Received: 27 September 2012, Revised: 14 June 2013, Accepted: 14 June 2013, Published online in Wiley Online Library: 19 July 2013

(wileyonlinelibrary.com) DOI: 10.1002/env.2221

Bayesian stable isotope mixing models

Andrew C. Parnell[a]*, Donald L. Phillips[b], Stuart Bearhop[c], Brice X. Semmens[d], Eric J. Ward[e], Jonathan W. Moore[f], Andrew L. Jackson[g], Jonathan Grey[h], David J. Kelly[g] and Richard Inger[i]

In this paper, we review recent advances in stable isotope mixing models (SIMMs) and place them into an overarching Bayesian statistical framework, which allows for several useful extensions. SIMMs are used to quantify the proportional contributions of various sources to a mixture. The most widely used application is quantifying the diet of organisms based on the food sources they have been observed to consume. At the centre of the multivariate statistical model we propose is a compositional mixture of the food sources corrected for various metabolic factors. The compositional component of our model is based on the isometric log-ratio transform. Through this transform, we can apply a range of time series and non-parametric smoothing relationships. We illustrate our models with three case studies based on real animal dietary behaviour. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: stable isotope analysis; mixing models; Bayesian hierarchical model; compositional data; time series

1. INTRODUCTION

Stable isotope analysis is an increasingly important tool in the study of ecological food webs. The technique utilises the fact that elements that exist in food sources are transferred to a consumer when eaten. Analysis of the tissues of the consumer will provide information as to the proportional contribution of the different food sources consumed. In practice, it is the part of the elements termed ‘stable isotopes’ (i.e. not radioactive) that are measured using a mass spectrometer and expressed as the ratio of heavy to light form. These measurements can be taken for both the food sources and the consumers.

As a consumer’s tissues are ultimately derived from the dietary sources they consume, it is possible to use stable isotope mixing models (SIMMs) to estimate the assimilated diet of an individual, or a group of individuals, given the isotopic ratios of the consumers’ tissues and food sources (Phillips, 2012). A number of recent papers have proposed models to analyse such data, gathering over 1500 citations since their first introduction. More recently, the models proposed for such data are Bayesian. In this paper, we review the different models proposed and bring them into an overarching framework. We include three case studies ranging from the simple to the complex, together with Just Another Gibbs Sampler (JAGS; Plummer, 2003) code for their implementation[†].

The stable isotope measurements are standardised against international reference samples and reported in the delta (δ) notation as parts per thousand or per mil (‰). Generally the isotopic ratios of a sample of a consumer’s tissues (e.g. blood, feathers and whiskers) are measured along with a representative sample of potential items from a consumer’s diet. The consumer isotopic values are represented in the model as the convex combination of the source values where the coefficients in the simplex are the ‘dietary proportions’; strictly speaking they are the proportion of the consumers’ dietary proteins obtained from the sources. Estimation of these dietary proportions is the main focus of our analysis. Most commonly the isotopic observations on the consumers and sources are multivariate with dimension 2. A thorough description of the uses of stable isotopes can be found in the studies of Inger and Bearhop (2008).

* *Correspondence to: Andrew C. Parnell, Room 500, Library Building, University College Dublin, Belfield, Dublin, Ireland. E-mail: Andrew.Parnell@ucd.ie*

a *School of Mathematical Sciences (Statistics), Complex and Adaptive Systems Laboratory, University College Dublin, Belfield, Dublin, Ireland*

b *Western Ecology Division, National Health & Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Corvallis, OR, U.S.A.*

c *Centre for Ecology and Conservation, School of Biosciences, University of Exeter, Exeter, U.K.*

d *Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, U.S.A.*

e *Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA, U.S.A.*

f *Earth2Ocean Research Group, Simon Fraser University, Burnaby, BC, Canada*

g *School of Natural Sciences, Trinity College Dublin, Dublin, Ireland*

h *School of Biological & Chemical Sciences, Queen Mary, University of London, London, U.K.*

i *Environment and Sustainability Institute, School of Biosciences, University of Exeter,, Exeter, U.K.*

[†] See mathsci.ucd.ie/~parnell_a/.

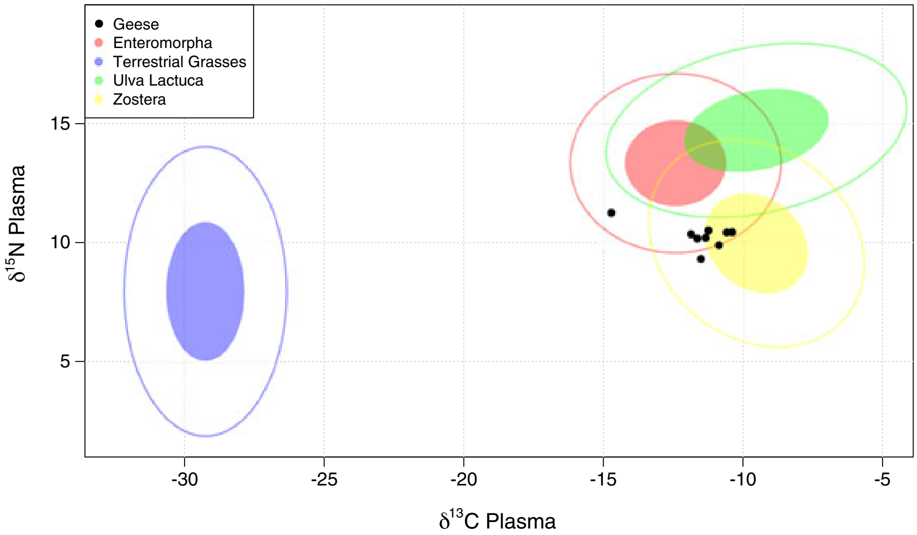

Figure 1. Iso-space plot of the geese data of case study 1 (Section 5.1). The consumer information is shown in the black-filled circles, whereas the sources are shown as contours (90% range) and filled ellipses (50% range). The consumers seem to lie close to the *Zostera* source so this is likely to form a substantial part of their diet

Once the isotopic data have been collected for both consumer and sources, it is usual to create an *iso-space* plot, which shows the consumer and source values. An example is shown in Figure 1. It is desirable for the consumer values to lie within the fuzzy convex hull of the sources. However, a further phenomenon is often observed here, that of *trophic enrichment*, whereby light isotopes are lost during the conversion of source proteins into consumer tissues. The isotopic values of the consumer (or equivalently the sources) are thus corrected by a *trophic enrichment factor* (TEF), which may vary by food source and consumer. These TEF corrections arise from laboratory studies and thus contribute another set of (uncertain) data to our analysis.

The inference challenge involved in a SIMM is to estimate the dietary proportions whilst taking account of the uncertainty in the source and TEF values. Clearly not all of the consumers will eat exactly the same diet, so it is common to use a hierarchical model. Furthermore, covariates such as time or age may be available, which are thought to influence the dietary proportions. The model we present in this paper takes account of these common features in a multivariate hierarchical Bayesian model.

The paper is organised as follows. In Section 2, we introduce our notation and outline our general SIMM. In Section 3, we discuss previous work in this area and show how each of these fits into our general SIMM. Section 4 outlines the statistical issues concerning the formulation of such models and how they can be fitted. We outline three case studies in Section 5, showing how the model works in different situations depending on the information available. We discuss future directions in Section 6. A technical appendix is included to outline some of our modelling assumptions in more detail.

2. MODEL FORMULATION

We first provide the notation we use to formulate our model. We suppose that there are consumer measurements taken from N consumers on J isotope values. These consumers are assumed to be eating a combination of K known sources only. There are further measurements on the sources themselves, totalling N_k^s measurements on source k, and N_k^c measurements on TEFs for source k. We outline the most important components of our model as follows:

1. Y_{ij} represents the consumer isotope measurement on observation i $(i = 1, \ldots, N)$ for isotope value j $(j = 1, \ldots, J)$. We write $\boldsymbol{Y}_i$ as the J-vector of isotope values for consumer i and $\boldsymbol{Y}$ for the full set of consumer data.
2. Y_{ijk}^s represents the source isotope measurement for observation i $\left(i = 1, \ldots, N_k^s\right)$, isotope value j $(j = 1, \ldots, J)$ and source k $(k = 1, \ldots, K)$. We write $\boldsymbol{Y}_{ik}^s$ as the J-vector of source measurements for observation i on source k and $\boldsymbol{Y}^s$ for the full set of source data.
3. Y_{ijk}^c represents the TEF isotope measurement (we use superscript c as the TEF represents a correction term) for observation i $(i = 1, \ldots, N^c)$, isotope value j $(j = 1, \ldots, J)$ and source k $(k = 1, \ldots, K)$. We write $\boldsymbol{Y}_{ik}^c$ as the J-vector of TEF measurements for observation i on source k and $\boldsymbol{Y}^c$ for the full set of TEF data.
4. s_{ijk} is the source random effect for consumer i on isotope value j and source k. We write $\boldsymbol{s}_{ik}$ to be the J-vector of isotope source values for consumer i on source k and $\boldsymbol{s}_i$ to be the $J \times K$ matrix of source values for consumer i.
5. c_{ijk} is the TEF random effect for consumer i on isotope value j and source k. We write $\boldsymbol{c}_{ik}$ to be the J-vector of TEF values for consumer i on source k and $\boldsymbol{c}_i$ as the $J \times K$ matrix of TEF values related to consumer i.

