calculate the posterior probability of *Paranthropus* true absence using the estimated parameters and resampled data (Equation 10). All this is repeated 1,000 times to get sampling distributions for $\hat{\psi}$, $\hat{\lambda}$, and the posterior probability of true absence for each of the 56 sites. We computed 95% confidence intervals for each sampling distribution using the `quantile` function in `R`.

2   Model selection

To assess our model's goodness of fit and whether it is overly complex (e.g., too many parameters), we compared our model to others using the small-sample, bias-corrected Akaike Information Criterion (AIC_c) (Burnham and Anderson, 2002). The AIC_c formula is:

$$\text{AIC}_c = -2\log(L(\hat{\theta})) + 2K\left(\frac{n}{n-K-1}\right),\tag{11}$$

where $L(\hat{\theta})$ is the likelihood of the fitted model, K is the number of model parameters, and n is sample size (in our case, the number of sites: 56).

AIC_c balances goodness of fit (the likelihood) with model complexity (penalties for extra parameters), and compared to AIC, AIC_c has an additional correction factor for small sample sizes (Burnham and Anderson, 2002). As sample size increases, AIC_c converges on the traditional AIC.

Lower scores of AIC_c indicate better model fit, and to facilitate interpretation of AIC_c scores, we transform AIC_c into weights (Burnham and Anderson, 2002):

$$w_i = \frac{\exp(-0.5\Delta_i)}{\sum_{r=1}^{R}\exp(-0.5\Delta_r)},\tag{12}$$

where i indicates the ith model considered, Δ_i is the AIC_c of the ith model minus the minimum AIC_c across all models, and R is the total number of investigated models. AIC_c weights sum to one across all models, and larger weights indicate better model support.

2.1   Other investigated models

We compared our main zero-inflated beta-binomial model (two parameters: ψ, λ) to four other models:

1. binomial model (one parameter: p)
2. beta-binomial model (one parameter: λ)
3. zero-inflated binomial (two parameters: ψ, p)
4. zero-inflated beta-binomial with a two-parameter beta distribution (three parameters: ψ, α, β).

All models are fit to the observed data using the L-BFGS-B optimization algorithm (Byrd et al., 1995).

2.1.1   Binomial model

This model has no zero inflation (i.e., no probability point mass on zero), and the *Paranthropus* sampling probability is fixed at a single value, p, which does not vary across sites. The PMF for the model is:

$$P(X_i = x) = \binom{n_i}{x}p^x(1-p)^{n_i-x}.\tag{13}$$

2.1.2   Beta-binomial model

This model also has no zero inflation and is a binomial model (Equation 13), wherein p is allowed to vary across sites as a random effect, according to a monotonic, one-parameter beta distribution: $Beta(1, 1+\lambda)$. The model PMF is found in Equation 6 but is replicated below:

$$P(X_i = x) = \binom{n_i}{x}\frac{B(x+1, n_i - x + 1 + \lambda)}{B(1, 1+\lambda)},\tag{14}$$

where B denotes the beta function.

2.1.3 Zero-inflated binomial model

This model is equivalent to the binomial model (Equation 13) but has an additional probability point mass on zero. The PMF is:

$$P(X_i = x) = I(x = 0)(1 - \psi) + \binom{n_i}{x} p^x (1 - p)^{n_i - x} \psi. \tag{15}$$

2.1.4 Zero-inflated beta-binomial model (three parameters)

This model is the most complex and is equivalent to our main model (Equation 7), but the beta distribution for the *Paranthropus* sampling probability has two parameters (α, β) instead of one. Therefore, the sampling probability distribution is allowed to be non-monotonic. The PMF is:

$$P(X_i = x) = I(x = 0)(1 - \psi) + \binom{n_i}{x} \frac{B(x + \alpha, n_i - x + \beta)}{B(\alpha, \beta)} \psi. \tag{16}$$

2.2 AIC$_c$ results

Table 1 shows the model selection results. We excluded results for our most complex model (three-parameter zero-inflated beta-binomial) for reasons described in Section 3.

Table 1: Model selection results using AIC$_c$.

Model	Log-likelihood	Number of parameters	AIC$_c$	ΔAIC$_c$	AIC$_c$ weight
Binomial	-343.04	1	688.15	492.73	0.0000
Beta-binomial	-106.36	1	214.79	19.38	0.0001
Zero-inflated binomial	-286.25	2	576.73	381.31	0.0000
Zero-inflated beta-binomial (two parameters)	-95.59	2	195.41	0.00	0.9999

Results show that our main model—the zero-inflated beta-binomial with two parameters—fits the data the best.

3 Simulations to compare our main model with a more complex one

When the most complex zero-inflated beta-binomial model (three parameters) is included in the model selection process (Section 2), it has the smallest AIC$_c$ score (190.68) and the largest AIC$_c$ weight (0.91). However, this model's parameter estimates seemed nonsensical to us ($\hat{\psi} = 1.0$, $\hat{\alpha} = 0.2$, $\hat{\beta} = 56.3$). The ψ estimate in particular implies that *Paranthropus* is truly present at all sites, regardless of whether it's been found yet. This contradicts what we know from ecological theory: a species is not present at all sites within its geographic range (Rapoport, 1982; Hurlbert and White, 2005, 2007). Therefore, we conducted simulations to assess how well the three-parameter zero-inflated beta-binomial (ZI-BB3) model estimated known parameter values. We repeated this process for our main model: the two-parameter zero-inflated beta-binomial (ZI-BB2).

3.1 Simulation details

For both sets of simulations, simulated ψ values are {0.1, 0.3, 0.5, 0.7, 0.9}. For the ZI-BB2 model, simulated λ values are {1, 3, 5, 10, 25, 50, 100, 200}. For the ZI-BB3 model, simulated α values are {0.1, 0.3, 0.5, 0.7, 0.9, 1.1} and simulated β values are {10, 25, 50, 100, 250, 500, 1000}. We chose small α values due to the small value of α estimated from the data ($\hat{\alpha} = 0.2$).

We simulate *Paranthropus* abundances across sites, given known parameter values: ψ and λ for the ZI-BB2 simulations, and ψ, α, and β for the ZI-BB3 simulations. We then estimate these parameters with their corresponding model and compare the estimates to their predetermined values. We use the observed number of sites and number of mammalian specimens across sites for the simulations, thereby varying the parameter values only. We conduct the simulations as follows:

1) To determine at which sites *Paranthropus* was truly absent or present (i.e., Z_i), we take a random draw from a Bernoulli distribution, where the number of trials is the number of sites, and the probability of success (i.e., *Paranthropus* is present) is the prespecified value of ψ. The output is a vector of 0s and 1s, where the length of the vector is the total number of sites, 0 denotes true *Paranthropus* absence, 1 denotes true *Paranthropus* presence, and the proportion of 1s equals ψ on average.

2) For those sites that have *Paranthropus* (i.e., $Z_i = 1$), we simulate *Paranthropus* abundances (i.e., X_i) as a random draw from the beta-binomial distribution, given the observed number of mammalian specimens at each site and either a prespecified λ value (Equation 6) or prespecified α and β values (right-hand side of plus sign in Equation 16, without the ψ). We do this using the `rbetabinom` function in the `rmutil` package (Swihart and Lindsey, 2020).

3) The appropriate zero-inflated beta-binomial model is fit to the simulated data, generating parameter estimates.

4) Steps 1–3 are iterated 1000 times, yielding 1000 estimates of each parameter.

3.2 Simulation results

3.2.1 Two-parameter, zero-inflated beta-binomial model (ZI-BB2)

Figure 2 shows that the ZI-BB2 model produces unbiased estimates of ψ at all simulated values of ψ and λ. The variation around $\hat{\psi}$—when considered across all values of λ—is at its maximum when $\psi = 0.5$. For a given simulated iteration of 0s and 1s determining *Paranthropus* true absence or presence at sites (i.e., Z_i) (see Step 1 in Section 3.1), the number of 1s follows a binomial distribution. It is known that the variance of a binomial random variable is at its maximum when the probability of success equals 0.5, so this likely explains the increased variation of $\hat{\psi}$ when $\psi = 0.5$.

