A Note on Bayesian Inference with Incomplete Multinomial Data with Applications for Assessing the Spatio-Temporal Variation in Pathogen-Variant Diversity

Kwang Woo Ahn[1,*], Kung–Sik Chan[1,], Ying Bai[2,3,***], and Michael Kosoy[2,****]**

[1]Department of Statistics and Actuarial Science, The University of Iowa, Iowa, U.S.A.

[2]Centers for Disease Control and Prevention, Colorado, U.S.A.

[3]University of Colorado at Boulder

**email:* kwangwoo–ahn@uiowa.edu

***email:* kung-sik-chan@uiowa.edu

****email:* bby5@cdc.gov

*****email:* mck3@cdc.gov

Summary: With recent advance in genetic analysis, it has become feasible to classify a pathogen into genetically distinct variants even though they apparently cause an infected subject similar symptoms. The availability of such data opens up the interesting problem of studying the spatio-temporal variation in the diversity of variants of a pathogen. Data on pathogen variants often suffer the problems of (i) low cell counts, (ii) incomplete classification due to laboratory problems, for example, contamination, and (iii) unseen variants. Shannon entropy may be employed as a measure of variant diversity. A Bayesian approach can be used to deal with the problems of low cell counts and unseen variants. Bayesian analysis of incomplete multinomial data may be carried out by Markov chain Monte Carlo techniques. However, for pathogen-variant data, it often happens that there is only one source of missingness, namely, some subjects are known to be infected by some unidentified pathogen variant. We point out that for incomplete data with disjoint sources of missingness, Bayesian analysis can be more efficiently done by an iid sampling scheme from the posterior distribution. We illustrate the method by analyzing a dataset on prevalence of bartonella infection among individual colonies of prairie dog at the study site in Colorado, from 2003 to 2006.

Key words: Shannon entropy, Bartonella, Bayes factor, Dirichlet distribution, Pathogen diversity, Spatial epidemiology.

1. Introduction

In the study of the epidemiology of an infectious disease, it is important to monitor the spatio-temporal variation of the prevalence rate of the disease. With the modern advance in genetics, it has been found that a disease-causing pathogen often admits multiple variants (Eames and Keeling, 2006; Kosoy et al., 2004; Read and Taylor, 2001). The effects of the multiplicity of pathogen strains on the epidemiology of an infectious disease have received much attention in the literature (Li et al., 2003; Read and Taylor, 2001). Here, we address several statistical problems encountered in studying the spatio-temporal variation in the pathogen-variant diversity of an endemic.

The main statistical problem concerns the analysis of incomplete multinomial data. Suppose that a subject can be classified into one of k categories, with the categories denoted by the symbols from 1 to k. In our epidemiological application, the first category stands for no detected bartonella in the blood sample of the subject, whereas the other categories signify that the subject is infected and record the corresponding type of bartonella strain infecting the subject; hence, there are altogether $k-1$ strains of the pathogen. A common laboratory problem is that the blood sample of a subject may not be usable due to contamination or other problems, so that while the subject is known to be infected, the exact nature of the infecting pathogen strain is unknown. Often, this constitutes the only source of incompleteness in the multinomial data. Here, we address the problem of analyzing such kind of incomplete multinomial data. (In fact, we shall consider slightly more general kind of incomplete multinomial data with disjoint sources of missingness.)

Besides the incomplete-data problem, epidemiological data often have low or zero cell counts, rendering large-sample asymptotics unreliable. The Bayesian approach with non-informative prior is, however, more appropriate for such cases. Yet another problem is that of unseen variants, not observed in the sample perhaps due to their low probabilities and/or inadequate sampling efforts. Ignoring the possibility of unseen variants may result in bias when estimating some functional, e.g. Shannon entropy (Shannon, 1948, and see also section 3) of the variant distribution. The latter problem may be tackled by parametric or nonparametric methods from a frequentist perspective, see Bunge and Fitzpatrick (1993), Chao and Shen (2003). Here, we propose a Bayesian approach to deal with the problems of low cell counts and unknown number of categories. In practice, some upper bound on the total number of categories is known, in which case a non-informative prior distribution may then be employed.

The posterior distribution with incomplete multinomial data is often intractable, and Markov chain Monte Carlo techniques have been proposed to draw inference based on dependent sample from the posterior distribution, see Gelman et al. (2003). A main purpose of this note is to point out that for incomplete multinomial data with disjoint sources of missingness and conditional on the number of categories, the posterior distribution has a simple representation that admits exact calculation of its lower moments, and that iid samples can be easily drawn from the posterior distribution. Furthermore, the posterior distribution of the number of categories has a closed-form solution. These results make it easy to study the posterior distribution of some nonlinear functional of the multinomial probability distribution.

Shannon entropy is widely used to quantify species richness and diversity in ecology (Holgate, 1981; Pielou, 1966; Lande, 1996; Jost, 2006). In our epidemiological application, we employ Shannon entropy to measure the diversity of bartonella strains among individual colonies of prairie dogs at the field site in Colorado, from 2003 to 2006. The bartonella monitoring dataset is incomplete as some data from 2004 are partially classified in that for some infected prairie dogs, the bartonella strains are unknown.

The outline of this note is as follows. In section 2, we state some useful results on the posterior distribution with incomplete multinomial data. All proofs are deferred to the appendices. In section 3, we illustrate the result with the aforementioned epidemiological application. We briefly conclude in section 4.

2. Bayesian Analysis of Incomplete Multinomial Data

To fix ideas, we start with an example taken from the bartonella monitoring data. In 2004, among the prairie dogs trapped at site 1, 14 of them had no detected bartonella, 5 were infected with bartonella variant A, 1 with variant B, 1 with variant C, and 1 missing observation at site 1. For the missing observation, it is known to be infected but by unknown bartonella variant. However, other bartonella variants were present in data at other sites. For example, in site 14, there were 11 undetected, 1 bartonella variant E, and 4 missing observations. This suggests there may be some unseen variants present at site 1, leading to the following scenarios:

(1) If no possibility of other variants besides A, B, and C is considered to be present in site 1, the observations there can be represented as a vector (14,5,1,1,1), with the first four counts being those of the 4 categories: undetected, variants A, B, and C, and the last count that of the missing category. Also the lone missing observation should be one of variants A, B, and C.

(2) If the possibility of variant E in site 1 is entertained, the data may be written as (14,5,1,1,0,1), with the counts augmented by a zero, the cell count of variant E. In this case, we have 5 categories: undetected, variants A, B, C, and E. Also the missing observation should be one of variants A, B, C, and E.

(3) In the case of c unseen variants, the observations have to be augmented with c zeros.

Thus, the choice of the number of unseen variants plays an important role in determining the format of data. Below, q stands for the number of observed categories, i.e., the number of categories with positive counts, and k equals the true number of (observed plus unseen) categories. In the above example, $q = k = 4$ in case (1), and $q = 4$, $k = 5$ in case (2).

We shall derive the posterior distribution by first deriving the conditional posterior distribution given the total number of categories k. Then, we provide the formula for the posterior probability of k. Let $\{W_1, \ldots, W_n\}$ be a random sample from a probability distribution taking values in the finite set $A = \{1, 2, \ldots, k\}$ with $\text{Prob}(W = j) = \theta_j$. We first consider the case of known k. Suppose that some of the Ws are incompletely observed and they are only known to belong to some non-singleton proper subsets of A, denoted by $A_j, j = 1, \ldots, m$; the sampling mechanism of missing at random will be assumed. Let X_i be the non-zero counts of Ws that equal i, for $i = 1, \ldots, q$ where $q \leqslant k$ and X_{A_j} be the count of incompletely observed Ws that belong to A_j. The observed data consist of $\boldsymbol{Z} = (X_1, \ldots, X_q, X_{A_1}, \ldots, X_{A_m})^T$ with the augmented counts $X_{q+1} = \cdots = X_k = 0$ under the assumption of a total number of k categories. In the epidemiological application discussed in the next section, there is only one kind of incomplete observations and $A_1 = \{1\}^c = \{2, 3, \ldots, k\}$. Let the prior distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ be the Dirichlet distribution with hyperparameter vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)^T$, that is, the prior pdf equals

$$\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}.$$

Below, the notation $\pi(\cdot|\cdot)$ denotes the conditional pdf of the the first expression given the second expression. Let $\boldsymbol{X} = (X_1, \ldots, X_k, X_{A_1}, \ldots, X_{A_m})^T$. The posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X} = \boldsymbol{x}$ equals

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto \prod_{i=1}^{k} \theta_i^{\alpha_i + x_i - 1} \prod_{i=1}^{m} \Big(\sum_{j \in A_i} \theta_j \Big)^{x_{A_i}}. \tag{1}$$

In general, the preceding posterior distribution is intractable and inference may have to be drawn by drawing dependent samples from the posterior distribution, via Markov chain Monte Carlo techniques, see Gelman et al. (2003). Interestingly, for the case that the A_j's are disjoint subsets, the posterior distribution has a tractable representation that admits closed-form analysis, for example, moment calculations. To describe the representation, let $A_0 = A - \cup_{j=1}^{m} A_j$. We show below that the posterior distribution is tractable by re-parameterizing the model using the parameters defined by the group probabilities $\boldsymbol{U} = (U_0, U_1, \ldots, U_m)^T$, where $U_j = \sum_{i \in A_j} \theta_i$, $j = 0, 1, \ldots, m$, and the conditional probabilities of individual cells within each group $\boldsymbol{V}_{A_j} = (\theta_i / U_j, i \in A_j)^T, 0 \leqslant j \leqslant m$. Clearly, the sum of the components in $\boldsymbol{U}$ and those of each $\boldsymbol{V}_{A_j}$ are constrained to be 1. Furthermore, it is readily checked that $\boldsymbol{\theta}$ and $(\boldsymbol{U}, \boldsymbol{V}_{A_j}, j = 0, \ldots, m)$ are equivalent parameterization as they bear a one-to-one relationship.

