An Ecological Latent Health Factor Index via a Random-Effects Model for Taxa Richness and Composition

University of Waterloo
Department of Statistics and Actuarial Science
2006 Working Paper Series #2006-02
Finalized June 1, 2007.

Grace S. Chiu 1 (corresponding author), Peter Guttorp,2 Shahedul A. Khan,1 Jun Liang,3 and Anton H. Westveld 4

1 Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.
2 Department of Statistics, University of Washington, Seattle, WA 98195, USA.
3 Canadian Institute for Health Information, 90 Eglinton Avenue East, Toronto, Ontario, M4P 2Y3, Canada.
4 Center for Empirical Research in the Law / Applied Statistics and Computation, Campus Box 1120, One Brookings Drive, Washington University, St. Louis, MO 63130, USA.
Author Footnote

Grace Chiu is Assistant Professor (Email: gchiu@uwaterloo.ca), Shahedul Khan is graduate student, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1; Peter Guttorp is Professor and Chair, Department of Statistics, University of Washington, Seattle, WA 98195; Jun Liang is Data Analyst, Canadian Institute for Health Information, 90 Eglinton Avenue East, Toronto, ON M4P 2Y3; and Anton Westveld is Postdoctoral Fellow, Center for Empirical Research in the Law / Applied Statistics and Computation, Campus Box 1120, One Brookings Drive, Washington University, St. Louis, MO 63130. This research in progress is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant to G. Chiu. The corresponding author thanks Professor Michael Dowd, Department of Mathematics and Statistics, Dalhousie University, for stimulating discussions on MCMC techniques and post-normal science.
Abstract and Keywords

We propose a statistical-model-based index as a powerful alternative to existing ecological health indices, such as the index of biotic integrity (IBI) and the stream health index for the Puget Sound Lowland (SHIPSL). Our *latent health factor index* (LHFI) is obtained by estimating an unobservable *health factor* term in a mixed-effects analysis of (co)variance regression that directly models the relationship among indicator variables (or *metrics*) and health. Various formulations of Poisson and logistic regression are used to construct the index in a Bayesian hierarchical framework, whose implementation requires Markov chain Monte Carlo techniques. Results indicate that the health index developed by our approach is no less effective than existing indices in reflecting underlying health conditions, and is far superior in its scientific integrity and versatility. The latter is due to the fact that (1) subject-matter expertise enters the LHFI developing scheme solely through the choice of metrics (as opposed to the conventional approach in which subjective calibration or *scoring* of metrics and definition of the index leads to inevitable loss of statistical information); (2) statistical properties of the LHFI are fully tractable based on the model itself via the index posterior distribution; (3) conditions from multiple geographical and/or temporal domains can be studied simultaneously without the need for painstaking metric recalibration to adjust for different spatio-temporal scales; and (4) when formulated as a hierarchical model where latent health is explained by covariates, our approach provides resource managers and policy makers with a formal quantitative mechanism for prioritizing the control of crucial factors that may impact ecosystem health. Furthermore, unlike other model-based approaches which often sacrifice interpretability to gain statistical integrity, our approach yields a scalar-valued index that is readily interpretable by the practitioner. Finally, our approach is applicable to developing health indices not only in ecological contexts, but also in medical contexts, for instance.

**Keywords:** analysis of covariance; analysis of variance; Bayesian models; hierarchical models; ecological health; ecosystem health; generalized linear mixed models; latent factor; metric scoring; mixed effects; multimetric index; random effects
1 Introduction

The development of conventional aquatic health indices such as the *benthic index of biotic integrity* (B-IBI) (Kerans and Karr 1994) and its variants (e.g. McCormick *et al.* 2001) involves studying and combining the ability of indicator variables, or *metrics*, to reflect the underlying health conditions of field sites in a single spatial domain. The general disadvantages of relying on conventional indices such as these have been well documented (e.g. Steedman and Regier 1990, Chiu and Guttorp 2006). Statistically speaking, the conventional approach may be regarded as *ad hoc* due to a high degree of arbitrariness involved in several stages of quantifying qualitative features. For instance, Stage 1 involves testing / analyzing / calibrating / validating each of a potentially enormous pool of metrics using existing or training data. In Stage 2, a final reduced set of metrics is agreed upon for use in forming the index. Stage 3 involves devising a standardization or *scoring* scheme so that all selected metrics share the same scale. This scoring is often geography-specific, and to this date, no obvious strategy exists for ensuring that metrics calibrated for one spatial domain would effectively reflect ecosystem health of another domain. Stage 4 involves deciding on a weighting scheme for the *metric scores* to form the scalar-valued multimetric index. Equal weighting is common, leading to a multimetric index that is the sum of all metric scores. However, each metric reflects a different aspect of health, while all of them include an overlap in informational content that is not easily quantifiable. Therefore, what scoring and weighting schemes may effectively reflect the overall health conditions is traditionally an open-ended question. Moreover, existing criteria for all four stages are often created through devising a numerical scheme based on qualitative ideas of health. Similar practices are common for assessing health for other ecosystems (e.g. see Jørgensen *et al.* 2005). Overall, this *ad hoc* approach poses extreme challenges to proper statistical assessment of the indices and comparison of ecosystem health from different spatial and/or temporal domains. Despite the issues, such indices are popular in policy-making contexts due to their structural simplicity, interpretability, and high biological content in the form of subject-matter expertise from numerous scientists involved in all four stages of the index development process.
To address the statistical inadequacies, Chiu and Guttrop (2006) propose the SHIPSL scheme for pooling metrics to form an index that is also scalar-valued. When compared to the conventional approach, this scheme reduces arbitrariness and therefore improves statistical tractability of the resulting index. However, the SHIPSL scheme retains some degree of qualitative involvement. Other environmetricians (Billheimer et al. 1997; Bunea et al. 1999) have directly modeled the metrics, thereby removing all intermediate stages, each of which masks valuable information on health. These statistical models, such as state-space models and graphical models, are often expressed in a multivariate framework which can effectively describe underlying conditions of an ecosystem; however, they may not be immediately useful to resource managers due to the complex messages embedded in a multivariate system.

In this article, we investigate an entirely new approach for ecological health assessment that combines the statistical integrity of model-based techniques and the interpretability of existing indices in policy making contexts. For metrics that have been identified to be informative, we model the interdependence among them via a multivariate response that is regressed on univariate factors, one of which is latent health. This single-number latent, unobservable factor can be estimated in a Bayesian hierarchical framework, thereby yielding a numerical assessment of ecological health. Other observable covariates may be included, such as demographic variables associated with the sampled field sites. The resulting latent health factor index (LHFI) can then be compared to existing indices to determine if both types of indices contain similar information about the underlying health conditions of the ecosystem. If so, the statistically based LHFI would be preferred in practice, since (1) the process undertaken to define the index is based almost entirely on standard modeling principles (hence, is much less arbitrary), (2) its performance is highly tractable in the statistical sense, (3) it is much less sensitive to random variability (field noise) because it removes all the intermediate ad hoc stages that mask valuable information, and (4) it is readily interpretable by resource managers. This last point is further enhanced when covariates appear

---

1In practice, one may adopt a dimension reduction modeling approach for variable selection. For instance, see Ding (2000). In the current article, we do not discuss what pertains to an appropriate variable selection scheme.
in the model in a latent regression to explain the health factor. In this case, the significance of their impact on health can be statistically assessed and classified, thus readily providing policy makers with unambiguous guidelines on how conservation measures may be prioritized. Indeed, ranking sites according to their LHFI values and classifying factors that have impact on health can be achieved simultaneously in one step from fitting the model.

We demonstrate the merits of our methodology in Sections 3.1, 5.4, and 6.2 by applying it to the 1997 Puget Sound Lowland taxonomic data that appear in Chiu and Guttorp (2006). SHIPSL and B-IBI values for these data are compared to those of our LHFI, and their statistical and biological performance contrasted.

To address the issue of comparing ecosystems over different spatial domains, we propose the addition of a spatial effect term in the latent factor model, and similarly for different temporal domains. We believe our methodology to be the first to overcome, in a scientifically sound manner, the age-old difficulty encountered in the assessment of ecosystem health over a variety of geographical locations and/or time periods. This approach is discussed in Section 7.

2 Modeling Preliminaries

Typically, benthic taxonomic data are collected by inserting a fixed-size shovel into the soil, separating the animals from the collected soil, then categorizing each animal into one of many categories; there are 80 for the Puget Sound Lowland (PSL). For each sampled site, the distribution of these 80 counts reveals the underlying “health” of the site. In general, a healthy site should have many non-zero counts, and no particular non-zero count should dominate the rest.

Instead of investigating the distribution of the 80 counts, a naive approach is to examine the sample size (total number of animals coming out of the shovel). Although not always biologically reasonable, one may expect healthier sites to yield larger sample sizes. We may also wish to model metrics whose values are computed based on the 80 counts. For the PSL, biologists have identified 10 useful metrics, 7 of which describe taxa richness (counts), and 3 describe relative abundance (percentages). (See Chiu and Guttorp 2006, Table 1 for a list
of the metrics.) Note that all 10 metrics are highly correlated due to their definitions.

To demonstrate the idea of a model-based LHFI, first consider modeling the sample size as a function of the unobservable *health*. For the PSL, each field site was sampled three times. Assume that the three replicates are independent. Let \( N_{ik} \) denote the sample size of the \( k \)th replicate sampled from site \( i \). Then, a possible mixed-effects analysis of variance (ANOVA) model is

\[
\ln N_{ik} = H_i + \varepsilon_{ik}, \quad H_i \overset{\text{iid}}{\sim} N(\mu, \sigma_H^2), \quad \varepsilon_{ik} \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon) \tag{1}
\]

where \( H_i \)'s and \( \varepsilon_{ik} \)'s are independent. Here, \( H_i - \mu \) is the random effect of site \( i \) on the overall mean log-sample size, \( \mu \). In other words, \( H_i \) can be viewed as the *latent health factor* associated with site \( i \). Note that the log-transformation is necessary for modeling sample size (count) data with a Gaussian model. A popular way to estimate the latent factor in a model such as (1) is via a Bayesian hierarchical framework, in which the parameters \( \mu, \sigma_H^2, \) and \( \sigma^2_\varepsilon \) are random variables with their own prior distributions. Hence, it becomes a strictly random-effects model, and can be formulated as follows:

\[
\ln N_{ik} = H_i + \varepsilon_{ik}, \quad [H_i|\mu, \sigma_H^2] \overset{\text{iid}}{\sim} N(\mu, \sigma_H^2), \quad [\varepsilon_{ik}|\sigma^2_\varepsilon] \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon) \tag{2}
\]

with prior distributions

\[
\mu \sim N(c_1, c_2), \quad \sigma_H^2 \sim \text{inverse-gamma}(c_3, c_4), \quad \sigma^2_\varepsilon \sim \text{inverse-gamma}(c_5, c_6) \tag{3}
\]

where \( c_1, \ldots, c_6 \) are chosen to reflect *a priori* notions of the parameter values.

An alternative model, also in the Bayesian framework, may be expressed in the form of a Poisson regression generalized linear mixed model (GLMM):

\[
[N_{ik}|H_i] \overset{\text{iid}}{\sim} \text{Poisson}(e^{H_i}), \quad [H_i|\alpha, \sigma_H^2] \overset{\text{iid}}{\sim} N(\alpha, \sigma_H^2), \tag{4}
\]

with prior distributions

\[
\alpha \sim N(c_1, c_2), \quad \sigma_H^2 \sim \text{inverse-gamma}(c_3, c_4) \tag{5}
\]
Note that (4) automatically addresses possible overdispersion of the sample size data, since

\[
E[N_{ik}] = E[E(N_{ik}|H_i)] = E(e^{H_i})
\]
\[
Var[N_{ik}] = E[Var(N_{ik}|H_i)] + Var[E(N_{ik}|H_i)] = E(e^{H_i}) + Var(e^{H_i}) \geq E[N_{ik}]
\]

For either (2)–(3) or (4)–(5), samples from the posterior distribution of each random effect \(H_i\) may be drawn by a Markov chain Monte Carlo (MCMC) algorithm. To this end, a standard Bayesian estimation software such as BUGS may be employed.\(^2\) The actual value of the LHFI for each site may be taken as the posterior mean for \(H_i\).