The variation around $\hat{\psi}$ also increases as λ increases (Figure 2). This is due to larger values of λ concentrating more of the sampling probability density function (p_i) at smaller values (Figure 1). As such, simulations with large λ produce more zero *Paranthropus* specimens across sites, even when *Paranthropus* is truly present (i.e., false absences). The model 'struggles' to distinguish between true and false absences, leading to greater variation around $\hat{\psi}$. Indeed, we see this problem is at its worst when $\psi = 0.1$ and $\lambda = 200$ (Figure 2). These are the parameter values that would produce the most zero *Paranthropus* specimens: small ψ means more true absences and large λ means more false absences.

Figure 3 shows that the model produces unbiased estimates of λ at most values of simulated ψ and λ; there is a slight positive bias in $\hat{\lambda}$ when $\psi = 0.1$ (+56 when averaged across all values of λ). Such a low value for ψ means that *Paranthropus* is simulated to be truly absent at most sites. The positive bias in λ suggests that the model is confusing true absences for false ones: the model is identifying many of the observed zeros as false absences and is estimating a larger λ (Figure 1) to account for these unsampled presences. This phenomenon also explains why the variation around $\hat{\lambda}$ increases as ψ decreases.

All in all, the ZI-BB2 model performs well when the simulated parameter values are near their counterparts, as estimated from the data ($\hat{\psi} = 0.52$, $\hat{\lambda} = 153$).
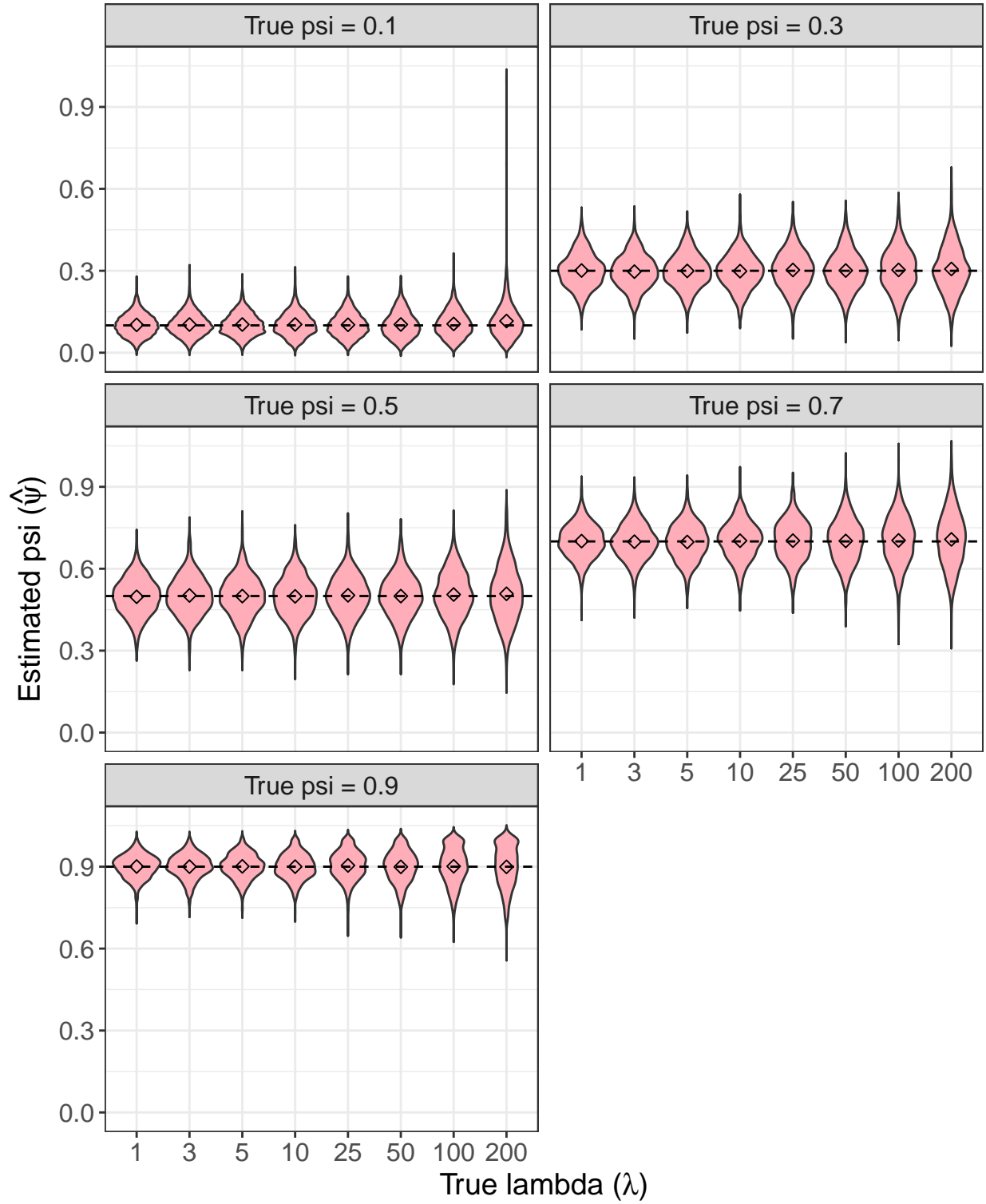
Figure 2: Simulation results showing the two-parameter, zero-inflated beta-binomial model's ability to estimate ψ for various values of ψ and λ. The horizontal dashed lines indicate true values of ψ, while the diamonds indicate the average estimated ψ values.
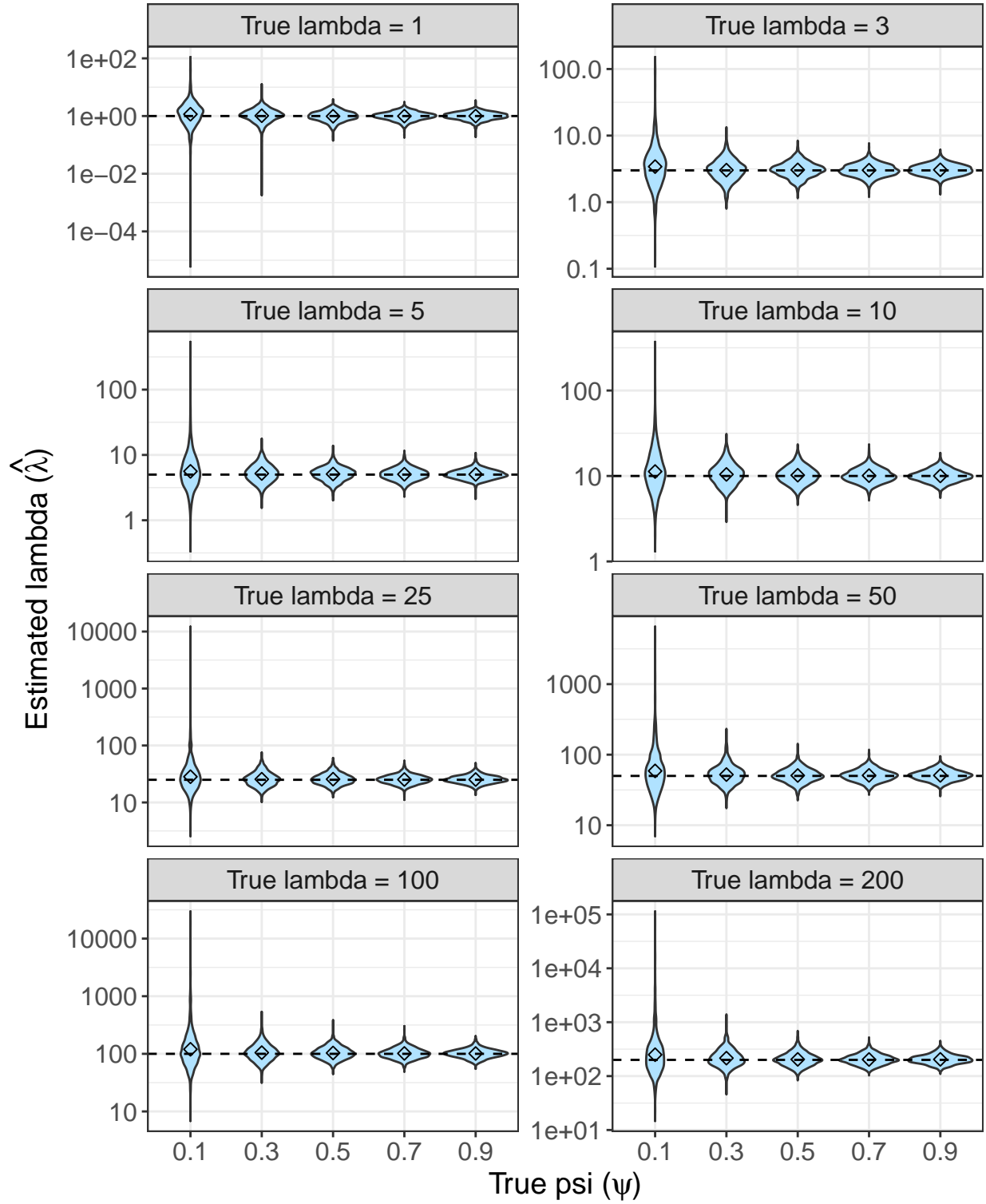
Figure 3: Simulation results showing the two-parameter, zero-inflated beta-binomial model's ability to estimate λ for various values of ψ and λ. The horizontal dashed lines indicate true values of λ, while the diamonds indicate the average estimated λ values.