It follows from (1) that

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{x}) &\propto \prod_{i=1}^{k} \theta_i^{\alpha_i + x_i - 1} \prod_{i=1}^{m} \Big(\sum_{j \in A_i} \theta_j \Big)^{x_{A_i}} \\
&\propto \prod_{i \in A_0} \Big(\frac{\theta_i}{\sum_{j \in A_0} \theta_j} \Big)^{\alpha_i + x_i - 1} \prod_{i \in A_1} \Big(\frac{\theta_i}{\sum_{j \in A_1} \theta_j} \Big)^{\alpha_i + x_i - 1} \\
&\quad \times \cdots \times \prod_{i \in A_m} \Big(\frac{\theta_i}{\sum_{j \in A_m} \theta_j} \Big)^{\alpha_i + x_i - 1} \\
&\quad \times \Big(\sum_{i \in A_0} \theta_i \Big)^{\sum_{j \in A_0} (\alpha_j + x_j - 1)} \Big(\sum_{i \in A_1} \theta_i \Big)^{x_{A_1} + \sum_{j \in A_1} (\alpha_j + x_j - 1)} \\
&\quad \times \cdots \times \Big(\sum_{i \in A_m} \theta_i \Big)^{x_{A_m} + \sum_{j \in A_m} (\alpha_j + x_j - 1)}.
\end{aligned}$$

Hence, $\boldsymbol{U}$ and $\boldsymbol{V}_{A_j}, j = 0, \ldots, m$ are jointly independent. Let $A_p = \{p_1, \ldots, p_{n_p}\}$, $p = 0, \ldots, m$. Then, $\boldsymbol{V}_{A_p} = (\frac{\theta_{p_1}}{\sum_{j \in A_p} \theta_j}, \ldots, \frac{\theta_{p_{n_p}}}{\sum_{j \in A_p} \theta_j})$ follows the Dirichlet distribution with parameter vector $(\alpha_{p_1} + x_{p_1}, \ldots, \alpha_{p_{n_p}} + x_{p_{n_p}})$, for $p = 0, \ldots, m$. Also, $\boldsymbol{U} = (\sum_{i \in A_0} \theta_i, \sum_{i \in A_1} \theta_i, \ldots, \sum_{i \in A_m} \theta_i)$ has the Dirichlet distribution with parameter vector $(\sum_{j \in A_0}(\alpha_j + x_j), x_{A_1} + \sum_{j \in A_1}(\alpha_j + x_j), \ldots, x_{A_m} + \sum_{j \in A_m}(\alpha_j + x_j))$. The above representation of the posterior distribution furnishes a simple way to draw random samples from the posterior distribution which admits analytical formulas for the lower moments. See Appendix A for proofs of the preceding claims.

We now address the problem of unknown k. We assume that an upper bound, say M, of k is known, and adopt a flat prior distribution for k over the range from 1 to M. It is shown in Appendix B that for $k \geqslant q$, the posterior probability of k given the observations $\boldsymbol{Z} = (X_1, \ldots, X_q, X_{A_1}, \ldots, X_{A_m})^T = \boldsymbol{z}$ equals (now $\boldsymbol{\theta}$ is written as $\boldsymbol{\theta}_k$ to emphasize its dependence on k)

$$\begin{aligned} \pi(k|\boldsymbol{z}) &\propto \pi(\boldsymbol{z}|k)\pi(k) \propto \int_{\mathcal{A}} \pi(\boldsymbol{x}|\boldsymbol{\theta}_k, k)\pi(\boldsymbol{\theta}_k|k)\pi(k)d\boldsymbol{\theta}_k \\ &\propto \frac{1}{M} \frac{(\sum_{i=1}^{q} x_i + \sum_{i=1}^{m} x_{A_i})!}{x_1! \cdots x_q! x_{A_1}! \cdots x_{A_m}!} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=0}^{m} \Big\{ \frac{\prod_{j \in A_i} \Gamma(\alpha_j + x_j)}{\Gamma(\sum_{j \in A_i}(\alpha_j + x_j))} \Big\} \\ &\times \frac{\prod_{i=0}^{m} \Gamma(\sum_{j \in A_i}(\alpha_j + x_j) + x_{A_i})}{\Gamma(\sum_{i=0}^{m}\{\sum_{j \in A_i}(\alpha_j + x_j) + x_{A_i}\})}, \end{aligned} \tag{2}$$

where $x_{A_0} = 0$ and $\Gamma(\cdot)$ denotes the Gamma function. If $k < q$, $\pi(k|\boldsymbol{z})$ is clearly zero. Since $\sum_{k=1}^{M} \pi(k|\boldsymbol{z}) = \sum_{k=q}^{M} \pi(k|\boldsymbol{z}) = 1$, we can obtain the closed form of the posterior probability of k by normalizing the left side of (2) to sum up to 1. Note that for complete data, i.e., $x_{A_1} = \cdots = x_{A_m} = 0$, the posterior probability can be simplified as follows:

$$\pi(k|x_1, \ldots, x_q) \propto \frac{1}{M} \frac{(\sum_{i=1}^{q} x_i)!(\sum_{i=1}^{k} \alpha_i - 1)!}{(\sum_{i=1}^{q} x_i + \sum_{i=1}^{k} \alpha_i - 1)!} \prod_{i=1}^{q} \Big\{ \frac{(x_i + \alpha_i - 1)!}{x_i!(\alpha_i - 1)!} \Big\}. \tag{3}$$

3. An Epidemiological Example: Spatio-temporal Variation of the Diversity of Bartonella Variants in a Prairie-Dog System

Data on temporal dynamics and spatial distribution of Bartonella in black-tailed prairie dogs (*Cynomys ludovicianus*) based on a longitudinal study conducted in 20 black-tailed prairie dog (BTPD) colonies in Boulder County, Colorado from 2003 to 2005, but only 9 sites were examined in 2006. Prevalence of bartonella in prairie dogs was determined by culturing blood samples in specific medium and strains infecting the trapped prairie dogs were identified by sequencing of the target gene of bartonella. Except in 2004, the disease status and the type of bartonella strain affecting the trapped prairie dogs were determined. However, results on the bartonella variant for some trapped prairie dogs in 2004 were incomplete because of technical problems. Two interesting epidemiological questions are (i) whether there is any spatial variation in the diversity of bartonella variants within each year, and (ii) whether the diversity of the bartonella variants change over time.

To study these issues, we first lay out the framework for analyzing complete data from a single year, and assuming that there are n sites. In particular, we can drop the year index. Let k_s be the true number of categories at site s. Let $\boldsymbol{\theta}_s = (\theta_{s,1}, \ldots, \theta_{s,k_s})$ be the true proportions of the k_s categories at site s, $s = 1, \ldots, n$. Recall that the first category stands for no detected bartonella in the blood sample of a random prairie dog, and category 2 stands for a random prairie dog being infected by bartonella variant 1, etc. Let $\boldsymbol{x}_s = (x_{s,1}, \ldots, x_{s,k_s})$ be the observed counts of the k_s categories at sampling site s. Assume the maximum number of categories is M for all sites. Assume a flat prior for k_s that are independent across sites. Given $\{k_s, s = 1, \ldots, n\}$, we adopt the non-informative prior, that is, the prior distribution of $\boldsymbol{\theta}_s$, $s = 1, \ldots, n$ is independent across site and non-informative at each site. We assume that the counts are conditionally independent multinomially distributed given the true proportions over all sites. Consequently the posterior distribution of $(k_s, \boldsymbol{\theta}_s^T)$ are independent across sites. Further, given $\{k_s, s = 1, \ldots, n\}$, the marginal posterior distribution of $\boldsymbol{\theta}_s$ equals

$$\begin{aligned} \boldsymbol{\theta}_s | x_{s,1}, \ldots, x_{s,k_s} &\sim \text{Dirichlet}(1 + x_{s,1}, \ldots, 1 + x_{s,k_s}), \\ &\propto \prod_{i=1}^{k_s} \theta_{s,i}^{x_{s,i}}, \quad \sum_{i=1}^{k_s} \theta_{s,i} = 1, \quad s = 1, \ldots, n, \end{aligned}$$

where the sign $\sim$ is read as "distributed as." Note that a realization of $\boldsymbol{\theta}_s$ from the Dirichlet distribution with parameter vector $(\alpha_{s,1}, \ldots, \alpha_{s,k_s})$ can be obtained by the following well-known sampling scheme:

(1) $Y_i \sim \text{Gamma}(\text{shape} = \alpha_{s,i},\ \text{scale} = 1),\ i = 1, \ldots, k_s,$

(2) $V = \sum_{i=1}^{k_s} Y_i \sim \text{Gamma}(\text{shape} = \sum_{i=1}^{k_s} \alpha_{s,i},\ \text{scale} = 1),$

(3) $(\frac{Y_1}{V}, \ldots, \frac{Y_{k_s}}{V}) \sim \text{Dirichlet}(\alpha_{s,i}, \ldots, \alpha_{s,k_s}),$

where the Y_i's are independent.