3 An LHFI via a Random-Effects Model for Taxa Richness

The models above demonstrate the notion of an LHFI based on the sample size. However, none of them utilizes all the information readily available in the taxonomic data. For the PSL, one improvement is to consider the 7 taxa richness count metrics in a framework similar to (4)–(5).

Let \(Y_{ijk}\) be the value of the \(j\)th metric for the \(i\)th site’s \(k\)th replicate, \(j=1,\ldots,7\). A possible model is

\[
[Y_{ijk}|\nu_{ij}] \sim \text{Poisson}(e^{\nu_{ij}}), \quad \nu_{ij} = H_i + \beta_j, \quad [H_i|\alpha, \sigma^2_H] \sim \text{N}(\alpha, \sigma^2_H), \quad (6)
\]

where \(\sum_j \beta_j = 0\), with prior distributions

\[
\alpha \sim \text{N}(c_1, c_2), \quad \beta_2, \ldots, \beta_7 \sim \text{N}(0, c_3), \quad \sigma^2_H \sim \text{inverse-gamma}(c_4, c_5) \quad (7)
\]

Here, we may regard \(\beta_j\) as the effect of variable \(j\) on the overall mean \(\alpha\). Thus, \(\beta_j\)’s account somewhat for dependence among metrics. Later, in Section 5, we shall improve the model to more appropriately represent the dependence structure among metrics.\(^2\)

\(^2\)All analyses in this article were performed on either WinBUGS version 1.4.1 or OpenBUGS version 2.2.0.
Note that in a non-Bayesian framework, $\alpha$ and $\beta_j$’s would be considered fixed effects. In the context of developing an ecological health index, the notion of fixed $\beta_j$’s seems to be appropriate, since the ecologists’ decision to use the particular metrics for forming the index was not made at random, but rather through a painstaking evaluation procedure. In contrast, field sites, i.e. different locations along the same stream for many streams, are chosen in a somewhat random manner; therefore, it seems reasonable to consider $H_i$’s as random effects. However, in Section 4, we discuss why one may prefer to also model the metric effects as random.

3.1 Application to the 1997 PSL Data

We apply (6)–(7) to the 1997 PSL taxonomic data as appear in Chiu and Guttorp (2006).

For the hyperparameters in (7), we take $(c_1, \ldots, c_5)$ to be (a) $(3, 10, 10, 1, 1)$ and (b) $(3, 100, 100, 0.5, 2)$. We justify the choice of hyperparameter values as follows.

For PSL taxonomic data, an animal in the sample may belong to one of 80 taxa. One of the metrics is total number of taxa, which cannot exceed 80. None of the other 6 taxa richness metrics under consideration can exceed the total taxa richness itself. Hence, 80 is an upper bound for all 7 metrics. Now, we consider that typically, no more than half of the taxa can appear simultaneously in a single sample. Due to the logarithmic link function of a Poisson regression model, it seems reasonable to take $c_1 = E(\alpha) \approx \ln 40 \approx 3.7 \approx 3$. For $c_2 = Var(\alpha)$ and $c_3 = Var(\beta_j)$, we take values that lead to diffuse distributions, as we have no definitive prior notions of the variability of $\alpha$ and $\beta_j$’s. Both 10 and 100 seem reasonable for the log-scale of taxa richness values. For the same reason, we take $(c_4, c_5)$ to be $(1,1)$ and $(0.5,2)$, respectively, so that the distribution of $\sigma_H^2$ is diffuse.

The PSL data with the above hyperparameter values were analyzed using BUGS. Note that the formulation of (6) utilizes hierarchical centering which often reduces MCMC mixing problems (Gelfand et al., 1995). In order to obtain samples from the posterior distribution of each of $\alpha, \beta_2, \ldots, \beta_7, H_1, \ldots, H_{18}$ and $\sigma_H^2$, two different chains were run based on different starting values. Each of the two chains for either Set (a) or (b) consists of 10,000 unthinned samples, not including a burn-in of 5,000 iterations for the former, and 8,000 for
the latter. The empirical posterior distribution for each parameter in both sets appears to be
smooth, and convergence criteria are reasonably met.

The upper-left 4×4 panels of Figure 1 show the comparison between the resulting LHFI (posterior mean of $H_i$ from both chains combined), B-IBI, and SHIPSL, in the form of scatterplots. We see that LHFI(a) and LHFI(b) are almost identical. This indicates that the different choices of hyperparameters virtually has no effect on the LHFI values. Thus, we need to only consider the performance of, say, LHFI(a) in its ability to reflect underlying health conditions. To this end, we consider the scatterplot between LHFI and either B-IBI or SHIPSL. It shows a high positive correlation (0.82 and 0.91, respectively), although the relationship is curvilinear. (In comparison, the correlation between B-IBI and SHIPSL is 0.97.) The curvature can be explained by the non-linearity of the model that produces LHFI, whereas both B-IBI and SHIPSL are linear combinations of the metric scores. Nonetheless, the high positive correlation indicates that LHFI is no less informative than the other two indices about the field sites’ health conditions. Note that the LHFI under consideration excludes 3 metrics that are included in both B-IBI and SHIPSL. Later, in Section 6, we propose an improved model that considers all 10 metrics, which undoubtedly improves the informational content of the LHFI.

4 Issues on Modeling Taxa Richness

Although our primary interest is not in $\sigma_H$, it is interesting to note that its posterior distribution indeed differs slightly between Sets (a) and (b), with the two posterior means averaging to be about 0.58. A possible explanation is that Model (6)–(7) does not fully address (i) the variance structure of the metric counts, nor (ii) the covariance structure among sites and among metrics.

To see this, let us first consider (i) via a Gaussian analog of (6):

$$W_{ijk} \equiv \ln Y_{ijk} = H_i + \beta_j + \varepsilon_{ijk}, \quad H_i \overset{iid}{\sim} N(\alpha, \sigma_H^2), \quad \sum_j \beta_j = 0, \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2),$$

$3$The posterior distributions for LHFI(a) and (b) differ only in the form of a slightly larger dispersion for (a).
Thus, we have $\text{Var}(W_{ijk}) = \sigma_H^2 + \sigma^2$, which is constant over $i$ and $j$. In contrast, Figure 2 demonstrates that $\text{Var}(Y_{ijk})$ — and hence, $\text{Var}(W_{ijk})$ — can differ vastly over $j$. Although taking $\varepsilon_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2_j)$ easily handles this issue in the log-linear formulation here, it has no counterpart in the Poisson GLMM formulation.

An alternative log-linear model that also addresses the heteroskedasticity of the response due to metric is a Gaussian random-effects analog of (6):

$$W_{ijk} \equiv \ln Y_{ijk} = H_i + \beta_j + \varepsilon_{ijk}, \quad H_i \overset{\text{iid}}{\sim} N(\alpha, \sigma^2_H), \quad \beta_j \overset{\text{iid}}{\sim} N(0, \sigma^2_j), \quad \varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2_\varepsilon) \quad (9)$$

which differs from (8) in the randomness of $\beta_j$ only. Model (9) indeed has a Poisson GLMM counterpart, namely,

$$[Y_{ijk} \mid \nu_{ij}] \overset{\text{iid}}{\sim} \text{Poisson}(e^{\nu_{ij}}), \quad \nu_{ij} = H_i + \beta_j, \quad [H_i \mid \alpha, \sigma^2_H] \overset{\text{iid}}{\sim} N(\alpha, \sigma^2_H), \quad [\beta_j \mid \sigma^2_j] \overset{\text{iid}}{\sim} N(0, \sigma^2_j) \quad (10)$$

Note that $H_i$ in (10) can be expressed as

$$H_i = \alpha + \varepsilon_i \quad \text{where} \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2_H) \quad (11)$$

Possible prior distributions are

$$\alpha \sim N(c_1, c_2), \quad \sigma^2_H \sim \text{inverse-gamma}(c_3, c_4), \quad \sigma^2_j \sim \text{inverse-gamma}(c_{5j}, c_{6j}) \quad (12)$$

Next, we discuss (ii) by considering (9). Here, the covariance structure of the data coming from sites $(i, i')$ and from metrics $(j, j')$ is:

$$\text{Cov}(W_{ijk}, W_{ij'k}) = \text{Cov}(H_i + \beta_j + \varepsilon_{ijk}, H_{i'} + \beta_{j'} + \varepsilon_{ij'k}) = \text{Var}(\beta_j) = \sigma^2_j \quad (13)$$

$$\text{Cov}(W_{ij}, W_{ij'}) = \text{Cov}(H_i + \beta_j + \varepsilon_{ijk}, H_i + \beta_{j'} + \varepsilon_{ij'k}) = \text{Var}(H_i) = \sigma^2_H \quad (14)$$

By (13)–(14), the correlation in the log-value of metric $j$ between any pair of sites is constant over sites (i.e. independent of $(i, i')$). Similarly, (15)–(16) implies that for site $i$, the correlation between any pair of metrics is constant over metrics (i.e. independent of $(j, j')$).
This assumption is somewhat foolish. Take \( i = \text{BB1} \) and \( i' = \text{BB2} \), for instance. Both sites are along the same stream, Big Bear. Thus, \( W_{ijk} \) and \( W_{i'jk} \) are highly correlated. Now take \( i = \text{BB1} \) and \( i' = \text{TH1} \). These sites are from entirely different streams (Big Bear and Thornton), so that \( W_{ijk} \) and \( W_{i'jk} \) are possibly uncorrelated.

Similarly, take \( j = 1 \) (total number of observed taxa (#Tx)) and \( j' = 2 \) (number of observed \textit{Ephemeroptera} taxa (#Eph)). As #Tx is obtained by adding #Eph to the number of other taxa, \( W_{ijk} \) and \( W_{i'jk} \) are linearly correlated. Now, take \( j = 2 \) and \( j' = 7 \) (number of observed clinger taxa (#Cl)). The correlation structure between \( W_{ijk} \) and \( W_{i'jk} \) is now intrinsically different and may not be linear, since some \textit{Ephemeroptera} taxa fall in the clinger category, while others may not.

Thus, a better model must address the dependence on \((i, i')\) and \((j, j')\) when considering \( \text{Cov}(Y_{ijk}, Y_{i'jk}) \) and \( \text{Cov}(Y_{ijk}, Y_{i'j'k}) \).

5 Improved Models, Improved LHFI’s

If heteroskedasticity of the response due to metric is the only concern, then one may apply Model (10)–(12).

To address the correlation of \((i, i')\) due to sites along the same streams, we can include urbanization and Global Positioning System (GPS) co-ordinates as covariates. Urbanization is the percentage area of the district / county that is recognized as urban. Such data for PSL streams are available from Morley (2000). For example, in 1997, the district through which the stream Big Bear runs was 41% urbanized. Thus, sites BB1 to BB5 all share the same covariate value of 0.41. Also available from Morley (2000) are the GPS co-coordinates of PSL study sites in the form of latitudes and longitudes. In principle, any variable that may influence the metric responses, \( Y_{ijk} \), may be considered a covariate, provided that data are available. Note that a covariate’s influence on \( Y_{ijk} \) may be modeled as indirect. For example, it may directly influence the field site’s latent health, \( H_i \), which in turn influences \( Y_{ijk} \). Figure 3 illustrates this concept.