3.2.2 Three-parameter, zero-inflated beta-binomial model (ZI-BB3)

Simulation results show that almost all ZI-BB3 parameter estimates are biased (Figures 4–9). For $\hat{\psi}$, this bias is proximately caused by the bimodal nature of its sampling distribution: one mode is located at 1.0, while the other is located near ψ's true value (Figures 4 & 5) (the bimodal sampling distribution of $\hat{\psi}$, or any parameter, indicates the existence of local optima in the likelihood function). The magnitude of $\hat{\psi}$'s bias is directly related to the number of simulated iterations where $\hat{\psi} = 1.0$ (Figures 4 & 5). The number of simulated iterations where $\hat{\psi} = 1.0$ is itself caused by those iterations where there are many simulated zeros, i.e., zero *Paranthropus* specimens across sites ($X_i = 0$ for most sites). Many simulated zeros occur when ψ is smaller, which produces more true absences across sites, and when α is smaller and β is larger, which produces more false absences across sites (e.g., expected *Paranthropus* sampling probability, $E(p_i)$, equals $\alpha/(\alpha + \beta)$); these parameter values cause the greatest degree of bias in $\hat{\psi}$ (Figures 4 & 5). With lots of simulated zeros across sites, the model sometimes interprets *all* zeros, incorrectly, as false absences: $\hat{\psi} = 1.0$ implies that *Paranthropus* is truly present at *all* sites, and the false absences are modeled with smaller $\hat{\alpha}$ and larger $\hat{\beta}$. This phenomenon also explains the sampling distribution modes at $\hat{\alpha} < 0.1$ (Figures 6 & 7) and $\hat{\beta} > 10,000$ (Figures 8 & 9) when ψ is small, α is small, and β is large. When ψ is larger (i.e., 0.7 or 0.9), those parameter values that generate more false absences (smaller α and larger β) create more variation in $\hat{\psi}$, which 'pulls' $\hat{\psi}$ away from 1.0 and generates a negative bias (Figures 4 & 5).

All estimated α and β values are biased upwards (Figures 6–9). The magnitude of these biases are larger when ψ is smaller, α is smaller, and β is larger, i.e., those parameter values that generate many simulated zeros across sites. These zeros 'confuse' the ZI-BB3 model, which sometimes treats the zeros as false absences (i.e., smaller $\hat{\alpha}$ and larger $\hat{\beta}$) or true absences (i.e., larger $\hat{\alpha}$ and $\hat{\beta}$ values). It is these larger $\hat{\alpha}$ and $\hat{\beta}$ values—which still result in small expected *Paranthropus* sampling probabilities (on average, ~0.005)—that generate the long, positive tails in these sampling distributions (Figures 6–9); this in turn drives the positive bias in these parameter estimates.

In sum, the biased parameter estimates of the ZI-BB3 model indicate that the model is misspecified and overly complex (i.e., too many parameters). Regarding $\hat{\psi}$, the erroneous sampling distribution modes at 1.0 (Figures 4 & 5) indicate that the result of $\hat{\psi} = 1.0$ when the ZI-BB3 model is fit to the data is also incorrect. This suggests that the α and β estimates are also biased and incorrect. As such, we view the ZI-BB3 model as inappropriate for our data, and we do not consider it any further.

4 Assessing model assumptions

Our model makes two assumptions regarding data independence:

1) Site data are independent from each other.
2) Within sites, specimen data are independent from each other.