It remains to derive the posterior probability of k_s, the number of categories at site s.

The data, which is not shown in this paper, indicates that there may be some unobserved categories, that is, unseen variants at each site. Let q_s be the number of categories with positive counts at site s. It follows form (3) that the posterior probability of the number of categories, k_s, is given by the formula

$$\pi(k_s|x_{s,1},\ldots,x_{s,q_s}) = \frac{\frac{(k_s-1)!}{(\sum_{i=1}^{q_s} x_{s,i}+k_s-1)!}}{\sum_{j=q_s}^{M}\left[\frac{(j-1)!}{(\sum_{i=1}^{q_s} x_{s,i}+j-1)!}\right]}$$

for $q_s \leqslant k_s \leqslant M$, and zero otherwise.

At each site, the diversity of the bartonella variants can be quantified by Shannon entropy which, at site s, is given by $D_s = -\sum_{i=1}^{k_s} \theta_{s,i} \log_2 \theta_{s,i}$. A larger entropy means more diversity. The posterior distribution of D_s can be easily studied by drawing a random sample from the joint posterior distribution of $\boldsymbol{\theta}_s$ and k_s and compute Shannon entropy for each realization of the probabilities. For a specific algorithm to obtain samples of Shannon entropy for site s,

(1) Draw k_s according to the posterior probabilities $\pi(k_s|x_{s,1},\ldots,x_{s,q_s})$.

(2) Draw $Y_i \sim \text{Gamma}(\text{shape} = x_{s,i}+1,\ \text{scale} = 1),\ i = 1,\ldots,k_s$.

(3) Draw $V = \sum_{i=1}^{k_s} Y_i \sim \text{Gamma}(\text{shape} = \sum_{i=1}^{k_s} x_{s,i} + k_s,\ \text{scale} = 1)$.

(4) Obtain $\boldsymbol{\theta}_s = (\frac{Y_1}{V},\ldots,\frac{Y_{k_s}}{V}) \sim \text{Dirichlet}(x_{s,1}+1,\ldots,x_{s,k_s}+1)$.

(5) Calculate D_s.

(6) Repeat (1)–(5) T times.

The posterior median Shannon entropy at each site can be used as a summary statistic of the bartonella-variant diversity at that site. The spatial variation of the within-year bartonella-variant diversity is displayed in Figure 1.

[Figure 1 about here.]

The sub-figure for 2004 in Figure 1, however, requires some modification of the method due to the presence of missing data in that year. The strain of some of the bartonella-positive prairie dogs trapped in 2004 were unknown. Consequently, a new category is introduced to account for such subjects. Specifically, let there be x_{s,A_1} prairie dogs at site s that were found to be bartonella-positive, but of unknown strain. Here $A_1 = \{1\}^c$, the complement of the undetected category. It follows from (2) that

$$\pi(k_s|x_{s,1},\ldots,x_{s,q_s},x_{s,A_1}) \propto \frac{1}{M}\frac{(k_s-1)!(\sum_{i=1}^{q_s} x_i + x_{s,A_1})!}{(\sum_{i=1}^{q_s} x_{s,i} + x_{s,A_1} + k_s - 1)!}\frac{(\sum_{i=2}^{q_s} x_{s,i} + x_{s,A_1} + k_s - 2)!}{x_{s,A_1}!(\sum_{i=2}^{q_s} x_{s,i} + k_s - 2)!},$$

for $q_s \leqslant k_s \leqslant M$ and zero otherwise. Now, we can apply the results from section 2 to derive the posterior distribution given k_s. The posterior cell probabilities at a site s is given as follows:

$$\begin{aligned} \theta_{s,1} &\sim \text{Beta}(x_{s,1}+1, \sum_{j=1}^{k_s} x_{s,j} + x_{s,A_1} + k_s - (x_{s,1}+1)), \\ \Big(\frac{\theta_{s,2}}{1-\theta_{s,1}},\ldots,\frac{\theta_{s,k_s}}{1-\theta_{s,1}}\Big) &\sim \text{Dirichlet}(x_{s,2}+1,\ldots,x_{s,k}+1). \end{aligned}$$

Hence, $\theta_{s,1}$ and $\Big(\frac{\theta_{s,2}}{1-\theta_{s,1}},\ldots,\frac{\theta_{s,k_s}}{1-\theta_{s,1}}\Big)$ are independent Dirichlet random variables. Therefore, a random sample of size T from the posterior distribution of $\boldsymbol{\theta}_s$, $s = 1,\ldots,n$ can be generated by the following sampling scheme:

(1) Draw k_s according to the posterior probabilities $\pi(k_s|x_{s,1},\ldots,x_{s,q_s},x_{s,A_1})$.

(2) Draw $\theta_{s,1} \sim \text{Beta}(x_{s,1}+1, \sum_{j=1}^{k_s} x_{s,j} + x_{s,A_1} + k_s - (x_{s,1}+1))$.

(3) Draw $(\frac{\theta_{s,2}}{1-\theta_{s,1}},\ldots,\frac{\theta_{s,k_s}}{1-\theta_{s,1}}) = (p_{s,2},\ldots,p_{s,k_s}) \sim \text{Dirichlet}(x_{s,2}+1,\ldots,x_{s,k_s}+1)$.

(4) Obtain $\theta_{s,i} = p_{s,i}(1-\theta_{s,1})$, $i = 2,\ldots,k_s$.

(5) Repeat (1)–(4) T times.

As before, the posterior distribution of the Shannon entropy can be easily obtained by computing Shannon entropy for each posterior probability vector.

The total number of variants observed from 2003 to 2006 is 9. Thus, we have a total of 10 observed categories including the category of no detected bartonella, which means $M \geqslant 10$. We tried $M = 10,\ 11,\ 12,\ 13$, and 14, and found the results to be robust against the choice of M. Figure 1 plots a circle centered at each site with its area proportional to the median posterior Shannon entropy of the bartonella strains at that site, with simulation size equal to $T = 3000$ and the maximum number of categories equal to $M = 13$.

Figure 1 suggests that there was spatial variation in the diversity of bartonella strains in each year over the study period. The spatial homogeneity hypothesis can be assessed by computing the Bayes factor (Appendix C) for the null hypothesis $H_0 : p_{s,i} \equiv p_i, \forall s$ versus the alternative hypothesis H_1 that H_0 is invalid. The Bayes factor is essentially the ratio of the posterior probabilities of the two hypotheses assuming equal prior probabilities for the two hypotheses. All Bayes factors are found to be less than 0.0001 for $M = 10,\ 11,\ 12,\ 13,$ and 14. Thus, there is strong statistical evidence for spatial heterogeneity in the distribution of the bartonella strains.

[Figure 2 about here.]

[Table 1 about here.]

[Table 2 about here.]

To assess the extent of temporal variation, we compute the yearly average Shannon entropy by computing the simple mean of the entropy across sites, based on the posterior distribution under the general hypothesis that the bartonella distribution may vary across site and year. Figure 2 shows the posterior distribution of the annual Shannon entropy of the bartonella variants, over the study period. The posterior distributions are fairly symmetric for 2004 and 2005 but somewhat skewed to the right for 2003 and 2006. Table 1 shows the posterior means of Shannon entropy. For $M = 10,\ 11,\ 12,\ 13,$ and 14, the results are almost identical. Table 2 reports the posterior 95% intervals of the annual Shannon entropy and the posterior mean entropies for $M = 13$. Based on Table 2, there is strong statistical evidence that the entropy increased gradually from 2003 to 2004, then jumped sharply in 2005 and decreased slightly in 2006.

4. Conclusion

We have pointed out that, for a particular kind of incomplete multinomial data that commonly occur in epidemiology, the posterior distribution can be studied based on drawing iid samples rather than the less efficient method of drawing dependent samples via Markov chain Monte Carlo. Using this approach, we studied the spatio-temporal variation in the diversity of bartonella strains among prairie dogs in a monitoring system in Colorado. We employed Shannon entropy in studying the bartonella-strain diversity, by analyzing the data from a Bayesian point of view with a uniform prior. We found that the diversity of bartonella variants increased from 2003 to 2005, but decreased in 2006. However, the results suggest that the bartonella-variant diversity in 2006 was still higher than that of 2003 and 2004. A biologically interesting question is to further probe the factors affecting the found temporal variation in the Bartonella-strain diversity. On the other hand, an interesting statistical problem concerns the development of more efficient Markov chain sampling from general, incomplete multinomial data based on the results reported in Section 2.