6. p_{ik} is the dietary contribution of source k for consumer i. $\boldsymbol{p}_i$ is the K-vector of dietary proportions for consumer i. Estimation of these dietary proportions is the main focus of our analysis.
7. ϵ_{ijk} is a random noise term representing residual variation. We write $\boldsymbol{\epsilon_i}$ as the J-vector of residual terms for consumer i and set $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ a covariance matrix, itself given an inverse-Wishart prior distribution.

We make a conditional independence assumption between the consumer, source and TEF data sets. Using the notation in the previous discussion, we can write out the likelihood for the data as

$$\begin{aligned}
\boldsymbol{Y}_i &\sim N\left(\boldsymbol{p}_i^T(\boldsymbol{s}_i + \boldsymbol{c}_i), \boldsymbol{\Sigma}\right),\ i = 1, \ldots, N, && (1)\\
\boldsymbol{Y}^s_{ik} &\sim N\left(\boldsymbol{\mu}^s_k, \boldsymbol{\Sigma}^s_k\right),\ i = 1, \ldots, N^s_k,\ k = 1, \ldots, K, && (2)\\
\boldsymbol{Y}^c_{ik} &\sim N\left(\boldsymbol{\mu}^c_k, \boldsymbol{\Sigma}^c_k\right),\ i = 1, \ldots, N^c_k,\ k = 1, \ldots, K. && (3)
\end{aligned}$$

We assume a hierarchical formulation so that $\boldsymbol{s}_{ik} \sim N\left(\boldsymbol{\mu}^s_k, \boldsymbol{\Sigma}^s_k\right)$ and $\boldsymbol{c}_{ik} \sim N\left(\boldsymbol{\mu}^c_k, \boldsymbol{\Sigma}^c_k\right)$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (with super- and sub-scripts added accordingly) represent the means and covariances of the source and TEF data sets, respectively.

Of particular interest is the modelling structure for the dietary proportions $\boldsymbol{p}$. We use an isometric log-ratio (ilr) approach as proposed by Egozcue *et al.* (2003), although other transformations are available (see next two sections for further discussion). The transformation is written as

$$\boldsymbol{\phi}_i = \text{ilr}(\boldsymbol{p_i}) = \boldsymbol{V}^T \log\left[\frac{p_{i1}}{g(\boldsymbol{p}_i)}, \ldots, \frac{p_{iK}}{g(\boldsymbol{p}_i)}\right] \text{ with } g(\boldsymbol{p}_i) = \left(\prod_{i=1}^{K} p_{ik}\right)^{1/K} \tag{4}$$

with $\boldsymbol{V}$ a $K - 1 \times K$ matrix of orthonormal basis functions on the simplex. The inverse transformation $\boldsymbol{p}_i = \text{ilr}^{-1}(\boldsymbol{\phi}_i)$ simply involves exponentiating and re-normalising the values. There are two consequences of working with the ilr. The first is that we now work in a $K - 1$ dimensional space. The second is that there is no obvious link between the elements of ϕ_{ik} and p_{ik}, so we lose some degree of interpretability. We further parameterise the transformed proportions so that $\phi_{ik} \sim N(\gamma_{ik}, \kappa_k)$ with κ_k quantifying a random effect variance and given a vague inverse gamma prior. An alternative to the ilr would be to use the centred log-ratio (clr) transform (equivalent to setting $\boldsymbol{V} = \boldsymbol{I}$) so that κ_k now represents the consumer-level variance in diet for source k. In Section 4, we explore why such a choice might not be a good idea. Finally, we set γ_{ik} to be a mean term through which we allow covariates to impact the dietary proportion behaviour. A multivariate prior for $\boldsymbol{\phi}_i$ would also be feasible, although we do not explore such a possibility here.

In situations where covariates $\boldsymbol{x}_i$ are available, they are usually linked to the model through the dietary proportions. The covariates may take the form of age, sex, time or any other variables upon which diet is expected to depend. In certain cases (such as our third case study) both the diet and the sources are expected to be functions of a covariate. In the simpler case, we apply the covariates by making γ_{ik} functions of $\boldsymbol{x}_i$. In the more advanced case, we additionally apply them to $\boldsymbol{\mu}^s_k$ and $\boldsymbol{\Sigma}^s_k$.

The posterior distribution can be written out in full as follows:

$$\begin{aligned}
\pi\left(\boldsymbol{p}, \boldsymbol{\phi}, \boldsymbol{\kappa}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}^s, \boldsymbol{\Sigma^c}, \boldsymbol{\mu}^s, \boldsymbol{\mu}^c, \boldsymbol{s}, \boldsymbol{c} | \boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{Y}^s, \boldsymbol{Y}^c\right) \propto & \left[\prod_{i=1}^{N} \pi(\boldsymbol{Y}_i | \boldsymbol{p}_i, \boldsymbol{s}_i, \boldsymbol{c}_i, \boldsymbol{\Sigma})\right]\\
& \times \left[\prod_{k=1}^{K} \prod_{i=1}^{N^s_k} \pi\left(\boldsymbol{Y}^s_{ik} | \boldsymbol{\mu}^s_k, \boldsymbol{\Sigma}^s_k\right)\right]\\
& \times \left[\prod_{k=1}^{K} \prod_{i=1}^{N^c_k} \pi\left(\boldsymbol{Y}^c_{ik} | \boldsymbol{\mu}^c_k, \boldsymbol{\Sigma}^c_k\right)\right]\\
& \times \left[\prod_{i=1}^{N} \pi(\boldsymbol{\phi}_i | \boldsymbol{x}_i, \boldsymbol{\kappa})\right]\\
& \times \left[\prod_{i=1}^{N} \prod_{k=1}^{K} \pi\left(\boldsymbol{s}_{ik} | \boldsymbol{\mu}^s_k, \boldsymbol{\Sigma}^s_k\right)\right]\\
& \times \left[\prod_{i=1}^{N} \prod_{k=1}^{K} \pi\left(\boldsymbol{c}_{ik} | \boldsymbol{\mu}^c_k, \boldsymbol{\Sigma}^c_k\right)\right]\\
& \times \left[\prod_{k=1}^{K} \pi\left(\boldsymbol{\mu}^s_k, \boldsymbol{\Sigma}^s_k\right)\right] \times \left[\prod_{k=1}^{K} \pi\left(\boldsymbol{\mu}^c_k, \boldsymbol{\Sigma}^c_k\right)\right]\\
& \times \left[\prod_{k=1}^{K} \pi(\kappa_k)\right] \times \pi(\boldsymbol{\Sigma})
\end{aligned} \tag{5}$$

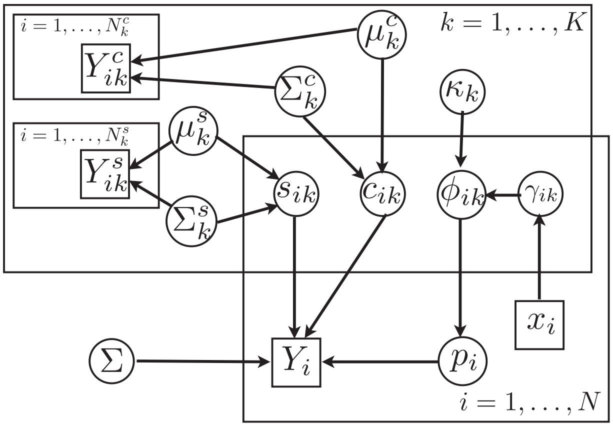

Figure 2. A directed acyclic graph of our model. Circles indicate parameters to be estimated, whereas squares indicate data. The arrows indicate the direction of information flow

Note that we have shortened notation in the posterior on the left-hand side here so that $\boldsymbol{\Sigma}^s = \left\{\boldsymbol{\Sigma}_1^s, \dots, \boldsymbol{\Sigma}_K^s\right\}$, $\boldsymbol{\Sigma}^c = \left\{\boldsymbol{\Sigma}_1^c, \dots, \boldsymbol{\Sigma}_K^c\right\}$, $\boldsymbol{\mu}^s = \left\{\boldsymbol{\mu}_1^s, \dots, \boldsymbol{\mu}_K^s\right\}$, and $\boldsymbol{\mu}^c = \left\{\boldsymbol{\mu}_1^c, \dots, \boldsymbol{\mu}_K^c\right\}$. A directed acyclic graph for the model is shown in Figure 2.