To address the different correlation of \((j, j')\) over different combinations of \( j \) and \( j' \), we must specify the overall \( 7 \times 7 \) variance-covariance matrix for \( (\beta_1, \ldots, \beta_7) \) as unknown,
and allow the data to estimate it. In particular, we do not have an obvious way to model the intrinsically different structure between \((j, j')=(1,2)\) and \((j, j')=(2,7)\), etc. (For instance, we cannot expect a simple structure such as AR(1) to reasonably reflect the true structure of the \(Cov(\beta)\) matrix.) Therefore, we impose an unstructured \(Cov(\beta)\) matrix in the model,\(^4\) which also automatically addresses the heteroskedasticity of the response due to metric.

Note that, together with covariates, we now have an analysis of covariance (ANOCOVA) model which has intercept and slope terms associated with the covariates. Below, we first consider covariates having a direct influence on \(Y_{ijk}\). Then, we consider a simpler, more practical model embedded in which is a latent regression of \(H_i\) on covariates.

### 5.1 Urbanization as a Covariate for Metric Response

For simplicity, let us currently consider a single covariate, urbanization.

First, fix \(k=1\), say, and consider the dependence of \(Y_{ijk}\) on \(x_i\)=urbanization for site \(i\) in the log-linear scenario of (9). Take \(j=1\) and plot \(W_{i11}\) against \(x_i\), and there should be a linear trend (or so we assume). Now, take \(j=2\) and plot \(W_{i21}\) against \(x_i\) — again, there should be a linear trend. However, the trend for \(W_{i11}\) should be different from the trend for \(W_{i21}\), since metrics \(j=1\) and \(j=2\) are entirely different features of the ecosystem. In other words, the intercept and slope are unique to each trend. Figure 4 demonstrates this idea.

This phenomenon calls for the specification of random intercept and slope terms associated with \(x_i\), where the randomness comes from the 7 different metrics. That is,

\[
\begin{align*}
\beta_0 &= (\beta_{01}, \beta_{02}, \ldots, \beta_{07}) & \sim & \text{MVN}(b_01, \Sigma_0) \\
\beta_1 &= (\beta_{11}, \beta_{12}, \ldots, \beta_{17}) & \sim & \text{MVN}(b_11, \Sigma_1)
\end{align*}
\]

where \(1\) is the vector \((1,1,\ldots,1)\).

\(^4\)In principle, one may impose a covariance structure that is based on the conceptual relationship among metrics. However, complications may arise in Bayesian estimation of this covariance. For example, it is often unclear how to efficiently sample from the joint posterior distribution of covariance parameters for certain covariance structures. See Westveld (2007), for instance.
Thus, we have the following Gaussian formulation:

\[
W_{ijk} \equiv \ln Y_{ijk} = (\beta_{0j} + \beta_{1j} x_i) + \tilde{H}_i + \varepsilon_{ijk} \quad (17)
\]

\[
\beta_0 \sim \text{MVN}(b_0 \mathbf{1}, \Sigma_0) , \quad \beta_1 \sim \text{MVN}(b_1 \mathbf{1}, \Sigma_1) \quad (18)
\]

\[
\tilde{H}_i \overset{\text{iid}}{\sim} N(0, \sigma_H^2) , \quad \varepsilon_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2) \quad (19)
\]

where \(\tilde{H}_i\)'s, \(\beta_0, \beta_1\), and \(\varepsilon_{ijk}\)'s are mutually independent. Hence,

\[
\text{Cov}(W_{ijk}, W_{i'jk}) = \text{Cov}[(\beta_{0j} + \beta_{1j} x_i) + \tilde{H}_i + \varepsilon_{ijk}, (\beta_{0j} + \beta_{1j} x_{i'}) + \tilde{H}_{i'} + \varepsilon_{i'jk}]
\]

\[
= \text{Var}(\beta_{0j}) + x_i x_{i'} \text{Var}(\beta_{1j})
\]

\[
= \Sigma_{0,ij} + x_i x_{i'} \Sigma_{1,ij}
\]

\[
\text{Cov}(W_{ijk}, W_{ij'k}) = \text{Cov}[(\beta_{0j} + \beta_{1j} x_i) + \tilde{H}_i + \varepsilon_{ijk}, (\beta_{0j'} + \beta_{1j'} x_i) + \tilde{H}_i + \varepsilon_{ij'k}]
\]

\[
= \text{Cov}(\beta_{0j}, \beta_{0j'}) + x_i^2 \text{Cov}(\beta_{1j}, \beta_{1j'}) + \text{Var}(\tilde{H}_i)
\]

\[
= \Sigma_{0,ij} + x_i^2 \Sigma_{1,ij} + \sigma_H^2
\]

Here, we see that both the correlation structure for \((i, i')\) and for \((j, j')\) depend on sites and metrics. Obviously, this is a more appropriate model than either (6) or (9), but it also requires estimation of many more parameters. Also note that the latent health factor here, \(\tilde{H}_i\), is formulated with a zero mean, thus differing from previous models in which \(H_i\) implicitly includes a non-zero intercept, \(\alpha\). The addition of a site-metric interaction, \(\beta_{1j} x_i\), in (17) makes it impossible to isolate a site-only intercept term.

In practice, one may prefer a GLMM for our taxa richness counts. To this end, we reformulate (17)–(19) as a Poisson GLMM:

\[
[Y_{ijk} | \nu_{ij}] \overset{\text{iid}}{\sim} \text{Poisson}(e^{\nu_{ij}}) , \quad \nu_{ij} = (\beta_{0j} + \beta_{1j} x_i) + \tilde{H}_i \quad (20)
\]

\[
[\beta_0 | b_0, \Sigma_0] \sim \text{MVN}(b_0 \mathbf{1}, \Sigma_0) , \quad [\beta_1 | b_1, \Sigma_1] \sim \text{MVN}(b_1 \mathbf{1}, \Sigma_1) \quad (21)
\]

\[
\begin{bmatrix}
\tilde{H}_i \sigma_H^2
\end{bmatrix} \overset{\text{iid}}{\sim} N(0, \sigma_H^2) \quad (22)
\]

where \(\beta_0 = (\beta_{01}, \ldots, \beta_{07})\), and similarly for \(\beta_1\), with prior distributions

\[
b_0 \sim N(c_1, c_2) , \quad b_1 \sim N(c_3, c_4) , \quad \sigma_H^2 \sim \text{inverse-gamma}(c_5, c_6) \quad (23)
\]

\[
\Sigma_0^{-1} \sim \text{Wishart}(S_0, c_7) , \quad \Sigma_1^{-1} \sim \text{Wishart}(S_1, c_8) \quad (24)
\]
where the hyperparameter values $c_1, \ldots, c_8, S_0$, and $S_1$ are chosen to reflect a priori notions of the parameters to be estimated.

### 5.2 Urbanization as a Covariate for Health

The large number of unknown parameters in Model (20)–(22) may require large datasets in order to produce reasonable estimates. This may not be practical, as field data are expensive to collect and assay. Instead, consider the following hierarchical model in the log-linear case:

\[
W_{ijk} = H_i + \beta_j + \varepsilon_{ijk} \quad (25)
\]

\[
H_i = \gamma_0 + \gamma_1 x_i + \delta_i \quad (26)
\]

where \( \gamma_0 \sim N(\mu_{\gamma_0}, \sigma_{\gamma_0}^2) \), \( \gamma_1 \sim N(\mu_{\gamma_1}, \sigma_{\gamma_1}^2) \), \( \beta = (\beta_1, \ldots, \beta_7) \sim \text{MVN}(0, \Sigma) \), \( \delta_i \overset{iid}{\sim} N(0, \sigma_H^2) \), \( \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2) \) \( (27) \)

are mutually independent random effects. (The intercept, \( \gamma_0 \), and slope, \( \gamma_1 \), could be fixed effects, in which case (27) may be regarded as prior distributions in a Bayesian context.) A possibly linear relationship between \( H_i \) and \( x_i \) is supported by Figure 5 (labeled ‘h’ in the graph), which shows a general downward sloping trend between urbanization and LHFI(a), with a correlation of \(-0.72\) (compared to \(-0.76\) for either B-IBI or SHIPSL). As noted by Chiu and Guttorp (2006), the high correlation between “health” and urbanization is predominantly driven by sites at the extremes of the health spectrum. Therefore, it remains to be seen if incorporating the information that \( x_i \) contains about \( H_i \) into the LHFI model would affect the sites’ relative health ratings, and if so, in what way.

In principle, the hierarchical GLMM (25)–(28) is preferable to the simpler two-way model of (6), since it addresses the heteroskedasticity of the response due to metric, and
the different dependence structures among observations from different sites and metrics:

\[
\text{Var}(W_{ijk}) = \text{Var}[\gamma_0 + \gamma_1 x_i + \delta_i + \beta_j + \varepsilon_{ijk}]
= \text{Var}(\gamma_0) + x_i^2 \text{Var}(\gamma_1) + \text{Var}(\delta_i) + \text{Var}(\beta_j) + \text{Var}(\varepsilon)
= \sigma_{\gamma_0}^2 + x_i^2 \sigma_{\gamma_1}^2 + \sigma_H^2 + \Sigma_{jj} + \sigma_{\varepsilon}^2
\]

\[
\text{Cov}(W_{ijk}, W_{i'j'k}) = \text{Cov}[\gamma_0 + \gamma_1 x_i + \delta_i + \beta_j + \varepsilon_{ijk}, \gamma_0 + \gamma_1 x_{i'} + \delta_{i'} + \beta_{j'} + \varepsilon_{i'j'k}]
= \text{Var}(\gamma_0) + x_i x_{i'} \text{Var}(\gamma_1) + \text{Var}(\beta_j) + \text{Var}(\beta_{j'}) + \text{Cov}(\delta_i, \delta_{i'}) + \text{Cov}(\beta_j, \beta_{j'})
= \sigma_{\gamma_0}^2 + x_i x_{i'} \sigma_{\gamma_1}^2 + \sigma_H^2 + \Sigma_{jj'} + \Sigma_{j'j}
\]

In addition, comparing this model to the previous Model (20)–(22), we see that the latent regression (26) reduces the need for a random slope due to metric, while retaining the random intercept in the outer level (25).

For reasons already given earlier, we prefer to reformulate (25)–(28) in a Poisson regression:

\[
[Y_{ijk}|\nu_{ij}] \sim \text{Poisson}(e^{\nu_{ij}}), \quad \nu_{ij} = H_i + \beta_j, \quad H_i = \gamma_{0i} + \gamma_1 (x_i - \bar{x}) \quad (29)
\]

\[
[\gamma_{0i}|\alpha, \sigma_H^2] \sim \text{N}(\alpha, \sigma_H^2), \quad [\beta|\Sigma] \sim \text{MVN}(0, \Sigma) \quad (30)
\]

Here, the latent health term in (29) can be expressed as \( H_i = \alpha + \gamma_1 (x_i - \bar{x}) + \delta_i \) where \( \delta_i \sim \text{N}(0, \sigma_H^2) \). Compare this formulation to (10)–(11) to see that we have further broken down \( \varepsilon_i \) from (11) into \( \gamma_1 (x_i - \bar{x}) + \delta_i \). Prior distributions are

\[
\alpha \sim \text{N}(c_1, c_2), \quad \gamma_1 \sim \text{N}(c_3, c_4)
\]

\[
\sigma_H^2 \sim \text{inverse-gamma}(c_5, c_6), \quad \Sigma^{-1} \sim \text{Wishart}(\mathbb{S}, c_7) \quad (32)
\]

where \( c_1, \ldots, c_7, \) and \( \mathbb{S} \) are \textit{a priori} values associated with the unknown parameters.
5.3 Latitude and Longitude as Additional Covariates for Health

The regression of $H_i$ in Equation (29) may be extended to include GPS co-ordinates of the field sites, thus becoming

$$H_i = \gamma_{0i} + \gamma_1(x_i - \bar{x}) + \gamma_2(t_i - \bar{t}) + \gamma_3(u_i - \bar{u}) \quad \text{(33)}$$

where $t_i$ and $u_i$ are the latitude and longitude, respectively, of site $i$. Again, $\gamma_{0i}$’s are random intercepts, and $\gamma_1, \gamma_2, \gamma_3$ are fixed slopes with a diffuse normal prior distribution.