Acknowledgements

We gratefully acknowledge Sharon Collinge of Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado and all field crews for trapping and collecting blood samples. The field investigation and laboratory analysis was supported by grants from the NSF/NIH joint program in Ecology of Infectious Diseases (DEB-0224328), the National Center for Environmental Research (NCER) STAR program of the US-EPA (R-82909101-0). KWA and KSC thank the National Science Foundation (NSF CMG-0620789) for partial support.

References

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of American Statistical Association* **88,** 364–73.

Chao, A. and Shen, T. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample, *Environmental and Ecological Statistics* **10,** 429–443.

Eames, K. and Keeling, M. (2006). Coexistence and specialization of pathogen strains on contact networks, *The American Naturalist* **168,** 230–241.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, 2nd edition. Boca Raton, Florida: Chapman and Hall.

Holgate, P. (1981). The statistical entropy of a sample from a community of species, *Biometrics* **37,** 795–799.

Jost, L. (2006). Entropy and diversity. *Oikos* **113,** 363–375.

Kosoy, M., Mandel, E., Green, D., Marston, E., Jones, D., and Childs, J. (2004). Prospective studies of bartonella of rodents. Part II. Diverse infections in a single community. *Vector-Borne and Zoonotic Diseases* **4,** 296–305.

Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* **76,** 5–13.

Li, Z., Ma, Z., Blythe, S., and Castillo-Chavez, C. (2003). Coexistence of pathogens in sexually-transmitted disease models. *Journal of Mathematical Biology* **47,** 547–568.

Pielou, E.C. (1966). Shannon's formula as a measure of specific diversity: its use and misuse. *The American Naturalist* **100,** 463–465.

Read, A. and Taylor, L. (2001). The ecology of genetically diverse infections. *Science* **292,** 1099–1102.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27,** 379–423.

Appendix

Appendix A: Proof of Distribution Results Claimed in Section 2

Let's consider the unnormalized posterior distribution of $\boldsymbol{\theta}$:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{x}) \;\propto\; & \prod_{i\in A_0}\Big(\frac{\theta_i}{\sum_{j\in A_0}\theta_j}\Big)^{\alpha_i+x_i-1}\prod_{i\in A_1}\Big(\frac{\theta_i}{\sum_{j\in A_1}\theta_j}\Big)^{\alpha_i+x_i-1}\times\cdots\times\prod_{i\in A_m}\Big(\frac{\theta_i}{\sum_{j\in A_m}\theta_j}\Big)^{\alpha_i+x_i-1}\\ & \times\Big(\sum_{i\in A_0}\theta_i\Big)^{\sum_{j\in A_0}(\alpha_j+x_j-1)}\Big(\sum_{j\in A_1}\theta_j\Big)^{x_{A_1}+\sum_{j\in A_1}(\alpha_j+x_j-1)}\times\cdots\times\Big(\sum_{j\in A_m}\theta_j\Big)^{x_{A_m}+\sum_{j\in A_m}(\alpha_j+x_j-1)}\end{aligned}$$

Without loss of generality, assume $A_0=\{1,2,\ldots,a_0\},\; A_1=\{a_0+1,a_0+2\ldots,a_1\},\ldots,A_m=\{a_{m-1}+1,a_{m-1}+2\ldots,a_m\}$ where $a_m=k$. Consider the transformation from $(\theta_1,\ldots,\theta_{k-1})$ to $(V_1,\ldots,V_{a_0-1},U_0,\ldots,V_{a_{m-2}+1},\ldots,V_{a_{m-1}-1},U_{m-1},V_{a_{m-1}+1},\ldots,V_{a_m-1})$ defined by the following formula:

$$\begin{gathered}V_1=\tfrac{\theta_1}{\sum_{j\in A_0}\theta_j},\ldots,V_{a_0-1}=\tfrac{\theta_{a_0-1}}{\sum_{j\in A_0}\theta_j},U_0=\textstyle\sum_{j\in A_0}\theta_j,\\ V_{a_0+1}=\tfrac{\theta_{a_0+1}}{\sum_{j\in A_1}\theta_j},\ldots,V_{a_1-1}=\tfrac{\theta_{a_1-1}}{\sum_{j\in A_1}\theta_j},U_1=\textstyle\sum_{j\in A_1}\theta_j,\\ \vdots\\ V_{a_{m-2}+1}=\tfrac{\theta_{a_{m-2}+1}}{\sum_{j\in A_{m-1}}\theta_j},\ldots,V_{a_{m-1}-1}=\tfrac{\theta_{a_{m-1}-1}}{\sum_{j\in A_{m-1}}\theta_j},U_{m-1}=\textstyle\sum_{j\in A_{m-1}}\theta_j,\\ V_{a_{m-1}+1}=\tfrac{\theta_{a_{m-1}+1}}{\sum_{j\in A_m}\theta_j},\ldots,V_{a_m-1}=\tfrac{\theta_{a_m-1}}{\sum_{j\in A_m}\theta_j}.\end{gathered}$$

Letting $a_{-1}=0$, $V_{a_i}=\frac{\theta_{a_i}}{\sum_{j\in A_i}\theta_j}, i=0,1,\ldots,m$ and $U_m=\sum_{j\in A_m}\theta_j$, we have the following constraints:

$$\begin{aligned}V_{a_i} &= 1-\sum_{r=a_{i-1}+1}^{a_i-1}V_r,\; i=0,\ldots,m,\\ U_m &= 1-\sum_{r=0}^{m-1}U_r.\end{aligned}$$

Then, the inverse transformation from $(V_1,\ldots,V_{a_0-1},U_0,\ldots,V_{a_{m-2}+1},\ldots,V_{a_{m-1}-1},U_{m-1},$ $V_{a_{m-1}+1},\ldots,V_{a_m-1})$ to $(\theta_1,\ldots,\theta_{k-1})$ is given by

$$\theta_1=V_1U_0,\ldots,\theta_{a_0-1}=V_{a_0-1}U_0,\theta_{a_0}=(1-\textstyle\sum_{r=1}^{a_0-1}V_r)U_0,$$

$$\theta_{a_0+1}=V_{a_0+1}U_1,\ldots,\theta_{a_1-1}=V_{a_1-1}U_1,\theta_{a_1}=(1-\textstyle\sum_{r=a_0+1}^{a_1-1}V_r)U_1,$$

$$\vdots$$

$$\theta_{a_{m-1}+1} = V_{a_{m-1}+1}(1 - \textstyle\sum_{r=0}^{m-1} U_r), \ldots, \theta_{a_m-1} = V_{a_m-1}(1 - \textstyle\sum_{r=0}^{m-1} U_r).$$

Then, the Jacobian matrix equals

$$\begin{aligned} \boldsymbol{J} &= \frac{\partial(\theta_1, \ldots, \theta_{k-1})}{\partial(V_1, \ldots, V_{a_0-1}, U_0, \ldots, V_{a_{m-2}+1}, \ldots, V_{a_{m-1}-1}, U_{m-1}, V_{a_{m-1}+1}, \ldots, V_{a_m-1})} \\ &= \begin{pmatrix} \boldsymbol{J}_0 & & & & \\ & \boldsymbol{J}_1 & & & \\ & & \ddots & & \\ & & & \boldsymbol{J}_{m-1} & \\ \boldsymbol{B}_0 & \boldsymbol{B}_1 & \cdots & \boldsymbol{B}_{m-1} & \boldsymbol{J}_m \end{pmatrix}, \end{aligned}$$

where the blank entries (here and below) are zeros and for $i = 0, \ldots, m-1$,

$$\boldsymbol{J}_i = \begin{pmatrix} U_i & & & & V_{a_{i-1}+1} \\ & U_i & & & V_{a_{i-1}+2} \\ & & \ddots & & \vdots \\ & & & U_i & V_{a_i-1} \\ -U_i & -U_i & \cdots & -U_i & 1 - \sum_{r=a_{i-1}+1}^{a_i-1} V_r \end{pmatrix}, \quad \boldsymbol{B}_i = \begin{pmatrix} 0 & \cdots & 0 & -V_{a_{m-1}+1} \\ 0 & \cdots & 0 & -V_{a_{m-1}+2} \\ 0 & \cdots & 0 & \vdots \\ 0 & \cdots & 0 & -V_{a_m-1} \end{pmatrix},$$

and

$$\boldsymbol{J}_m = \begin{pmatrix} 1 - \sum_{r=0}^{m-1} U_r & & \\ & \ddots & \\ & & 1 - \sum_{r=0}^{m-1} U_r \end{pmatrix}.$$