In many SIMM situations, the source and TEF data sets ($\boldsymbol{Y}^s$ and $\boldsymbol{Y}^c$) are not fully given, possibly having been previously published by other authors who have neglected to include the full data and instead only quoted means, variances and, occasionally, covariances. In all three of our case studies, we do not have access to the TEF data set $\boldsymbol{Y}^c$. In such cases, the values given by the previously published sources $\left(\text{i.e. } \boldsymbol{\Sigma}_k^c \text{ and } \boldsymbol{\mu}_k^c\right)$ are treated as fixed. By contrast, we often do have access to the full-source data sets (as often this is collected concomitantly with that of the consumers), and so these can be used in a more complete Bayesian model. However, as is clear from the directed acyclic graph and the model presented in Equation 1, the information on $\boldsymbol{\Sigma}_k^s$, $\boldsymbol{\mu}_k^s$, $\boldsymbol{\Sigma}_k^c$ and $\boldsymbol{\mu}_k^c$ is ancestral (in the sense of being a grandfather node) to the consumer data $\boldsymbol{Y}$; indeed were the source and TEF data sets not used at all, both s_{ik} and c_{ik} would be unidentifiable. Furthermore, when we have $\sum_k N_k^s \gg N$ and $\sum_k N_k^c \gg N$, the consumer data $\boldsymbol{Y}$ contributes little if anything to the posterior distributions of source, and TEF means and covariances. Given the form of the model, we are happy to treat the source and TEF data sets as a separate source of information and use the posterior means $\bar{\boldsymbol{\Sigma}}_k^s, \bar{\boldsymbol{\mu}}_k^s | \boldsymbol{Y}^s$ and $\bar{\boldsymbol{\Sigma}}_k^c, \bar{\boldsymbol{\mu}}_k^c | \boldsymbol{Y}^c$ as fixed in place of a more complete yet computationally slower model. We explore the effect of fixing these ancestral nodes from the external data in Appendix A.

3. PREVIOUS WORK

The earliest attempts at applying a mixing model framework to stable isotope data did not use probability as the basis for estimation. Initially, SIMMs were restricted to systems involving a single consumer (or the mean of multiple consumers), and where the number of isotope values and sources was arranged such that $J+1=K$. Such an arrangement yields a linear system with a single solution. Phillips and Gregg (2001) provided propagation of error calculations for such a system in their IsoError model to establish confidence intervals around the estimates based on the variances of the consumer and source isotopic measurements. The method was expanded upon in IsoSource (Phillips and Gregg, 2003) to relax the $J+1=K$ restriction and allow for multiple sources but without explicit incorporation of source and consumer variability. IsoSource works by simulating values of the dietary proportions p on a grid to produce multiple valid solutions to the linear system. The valid solutions could be plotted in a histogram-like fashion, although they did not represent probability distributions, rather simply the range of values that might be plausible given the geometry of the system.

These initial attempts were formalised in a Bayesian fashion in the models MixSIR (Moore and Semmens, 2008) and SIAR (Parnell *et al.*, 2008). The MixSIR model can be thought of as a simplification of Equation 1 without explicit random effects across the dietary proportions and with $\boldsymbol{\Sigma}$ set to zero. The sources and TEFs are treated as independent across isotopes and are given fixed values for their mean and variance. The dietary proportions are given independent Beta-distributed priors or, in a later version, a Dirichlet distribution. Because the dietary proportions are the only parameters in the model, it can be fitted extremely efficiently using importance resampling (e.g. Robert and Casella, 2005) on a grid encompassing the range of proportion values. Updated versions of the MixSIR model have included random effects in the dietary proportions through the clr transform, and also have allowed for hierarchical models to be fitted, most elegantly in capturing familial relationships affecting the diet of gray wolves in British Columbia (Semmens *et al.*, 2009). These latter advanced versions of MixSIR are fitted using the JAGS software (Plummer, 2003).

The SIAR model (Parnell *et al.*, 2010) is in many ways similar to the basic MixSIR model (and thus still a simplification of Equation 1) although includes a residual component, which is treated as independent between isotope values. The model also allows for concentration dependencies (q_{kj}, vectorised as $\boldsymbol{q}$) that quantify the concentration of the chemical element (e.g. carbon or nitrogen) in the given food source (Phillips and Koch, 2002). They can be added in to our model by replacing $\boldsymbol{p}_i$ in Equation 1 with $\boldsymbol{p}_i \oplus \boldsymbol{q}$ where $\oplus$ is the simplex perturbation (Egozcue *et al.*, 2003). Usually the elements of $\boldsymbol{q}$ are either fixed or given a suitably informative prior distribution. The SIAR model is fitted using standard Markov chain Monte Carlo (MCMC) with Metropolis–Hastings steps to update the dietary proportions.

More recently, the IsotopeR model (Hopkins and Ferguson, 2012) has been introduced, which extends the SIAR/MixSIR models to a multivariate setting (both sources and TEFs are multivariate normal) and partitions the residual covariance $\mathbf{\Sigma}$ into a mass spectrometer calibration error and that of residual error. They further allow for the sources, TEFs and dietary proportions to be random effects with the latter obtained through the clr transform (see next section for further discussion). The model is fitted in JAGS (Plummer, 2003) but does not allow for covariate information or for the estimation of the dietary random effect variance.

Aside from explicit SIMM model development, recent focus has also been on the performance of SIMMs in non-ideal conditions, most notably with respect to the characterisation of source and TEF values by Bond and Diamond (2011). Clearly, it is absolutely vital that food sources are not excluded from the model as they will yield biased dietary proportions. Similarly, the estimation of the TEF values must be conducted appropriately or there will be some extra uncertainty in the estimated dietary proportions. In related work, Ward *et al.* (2011) considered the problem of (dis)aggregating sources and its effect on the resulting estimates. For example, if a consumer only eats part of a food source, it may be hard to obtain isotopic values from just that part. Similarly if two sources, although different species, lie in the same location in iso-space, it may be impossible for the model to determine the difference in their dietary consumptions. Thus, on occasion, it may be pertinent to aggregate sources without any loss of information. Alternatively, the aggregation can be accomplished with fewer assumptions if it is calculated *a posteriori* by combining the negatively correlated dietary proportions.

A number of other application areas link with the SIMM formulation we have presented previously. These go by various names such as ‘end member analysis’, ‘receptor modelling’, ‘mass balance analysis’ and ‘source apportionment’. However, all follow similar data formats with the aim of estimating the unknown (or only partially known) proportional contributions. In some of the earlier work on receptor models, Henry (1997) creates a multivariate air quality model where both the sources and the compositions are unknown, and discusses the determinacy of the model in the absence of both model and observational error. Billheimer (2001) extends such a model in a Bayesian fashion to include uncertainty, although the data here are also compositions. Park *et al.* (2001) allow for a model where both sources and proportions are unknown, and thus becomes a factor analysis model (also identified and used by Christensen *et al.*, 2006), and even allow for time-moving sources. They use a truncated normal prior on the proportions and restrict the source matrix to allow for its estimation. Lastly, Lingwall *et al.* (2008) build an alternative Bayesian version of the model with a generalised Dirichlet prior on the unknown proportions.

In the end member analysis literature, such models are used by geologists to determine the composition of, e.g. river sediment. In such cases the sources are usually known with minimal error, and the challenge is to estimate the proportional contributions of different sediment sources. A number of different methods have been proposed, e.g. Soulsby *et al.* (2003) and Brewer *et al.* (2011) (and references therein). Ours most closely resembles that of Palmer and Douglas (2008), although they use the additive log-ratio (alr) transformed proportions (see below for definition), which are given a spatial prior distribution.

Lastly, mass balance analysis is often used to quantify sources of, e.g. air pollution. Christensen (2004) fits a non-Bayesian model where all but one of the sources is known. They include an unknown intercept to determine whether an extra source is required. An alternative model specification is given by Bandeen-Roche and Ruppert (1991) with the Dirichlet distribution used to model the proportions. The focus on this latter paper is the attainment of consistent estimators of the unknown proportions. A nice summary of such models is given by Christensen and Gunst (2004).

4. STATISTICAL ISSUES IN STABLE ISOTOPE MIXING MODELS

The models we fit use ideas from Bayesian hierarchical modelling (BHM) (e.g. Gelman *et al.*, 2003) and compositional data analysis (Aitchison, 1986). BHM is now part of the standard toolbox of Bayesian statistics, and we do not discuss them further here. Of more interest however is the compositional structure applied to the dietary proportions as this can strongly affect the behaviour of the posterior distribution. A recent review of the state of compositional data analysis can be found in the study of Pawlowsky-Glahn and Buccianti (2011). Our models differ fundamentally from many standard problems as our compositions are not observed directly but are latent parameters to be estimated, constrained by their geometrical position in iso-space and the covariates upon which they may depend.

