UNIVERSITY OF WATERLOO
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE
Working Paper

November 20, 2008

A Latent Health Factor Index Modelling Approach via Generalized Linear Mixed Models, with Application to Ecological Health Assessment

Grace S. Chiu,
University of Waterloo, Waterloo, Canada
Peter Guttorp,
University of Washington, Seattle, USA
Anton H. Westveld,
University of Nevada, Las Vegas, USA
Shahedul A. Khan
University of Waterloo, Waterloo, Canada
and Jun Liang
Canadian Institute for Health Information, Toronto, Canada

Summary. Multimetric indices (MMI's) are appealing scalar-valued tools for rating a set of systems with respect to conditions that are not directly measurable. In ecological health assessment, conventional MMI's often involve qualitative and arbitrary defintions that are restricted geographically and temporally. We propose a statistical-model-based approach for constructing MMI's. Our latent health factor index (LHFI) is obtained by estimating an unobservable *health factor* term in a mixed-effects analysis of covariance regression that directly models the relationship among metrics, health, and factors that can influence health. Our approach aims to address concerns about loss of information from data, spatio-temporal restrictions, and scientific integrity that are common to some conventional indices. To illustrate our methodology, we construct an LHFI for Puget Sound Lowland benthic systems. Various formulations of Poisson and logistic regression are considered in a Bayesian hierarchical framework, implemented by Markov chain Monte Carlo techniques. Our approach is also applicable to medical and other contexts.

Keywords: Analysis of covariance; Bayesian hierarchical models; Ecosystem health; Latent factor; Multimetric index; Random effects

Address for correspondence: Grace S. Chiu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada.
E-mail: gchiu@uwaterloo.ca

1. Introduction

Many disciplines from the biological sciences involve assessing underlying conditions with a single number computed based on various measurable characteristics. We generically refer to these conditions as *health* throughout this article. A familiar example is the *body mass index* (BMI), which combines a person's height and weight measurements to yield a scalar-valued quantification of obesity (perhaps a form of poor health). Scalar-valued assessments are naturally appealing for their structural simplicity and supposed ease of interpretation, particularly in decision making contexts such as disease diagnosis. However, how one should interpret BMI values in different situations has long been a contentious issue (e.g. López-Alvarenga *et al.*, 2003). To assess aquatic ecosystem health, conventional indices such as the *benthic index of biotic integrity* (B-IBI) (Kerans and Karr, 1994) and its variants (e.g. McCormick *et al.*, 2001) are similarly constructed by studying and combining indicator variables, or *metrics*, to reflect underlying health conditions of field sites in a single spatial domain. Disadvantages of relying on these conventional indices have been well documented (e.g. Steedman and Regier, 1990; Chiu and Guttorp, 2004, 2006). IBI variants are a type of reference-based health indices; *observed-to-expected* (O/E) indices via the *river invertebrate prediction and classification system* (RIVPACS) approach are another (see Hawkins *et al.*, 2000). For reference-based indices, test sites (sites whose health is under scrutiny) are gauged against reference sites that are ideally comparable to the test sites in every aspect except for their pristine "reference" conditions. Sadly and realistically, however, so-called pristine sites either no longer exist due to widespread environmental degradation across the globe, or are inaccessible to scientists due to their remoteness. Consequently, the definition of reference criteria are often admittedly arbitrary (e.g. Hawkins *et al.*, 2000). For these three and perhaps other existing scalar-valued health indices, the fundamental issue lies in the ambiguity of the scheme used to construct an index that allegedly reflects unobservable conditions of interest.

Statistically speaking, conventional schemes could appear *ad hoc* due to a high degree of arbitrariness and the lack of a unified approach in several stages of quantifying qualitative features. Take the aforementioned reference-based indices, for instance. Stage 1 involves testing, analyzing, and/or 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. Discriminant analysis is popular for the screening of candidate metrics (e.g. Hawkins *et al.*, 2000; Stoddard *et al.*, 2005), although some scientists caution against the use of standard multivariate statistical techniques for this stage (e.g. see Brinck (2002), page 7), preferring less systematic (thereby potentially arbitrary) approaches. However, how best to define and choose metrics is beyond the scope of our article. The main concern of our work is the potential statistical inadequacies of an index that results from the following stages.

In the calibration stage, common attempts to overcome the lack of truly pristine reference sites is to use the best sites available in the context of the study, which themselves vary in quality (Clarke *et al.*, 2003). This is true both within and across geographical-temporal domains. Yet, this variation is not formally or systematically accounted for, as reference conditions are often externally standardized then treated as invariant (e.g. see MMI and O/E index in Stoddard *et al.*, 2005). Admittedly, such practice leads to arbitrariness / ambiguity / inconsistencies in the definition of reference-based health indices. In particular, for the RIVPACS approach, the last stage of forming the O/E index involves predicting the number of certain taxa, conditioned on the currently observed reference data that often span across several ecoregions and time periods. Consequently, test sites from different geographical-temporal domains could be identified as belonging to the same group with respect to health. This may appear unsatisfying, as recognizable difference in space and time among sites either plays no explicit role in how sites are categorized, or is *a priori* assumed insignificant.

For IBI variants, the calibration stage further involves devising a standardization or *scoring* scheme for converting the selected metrics into scores that share the same scale. However, there is no universally accepted scheme here. For instance, the B-IBI scheme relies on qualitative ideas of health to map metrics onto a discrete scale of $\{1, 3, 5\}$, while a more mathematical approach is favoured by Stoddard *et al.* (2005) to map metrics onto a continuous scale of [0, 10] for their multimetric index (MMI). Moreover, to this date, no obvious strategy exists for ensuring that metrics calibrated against reference sites from one domain would effectively reflect ecosystem health of another. The last stage involves deciding on a weighting scheme for the *metric scores* to form the index. What constitutes an effective weighting scheme is traditionally an open-ended question. Equal weighting is common, leading to an index that is the sum of all metric scores. However, each metric reflects a different but not necessarily disjoint aspect of health, and the overlap in informational content is not easily quantifiable (Ter Braak, 1987). Similar semi-qualitative practices are common for gauging the health of ecosystems in general (e.g. see Jørgensen *et al.* (2005)). This type of statistically *ad hoc* approach poses extreme challenges to proper statistical assessment of the indices and comparison of health across spatial and/or temporal domains. Despite the issues, IBI variants remain popular among policy makers 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 index construction. To address the statistical inadequacies of the latter two stages, Chiu and Guttorp (2006) propose the SHIPSL scheme for pooling metrics to form an index that is also scalar. 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 prefer not to assess health explicitly. Billheimer *et al.* (1997) and Bunea *et al.* (1999) directly model the metric-to-metric relationship, thereby removing all intermediate stages, each of which could mask valuable information on health, resulting in artificial precision (as in metric scores being restricted to 3 discrete values only) or unnecessary variability (as in sensitivity to field noise discussed by Chiu and Guttorp, 2006). These statistical models, e.g. state-space and graphical models, often expressed in a multivariate framework, can effectively describe underlying conditions of an ecosystem. Ter Braak (1986) prefers eigenvalue-based multivariate analyses to describe the relationship between species abundance and environmental variables. However, none of these may be immediately useful to resource managers due to the complex messages embedded in a multivariate system. Here, scalar-valued MMI's appear more appealing.

In this article, we investigate an entirely new approach for constructing MMI's. It combines the statistical integrity of model-based techniques and the interpretability of conventional indices such as the BMI and B-IBI. We focus on the freshwater ecosystem health assessment problem to demonstrate our approach. For metrics that have been identified from Stage 2 as informative, we model their interdependence by regressing them on univariate factors, one of which is latent health. This latent, unobservable factor can be estimated statistically, thereby yielding a scalar, numerical assessment of underlying health conditions. Other observable covariates may be included, such as physical traits of the streams (e.g. stream order), the spatial and/or temporal locations of sites, and human demographic variables that could directly influence site health. The resulting *latent health factor index* (LHFI) can then be compared to existing indices for the same data. If both are deemed to contain similar information about health, then the model-based LHFI would be preferred in general. This is because (1) the process undertaken to define the index is based almost entirely on standard modelling principles (hence, is much less arbitrary), (2) its performance is highly tractable in the statistical sense, (3) it is expected to be much less sensitive to random variability (field noise) since no intermediate *ad hoc* stages are present to distort valuable information, and (4) when covariates appear in the model in a latent regression to explain the health factor, then (a) prediction of site

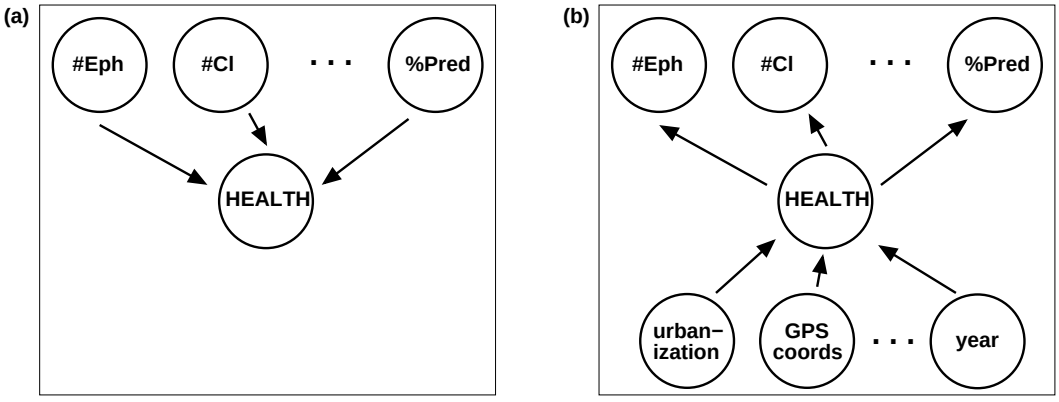

Fig. 1. (a) Conventional indices: health measurement is driven by observable metrics; (b) latent health factor modelling: a hierarchical framework in which metric responses are driven by unobservable health, which, in turn, is influenced by auxiliary covariates

health and its inference are straightforward and unambiguous (unlike conventional methods), and (b) the fitted model can help resource managers to identify external factors that have a direct influence on health. Specifically, the significance of their impact on health can be statistically assessed and classified, thus readily providing policy makers with unambiguous guidelines for prioritizing conservation measures. Indeed, both ranking sites with respect to health and classifying factors according to their impact on health can be achieved in a single step of fitting the model.

The rationale and basic principles of our methodology are given in Section 2. Statistical inference for health, including its prediction, is discussed in Section 3. We illustrate our methodology in Sections 4 and 5 by applying it to the 1997 Puget Sound Lowland (PSL) taxonomic data that appear in Chiu and Guttorp (2006). Corresponding values of SHIPSL and B-IBI are compared to those of LHFI, and their statistical and biological performance contrasted. (No documented O/E index values exist for these data.) In Section 6, we suggest how in practice one may make biological interpretations from a model-based LHFI without relying on external or prescribed reference conditions. Overall findings and advice are summarized in Section 7. Some technical details omitted in the text appear in the Appendices. The reader should note that the purpose of our article is to motivate and demonstrate the methodology of latent factor modelling for assessing health; it is not to use the particular 1997 PSL data to build an ecosystem health index.

