MODEL SUPPLEMENT: A BAYESIAN MULTILEVEL MIXTURE MODEL FOR ZOOARCHAEOLOGICAL MEASUREMENTS

Jesse Wolfhagen[1]

[1]Department of Anthropology, Purdue University, West Lafayette, Indiana, USA

email: jl.wolfhagen@gmail.com

R Markdown version last compiled on 2023-03-02

Model Supplement: A Bayesian Multilevel Mixture Model for Zooarchaeological Measurements

The Bayesian model developed for this paper describes assemblages of faunal measurements as a mixture of immature animals, (adult-sized) females, and (adult-sized) males that have distinct average body sizes and expected variation around that average size. The model uses multiple measured dimensions (e.g., humerus distal breadth "humerus Bd," radius proximal breadth "radius Bp," abbreviations following Driesch 1976), which are first converted into a logarithmic size index, or LSI, values with a natural logarithm base (Meadow 1999; Wolfhagen 2020). LSI observations from measurement sets are grouped together within a specimen to create individuals grouped into defined "element portions" that serve as the basis for the mixture model analysis. Element portions relate to categories used for element fusion (e.g., distal humerus) to relate biometry and mortality profiles; specimens that contain multiple element portions—like complete limb bones—are grouped into the latest-fusing element portion (compare to "skeletal part type" in Breslawski 2022).

The multilevel structure of the model uses partial pooling to allow the parameters of the mixture model to vary between element portions while resisting overfitting. These element portion-specific parameters are related to each other through hyper-parameters, which describe the average value of the model parameters and the variability of model parameters across element portions (Wolfhagen 2020). The following sections describe the details of the multilevel mixture model, including the model's likelihood, the ways that the direct observations of measurements and demographic data are used by the model to account for measurement error, and the development of prior distributions for the model's hyper-parameters and for parameters that govern the model's multilevel structure. Finally, this supplement provides the full sets of equations for the single-site and multisite Bayesian multilevel mixture models and the prior distribution definitions used in the model applications described in the main text.

1. Definition of the Bayesian Multilevel Model

The central likelihood of the mixture model uses parameters that are specific to each element portion. These parameters include the relative proportions for the different animal groups: immature animals, females, and males (π_1, π_2, π_3), the average size for each group (μ_1, μ_2, μ_3), and the standard deviation for each group (σ_1, σ_2, σ_3). For each element portion, immature animals are described with the first set of parameters (π_1, μ_1, σ_1), adult-sized females with the second set of parameters (π_2, μ_2, σ_2), and adult-sized males with the third set of parameters (π_3, μ_3, σ_3). This results in both a set of parameters that describe the composition

of the assemblage (of measurements from that element portion) and an equation to estimate the probability that a particular specimen comes from an immature, adult female, or adult male individual.

Mixture Model Likelihood Equation:

$$\begin{aligned} P(x|\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3) = \\ \pi_1 * \text{Normal}(x, \mu_1, \sigma_1) + \\ \pi_2 * \text{Normal}(x, \mu_2, \sigma_2) + \\ \pi_3 * \text{Normal}(x, \mu_3, \sigma_3) \end{aligned} \tag{1}$$

In addition to a specimen's LSI value, the model needs two additional observed variables to address the potential presence of immature animals in the assemblage. First, an indicator variable Immature[specimen] describes whether the specimen could be from an immature animal based on the body part and the fusion characteristics (1 = potentially immature, 0 = cannot be immature). Data from known-age Shetland sheep show that specimens killed at younger than one year of age are significantly smaller than those killed at older ages, regardless of fusion status (Popkin et al. 2012). Thus, any measurement from an element with an unfused epiphysis or from an element that does not fuse or could fuse before one year of age is considered potentially immature. Measurements from specimens with fused epiphyses that fuse after one year of age are considered ineligible to be immature, so the model does not consider that probability (it considers $\pi_1 = 0$ for fitting that specimen).

Second, the proportion of specimens from an element portion that could be immature ($\text{proportion}_{\text{immature}}$) determines how to re-weight the mixture components (π_1, π_2, and π_3) for potentially-immature specimens from that element portion. The mixture components describe the entire assemblage for an element portion (a combination of potentially-immature and non-immature specimens), meaning that if $\pi_1 = 0.25$ we should expect 25% of the specimens to be from immature animals. If every specimen could be immature—say, for specimens from an early-fusing element—then the mixture components do not need to be re-weighted. If, however, only half of the specimens could be immature, then the mixture components must be re-weighted to ensure that the whole-assemblage proportions are correct. In such a case, we would expect half of the potentially-immature animals to be from immature animals if $\pi_1 = 0.25$ for the whole assemblage and we know that $\text{proportion}_{\text{immature}} = \left(\frac{0.25}{0.50} = 0.50\right)$; this same re-weighting would make it less likely that potentially-immature animals are from adult-sized female or adult-sized male animals. The code includes checks to ensure that π_1 can never exceed 1.00 after accounting for the proportion of immature specimens in cases where there are very few potentially-immature specimens and/or a relatively high expected proportion of immature animals in the assemblage.

2. Measurement Error and Observations

The model estimates measurement error for different observed quantities that are input into the model. Measurements on both the archaeological specimens and the standard values used to calculate LSI values are assumed to have a 1% measurement error (Breslawski and Byers 2015; Popkin et al. 2012: Figure 6). This 1% value comes from an evaluation of the Breslawski and Byers (2015) measurement data, where the average standard deviation of repeated measurements on bison radius proximal breadth measurements was 1.1% the average value of the measurement. This means that each measurement is given a standard deviation based on the observed value, which is used to estimate the "modeled" measurement value based on the observation. These modeled measurements are then used to calculate the LSI value for that observation ($\text{LSI}_{\text{Measurement}}$).

Because specimens can have multiple measured dimensions that are included in the mixture model on them (e.g., a distal humerus with both Bd and BT observations or a complete radius with Bp and Bd observations), the mixture model calculates specimen-specific LSI values ($\text{LSI}_{\text{Specimen}}$) that are related to the observed measurement-specific LSI values ($\text{LSI}_{\text{Measurement}}$). $\text{LSI}_{\text{Measurement}}$ values are the "observations" with a standard deviation of 0.01 (in LSI_e scale) based on intra-individual variation of LSI_e values for the Popkin et al. (2012) sheep using the *Ovis orientalis* female standard animal (FMC 57951) from Uerpmann and Uerpmann (1994, Table 12).