The starting point for modelling proportions in the early Bayesian SIMMs was that of the Dirichlet distribution, being perhaps the simplest valid distribution on the simplex. The usual prior distributions used were either flat where all Dirichlet parameters are set to 1 or the Jeffreys prior where all are set to K^{-1}. Unfortunately, as is well known (e.g. Aitchison, 1986), the Dirichlet distribution suffers from a very rigid sub-compositional independence assumption. This is not necessarily a problem when used as a prior distribution with fixed components as the posterior distribution may well show interesting sub-compositional properties. However, if dependence is to be modelled through hyper-parameters of the Dirichlet distribution (e.g. with covariates), this restriction will remain in the posterior.

A number of extensions to the Dirichlet have been proposed (e.g. Wong, 1998), but we focus here on the logistic-normal transformations of Aitchison (1986) and Egozcue *et al.* (2003), through which more flexible sub-compositional dependence can be obtained. The simplest of these is perhaps the additive log-ratio (alr), where $g(\boldsymbol{p}_i)$ in Equation 4 is set to be one of the chosen proportions (which is then removed from the composition) and $\boldsymbol{V} = \boldsymbol{I}$. However, this can perform poorly when there is no obvious choice of denominator and is not permutation invariant. The clr (defined in Equation 4 when $\boldsymbol{V} = \boldsymbol{I}$) removes the need for a choice of denominator in the log ratio but produces a covariate matrix of rank $K - 1$. It is also not sub-compositionally coherent (see Pawlowsky-Glahn and Buccianti, 2011, for further discussion of these terms). Finally, ilr uses orthonormal basis functions in the simplex to obtain coordinates that are isometric and satisfy the usual compositional requirements of coherence and permutation/sub-composition invariance. The choice of basis functions is somewhat subjective; we follow the method in the study of Egozcue *et al.* (2003). From a Bayesian perspective, the latent compositional parameters are more easily identifiable when working with the ilr.

Once a suitable transform has been chosen, it is feasible to include covariates or perform any of the traditional multivariate analysis techniques. Numerous examples can be found, including a geo-statistical framework (Tolosana-Delgado *et al.*, 2011), discrete time series (Barceló-Vidal *et al.*, 2011) or spectral methods (Pardo-Igúzquiza and Heredia, 2011). A popular topic is that of zero compositions or zero inflation (e.g. Butler and Glasbey, 2008) which is not a severe issue in SIMMs because we know from experimental observation that all dietary sources are consumed.

5. CASE STUDIES

We now present three case studies and show how our general model can be used in each of the scenarios. In the first case study, we analyse the diet of a small sample of geese from data previously studied by Inger *et al.* (2006). This first case study uses the model as proposed in Section 2 and can be seen as a small extension to that of Hopkins and Ferguson (2012). The second case study extends the geese model to allow for the inclusion of covariates or basis function models. Our final case study includes a compositional time series component where the sources and consumers are observed at different time points. In this final case study, the data are swallows consuming chironomid midges, other freshwater invertebrates and terrestrial invertebrates. In all cases, we run the models using the JAGS software and check convergence using the coda package (Plummer *et al.*, 2006) and the Brooks–Gelman–Rubin diagnostic (Gelman and Rubin, 1992; Brooks and Gelman, 1998).

5.1. Case study 1

Our first case study contains nine $\left(\delta^{13}\text{C}, \delta^{15}N\right)$ pairs of isotopic values taken from the blood plasma of Brent geese sampled on 26th October 2003. The food sources are *Zostera* spp, terrestrial grasses, *Ulva lactuca*, and *Enteromorpha* spp. Our posterior mean estimates for the sources are given in Table 1 from a set of 42 source measurements. The TEFs were taken from values in the literature (see references in Inger *et al.*, 2006, for more details) so that $\boldsymbol{\mu}_k^c = (1.63, 3.54)^T$ and $\boldsymbol{\Sigma}_k^c = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ for all k. For a simple random effects formulation, we set $\gamma_{ik} = 0$ over all i and k. An iso-space plot for the data (where the sources have been corrected by the TEFs) is shown in Figure 1. The JAGS model was run for three chains over 50 000 iterations, removing 10 000 for burn-in and thinning by a factor of 20.

A density plot of the mean dietary proportions is shown in the left panel of Figure 3. They can be seen to compare favourably with the simpler SIAR model (see the function `siardemo` in Parnell *et al.*, 2008), although in this case, we have extra information covering individual dietary estimates, as well as improved estimation of the source and TEF random effects. In particular, the flexibility of the hierarchical

Table 1. Estimates of the source means and covariance matrices for the geese data of case study 1

Source	*Enteromorpha*	Terr grasses	*Ulva lactuca*	*Zostera*
$\boldsymbol{\mu}_s^k$	$(-14.06, 9.82)^T$	$(-30.88, 4.43)^T$	$(-11.17, 11.2)^T$	$(-11.17, 6.45)^T$
$\boldsymbol{\Sigma}_s^k$	$\begin{bmatrix} 1.37 & 0 \\ 0 & 1.37 \end{bmatrix}$	$\begin{bmatrix} 0.41 & 0 \\ 0 & 5.15 \end{bmatrix}$	$\begin{bmatrix} 3.83 & 0.85 \\ 0.85 & 1.24 \end{bmatrix}$	$\begin{bmatrix} 1.48 & -0.56 \\ -0.56 & 2.16 \end{bmatrix}$

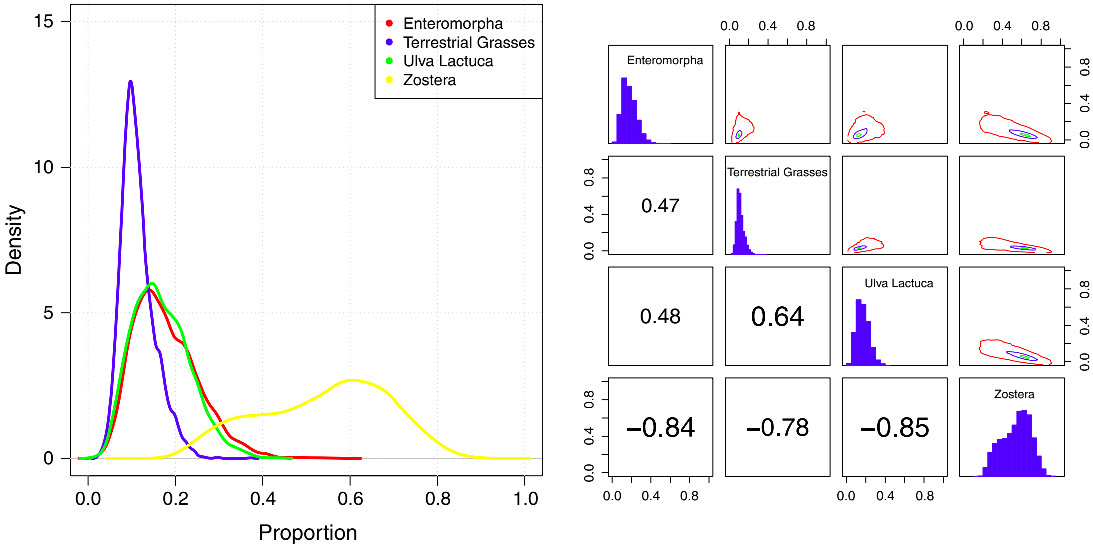

Figure 3. Left panel: density plot of mean dietary proportions for the geese data of case study 1. Right panel: matrix plot of the posterior dietary proportions obtained from the geese data. The upper-diagonal shows a contour plot; the diagonal, a histogram; and the below-diagonal, the correlation between the different sources

formulation through the ilr transform allows for some multimodality in the posterior distributions. The right panel of Figure 3 shows a matrix plot of the joint behaviour of the dietary proportions. These can be useful in determining unavoidable model inadequacy, for example, when it is impossible to ascertain which food sources are being consumed together. A strong negative correlation indicates that the food sources are indistinguishable. For example, the strong negative correlation between *Zostera* and *U. lactuca* indicates that, whilst it is clear they are consuming mainly *Zostera*, the balance between the two cannot be exactly determined.

There are many other useful statistics we can calculate here quite simply from the posterior distributions. In particular, we can focus on individual level variation by calculating, for example, the probability that an individual consumes more *Zostera* than another. Similarly, we can estimate the within source variation to determine whether there is more variation amongst consumers in some sources rather than others. Lastly, it is often desirable to perform model comparison diagnostics to determine whether certain parts of the model can be removed without a detrimental effect on prediction. However, we do not perform any further analysis on this data set, preferring instead to use these tools when analysing the more sophisticated data in the succeeding discussion.

5.2. Case study 2

We now extend our Goose model to a larger data set of 248 observations collected over the period of October 2003 to April 2005 of which the previous case study was just a small part. The sources and TEFs are believed to be stable over the course of the study so the source means and covariances are the same as Table 1. The diet of the geese will however vary during the season due to variations in abundance of food sources along with social and demographic factors. An iso-space plot for the full data is shown in Figure 4. The iso-space plot appears to show the diet during October to be focussed mainly on *Zostera*, moving on to *U. lactuca* and/or *Enteromorpha* during November/December. In January and February, the diet appears to be relatively mixed, but focussing almost solely on terrestrial grasses around April. In addition to the stable isotope information, we have covariates that state the goose's sex and whether they are juvenile or adult.