2. A fully quantitative modelling framework for the LHFI

Consider assessing freshwater ecosystem health, such as for streams. Typically, benthic taxonomic data are collected by inserting some form of fixed-size shovel into the mud, separating the animals from the collected mud, then sorting each animal into one of many taxa. This collection of animals forms the field sample. It is common to collect replicate field samples per site. Ecologists identify various numerical aspects of the field sample composition to reflect ecosystem health. For example, an abundance of animals from predatory taxa would reflect a healthy ecosystem that can sustain a large number of predators. Similarly, a field sample rich in stress-sensitive taxa would point towards an ecosystem that has been subject to minimal stress. These numerical indicators of health are known as metrics when used to construct health indices. We rely on subject-matter expertise for what constitutes a biologically meaningful metric. On the other hand, our statistical expertise may

Table 1. Sites sampled from the PSL in 1997, and metrics identified in ecological studies to be effective indicators of stream health for the PSL

site		*metric*		
name	*location*	*label*	*characteristic*	*type*
BB1	Big Bear Creek	#Tx	all taxa	richness* (count)
BB2		#Eph	*Ephemeroptera* taxa	richness
BB3		#Ple	*Plecoptera* taxa	richness
BB4		#Tri	*Trichoptera* taxa	richness
BB5		#LL	long-lived taxa	richness
BS1	Big Soos Creek	#Intol	intolerant taxa	richness
JE1	Jenkins Creek	#Cl	clinger taxa	richness
LB1	Little Bear Creek	%Tol	tolerant taxa	abundance† (%)
LB2		%Pred	predatory taxa	abundance
LB3		%Dom3	3 most dominant taxa	abundance
LB4				
MA1	May Creek			
MI1	Miller Creek	*# distinct taxa of given characteristic appearing in field sample		
RO1	Rock Creek			
SW1	Swamp Creek	$^{\dagger}\frac{100 \times \text{\# animals of given characteristic in field sample}}{\text{total \# animals in field sample}}$		
SW2				
SW3				
TH1	Thornton Creek			

play an important role in using these metrics to construct a biologically meaningful health index, as the construction scheme must effectively account for the information that the metrics contain.

By combining metrics (scored or raw) to form a health index, conventional schemes for IBI variants and O/E indices essentially regard health as the response variable and metrics as covariates or driving factors of the health measurements (Fig. 1(a)). In reality, metrics are indicators of health, i.e. the underlying health is what drives the metric measurements. Thus, in a statistical model, metrics would appear more naturally as response variables, to be explained by health in the form of a latent covariate (top two tiers of Fig. 1(b)). This role reversal is fundamental to the scientific integrity of our index construction approach, as the relationship between health and metrics is directly modelled in an intrinsically quantitative framework without any ambiguous variable manipulation. Moreover, our approach allows health to appear hierarchically in a latent regression on auxiliary variables (e.g. urbanization, geography, year — bottom tier of Fig. 1(b)) that have a potential impact on the field site's overall health conditions. Altogether, this framework constitutes a hierarchical model that relates health, metrics, and auxiliary covariates simultaneously.

To illustrate the concepts of our LHFI modelling approach for gauging ecosystem health, we consider the 1997 PSL benthic taxonomic data as appear in Chiu and Guttorp (2006). These data were collected from 18 sites scattered over 9 streams (Table 1). Each site yielded 3 replicate field samples. For the PSL, an animal in the field sample could belong to one of 80 taxa, and the animal count per taxon could range from 0 to more than 1,000, but is typically equal or close to 0. Biologists have previously identified 10 useful metrics for the PSL (see Table 1), whose values are computed based on the 80 counts. Here, all 10 metrics are highly correlated due to their definitions: 7 describe taxa richness (count), and 3 describe relative abundance (%). To consider the relationship between metrics and health, non-Gaussian multivariate-response models can easily account for both count and % data types simultaneously, but such models often require complex parametrizations. Instead, we take a simpler approach by using the principles of analysis of (co)variance (ANO(CO)VA).

2.1. Building the latent health factor model

Researchers in various disciplines have employed the explicit statistical estimation of latent quantities to assess unobservable traits of interest (e.g. Hays *et al.*, 2000; Pietrobon *et al.*, 2004; Rosas, 2008; Stock and Watson, 1989; Ward and Hoff, 2007). Here, we exploit the practical appeal of this approach in biomonitoring and environmental policy making. Specifically, we aim to retain the widely accepted soundness of the multimetric approach, and the popularity of scalar-valued assessment among policy makers.

For benthic data, let Y_{ijk} denote the (possibly transformed) value of the jth metric for the ith site's kth independent replicate, where i=1,...,n, j=1,...,J, and k=1,2,...,K. For the PSL, we have n=18, J=7 or 9 (explained at the end of Section 2.2), and K=3. Naturally, the response Y_{ijk} can be explained by H_i — the underlying health of site i, and β_j — the block effect due to metric j, in an analysis-of-variance (ANOVA) generalized linear mixed model (GLMM):

$$\nu_{ij} = H_i + \beta_j \tag{1}$$

where ν_{ij}=$g\big(E[Y_{ijk}]\big)$, and $g(\,)$ is an appropriate link function. For randomly chosen sites, the unobservable health factor H_i is considered random, and in turn can be explained in a latent regression:

$$H_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + \varepsilon_i \tag{2}$$

where $\boldsymbol{x}_i$ is a vector of observable auxiliary covariates that may influence site health, $f_{\boldsymbol{\theta}}(\,)$ is the regression function with coefficients $\boldsymbol{\theta}$, and ε_i's are independent and identically distributed (iid) 0-mean errors. Our main interest is in H_i; its estimate $\widehat{H}_i$ is obtained by fitting the model to the observed Y_{ijk}'s and $\boldsymbol{x}_i$'s. Although health itself is latent, $\widehat{H}_i$ is an explicit quantification of site health. For the β_j's, we model them as 0-mean random-effects with an appropriate covariance structure. Altogether, (1)–(2) constitute a hierarchical ANOCOVA GLMM.

2.2. Model for combining spatial and other types of domains

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

Suppose our J metrics are deemed adequate for spatial domains A and B, one or both of which could lie within the PSL. 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 approach for IBI variants would require recalibration of all J metrics to account for the different spatial scales. (Spatial differences are simply unaccounted for with O/E indices.) Painstaking effort aside, personal preferences could play a heavy role in this recalibration, 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 pooled sites. However, a simple arithmetic mean overlooks the fact that sites a_i's are more similar among themselves than when compared to b_i's. Chiu and Guttorp (2006) advocate the *gold standard* scheme with pre-determined region-specific values to replace the sample mean and SD, but warn that implementation could be challenging in practice.

With our latent factor model, we can handle this issue properly, while avoiding the complexity of so-called spatial models that directly address the spatial correlation patterns. In fact, based on our experience and communications with various ecologists, the sparsity / large variability of ecological data of this sort typically prevents any underlying spatial correlation pattern to be statistically detectable. As formal applications of spatial models are impractical, a reasonable compromise, without loss of biological or statistical integrity, may be the following.

Let $Y_{ijk\ell}$ denote the value of the jth metric from replicate sample k in the ith site within spatial domain ℓ. In the presence of auxiliary covariates, a spatial effect term λ_ℓ can be an added as follows:

$$\nu_{ij\ell} = H_{i(\ell)} + \beta_j\,, \qquad H_{i(\ell)} = \lambda_\ell + f_{\boldsymbol{\theta}}(\boldsymbol{x}_{i\ell}) + \varepsilon_{i(\ell)}$$

with i nested in ℓ. Here, the λ_ℓ's may be modelled as random or fixed depending on the context. If covariates are absent, then the term $f_{\boldsymbol{\theta}}(\boldsymbol{x}_{i\ell})$ can be removed from the latent regression. In either case, 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.

The same principles may be applied to contexts of multiple temporal domains, whereby one would introduce a temporal effect term (e.g. *year*), possibly ordinal, to the model in a similar manner as above. To account for both types of domains, a spatio-temporal interaction term may be included. Similarly, *stream order* (size category of stream) could be considered a type of domain, and modelled in a similar fashion.

Since we have experienced much difficulty in obtaining quality taxonomic data that span several domains, we currently cannot demonstrate an application of this approach. Nevertheless, we are unaware of existing articles on biomonitoring that account for spatial and temporal differences in a truly sound manner; ours attempts to do so. Application of our methodology to inter-regional and -temporal data are currently in preliminary stages, conducted in collaboration with aquatic ecologists. In the remainder of the article, we will focus on fitting (1)–(2) only. In Section 4, we use Poisson ANO(CO)VA models to construct LHFI's for the PSL based on the 7 taxa richness metrics only. In Section 5, we apply a natural transformation to these 7 metrics to form 6 new relative richness metrics, combine them with the 3 relative abundance metrics, then model them altogether as a logistic ANOCOVA in two different formulations.

3. Computing the LHFI: model inference by Bayesian estimation

According to Gelman and Hill (2007), the hierarchical Bayesian framework is the most direct way to handle models with latent structures, as each level of latent regression in the *model hierarchy* has a direct correspondence to a specific level in the *parameter hierarchy*. Indeed, many existing works on modelling latent quantities utilize Bayesian inference. As a bonus, unlike some classical techniques, this framework does not rely on asymptotics that may be inappropriate due to small sample sizes and/or unbalanced designs that are common in ecological and other contexts. Here, we apply Bayesian inference to H_i's and other nuisance quantities in our latent health factor model (1)–(2).

Let $\boldsymbol{H} = (H_1, \ldots, H_n)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$. Let $\boldsymbol{\nu}$ denote the vector of ν_{ij}'s, $\boldsymbol{Y}$ denote the vector of Y_{ijk}'s, $\mathbb{X}$ denote the design matrix whose rows are $\boldsymbol{x}_i$'s, and $\boldsymbol{\Omega}$ denote the vector of remaining model parameters, including $\boldsymbol{\theta}$ and those from the distributions of $\boldsymbol{\beta}$ and ε_i's. For our model in a Bayesian context, all but $\mathbb{X}$ are considered random quantities.