Here, $\boldsymbol{J}_i$ is a $(a_i - a_{i-1}) \times (a_i - a_{i-1})$ matrix for $i = 0, \ldots, m-1$, $\boldsymbol{B}_i$ is a $(a_m - a_{m-1} - 1) \times (a_i - a_{i-1})$ matrix, and $\boldsymbol{J}_m$ is a $(a_m - a_{m-1} - 1) \times (a_m - a_{m-1} - 1)$ matrix. Now, for $i = 0, \ldots, m-1$, it can be readily checked that the determinant

$$\begin{aligned} |J_i| &= U_i^{a_i - a_{i-1} - 1}\{(1 - \sum_{r=a_{i-1}+1}^{a_i-1} V_r) + V_{a_i-1}\} + U_i^{a_i - a_{i-1} - 1}(\sum_{r=a_{i-1}+1}^{a_i-2} V_r) \\ &= U_i^{a_i - a_{i-1} - 1}. \end{aligned}$$

Thus, $|J| = \prod_{i=0}^{m} |J_i| = \prod_{i=0}^{m} U_i^{a_i - a_{i-1} - 1}$ where $U_m = 1 - \sum_{r=0}^{m-1} U_r$. Letting $x_{A_0} = 0$, $\boldsymbol{U}$ be the vector of U_i's and $\boldsymbol{V}$ the vector of V_j's, we have

$$\begin{aligned}
\pi(\boldsymbol{U}, \boldsymbol{V}|\boldsymbol{x}) &\propto \prod_{i \in A_0} V_i^{\alpha_i + x_i - 1} \prod_{i \in A_1} V_i^{\alpha_i + x_i - 1} \cdots \prod_{i \in A_m} V_i^{\alpha_i + x_i - 1} \\
&\quad \times \prod_{i=0}^{m} U_i^{x_{A_i} + \sum_{j \in A_i} (\alpha_j + x_j - 1)} \prod_{i=0}^{m} U_i^{a_i - a_{i-1} - 1} \\
&\propto \prod_{i \in A_0} V_i^{\alpha_i + x_i - 1} \prod_{i \in A_1} V_i^{\alpha_i + x_i - 1} \cdots \prod_{i \in A_m} V_i^{\alpha_i + x_i - 1} \\
&\quad \times \prod_{i=0}^{m} U_i^{a_i - a_{i-1} - 1 + x_{A_i} + \sum_{j \in A_i} (\alpha_j + x_j - 1)} \\
&\propto \prod_{i \in A_0} V_i^{\alpha_i + x_i - 1} \prod_{i \in A_1} V_i^{\alpha_i + x_i - 1} \cdots \prod_{i \in A_m} V_i^{\alpha_i + x_i - 1} \\
&\quad \times \prod_{i=0}^{m} U_i^{x_{A_i} + \sum_{j \in A_i} (\alpha_j + x_j) - 1}.
\end{aligned}$$

This proves that the random vectors $(V_1, \ldots, V_{a_0}), \ldots, (V_{a_{m-1}+1}, \ldots, V_{a_m})$, and $(U_0, \ldots, U_m)$ are jointly independent. Upon noting that $a_i - a_{i-1}$ equals the cardinality of A_i, $i = 1, \ldots, m$ and that $\boldsymbol{V}_{A_j} = (V_{1+a_{j-1}}, \ldots, V_{a_j})^T$ for all j, this completes the proof of the distributional results claimed in Section 2.

We now derive the first and second posterior moments of $\boldsymbol{\theta}$. Let i, j be such that they belong to some A_p. Based on the distributional results in Section 2, we obtain

$$\begin{aligned}
\mathrm{E}\Big(\frac{\theta_i}{\sum_{r \in A_p} \theta_r} |\boldsymbol{x}\Big) &= \frac{\alpha_i + x_i}{\sum_{r \in A_p} (\alpha_r + x_r)}, \\
\mathrm{Var}\Big(\frac{\theta_i}{\sum_{r \in A_p} \theta_r} |\boldsymbol{x}\Big) &= \frac{(\alpha_i + x_i)(\sum_{r \in A_p} (\alpha_r + x_r) - \alpha_i + x_i)}{(\sum_{r \in A_p} (\alpha_r + x_r))^2 (\sum_{r \in A_p} (\alpha_r + x_r) + 1)}, \\
\mathrm{Cov}\Big(\frac{\theta_i}{\sum_{r \in A_p} \theta_r}, \frac{\theta_j}{\sum_{r \in A_p} \theta_r} |\boldsymbol{x}\Big) &= -\frac{(\alpha_i + x_i)(\alpha_j + x_j)}{(\sum_{r \in A_p} (\alpha_r + x_r))^2 (\sum_{r \in A_p} (\alpha_r + x_r) + 1)}.
\end{aligned}$$

Let

$$\begin{aligned}
\beta_i &= x_{A_i} + \sum_{r \in A_i} (\alpha_r + x_r), \ i = 0, \ldots, m, \\
\beta_{00} &= \sum_{\ell=0}^{m} \{x_{A_\ell} + \sum_{r \in A_\ell} (\alpha_r + x_r)\},
\end{aligned}$$

where $x_{A_0} = 0$, for $i = 1, \dots, m$. Then, we have

$$\begin{aligned}
\mathrm{E}\Big(\sum_{r\in A_i} \theta_r|\boldsymbol{x}\Big) &= \frac{\beta_i}{\beta_{00}}, \\
\mathrm{Var}\Big(\sum_{r\in A_i} \theta_r|\boldsymbol{x}\Big) &= \frac{\beta_i(\beta_{00}-\beta_i)}{\beta_{00}^2(\beta_{00}+1)}, \\
\mathrm{Cov}\Big(\sum_{r\in A_i} \theta_r, \sum_{r\in A_j} \theta_r|\boldsymbol{x}\Big) &= -\frac{\beta_i\beta_j}{\beta_{00}^2(\beta_{00}+1)}.
\end{aligned}$$

Consequently, the posterior mean

$$E(\theta_i|\boldsymbol{x}) = \mathrm{E}\Big(\frac{\theta_i}{\sum_{r\in A_p}\theta_r}|\boldsymbol{x}\Big)\mathrm{E}\Big(\sum_{r\in A_p}\theta_r|\boldsymbol{x}\Big),$$

whereas the posterior covariance of θ_i and θ_j is given by

$$\begin{aligned}
\mathrm{Cov}(\theta_i,\theta_j|\boldsymbol{x}) &= \Big[\mathrm{Cov}\Big(\frac{\theta_i}{\sum_{r\in A_p}\theta_r}, \frac{\theta_j}{\sum_{r\in A_p}\theta_r}|\boldsymbol{x}\Big) + \mathrm{E}\Big(\frac{\theta_i}{\sum_{r\in A_p}\theta_r}|\boldsymbol{x}\Big)\mathrm{E}\Big(\frac{\theta_j}{\sum_{r\in A_p}\theta_r}|\boldsymbol{x}\Big)\Big] \\
&\quad \times\Big[\mathrm{Var}\Big(\sum_{r\in A_p}\theta_r|\boldsymbol{x}\Big) + \Big\{\mathrm{E}\Big(\sum_{r\in A_p}\theta_r|\boldsymbol{x}\Big)\Big\}^2\Big] \\
&\quad -\mathrm{E}\Big(\frac{\theta_i}{\sum_{r\in A_p}\theta_r}|\boldsymbol{x}\Big)\mathrm{E}\Big(\frac{\theta_j}{\sum_{r\in A_p}\theta_r}|\boldsymbol{x}\Big)\Big\{\mathrm{E}\Big(\sum_{r\in A_p}\theta_r|\boldsymbol{x}\Big)\Big\}^2,
\end{aligned}$$

Similarly, it can be shown that for $i \in A_p$, $j \in A_q$, and $p \neq q$, the posterior covariance of θ_i and θ_j equals

$$\mathrm{Cov}(\theta_i,\theta_j|\boldsymbol{x}) = \mathrm{E}\Big(\frac{\theta_i}{\sum_{r\in A_p}\theta_r}|\boldsymbol{x}\Big)\mathrm{E}\Big(\frac{\theta_j}{\sum_{r\in A_q}\theta_r}|\boldsymbol{x}\Big)\mathrm{Cov}\Big(\sum_{r\in A_p}\theta_r, \sum_{r\in A_q}\theta_r|\boldsymbol{x}\Big).$$