Observation Error Equations for Measurements:

$$
\begin{aligned}
\sigma_{\text{measurement}} &= \text{Measurement}_{\text{observed}} * 0.01 \\
\text{Measurement}_{\text{observed}} &\sim \text{Normal}(\text{Measurement}_{\text{modeled}}, \sigma_{\text{measurement}}) \\
\sigma_{\text{reference}} &= \text{Reference}_{\text{observed}} * 0.01 \\
\text{Reference}_{\text{observed}} &\sim \text{Normal}(\text{Reference}_{\text{modeled}}, \sigma_{\text{reference}}) \\
\text{LSI}_{\text{measurement}} &= \log_e(\text{Measurement}_{\text{modeled}}) - \log_e(\text{Reference}_{\text{modeled}}) \\
\text{LSI}_{\text{measurement}} &\sim \text{Normal}(\text{LSI}_{\text{specimen}}, 0.01)
\end{aligned}
\tag{2}
$$

The model also uses observations of sex ratios and fusion rates to estimate assemblage-level demographic proportions. This allows relevant data to inform the model about the expected relative proportions of different animal groups while still allowing these proportions to vary across different element portions. These observations are interpreted as binomial data: counts of some quantity (e.g., immature specimens) out of a total count of relevant specimens (e.g., total ageable specimens); this approach lets the model incorporate the uncertainty caused by small sample sizes. The observation of the average proportion of immature specimens

(μ_{π_1}) is based on the fusion rate of proximal and middle phalanges (the number of unfused phalanges $\mathrm{N_{Unfused}}$ out of the total number of phalanges with fusion data $\mathrm{N_{Ageable}}$), which fuse at around the same time as the estimated time that animals reach adult body size in the Shetland sheep population (Popkin et al. 2012). The observation of the average adult sex ratio—the proportion of females among mature animals $\left(\frac{\mu_{\pi_2}}{\mu_{\pi_2}+\mu_{\pi_3}}\right)$—is based on the sex ratio of fused pelvises (the number of female pelvises $\mathrm{N_{Female}}$ out of the total number of pelvises with a sex assignment $\mathrm{N_{Sexable}}$). In each case, the number of observable specimens determines the measurement error using the binomial distribution. While this paper uses these quantities to estimate the relevant hyper-parameters, relevant observations from other elements can be incorporated into the model in the same fashion if there is a clear sense of the total number of specimens that could have potentially been immature or female.

Observation Error Equations for Demographic Estimates:

$$\begin{aligned} N_{\text{unfused}} &\sim \text{Binomial}(N_{\text{ageable}}, \mu_{\pi_1}) \\ N_{\text{female}} &\sim \text{Binomial}(N_{\text{sexable}}, \frac{\mu_{\pi_2}}{\mu_{\pi_2} + \mu_{\pi_3}}) \end{aligned} \tag{3}$$

3. Prior Distributions

The prior distributions in this model are focused on describing previous beliefs about the value of the mixture model hyper-parameters, as the element portion-specific parameters are derived from these distributions. Mixture modeling performs well with unconstrained parameters values because it is more straightforward to estimate variation across element portions, meaning that constrained parameters-those where the range of possible values depends on the values of other parameters—must first be transformed into related unconstrained parameters (Betancourt 2017) . The following sub-sections describe the necessary transformations for different sets of the mixture model parameters, describing the unconstrained parameters that can be modeled and the transformations that result in the mixture model parameters. While these sections use the mixture model parameter notations, prior distributions are for the 'central tendency' hyper-parameter for the described unconstrained parameter.

It is important to remember that for all of these prior distributions are arbitrary choices made by the researcher, regardless of whether the distributions are based on specific animal populations or on reference priors. Other researchers could and should use different prior distributions to best reflect their intuition about likely parameter values for particular case studies. This also highlight the importance of reporting the prior distributions used in a Bayesian analysis to ensure replicability. Examining the implications of different prior distributions is an important step in the development of Bayesian models, one that should be

regularly tested even before models are fit to datasets (Gelman et al. 2020).

3.1 Mixture Proportion Priors

Prior distributions for the mixture proportions reflect our prior beliefs about the relative proportions of the three animal groups in the assemblage (immature, adult-sized females, and adult-sized males). The three mixture proportions (π_1, π_2, π_3) are a three-value unit simplex, meaning that the values are constrained as a group to sum up to one. Thus, the simplex can be described by only two variables because the third value cannot vary once those two values are known. The model uses two unconstrained variables (θ_1 and θ_2) to describe the π values. These θ values are related back to π values using a 'stick-breaking' transformation that iteratively estimates the relative proportions of the simplex taken up by each θ value (Team 2022: Section 10.7).

Stick-Breaking Transformations:

$$\begin{aligned} \pi_1 &= \text{logit}^{-1}(\theta_1 + \log(0.5)) \\ \pi_2 &= (1 - \pi_1) * \text{logit}^{-1}(\theta_2 + \log(1)) \\ \pi_3 &= 1 - (\pi_1 + \pi_2) \end{aligned} \tag{4}$$

The θ_1 value can be directly related to the π_1 value using the first line of the stick-breaking transformation, meaning that one can examine the associated π_1 estimate for a given θ_1 value. Within the stick-breaking transformation, θ_2 relates to the relative proportions of π_2 and π_3 after π_1 has been estimated, which is effectively the adult sex ratio. Just as we could examine the expected π_1 estimates from a distribution of θ_1 values, we can thus use expected adult sex ratios $\left(\frac{\pi_2}{\pi_2+\pi_3}\right)$ estimates from a particular prior distribution for θ_2. Relating these θ values back to observable phenomena makes it easier to define reasonable prior distribution definitions for the parameters from domain expertise (see Section 2.2 and Figure 2 of the main text).