4.1 Assumption #1: independence of site data

4.1.1 Regarding ψ

Regarding assumption #1 in estimating ψ, this would be violated if true *Paranthropus* presence/absence is spatiotemporally autocorrelated across sites. That is, sites that are closer together in space and/or time are more likely to exhibit the same state of *Paranthropus* presence/absence (e.g., closer sites all have *Paranthropus* present). To assess this assumption, we first take observed *Paranthropus* presence/absence data across sites and assume that they reflect true presence/absence. We know that this is strictly not true as an observed absence can reflect an unsampled presence (after all, this is the general problem we are trying to address), but it can also be reasonably assumed that observed presence/absence is more reflective of true presence/absence, rather than complete noise. We then consider all pairwise comparisons of sites in terms of (1) whether their observed *Paranthropus* presence/absence states match (1 = match, 0 = otherwise) and (2) their temporal and spatial distances from each other. We finally use multiple logistic regression to model (1) as a function of (2) with an interaction term between temporal and spatial distance (after centering and scaling the independent
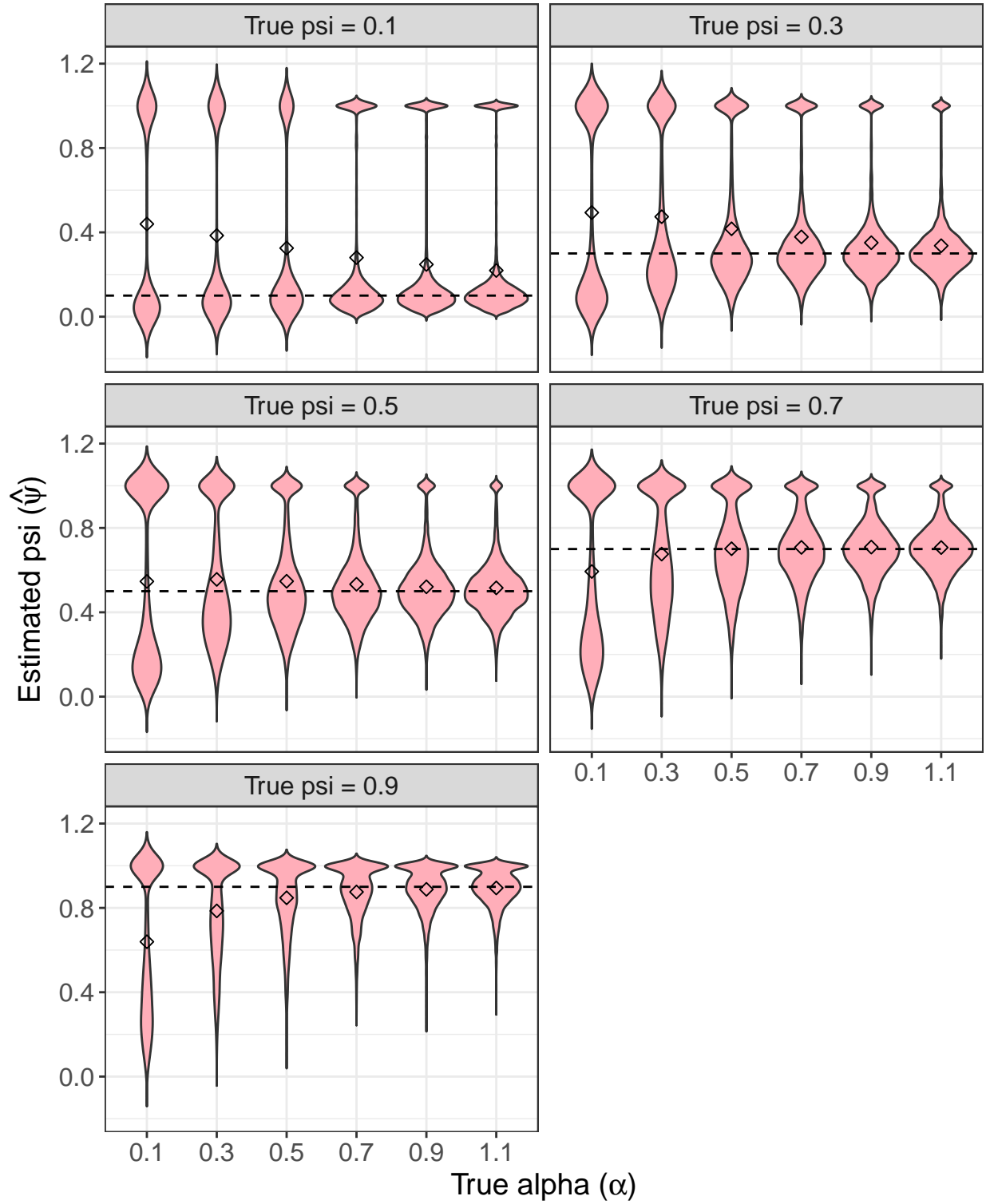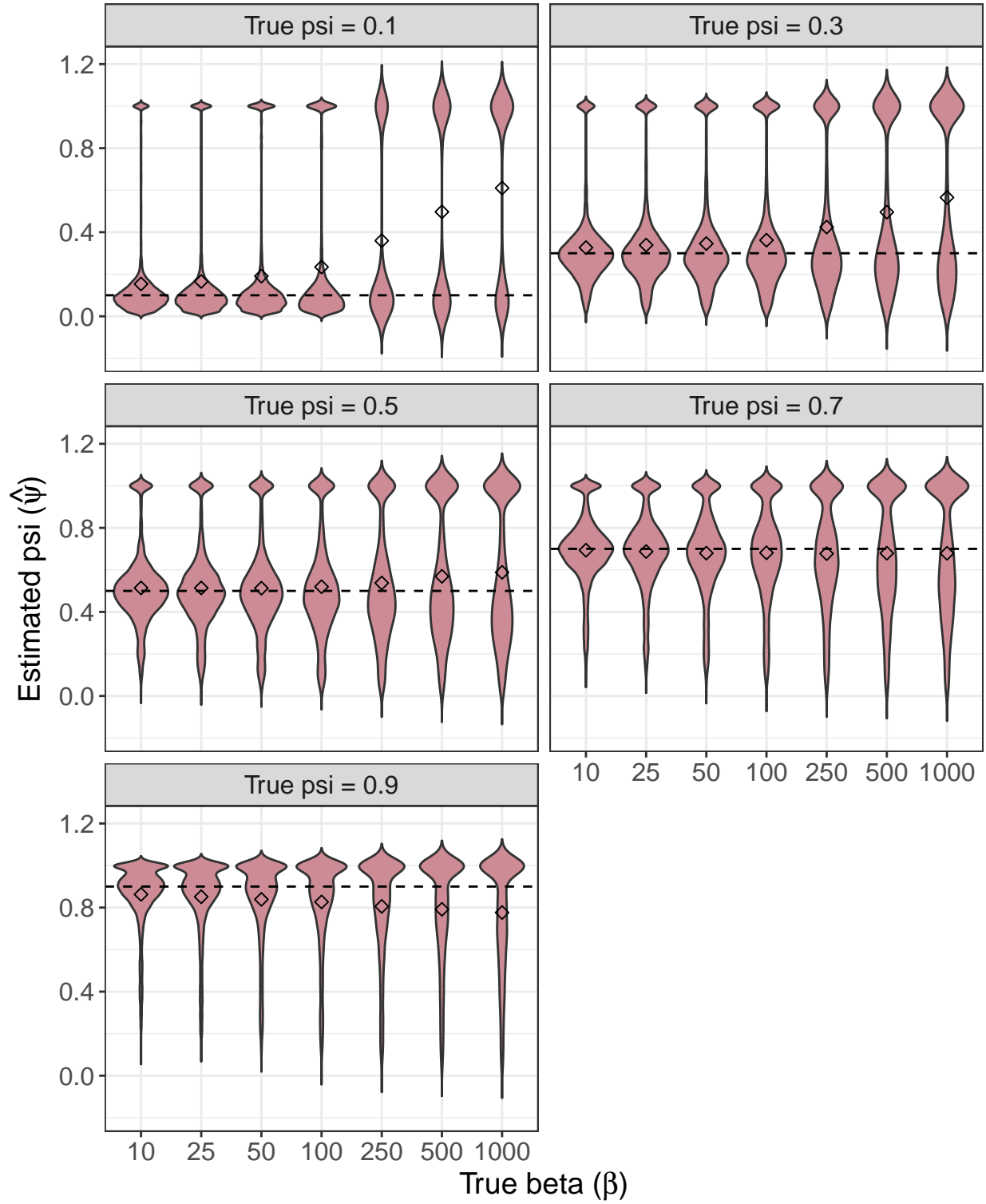
Figure 4: Simulation results showing the three-parameter, zero-inflated beta-binomial model's ability to estimate ψ for various values of ψ and α. The horizontal dashed lines indicate true values of ψ, while the diamonds indicate the average estimated ψ values.
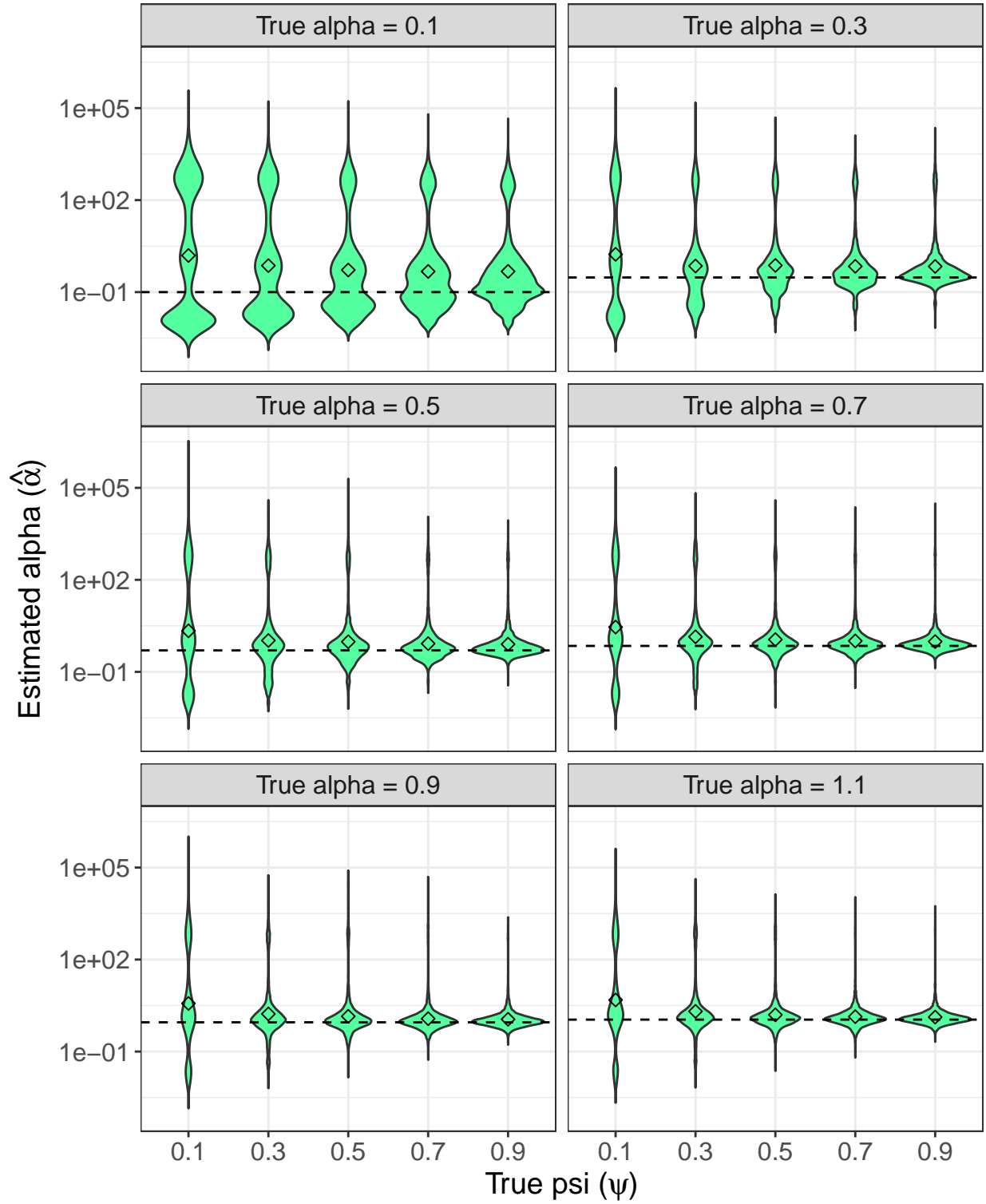
Figure 5: Simulation results showing the three-parameter, zero-inflated beta-binomial model's ability to estimate ψ for various values of ψ and β. The horizontal dashed lines indicate true values of ψ, while the diamonds indicate the average estimated ψ values.
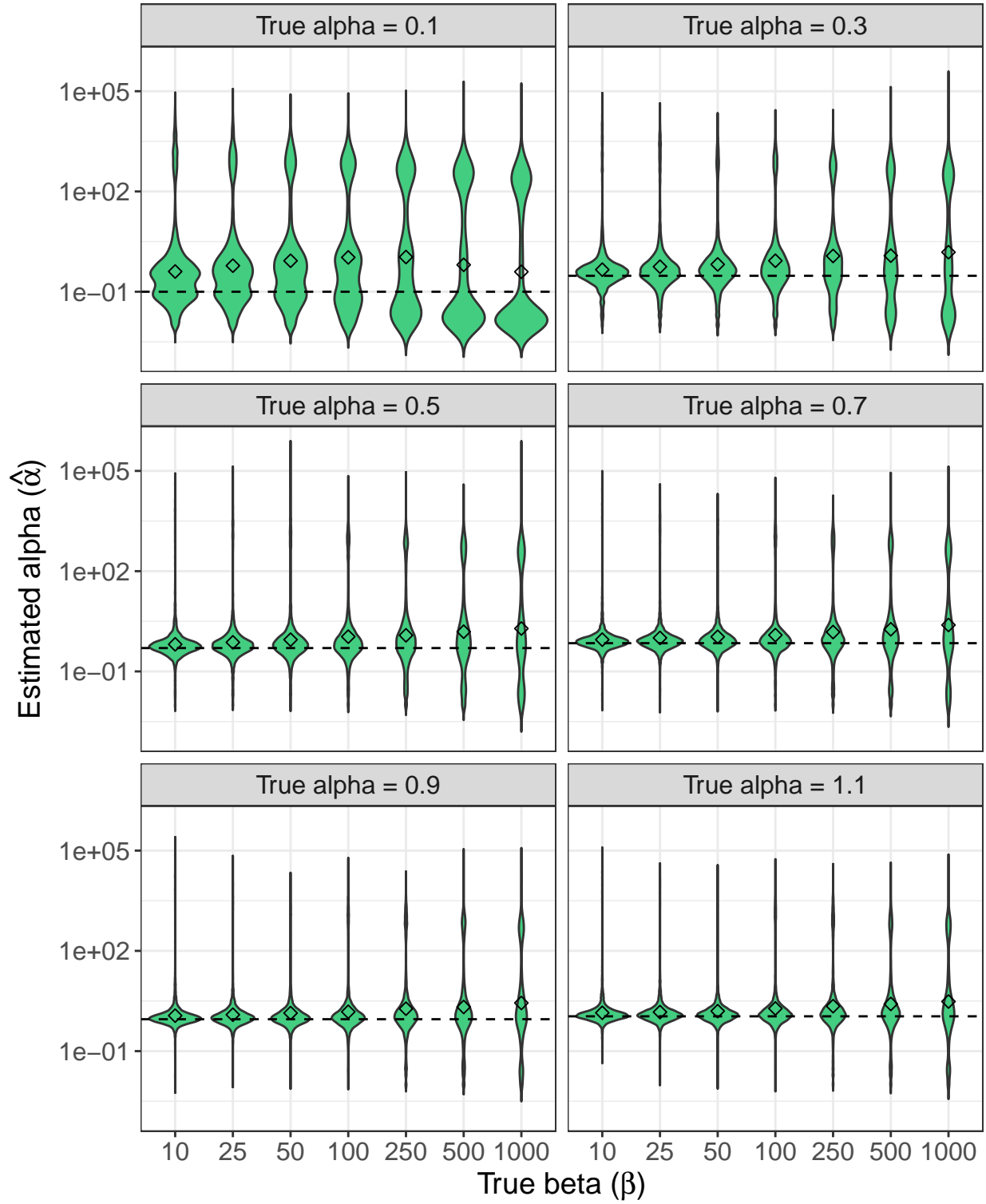
Figure 6: Simulation results showing the three-parameter, zero-inflated beta-binomial model's ability to estimate α for various values of α and ψ. The horizontal dashed lines indicate true values of α, while the diamonds indicate the average estimated α values.