We consider six possible models for the dietary behaviour, accounting for the covariates. In each case, we set $\boldsymbol{\gamma}_{ik} = \boldsymbol{X}_i^T \boldsymbol{\beta}_k$ where $\boldsymbol{X}_i$ is an L-vector of covariates or basis functions for consumer i, and $\boldsymbol{\beta}_k$ is an L-vector of regression parameters associated with ϕ_k. Note that, when using the ilr transform, the parameters $\boldsymbol{\beta}_k$ do not have any association with source k. With the extra parameters in the model, convergence of the MCMC algorithm is a problem because the parameters are often highly correlated across sources, although this is somewhat lessened with appropriate choice of $\boldsymbol{V}$ in the ilr. We find it helpful to re-parameterise with Helmert contrasts across sources so that $\boldsymbol{\gamma}_{i1} = \boldsymbol{X}_i^T \boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_{ik} = \boldsymbol{X}_i^T(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_k)$ for $k \geqslant 2$. We use the first source (*Enteromorpha*) as $k = 1$ as this seems to be consumed throughout the season, although this is obviously not known in advance.

In all cases, the parameters β are given vaguely informative $N(0, 10)$ distributions as a value of $|\phi|$ in excess of 10 is likely to yield dietary proportions near 100% (although again this depends on the correlation structure in the data set). We compare the different models using the deviance information criterion (DIC; *Spiegelhalter et al.*, 2002) and, for the final chosen model, posterior predictive distributions of the data. We use two versions of the DIC, the standard p_D method of Spiegelhalter *et al.* (2002), and the p_V method of Plummer (2008) that estimates the optimism as half the variance of the deviance, and thus penalises complex models more harshly. Due to the extra complexity in the models, we run them with three chains for 200 000 iterations, removing 20 000 for burn-in and thinning by 90.

The first model we try involves no covariates and is thus the same as that in case study 1. The second model includes an intercept term and time as a simple linear covariate. The third model replaces linear time with a single harmonic component so that $X_i^T = \left[1, \cos\left(\frac{2\pi t_i}{365}\right), \sin\left(\frac{2\pi t_i}{365}\right)\right]$ where t_i is the Julian day. The fourth, fifth and sixth models are expansions of model 3 to include

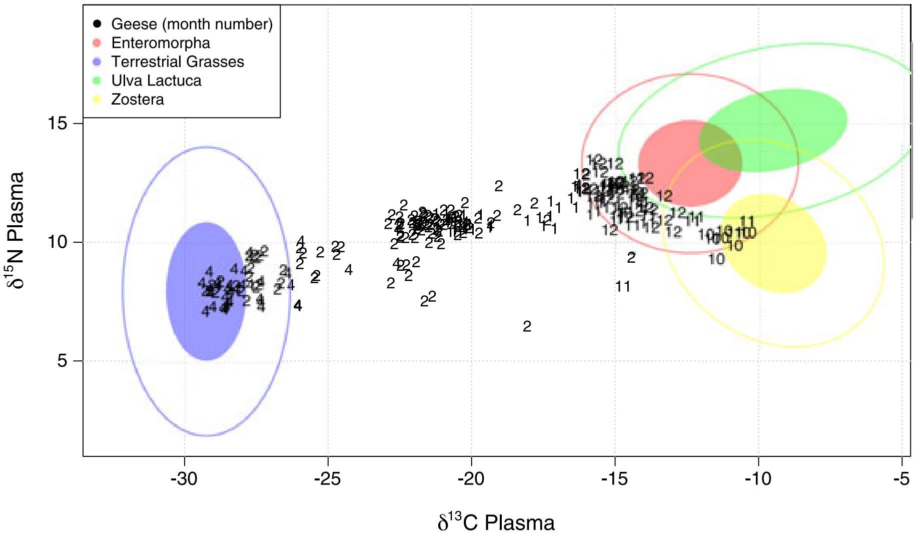

Figure 4. Iso-space plot of the full goose data set where the consumers are labelled by month. The data in case study 1 correspond to the data shown for month 10. Note that the sources and the TEFs are unchanged from Figure 1

juvenile/adulthood (model 4), sex (model 5) and also their interaction (model 6) as covariates. Table 2 shows the different models and the associated DIC values. Both versions of DIC seem to prefer models with the harmonic covariate, and also the addition of either sex or adulthood.

Figure 5 shows the relationship between Julian day and dietary proportion for the different sources for model 4. The switch from *Zostera*, to *Enteromorpha*, to terrestrial grasses is very clearly seen in both juveniles and adults, although the uncertainty in juveniles is slightly higher, especially around December 1 to February 1 when *Enteromorpha* dominates the other species, but by an uncertain amount. Figure 6 shows

Table 2. Table of models and model selection criteria. The models with the lowest deviance information criterion (DIC) are shown in bold

Model	Covariate(s)	DIC (using p_V)	DIC (using p_D)
1	None	26907	1210.0
2	Julian day (linear)	16957	524.3
3	Julian day (harmonic)	16683	385.1
4	**Julian day (harmonic), juvenile/adult**	**7551**	393.6
5	**Julian day (harmonic), sex**	8600	**382.8**
6	Julian day (harmonic), juvenile/adult, sex, interaction	10812	382.9

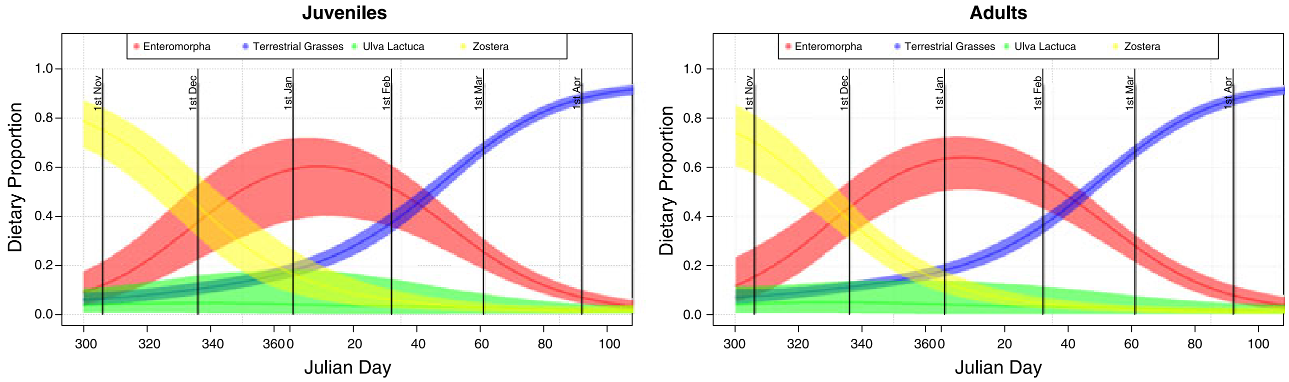

Figure 5. Plot of mean proportion values against Julian day for model 3. The left panel shows estimates for juveniles, whereas the right panel shows adults. The solid lines show median estimates, the outer of the polygons show 90% credibility intervals. The geese appear to focus on *Zostera* around November before moving on to *Enteromorpha* and then terrestrial grasses. There is slightly more uncertainty in the juveniles than the adults

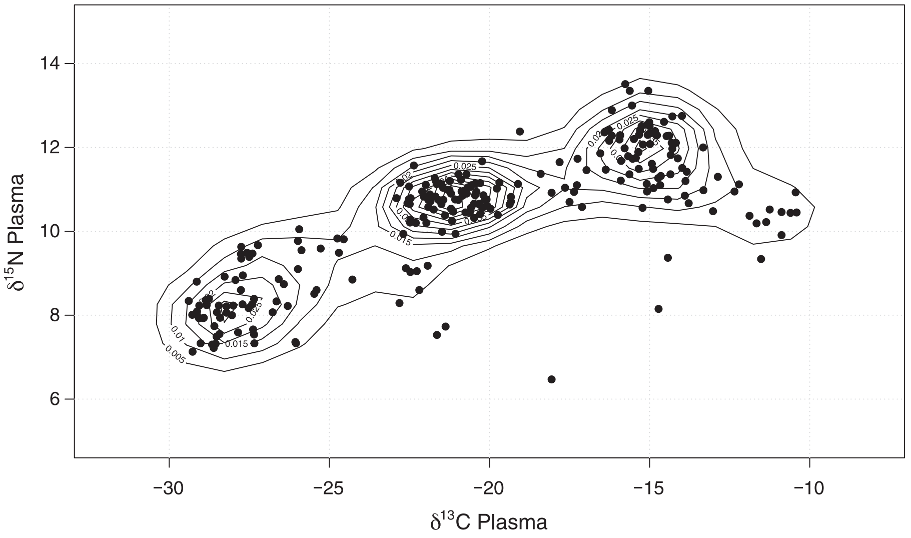

Figure 6. Plot of the predictive distribution of the data from model 3. The observations are shown as filled circles, whereas the posterior predictive density is represented by the contours

the posterior predictive distribution of the data under model 4. The model seems to predict the data well, the three modes corresponding to the main sampling times of October, February and April.