3.2 Average Body Size and Size Variability Priors

While the average body sizes of the different components (μ_1, μ_2, μ_3) are not intrinsically linked in the same way that π values are, the model still requires some structure to aid interpretability. Bayesian mixture models that are fit using Markov Chain Monte Carlo (MCMC) methods, like the model in this paper, can suffer from an issue called "label switching" if μ values are not ordered in some way (Jasra, Holmes, and Stephens 2005). MCMC methods rely on running multiple "chains"—separate iterations of the model that are independently fit and then combined together—to show that the results are independent of the initial

conditions. Label switching describes a scenario where different chains fit the data well, but the parameter labels relate to different specimens (e.g., smaller specimens are assigned to μ_1 in one chain and to μ_2 in another). To avoid label switching, the average body sizes are strictly ordered, meaning that $\mu_1 < \mu_2 < \mu_3$ must be maintained. Note that this only affects average values, individual immature specimens can still be larger than female specimens or male specimens and individual female specimens can be larger than male specimens. This ordering is achieved by only estimating μ_2 directly (average LSI_e value for females) and estimating the average LSI_e value for immature and male animals with offsets (δ_1, δ_2) from the female average. The δ values must be positive to maintain the ordering of the μ values, so each δ is modeled in a log-transformed space.

Offset Equations for μ Values:

$$\begin{aligned} \mu_1 &= \mu_2 - \delta_1 \\ \mu_3 &= \mu_2 + \delta_2 \end{aligned} \tag{5}$$

Conceptually, this expression of animal body size defines female animals as the standard "body size" that is subject to various selective pressures, with the offset for male animals reflecting sex-specific pressures on males. This interpretation of body size broadly fits the general pressures affecting adult body size in females and males across many ungulate taxa, including domestic herd animals (Pérez-Barbería, Gordon, and Pagel 2002; Tchernov and Horwitz 1991). The body size offset between immature animals and adult-sized females (δ_1) is admittedly an *ad hoc* definition rather than one under strict biological constraints, as it can be affected by the age immature animals reach before being killed (Gillis et al. 2014). The computational advantages of this definition arguably outweigh the awkwardness of the definition, however. Further, evaluation of δ_1 and δ_2 values across different sites could conceivably highlight variation in the timing of the killing of immature animals (δ_1) and the degree of adult sexual dimorphism (δ_2); both variables can be related to models of hunting intensity, animal domestication, and herd management (Gillis et al. 2014; Marom and Bar-Oz 2013; Zeder and Hesse 2000).

The LSI_e size variability of animals within a group (σ_1, σ_2, σ_3) is a key variable in the Bayesian mixture model. The values of these standard deviation parameters play a major role in ensuring that the mixture components reflect biological entities rather than overfitting to specific sample noise. Coefficients of variance (CVs) for raw mammal bone measurements from a single sex have been found to be relatively consistent (Davis 1996; Popkin et al. 2012). When transforming these measurements using a logarithm, this produces consistent standard deviations of the transformed measurement values (Wolfhagen 2020: Figure 1), suggesting that σ values should be relatively stable across elements. While σ parameters from different

animal groups within the model are not directly related to each other, the values still need some transformations to be modeled consistently by the multilevel model. These values must be positive, which conflicts with the multilevel model's need for unconstrained variables. To achieve this, the model uses the same log-transformation technique used for size offsets to create an unconstrained parameter, $\log_e \sigma$, that is then transformed to actual σ values after estimating variation across element portions.

3.3 Developing Priors from a Simulated Prior Assemblage

Relevant prior estimates of the biologically-relevant parameters were derived from the biometry of 356 known-age and known-sex Shetland sheep described in Popkin et al. (2012) (see Section 2.2). Castrated individuals were included as males and animals killed under one year of age were considered immature; this resulted in an assemblage of 48 immature animals, 164 females, and 144 males/castrates. The element portion definitions and included measurements are shown Table 4 of the main text. From the 2848 element portions in the full assemblage, 150 immature, female, and male element portions were randomly selected to create an assemblage of 450 element portions for analysis. LSI_e values are calculated using the *Ovis orientalis* female standard animal (FMC 57951) from Uerpmann and Uerpmann (1994: Table 12). These specimens were modeled using a Bayesian multilevel mixture model that used their known identities to estimate the biologically relevant parameters directly (μ_2, δ_1, δ_2, σ_1, σ_2, σ_3). The resulting hyper-parameters, which average across anatomical variation in the parameter values, are used as a baseline for defining prior distributions of the parameter in the archaeological model (see Figure 3 in the main text).

The prior distributions used in this Bayesian multilevel mixture model on the reference population are more straightforward. While the same transformations to create unconstrained parameters are necessary (e.g., modeling average size as μ_2 with offsets for immature and male animals), the definition of these prior distributions can be broadly described as weakly-informative priors (Gelman et al. 2008). These weakly-informative prior distributions are reasonable in this case—and not in the archaeological case—because all the mixture model parameters have direct observations rather than relying on latent state estimations. That is, parameters like the size difference between males and females (δ_2) and the size variability in female animals (σ_2) can be directly observed because the group identities of every specimen are known. With these direct observations, the prior distributions have a more muted influence on the resulting posterior distributions. This does not mean that the prior distributions have no effect, however, which is why objective priors can have undesirable impacts on modeling results (Gabry et al. 2019).

Prior Distribution Definitions for the Known-Identity Model Hyper-Parameters

$$
\begin{aligned}
\mu_2 &\sim \text{Normal}(-0.1, 0.1) \\
\log \delta_1 &\sim \text{Normal}(-3, 0.5) \\
\log \delta_2 &\sim \text{Normal}(-3, 0.5) \\
\log \sigma_1 &\sim \text{Normal}(-3, 0.1) \\
\log \sigma_2 &\sim \text{Normal}(-3, 0.1) \\
\log \sigma_3 &\sim \text{Normal}(-3, 0.1) \\
\sigma_{\text{Element}}[1, 2, 3, 4, 5, 6] &\sim \text{Half-Normal}(0, 0.05)
\end{aligned}
\tag{6}
$$

The average LSI_e value for females (μ_2) is likely to vary across contexts in reaction to different selective pressures, both anthropogenic and ecological (e.g., Davis 1981; Manning et al. 2015; Wright and Viner-Daniels 2015). While the posterior distribution is extremely focused on this specific population, there is no reason to think that this value should be centered at any particular value since that relates to the standard animal used (Meadow 1999; Wolfhagen 2020). Therefore, the prior distribution used in archaeological models for μ_2 uses a larger standard deviation, $\mu_2 \sim \text{Normal}(0, 0.1)$, to encompass likely LSI_e values (Figure 3A in the main text). Under this definition, there is a 95% probability that the μ_2 value lies within the range of -0.20 and 0.20 on the LSI_e scale, which translates to roughly 82-122% the size of the standard animal's measurement.