5.4 Application to the 1997 PSL Data

Due to its complexity, Model (20)–(22) was not implemented. However, Model (29)–(32) was implemented in BUGS to analyze the PSL data from 1997. The hyperparameters were $c_1 = c_3 = 0$, $c_2 = c_4 = 100$, $c_5 = c_6 = 1$, $c_7 = 7$, and $\mathbf{S}$ taken to have diagonal values 1 and off-diagonal values 0.5. This reflects the notion that all 7 metrics are positively associated with the latent health factor, and hence, with each other.

Two Markov chains of 200,000 samples were generated from the joint posterior distribution based on different initial values. Despite removing a burn-in of 100,000 for each chain, and thinning the remainder by a lag of 20 to somewhat reduce autocorrelation, the (thinned) chains still mixed very poorly. To improve mixing and convergence, Model (29)–(30) was reformulated using partial hierarchical centering (HC), so that

$$Y_{ijk} \sim \text{Poisson}(e^{\nu_{ij}}), \quad \nu_{ij} = \nu_{i1} - (b_1 - b_j) \quad \forall j > 1 \quad \text{(34)}$$
$$b_j = \alpha + \beta_j, \quad b = (b_1, \ldots, b_7), \quad [b|\alpha, \Sigma] \sim \text{MVN}(\alpha \mathbf{1}, \Sigma) \quad \text{(35)}$$
$$[\nu_{i1}|b_1, \gamma_1, \mathbf{x}, \sigma_H^2] \sim N(b_1 + \gamma_1(x_i - \bar{x}), \sigma_H^2), \quad H_i = \alpha + \nu_{i1} - b_1 \quad \text{(36)}$$

The rationale behind (34)–(36) appears in the Appendix. Mixing improved drastically with this reformulation, where two chains of 25,000 samples were drawn using different initial values. After excluding a burn-in of 5,000 and thinning the remainder by a lag of 5, we have two chains of 4,000 posterior draws each. From the frequentist’s perspective, the primary parameters of interest are $\alpha, \gamma_1, \sigma_H^2$, and $\Sigma$, and thus we focus on the MCMC mixing for
these. However, the posterior samples of the matrix components of $\Sigma$ were particularly volatile,\(^5\) although the first chain was less so (much smaller posterior dispersion), possibly due to a better starting value. Furthermore, minor mixing problems remained for $\alpha$ and $\gamma_1$.\(^6\)

As the values of $H_i$ depend on $\alpha$ and $\gamma_1$ which yielded somewhat ambiguous estimates, it may appear unclear how one would define an LHFI in this case.

However, upon closer inspection, the chains for $\nu_{ij} = H_i + \beta_j$ as well as $\sigma^2_{H_i}$ mixed very well for either formulation, but more so for (34)–(36). MCMC mixing problems often arise due to correlation among and/or weak identifiability of parameters (Gelfand et al. 1995). Thus, it appears that in either formulation, $\nu_{ij}$’s are well identified but $H_i$’s and $\beta_j$’s are not. For the purpose of rating the sites relative to each other’s health conditions, strong identifiability of $H_i$ may not be necessary, provided that (i) $\nu_{ij}$’s are well estimated, and (ii) for each $i$, the different initial values for the Markov chains lead to empirical posterior distributions of $H_i$ that only differ by a location shift. Since both are satisfied here, we prefer the chain with larger posterior variability for $\Sigma$, and define LHFI(c) to be the mean of the $H_i$ samples from this chain. The smaller posterior variability for the rejected chain could have resulted from a bad initial value that confined the Markov chain to a small subset of the sample space.

Another possible explanation for the poor mixing is that our dataset does not contain enough information to produce stable parameter estimates for such a large model. To further investigate this problem, we replaced $[\beta|\Sigma] \sim \text{MVN}(0, \Sigma)$ in (30) by $[\beta_j|\sigma^2_{\beta_j}] \sim \text{ind} N(0, \sigma^2_{\beta_j})$, and applied the prior of $\sigma^2_{\beta_j} \sim \text{id} \ \text{inverse-gamma}(1,1)$. The resulting model is the hierarchical version of Model (10)–(12), which was also implemented for comparison purposes, with hyperparameters $c_1 = 0, c_2 = 100, c_3 = c_4 = c_{5j} = c_{6j} = 1$ for all $j$. Each version was implemented in BUGS with an HC formulation similar to (34)–(36), where two chains of posterior samples were generated using different initial values. For the hierarchical version — call it (d) — each chain consisted of 10,000 samples, excluding a burn-in

\(^5\)Besides poor mixing, both chains show extreme skewness with tail values on the order of $10^5$. However, the posterior medians are comparable between chains.

\(^6\)Somewhat high autocorrelation within a chain and a very slight difference in location for the empirical distributions between two chains were observed.
of 1,000. For the simple two-way ANOVA — call it (e) — the first 1,000 samples of each chain were discarded as burn-in, then the subsequent 50,000 samples were thinned by a lag of 10, resulting in a final sample of 5,000 per chain. In both cases, the resulting chains mixed exceptionally well. We combine the two (d) chains to form a sample of 20,000, and the two (e) chains to form a sample of 10,000. We let LHFI(d) and LHFI(e) denote the mean of the respective combined sample for $H_i$.

Another model was considered to investigate the influence of taking $\beta_j$ to be a fixed effect (such as for LHFI(a) and (b)) but in a hierarchical context. This new model replaces $[\beta|\Sigma] \sim \text{MVN}(0, \Sigma)$ in (30) by fixed effects $\beta_1 \equiv 0$ and $\beta_2, \ldots, \beta_7$ unknown with common prior distribution $N(0, 100)$. Despite not employing HC here, two chains of 10,000 unthinned posterior samples (excluding a burn-in of 1,000) mixed very well, except for minor problems\footnote{See Footnote 6} with the univariate chains of $\gamma_1$. All of $\alpha, H_i, \beta_j, \nu_{ij}, \sigma^2_H$, and $\sigma^2_\beta$ were well estimated. For the purpose of defining an LHFI, the unambiguity of the $H_i$’s encourages us to combine the two chains to form a sample of 20,000. The combined empirical posterior mean is denoted by LHFI(f).

Additional covariates for $H_i$ were also considered — see (33). For the 1997 PSL data, urbanization is highly correlated with longitude ($r=0.91$), but very weakly with latitude ($r=0.19$). To avoid regressing the latent health factor on redundant variables, we included only latitude as an additional covariate to urbanization. Results (not shown) indicate that including latitude in the model has virtually no impact on the posterior distributions of the $H_i$’s (or of other model parameters). In fact, a typical 95\% credible interval (C.I.) for the coefficient for latitude approximately ranges from $-1.3$ to $1.5$, suggesting statistically insignificant effect on health due to latitude. Therefore, LHFI’s derived from such models are not considered.

5.4.1 Discussion of Results

The top half of Table 1 lists the models that correspond to LHFI(a)–(f). As it turns out, the pairwise correlations among the six LHFI’s are all equal to 1.00. Of course, each in-
dex has a slightly different scale, as can be seen in Figure 6. Interestingly, therefore, it appears that besides a different scale, all of LHFI(a)–(f) contain virtually identical information regarding the health conditions of the 18 PSL sites. However, while we define the index to be the posterior mean of $H_i$, we must also consider its variability as a measure of reliability for the assessment of health. From Figure 6, we see that LHFI(f) — based on the hierarchical model with fixed $\beta_j$’s — has the least variability, followed by (a), then (d) or (e), and finally, (c). It is also interesting to note that LHFI(d) is substantially more variable than (f) despite its being associated with only one extra model parameter. At this point, it would appear that LHFI(f) wins the race of precision and power, followed by (a). Both correspond to a model taking metric effects as fixed. However, we must recall that such models\(^8\) ignore heteroskedasticity of the response due to metric, and may yield a false sense of precision as a result of model misspecification. Therefore, LHFI(f) and (a) should be disqualified in the race. Among the remaining indices, LHFI(d) and (e) appear to be equal winners. Thus, despite the earlier argument for the hierarchical models leading to (c) and (d), the simple two-way model apparently produces no less precision when applied to the 1997 PSL data. However, this does not necessarily imply that one should prefer the simpler model in practice.

Indeed, the two models leading to (c) and (d) each yields a 95% C.I. for $\gamma_1$ that is below zero (approximately $-3.5$ to $-0.7$ — see Table 3). That is, there is statistical evidence that urbanization has a negative impact on stream health. While this might have been a foregone conclusion from a biological point of view, the hierarchical models considered here provide direct quantitative evidence to support this biological notion. Such results indeed have profound implications in practice. A policy maker may be presented with several factors that have potential impact on ecosystem health. Meanwhile, due to limited resources, s/he may be forced to devise conservation policies to control some of these factors only. Our latent factor hierarchical modeling approach thus provides the policy maker with a scientific mechanism to prioritize the control of factors: a negative C.I. indicates detrimental effects, one that spans 0 indicates insignificant impact, and a positive C.I. implies positive impact.

---

\(^8\)... and also for $\beta_j \sim N(0, \sigma_\beta^2)$. 

20
Take the model that includes both urbanization and latitude, for instance. As mentioned earlier, the 95% C.I. for the latitude coefficient contains 0, while that corresponding to urbanization does not. These C.I.’s suggest that latitude has much less impact on latent health.\(^9\)

We now return our focus to LHFI’s from models with no latitude term. Again, (c) is less preferable due to a lower precision of the health assessment. Indeed, each level in the hierarchy of a model could introduce larger variability to the latent health factor due to the extra unknown parameters involved. Yet, when both models (d) and (e) — hierarchical and otherwise — yield equally variable \(H_i\)’s, how might one proceed with the comparison?

### 5.4.2 Quantitative Comparison of (d) and (e)

Since precision of the health assessment is of concern, one approach is to compare the posterior distributions for \(\sigma_H\), whose mean is 0.50 for the the hierarchical model for \textbf{LHFI(d)}, and 0.60 for (e). The corresponding 95% C.I.’s are [0.36, 0.71] and [0.43, 0.86], respectively, with a huge overlap. These summaries (see Table 3) indicate that there is little statistical difference in the precision parameter for \(H_i\)’s between the two models. (Note that \(\sigma_H\) is different from the dispersion of the \(H_i\) posterior distributions.)

Another approach is to consider cross validation by out-of-sample predictions for each model. Intuitively, breaking down \(\varepsilon_i\) from (10) of (e) into \(\gamma_1(x_i - \bar{x}) + \delta_i\) of (d) could improve predictive power, given that the covariate \(x_i\) is useful for explaining latent health \(H_i\), which in turn explains the indicators \(y_{ijk}\)’s. To investigate this empirically, we randomly divided the 1997 PSL data \((18 \times 7 \times 3 = 378\) observations in total) into nine disjoint subsets of 42 values each. The following procedure was then performed on each subset:

- The 42 values were removed from the data and treated as missing values.
- Each of the models for \textbf{LHFI(d)} and (e) was performed on the remaining \(8 \times 42 = 336\) observations (as described under Section 5.4), with the 42 missing values imputed as unknown parameters within the Bayesian framework.