Next, let $P(\,)$ be the generic label for a probability distribution. Then, $P(\boldsymbol{\Omega})$ is the prior distribution of $\boldsymbol{\Omega}$, $P(\boldsymbol{Y}|\boldsymbol{\nu})$ or $P(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{\beta})$ is the likelihood, $P(\boldsymbol{H}|\boldsymbol{\Omega},\mathbb{X})$ is the distribution of $\boldsymbol{H}$, and $P(\boldsymbol{\beta}|\boldsymbol{\Omega})$ is the distribution of $\boldsymbol{\beta}$. In the absence of concrete preconceptions of $\boldsymbol{\Omega}$, a diffuse (nearly flat) prior $P(\boldsymbol{\Omega})$ is commonly applied. Assumptions about $[\boldsymbol{Y}|\boldsymbol{\nu}]$, $f(\,)$, $\boldsymbol{\beta}$, and ε_i's determine the remaining distributions. Bayesian inference for $\boldsymbol{H}$, $\boldsymbol{\beta}$, and $\boldsymbol{\Omega}$ is made based on $P(\boldsymbol{H},\boldsymbol{\beta},\boldsymbol{\Omega}|\boldsymbol{Y},\mathbb{X})$, the joint posterior distribution of $\boldsymbol{H},\boldsymbol{\beta}$, and $\boldsymbol{\Omega}$. We assume independence of $\boldsymbol{H}$ and $\boldsymbol{\beta}$, so that

$$\begin{aligned} P(\boldsymbol{H},\boldsymbol{\beta},\boldsymbol{\Omega}|\boldsymbol{Y},\mathbb{X}) &\propto P(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{\beta},\boldsymbol{\Omega},\mathbb{X})\,P(\boldsymbol{H},\boldsymbol{\beta}|\boldsymbol{\Omega},\mathbb{X})\,P(\boldsymbol{\Omega},\mathbb{X}) \\ &= P(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{\beta})\,P(\boldsymbol{H}|\boldsymbol{\Omega},\mathbb{X})\,P(\boldsymbol{\beta}|\boldsymbol{\Omega})\,P(\boldsymbol{\Omega})\,. \end{aligned} \tag{3}$$

Estimating H_i's is our main interest; here, we take the posterior mean to be our LHFI (although other relevant posterior statistics are possible, e.g. posterior mode). That is,

$$\widehat{H}_i \equiv \widehat{H}_i(\boldsymbol{Y}, \mathbb{X}) = E(H_i|\boldsymbol{Y}, \mathbb{X}) = \int H_i \int\int\int P(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega}|\boldsymbol{Y}, \mathbb{X})\, d\boldsymbol{\beta}\, d\boldsymbol{\Omega}\ \ d\boldsymbol{H}_{-i} \qquad (4)$$

where $\boldsymbol{H}_{-i}$ is obtained by removing H_i from $\boldsymbol{H}$. Uncertainty in the estimation can be assessed by *highest posterior density* (HPD) intervals, available through statistical packages such as `boa` in R (Smith, 2007). Once the posterior in (3) is determined, obtaining HPD intervals is straightforward and unambiguous. In contrast, confidence intervals for existing indices such as IBI and SHIPSL variants rely on the non-parametric bootstrap, and are negatively biased in location and width in general (Chiu and Guttorp, 2006).

Note that closed forms may not exist for (3) or (4). In this case, one can simulate samples from (3) by numerical methods such as Markov chain Monte Carlo (MCMC), widely available in software packages such as OpenBUGS (Thomas *et al.*, 2006). Approximating (4) based on drawn posterior samples is then trivial. The remaining nuisance quantities can be estimated in a similar fashion.

3.1. Predicting site health

Another disadvantage to monitoring ecosystem health with common indices is the inability to make reasonable inference on the predictions of site health. For IBI / SHIPSL variants and O/E indices, one might predict an index value by inputting hypothetical raw richness / abundance counts or the corresponding metric values. However, as metric values themselves indicate health, the logic behind this prediction appears to be circular. Alternatively, one may first compute the index values, then regress them on auxiliary variables and make predictions of future site health via the fitted regression. For this two-step approach, inference on predicted values depends on the assumptions about the distribution of the index values. How might one incorporate into these assumptions the variability of metrics that form the index? The answer is far from being clear.

In contrast, prediction of the LHFI at site i and its inference is much more straightforward with our hierarchical ANOCOVA model (1)–(2), via the posterior predictive distribution $P(H^*|\boldsymbol{Y}, \mathbb{X}, \boldsymbol{x}^*)$, where a "$*$" denotes a future value. One can take the predicted LHFI for this site to be $\widehat{H}^* = E(H^*|\boldsymbol{Y}, \mathbb{X}, \boldsymbol{x}^*)$. Specifically, first consider a single Monte Carlo sample from the joint posterior (3). Extract from this sample those components of $\boldsymbol{\Omega}$ that are relevant to (2). Now, substitute these components together with $\boldsymbol{x}^*$ into (2) to simulate a Monte Carlo draw from $P(H^*|\boldsymbol{Y}, \mathbb{X}, \boldsymbol{x}^*)$. Repeat this process until a collection of simulated draws are obtained from $P(H^*|\boldsymbol{Y}, \mathbb{X}, \boldsymbol{x}^*)$. Then, $\widehat{H}^*$ is approximated by the mean of this collection. Predictive HPD intervals based on $P(H^*|\boldsymbol{Y}, \mathbb{X}, \boldsymbol{x}^*)$ are also easily approximated using appropriate quantiles of the simulated posterior predictive draws.

The advantage of our predictive inference approach is that it accounts for the modelled relationship among metrics, health, and auxiliary variables simultaneously in an unambiguous fashion.

4. Three LHFI's for the PSL based on taxa richness

We apply our modelling methodology described in previous sections to construct LHFI's for the PSL. To avoid handling metrics on different scales, we first restrict our attention to the J=7 count-valued richness metrics (Table 1). A Poisson likelihood appears to be appropriate here. Thus, a possible initial model can be the simple ANOVA (1) that assumes site and metric effects to be independent and normally distributed:

$$[Y_{ijk}|\nu_{ij}] \overset{\text{ind}}{\sim} \text{Poisson}(e^{\nu_{ij}})\,,\ \ \nu_{ij} = H_i + \beta_j\,,\ \ [H_i|\alpha, \sigma_H] \overset{\text{iid}}{\sim} N(\alpha, \sigma_H^2)\,,\ \ [\beta_j|\sigma_j] \overset{\text{ind}}{\sim} N(0, \sigma_j^2)\,. \qquad (5)$$

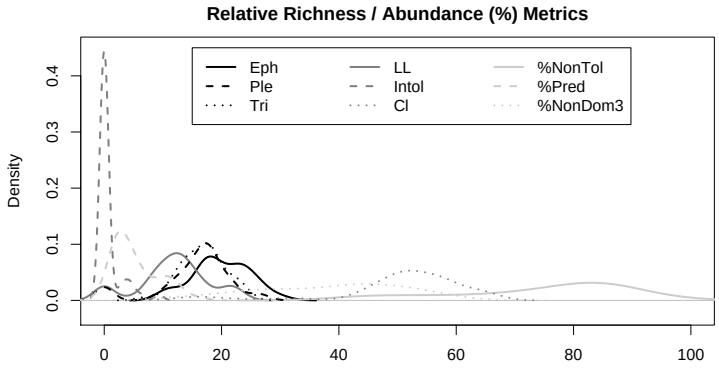

Fig. 2. Distributions of relative richness and abundance metrics from the 1997 PSL benthic taxonomic data

Note that (5) implicitly addresses possible overdispersion of taxa richness counts at the across-site and across-metric level (McCulloch and Searle, 2001), which is the level of main concern for overdispersion in biomonitoring studies. (Biomonitoring data are typically too sparse for modelling overdispersion on a finer scale.) Note also that the vast range of Var(Y_{ijk}) over j may be attributed to a combination of non-constant σ_j's and the Poisson link (see Fig. 2, curves in dark grey and black, corresponding to $100\,Y_{ijk}/Y_{1jk}$ for i=2,...,7, where Y_{1jk} is #Tx; i.e. they are the percentage counterparts of the richness metrics currently being considered — see Section 5; although not shown, the distributions of the Y_{ijk}'s are comparable to these six in shape and relative location). The former assumption of metric effects heterogeneity can be removed in subsequent reduced fits if it is deemed unncessary based on model diagnosis for (5).

Diffuse priors are given to the elements of $\boldsymbol{\Omega} = (\alpha, \sigma_H, \sigma_1, \ldots, \sigma_7)$:

$$\alpha \sim N(c_1, c_2)\,, \qquad \sigma_H^2, \sigma_j^2 \sim \text{inverse-gamma}(c_3, c_4)\ \ \forall j \tag{6}$$

where c_1=0, c_2=100, and c_3=c_4=1 are hyperparameter values chosen to impose diffuseness. Other values of c's that correspond to more diffuseness were also used, but they led to minimal change in the model fit, and will not be discussed further. Similarly, we chose normal and inverse-gamma priors for ease of implementation, but normality or otherwise generally plays little to no role in the inference provided that the prior $P(\boldsymbol{\Omega})$ is diffuse.

Priors in (6) and Model (5) (and all subsequent models) were implemented with OpenBUGS after partial hierarchical centring (see Appendix A). Based on two Markov chains of posterior draws of $(\boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\Omega})$ generated from different initial values, all unknown quantities were well estimated. We combine both chains to obtain $\widehat{H}_i$'s, labelled as LHFI(5). These index values and corresponding 95% HPD intervals appear in black in Fig. 3, top panel. Posterior summary statistics for $\boldsymbol{\Omega}$ appear in Table 2. Details of MCMC sampling appear in Appendix A.

In addition to metrics data for the 18 PSL sites, associated with each stream are data for auxiliary covariates, taken from Morley (2000), that include *urbanization* and *Global Positioning System (GPS) co-ordinates*. Urbanization is the percentage of total impervious area in the sub-basin to which the stream belongs. Thus, some sites share the same urbanization value. GPS co-ordinates recorded with sensitive instruments appear as latitudes and longitudes that are unique to each site. To additionally account for the potential influence of these covariates on health, we consider them and LHFI(5) in scatterplots in Fig. 4. (Note that the latitude scale shown has been shifted by -47 and

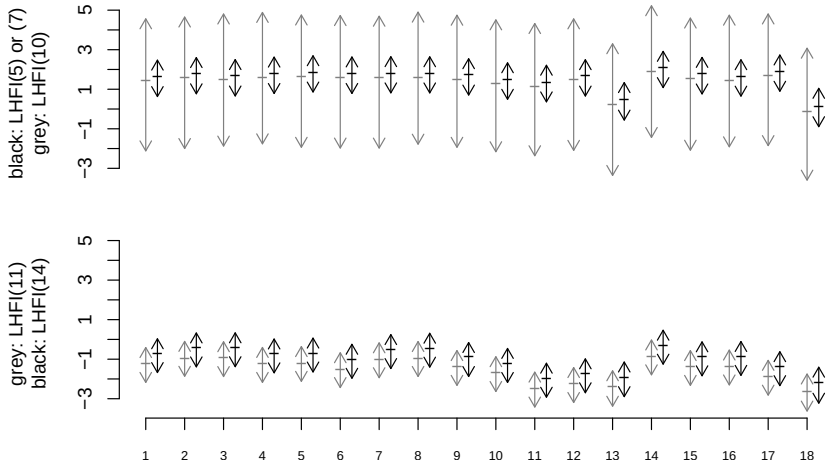

Fig. 3. Ninety-five percent HPD intervals for Poisson-based LHFI's (top panel) and logit-based LHFI's (bottom panel); a '—' denotes the observed index value