For the average size difference between immature and female animals (δ_1), the narrowness of the posterior distribution likely reflects the fact that immature animals in the sample cover a narrow age range. Animals killed under one year of age span only 36 days and the youngest animals are nearly half a year old (178-214 days: Popkin et al. 2012). Thus while the posterior results provide a useful starting point for estimating this offset, there is a good potential for larger δ_1 values (i.e., greater size differences between immature and adult-sized female animals) in other contexts that could include animals killed at a younger age (Figure 3B in the main text). To capture this possibility, the archaeological model uses a prior distribution with a larger standard deviation and a slightly higher center, $\log \delta_1 \sim \text{Normal}(-3.5, 0.4)$, which results in an average size difference of 0.03 on the LSI_e scale and a 95% probability that the size difference is between 0.01 and 0.07. This translates into expecting the average body size of immature animals in an assemblage being 3% smaller than the average body size of adult-sized female animals, but also plausibly believing that this size difference could range from 1-7% smaller.

The average size difference between adult males and females (δ_2), also known as the index of sexual

dimorphism (Fernández and Monchot 2007), is likely to be under stricter biological control than the other "average body size" parameters in the model. This does not mean that this difference could not vary between contexts, however. Some models of animal domestication argue that initial domestication removed sexual selective pressures on male body size, reducing sexual dimorphism (e.g., Tchernov and Horwitz 1991). In a similar fashion, specialized hunting strategies could also reduce sexual dimorphism by targeting large-bodied males, for example (Zeder 2012; Proaktor, Coulson, and Milner-Gulland 2007; Milner, Nilsen, and Andreassen 2007). Again, the posterior distribution of the extent of sexual dimorphism in the Shetland sheep population provides a useful starting point to describe a prior distribution for the model (Figure 3C in the main text). Increasing the standard deviation of the distribution slightly, $\log \delta_2 \sim \text{Normal}(-2.7, 0.1)$, produces a distribution centered at 0.07 LSI_e units with a 95% probability that the value is between 0.06-0.08, translating to the average male being 6-9% larger than the average female relative to a standard measurement. The smaller standard deviation in the prior distribution of δ_2 than for δ_1 reflects our understanding that the extent of sexual dimorphism, as a biological phenomenon, is less likely to have extreme values than the average size difference between immature and female animals, since δ_2 is unaffected by the specific age structure of the assemblage.

As in the average body size parameters, prior distributions for the size variability model parameters are developed from the Bayesian model of known-identity Shetland sheep measurements. The resulting σ hyper-parameters provide a baseline for establishing hyper-parameter prior distributions in archaeological cases. Figure 3D-F of the main text shows the posterior distributions of these σ hyper-parameters in both the log-transformed values and associated LSI_e values. Average size variability within an element portion for immature animals (σ_1) is higher, on average, than for females (σ_2) and males (σ_3). The immature category includes both male and female animals, so larger size variability makes sense; again, it is possible that σ_1 is relatively low in this population relative to other contexts given the narrow age range of immature animals in the Shetland sheep population. Unlike the average body size parameters, there are not compelling reasons to believe that size variability parameters for females and males (σ_2 and σ_3) should vary widely in different contexts given the consistency of coefficients of variation in mammals broadly (Davis 1996). Thus, the results of this analysis are used for the prior distributions of $\log \sigma_2$ and $\log \sigma_3$, while the prior distribution of $\log \sigma_1$ is given an increased standard deviation and slightly increased average value. Overall, however, these prior distributions suggest that the average size variability within an element portion is between 0.04-0.06 for females and males and is between 0.04-0.05 for immature animals. Note that even though σ_2 and σ_3 have the same prior distributions, these values can still vary from each other in different contexts.

4. Multilevel Structure of the Model

The previous section described prior distributions that describe the *average* value for different mixture model parameters across all element portions. To create parameter estimates that are specific to different element portions, it is necessary to estimate the *variation* around these average values that different parameters can have among different element portions. The model uses a Multivariate Normal definition of the model parameters to allow for correlations between different parameters; effectively, the possibility that multiple model parameters will covary from element portion to element portion. To do this, each hyper-parameter has an associated σ_{element} parameter that describes inter-element variation in parameter values. The model uses a non-centered parameterization, wherein the Multivariate Normal distribution is centered at zero to calculate offsets, ν_{element}, that are added to the average hyper-parameters to calculate model parameters for each element portion. This definition provides computational stability and makes it more straightforward to incorporate other levels of multilevel structure.

Equations for Defining Inter-Element Variation (Multilevel Modeling):

$$
\begin{aligned}
\nu_{\text{Element}} &\sim \text{MultivariateNormal}\left(\begin{bmatrix}0\\ \vdots \\ 0\end{bmatrix}, \Sigma_{\text{Element}}\right)\\
\Sigma_{\text{Element}} &= \begin{pmatrix}\sigma_{\text{Element}}[1] & \dots & 0\\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{\text{Element}}[8]\end{pmatrix} \rho_{\text{Element}} \begin{pmatrix}\sigma_{\text{Element}}[1] & \dots & 0\\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{\text{Element}}[8]\end{pmatrix}\\
\rho_{\text{Element}} &= LKJcorr(2)\\
\theta_1[\text{Element}] &= \theta_1 + \nu_{\text{Element}}[1]\\
\theta_2[\text{Element}] &= \theta_2 + \nu_{\text{Element}}[2]\\
\mu_2[\text{Element}] &= \mu_2 + \nu_{\text{Element}}[3]\\
\log \delta_1[\text{Element}] &= \log \delta_1 + \nu_{\text{Element}}[4]\\
\log \delta_2[\text{Element}] &= \log \delta_2 + \nu_{\text{Element}}[5]\\
\log \sigma_1[\text{Element}] &= \log \sigma_1 + \nu_{\text{Element}}[6]\\
\log \sigma_2[\text{Element}] &= \log \sigma_2 + \nu_{\text{Element}}[7]\\
\log \sigma_3[\text{Element}] &= \log \sigma_3 + \nu_{\text{Element}}[8]
\end{aligned}
\tag{7}
$$