---
\(^9\)A technical note on this ranking scheme for multiple covariates is that the credible level may require adjusting in the context of multiple testing; see Westfall \textit{et al.} (1997), for instance.
• The mean of the posterior distribution for each imputed value was taken to be the “estimate” for the missing observation. The mean is preferred over the median or mode here since a discrete likelihood (Poisson) is considered for the observations, so that different models can yield the same median or mode even if their means are quite different.

• An empirical root-mean-square-error (RMSE) is computed based on

\[
\text{MSE} = \text{RMSE}^2 = \sum_{\{y: \text{missing values}\}} (y^{\text{estimated}} - y^{\text{true}})^2
\] (37)

The histogram of the nine differences between these RMSE’s appears in Figure 7. While the hierarchical model produced larger RMSE’s for all but one set, the difference is minimal across all sets, ranging from −0.02 to 0.07. Therefore, both models perform almost equally well with respect to out-of-sample predictions for these nine datasets. In theory, one could compare the predictive power between models by considering the respective posterior predictive distributions over all possible combinations of missing values. This approach is practically infeasible, and thus the comparison based on nine randomly generated datasets serves as a pilot study.

In our pilot study, we noticed that the Monte Carlo error for the imputed values is on the order of $10^{-2}$ for both models. Therefore, could the theoretical posterior predictive distributions for the imputed values be identical between models, so that the observed difference in RMSE is entirely due to Monte Carlo error? This is an unlikely scenario, for several reasons.

Firstly, the posterior predictive depends on the specified priors. In essence, the model for (e) specifies a prior for $\gamma_1$ that is a point mass of 1 at the value 0. In contrast, the $\gamma_1$ prior corresponding to (d) is diffuse, and in practice can be centered at any value including 0.

Secondly, one may consider the case of imputing a single missing value — say, $y_{111}$ — and examine the posterior predictive distributions analytically. To this end, let $W$ denote the posterior predictive mean for $y_{111}$. Suppose that the posterior predictive distribution for $y_{111}$ were identical between (d) and (e). Then $W$, and hence $E(W)$ also, would be equal between models. However, below we show that this is not true.
Let \( \mathbf{y}_{-111} \) denote the vector of all available observations. Consider the case for \( \text{LHFI(e)} \), where

\[
E_{(e)}(W) = E [E_{(e)}(Y_{111} | \mathbf{y}_{-111})] = E_{(e)}(Y_{111}) = E [E (Y_{111} | H_1, \beta_1)] = E [e^{H_1 + \beta_1}] = E [e^{H_1}] \times E [e^{\beta_1}]\tag{38}
\]

\[
= E [E (e^{H_1} | \alpha, \sigma_H^2)] \times E [E (e^{\beta_1} | \sigma_1^2)]\tag{39}
\]

where the subscript \( (e) \) in (38)–(40) has been (and will continue to be) suppressed to reduce clutter. Now, based on a two-term Taylor expansion, we have

\[
E [e^{H_1} | \alpha, \sigma_H^2] \approx e^\alpha \left( 1 + \frac{\sigma_H^2}{2} \right), \quad E [e^{\beta_1} | \sigma_1^2] \approx 1 + \frac{\sigma_1^2}{2}
\]

Recall the hyperparameters from (10)–(12) and substitute terms into (40) to get

\[
E [E (e^{\beta_1} | \sigma_1^2)] \approx 1 + \frac{1}{2} E[\sigma_1^2] = 1 + \frac{c_{51}}{2c_{61}}
\]

\[
E [E (e^{H_1} | \alpha, \sigma_H^2)] \approx E \left[ e^\alpha \left( 1 + \frac{\sigma_H^2}{2} \right) \right]
\]

\[
= E [e^\alpha] \times \left( 1 + \frac{1}{2} E [\sigma_H^2] \right) = E [e^\alpha] \times \left( 1 + \frac{c_3}{2c_4} \right)
\]

\[
\implies E_{(e)}(W) \approx E_{(e)} [e^\alpha] \times \left( 1 + \frac{c_3}{2c_4} \right) \left( 1 + \frac{c_{51}}{2c_{61}} \right)\tag{41}
\]

Similarly, it can be shown that the corresponding mean for \( \text{LHFI(d)} \) is

\[
E_{(d)}(W) \approx E_{(d)} [e^\alpha] \times E [e^{\gamma_1 x_1}] \left( 1 + \frac{c_3'}{2c_4'} \right) \left( 1 + \frac{c_{51}'}{2c_{61}'} \right)\tag{42}
\]

where \( c_3', c_4', c_{51}', \) and \( c_{61}' \) are the hyperparameters in the prior distributions for \( \sigma_H^2 \) and \( \sigma_1^2 \) for \( \text{(d)} \), and are possibly different from the corresponding values for \( \text{(e)} \).

It is now easy to see that generally, (41) and (42) should take on values that are quite different. Even for identical hyperparameters between both models for the prior distributions of \( \alpha, \sigma_H^2 \) and \( \sigma_1^2 \), we would still have

\[
\frac{E_{(d)}(W)}{E_{(e)}(W)} \approx E [e^{\gamma_1 x_1}]\tag{43}
\]

23
which cannot be close to 1 since the hypervariance for \( \gamma_1 \) is large and \( x_1 \neq 0 \). Therefore, (43) demonstrates by contradiction that the posterior predictive distribution for \( y_{111} \) should be quite different between models. The same argument applies to imputing any combination of \( y_{ijk} \)’s.

To summarize, we have argued that theoretically, the two models corresponding to LHFI(d) and (e) yield different posterior predictive distributions for the missing values. However, the observed difference in the out-of-sample predictions between the two models in our pilot study is too minor to be considered significant. That is, judging empirically, we do not have evidence that one model performs better than the other in terms of reliability for the sole assessment of latent health. Of course, this argument does not account for the merit of including relevant covariates to provide guidelines on conservation measures, which we have discussed in Section 5.4.1.

6 A Comprehensive Model

Although not highly informative on its own, Model (2)–(3) was fitted to determine the possibility of including \( N_{ik} \) in addition to the 10 metrics that already form B-IBI and SHIPSL. Results (not shown here) indicate that the \( H_i \)’s from this simple model are positively correlated, although not highly, with the LHFI values from Section 3.1.

However, \( N_{ik} \) already appears implicitly as the denominator in the definition of relative abundance metrics; for the PSL, they are \% animals from tolerant taxa\(^{10} \) (%Tol), \% animals from predatory taxa (%Pred), and \% animals from the three most dominant taxa (%Dom3). Note that %Tol and %Dom3 are negatively associated with health (Morley, 2000), and must be transformed so that higher values of the index (LHFI/B-IBI/SHIPSL) correspond to higher values of any metric. An obvious transformation is to take

\[
\%\text{NonTol} = 100\% - \%\text{Tol}, \quad \%\text{NonDom3} = 100\% - \%\text{Dom3}
\]

Indeed, Chiu and Guttorp (2004) show that it is beneficial, at least statistically, to con-

\(^{10}\)Some taxa are classified as neither tolerant nor intolerant. Therefore, non-tolerant taxa are not necessarily intolerant.
vert taxa richness (count) metrics to percentages also, before combining them with relative abundance metrics to form a health index. They suggest taking \#Tx as the denominator for transforming the other 6 count metrics into *relative richness* percentages. This way, all 9 variables now share the same scale, and a random-effects model similar in principle to (6)–(7), (20)–(24), or (29)–(32) may be used to construct a comprehensive LHFI.

To this end, each metric may be considered an observed *probability of success*, where “success” is an occurrence of the taxon or animal indicative of a healthy stream.\(^{11}\) Therefore, it appears that logistic regression models are appropriate for constructing a comprehensive LHFI. Below, we will first consider one that is entirely binomial-based. We will then make use of the disjointness of three of the richness metrics to formulate a binomial-multinomial model.

### 6.1 Two Logistic Regression Models for Relative Richness and Abundance

We focus on hierarchical models similar to Model (29)–(32) with dependent metric effects.

Two groups of variables form our 9 metrics: 6 pertaining to richness, and 3 pertaining to abundance. Let \(s=1\) represent the richness group, and \(s=2\) the abundance group. Furthermore, for replicate \(k\) from site \(i\), let \(Y_{isjk}\) denote the total number of successes for metric \(j\) in group \(s\), each success occurring with probability \(p_{isj}\). Hence,

\[
[Y_{isjk}|T_{is,k},p_{isj}] \overset{\text{ind}}{\sim} \text{Binomial}(T_{is,k},p_{isj}) \quad \forall \; k = 1, 2, 3 \tag{44}
\]

where

\[
T_{is,k} = \begin{cases} 
(#Tx)_{ik} & \text{if } s = 1 \\
N_{ik} & \text{if } s = 2
\end{cases} \tag{45}
\]

and \(i \in \{1, \ldots, 18\}, \; j \in \{1, \ldots, m_s\}, \; m_1=6, \; m_2=3\). Now, let \(\nu_{isj}\), denote the logit-transformed

\(^{11}\)For instance, consider *relative clinger richness* (%Cl). Suppose \#Tx=30 and \#Cl=24. Hence, with respect to this metric, we observe 24 successes (80%) out of 30 binomial trials. Similarly, if \(N=1,000\) and \%NonTol=60, then there are 600 successes out of 1,000 binomial trials with respect to abundance of non-tolerant taxa.
\[ p_{i,sj,k} \text{, and consider the GLMM} \]

\[ \ln \frac{p_{i,sj,k}}{1 - p_{i,sj,k}} \equiv \nu_{isj,k} = H_i + \theta_s + \beta_{j(s)} \quad \text{for } s = 1, 2 \quad (46) \]

\[ H_i = \gamma_{0i} + \gamma_1(x_i - \bar{x}) \quad (47) \]

\[ [\gamma_{0i}|\alpha, \sigma^2_H] \overset{\text{iid}}{\sim} N(0, \sigma^2_H), \quad \theta_2 = 0, \quad [\beta|\Sigma] \sim \text{MVN}(0, \Sigma) \quad (48) \]

where \( \beta = [\beta_{(1)}, \beta_{(2)}]^T \), \( \beta_{(s)} = [\beta_{1(s)}, \ldots, \beta_{m(s)}]^T \), and the indexing notation of “\( j(s) \)” is adopted from Montgomery (2005) for level \( j \) of a factor nested within level \( s \) of a parent factor. Here, Model (44)–(48) stipulates that the probability of success is affected by site health, metric type (richness or abundance), and the actual metric within type. As in previous models, site health is regressed linearly on urbanization, and metric effects have a multivariate normal distribution. Note that with only two levels, \( \theta_s \) is modeled as a fixed effect with \( \theta_2=0 \) (abundance) taken as baseline. Prior distributions are as in (31)–(32), and additionally \( \theta_1 \sim N(0, c_s) \).

The formulation of this logistic regression model is based entirely on binomial distributions associated with the nine metrics. The covariance matrix \( \Sigma \) is meant to capture the dependence among metrics due to overlap of definitions (see Section 4). However, the dependence among \( \beta_{1(s)}, \beta_{2(s)}, \) and \( \beta_{3(s)} \) could be adjusted for the disjoint nature of the \( Ephemeroptera \) (Eph), \( Plecoptera \) (Ple), and \( Trichoptera \) (Tri) taxa, often known collectively as the EPT taxa. Hence, \#Eph, \#Ple, and \#Tri define a quadrinomial variate.