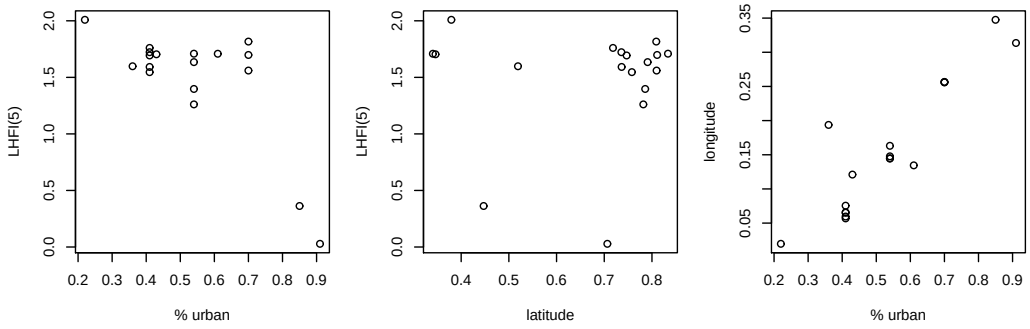

Fig. 4. The relationship among Poisson-based latent health and covariates for the PSL in 1997

Table 2. Summary statistics of posterior draws for Poisson counts models

	mean	*median*	*2.5th HPD %-ile*	*97.5th HPD %-ile*	*MC error*	*# draws*
			Model (5): DIC=1354.0			
α	1.51	1.51	0.57	2.46	0.01	20 000
σ_1	1.71	1.38	0.46	3.66	0.01	
σ_2	1.20	0.97	0.35	2.56	0.01	
σ_3	1.19	0.98	0.36	2.52	0.01	
σ_4	1.18	0.96	0.40	2.56	0.01	
σ_5	1.24	1.01	0.37	2.64	0.01	
σ_6	2.75	2.25	0.77	5.92	0.02	
σ_7	1.39	1.12	0.42	3.02	0.01	
σ_H	0.60	0.59	0.41	0.80	0.00	
			Model (7): DIC=1353.0			
α^*	1.56	1.55	0.65	2.52	0.01	10 000
γ_1	−2.06	−2.05	−3.44	−0.70	0.01	20 000
σ_1	1.71	1.38	0.44	3.75	0.01	
σ_2	1.19	0.97	0.38	2.52	0.01	
σ_3	1.21	0.98	0.36	2.62	0.01	
σ_4	1.20	0.97	0.34	2.55	0.01	
σ_5	1.26	1.01	0.37	2.72	0.01	
σ_6	2.74	2.24	0.81	5.93	0.01	
σ_7	1.39	1.12	0.39	3.00	0.01	
σ_H	0.50	0.48	0.34	0.68	0.00	
			Model (10): DIC=1353.0 *(only diagonal elements of* $\boldsymbol{\Sigma}$ *are given)*			
α	1.27	1.28	−2.11	4.57	0.02	10 000
γ_1	−2.07	−2.06	−3.46	−0.67	0.01	
σ_1^{2*}	28.63	3.39	0.09	62.91	7.26	5 000
σ_2^{2*}	17.93	1.87	0.09	34.32	5.47	
σ_3^{2*}	16.32	1.77	0.08	34.90	4.20	
σ_4^{2*}	17.85	1.79	0.07	36.00	4.59	
σ_5^{2*}	17.82	1.78	0.07	37.88	5.21	
σ_6^{2*}	87.49	6.66	0.08	114.80	37.13	
σ_7^{2*}	17.11	2.40	0.06	44.96	2.08	
σ_H	0.50	0.49	0.34	0.68	0.00	10 000

Note: Values for parameters with a '*' are based on one Markov chain only.

longitude by -122.) An approximately linear relationship is seen between health and urbanization, which is highly correlated with longitude. In contrast, latitude has little association with health. Note the two most degraded sites (TH1 and MI1) with LHFI(5) values below 0.4. They appear to drive the linearity between health and urbanization, and the low correlation between health and latitude. Nevertheless, disregarding these sites for the latent regression in (2) may be unwise in contexts of, say, habitat conservation — one indeed expects to find unhealthy ecosystems in highly urbanized sites, making urbanization (at least on the high end of the spectrum) an important factor in determining site health. To avoid regressing health on redundant or unnecessary variables, we consider a single covariate x_i, namely, urbanization. Altogether, we have the ANOCOVA model

$$\begin{aligned}[Y_{ijk}|\nu_{ij}] \overset{\text{ind}}{\sim} \text{Poisson}(e^{\nu_{ij}})\,, \quad \nu_{ij} = H_i + \beta_j\,, \quad H_i = \gamma_{0i} + \gamma_1(x_i - \overline{x})\,,\\ \left[\gamma_{0i}|\alpha, \sigma_H^2\right] \overset{\text{iid}}{\sim} N(\alpha, \sigma_H^2)\,, \quad \left[\beta_j|\sigma_j^2\right] \overset{\text{ind}}{\sim} N(0, \sigma_j^2)\,.\end{aligned} \tag{7}$$

Here, the covariate in the latent regression is centred by subtracting the mean $\overline{x}$, to remove dependence between α and γ_1. For γ_1, we take the same diffuse prior for α as in (6). That is,

$$\gamma_1 \sim N(c_1, c_2)\,. \tag{8}$$

Model (7) with priors from (6) and (8) for $\mathbf{\Omega} = (\alpha, \gamma_1, \sigma_H, \sigma_1, \ldots, \sigma_7)$ were fitted to the data via two Markov chains of posterior draws. The chains mixed exceptionally well except for minor mixing problems for α; we combined the chains to produce LHFI(7). Posterior summary statistics for $\mathbf{\Omega}$ appear in Table 2. See Appendix A for sampling details. Incidentally, LHFI(7) and LHFI(5) are virtually identical in value and 95% HPD interval (Fig. 3).

Finally, one might wish to consider as part of the model the dependency of the richness counts over sites and metrics, for the following reason. The nature of the dependence between pairs of richness counts is expected to vary by site and metric. Take i=BB1 and i'=BB2, for instance. Both sites are located along Big Bear Creek, and therefore Y_{ijk} and $Y_{i'jk}$ are highly dependent. Now, take i=BB1 and i'=TH1, the latter located along Thornton Creek; hence, Y_{ijk} and $Y_{i'jk}$ are possibly independent. Similarly, take j=1 (#Tx) and j'=2 (#Eph). As #Tx is obtained by adding #Eph to the number of other taxa, Y_{ijk} and $Y_{ij'k}$ are linearly correlated. Now, take j=2 and j'=7 (#Cl). Then, the covariance structure between Y_{ijk} and $Y_{ij'k}$ is intrinsically different and may not be linear, since some *Ephemeroptera* taxa fall in the clinger category, while others do not.

In Appendix B, we show that the dependence of pairwise covariance on (i, i') is already reflected by the latent regression of (2), and that having correlated β_j's can further allow us to account for the dependence on (j, j'). In particular, we replace the β_j distributions from (7) with $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_7)^T \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the variance-covariance matrix whose jth diagonal element is σ_j^2 and off-diagonal (j, j')th element is $\sigma_{jj'}$. In principle, one may wish to impose a covariance structure that is based on the conceptual relationship among metrics. However, except for some special structures, it is often challenging to efficiently sample from the posteriors of the covariance parameters (Westveld, 2007). Thus, we assume an unstructured $\mathbf{\Sigma}$ here. Then, the inverse-Wishart distribution is a popular choice for the prior of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma}^{-1} \sim \text{Wishart}(\mathbb{S}, c_5)\,, \tag{9}$$

parametrized in such a way that $E(\mathbf{\Sigma}) \propto \mathbb{S}$. Altogether, our third model is

$$\begin{aligned}[Y_{ijk}|\nu_{ij}] \overset{\text{ind}}{\sim} \text{Poisson}(e^{\nu_{ij}})\,, \quad \nu_{ij} = H_i + \beta_j\,, \quad H_i = \gamma_{0i} + \gamma_1(x_i - \overline{x})\,,\\ \left[\gamma_{0i}|\alpha, \sigma_H^2\right] \overset{\text{iid}}{\sim} N(\alpha, \sigma_H^2)\,, \quad \boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma})\,,\end{aligned} \tag{10}$$

with priors from (6), (8), and (9) for $\mathbf{\Omega} = (\alpha, \gamma_1, \sigma_H, \mathbf{\Sigma})$. For the hyperparameters in (9), we take c_5=7 and $\mathbb{S}$ to have diagonal values 1 and off-diagonal values 0.5. This reflects the prior notion that all 7 metrics are positively associated with the latent health factor, and hence, with each other. These hyperparameter values yield reasonably diffuse proper priors.

Again, two Markov chains of posterior draws were generated. However, despite hierarchical centring, we encountered mixing problems for many matrix entries of $\mathbf{\Sigma}$ (see Appendix A). Nevertheless, H_i's mixed well marginally, and hence, we can define LHFI(10) based on the combined chain. Index values and 95% HPD intervals are shown in grey in Fig. 3, top panel. Posterior summary statistics for selected elements of $\mathbf{\Omega}$ are in Table 2.

4.1. Discussion of results

As it turns out, pairwise correlations among the three LHFI's are all equal to 1.00. However, while we define an LHFI to be the posterior mean for H_i, we must also consider the reliability of this assessment of health. From Fig. 3, top panel, we see that latent health corresponding to LHFI(10) has substantially more uncertainty (longer HPD intervals) than LHFI(5) and (7), the latter two showing almost identical properties. In other words, although Model (10) in principle incorporates the natural correlation among metric values over sites and over metrics into the latent health factor model, the extra complexity of the model did little in practice to improve our inference. Of course, this larger model may prove to be beneficial when applied to other datasets.

For our 1997 PSL data, we prefer LHFI(5) and (7) based on simpler models. Despite nearly identical properties between the two indices, LHFI(7) from the hierarchical ANOCOVA model is more appropriate in practice, as Model (7) (as well as (10)) yields statistical evidence that urbanization has a negative impact on stream health: the 95% HPD interval for γ_1 is below zero (approximately -3.4 to -0.7; see Table 2). While this negative effect might have been a foregone conclusion from a biological point of view, our ANOCOVA models 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 in response to selected factors only. For instance, consider a model that regresses latent health on both urbanization and latitude. We fitted such a model in the framework of Model (7), but will not discuss the details except for the inference on the latitude effects. The inclusion of latitude has virtually no impact on the posterior distributions of the H_i's (or of other unknown quantities). In fact, a typical 95% HPD interval for the corresponding coefficient includes 0, suggesting statistically insignificant effect on health due to latitude. Thus, our latent factor hierarchical modelling approach provides the policy maker with a scientific mechanism to classify factors according to their impact on health: a negative HPD interval indicates detrimental effects, one that covers 0 indicates undetectable impact, and a positive HPD interval implies positive impact. (A technical note on this ranking scheme for multiple covariates is that the HPD credible level may require adjusting in the context of multiple testing; see Westfall *et al.* (1997), for instance.) Of course, as in any subject area, caution is required when interpreting statistical results: a statistically significant impact may result from very dense data in the absence of a true impact, although it is highly unlikely for ecological data as they are typically sparse; and a statistically undetectable impact (e.g. due to sparsity) does not preclude an actual impact. Nevertheless, a hierarchical ANOCOVA modelling approach for constructing health indices indeed provides some practical guidelines in cases where the effects of a number of factors must be assessed.