Appendix B: Posterior Distribution of k

Case I: Complete Data

Let $\boldsymbol{z} = (x_1, x_2, \dots, x_q)$ be the observed data where $x_i > 0$, $i = 1, \dots, q$. And let $\boldsymbol{x} = (x_1, x_2, \dots, x_q, 0, \dots, 0) = (x_1, x_2, \dots, x_k)$ with $k \geqslant q$. Assume the maximum number of categories is M. Let k be the true number of categories. We assume a flat prior for k. Let $\boldsymbol{\theta}_k = (\theta_1, \dots, \theta_k)$. Then, for $k \geqslant q$,

$$\pi(k|\boldsymbol{z}) \propto \pi(\boldsymbol{z}|k)\pi(k) \propto \int_{\mathcal{A}} \pi(\boldsymbol{x}|\boldsymbol{\theta}_k, k)\pi(\boldsymbol{\theta}_k|k)\pi(k)d\boldsymbol{\theta}_k,$$

where $\mathcal{A}$ is the support of $\boldsymbol{\theta}_k$. The prior probability density of $\boldsymbol{\theta}_k$ is set to be the Dirichlet distribution with $(\alpha_1, \dots, \alpha_k)$. Then, we have

$$\begin{aligned}
\int_{\mathcal{A}} \pi(\boldsymbol{x}|\boldsymbol{\theta}_k, k)\pi(\boldsymbol{\theta}_k|k)\pi(k)d\boldsymbol{\theta}_k &= \int_{\mathcal{A}} \Big\{ \frac{(\sum_{i=1}^{q} x_i + \sum_{i=q+1}^{k} 0)!}{x_1! \cdots x_q! 0! \cdots 0!} \prod_{i=1}^{q} \theta_i^{x_i} \prod_{i=q+1}^{k} \theta_i^{0} \\
&\quad \times \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \prod_{i=1}^{q} \theta_i^{\alpha_i - 1} \prod_{i=q+1}^{k} \theta_i^{\alpha_i - 1} \frac{1}{M} \Big\} d\boldsymbol{\theta}_k \\
&= \frac{1}{M} \frac{(\sum_{i=1}^{q} x_i)!}{x_1! \cdots x_q!} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \\
&\quad \times \frac{\Gamma(x_1 + \alpha_1)\cdots\Gamma(x_q + \alpha_q)\Gamma(\alpha_{q+1})\cdots\Gamma(\alpha_k)}{\Gamma(\sum_{i=1}^{q} x_i + \sum_{i=1}^{k} \alpha_i)} \\
&= \frac{1}{M} \frac{(\sum_{i=1}^{q} x_i)!(\sum_{i=1}^{k} \alpha_i - 1)!}{(\sum_{i=1}^{q} x_i + \sum_{i=1}^{k} \alpha_i - 1)!} \prod_{i=1}^{q} \Big\{ \frac{(x_i + \alpha_i - 1)!}{x_i!(\alpha_i - 1)!} \Big\}.
\end{aligned}$$

If $k < q$, then $\pi(k|\boldsymbol{z}) = 0$.

For the bartonella application, $\alpha_1 = \cdots = \alpha_k = 1$. Thus, for $k \geqslant q$ we obtain

$$\pi(k|\boldsymbol{z}) \propto \frac{1}{M} \frac{(\sum_{i=1}^{q} x_i)!(k-1)!}{(\sum_{i=1}^{q} x_i + k - 1)!} \prod_{i=1}^{q} \Big\{ \frac{x_i!}{x_i!} \Big\} \propto \frac{1}{M} \frac{(\sum_{i=1}^{q} x_i)!(k-1)!}{(\sum_{i=1}^{q} x_i + k - 1)!}.$$

Case 2: Incomplete Data

Let $\boldsymbol{z} = (x_1, x_2, \dots, x_q, x_{A_1}, \dots, x_{A_m})$ be the observed data where $x_i > 0$ and $x_{A_j} > 0$ for $i = 1, \dots, q$, $j = 1, \dots, m$. The symbols M and k are as defined in Case 1. A flat prior for k is assumed. Let $\boldsymbol{x} = (x_1, x_2, \dots, x_q, 0, \dots, 0, x_{A_1}, \dots, x_{A_m}) = (x_1, x_2, \dots, x_k, x_{A_1}, \dots, x_{A_m})$ with $k \geqslant q$.

$$\pi(k|\boldsymbol{z}) \propto \pi(\boldsymbol{x}|k)\pi(k) \propto \int_{\mathcal{A}} \pi(\boldsymbol{x}|\boldsymbol{\theta}_k, k)\pi(\boldsymbol{\theta}_k|k)\pi(k)d\boldsymbol{\theta}_k.$$

Let x_{A_0} be 0. Using the notations and results of Appendix A, we have

$$\begin{aligned}
\int_{\mathcal{A}} \pi(\boldsymbol{x}|\boldsymbol{\theta}_k, k)\pi(\boldsymbol{\theta}_k|k)\pi(k)d\boldsymbol{\theta}_k &= \int_{\mathcal{A}} \Big\{ \frac{(\sum_{i=1}^{k} x_i + \sum_{i=1}^{m} x_{A_i})!}{x_1! \cdots x_k! x_{A_1}! \cdots x_{A_m}!} \prod_{i=1}^{k} \theta_i^{x_i} \prod_{i=1}^{m} \big(\sum_{j \in A_i} \theta_j \big)^{x_{A_i}} \\
&\quad \times \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \frac{1}{M} \Big\} d\boldsymbol{\theta}_k \\
&= \frac{1}{M} \frac{(\sum_{i=1}^{q} x_i + \sum_{i=1}^{m} x_{A_i})!}{x_1! \cdots x_q! x_{A_1}! \cdots x_{A_m}!} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \\
&\quad \times \int_{\mathcal{A}} \Big\{ \prod_{i \in A_0} \big(\frac{\theta_i}{\sum_{j \in A_0} \theta_j} \big)^{\alpha_i + x_i - 1} \cdots \prod_{i \in A_m} \big(\frac{\theta_i}{\sum_{j \in A_m} \theta_j} \big)^{\alpha_i + x_i - 1}
\end{aligned}$$

$$
\begin{aligned}
&\times \prod_{i=0}^{m} \Big(\sum_{j \in A_i} \theta_j \Big)^{x_{A_i} + \sum_{j \in A_i} (\alpha_j + x_j - 1)} \Big\} d\boldsymbol{\theta}_k \\
= \; & \frac{1}{M} \frac{(\sum_{i=1}^{q} x_i + \sum_{i=1}^{m} x_{A_i})!}{x_1! \cdots x_q! x_{A_1}! \cdots x_{A_m}!} \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \\
&\times \prod_{i=0}^{m} \Big\{ \frac{\prod_{j \in A_i} \Gamma(\alpha_j + x_j)}{\Gamma(\sum_{j \in A_i} (\alpha_j + x_j))} \Big\} \frac{\prod_{i=0}^{m} \Gamma(\sum_{j \in A_i} (\alpha_j + x_j) + x_{A_i})}{\Gamma(\sum_{i=0}^{m} \{ \sum_{j \in A_i} (\alpha_j + x_j) + x_{A_i} \})}.
\end{aligned}
$$

If $k < q$, then $\pi(k|\boldsymbol{z}) = 0$.

In the bartonella application, $A_0 = \{1\}$, $A_1 = A_0^c$, and $\alpha_1 = \cdots = \alpha_k = 1$. Thus, for $k \geqslant q$, we obtain

$$
\pi(k|x_1, \ldots, x_q, x_{A_1}) \;\; \propto \;\; \frac{1}{M} \frac{(k-1)!(\sum_{i=1}^{q} x_i + x_{A_1})!}{(\sum_{i=1}^{q} x_i + x_{A_1} + k - 1)!} \frac{(\sum_{i=2}^{q} x_i + x_{A_1} + k - 2)!}{x_{A_1}!(\sum_{i=2}^{q} x_i + k - 2)!}.
$$

Appendix C: Bayes Factors

Case 1: Complete Data

Assume we have n sites, and k_s denotes the true number of categories for site s. Let $\mathbf{X}_s = (X_{s,1}, X_{s,2}, \ldots, X_{s,q_s}, 0, \ldots, 0) = (X_{s,1}, X_{s,2}, \ldots, X_{s,k_s})$ be a vector of counts at site s where $X_{s,i} > 0$ for $i = 1, \ldots, q_s$. And let $\boldsymbol{\theta}_s = (\theta_{s,1}, \theta_{s,2}, \ldots, \theta_{s,k_s})$ be the corresponding cell probabilities, $s = 1, \ldots, n$. Let q be the number of observed and distinct categories in all sites. Note that $q \geqslant q_s$ for $s = 1, \ldots, n$. Assume the maximum number of observed and unobserved categories is M. We want to test the following hypothesis:

$$
\begin{aligned}
H_0 \;\; &: \;\; \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_n = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k), \quad \text{versus} \\
H_1 \;\; &: \;\; H_0 \text{ is not true.}
\end{aligned}
$$

Let the prior probabilities of the two hypotheses be $\pi(H_0) = \frac{1}{2}$ and $\pi(H_1) = \frac{1}{2}$. The Bayes factor equals

$$
\begin{aligned}
B \;\; &= \;\; \frac{\pi(H_0|\mathbf{X}_1, \ldots, \mathbf{X}_n)}{\pi(H_1|\mathbf{X}_1, \ldots, \mathbf{X}_n)} = \frac{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_n|H_0)\pi(H_0)/\pi(\mathbf{X}_1, \ldots, \mathbf{X}_n)}{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_n|H_1)\pi(H_1)/\pi(\mathbf{X}_1, \ldots, \mathbf{X}_n)} \\
&= \;\; \frac{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_n|H_0)}{\pi(\mathbf{X}_1, \ldots, \mathbf{X}_n|H_1)}.
\end{aligned}
$$