5.3. Case study 3

Our last case study concerns the dietary behaviour of barn swallows (unpublished data collected by the authors). The data are stable isotope ratios of blood plasma samples from birds captured between May and August in 2009. Again, we use Julian day to determine their behaviour over time. However, in this scenario, the sources (chironomid midges, other freshwater invertebrates and terrestrial invertebrates) are also expected to change over time and may have been observed on different days to that of the swallows. We thus expand our model to include a temporal component on the source vector as well as that of the dietary proportions. In each case, we use P-splines (Eilers and Marx, 1996) to explain the suitably flexible behaviour. Other time series methods, such as random walks or Levy processes, may also be appropriate. We write continuous time as t_i so that the consumers are now $\boldsymbol{Y}(t_i)$. We set $\gamma_k(t_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}_k$ where now $\boldsymbol{X}_i$ is a an L-vector of cubic B-spline basis functions evaluated at time point t_i and β_k are weights for each basis function on source k. The P-spline formulation is completed by giving a random walk prior such that $\beta_{lk} - \beta_{l-1,k} \sim N\left(0, \tau_k^{-1}\right)$ where τ_k is a roughness parameter associated with source k and given a weakly informative gamma $Ga(2, 1)$ prior.

The sources are now described by a multivariate spline model so that source data pairs, denoted $\boldsymbol{Y}^s_{ik}(t)$ for the source experimental J-vector at time t on source k, are distributed as $N\left(\left[\boldsymbol{X}^T\boldsymbol{\beta}^s_{1k}, \ldots, \boldsymbol{X}^T\boldsymbol{\beta}^s_{Jk}\right]^T, \boldsymbol{\Sigma}^s_k(t)\right)$ independently for each source k. The number of observations for each source N^s_k is likely to be variable and certainly not equal to the number of consumer observations N. Here, the spline parameters $\boldsymbol{\beta}^s_{jk}$ determine the mean behaviour of source k over time on isotope j. The $J \times J$ variance matrix $\boldsymbol{\Sigma}^s_k(t)$ is also allowed to change over time with diagonal elements given log splines: $N\left(\boldsymbol{X}^T\boldsymbol{\beta}_\Sigma, \kappa_\Sigma\right)$. The cross isotope covariance is parameterised through a single correlation parameter for each source, denoted ρ_k, and does not change over time. A spline could also be used here (e.g. on the arctangent of ρ_k) but this was not found to improve the fit. The source spline model was run for each of the three sources in turn and used to calculate maximum a *posteriori* estimates of $\boldsymbol{\mu}^s_k(t)$ and $\boldsymbol{\Sigma}^s_k(t)$ for all times t at which consumer data were available. An iso-space plot of the swallows data showing the output of the source spline model is shown in Figure 7.

We use the source model predictions in our standard SIMM with the spline formulation on the proportions as outlined in the previous discussion. An alternative model where all parameters are estimated simultaneously would be possible, if rather slow. Instead, we fix these parameters so that the source and dietary proportions are run separately (using the ideas presented in Appendix A). For both the source and the dietary proportion models, we run for 200 000 iterations, removing 20 000 for burn-in and thinning by 90. For both models, we use 25 knots, which seems to cover the flexibility in the data adequately.

Figure 8 shows the posterior dietary proportion estimates over Julian day for the swallows, predicted from the resulting spline parameter estimates. The results clearly indicate that they are feeding on mainly fresh water invertebrates during the early part of the data before concentrating on chironomids around the start of August. Figure 9 shows the predictive distribution of the data under this model. The fit appears satisfactory.

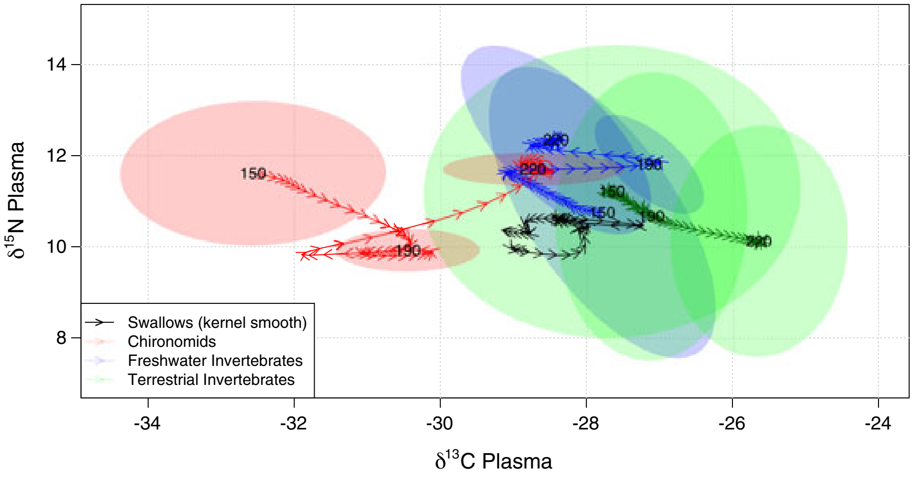

Figure 7. Iso-space plot of the swallows data. The source spline model has been run to obtain estimates of the source means and covariances throughout the study period. Arrows indicate the direction of movement over time for the sources and the swallows. The Julian day and the 50% standard ellipses are given for Julian days 150, 190 and 220. Note that the chironomids' δ^{13}C values increase over time from the start of the study period. A similar occurrence can be seen in the terrestrial invertebrates. The data are shown as kernel-smoothed estimates of the swallows' isotope data again over Julian day, starting at 150 and ending on day 220. The swallows can be seen to also increase their δ^{13}C values up until around day 180 whereupon the δ^{13}C returns towards its original value. This plot should be read in conjunction with Figure 8

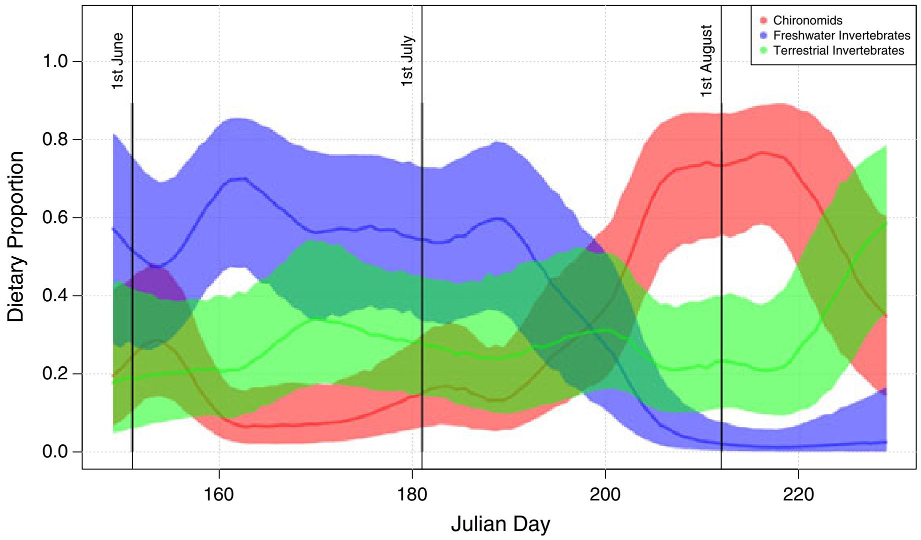

Figure 8. Plot of dietary proportion values against Julian day for the swallows data of case study 3. The solid lines show median estimates, whereas the filled polygons show 90% credibility intervals for each source

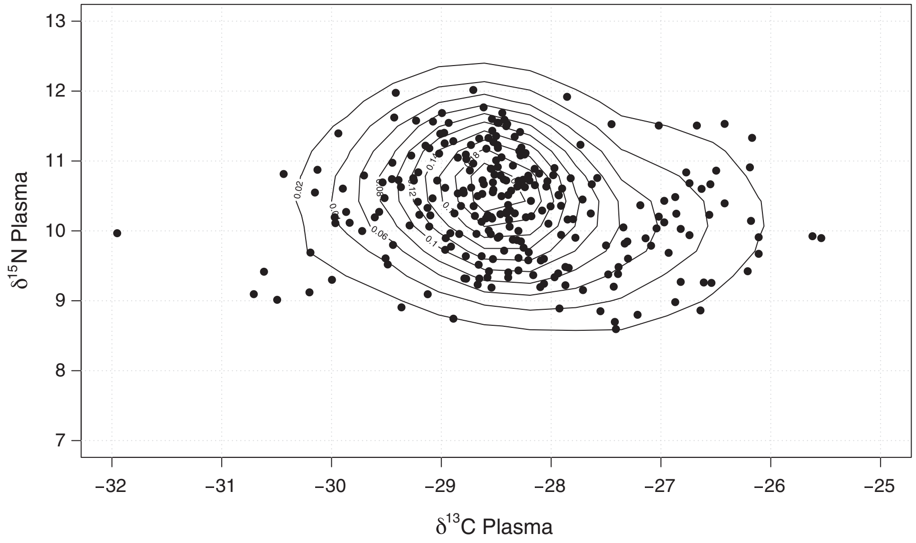

Figure 9. Predictive distribution of the data for the swallows example. The observations are shown as filled circles, whereas the posterior predictive density is represented by the contours

6. DISCUSSION AND FUTURE DIRECTIONS

The SIMM formulation outlined in Section 2 allows for a rich framework upon which to include a variety of other statistical structures. The basis of such models is a mixture compositional structure applied to the dietary sources. In the case studies in the previous discussion, we have illustrated how some simple regression and smoothing models might be included. The results seem useful and allow for many interesting findings, which would not have otherwise been possible without, e.g. the time series or spline components.