Figure 7: Simulation results showing the three-parameter, zero-inflated beta-binomial model's ability to estimate α for various values of α and β. The horizontal dashed lines indicate true values of α, while the diamonds indicate the average estimated α values.
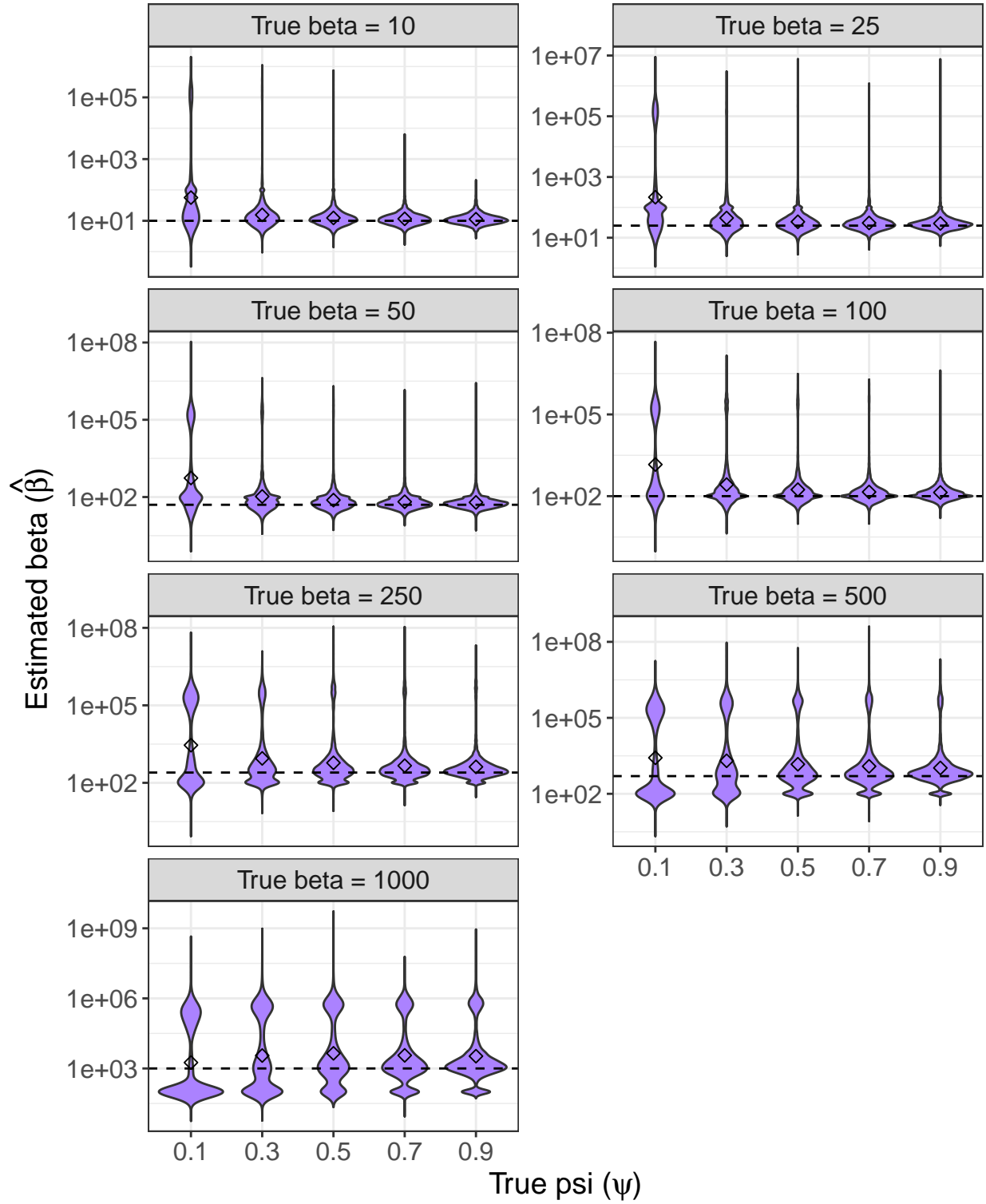
Figure 8: Simulation results showing the three-parameter, zero-inflated beta-binomial model's ability to estimate β for various values of β and ψ. The horizontal dashed lines indicate true values of β, while the diamonds indicate the average estimated β values.
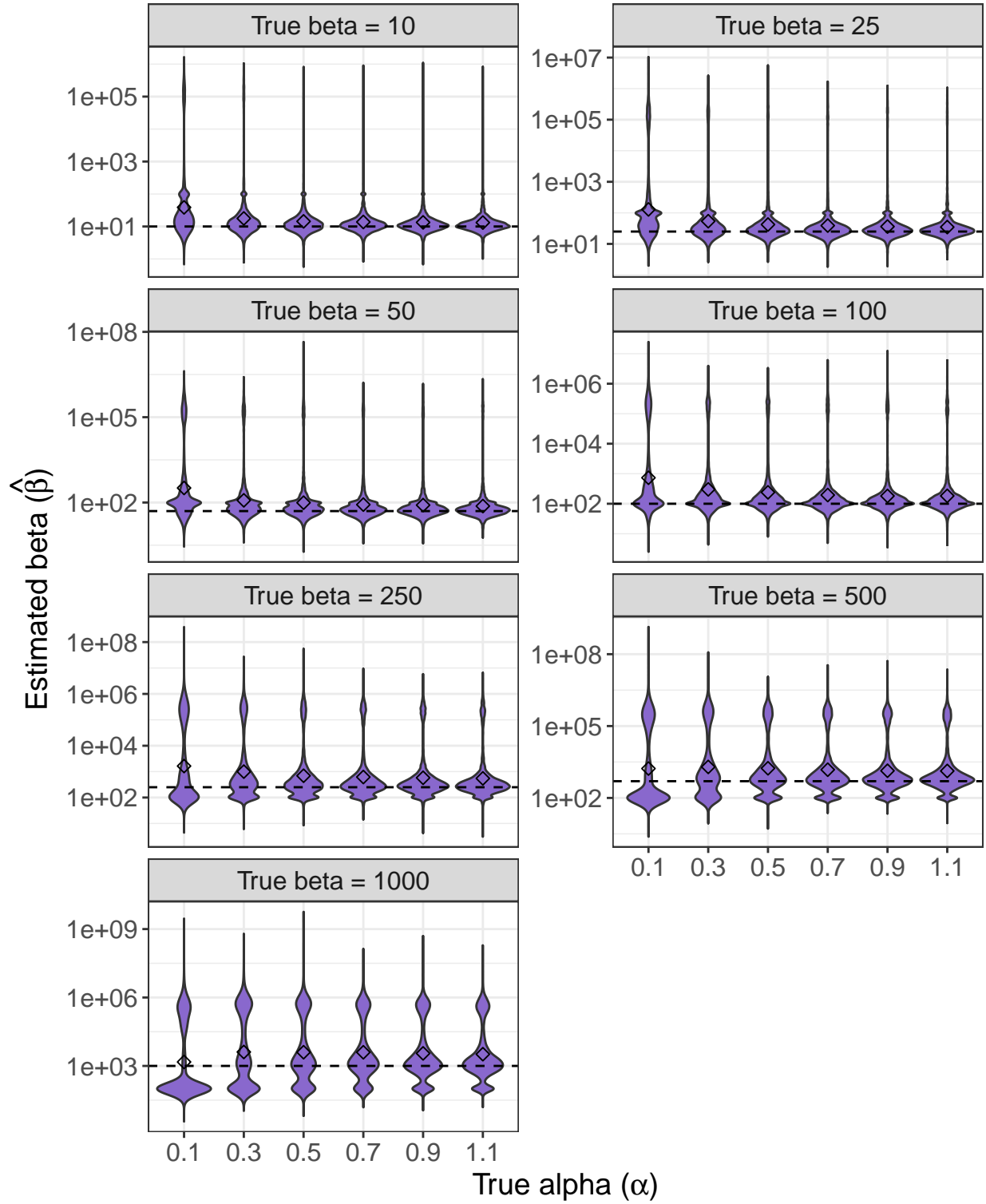
Figure 9: Simulation results showing the three-parameter, zero-inflated beta-binomial model's ability to estimate β for various values of β and α. The horizontal dashed lines indicate true values of β, while the diamonds indicate the average estimated β values.

