**A Bayesian Multinomial Model for Estimating Artifact Class Diversity**

**Ryan Breslawski**

**Motivation**

Archaeologists are often interested in estimating the *diversity*[[1]](#footnote-1) of artifact classes in bounded geographic regions, with the goal of comparing artifact class diversity between those regions (e.g., Buchanan et al. 2017; Eren et al. 2016). In these comparisons, geographic regions are generally defined by an archaeologically significant research question that outlines the relative levels of diversity that we should expect. Evaluating those expectations—and the research question(s) to which they correspond—requires the ability to measure diversity reliably. Unfortunately, obtaining such a reliable measure can be challenging due to sampling biases, varying sample sizes, and decisions about how to define artifact classes.

Here, I will consider only paradigmatic classification schemes (*sensu* Dunnell 1971). In these schemes, the number of artifact classes *K* is researcher-defined, imposing a natural upper bound on classes. This differs from other classification schemes in which there may be no upper bound on *K*. Paradigmatic classification makes it possible to discuss artifact classes in terms of a proportions, where each class *k* in *K* has proportions *p1*, *p2*, …, *pK*. With infinite sampling effort, we can know these proportions exactly, and therefore calculate a measure of diversity exactly. Unfortunately, archaeologists must work with samples, which means that class proportions and measures of diversity are estimates. If we want to evaluate our archaeological expectations, we need to obtain reliable estimates and honestly communicate the uncertainty around those estimates.

I outline a method for obtaining class proportion estimates and the uncertainty around those estimates, from which we can then calculate measures of diversity and their attendant uncertainties. This is done through a Bayesian multinomial model that expresses artifact class proportions as posterior probability densities. These posterior proportion densities are then used to calculate a posterior distribution for a diversity measure of interest. A later section describes a method for handling cases where artifact class observations have uncertain membership in geographic regions. I do not review measures of diversity (e.g., Simpson’s *D* and Shannon’s *H*), as many different measures could be calculated from the posterior class proportion distribution. The measure of interest should vary with the research question and is therefore outside the scope of this overview.

**Artifact Class Distributions**

An observed sample of artifact counts across classes follows a *multinomial distribution*, where the observed count for each class *k* is a function of the total counts across classes *K* and the probability *pi* of observing class *k*. The probability vector *θ1*, *θ2*, …, *θK* (henceforth Θ) corresponds to the true proportions of each artifact class, summing to 1. Our goal is to estimate Θ. Since we can never know Θ exactly, we will be working with a probability distribution that describes our uncertainty around each *θi* in Θ: the *Dirichlet distribution*. The Dirichlet distribution can be thought of as a “distribution of distributions”, with randomly sampled Θ vectors of length *K* that each sum to 1. As we observe more data, uncertainty in the Dirichlet distribution reduces and the proportion estimates approach Θ.

**The Model**

The model outcome consists of artifact counts *C* across *K* artifact classes in *J* unique regions. As such, observations take the form of a *[J, K]* matrix. We can express the multinomial likelihood for counts across *K* artifact classes in region *jj* of *J* as

.

To make model specification easier, Θ*j* is estimated with parameters on the normal scale using the softmax function,

,

where Χ*j* is a vector of real parameter values that, once exponentiated and standardized, sum to 1 (i.e., the target Θj vector). Fitting the model to observed artifact count data provides posterior distributions for Χ*j*, and therefore Θj.

***Prior distributions for parameter values***

Values in each Χ*j* vector were assigned a prior probability of Ν(0, σ*j*), and each σ*j* received its own prior probability distribution. This “hyper prior” takes the form

|  |  |  |
| --- | --- | --- |
|  |  | . |

*σj* controls the degree of dispersion in the parameter values contained in the vector Χ*j*. Therefore, it is responsible for the amount of heterogeneity in artifact class proportions in region *jj*. More positive values in Χ*j* correspond to artifact classes that dominate the region, whereas more negative values correspond to relatively rare artifact classes. *ω* is the average log dispersion in Χ*j* across all regions *J*. *φj* is an offset from *ω* specific to region *jj*.

The fixed parameter values for *ω* and *φj* were chosen based on trial simulations from prior parameter values, as well as model sensitivity tests for simulated datasets with known parameter values. The selected values allow for a wide degree of scenarios, ranging from regions dominated by a single artifact class (e.g., where class *k* has a proportion greater than 0.9) to regions where artifact proportions are uniform across classes. In cases where region *jj* has few data, the degree of dispersion in artifact class proportions shrinks toward the average degree of dispersion estimated across all regions *J* (*ω*). In other words, when data are sparse for region *jj*, the model assumes that the dispersion of artifact class proportions in region *jj* is more likely to look like the average dispersion of artifact class proportions for the other observed *J*-1 regions.

***Model fitting***

Models were fitted with Hamiltonian Monte Carlo simulation in Stan with four chains. Each chain ran with 5000 warmup iterations and 5000 sampling iterations, producing 20,000 total sampling iterations. Model convergence was checked by visual inspection of trace plots, confirming the absence of divergent iterations, and ensuring that all R-hat values were below 1.01. Effective sample sizes for all parameters are above 1500, ensuring reasonably precise posterior distributions.