The multilevel structure used to allow variation in parameter estimates across element portions can

also be expanded to create multisite models that can directly compare sex-specific biometric estimates alongside the age/sex composition of different assemblages. Such comparisons can highlight variation in herd management strategies or diachronic body size trends related to population turnover (e.g., Arbuckle and Atici 2013; Arbuckle et al. 2016). To do this, an additional multilevel structure can be applied to the same mixture model parameters, using σ_{Site} rather than σ_{Element} parameters. However, an additional set of multilevel structure parameters, $\sigma_{Interaction}$, are also necessary to ensure that elemental variation is different at different sites (e.g., the difference between μ_2 for the distal humerus and μ_2 for the distal radius is not necessarily the same at different sites). Again, weakly-informative priors are appropriate for both sets of parameters. Each additional term is included in the sum to create specific mixture model parameter values.

Example of Parameter Definition for Inter-Site and Inter-Element Variation:

$$\theta_1[\text{Site}, \text{Element}] = \theta_1 + \nu_{\text{Site}}[\text{Site}] + \nu_{\text{Element}}[\text{Element}] + \nu_{\text{Interaction}}[\text{Site}, \text{Element}]$$

The inclusion of multiple sites changes the definition of the 'grand mean' variable (θ_1 in the example equation) from a site-level estimate to an overall mean across the sites and elements. These parameter estimates thus describe the average composition of the entire set of assemblages. The details of the assemblages included in the analyses would affect how useful these estimates are for interpretation. Assemblage-specific estimates can be calculated for each model parameter by adding the relevant ν_{site} estimate to the 'grand mean' parameter, which would again act to describe the average composition of the assemblage regardless of its elemental composition.

Prior distributions for σ_{Element} values (and σ_{Site} and $\sigma_{\text{Interaction}}$ values in multisite models) are weakly-informative priors based on the scale of the parameter and the expectation for variation for the parameter. For example, there is likely more variation in θ parameters—that govern the relative composition of immature, female, and male animals—among element portions than variation in σ parameters that govern size variability within each group. Similarly, it is expected that average body sizes of females μ_2 will vary more between sites $\sigma_{\text{Site}}[3]$ than between elements within a site $\sigma_{\text{Element}}[3]$. The impacts of these prior distribution definitions were evaluated using prior predictive checking, suggesting that these prior distributions allow enough variability to encompass reasonable size estimates without providing too much prior support to implausible or impossible values (see Section 6 of the Model Supplement).

Prior Distributions for Element-level Variation (Multilevel Component):

$$\begin{aligned} \sigma_{\text{Element}}[1, 2] &\sim \text{Half-Normal}(0, 0.5) \\ \sigma_{\text{Element}}[3, 4, 5, 6, 7, 8] &\sim \text{Half-Normal}(0, 0.05) \end{aligned} \tag{8}$$

Prior Distributions for Site-level Variation (Multilevel Component):

$$\begin{aligned}
\sigma_{\text{Site}}[1,2] &\sim \text{Half-Normal}(0, 0.5) \\
\sigma_{\text{Site}}[3,4,5] &\sim \text{Half-Normal}(0, 0.1) \\
\sigma_{\text{Site}}[6,7,8] &\sim \text{Half-Normal}(0, 0.05) \\
\rho_{\text{Site}} &= LKJcorr(2)
\end{aligned} \tag{9}$$

Prior Distributions for Interaction Effect (Multilevel Component):

$$\begin{aligned}
\sigma_{\text{Interaction}}[1,2] &\sim \text{Half-Normal}(0, 0.25) \\
\sigma_{\text{Interaction}}[3,4,5] &\sim \text{Half-Normal}(0, 0.1) \\
\sigma_{\text{Interaction}}[6,7,8] &\sim \text{Half-Normal}(0, 0.05) \\
\rho_{\text{Interaction}} &= LKJcorr(2)
\end{aligned} \tag{10}$$

5. Prior Distributions for the Model Hyper-Parameters (Simulations and Archaeological Cases)

Prior Distribution Definitions for the Single Assemblage Simulation Model Hyper-Parameters

$$\begin{aligned}
\theta_1 &\sim \text{Normal}(-0.5, 1.5) \\
\theta_2 &\sim \text{Normal}(0.0, 1.5) \\
\mu_2 &\sim \text{Normal}(0.0, 0.1) \\
\log \delta_1 &\sim \text{Normal}(-3.5, 0.4) \\
\log \delta_2 &\sim \text{Normal}(-2.7, 0.1) \\
\log \sigma_1 &\sim \text{Normal}(-3.05, 0.1) \\
\log \sigma_2 &\sim \text{Normal}(-3.10, 0.1) \\
\log \sigma_3 &\sim \text{Normal}(-3.10, 0.1)
\end{aligned} \tag{11}$$

Prior Distribution Definitions for the Multisite Simulation Model Hyper-Parameters

$$
\begin{aligned}
\theta_1 &\sim \text{Normal}(-0.5, 1.5) \\
\theta_2 &\sim \text{Normal}(0.0, 1.5) \\
\mu_2 &\sim \text{Normal}(0.0, 0.2) \\
\log \delta_1 &\sim \text{Normal}(-3.5, 0.5) \\
\log \delta_2 &\sim \text{Normal}(-2.7, 0.2) \\
\log \sigma_1 &\sim \text{Normal}(-3.05, 0.1) \\
\log \sigma_2 &\sim \text{Normal}(-3.10, 0.1) \\
\log \sigma_3 &\sim \text{Normal}(-3.10, 0.1)
\end{aligned}
\tag{12}
$$