4.1.1. Quantitative comparison of (5) and (7)

Besides the practical perspective, a quantitative comparison between the simple ANOVA approach for LHFI(5) and the hierarchical ANOCOVA approach for LHFI(7) may be of interest.

A common basis of comparison is the use of the deviance information criterion (DIC). It assesses how well the Y_{ijk}'s are predicted, and can be used to compare performance among models for identical data (Spiegelhalter *et al.*, 2002). For our models, DIC values are readily available from the OpenBUGS output, and are shown in Table 2. (Theory behind the DIC is beyond the scope of our article.) Here, the DIC is 1354.0 for Model (5) and 1353.0 for Model (7) (and (10); see Table 2). Thus, there may be a slight gain in posterior predictive power by explaining health with a relevant covariate.

Alternatively, a model's predictive ability may be more concretely assessed by cross-validation with out-of-sample predictions. To this end, we randomly divided the 1997 PSL data ($18\times7\times3$=378 observations in total) into nine disjoint subsets of 42 values each. Then, on each subset, the following three-step procedure was performed: (i) The 42 values were removed from the data and treated as missing values. (ii) Each of Models (5) and (7) was fitted to the remaining 8×42=336 observations, with the 42 missing values imputed as unknown quantities within the Bayesian framework. (iii) 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.

To pool this information from all nine subsets, an empirical sum-of-squares-error (SSE) ratio, defined as SSE(5)/SSE(7), is computed based on

$$\text{SSE} = \sum_{\{\text{nine subsets}\}} \sum_{\{y\text{: missing values}\}} \left(y^{\text{estimated}} - y^{\text{true}}\right)^2 .$$

Using a type of SSE ratio to assess predictive ability is favoured by Ward and Hoff (2007), for instance. In theory, one could compare the two models by considering the respective posterior predictive distributions over all possible combinations of missing values. As this approach is practically infeasible, the comparison based on nine randomly generated datasets thus serves as a pilot study. Here, the SSE ratio is 0.97. Considering the variability inherent in a small pilot study, a 3% difference in SSE is not enough evidence that Model (5) predicts better than Model (7). Combining this result with that of the DIC, we conclude that both models predict observations equally well.

Aside from a model's predictive ability, perhaps the precision of the health assessment is of ultimate concern. Therefore, one may also compare the posterior distributions for σ_H, whose median is 0.59 with a 95% HPD interval of [0.41, 0.80] for the the simpler Model (5), and 0.48 and [0.34, 0.68], respectively, for the hierarchical Model (7) (see Table 2). That is, there is some (weak) evidence that Model (5) contains more uncertainty, despite the substantial overlap between HPD intervals. This makes intuitive sense, as the variability in H_i unexplained by α in Model (5) is further addressed by γ_1 and $\boldsymbol{x}$ in Model (7). (Note that σ_H is different from the dispersion of the H_i posterior distributions.)

In summary, while both Models (5) and (7) have very comparable predictive ability, the hierarchical Model (7) seems to contain slightly less uncertainty, and can provide guidelines on conservation measures, as already discussed in Section 4.1. Therefore, we will focus on LHFI(7) in subsequent discussions of taxa-richness-based LHFI's.

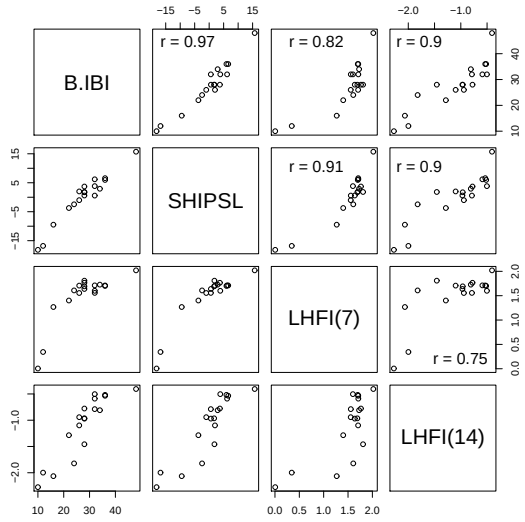

Fig. 5. Scatterplots among various health indices for the 1997 PSL data

4.1.2. Comparing LHFI's to existing health indices

For comparisons with documented values of B-IBI and SHIPSL, we refer to scatterplots in Fig. 5: the upper-left 3×3 panels show pairwise relationships among LHFI(7), B-IBI, and SHIPSL. There is a strong positive correlation ($r>0.8$) between our LHFI and either existing index, but the relationship is curvilinear. The curvature can be explained by the non-linearity of the model that produces LHFI(7), whereas both B-IBI and SHIPSL are linear combinations of metric scores. The strong correlation demonstrates that LHFI(7) is no less informative about the sites' health conditions, despite our excluding 3 metrics in its definition. In the next section, we propose a comprehensive model that accounts for all 10 metrics, thus improving the informational content of the LHFI.

5. Two comprehensive logistic LHFI's based on relative richness and abundance

Among various models, Chiu *et al.* (2007) consider a one-way ANOVA with latent health as the main factor for explaining sample cardinality, N_{ik} (total number of animals in the field sample), as the sole response. When fitted, this simple model demonstrates that N_{ik} contributes some information towards latent health, although it is traditionally not a metric in its own right. However, N_{ik} already appears implicitly as the denominator in the definition of relative abundance metrics. For the PSL, these metrics are %Tol,%Pred, and %Dom3 (Table 1). Note that some taxa are classified as neither tolerant nor intolerant; therefore, non-tolerant taxa are not necessarily intolerant. Furthermore, %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 %NonTol=100%$-$%Tol and %NonDom3=100%$-$%Dom3.

Indeed, Chiu and Guttorp (2004) show that it is beneficial, at least statistically, to convert taxa richness (count) metrics to percentages also, before combining them with relative abundance metrics to form a health index. They suggest removing #Tx from the metric list, but incorporating it as the denominator for transforming the other 6 count metrics into *relative richness* percentages, just as

Table 3. Summary statistics of posterior draws for logit models

	mean	*median*	*2.5th HPD %-ile*	*97.5th HPD %-ile*	*MC error*	*# draws*
			Model (11): DIC=4651.0			
α	−1.62	−1.63	−2.51	−0.68	0.01	20 000
γ_1	−2.03	−2.03	−3.59	−0.43	0.01	
$\sigma_{1(1)} = \sigma_{2(1)} = \sigma_{3(1)}$	0.87	0.58	0.10	2.36	0.01	
${\sigma_{4(1)}}^*$	2.31	1.03	0.12	7.21	0.08	10 000
${\sigma_{5(1)}}^*$	12.92	5.50	0.64	37.76	0.56	
${\sigma_{6(1)}}^*$	4.35	1.92	0.16	13.91	0.18	
${\sigma_{1(2)}}^*$	10.11	4.30	0.37	28.62	0.77	
${\sigma_{2(2)}}^*$	4.30	1.85	0.15	12.99	0.17	
${\sigma_{3(2)}}^*$	3.12	1.31	0.12	9.16	0.20	
σ_H	0.58	0.56	0.40	0.79	0.00	20 000
			Model (14): DIC=4606.0			
α	−1.07	−1.06	−1.97	−0.20	0.01	20 000
γ_1	−2.08	−2.09	−3.64	−0.49	0.00	
$\sigma_{1(0)} = \sigma_{2(0)} = \sigma_{3(0)}$	0.86	0.57	0.10	2.37	0.01	
${\sigma_{1(1)}}^*$	3.22	1.32	0.13	9.36	0.15	10 000
${\sigma_{2(1)}}^*$	17.41	7.02	0.78	47.77	0.85	
${\sigma_{3(1)}}^*$	3.28	1.38	0.14	9.42	0.15	
${\sigma_{1(2)}}^*$	7.55	3.22	0.25	22.43	0.39	
${\sigma_{2(2)}}^*$	5.85	2.58	0.20	18.33	0.16	
${\sigma_{3(2)}}^*$	2.71	1.00	0.12	7.08	0.22	
σ_H	0.58	0.57	0.40	0.79	0.00	20 000

Note: Values for parameters with a '*' are based on one Markov chain only.

how N_{ik} is used to define relative abundance. This way, all 9 variables now share the same scale; their distributions are shown in Fig. 2. Now, a GLMM similar in principle to those of Section 4 may be used to construct a comprehensive LHFI from these 9 metrics. To do so, each metric may be considered an observed *probability of success*, where "success" is an occurrence of the taxon (towards richness) or animal (towards abundance) indicative of a healthy stream. 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.

We focus on hierarchical models involving urbanization as a covariate for latent health. With three extra metrics here, we have $3\times3\times18$=162 additional observations for model fitting. However, $\mathbf{\Sigma}$ — the dependence among the nine metric effects — now also involves more unknown parameters. Indeed, Chiu *et al.* (2007) fit logistic extensions of Model (10) with the extra metrics, and demonstrate that the estimation of $\mathbf{\Sigma}$ remains difficult. Therefore, here we will only discuss logistic extensions of Model (7), with independently distributed metric effects.

Two groups of variables form our 9 metrics: J_1=6 pertaining to richness, and J_2=3 pertaining to abundance. Let s=1 denote 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 $p_{isj\bullet}$, where j=1,...,J_s. Finally, let $\nu_{isj\bullet}$ denote the logit-

transformed $p_{isj\bullet}$. Now, consider the GLMM

$$[Y_{isjk}|T_{is\bullet k}, p_{isj\bullet}] \overset{\text{ind}}{\sim} \text{Binomial}(T_{is\bullet k}, p_{isj\bullet}) \quad \forall\, k = 1,2,3, \qquad s = 1,2\,,$$
$$\ln \frac{p_{isj\bullet}}{1-p_{isj\bullet}} \;\equiv\; \nu_{isj\bullet} = H_i + \beta_{j(s)}\,, \qquad H_i = \gamma_{0i} + \gamma_1(x_i - \overline{x})\,, \tag{11}$$
$$[\gamma_{0i}|\alpha, \sigma_H] \overset{\text{iid}}{\sim} N(0, \sigma_H^2)\,, \quad [\beta_{j(s)}|\sigma_{j(s)}] \overset{\text{ind}}{\sim} N(0, \sigma^2_{j(s)})\,, \quad T_{is\bullet k} = \begin{cases} (\#\text{Tx})_{ik} & \text{if} \quad s = 1 \\ N_{ik} & \text{if} \quad s = 2 \end{cases}.$$

This model stipulates that the probability of success is affected by site health and the metric, but not by metric type (richness or abundance). Indeed, evidence from Chiu *et al.* (2007) (based on the more complex model with dependent metric effects) suggests that metric type is insignificant. Furthermore, their analyses also indicate evidence for a common variance for the metric effects of *Eph.*, *Ple.*, and *Tri.* taxa, often known collectively as EPT taxa. This agrees with the distributions of the EPT metrics (black in Fig. 2). Therefore, we additionally assume $\sigma_{1(1)}$=$\sigma_{2(1)}$=$\sigma_{3(1)}$. The prior distribution for $\mathbf{\Omega} = (\alpha, \gamma_1, \sigma_H, \sigma_{1(1)}, \sigma_{4(1)}, \sigma_{5(1)}, \sigma_{6(1)}, \sigma_{1(2)}, \sigma_{2(2)}, \sigma_{3(2)})$ and hyperparameters are as for Model (7).