variables) to see if sites that are closer together in time and/or space are more likely to exhibit the same observed *Paranthropus* presence/absence state.

Table 2 shows that the logistic regression coefficient estimates are all small. We will not discuss the interaction term given its negligible magnitude (indeed, coefficient estimates are virtually identical when no interaction term is included). Firstly, the coefficient estimates are expected to be negative: an increase in spatial and/or temporal distance between sites should lead to differences in *Paranthropus* presence/absence states (i.e., *Paranthropus* is present at one site but absent at the other); recall that we coded different states as 0s, while matches are 1s. We can exponentiate the logistic regression coefficients to make them more interpretable using the language of odds. Odds are defined, in our case, as the probability of *Paranthropus* presence/absence states matching divided by the probability that they do not match. Therefore, an odds of 2.0 means that it is twice as likely that *Paranthropus* presence/absence states will match rather than not. When exponentiated, the intercept is 1.19, meaning that the probability of *Paranthropus* states matching is only 19% more likely than them not matching when the independent variables are set to zero (zero values in our case indicate the average pairwise temporal and spatial distance in our dataset because we centered the variables prior to analysis). The exponentiated coefficients for "scale(temporal distance)" and "scale(spatial distance)" are 1.08 and 0.95, respectively. Thus, when holding spatial distance constant, a one standard deviation increase in temporal distance results in a 8% *increase* in the odds of *Paranthropus* presence/absence states matching, opposite the direction we would expect. Holding temporal distance constant, a one standard deviation increase in spatial distance leads to a 5% decrease in the odds of *Paranthropus* presence/absence states matching. This translates to a decrease in odds from 1.19 to 1.13 when the variable "scale(spatial distance)" increases from 0 to 1 (while holding temporal distance constant). However, because we centered and scaled the data prior to analysis, this unit increase is quite substantial (Figure 10). In sum, there appears to be no temporal autocorrelation in *Paranthropus* presence/absence states, but there might be some very slight spatial autocorrelation.

4.1.2 Regarding λ

Regarding assumption #1 in estimating λ, this would be violated if *Paranthropus* sampling probability is spatiotemporally autocorrelated across sites. That is, sites that are closer together in space and/or time are more likely to have similar *Paranthropus* sampling probability values. As mentioned in the main text, *Paranthropus* sampling probability at site i can be equated to its observed relative abundance at that site. We consider all pairwise comparisons of sites in terms of their (1) differences in observed *Paranthropus* relative abundance and (2) temporal and spatial distances from each other. As with the the presence/absence logistic regression, we model (1) as a function of (2) with an interaction between the centered and scaled independent variables, but this time using multiple ordinary least squares regression. Again, the goal is to assess whether sites that are closer together in time and/or space are more similar in their observed *Paranthropus* relative abundances.