Prior Distribution Definitions for the Pinarbaşı B Sheep Model Hyper-Parameters

$$
\begin{aligned}
\theta_1 &\sim \text{Normal}(-0.5, 1.5) \\
\theta_2 &\sim \text{Normal}(0.0, 1.5) \\
\mu_2 &\sim \text{Normal}(0.0, 0.1) \\
\log \delta_1 &\sim \text{Normal}(-3.5, 0.4) \\
\log \delta_2 &\sim \text{Normal}(-2.7, 0.1) \\
\log \sigma_1 &\sim \text{Normal}(-3.05, 0.1) \\
\log \sigma_2 &\sim \text{Normal}(-3.10, 0.1) \\
\log \sigma_3 &\sim \text{Normal}(-3.10, 0.1)
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
\theta_1 &\sim \text{Normal}(-0.5, 1.5) \\
\theta_2 &\sim \text{Normal}(0.0, 1.5) \\
\mu_2 &\sim \text{Normal}(-0.1, 0.1) \\
\log \delta_1 &\sim \text{Normal}(-3.5, 0.5) \\
\log \delta_2 &\sim \text{Normal}(-2.0, 0.5) \\
\log \sigma_1 &\sim \text{Normal}(-3.05, 0.1) \\
\log \sigma_2 &\sim \text{Normal}(-3.10, 0.1) \\
\log \sigma_3 &\sim \text{Normal}(-3.10, 0.1)
\end{aligned}
\tag{14}
$$

6. Simulating Assemblages from the Prior Distributions (Prior Predictive Checking)

Prior predictive checks are a critical component of Bayesian model development workflows, ensuring that reasonable prior definitions are chosen (Gabry et al. 2019; Gelman et al. 2020). This process uses the model's prior distribution definitions to simulate data, which can then be evaluated against domain knowledge and observed data. This is particularly important when dealing with model parameters that are difficult to examine in isolation, like parameters that govern inter-element variation (σ_{Element}) in a multilevel model structure. Prior predictive checking is an iterative process, informing researchers about the potential consequences of their prior distribution definitions; in particular, it can highlight how excessively imprecise definitions can provide considerable prior weight on implausible and even impossible values for data (Gabry et al. 2019: Figure 4). Thus, prior predictive checking allows researchers to create more accurate summaries of their domain knowledge but also produces more efficient MCMC performance because less time is spent evaluating parameter values that are inconsistent with even cursory prior knowledge about the problem being modeled.

Prior predictive checks were developed for both the single-assemblage and multisite model fits, using the prior distribution definitions used in the sheep simulations in the main text (see Section 5 of the Model Supplement). In each simulation, 25 $\text{LSI}_{\text{Specimen}}$ values were calculated for each of the 5 element portions based on the relevant prior distributions and model structures; for the multisite simulation, this was done for 3 assemblages. To evaluate the feasibility of these prior distribution definitions, the $\text{LSI}_{\text{Specimen}}$ were converted into simulated measurement values based on the reference values of the *Ovis orientalis* female standard animal (FMC 57951) from Uerpmann and Uerpmann (1994: Table 12). Table 1 shows how the

Table 1: Values used to convert simulated LSI values into measurements in the prior predictive simulations. Dimension definitions follow von den Driesch (1976). Reference value refers to the female standard mouflon FMC 57951 (Uerpmann and Uerpmann 1994: Table 12).

Element Portion	Measured Dimension	Reference Value (mm)
Humerus	Bd	33.0
Humerus	BT	29.5
Radius	Bp	33.5
Metacarpus	Bp	25.0
Metatarsus	Bp	22.5
Astragalus	Bd	19.6

five element portions were converted into measurements from specific dimensions, following the equations in Section 2 of this model supplement. Each simulation was run 1000 times, producing 1000 assemblages of relevant simulated measurement values.

Figure 1 shows two results for the single assemblage (left) and multisite (right) prior prediction simulations. The top row shows the simulated proportions of immature, female, and male specimens in the simulated assemblages (Figure 1A). These proportions were sampled directly from the element-specific mixture proportion variables, showing the range of potential distributions the model is expecting before seeing any data. The 95% quantiles of the single assemblage model's proportion of immature specimens ranges from 2-84%, for female specimens, and for male specimens. It is noteworthy that the simulated proportion of immature specimens for the single assemblage model does not reach the extreme value seen in the Pınarbaşı B sheep data, which may explain some of the long tails in the element-specific compositional estimates (Figure 9 of the main text). The multisite model has similar expected ranges (immature: 1-83%, female: , male:).

The bottom row shows the distribution of simulated humerus Bd measurements for the assemblages (Figure 1B). These plots include all simulated specimens, showing the range of sheep humerus Bd values the model expects before seeing any data. Several vertical lines on the plots give a sense of domain knowledge about sheep humerus Bd values. First, the standard value (33.0 mm) from the *Ovis orientalis* female standard animal (FMC 57951) from Uerpmann and Uerpmann (1994: Table 12) is shown in the red dashed line. Second, vertical blue lines show ranges of observed archaeological sheep humerus Bd values: the largest sheep (technically *Ovis orientalis*) humerus Bd (39.0 mm) from the 10th millennium BP site of Körtik Tepe, in southeastern Anatolia (Arbuckle and Özkaya 2006: Table b) and the smallest sheep (*Ovis aries*) humerus Bd (25.0 mm) from the fifth-sixth century CE site of West Stow, United Kingdom (Crabtree 1990: Table 29). These plots highlight how the multisite model's structure allows for much more variation in body