To incorporate this multinomial distribution into a latent health factor model, we break down the group of richness metrics into two subgroups by letting \( s=0 \) represent relative richness for each EPT taxon, and \( s=1 \) for the remaining three richness metrics. The group of abundance metrics remains as \( s=2 \). Therefore, each group consists of three metrics. As before, \( Y_{i,sj,k} \) are binomial counts for \( s=1, 2 \). However, for \( s=0 \), we have

\[ \left[ \begin{array}{c} Y_{i,01k} \\ Y_{i,02k} \\ Y_{i,03k} \\ T_{i1,k} = \sum_{j=1}^3 Y_{i0jk} \end{array} \right] \sim \text{Multinomial}(T_{i1,k}, \left[ \begin{array}{c} p_{i01} \\ p_{i02} \\ p_{i03} \\ 1 - \sum_{j=1}^3 p_{ij0} \end{array} \right]) \quad (49) \]

where \( p_{i0j} \) is the probability of an observed taxon from site \( i \) falling in the \( j \)th EPT category. Note that all 6 richness metrics share the same margin, namely, \( T_{i1,k} \), irrespective of \( s=0 \)
or $s=1$. As large values of $p_{i01}$, $p_{i02}$, and $p_{i03}$, are indicative of good stream health, we model them via a multinomial logit link, so that

$$\ln \frac{p_{i0j}}{1 - \sum_{j=1}^{3} p_{i0j}} \equiv \nu_{i0j} \equiv H_i + \theta_0 + \beta_{j(0)} \quad \text{for} \quad j = 1, 2, 3 \quad (50)$$

We retain (46) for $s=1,2$ in our model, and continue to take $s=2$ as baseline. In other words, our binomial-multinomial mixture logit model consists of (44)–(48) combined with (49)–(50), except that $m_s=3$ for all $s=1,2,3$, and $\beta = [\beta_{(0)}, \beta_{(1)}, \beta_{(2)}]^T$. We also consider the same prior distributions as the binomial-only model, but add the extra prior $\theta_0 \sim N(0, c_0).^{12}$

### 6.2 Application to the 1997 PSL Data

Both logit models were implemented in BUGS and applied to the 1997 PSL data, with hyperparameters $c_1 = c_3 = 0, c_2 = c_4 = c_5 = c_6 = 100, c_7 = 9$, and $\mathbb{S}$ taken to have diagonal values 1 and off-diagonal values 0.5. For such models, a truly effective hierarchically centered formulation is difficult to construct, due to the extra $\theta_s$’s that result in a three-way dependence among the intercepts, site health, and metric effects. One of the formulations we explored performed somewhat satisfactorily:

$$\nu_{isj} = \tilde{H}_i + b_{js} \quad , \quad [\tilde{H}_i | \gamma_1, \alpha, \sigma_{\tilde{H}}^2] \sim N(\gamma_1(x_i - \bar{x}), \sigma_{\tilde{H}}^2)$$

$$[b | \psi^*, \Sigma] \sim \text{MVN}(\psi^*, \Sigma) \quad , \quad \psi_s = \alpha + \theta_s \quad H_i = \alpha + \tilde{H}_i \quad , \quad \beta = b - \psi^*$$

where $\psi^* = [\psi_1, \ldots, \psi_1, \psi_2, \psi_2]^T$ for the binomial-only model, and $\psi^* = [\psi_0, \psi_0, \psi_0, \psi_1, \psi_1, \psi_2, \psi_2]^T$ for the mixture model.

For the binomial-only model, two Markov chains of 155,000 samples were drawn from the joint posterior distribution. The first 3,000 draws were removed from each chain as burn-in, and the remainder was thinned by a lag of 10 to reduce autocorrelation. The final chains of 15,200 each mixed reasonably well as a whole, but particularly so for $p_{isj}$’s, $\gamma_1$, $\alpha$, $\sigma^2_{\tilde{H}}$.

---

12 Although one may desire to impose a structure on $\Sigma$ — particularly after incorporating the multinomial distribution of $Y_{i0jk}$’s — Bayesian estimation of structured covariances is difficult in general. See Footnote 4.
and $\sigma_H^2$. However, we encountered very similar mixing problems here for $\Sigma$, $\alpha$, and $\theta_s$’s as we did in fitting Model (29)–(30). Again, we focused on the chain that produced a more dispersed posterior for $\Sigma$. The same applies to the mixture model, except for a burn-in of 5,000 removed from a total of 305,000 samples, then subsequently thinned by a lag of 60 to produce two chains of 5,000 samples each. Both models resulted in 95% C.I.’s for $\theta_s$’s that span 0, namely, $[-3.81, 1.52]$ and $[-4.87, 1.81]$ for $\theta_1$ for the two models, respectively, and $[-3.43, 3.11]$ for $\theta_0$ for the mixture model. The C.I.’s suggest that richness and abundance as metric types do not significantly affect the binomial / multinomial probabilities. This prompted us to fit reduced models without $\theta_s$’s.

The same MCMC procedure was undertaken, where, for the reduced binomial-only model, a burn-in of 1,000 was removed from each of two chains of 51,000 samples, then thinned by a lag of 10 to produce 5,000 samples per chain; for the other model, the figures were 5,000 for burn-in, 255,000 for total, 50 for thinning, and 5,000 for final chains. Despite model reduction, mixing problems remained for $\Sigma$, and very slightly for $\alpha$, $H_i$’s, and $\beta_j$’s. To be conservative, we chose the chain from the binomial-only model that produced a more dispersed posterior for $\Sigma$, and let $\text{LHFI}(g)$ denote the empirical mean for $H_i$ from this chain. Mixing problems were much less severe for $\Sigma$ in the other model, but slightly more so for the intercept and random effects. There, we arbitrarily chose one chain and labeled the corresponding health index $\text{LHFI}(h)$.

Despite the need to capture metric dependence, our results have again demonstrated its infeasibility for the 1997 PSL dataset. Therefore, we further reduced the logit models by replacing a general $\Sigma$ with a diagonal covariance matrix, such as for $\text{LHFI}(d)$. However, mixing problems remained for this drastically simplified $\Sigma$. We considered a last step towards model reduction by taking a common variance for the EPT metric effects. This is a reasonable assumption both biologically and statistically. For the latter, refer to $\text{LHFI}(g)$ and (h) in Table 4. There, we see that the posterior distributions of the EPT metric effect variances within each chain have almost identical summary statistics, despite severe mixing problems between chains.

This new assumption was subsequently implemented. For the binomial-only model,
we removed from each of two chains 5,000 draws as burn-in, and generated another 20,000 samples that were thinned by a lag of 2. For the mixture model, a longer burn-in was required: 10,000 draws were removed, then another 30,000 samples were generated but thinned by a lag of 3. In both cases, the final samples consisted of two chains of 10,000 samples each.

In the former case, the chains mixed exceptionally well for $\gamma_1$, $\sigma_H$, $\sigma_1(1) = \sigma_2(1) = \sigma_3(1)$, and all $p_{i(s)}$’s. Minor mixing problems were observed for $\alpha$, $H_i$’s, and $\beta_j$’s. Severe mixing problems remained for the non-EPT $\sigma_{j(s)}$’s, largely due to the extreme skewness of the posterior with some tail values on the order of $10^3$; however, the posterior medians were comparable between chains. Nevertheless, we arbitrarily chose one chain and defined LHFI(j) to be the posterior mean of $H_i$.

Finally, in the latter case, we encountered similar issues with the non-EPT $\sigma_{j(s)}$’s as above, but all other chains mixed exceptionally well. Therefore, we combined both chains for the purpose of defining LHFI(i) as the posterior mean of $H_i$.

### 6.2.1 Discussion of Results

Since LHFI(g)–(j) are based on small variations of what is essentially the same model (see bottom half of Table 1), it is not surprising that all four indices have perfect correlation empirically. Such was also the case for LHFI(a)–(f).

Figure 1 compares LHFI(g) (representing the logit-model-based indices) to LHFI(a) and (b) (representing the Poisson-model-based indices), as well as B-IBI and SHIPSL. Table 2 lists the correlations among them. Note that B-IBI is more linearly associated with the logit-based LHFI than the Poisson-based counterpart. The higher correlation may be explained by all metrics sharing the same scale for B-IBI and the logit models. Interestingly, our logit models have yielded LHFI’s that are (essentially) equally correlated with B-IBI, SHIPSL, and the Poisson-based LHFI’s.

Having metrics share a common scale may also have increased the power of the LHFI to distinguish among sites. Consider Figures 6 and 8. There, LHFI(g) and (h) show bigger differences among sites than does LHFI(c), although the three have comparable ranges.
and are derived from a model that considers metric overlap via the estimation of $\Sigma$. This common-scale principle has been used to develop all versions of the IBI as well as SHIPSL. However, for these latter types of indices and particularly for the IBI, the scheme used to map indicators to a single scale undoubtedly is more controversial and causes more loss of information, when compared to the minimal manipulation of metrics before they are incorporated into the logistic regression model for constructing an index.

Comparing the four logit models corresponding respectively to LHFI(g)–(j), all have identified a significant dependence of health on urbanization; the 95% C.I.’s for $\gamma_1$ range approximately from $-3.6$ to $-0.5$ (see Table 4). These figures are comparable to those observed for LHFI(c) and (d). Indeed, the additional multinomial structure incorporated to account for the dependence structure among the EPT metrics does not appear to affect the parameter estimates, except perhaps for slightly improved mixing in the case of LHFI(j). The little difference is demonstrated by the top four panels of Figure 8, where the layout and vertical scales for the boxplots are essentially identical between (g) and (h), and between (i) and (j). However, the third and fourth panels demonstrate that the drastic reduction in the structure of $\Sigma$ led to a highly noticeable improvement in the power of LHFI(i) and (j) to distinguish among sites. Of course, with a large enough dataset (which was not the case here), one would ideally retain a non-trivial dependence structure for the $\beta$’s to account for the actual overlap among metrics.

6.2.2 Quantitative Comparison among (i), (j), (d), and (e)

To compare the performance between the binomial-only and mixture logit models that yielded LHFI(i) and (j), we take the approaches of Section 5.4.2, and first consider the posterior distribution of $\sigma_H$. Both posterior means are 0.57, and respective 95% C.I.’s are $[0.41, 0.79]$ and $[0.41, 0.81]$ (see Table 4). With the striking similarities between the two posterior distributions, it appears that both models perform equally well.

However, we may again wish to consider out-of-sample predictions to further the comparison. We ran the models for LHFI(i) and (j) on 9 sets of incomplete data generated similarly as in Section 5.4.2: with an additional three abundance metrics and the removal
of one richness metric, we have a total of 486 observations, so that each incomplete subset involved 54 missing observations. The same imputation procedure from before was employed.

In this case, our primary concern is “estimating” the missing \( y_{isjk} \)’s by the posterior predictive mean. Recall how well the chains in each model mixed for \( p_{isj} \)’s. Hence, chains would mix just as well for the imputed counts — should different ones be generated — because within each scan of the Markov chain, the predicted \( y_{isjk} \)’s are directly based on the estimated \( p_{isj} \)’s. Thus, generating separate chains would be unnecessary for our purpose here. Instead, we generated a single chain of 20,000 (excluding a burn-in of 5,000) with no thinning.