Table 2 shows that all regression coefficients are small, and the multiple R^2 of the entire model is only 0.02. Considering all pairwise comparisons between sites, Figure 11 plots observed *Paranthropus* relative abundance differences as a function of temporal (Figure 11A) or spatial (Figure 11B) distances. We expect the relationships to be positive, where sites that are further away temporally or spatially are expected to have more dissimilar relative abundance values. The observed relationships are in fact *negative* (Figure 11), opposite the expected direction. Therefore, it appears that observed *Paranthropus* relative abundances across sites are not spatiotemporally autocorrelated.
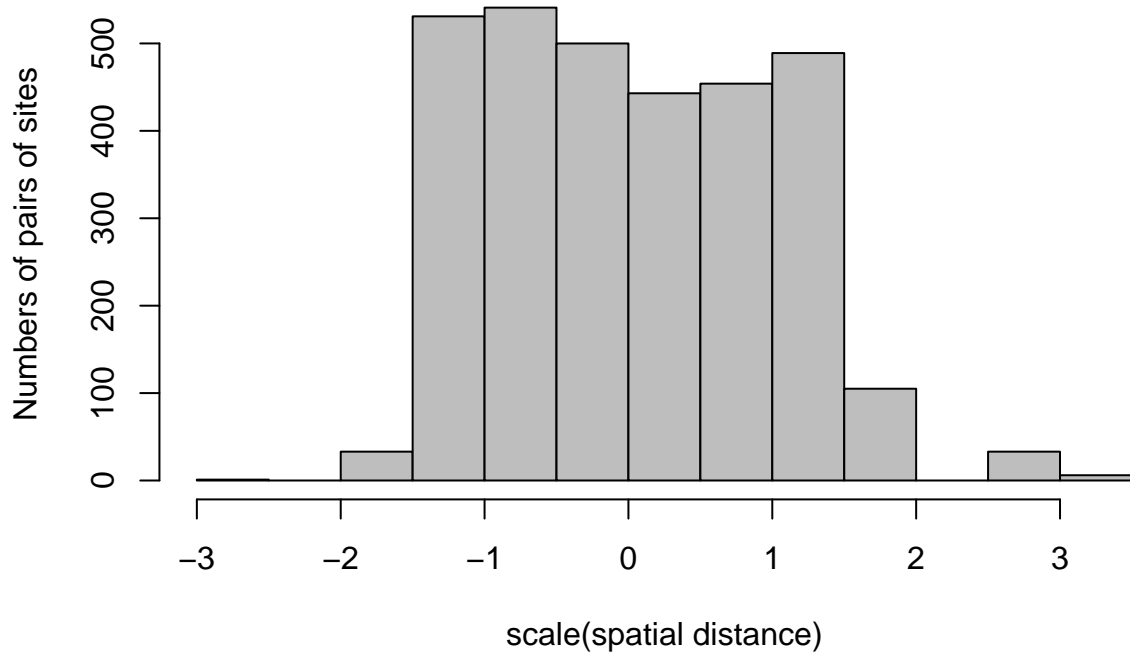
Figure 10: Histogram of the centered and scaled spatial distances between every pairwise combination of sites. Centering the data results in a mean of zero, while scaling transforms the data into standard deviation units.

Table 2: Coefficient estimates for the multiple logistic (presence-absence) and multiple ordinary least squares (relative abundance) regression models. "scale" indicates that these independent variables were centered and scaled prior to fitting the models.

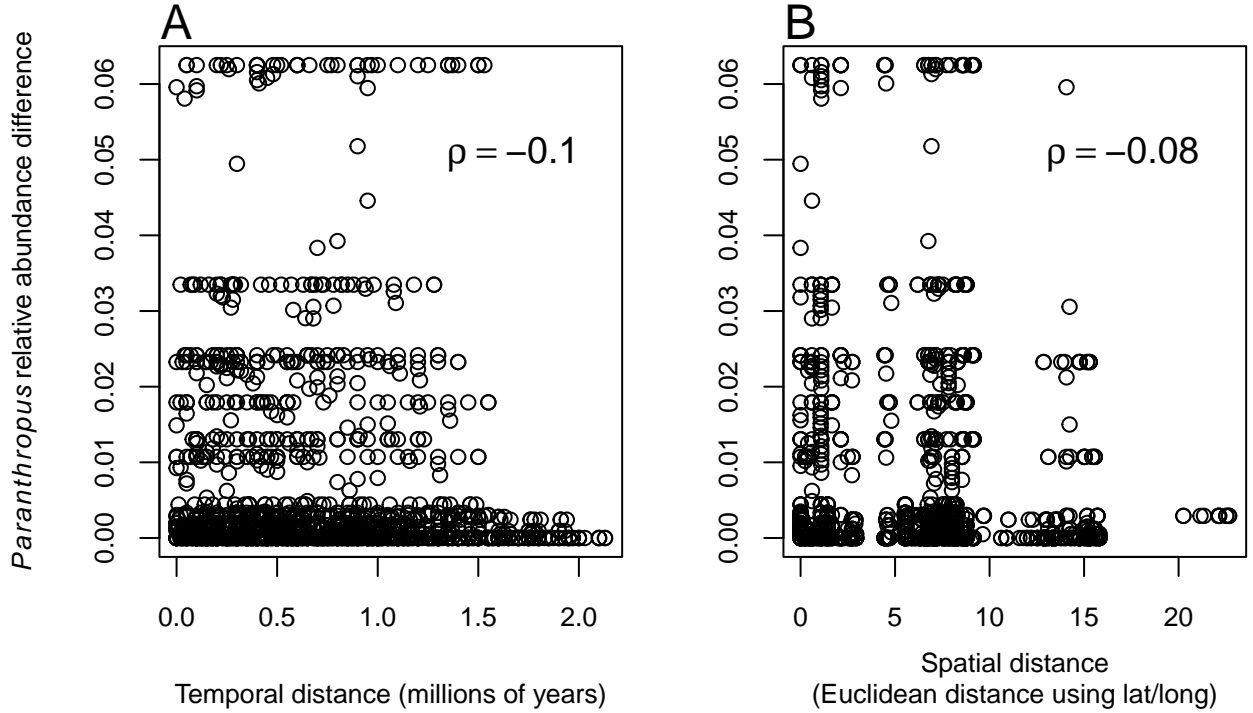|	Presence-absence logistic regression	Relative abundance OLS
Intercept	0.170	0.007
scale(temporal distance)	0.073	-0.001
scale(spatial distance)	-0.049	-0.001
Interaction term	0.067	0.000

Figure 11: Scatter plots of observed *Paranthropus* relative abundance differences as a function of (A) temporal distance and (B) spatial distance for all pairwise comparisons of sites. ρ is the Spearman's rank correlation coefficient.

It should be noted that our models exhibit negligible multicollinearity: the Spearman's rank correlation between temporal and spatial distance between pairs of sites is only $\rho = 0.08$.

4.2 Assumption #2: independence of specimen data

Assumption #2 would be violated if multiple specimens belonged to the same individual (e.g., a partial skeleton), and this was the case for a large proportion of specimens at each site. Non-independence would artificially inflate the number of independent specimens for each site, which would bias posterior probabilities