Under H_0, the prior distribution of the common $\boldsymbol{\theta}$ given k is set to be Dirichlet distribution with parameter vector $(\alpha_{k,1}, \ldots, \alpha_{k,k})$. Similarly, under H_1, the prior distribution of $\boldsymbol{\theta}_s$'s given the k_s's are jointly independent, with marginal Dirichlet distribution with parameter vector $(\alpha_{k_s,1}, \ldots, \alpha_{k_s,k_s})$, $s = 1, \ldots, n$; The prior distribution of the k_s's are also jointly independent, with k_s being uniform between 1 and M. By the independence of $\mathbf{X}_s$'s given k and $\boldsymbol{\theta}$,

$$\begin{aligned}\pi(\mathbf{X}_1, \ldots, \mathbf{X}_n | H_0) &= \sum_{k=1}^{M} \Big[\int \pi(\mathbf{X}_1, \ldots, \mathbf{X}_n | \boldsymbol{\theta}, k, H_0)\pi(\boldsymbol{\theta}|k, H_0)\pi(k|H_0)d\boldsymbol{\theta}\Big] \\ &= \sum_{q \leqslant k \leqslant M} \Big[\int \prod_{s=1}^{n} \Big\{\pi(\mathbf{X}_s|\boldsymbol{\theta}, k, H_0)\Big\}\pi(\boldsymbol{\theta}|k, H_0)d\boldsymbol{\theta}\Big]\pi(k|H_0).\end{aligned}$$

Similarly,

$$\begin{aligned}&\pi(\mathbf{X}_1, \ldots, \mathbf{X}_n | H_1) \\ =\ & \sum_{k_1=1}^{M} \cdots \sum_{k_n=1}^{M} \Big[\int \cdots \int \{\prod_{s=1}^{n} \pi(\mathbf{X}_s|\boldsymbol{\theta}_s, k_s, H_1)\pi(\boldsymbol{\theta}_s|k_s, H_1)\pi(k_s|H_1)\}d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_s\Big] \\ =\ & \sum_{q_1 \leqslant k_1 \leqslant M} \cdots \sum_{q_n \leqslant k_n \leqslant M} \Big[\prod_{s=1}^{n} \Big\{\int \pi(\mathbf{X}_s|\boldsymbol{\theta}_s, k_s, H_1)\pi(\boldsymbol{\theta}_s|k_s, H_1)\pi(k_s|H_1)d\boldsymbol{\theta}_s\Big\}\Big] \\ =\ & \prod_{s=1}^{n} \Big[\sum_{q_s \leqslant k_s \leqslant M} \Big\{\int \pi(\mathbf{X}_s|\boldsymbol{\theta}_s, k_s)\pi(\boldsymbol{\theta}_s|k_s)d\boldsymbol{\theta}_s\Big\}\pi(k_s|H_1)\Big].\end{aligned}$$

Note that in the first equality the region over which the θ's are identical has zero contribution to the integrals, and hence need not be removed from the integration. It follows from the results derived in Appendices A and B that, for $(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, the Bayes factor equals

$$\begin{aligned}B &= \frac{\pi(\mathbf{x}_1, \ldots, \mathbf{x}_n | H_0)}{\pi(\mathbf{x}_1, \ldots, \mathbf{x}_n | H_1)} \\ &= \frac{1}{M} \sum_{q \leqslant k \leqslant M} \Big[\prod_{s=1}^{n} \Big\{\frac{(\sum_{i=1}^{q} x_{s,i})!}{x_{s,1}! \cdots x_{s,q}!}\Big\} \frac{(\sum_{i=1}^{k} \alpha_{k,i} - 1)!}{\prod_{i=1}^{k} (\alpha_{k,i} - 1)!} \frac{\prod_{i=1}^{k} (\sum_{s=1}^{n} x_{s,i} + \alpha_{k,i} - 1)!}{(\sum_{s=1}^{n} \sum_{i=1}^{q} x_{s,i} + \sum_{i=1}^{k} \alpha_{k,i} - 1)!}\Big] \Big/ \\ &\quad \prod_{s=1}^{n} \Big[\sum_{q_s \leqslant k_s \leqslant M} \Big\{\frac{1}{M} \frac{(\sum_{i=1}^{q_s} x_{s,i})!(\sum_{i=1}^{k_s} \alpha_{k_s,i} - 1)!}{(\sum_{i=1}^{q_s} x_{s,i} + \sum_{i=1}^{k_s} \alpha_{k_s,i} - 1)!} \prod_{i=1}^{q_s} \Big(\frac{(x_{s,i} + \alpha_{k_s,i} - 1)!}{x_{s,i}!(\alpha_{k_s,i} - 1)!}\Big)\Big\}\Big].\end{aligned}$$

A large B serves as strong evidence for H_0 against H_1.

In the bartonella application, assuming a flat prior under H_0 and H_1. Then, the Bayes factor equals

$$B = \frac{\frac{1}{M}\sum_{q\leqslant k\leqslant M}\Big[\prod_{s=1}^{n}\Big\{\frac{(\sum_{i=1}^{q} x_{s,i})!}{x_{s,1}!\cdots x_{s,q}!}\Big\}\frac{(k-1)!\prod_{i=1}^{q}(\sum_{s=1}^{n} x_{s,i})!}{(\sum_{s=1}^{n}\sum_{i=1}^{q} x_{s,i}+k-1)!}\Big]}{\prod_{s=1}^{n}\Big[\sum_{q_s\leqslant k_s\leqslant M}\Big\{\frac{1}{M}\frac{(\sum_{i=1}^{q_s} x_{s,i})!(k_s-1)!}{(\sum_{i=1}^{q_s} x_{s,i}+k_s-1)!}\Big\}\Big]}.$$

Case 2: Incomplete Data

The definition of q is the same as that of Case 1. For simplicity, assume all sites have the same disjoint sets of missingness, i.e., $(A_{s,1},\ldots,A_{s,m}) = (A_1,\ldots,A_m)$, $s = 1,\ldots,m$. Let x_{s,A_0} be 0, $s = 1,\ldots,n$, as assumed earlier. Thanks to the results in Appendices A and B,

$$\begin{aligned}
&\pi(\mathbf{x}_1,\ldots,\mathbf{x}_n|H_0)\\
=\;& \sum_{q\leqslant k\leqslant M}\Big[\int\prod_{s=1}^{n}\Big\{\pi(\mathbf{X}_s|\boldsymbol{\theta},k,H_0)\Big\}\pi(\boldsymbol{\theta}|k,H_0)d\boldsymbol{\theta}\Big]\pi(k|H_0)\\
=\;& \frac{1}{M}\sum_{q\leqslant k\leqslant M}\Big[\prod_{s=1}^{n}\Big\{\frac{(\sum_{i=1}^{q} x_{s,i}+\sum_{i=1}^{m} x_{s,A_i})!}{x_{s,1}!\cdots x_{s,q}!x_{s,A_1}!\cdots x_{s,A_m}!}\Big\}\frac{\Gamma(\sum_{i=1}^{k}\alpha_{k,i})}{\prod_{i=1}^{k}\Gamma(\alpha_{k,i})}\prod_{i=0}^{m}\Big\{\frac{\prod_{j\in A_i}\Gamma(\alpha_{k,j}+\sum_{s=1}^{n} x_{s,j})}{\Gamma(\sum_{j\in A_i}(\alpha_{k,j}+\sum_{s=1}^{n} x_{s,j}))}\Big\}\\
&\times\frac{\prod_{i=0}^{m}\Gamma(\sum_{j\in A_i}(\alpha_{k,j}+\sum_{s=1}^{n} x_{s,j})+\sum_{s=1}^{n} x_{s,A_i})}{\Gamma(\sum_{i=0}^{m}\{\sum_{j\in A_i}(\alpha_{k,j}+\sum_{s=1}^{n} x_{s,j})+\sum_{s=1}^{n} x_{s,A_i}\})}\Big],
\end{aligned}$$

and

$$\begin{aligned}
&\pi(\mathbf{x}_1,\ldots,\mathbf{x}_n|H_1)\\
=\;& \prod_{s=1}^{n}\Big[\sum_{q_s\leqslant k_s\leqslant M}\Big\{\int\pi(\mathbf{x}_s|\boldsymbol{\theta}_s,k_s)\pi(\boldsymbol{\theta}_s|k_s)\pi(k_s)d\boldsymbol{\theta}_s\Big\}\Big]\\
=\;& \prod_{s=1}^{n}\Big[\sum_{q_s\leqslant k_s\leqslant M}\Big\{\frac{1}{M}\frac{(\sum_{i=1}^{q_s} x_{s,i}+\sum_{i=1}^{m} x_{s,A_i})!}{x_{s,1}!\cdots x_{s,q_s}!x_{s,A_1}!\cdots x_{s,A_m}!}\frac{\Gamma(\sum_{i=1}^{k_s}\alpha_{k_s,i})}{\prod_{i=1}^{k_s}\Gamma(\alpha_{k_s,i})}\prod_{i=0}^{m}\Big\{\frac{\prod_{j\in A_i}\Gamma(\alpha_{k_s,j}+x_{s,j})}{\Gamma(\sum_{j\in A_i}(\alpha_{k_s,j}+x_{s,j}))}\Big\}\\
&\times\frac{\prod_{i=0}^{m}\Gamma(\sum_{j\in A_i}(\alpha_{k_s,j}+x_{s,j})+x_{s,A_i})}{\Gamma(\sum_{i=0}^{m}\{\sum_{j\in A_i}(\alpha_{k_s,j}+x_{s,j})+x_{s,A_i}\})}\Big\}\Big].
\end{aligned}$$