Note that the current version of BUGS does not allow missing values when multinomial likelihoods are specified. To circumvent this problem for our mixture model, the multinomial part was reformulated as Poisson log-linear (see, for instance, Dobson 2001):

\[
\mu_{ijk} \equiv E[Y_{i0jk} | N_{i0k}, p_{i0j}] = N_{i0k} p_{i0j}, \quad j = 1, 2, 3, 4
\]

\[
p_{i04} \equiv 1 - \sum_{j=1}^{3} p_{i0j}, \quad Y_{i04k} \equiv N_{i0k} - \sum_{j=1}^{3} Y_{i0jk}
\]

\[
\ln \mu_{ijk} = \ln N_{i0k} + \ln p_{i0j} = H_i + \omega_k + \varepsilon_{ik} + \tilde{\beta}_{j(0)} + \xi_{ij}
\]

\[
\tilde{H}_{i} = \alpha + \gamma_1(x_i - \overline{x}) + \delta_i, \quad \tilde{\beta}_{j(0)} \equiv 0, \quad \xi_{i4} \equiv 0 \forall i
\]

\[
\begin{align*}
\left[ \tilde{\delta}_i | \sigma_H^2 \right] & \iid N(0, \sigma_H^2), \\
\left[ \tilde{\beta}_{j(0)} | \sigma_j^2 \right] & \iid N(0, \sigma_j^2) \quad \forall j = 1, 2, 3
\end{align*}
\]

\[
\begin{align*}
\left[ \omega_k | \sigma_\omega^2 \right] & \iid N(0, \sigma_\omega^2), \\
\left[ \varepsilon_{ik} | \sigma_\varepsilon^2 \right] & \iid N(0, \sigma_\varepsilon^2), \\
\left[ \xi_{ij} | \sigma_\xi^2 \right] & \iid N(0, \sigma_\xi^2) \quad \forall j = 1, 2, 3
\end{align*}
\]

Equation (53) specifies that \( \ln N_{i0k} \) is modeled by \( \tilde{H}_i, \omega_k, \) and \( \varepsilon_{ik} \); and the term \( \ln p_{i0j} \) by \( \tilde{H}_i, \tilde{\beta}_{j(0)}, \) and \( \xi_{ij} \). Here, \( \xi_{ij} \) is necessary so that \( \ln p_{i0j} \) from this Poisson formulation corresponds to that for the original multinomial formulation (50) with \( \theta_0=0 \). In fact, a healthier site is expected to have a richer composition of EPT taxa \( (j = 1, 2, 3) \); hence, \( p_{i0j} \) is not a simple product of a site-related probability and a metric-related probability (in which case \( \xi_{ij} \) would be identically 0).

There are several issues associated with the reformulation (51)–(56):

- While the terms with a tilde (“\( \sim \”) resemble those in the multinomial formulation,
there is no direct correspondence between the two sets. This is because linearizing the logarithm of $p_{i0j}$, as above is different from linearizing the logit-transformed $p_{i0j}$ from earlier.

- Extra parameters must be estimated, namely, $\sigma_w, \sigma_\varepsilon,$ and $\sigma_\xi$. This inherently affects the precision in estimating the other parameters.

- The posterior predictive distribution used to impute the missing observations depends on the specified priors. Of course, the new formulation involves different parameters, and hence, different priors. Therefore, the procedure undertaken here may not yield results that closely represent the predictive power of the mixture model. Ideally, one would employ or create software that allows the specification of multinomial likelihoods with missing observations.

Nevertheless, huge differences in predictive power between the binomial-only and mixture models, if present, should be observed regardless of the actual parametrization of the latter.

The bottom panel of Figure 7 shows the distribution of the difference in RMSE between the two models for the nine datasets. Generally, the binomial-only model appears to yield somewhat smaller RMSE’s for the imputed values. The actual simulated differences range from $-0.61$ to $0.24$. Although possibly due to the Poisson reparametrization for the multinomial part of the mixture model, the small RMSE differences demonstrate that accounting for the multinomial nature of the EPT metrics has little impact on the performance of the logit model. These results agree with what was discussed in Section 6.2.1 earlier.

One question remains: which type of model — a Poisson GLMM for richness metrics only, or a logit GLMM for richness and abundance metrics — is preferable for monitoring stream health based on the 1997 PSL data? To make a sensible comparison between types, we only consider the hierarchical models that involve urbanization as a covariate for latent health, and metric effects as independent. Therefore, we consider the models corresponding to LHFI(d), (e), (i), and (j). As little difference is shown to exist between the former two, and similarly, the latter two, we focus on LHFI(d) and (j) only without loss of generality.

In out-of-sample predictions, imputations are made for the responses, i.e. observed
metric values. However, this approach would be inappropriate for comparing the two model types as they involve different sets of response variables. Had site health been observable instead of latent, then out-of-sample predictions for $H_i$’s could be useful. This being not the case, we instead compare the posterior distribution for $H_i$’s and $\sigma_H$ between the two model types.

Consider the last two panels of Figure 8. The range of the posterior distribution for any given site appears to be comparable between the two models. However, the 50% C.I. (delimited by the “box” of the boxplot) is slightly shorter for the logit model. More noticeable for the logit model is the fluctuation across sites for the location of the posterior distribution, as opposed to the apparent flatness corresponding the Poisson model. In other words, the inclusion of abundance metrics in the logit model clearly led to additional ability of the LHFI to distinguish among sites.

7 Model for Combining Spatial and Other Types of Domains

For the purpose of developing a health index, neighboring geographical domains may be similar enough to share the same set of metrics yet differ enough that traditional metric calibration devised for one region may not effectively reflect the health conditions of another.

Consider a single set of $J$ metrics that is deemed adequate for spatial domains A and B. The goal is to assess in one combined study the ecological health of sites $a_1, a_2, \ldots, a_m$ from Domain A and $b_1, b_2, \ldots, b_n$ from Domain B. The traditional ad hoc approach would require recalibration of all $J$ metrics to account for the different spatial scales. Painstaking effort aside, subjectivity plays a heavy role in this recalibration, thus reducing scientific integrity of the resulting index. The SHIPSL scoring scheme handles this problem by way of metric standardization against a mean and SD computed from all $m + n$ sites combined. However, a simple arithmetic mean overlooks the fact that sites $a_i$’s are more similar among themselves than when compared to sites $b_i$’s. Subjectivity would again play a role if a weighted mean and SD for each metric were to be devised.
With our latent factor model, we can, for the first time, handle this issue with biological and statistical integrity. Take the Poisson counts model, for instance. Let \( Y_{ijk\ell} \) denote the value of the \( j \)th count metric from replicate sample \( k \) in the \( i \)th site within spatial domain \( \ell \). Assume a simple non-hierarchical ANOVA similar to (6), except for an additional spatial effect term \( \lambda_\ell \), that is,

\[
[Y_{ijk\ell}|\nu_{ij}] \sim \text{Poisson}(e^{\nu_{ij} \ell}), \quad \nu_{ij\ell} = \alpha + H_{i(\ell)} + \beta_j + \lambda_\ell, \quad [H_{i(\ell)}|\sigma^2_\ell] \sim N(0, \sigma^2_\ell)
\]

with \( i \) nested in \( \ell \). Here, \( \lambda_\ell \)'s may be modeled as random or fixed depending on the context of the study.

The simple addition of a spatial effect term in the latent factor model allows us to study the health conditions (by “estimating” \( H_{i(\ell)} \)) over all sites simultaneously and without ambiguity. This approach can be extended to hierarchical models such as (29)–(30), by regressing \( H_{i(\ell)} \) on \( \lambda_\ell \) and other relevant covariates.

The same principles may also be applied to contexts of multiple temporal domains, whereby one would introduce a temporal effect term (e.g. year) to the model in the same manner as above. To account for both types of domains, a spatio-temporal interaction term may be included. This further demonstrates that our model-based approach for defining an LHFI is a powerful alternative to conventional index calibration that remains widely popular to date despite the lack of spatial or temporal transferability. In the near future, we hope to obtain inter-regional and temporal data from field scientists to illustrate our approach proposed in this section.

8 Handling Scaling Issues for LHFI

The numerical value of an IBI-type index is often believed to be absolute, in the sense that a site’s IBI is supposed to indicate its degree of degradation without the need for comparison to another site. This is one of the reason’s for the IBI’s popularity. However, one must not overlook the calibration scheme that brings about this apparent advantage. This scheme relies heavily on the availability of so-called pristine ecosystems used as references to provide the metric calibration to be applied to any given site. In practice, pristine ecosystems
rarely exist. Thus, it is arguably more sensible to gauge health against a heavily degraded ecosystem, which, sadly, is not difficult to locate these days. Of course, reference-based-calibration also suffers from the disadvantage of non-transferability between geographical domains. Hence, one may wish to abandon the use of reference-based-calibration altogether, and rely on a scheme of relative rating among several sites included in a single study. This concept was originally proposed by Chiu and Guttrop (2006).

Let us demonstrate this idea in the context of the LHFI using the 1997 PSL study. It has been well documented (e.g. local news, residents’ forums) that the health conditions of Thornton Creek (Site #18) is commonly considered “extremely poor,” even without the need to record physical measurements. For instance, consider LHFI(j). Thus, its value for Thornton can act as a baseline value for other sites. Let us now take the LHFI(j) for Site #1 situated along Big Bear Creek. It is −0.79, which is more than 3.2 times the (posterior) standard deviation above −2.27 for Thornton. A subject-matter expert may now translate this quantitative comparison back to practical terms, and decide on the overall degree of degradation for Site #1.

9 Conclusion

We have seen that the biological information contained in an LHFI based on any of the above models is highly comparable to that contained in existing ecological health indices. One of the reasons for the popularity of IBI and its variants among policy makers is that it is scalar-valued and thus readily interpretable. However, the disadvantages of their usage must not be overlooked. Major ones include (1) the loss of precision as well as statistical tractability through the various stages of a semi-quantitative procedure when forming the index, and (2) the lack of an obvious calibration scheme that would allow the simultaneous comparison between spatial domains. For (1), both the metric scoring stage and score weighting stage of forming an IBI mask information contained in the raw metrics. In contrast, LHFI is model-based and involves virtually no qualitative procedures and deals directly with the raw metrics; hence, its statistical properties can be readily determined, and it retains all the information directly available from the metrics. For (2), the simple addition
of effect terms in an ANO(CO)VA model is a standard practice in many scientific contexts for comparison among strata, and can be easily adapted to the LHFI model to account for field sites coming from different geographical domains. Furthermore, the largely statistical principles used to construct the LHFI by no means diminish the biological worthiness of the resulting index, as subject-matter expertise remains vital and necessary in the metric selection process prior to model fitting. In this sense, here statisticians play the role of non-experts in the “extended peer communities” of ecologists, and because they are naturally “closer to the problem” of developing quantitative methods, their contribution can only enhance the overall value of the methodology in ecological applications.\textsuperscript{13} Desirable statistical properties and scientific integrity aside, LHFI is also a single-number assessment of ecological health, which makes a user-friendly yet powerful alternative to IBI variants and SHIPSL.

For the 1997 PSL data, we propose two types of models: (A) Poisson models for taxa richness count metrics only, and (B) logit models for relative richness and abundance metrics that reside on a common scale (0–100%) involving no ambiguous metric calibration whatsoever. Not surprisingly, the more comprehensive models from Type (B) perform better in their ability to distinguish sites according to health. Ideally, a latent health factor model would account for informational overlap carried among metrics. To this end, we fitted models of each type that involved a general covariance structure for the random metric effects. However, the covariance was poorly estimated, likely due to the lack of data. To reduce the number of covariance parameters, we assumed independent metric effects with unequal variances, and all model parameters were generally well estimated. Imposing a structure on a non-diagonal covariance matrix would have been difficult in a Bayesian framework, and was therefore unattempted. Assuming fixed metric effects led to the most stable model fits, although such model misspecification ignores obvious differences in response variability due to metric, and would inevitably lead to improper inference of latent health and other parameters.