Based on two Markov chains of posterior samples, all unknown quantities (including $p_{isj\bullet}$'s) were very well estimated, except for non-EPT $\sigma_{j(s)}$'s with similar mixing problems as for $\mathbf{\Sigma}$ entries from Model (10); see Appendix A. As no such problem was encountered for H_i's, we define LHFI(11) as the mean of the H_i draws from both chains combined. Index values and corresponding 95% HPD intervals appear in grey in Fig. 3, bottom panel. Posterior summary statistics for $\mathbf{\Omega}$ are given in Table 3. To investigate some assumptions about the latent regression in (11), we refer to Fig. 6. As for the Poisson-based LHFI's, we see no obvious violations of linearity between urbanization and the logit-based latent health, nor do we see the need to regress health on latitude.

Note that the formulation of Model (11) is based entirely on binomial distributions associated with the nine metrics. However, one could fine-tune the dependence among Y_{isjk}'s based on the disjoint nature of #Eph, #Ple, and #Tri that 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 EPT richness metrics, 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_{isjk}'s are binomial for s=1, 2. However, for s=0, we have

$$\left[\begin{pmatrix} Y_{i01k} \\ Y_{i02k} \\ Y_{i03k} \\ T_{i1\bullet k} - \sum_{j=1}^3 Y_{i0jk} \end{pmatrix} \middle| T_{i1\bullet k}, \begin{pmatrix} p_{i01\bullet} \\ p_{i02\bullet} \\ p_{i03\bullet} \end{pmatrix}\right] \sim \text{Multinomial}(T_{i1\bullet k}, \begin{bmatrix} p_{i01\bullet} \\ p_{i02\bullet} \\ p_{i03\bullet} \\ 1 - \sum_{j=1}^3 p_{i0j\bullet} \end{bmatrix})\,, \tag{12}$$

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

$$\ln \frac{p_{i0j\bullet}}{1 - \sum_{j=1}^3 p_{i0j\bullet}} \equiv \nu_{i0j\bullet} = H_i + \beta_{j(0)}\,. \tag{13}$$

Altogether, our binomial-multinomial mixture logit model is

$$\text{Equations (12)–(13) for } s\text{=0}, \quad \text{Equation (11) for } s\text{=1,2}, \quad J_s\text{=3 for all } s. \tag{14}$$

The prior for $\mathbf{\Omega} = (\alpha, \gamma_1, \sigma_H, \sigma_{1(0)}, \sigma_{1(1)}, \sigma_{2(1)}, \sigma_{3(1)}, \sigma_{1(2)}, \sigma_{2(2)}, \sigma_{3(2)})$ and hyperparameters are all as for Model (11) above. Mixing for the two Markov chains of posterior draws here is virtually

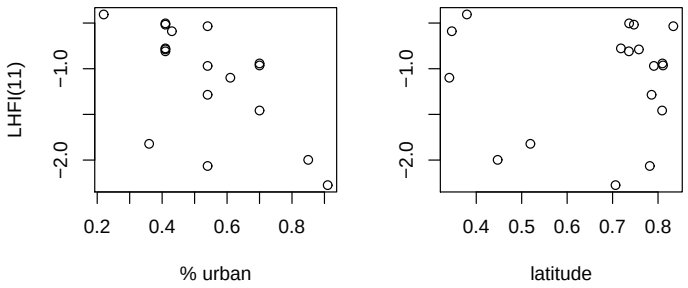

Fig. 6. The relationship between logit-based latent health and covariates for the PSL in 1997

identical as for the binomial-only model (see Appendix A). In the absence of mixing problems for the H_i chains, we combine both chains to form LHFI(14). Index values and corresponding 95% HPD intervals are given in black in Fig. 3, bottom panel. Posterior summary statistics for $\mathbf{\Omega}$ are in Table 3.

5.1. Discussion of results

Since LHFI(11) and (14) are based on small variations of what is essentially the same model, it is not surprising that they exhibit perfect correlation empirically. (Such was also the case for the Poisson-based LHFI's.) Aside from being perfectly correlated, their posterior distribution of latent health is also virtually identical in every aspect but for a location shift, as is apparent from the index values and 95% HPD intervals in Fig. 3, bottom panel. However, for ranking sites relative to each other's health, this location shift plays no role whatsoever.

Fig. 5 shows a comparison among Poisson-based LHFI's (represented by LHFI(7)), logit-based LHFI's (represented by LHFI(14)), and the existing B-IBI and SHIPSL. We see that B-IBI has a stronger linear association with the logit-based LHFI's than the Poisson-based counterparts. This may be due to the use of all available metrics, and that the metrics share the same scale for B-IBI or the logit models. Interestingly for these data, our logit-based LHFI's are equally correlated with B-IBI and SHIPSL, and SHIPSL is close to being equally correlated with either type of LHFI's.

Similar to the earlier Poisson ANOCOVA models, both logit Models (11) and (14) have identified a significant dependence of health on urbanization; the 95% HPD intervals for γ_1 range approximately from -3.7 to -0.4 (Table 3). In Section 5.1.1 below, we discuss the evidence for preferring Model (14), despite the additional multinomial structure incorporated into (14) appearing to have little effect on the parameter estimates. The reader should also note that with a large enough dataset (which was not the case here), one would ideally retain a non-trivial dependence structure $\mathbf{\Sigma}$ for the β's to account for metric overlap.

5.1.1. Quantitative comparison of (7), (11), and (14)

To compare the performance between the binomial-only Model (11) and the mixture Model (14), we first consider the posterior distribution of σ_H. Respectively, the posterior medias are 0.56 and 0.57, and both 95% HPD intervals are [0.40, 0.79] (Table 3). In this regard, both models appear equally effective. Next, we compute the SSE ratio from cross-validation; the ratio is 1.00. To obtain this, we ran both models on 9 sets of incomplete data generated similarly as in Section 4.1.1; with an additional three abundance metrics and the removal of one richness metric, we had a total

of 486 observations, so that each incomplete subset involved 54 missing observations. The same imputation procedure from before was employed, except that the multinomial part of our mixture model was re-expressed as Poisson log-linear (see, for instance, Dobson (2001)) to allow easier implementation in OpenBUGS. As before, a ratio so close to 1 is inconclusive due to the inherent variability of a small-scale cross-validation exercise. Finally, we consider the DIC (Table 3), which is substantially lower (with a difference of -45) for Model (14). In general, a larger DIC value may not imply an inappropriate model but one that could be improved upon perhaps by additional constraints to parameters (Gelman and Hill, 2007). Thus, in this case, there appears to be a clear advantage in placing the quadrinomial constraint on the EPT metrics.

However, one question remains: which type of GLMM — Poisson for richness metrics only, or logit for richness and abundance metrics — is preferable for monitoring stream health based on the 1997 PSL data? We focus on Models (7) and (14) as representatives of their respective groups. Note that neither the DIC nor out-of-sample predictions can be used to compare models that involve different data. To make a sensible comparison between types, we consider the posterior distribution for H_i's between the two model types. Specifically, consider the HPD intervals from both panels in Fig. 3, where Model (14) shows a larger fluctuation across sites for the posterior location (i.e. the LHFI value), as opposed to the apparent flatness corresponding to LHFI(7). Thus, the logit models demonstrate better distinction of sites than their Poisson counterpart. Moreover, the width of a 95% HPD interval is slightly less for the logit model (average width is 1.72 for (14), but 1.77 for (7)), indicating a higher precision for latent health. Altogether, the inclusion of abundance metrics by the logit models clearly led to additional ability for the LHFI to distinguish among sites. This increased ability may have also resulted from having metrics share a common scale. This common-scale principle has been used to develop all IBI and SHIPSL variants. However, for these earlier types of indices and particularly for the IBI, the scheme used to map indicators to a single scale is more controversial and causes potential loss of information, when compared to the minimal manipulation of metrics before they are incorporated into the logit model for constructing a health index.

5.2. Remarks

The reader may notice that the models thus far considered for the PSL data exhibit model unidentifiability at the level of metric effects. Specifically, their (co)variances cannot be estimated based on $P(\boldsymbol{\beta}|\boldsymbol{\Omega})$ alone. However, unlike the frequentist paradigm in which unidentified parameters are inestimable, Bayesian modelling allows proper estimation of parameter, identified or otherwise, provided that the posterior is proper. In addition, Chiu (2008) shows that for the Poisson-based analyses of the PSL data here, substantial *Bayesian learning* is achieved for the (co)variances despite unidentifiability. As the logit-based models are structurally identical to their Poisson counterparts, the results by Chiu (2008) are expected to extend to Models (11) and (14).

6. Interpreting the LHFI in the absence of prescribed reference conditions

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 reasons for the IBI's popularity. However, one must not overlook the calibration scheme that brings about this apparent advantage. As discussed in Section 1, this reference-based scheme suffers from non-transferability between geographical and temporal domains, and relies heavily on the availability of so-called pristine ecosystems, although they rarely exist. 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. As a compromise for the lack of a full

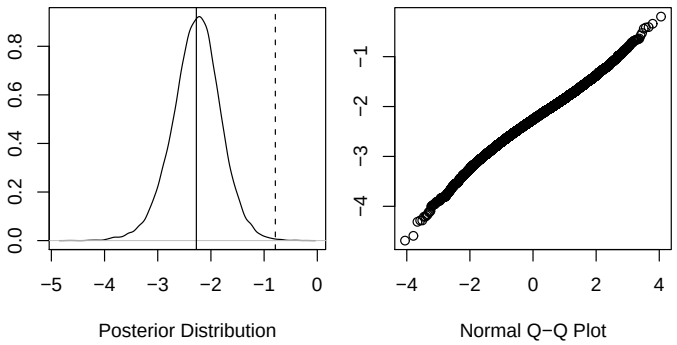

Fig. 7. Posterior distribution of health for TH1 (H_{TH1}), and corresponding normal QQ plot

spectrum of health conditions, it is perhaps more sensible to gauge health against a heavily degraded ecosystem; sadly, it is not difficult to locate these days. When included in the study, a badly degraded site then serves as the baseline for "internal referencing." This concept was originally proposed by Chiu and Guttorp (2006). Much in the same way as a one-way ANOVA assesses the *effectiveness* among several treatments relative to the least effective treatment, an inference-based comparison among all sites can be conducted using the LHFI to assess their *health* relative to the least healthy site; externally defined baseline or reference conditions thus become less relevant.