For the case that $m = 1$, the Bayes factor can be readily computed upon noting that

$$\begin{aligned}
&\pi(\mathbf{x}_1,\ldots,\mathbf{x}_n|H_0)\\
=\;& \frac{1}{M}\sum_{q\leqslant k\leqslant M}\Big[\prod_{s=1}^{n}\Big\{\frac{(\sum_{i=1}^{q} x_{s,i}+x_{s,A_1})!}{x_{s,1}!\cdots x_{s,q}!x_{s,A_1}!}\Big\}\frac{(k-1)!\prod_{i=1}^{q}(\sum_{s=1}^{n} x_{s,i})!}{(\sum_{i=2}^{q}\sum_{s=1}^{n} x_{s,i}+k-2)!}\\
&\times\frac{(\sum_{i=2}^{q}\sum_{s=1}^{n} x_{s,i}+\sum_{i=1}^{n} x_{s,A_1}+k-2)!}{(\sum_{i=1}^{q}\sum_{s=1}^{n} x_{s,i}+\sum_{i=1}^{n} x_{s,A_1}+k-1)!}\Big],\\
&\pi(\mathbf{x}_1,\ldots,\mathbf{x}_n|H_1)\\
=\;& \prod_{s=1}^{n}\Big[\sum_{q_s\leqslant k_s\leqslant M}\Big\{\frac{1}{M}\frac{(k_s-1)!(\sum_{i=1}^{q} x_{s,i}+x_{s,A_1})!}{(\sum_{i=1}^{q} x_{s,i}+x_{s,A_1}+k_s-1)!}\frac{(\sum_{i=2}^{q} x_{s,i}+x_{s,A_1}+k_s-2)!}{x_{s,A_1}!(\sum_{i=2}^{q} x_{s,i}+k_s-2)!}\Big\}\Big].
\end{aligned}$$

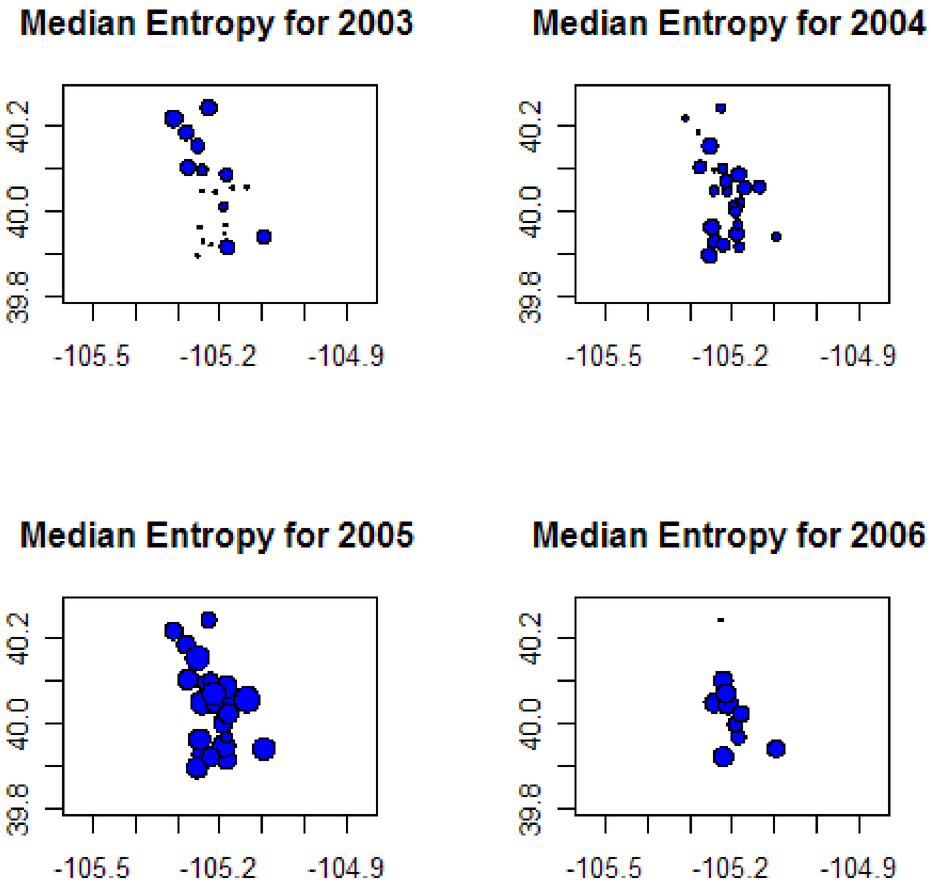

Figure 1. Spatio–temporal distribution of the median posterior Shannon entropy for $M = 13$. Each circle is drawn with center at the (latitude, logitude) of the site and area proportional to the median posterior Shannon entropy.

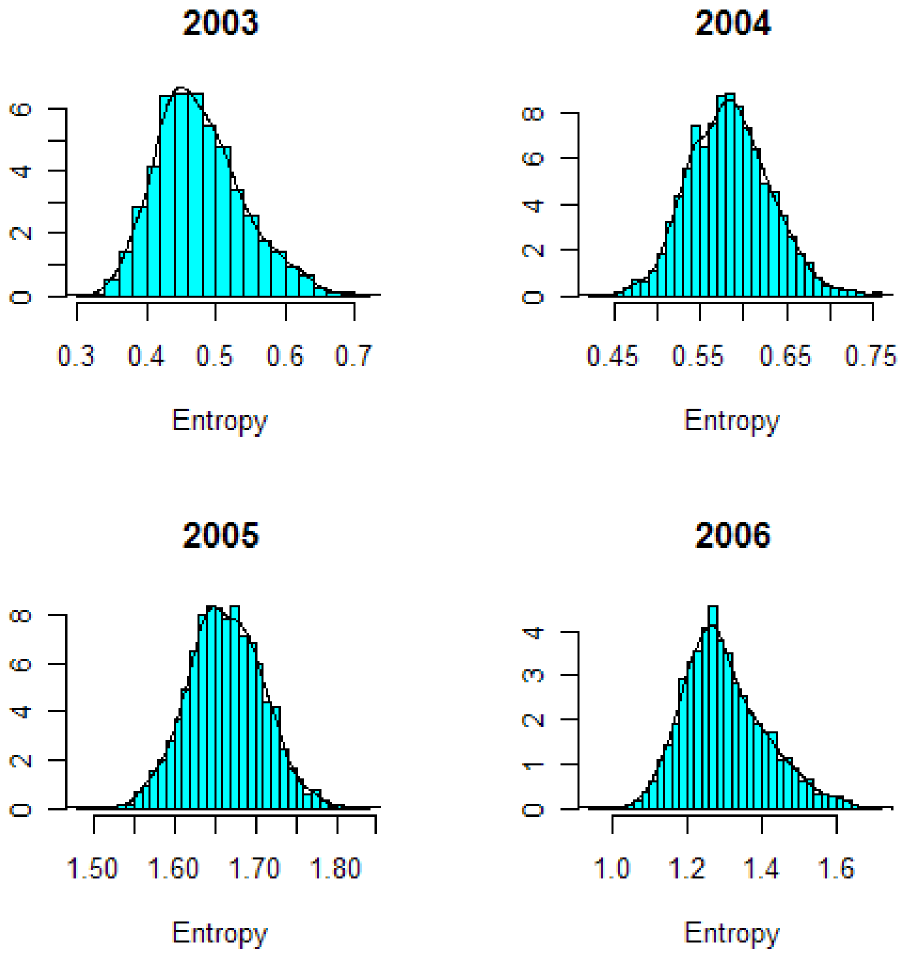

Figure 2. Histogram of Bootstrap for $M = 13$

Table 1
Posterior Means of Shannon Entropy

	M = 10	M = 11	M = 12	M = 13	M = 14
2003	0.470	0.472	0.475	0.476	0.474
2004	0.581	0.583	0.583	0.583	0.583
2005	1.662	1.663	1.662	1.663	1.662
2006	1.293	1.297	1.299	1.300	1.303

Table 2
Posterior 95% intervals of annual Shannon entropy for $M = 13$

	Posterior Mean	Posterior 95% Prediction Interval
2003	0.476	(0.372 0.619)
2004	0.583	(0.498 0.679)
2005	1.663	(1.573 1.755)
2006	1.300	(1.120 1.552)