The methodology for constructing an LHFI demonstrated in this article is rooted in a

\textsuperscript{13}Terminology in quotes is taken from Fjelland (2002), page 168.
simple statistical concept of ANO(CO)V A model building, and may be easily adapted to any context of health assessment, be it ecological, medical, or otherwise. (For example, our methodology could be applied to create an LHFI to assess a person’s health, similar to the idea of a body mass index (BMI).) Once a list of relevant observable variables has been identified, constructing an LHFI reduces to an exercise of forming a statistical model that efficiently describes the relationship among these variables and the unobservable or latent health factor. Some variables may be explanatory to health, and vice versa for others. By applying such statistical modeling principles, the age-old hurdle of having different spatial and/or temporal domains is eliminated, and an effective comparison of health associated with multiple geographical and temporal regions may be achieved. Simply through fitting the model, one can also readily and simultaneously determine (1) the statistical properties of the health assessment, as well as (2) the significance of the impact on health for the observable factors under consideration. Therefore, our methodology is a simple but universal scientific approach that is potentially far reaching to any research discipline that may involve any type of health assessment.
APPENDIX

The formulation of Model (34)–(36) stems from the presence of non-nested random-effect terms in some levels of the Bayesian hierarchy. In a regression model, a nested random effect has a one-to-one correspondence to the response. For example, consider the regression

\[ W_{ij} = \tau_i + r_{ij}, \quad [\tau_i | \tau^*, \sigma^2_\tau] \sim N(\tau^*, \sigma^2_\tau), \quad [r_{ij} | \sigma^2_r] \sim N(0, \sigma^2_r) \]

Here, \( r_{ij} \) is a nested random effect since each \( r_{ij} \) leads to a single \( W_{ij} \). In contrast, each \( \tau_i \) yields several \( W_{ij} \)’s, and is therefore a non-nested random effect. The idea of a nested random effect is related to hierarchical centering in a multi-level Bayesian structure (Gelfand et al., 1995).

For (29), neither \( H_i \) nor \( \beta_j \) is nested, and mixing problems are encountered when we sample according to (30). To circumvent this problem, we rearrange terms in the regression, define \( b_j = \alpha + \beta_j \), and define an artificially nested random effect

\[ \psi_{ij} \equiv (\gamma_0_i - \alpha) + \gamma_1(x_i - \bar{x}) \quad \text{for all } j \text{ for each } i \]

so that (29) becomes

\[
\nu_{ij} = b_j + \psi_{ij} \quad \text{subject to} \quad \psi_{i1} = \psi_{i2} = \ldots = \psi_{i7} \\
\implies \psi_{i1} = \nu_{i1} - b_1 = \nu_{i2} - b_2 = \ldots = \nu_{i7} - b_7 \\
\implies \nu_{ij} = \nu_{i1} - (b_1 - b_j) \quad \forall \; j = 2, \ldots, 7
\]

Hence, for each \( i \), only \( \nu_{i1} \) needs to be sampled, and the remaining \( \nu_{ij} \)’s are determined by (57). Overall, the sampling hierarchy for random-effect terms in this formulation only involves

\[
[b | \alpha, \Sigma] \sim \text{MVN}(\alpha \mathbf{1}, \Sigma) \\
[\nu_{i1} | b_1, \gamma_1, \mathbf{x}, \sigma^2_H] \sim N(b_1 + \gamma_1(x_i - \bar{x}), \sigma^2_H) \\
[Y_{ijk} | \nu_{ij}] \sim \text{Poisson}(e^{\nu_{ij}})
\]

Then, \( H_i \)’s and \( \beta_j \)’s can be retrieved by manipulating the relations among the variables.
REFERENCES


Table 1: Models that correspond to **LHFI**(a)–(j). Empirically, all of (a)–(f) are perfectly correlated with each other, and similarly for (g)–(j).

<table>
<thead>
<tr>
<th>LHFI</th>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>Equation (6)</td>
<td>simple two-way ANOVA with fixed metric effects $\beta_j$’s</td>
</tr>
<tr>
<td>(b)</td>
<td>same as (a)</td>
<td>minor difference in prior distributions</td>
</tr>
<tr>
<td>(c)</td>
<td>Equations (29)–(30)</td>
<td>hierarchical ANOCOVA with $[\beta</td>
</tr>
<tr>
<td>(d)</td>
<td>same as (c), except ... $[\beta_j</td>
<td>\sigma_j^2] \overset{\text{ind}}{\sim} N(0, \sigma_j^2)$ instead of $[\beta</td>
</tr>
<tr>
<td>(e)</td>
<td>same as (a), except ... $[\beta_j</td>
<td>\sigma_j^2] \overset{\text{ind}}{\sim} N(0, \sigma_j^2)$ instead of fixed $\beta_j$’s</td>
</tr>
<tr>
<td>(f)</td>
<td>Same as (e), except ... fixed $\beta_j$’s instead of $[\beta</td>
<td>\Sigma] \sim MVN(0, \Sigma)$</td>
</tr>
</tbody>
</table>

**Logistic**

| (g)  | Equations (44)–(48) | hierarchical ANOCOVA with $[\beta|\Sigma] \sim MVN(0, \Sigma)$, but $\theta_s=0$ for $s=1,2$ |
| (h)  | Equations (44)–(48), (49)–(50) | same as (g), except $\beta = [\beta_1(1), \ldots, \beta_6(1), \beta_1(2), \beta_2(2), \beta_3(2)]^T$ |
| (i)  | same as (g), except ... $[\beta_j|\sigma_j^2] \overset{\text{ind}}{\sim} N(0, \sigma_j^2)$, $\sigma_1(1)=\sigma_2(1)=\sigma_3(1)$ for EPT |
| (j)  | same as (h), except ... $[\beta_j|\sigma_j^2] \overset{\text{ind}}{\sim} N(0, \sigma_j^2)$, $\sigma_1(0)=\sigma_2(0)=\sigma_3(0)$ for EPT |

Table 2: Correlation matrix for various health indices for the PSL in 1997.

<table>
<thead>
<tr>
<th></th>
<th>B-IBI</th>
<th>SHIPSL</th>
<th>LHFI(a)</th>
<th>LHFI(g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-IBI</td>
<td>1.00</td>
<td>0.97</td>
<td>0.82</td>
<td>0.90</td>
</tr>
<tr>
<td>SHIPSL</td>
<td>0.97</td>
<td>1.00</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>LHFI(a)</td>
<td>0.82</td>
<td>0.91</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>LHFI(g)</td>
<td>0.90</td>
<td>0.90</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
<td>2.5th %-ile</td>
<td>97.5th %-ile</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>--------</td>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td><strong>LHFI(c)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1.35</td>
<td>1.34</td>
<td>−2.21</td>
<td>4.78</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>−2.07</td>
<td>−2.07</td>
<td>−3.44</td>
<td>−0.68</td>
</tr>
<tr>
<td>$\sigma_1^2$</td>
<td>77.46</td>
<td>3.39</td>
<td>0.27</td>
<td>156.30</td>
</tr>
<tr>
<td>$\sigma_2^2$</td>
<td>43.75</td>
<td>1.93</td>
<td>0.21</td>
<td>80.84</td>
</tr>
<tr>
<td>$\sigma_3^2$</td>
<td>45.12</td>
<td>1.92</td>
<td>0.21</td>
<td>84.41</td>
</tr>
<tr>
<td>$\sigma_4^2$</td>
<td>51.42</td>
<td>1.83</td>
<td>0.21</td>
<td>83.16</td>
</tr>
<tr>
<td>$\sigma_5^2$</td>
<td>47.36</td>
<td>1.92</td>
<td>0.20</td>
<td>83.09</td>
</tr>
<tr>
<td>$\sigma_6^2$</td>
<td>124.30</td>
<td>6.53</td>
<td>0.44</td>
<td>292.50</td>
</tr>
<tr>
<td>$\sigma_7^2$</td>
<td>56.49</td>
<td>2.42</td>
<td>0.23</td>
<td>110.40</td>
</tr>
<tr>
<td>$\sigma_H$</td>
<td>0.50</td>
<td>0.49</td>
<td>0.36</td>
<td>0.72</td>
</tr>
<tr>
<td><strong>LHFI(d)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1.50</td>
<td>1.49</td>
<td>0.56</td>
<td>2.44</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>−2.07</td>
<td>−2.07</td>
<td>−3.47</td>
<td>−0.69</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>1.71</td>
<td>1.38</td>
<td>0.63</td>
<td>4.66</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>1.18</td>
<td>0.96</td>
<td>0.48</td>
<td>3.19</td>
</tr>
<tr>
<td>$\sigma_3$</td>
<td>1.20</td>
<td>0.97</td>
<td>0.48</td>
<td>3.31</td>
</tr>
<tr>
<td>$\sigma_4$</td>
<td>1.19</td>
<td>0.97</td>
<td>0.48</td>
<td>3.18</td>
</tr>
<tr>
<td>$\sigma_5$</td>
<td>1.24</td>
<td>1.01</td>
<td>0.49</td>
<td>3.38</td>
</tr>
<tr>
<td>$\sigma_6$</td>
<td>2.72</td>
<td>2.22</td>
<td>1.04</td>
<td>7.40</td>
</tr>
<tr>
<td>$\sigma_7$</td>
<td>1.39</td>
<td>1.13</td>
<td>0.54</td>
<td>3.87</td>
</tr>
<tr>
<td>$\sigma_H$</td>
<td>0.50</td>
<td>0.49</td>
<td>0.36</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>LHFI(e)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>1.51</td>
<td>1.50</td>
<td>0.58</td>
<td>2.50</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>1.70</td>
<td>1.39</td>
<td>0.64</td>
<td>4.63</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>1.17</td>
<td>0.95</td>
<td>0.48</td>
<td>3.16</td>
</tr>
<tr>
<td>$\sigma_3$</td>
<td>1.21</td>
<td>0.98</td>
<td>0.48</td>
<td>3.33</td>
</tr>
<tr>
<td>$\sigma_4$</td>
<td>1.18</td>
<td>0.97</td>
<td>0.47</td>
<td>3.18</td>
</tr>
<tr>
<td>$\sigma_5$</td>
<td>1.24</td>
<td>1.02</td>
<td>0.50</td>
<td>3.39</td>
</tr>
<tr>
<td>$\sigma_6$</td>
<td>2.82</td>
<td>2.26</td>
<td>1.06</td>
<td>8.07</td>
</tr>
<tr>
<td>$\sigma_7$</td>
<td>1.39</td>
<td>1.14</td>
<td>0.54</td>
<td>3.88</td>
</tr>
<tr>
<td>$\sigma_H$</td>
<td>0.60</td>
<td>0.59</td>
<td>0.43</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Table 3: Summary statistics of posterior draws for Poisson counts models corresponding to the given LHFI’s. Only diagonal elements of $\Sigma$ are given for (c). Note that the 2.5th and 97.5th percentiles form 95% credible intervals. All statistics are based on combining two Markov chains, except when marked with a ‘∗’ indicating just one chain due to mixing difficulties.
Table 4: Summary statistics of posterior draws for logit models corresponding to the given LHFI's. Only diagonal elements of $\Sigma$ are given for (g) and (h). See the Table 3 caption for more details.
Figure 1: Scatterplots among B-IBI, SHIPSL, and LHFI(a), (b), and (g) for the 1997 PSL benthic taxonomic data. The values of B-IBI and SHIPSL are available from Chiu and Guttorp (2006).
Figure 2: Distribution of the 1997 PSL count responses, broken down by metric.
Figure 3: Covariates’ indirect influence on metric responses through latent health. This dependence structure among the variables constitutes a hierarchical ANOCOVA model.
Figure 4: An illustration of the need for random intercept and slope terms in an ANO(CO)VA model for taxa richness.
Figure 5: A scatterplot showing the dependence between urbanization and each of B-IBI ('b'), SHIPSL ('s'), and LHFI(a) ('h') for the 1997 PSL data. The index values have been standardized for easy visual comparison.
Figure 6: Boxplots of posterior samples that produce LHFI(a)–(f) except (b). Note the different vertical scales for all plots besides (d) and (e).
Figure 7: Difference in root-mean-square-error of imputed values between the models producing $\text{LHFI}(d)$ and (e) (top panel), and $\text{LHFI}(i)$ and (j) (bottom panel). Each pair of models involve nine incomplete datasets generated randomly from the 1997 PSL data.
Figure 8: Boxplots of posterior samples that produce LHFI(g)–(j), with LHFI(d) shown for reference.