The main challenges in such models are that the source and TEF values are fully and correctly characterised. It is a simple geometrical exercise to verify that if sources are missing from the data set, or that the TEF means and errors are poorly estimated, the dietary proportion estimates will be biased. Such problems have been considered before, although not directly in the SIMM literature; Christensen (2004); Bandeen-Roche and Ruppert (1991). Missing source biases are compounded by the compositional nature of the problem, where if one dietary proportion is poorly estimated, then the others may well be too. There is no statistical test for missing sources; we rely on the ability of the ecologist to observe the system adequately. This may be particularly important if multiple organisms and sources are to be analysed simultaneously in a dietary network. Such models are not to be encouraged unless there is very strong evidence that no sources are missing and that TEFs are estimated correctly. It is up to the researcher to know their system and which sources are being consumed before attempting to use a SIMM. By knowing the system and studying it in detail, we can restrict the source pool to the relevant set. We may identify candidate sources from: behavioural observation, faecal analysis, gut content analysis, and so on. Of course, one may miss a rare source, but if it is that rare, then its contribution to diet is negligible, and its omission from the mixing model will make little or no difference.

It is reasonably straightforward to expand our approach for other values of J where differing numbers of isotopes may be available for analysis. It should be noted, however, that if $J \geqslant 3$, it becomes harder to determine model fit, especially as no obvious iso-space plot can be created. The solutions to this problem may lie in closer scrutiny of predictive distributions, and some subjective judgement over the size of the residual covariance matrix Σ. In such scenarios extra care is required in reporting the output of the SIMM.

There are further opportunities for expansion of the models. Some of these may include:

1. The simultaneous inclusion of multiple tissue varieties. It is often the case that different tissues are sampled on the same consumer as they are captured. These differentially replenished tissues will represent the dietary proportions consumed over different periods. For example, whereas blood plasma might represent the immediately sampled food sources, feathers might represent the diet consumed over the previous few months. A long-term data set where multiple tissues are analysed simultaneously may allow for increased precision in the dietary proportions. Alternatively, it may be possible to estimate the time scale over which the tissues are responding.
2. Clustering/mixture models to determine groupings. If there are hidden groupings amongst the organisms, it may be possible to discern them using a model-based clustering approach (sensu Fraley and Raftery, 2002). Even without such groupings, increased flexibility can be obtained by using mixtures of Gaussian distributions to model non-parametric behaviour.
3. Long-tailed multivariate distributions to account for outliers or small sample sizes. Clearly where sample sizes are small the multivariate Gaussian assumption (especially for sources) may be invalid, and thus, heavier-tailed distributions may be required. One such which seems to be most easily fitted is the multivariate normal-inverse Gaussian (Barndorff-Nielsen, 1997), which has been used previously in clustering and financial settings.

These are just three of the active areas of research to which SIMMs are being applied by our group. Many others are likely to appear as the field advances. We hope to report soon on these exciting new developments.

Acknowledgements

The authors would like to thank John Connolly and two anonymous reviewers for their insightful comments on an earlier draft. The information in this document has been funded in part by the US Environmental Protection Agency. It has been subjected to the Agency's peer and administrative review, and it has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

REFERENCES

Aitchison J. 1986. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd.: London, UK.

Bandeen-Roche K, Ruppert D. 1991. Source apportionment with one source unknown. *Chemometrics and Intelligent Laboratory Systems* **10**(1–2): 169–184.

Barceló-Vidal C, Aguilar L, Martín-Fernández JA. 2011. Compositional VARIMA time series. In *Compositional Data Analysis: Theory and Applications*, Pawlowsky-Glahn V, Buccianti A (eds). John Wiley & Sons: Chichester; 87–102.

Barndorff-Nielsen OE. 1997. Normal inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics* **24**(1): 1–13.

Billheimer D. 2001. Compositional receptor modeling. *Environmetrics* **12**(5): 451–467.

Bond AL, Diamond AW. 2011. Recent Bayesian stable-isotope mixing models are highly sensitive to variation in discrimination factors. *Ecological Applications* **21**(4): 1017–1023.

Brewer MJ, Tetzlaff D, Malcolm IA, Soulsby C. 2011. Source distribution modelling for end-member mixing in hydrology. *Environmetrics* **22**: 921–932.

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

Butler A, Glasbey C. 2008. A latent Gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **57**(5): 505–520.

Christensen WF. 2004. Chemical mass balance analysis of air quality data when unknown pollution sources are present. *Atmospheric Environment* **38**(26): 4305–4317.

Christensen WF, Gunst RF. 2004. Measurement error models in chemical mass balance analysis of air quality data. *Atmospheric Environment* **38**(5): 733–744.

Christensen WF, Schauer JJ, Lingwall JW. 2006. Iterated confirmatory factor analysis for pollution source apportionment. *Environmetrics* **17**(2004): 663–681.

Egozcue J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. 2003. Isometric log-ratio transformations for compositional data analysis. *Mathematical Geology* **35**(3): 279–300.

Eilers PHC, Marx BD. 1996. Flexible smoothing with B-splines and penalties. *Statistical Science* **11**(October 2): 89–121.

Fraley C, Raftery AE. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**(458): 611–631.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. *Bayesian Data Analysis*, (2nd edn). Chapman and Hall/CRC: Boca Raton, Florida.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* **2**: 457–472.

Henry RC. 1997. History and fundamentals of multivariate air quality receptor models. *Chemometrics and Intelligent Laboratory Systems* **37**(1): 37–42.

Hopkins JB, Ferguson JM. 2012. Estimating the diets of animals using stable isotopes and a comprehensive Bayesian mixing model. *PLoS ONE* **7**(1): e28478.

Inger R, Bearhop S. 2008. Applications of stable isotope analyses to avian ecology. *Ibis* **150**: 447–461.

Inger R, Ruxton GD, Newton J, Colhoun K, Robinson JA, Jackson AL, Bearhop S. 2006. Temporal and intrapopulation variation in prey choice of wintering geese determined by stable isotope analysis. *Journal of Animal Ecology* **75**: 1190–1200.

Lingwall JW, Christensen WF, Reese CS. 2008. Dirichlet based Bayesian multivariate receptor modeling. *Environmetrics* **19**(6): 618–629.

Liu F, Bayarri MJ, Berger JO. 2009. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis* **4**(1): 119–150.

Moore JW, Semmens BX. 2008. Incorporating uncertainty and prior information into stable isotope mixing models. *Ecology Letters* **11**(5): 470–480.

Palmer MJ, Douglas GB. 2008. A Bayesian statistical model for end member analysis of sediment geochemistry, incorporating spatial dependences. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **57**(3): 313–327.

Pardo-Igúzquiza E, Heredia J. 2011. Spectral analysis of compositional data in cyclostratigraphy. In *Compositional Data Analysis: Theory and Applications*, Pawlowsky-Glahn V, Buccianti A (eds). John Wiley & Sons: Chichester; 282–288.

Park ES, Guttorp P, Henry RC. 2001. Multivariate receptor modeling for temporally correlated data by using MCMC. *Journal of the American Statistical Association* **96**(456): 1171–1183.

Parnell AC, Inger R, Bearhop S, Jackson AL. 2008. SIAR: stable isotope analysis in R. http://cran.r-project.org/web/packages/siar/index.html.

Parnell AC, Inger R, Bearhop S, Jackson AL. 2010. Source partitioning using stable isotopes: coping with too much variation. *PLoS ONE* **5**(3): e9672.

Pawlowsky-Glahn V, Buccianti A. 2011. *Compositional Data Analysis: Theory and Applications*. Wiley-Blackwell: Chichester, UK.