of absence upwards. We assessed independence of specimen data for those sites where data on the number of specimens per individual are available (i.e., databases). Results show that the majority of individuals at each site are comprised of one specimen (Figure 12). In fact, for each site, the proportion of individuals made up of one specimen exceeds 0.8, except for Member G of the Shungura Formation (0.68) (Figure 13). Because all sites in our dataset are taphonomically similar in a general sense (e.g., eastern African large mammalian fossil assemblages that were preserved in fluvio-lacustrine settings and predominantly surface collected), we assume that the dominance of single-specimen individuals is a general pattern that can be extrapolated to those sites not represented in Figures 12 and 13. Thus, given the low proportion of individuals with non-independent specimen data at each site, we view Assumption #2 as not grossly violated, and the estimated posterior probabilities of true absence are robust.
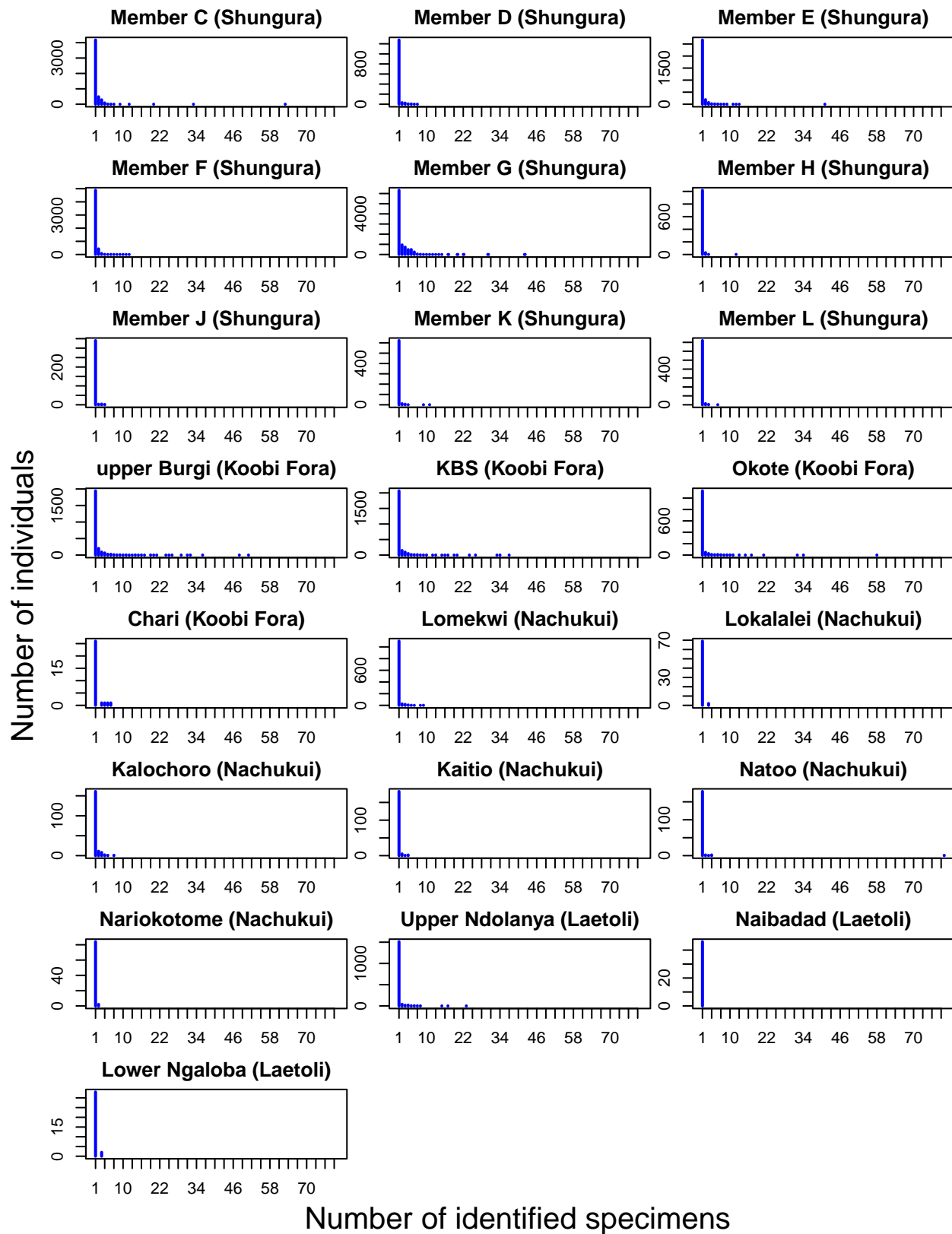
Figure 12: Frequency distributions depicting the number of individuals with a given number of specimens for each site, where those data are available (i.e., databases).
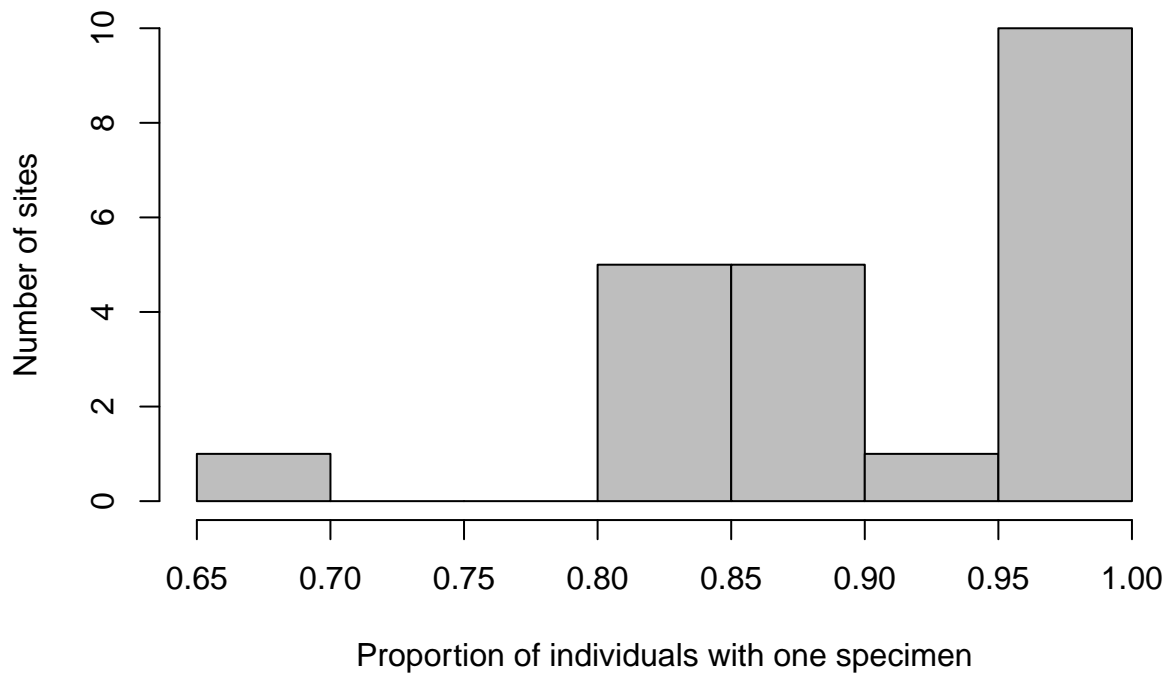
Figure 13: Histogram of the number of sites with a given proportion of individuals with one specimen. The proportion was calculated for each site as the number of individuals represented by one specimen divided by the total number of individuals.

References

Brown, J.H., Mehlman, D.W., Stevens, G.C., 1995. Spatial variation in abundance. Ecology. 76, 2028–2043.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: A practical information-theoretic approach, 2nd ed. Springer, New York.

Byrd, R.H., Lu, P., Nocedal, J., Zhu, C., 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 16, 1190–1208.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes methods for data analysis, 2nd ed. ed, Chapman & Hall/CRC texts in statistical science series. Chapman & Hall/CRC, Boca Raton.

Hobbs, N.T., Hooten, M.B., 2015. Bayesian Models: A Statistical Primer for Ecologists. Princeton University Press, Princeton.

Hurlbert, A.H., White, E.P., 2005. Disparity between range map- and survey-based analyses of species richness: Patterns, processes and implications. Ecology Letters. 8, 319–327.

Hurlbert, A.H., White, E.P., 2007. Ecological correlates of geographical range occupancy in North American birds. Global Ecology and Biogeography. 16, 764–773.

Murray, B.R., Rice, B.L., Keith, D.A., Myerscough, P.J., Howell, J., Floyd, A.G., Mills, K., Westoby, M., 1999. Species in the tail of rank-abundance curves. Ecology. 80, 1806–1816.

Rapoport, E.H., 1982. Areography: Geographical Strategies of Species. Pergamon Press, Oxford.

Swihart, B., Lindsey, J., 2020. Rmutil: Utilities for nonlinear regression and repeated measurements models.

Wang, S.C., 2010. Principles of statistical inference: Likelihood and the Bayesian paradigm. In: Alroy, J., Hunt, G. (Eds.), Quantitative Methods in Paleobiology, Paleontological Society Papers. Paleontological Society, New Haven, pp. 1–18.

Wang, S.C., Everson, P.J., Zhou, H.J., Park, D., Chudzicki, D.J., 2016. Adaptive credible intervals on stratigraphic ranges when recovery potential is unknown. Paleobiology. 42, 240–256.