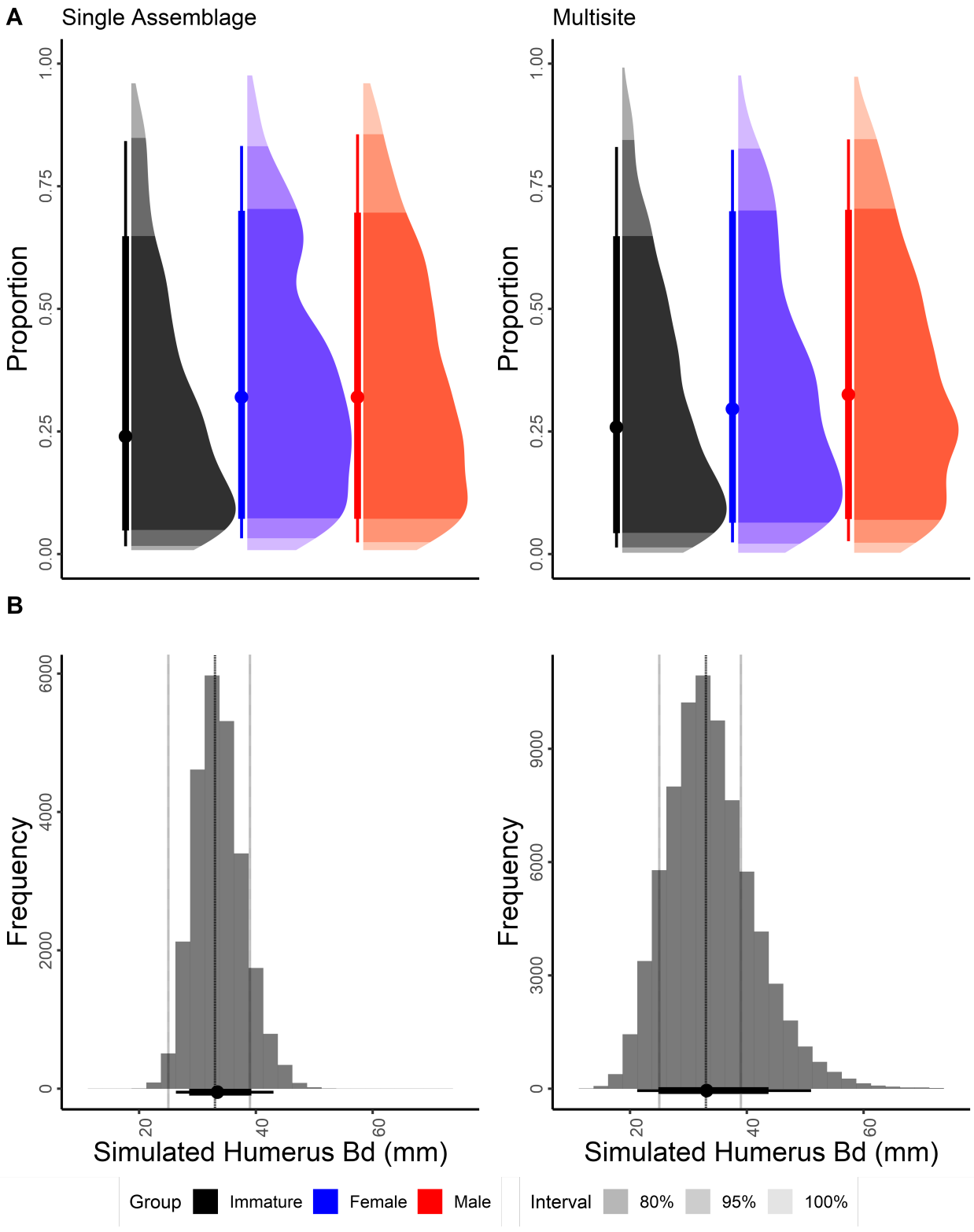

Figure 1: Single Assemblage and Multisite Prior Predictive Checks. Top row (A): Estimates of the proportion of immature, female, and male specimens. Bottom row (B): Histograms of simulated humerus Bd measurements from the simulations. Vertical lines show the value of the standard reference value (dashed) and two extreme archaeological samples (dotted).

size, especially large measurements. While the 95% quantiles of the single assemblage model's simulated measurements (26-43 mm) do not exactly encapsulate the range of the observed extreme measurements, the multisite model's simulated measurements go well beyond the range (21-51 mm). While one may not expect this full range of measurements in a single assemblage, the results of the prior predictive checks show that the multisite model could plausibly encapsulate variation in body size across diverse assemblages, though possibly at the cost of being somewhat inefficient (i.e., evaluating parameter values that are somewhat beyond reasonable expectations).

These prior predictive checks show that the Bayesian multilevel mixture model's structure is robust for diverse archaeological applications. The chosen prior distribution definitions for the multilevel variation components of the model encapsulate a reasonable range of expected variation and thus do not require extensive retooling as they are applied to new archaeological situations or taxa. Moreover, the multilevel model appears to have enough variability to model drastic changes in body size, making the models' structure relevant for examining broad spatial and temporal variation in animal biometry. While experts are encouraged to use domain knowledge to update and adjust the models to better fit the questions they ask, these models are widely-applicable tools that are suitable for asking many questions about animal body size and the composition of zooarchaeological assemblages.

References

Arbuckle, Benjamin S., and Levent Atici. 2013. "Initial Diversity in Sheep and Goat Management in Neolithic South-Western Asia." Journal Article. *Levant* 45 (2): 219–35.

Arbuckle, Benjamin S., and V. Özkaya. 2006. "Animal Exploitation at Körtik Tepe: An Early Aceramic Neolithic Site in Southeastern Turkey." Journal Article. *Paléorient* 32 (2): 113–36.

Arbuckle, Benjamin S., Max D. Price, Hitomi Hongo, and Banu Öksüz. 2016. "Documenting the Initial Appearance of Domestic Cattle in the Eastern Fertile Crescent (Northern Iraq and Western Iran)." Journal Article. *Journal of Archaeological Science* 72: 1–9. https://doi.org/10.1016/j.jas.2016.05.008.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." Journal Article. *arXiv* 1701.02434 [Preprint]: Available from https://arxiv.org/abs/1701.02434v1.

Breslawski, Ryan P. 2022. "Minimum Animal Units and the Standardized Count Problem." Journal Article. *Journal of Archaeological Method and Theory*. https://doi.org/10.1007/s10816-022-09563-9.

Breslawski, Ryan P., and David A. Byers. 2015. "Assessing Measurement Error in Paleozoological Osteometrics with Bison Remains." Journal Article. *Journal of Archaeological Science* 53: 235–42. https://doi.org/10.1016/j.jas.2014.10.001.