Phillips DL. 2012. Converting isotope values to diet composition: the use of mixing models. *Journal of Mammalogy* **93**(2): 342–352.

Phillips DL, Gregg JW. 2001. Uncertainty in source partitioning using stable isotopes. *Oecologia* **127**: 171–179.

Phillips DL, Gregg JW. 2003. Source partitioning using stable isotopes: coping with too many sources. *Oecologia* **136**(2): 261–269.

Phillips DL, Koch PL. 2002. Incorporating concentration dependence in stable isotope mixing models. *Oecologia* **130**(1): 114–125.

Plummer M. 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling.

Plummer M. 2008. Penalized loss functions for Bayesian model comparison. *Biostatistics (Oxford, England)* **9**(3): 523–539.

Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC.

Robert C, Casella G. 2005. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer: New York, USA.

Semmens BX, Ward EJ, Moore JW, Darimont CT. 2009. Quantifying inter- and intra-population niche variability using hierarchical Bayesian stable isotope mixing models. *PLoS ONE* **4**(7): 9.

Soulsby C, Petry J, Brewer M, Dunn S, Ott B, Malcolm I. 2003. Identifying and assessing uncertainty in hydrological pathways: a novel approach to end member mixing in a Scottish agricultural catchment. *Journal of Hydrology* **274**(1-4): 109–128.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society (Series B)* **64**: 583–639.

Tolosana-Delgado R, van den Boogaart KG, Pawlowsky-Glahn V. 2011. Geostatistics for compositions. In *Compositional Data Analysis: Theory and Applications*, Pawlowsky-Glahn V, Buccianti A (eds), Chapter 6. John Wiley & Sons: Chichester, UK; 73–84.

Ward EJ, Semmens BX, Phillips DL, Moore JW, Bouwes N. 2011. A quantitative approach to combine sources in stable isotope mixing models. *Ecosphere* **2**(2): art19.

Wong TT. 1998. The generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation* **97**: 165–181.

APPENDIX A. MODEL ISSUES WHEN THE SOURCE OR TROPHIC ENRICHMENT FACTOR RAW DATA ARE NOT UPDATED AS PART OF THE MODEL

When all the source and TEF data ($\boldsymbol{Y}^s$ and $\boldsymbol{Y}^c$) are available we can fit the model outlined in Equation 5 where the source and TEF mean and variance terms are estimated simultaneously from the three different data sets. In such a scenario, the complete conditionals for the mean terms are as follows:

$$\pi\left(\boldsymbol{\mu}_k^s|\ldots\right) \propto \left[\prod_{i=1}^{N_k^s} \pi\left(\boldsymbol{Y}_{ik}^s|\boldsymbol{\mu}_k^s, \boldsymbol{\Sigma}_k^s\right)\right] \underline{\left[\prod_{i=1}^{N} \pi\left(\boldsymbol{s}_{ik}|\boldsymbol{\mu}_k^s, \boldsymbol{\Sigma}_k^s\right)\right]} \pi\left(\boldsymbol{\mu}_k^s\right)$$

$$\pi\left(\boldsymbol{\mu}_k^c|\ldots\right) \propto \left[\prod_{i=1}^{N_k^c} \pi\left(\boldsymbol{Y}_{ik}^c|\boldsymbol{\mu}_k^c, \boldsymbol{\Sigma}_k^c\right)\right] \underline{\left[\prod_{i=1}^{N} \pi\left(\boldsymbol{c}_{ik}|\boldsymbol{\mu}_k^c, \boldsymbol{\Sigma}_k^c\right)\right]} \pi\left(\boldsymbol{\mu}_k^c\right)$$

where all terms are Gaussian. Similarly, the $J \times J$ variance matrices have complete conditionals such that

$$\pi\left(\boldsymbol{\Sigma}_k^s|\ldots\right) \propto \left[\prod_{i=1}^{N_k^s} \pi\left(\boldsymbol{Y}_{ik}^s|\boldsymbol{\mu}_k^s, \boldsymbol{\Sigma}_k^s\right)\right] \underline{\left[\prod_{i=1}^{N} \pi\left(\boldsymbol{s}_{ik}|\boldsymbol{\mu}_k^s, \boldsymbol{\Sigma}_k^s\right)\right]} \pi\left(\boldsymbol{\Sigma}_k^s\right)$$

$$\pi\left(\boldsymbol{\Sigma}_k^c|\ldots\right) \propto \left[\prod_{i=1}^{N_k^c} \pi\left(\boldsymbol{Y}_{ik}^c|\boldsymbol{\mu}_k^c, \boldsymbol{\Sigma}_k^c\right)\right] \underline{\left[\prod_{i=1}^{N} \pi\left(\boldsymbol{c}_{ik}|\boldsymbol{\mu}_k^c, \boldsymbol{\Sigma}_k^c\right)\right]} \pi\left(\boldsymbol{\Sigma}_k^c\right)$$

which will be inverse Wishart when the prior distributions are inverse Wishart too. A simplified version of the model can be created by removing the underlined terms above so that they no longer contribute information about these parameters. In such a scenario the model has become *modularised* (in the sense of Liu *et al.*, 2009). Indeed, provided N_k^s is large in comparison with N, the information from s_{ik} and c_{ik} about these terms will be minimal. When the model is modularised, we can create individual posterior distributions for the source and TEF means and covariances during a separate offline modelling stage. This has the added advantage of greatly speeding up the posterior convergence of MCMC chains.

More commonly, we do not have full access to the source and TEF data sets, usually because only part of the information has been published. In such a scenario, the complete conditionals given earlier are set as point masses on the published values given (similar to that shown in Table 1). For situations where we have some information (e.g. on the sources but not the TEFs), we set these values at the posterior mean of the modularised model. Thus, for example, we use $\hat{\boldsymbol{\mu}}_k^s = \mathbb{E}\left[\boldsymbol{\mu}_k^s|\boldsymbol{Y}^s\right]$ and $\hat{\boldsymbol{\Sigma}}_k^s = \mathbb{E}\left[\boldsymbol{\Sigma}_k^s|\boldsymbol{Y}^s\right]$. As these parameters are ancestral to the parameters of interest (the dietary proportions $\boldsymbol{p}$), we expect the influence of the removal of this uncertainty to be minimal.

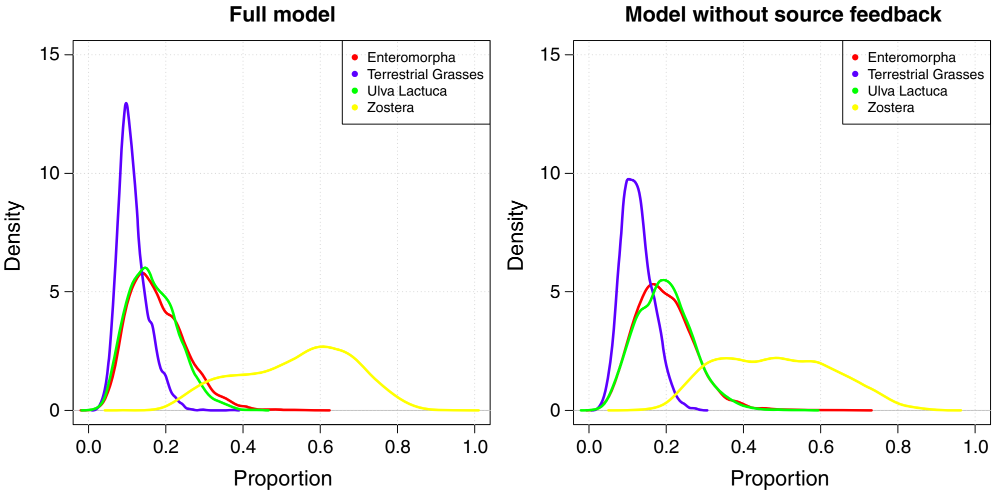

Figure A.1. Plot of the posterior densities for the dietary proportion values under two different models. The left-hand plot shows the posterior mean dietary proportions for the model as given in Equation 5, whereas the right-hand plot shows the posterior proportions when the source mean and covariance parameters are assumed fixed at their posterior mean values

We explore the effect of such a change on our first case study data set by comparing the posterior distribution of the proportions under each scenario. Figure A.1 shows the results. Whereas the mean for *Zostera* is very slightly higher under the full model, there is little difference in the shape of the posterior distributions between the two scenarios. Closer inspection of Figure A.1 reveals that the fixed-parameter model (right panel) has slightly increased uncertainty compared with that of the full model. This is most likely because the full model incorporates more data and so gives tighter posterior distributions on $\boldsymbol{\mu}_k^s$ and $\boldsymbol{\Sigma}_k^s$. A more complete description of the effects of modularisation is given by Liu *et al.* (2009). For the richer models (such as that of case studies 2 and 3), the full model cannot be fitted without considerable extra computing resources. We thus do not make any further comparisons here.