In particular, ratings can be defined relative to the posterior distribution of health H_{worst} for the site identified (before or after fitting the LHFI model) as the worst degraded in the region. Let us demonstrate this idea in the context of the 1997 PSL study. It has been well documented (e.g. local news, residents' forums) that the health conditions of Thornton Creek is commonly considered "extremely poor," even without the need to record physical measurements. Thus, H_{worst}=H_{TH1} here, and it can act as a baseline value for other sites. To assess Site BB1 situated along Big Bear Creek, a simple approach then is to compute a z-score for its LHFI value (posterior mean of H_{BB1}) relative to the posterior $P(H_{\text{TH1}}|\boldsymbol{Y}, \mathbb{X})$. Assuming Model (14), we have

$$z_{\text{BB1}} = \frac{E(H_{\text{BB1}}|\boldsymbol{Y}, \mathbb{X}) - E(H_{\text{TH1}}|\boldsymbol{Y}, \mathbb{X})}{\sqrt{Var(H_{\text{TH1}}|\boldsymbol{Y}, \mathbb{X})}} = \frac{-0.788 - (-2.272)}{0.461} = 3.22\,.$$

We can visualize this comparison in Fig. 7, left panel: the LHFI(14) value for TH1 is marked by a solid line, and that for BB1 by a dashed line. A subject-matter expert may now translate z=3.22 back to practical terms, and decide on the overall degree of degradation for BB1. Note that the z-score is appropriate here, as $[H_{\text{TH1}}|\boldsymbol{Y}, \mathbb{X}]$ is approximately normally distributed (Fig. 7, right panel).

Occasionally, the study may include healthier sites that are recognized as "nearly pristine." In this case, the above gauge could be replaced or used alongside its "mirror image," i.e. the same procedure but applied to the best site in the study. One could extend this principle further by using posterior quantiles for the best site in the study to define future "pristine" sites. For instance, a new site may be added to the current study, and the LHFI model re-fitted. The earlier best site will now have an updated posterior distribution $P(\tilde{H}_{\text{best}}|\tilde{\boldsymbol{Y}}, \tilde{\mathbb{X}})$ due to the inclusion of the new site, where the tilde '$\sim$' indicates the update; but qualitatively, the site remains "nearly pristine." Now, one may declare that the new site is "pristine" if its LHFI value falls above, say, the 90th percentile of $P(\tilde{H}_{\text{best}}|\tilde{\boldsymbol{Y}}, \tilde{\mathbb{X}})$. Similarly, the site could be labelled as "exceedingly degraded" if its index value falls below, say, the 10th percentile of $P(\tilde{H}_{\text{worst}}|\tilde{\boldsymbol{Y}}, \tilde{\mathbb{X}})$. Note that this approach is not restricted to new sites taken from the same spatial or temporal domain as the others, so long as the model from

Section 2.2 is applicable (i.e. extra domain-specific effect terms are estimable).

Gauging ecosystem health with this "internal referencing" scheme can reduce ambiguity in the absolute definition of health that is inherent in popular reference-based methods. Of course, a minor level of ambiguity is inevitable, such as in the percentile cut-offs used to define categories of health, which should be left to subject-matter experts to decide. However, our proposed method reduces the amount of ambiguity involved in as many stages of health index construction as it is feasible.

7. Conclusion

The methodology for constructing an LHFI demonstrated in this article is rooted in a simple statistical concept of ANOCOVA model building, and may be easily adapted to any context of health assessment, be it ecological, medical, or otherwise. 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 modelling principles, an effective comparison of health among sites can be achieved. Simply through fitting the model, one can also readily and simultaneously determine (1) the statistical properties of the health assessment for current and/or future sites, as well as (2) the significance of the impact on health for the observable factors under consideration. Although latent variable modelling techniques have become widely popular in many sciences, its use to produce a direct quantitative "report card" composite measure of overall health is apparently uncommon. Therefore, our methodology is a simple but universal and versatile scientific approach that is potentially far reaching to any research discipline in which a scalar assessment of health is desirable.

In the ecological context, LHFI modelling attempts to retain the user-friendliness of conventional scalar health indices, but overcome several hurdles not clearly addressed by conventional index building approaches. Specifically, to address the age-old difficulty encountered in inter-regional and -temporal studies, we proposed the addition of a *domain effect* term in the latent factor model, which is a standard practice in many scientific contexts for comparison among strata. This also avoids the complexity and impracticality of formal spatio-temporal models in biomonitoring studies. Building an LHFI for ecosystems involves virtually no qualitative procedures and deals directly with the raw metrics and associated covariates; hence, it can easily incorporate auxiliary information into the index and, unlike some others, can retain all the information directly available from the metrics. Through the LHFI model, proper and tractable inference of current and predicted site health is also practical and unambiguous. This is certainly not the case for common ecological indices. When pristine conditions are unavailable or inaccessible, the construction and interpretation of reference-based health indices may become arbitrary. To address this, we proposed "internal referencing" against badly degraded sites that can be easily included in a study. Scientific comparisons via statistical modelling is universal, and constructing a health index as such is intended to achieve the same purpose as reference-based techniques, but with few of the associated disadvantages.

In fact, the *statistical* principles used to construct the LHFI by no means diminish the *biological* worthiness of the resulting index, as subject-matter expertise remains vital in variable selection and results interpretation before and after model fitting. To use the terminology of Fjelland (2002) page 168, 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.

For the 1997 PSL data, we explored two types of LHFI 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 Type (B) models perform better in their ability to distinguish sites according to health. Generally, either type of LHFI contains biological information that is highly comparable to that in the existing B-IBI and SHIPSL, but it carries the extra advantages as discussed above. Of course, the same methodology could lead to very different results and conclusions when applied to another dataset. For example, a latent health factor model ideally would account for informational overlap carried among metrics. However, for the 1997 PSL data, the covariance among metric effects was poorly estimated, likely due to data sparsity. To reduce model complexity, we assumed independent metric effects with unequal variances, and 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. Another example is, despite statistical significance of the extra level of model hierarchy (i.e. the latent regression) for the Poisson model, it appears to have little impact on predictive ability. One can imagine that given another dataset with more relevant covariates, predictive ability will be improved. Nevertheless, it is always advisable to keep in the model any statistically significant covariate that subject-matter experts have previously identified as potentially influential to health. Such a model incorporates expert knowledge in an unambiguous fashion, and it certainly provides a more comprehensive picture of the relationship among metrics, factors, and latent health.

Irrespective of the dataset, the latent factor modelling methodology itself is systematic and unambiguous for any study from a suitable health assessment context. The associated modelling principles give our approach the versatility and adaptability to studies that involve multiple data types observed on different macroscopic scales.

Acknowledgements

We thank Dr. John van Sickle, Environmental Protection Agency Western Ecology Division, for sharing his ideas and providing some of his publications and data as reference; the editor, associate editor, and referees on an earlier version of this article for their invaluable comments that led to further developments of our work. The corresponding author thanks Dr. Melissa Dobbie, Mathematical and Information Sciences Division, Commonwealth Scientific and Industrial Research Organisation, for sharing her experience in ecological health indices; and Professor Michael Dowd, Department of Mathematics and Statistics, Dalhousie University, for stimulating discussions on MCMC techniques and post-normal science.

Appendix A: Details of MCMC posterior sampling

To minimize MCMC mixing problems, we employed partial hierarchical centring to reformulate parts of each model before implementation (see Appendix in Chiu *et al.*, 2007, for the full rationale). For example, the relevant parts of Model (5) become

$$[b_j|\alpha,\sigma_j] \overset{\text{ind}}{\sim} N(\alpha,\sigma_j^2)\,, \qquad [\nu_{i1}|b_1,\sigma_H] \overset{\text{iid}}{\sim} N\left(b_1,\sigma_H^2\right)\,,$$
$$\nu_{ij} = \nu_{i1} - (b_1 - b_j)\;\forall\; j > 1\,, \qquad H_i = \alpha + \nu_{i1} - b_1\,, \qquad \beta_j = b_j - \alpha\,.$$

The same principle is applied to Models (7) and (10), except that

$$[\nu_{i1}|b_1,\gamma_1,\boldsymbol{x},\sigma_H] \sim N\left(b_1 + \gamma_1(x_i - \overline{x}),\sigma_H^2\right)$$

to account for the latent regression. For implementation of Models (11) and (14), we explored several formulations of partial hierarchical centring, one of which performed satisfactorily:

$$[\tilde{H}_i|\gamma_1, \boldsymbol{x}, \sigma_H] \sim N(\gamma_1(x_i - \overline{x}), \sigma_H^2)\,, \quad \left[b_{j(s)}|\alpha, \sigma_{j(s)}\right] \overset{\text{ind}}{\sim} N(\alpha, \sigma_{j(s)}^2)\,,$$
$$\nu_{isj\bullet} = \tilde{H}_i + b_{j(s)}\,, \quad H_i = \alpha + \tilde{H}_i\,, \qquad \beta_{j(s)} = b_{j(s)} - \alpha\,.$$

For each of the models fitted to the 1997 PSL data, the posterior samples used for inference were generated from two Markov chains. Each chain consisted of M draws, reduced from removing a burn-in of b draws then thinning by a lag of ℓ. The chains started at two different randomly generated initial values.

For Model (5), M=10,000, b=30,000 and ℓ=2, and for Model (7), M=10,000, b=15,000 and ℓ=1 (i.e. no subsequent thinning). Minor mixing problems for α in Model (7) appeared in the form of a slight difference in location (mean / median / mode) between chains. Nevertheless, all other features of the chains for α were highly comparable. Although the distribution of each H_i depends on α, no mixing problems were observed for H_i's. Nor were mixing problems present for any of the parameters in Model (5). Thus, in either case, the two chains were combined to form one posterior sample of size $2M$ to define the corresponding LHFI measures without ambiguity.

For Model (10), M=5,000, b=7,000 and ℓ=20. The posterior samples for $\boldsymbol{\Sigma}$ were somewhat volatile, in that both chains showed extreme skewness for the $\boldsymbol{\Sigma}$ entries with tail values on the order of 10^5, and the chain dispersions were noticeably different; however, marginal posterior medians were comparable between chains. Nor did the Brooks-Gelman-Rubin convergence diagnostic plots (Brooks and Gelman, 1998) exhibit patterns that would cause a great deal of concern. Indeed, it is understandable that extensive coverage of the support of an exceedingly diffuse posterior may require an impractical number of simulated samples. Chiu (2008) also observes that this diffuseness of the posterior is a direct result of the extreme diffuseness of the prior, combined with the limited amount of data for estimating many parameters. In other words, our mixing problems could well be an artefact of this phenomenon. In light of how well all non-$\boldsymbol{\Sigma}$ parameters were estimated, how similar the $\boldsymbol{\Sigma}$ entries' medians were between chains, and how well all parameters were estimated for the smaller models (5) and (7), it appears that the mixing problems did not arise from an intrinsically incorrect model. Thus, for the purpose of estimating $\boldsymbol{\Sigma}$, we restrict our attention to the chain that yielded a larger posterior variability in $\boldsymbol{\Sigma}$. The smaller variability for the rejected chain could have resulted from an initial value that confined the Markov chain to a smaller subset of the parameter space. However, no mixing problems were encountered for H_i's between chains. Hence, we define LHFI(10) to be the mean of the posterior H_i samples based on the two chains combined.