Crabtree, Pam J. 1990. *West Stow: Early Anglo-Saxon Animal Husbandry.* Book. East Anglian Archaeology. Suffolk: Suffolk County Planning Dept.

Davis, Simon J. M. 1981. "The Effects of Temperature Change and Domestication on the Body Size of Late Pleistocene to Holocene Mammals of Israel." Journal Article. *Paleobiology* 7 (1): 101–14.

———. 1996. "Measurements of a Group of Adult Female Shetland Sheep Skeletons from a Single Flock: A Baseline for Zooarchaeologists." Journal Article. *Journal of Archaeological Science* 23: 593–612.

Driesch, Angela von den. 1976. *A Guide to the Measurement of Animal Bones from Archaeological Sites.* Book. Peabody Museum Bulletins. Cambridge, MA: Harvard University.

Fernández, Héléna, and Hervé Monchot. 2007. "Sexual Dimorphism in Limb Bones of Ibex (Capra Ibex l.): Mixture Analysis Applied to Modern and Fossil Data." Journal Article. *International Journal of Osteoarchaeology* 17: 479–91. https://doi.org/10.1002/oa.876.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." Journal Article. *Journal of the Royal Statistical Society, Series A* 182 (2): 389–402.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." Journal Article. *The Annals of Applied Statistics* 2 (4): 1360–83. https://doi.org/10.1214/08-aoas191.

Gelman, Andrew, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. "Bayesian Workflow." Journal Article. *ArXiv*.

Gillis, R., I. Carrère, M. Saña Seguí, G. Radi, and J. D. Vigne. 2014. "Neonatal Mortality, Young Calf Slaughter and Milk Production During the Early Neolithic of North Western Mediterranean." Journal Article. *International Journal of Osteoarchaeology* 26 (2): 303–13. https://doi.org/10.1002/oa.2422.

Jasra, A., C. C. Holmes, and D. A. Stephens. 2005. "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling." Journal Article. *Statistical Science* 20 (1): 50–67. https://doi.org/10.1214/088342305000000016.

Manning, Katie, Adrian Timpson, Stephen Shennan, and Enrico Crema. 2015. "Size Reduction in Early European Domestic Cattle Relates to Intensification of Neolithic Herding Strategies." Journal Article. *PLoS One* 10 (12): e0141873. https://doi.org/10.1371/journal.pone.0141873.

Marom, Nimrod, and Guy Bar-Oz. 2013. "The Prey Pathway: A Regional History of Cattle (Bos Taurus) and Pig (Sus Scrofa) Domestication in the Northern Jordan Valley, Israel." Journal Article. *PLoS One* 8 (2): e55958. https://doi.org/10.1371/journal.pone.0055958.g001.

Meadow, Richard. 1999. "The Use of Size Index Scaling Techniques for Research on Archaeozoological

Collections from the Middle East." Book Section. In *Historia Animalium Ex Ossibus: Beiträge Zur Paläoanatomie, Archäologie, Ägyptologie, Ethnologie Und Geschichte Der Tiermedizin*, edited by Cornelia Becker, Henriette Manhart, Joris Peters, and Jörg Schibler, 285–300. Leidorf: Verlag Marie.

Milner, J. M., E. B. Nilsen, and H. P. Andreassen. 2007. "Demographic Side Effects of Selective Hunting in Ungulates and Carnivores." Journal Article. *Conserv Biol* 21 (1): 36–47. https://doi.org/10.1111/j.1523-1739.2006.00591.x.

Pérez-Barbería, F. Javier, Iain J. Gordon, and M. Pagel. 2002. "The Origins of Sexual Dimorphism in Body Size in Ungulates." Journal Article. *Evolution* 56 (6): 1276–85.

Popkin, Peter R. W., Polydora Baker, Fay Worley, Sebastian Payne, and Andy Hammon. 2012. "The Sheep Project (1): Determining Skeletal Growth, Timing of Epiphyseal Fusion and Morphometric Variation in Unimproved Shetland Sheep of Known Age, Sex, Castration Status and Nutrition." Journal Article. *Journal of Archaeological Science* 39 (6): 1775–92. https://doi.org/10.1016/j.jas.2012.01.018.

Proaktor, G., T. Coulson, and E. J. Milner-Gulland. 2007. "Evolutionary Responses to Harvesting in Ungulates." Journal Article. *J Anim Ecol* 76 (4): 669–78. https://doi.org/10.1111/j.1365-2656.2007.01244.x.

Tchernov, Eitan, and Liora Kolska Horwitz. 1991. "Body Size Diminution Under Domestication: Unconscious Selection in Primeval Domesticates." Journal Article. *Journal of Anthropological Archaeology* 10: 54–75.

Team, Stan Development. 2022. "Stan Modeling Language Users Guide and Reference Manual, Version 2.29." Generic. https://mc-stan.org.

Uerpmann, Margarethe, and Hans-Peter Uerpmann. 1994. "Animal Bones." Book Section. In *Qala'at Al-Bahrain*, edited by Flemming Hojlund and H. Hellmuth Andersen, 1: The Northern City Wall and the Islamic Fortress:417–54. Jutland Archaeological Society Publications. Aarhus: Aarhus University Press.

Wolfhagen, Jesse. 2020. "Re-Examining the Use of the LSI Technique in Zooarchaeology." Journal Article. *Journal of Archaeological Science* 123: 105254. https://doi.org/10.1016/j.jas.2020.105254.

Wright, Elizabeth, and Sarah Viner-Daniels. 2015. "Geographical Variation in the Size and Shape of the European Aurochs (Bos Primigenius)." Journal Article. *Journal of Archaeological Science* 54: 8–22. https://doi.org/10.1016/j.jas.2014.11.021.

Zeder, Melinda A. 2012. "Pathways to Animal Domestication." Book Section. In *Biodiversity in Agriculture: Domestication, Evolution, and Sustainability*, edited by P. Gepts, T. R. Famula, and R. L. Bettinger, 227–59. Cambridge: Cambridge University Press.

Zeder, Melinda A., and Brian Hesse. 2000. "The Initial Domestication of Goats (Capra Hircus) in the Zagros Mountains 10,000 Years Ago." Journal Article. *Science* 287 (5461): 2254–57.