***An assumption***

The model assumes that there are not biases in how classes are sampled between regions. For example, if artifact class A is present at 50% in regions 1 and 2, the model should be able to estimate these proportions from samples drawn from those regions. However, if region 1 artifact collectors really love A artifacts while region 2 collectors really dislike A artifacts, artifact class A will not be sampled in proportion to its true proportion in each region. In other words, for unbiased estimates of artifact class proportions, the model assumes that all classes are equally as interesting between each other and between regions (that is, interesting to collectors and archaeologists). They should be sampled roughly in accordance with their true proportions.

**Uncertain Geographic Region Membership**

A recurring problem is that artifact provenience is often specific to some modern administrative unit, such as a county or state, rather than a geographic region of specific archaeological interest, such as a biome or physiographic province. In these cases, archaeologists have often measured diversity in those modern administrative units under the assumption that those units are a useful proxy for roughly spatially congruent regions of archaeological interest. Here, I outline a more principled way to handle this issue.

First, we can fit the above model to those data from modern administrative units contained entirely within individual regions of interest. In these cases, if we know with certainty that an artifact was recovered from administrative unit *aa* that is contained entirely within region *jj*, we know with certainty that it was recovered from *jj*. However, there will be cases when *aa* overlaps portions of *j1*, *j2*, etc., making the spatial relationship between the artifact and *j1*, *j2*, … uncertain. Fitting the model to artifact count data only from those administrative units contained entirely within individual regions will provide posterior probabilities for each artifact class *k* within each region.

After the model has been fitted, we can simulate dataset permutations that probabilistically assign the remaining artifacts to regions. For *N* dataset permutations, *N* posterior Θ matrices are first sampled from the fitted model. Recall that each row *j* of Θ contains *K* artifact class probabilities that sum to one. With this information, we then calculate the probability of each region *jj* membership for every regionally unassigned artifact *z*:

|  |  |
| --- | --- |
| *prob*(*jj* | *z*) | The probability of region *jj*, given artifact *z*. |
| *θz[j,k]* | The probability of observing artifact *z*, which belongs to class *k*, given region *jj* (*θ*). This value is sampled from the posterior of the model fitted to all artifacts located in administrative units contained entirely within individual regions. |
| *areaz[j]* | The area of region *jj* in the administrative unit containing artifact *z*. |
| *J* | The number of regions. The *w* index in the denominator is substituted for the *j* index in the numerator. |

In simpler terms, the probability that artifact *z* is associated with region *jj* is equal to: the product of the probability of observing artifact *z* given *jj* and the area of *jj* in the administrative unit containing artifact *z*, standardized by the sum of all such products for every region *jw* in the administrative unit.

Once region probabilities are obtained for every unassigned artifact, regions are probabilistically assigned to those artifacts. This is repeated over *N* permutations. The model is then fitted to every dataset permutation, and *N* posterior diversity densities are obtained for each region. Finally, the *N* posterior diversity densities are summed and standardized within each region to obtain a posterior diversity value distribution that averages across the dataset permutations. This builds regional membership uncertainty into each posterior diversity density.

The role of uncertain regional membership will vary depending on the dataset at hand. If most administrative units are divided up among multiple regions, then the dataset permutations will vary greatly, and the posterior diversity densities for each region will be wide. In cases where most administrative units are contained entirely within regions, dataset permutations will be similar, and the posterior diversity densities will not change much during this step. In the former scenario, many permutations may be necessary to capture the uncertainty in regional diversity values that is driven by uncertain artifact memberships in those regions. In the latter scenario, very few permutations may be necessary to capture this uncertainty (or, permutations will quickly become redundant and the summed and standardized posterior diversity densities across permutations will remain stable).

**References**

Buchanan, Briggs, Anne Chao, Chun-Huo Chiu, Robert K. Colwell, Michael J. O’Brien, Angelia Werner, and Metin I. Eren

2017 Environment-Induced Changes in Selective Constraints on Social Learning During the Peopling of the Americas. *Scientific Reports* 7:44431.

Dunnell, Robert C.

1971 *Systematics in Prehistory*. Free Press, New York.

Eren, Metin I., Anne Chao, Chun-Huo Chiu, Robert K. Colwell, Briggs Buchanan, Matthew T. Boulanger, John Darwent, and Michael J. O’Brien

2016 Statistical Analysis of Paradigmatic Class Richness Supports Greater Paleoindian Projectile-Point Diversity in the Southeast. *American Antiquity* 81(1):174—192.

1. Here, I use the term *diversity* to refer to any measure of richness, evenness, or combination of richness and evenness. Any value that can be calculated from a vector of proportions is applicable here. Given the bounded nature of paradigmatic classes, evenness measures have the potential to highlight more variation than richness measures. It is easy to imagine a scenario where multiple geographic regions have the maximum class richness allowable under a classification scheme (especially if the number of potential classes is low) yet vary dramatically in terms of class evenness. Indeed, if every class has a non-zero probability of discovery, all regions will eventually reach maximum richness under infinite sampling effort, even if the probability of discovery for a class is extraordinarily low. Since the probability of discovery for any given class is unknown, it should not be assumed to be zero. As such, we should assume that any geographic region will eventually reach maximum richness with infinite sampling effort. This renders measures of richness less useful than measures of evenness for a paradigmatic classification scheme. In contrast to richness, there is no *a priori* reason to expect true evenness to be any particular value after infinite sampling effort. [↑](#footnote-ref-1)