We now come to the logistic Models (11) and (14), for each of which M=10,000, b=5,000, and ℓ=5. In both cases, mixing problems were encountered only for the non-EPT $\sigma_{j(s)}$'s. The problems resemble those for the $\boldsymbol{\Sigma}$ entries from Model (10), with tail values on the order of 10^3 here. Again, they may be explained by posterior diffuseness and limited data. Finally, as the H_i chains mixed exceptionally well for either model, combining them to form the LHFI was justified.

Appendix B: The consequences of incorporating a latent regression and the dependence among metric effects

In this appendix, we show that having (a) health regressed on covariates and (b) correlated metric effects in the latent health factor model can address the natural correlation among metric values over sites and over metrics.

First, consider the potential shortcomings of Models (5) and (7) which assume independence among metric effects. We do so through a Gaussian analogue of Model (7):

$$\begin{aligned} W_{ijk} \equiv \ln Y_{ijk} = H_i + \beta_j + \varepsilon_{ijk}\,, \qquad & [\varepsilon_{ijk}|\sigma_\varepsilon] \overset{\text{iid}}{\sim} N(0,\sigma_\varepsilon^2)\,, \\ [H_i|\alpha,\gamma_1,\boldsymbol{x},\sigma_H] \overset{\text{iid}}{\sim} & N\left(\alpha+\gamma_1(x_i-\overline{x}),\sigma_H^2\right)\,, \qquad (15) \\ [\beta_j|\sigma_j] \overset{\text{iid}}{\sim} & N(0,\sigma_j^2)\,. \qquad (16) \end{aligned}$$

Conditioned on $\boldsymbol{\Omega} = (\alpha, \gamma_1, \sigma_H, \sigma_1, \ldots, \sigma_7)$, the mean and covariance structures of the data coming from sites (i, i') and from metrics (j, j') are

$$E(W_{i'jk}|\boldsymbol{\Omega}) = E(H_{i'} + \beta_j + \varepsilon_{i'jk}|\boldsymbol{\Omega}) = \alpha + \gamma_1(x_{i'} - \overline{x})\,, \qquad (17)$$

$$E(W_{ijk}|\boldsymbol{\Omega}) = E(H_i + \beta_j + \varepsilon_{ijk}|\boldsymbol{\Omega}) = \alpha + \gamma_1(x_i - \overline{x}) = E(W_{ij'k}|\boldsymbol{\Omega})\,, \qquad (18)$$

$$Cov(W_{ijk}, W_{i'jk}|\boldsymbol{\Omega}) = Cov(H_i + \beta_j + \varepsilon_{ijk},\ H_{i'} + \beta_j + \varepsilon_{i'jk}|\boldsymbol{\Omega}) = Var(\beta_j|\sigma_j) = \sigma_j^2\,, \qquad (19)$$

$$Cov(W_{ijk}, W_{ij'k}|\boldsymbol{\Omega}) = Cov(H_i + \beta_j + \varepsilon_{ijk},\ H_i + \beta_{j'} + \varepsilon_{ij'k}|\boldsymbol{\Omega}) = Var(H_i|\sigma_H) = \sigma_H^2\,. \qquad (20)$$

Now, take the priors from (6) and (8). Then, by the law of total covariance, one can easily show that the marginal covariances become

$$\begin{aligned} Cov(W_{ijk}, W_{ij'k}) &= \psi + c_2[1 + (x_i - \overline{x})^2]\,, \qquad (21) \\ Cov(W_{ijk}, W_{i'jk}) &= \psi + c_2[1 + (x_i - \overline{x})(x_{i'} - \overline{x})] \qquad (22) \end{aligned}$$

where ψ depends on c_3 and c_4 only. Thus, given site i, (21) implies that the correlation between (the log-values of) any pair of metrics is constant over metrics (i.e. independent of (j, j')). However, as discussed in Section 4, metric values could be naturally correlated over metrics and over sites. Conveniently, dependency over sites is addressed by regressing latent health on site-specific covariates according to (22): given metric j, the correlation of metric values between any pair of sites depends on (i, i'). However, this dependence would have been lost should the latent regression be removed from (15), leaving (17), (18), (21), and (22) simply as

$$\begin{aligned} E(W_{ijk}|\boldsymbol{\Omega}) = E(W_{i'jk}|\boldsymbol{\Omega}) = E(W_{ij'k}|\boldsymbol{\Omega}) = \alpha\,, \\ Cov(W_{ijk}, W_{i'jk}) = Cov(W_{ijk}, W_{ij'k}) = \psi + c_2\,. \end{aligned}$$

Just as the latent regression introduces correlation over sites, dependence among metric effects β_j's conveniently incorporates correlation over metrics into the model, by replacing (16) with $\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$. Adding this to the latent regression turns (21) and (22) into

$$\begin{aligned} Cov(W_{ijk}, W_{i'jk}) &= c_2\left[1 + (x_i - \overline{x})(x_{i'} - \overline{x})\right] + E(\sigma_j^2)\,, \\ Cov(W_{ijk}, W_{ij'k}) &= c_2\left[1 + (x_i - \overline{x})^2\right] + \psi + E(\sigma_{jj'})\,. \end{aligned}$$

The hyperparameter $\mathbb{S}$ in the inverse-Wishart prior (9) can be specified such that $E(\sigma_j^2)$ and $E(\sigma_{jj'})$ — and hence, the covariances — depend on j and (j, j'), respectively. For the PSL data, we tried various such priors, but all of them yielded somewhat ambiguous estimates due to mixing problems as described in Appendix A. Finally, to reduce extra model complexity, we settled for an exchangeable structure for $\mathbb{S}$ as described in Section 4. As it turns out, the posterior distribution of $\boldsymbol{\Sigma}$ indeed provides some evidence, albeit weak, that β_j's are correlated.

References

Billheimer, D., Cardoso, T., Freeman, E., Guttorp, P., Ko, H.-W. and Silkey, M. (1997) Natural variability of benthic species composition in the Delaware Bay. *Environmental and Ecological Statistics*, **4**, 95–115.

Brinck, K. W. (2002) Comparing methods for inferring site biological condition from a sample of site biota. *MS Thesis*. University of Washington, Seattle.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.

Bunea F., Guttorp, P. and Richardson, T. (1999) Ecological indices and graphical modeling of factors influencing benthic populations in streams. *NRCSE Technical Report 036*. University of Washington, Seattle.

Chiu, G. S. (2008) On identifiability of covariance components in hierarchical generalized analysis of covariance models. *Technical Report 2008-09*. University of Waterloo, Waterloo.

Chiu, G. and Guttorp, P. (2004). New developments involving the stream health index for the Puget Sound Lowland. *NRCSE Technical Report 079*. University of Washington, Seattle.

Chiu, G. and Guttorp, P. (2006) Stream health index for the Puget Sound Lowland. *Environmetrics*, **17**, 285–307.

Chiu, G. S., Guttorp, P., Liang, J., Khan, S. A. and Westveld, A. H. (2007) An ecological latent health factor index via a random-effects model for taxa richness and composition. *Technical Report 2006-02*. University of Waterloo, Waterloo.

Clarke, R. T., Wright, J. F., Furse, M. T. (2003) RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling*, **160**, 219–233.

Dobson, A. (2001) *An Introduction to Generalized Linear Models*, 2nd edn. London: Chapman and Hall / CRC.

Fjelland, R. (2002) Facing the problem of uncertainty. *Journal of Agricultural and Environmental Ethics*, **15**, 155–169.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parametrisation for normal linear mixed models. *Biometrika*, **82**, 479–488.

Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel / Hierarchical Models*. New York: Cambridge University Press.

Hawkins, P. H., Norris, R. H., Hogue, J. N. and Feminella, J. W. (2000) Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications*, **10**, 1456–1477.

Hays, R. D., Morales, L. S., Reise, S. P. (2000) Item response theory and health outcomes measurement in the 21st Century. *Medical Care*, **38**, II-28–II-42.

Jørgensen, S.E., Xu, F.-L. and Costanza, R. (2005) *Handbook of Ecological Indicators for Assessment of Ecosystem Health*. New York: CRC Press.

Kerans, B. L. and Karr, J. R. (1994) A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. *Ecological Applications*, **4**, 768–785.

López-Alvarenga J. C., Montesinos-Cabrera, R. A., Velázquez-Alva, C. and González-Barranco, J. (2003) Short stature is related to high body fat composition despite body mass index in a Mexican population. *Archives of Medical Research*, **34**, 137–140.

McCormick, F. H., Hughes, R. M., Kaufmann, P. R., Peck, D. V. and Stoddard, J. L. (2001) Development of an index of biotic integrity for the Mid-Atlantic Highlands region. *Transactions of the American Fisheries Society*, **130**, 857–877.

McCulloch, C. E. and Searle, S. R. (2001) *Generalized, Linear, and Mixed Models*. New York: Wiley and Sons.

Morley, S. A. (2000) Effects of urbanization on the biological integrity of Puget Sound Lowland streams: restoration with a biological focus. *MS Thesis*. University of Washington, Seattle.

Pietrobon, R., Taylor, M., Guller, U., Higgins, L. D., Jacobs, D. O. and Carey, T. (2004) Predicting gender differences as latent variables: summed scores, and individual item responses: a methods case study. *Health and Quality of Life Outcomes*, **2**: 59.

Rosas, G. (2008) Dynamic latent trait models: an application to Latin American banking crises. *Department of Politcal Science Working Paper*. Washington University in St. Louis.

Smith, B. J. (2007) An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, **21**.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Statistical Methodology*, **64**, 583–639.

Steedman, R. J. and Regier, H. A. (1990) Ecological bases for an understanding of ecosystem integrity in the Great Lakes Basin. In *An Ecosystem Approach to the Integrity of the Great Lakes in Turbulent Times* (eds C. J. Edwards and H. A. Regier), Special Publication No. 90-4, pp. 257–270. Ann Arbor: Great Lakes Fishery Commission.

Stock, J. H. and Watson, M. W. (1989) New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual*, **4**, 351–394.

Stoddard, J. L., Peck, D. V., Olsen, A. R., Larsen, D. P., van Sickle, J., Hawkins, C. P., Hughes, R. M., Whittier, T. R., Lomnicky, G., Herlihy, A. T., Kaufmann, P. R., Peterson, S. A., Ringold, P. L., Paulsen, S. G. and Blair, R. (2005) Western streams and rivers statistical summary. EPA 620/R-05/006. Washington: U.S. Environmental Protection Agency, Office of Research and Development.

Ter Braak, C. J. F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

Ter Braak, C. J. F. (1987) The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, **69**, 69–77.

Thomas, A., O'Hara, B., Ligges, U. and Sturtz, S. (2006) Making BUGS open. In *R News* (eds M. Plummer and P. Murrell), Vol. 6/1, pp. 12–17. Vienna: R Foundation for Statistical Computing.

Ward, M. D. and Hoff, P. D. (2007) Persistent patterns of international commerce. *Journal of Peace Research*, **44**, 157–175.

Westfall, P. H., Johnson, W. O. and Utts, J. M. (1997) A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, **84**, 419–427.

Westveld, A. H. (2007) Statistical methodology for longitudinal social network data. *PhD Dissertation*. University of Washington, Seattle.