A quantitative approach to combine sources in stable isotope mixing models

ERIC J. WARD,† BRICE X. SEMMENS, DONALD L. PHILLIPS, JONATHAN W. MOORE, and NICOLAAS BOUWES

1National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, Washington 98112 USA
2U.S. Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Western Ecology Division, Corvallis, Oregon 97333 USA
3Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, California 95060 USA
4Eco Logical Research, Inc., 456 South 100 West, Providence, Utah 84332 USA

Abstract. Stable isotope mixing models, used to estimate source contributions to a mixture, typically yield highly uncertain estimates when there are many sources and relatively few isotope elements. Previously, ecologists have either accepted the uncertain contribution estimates for individual sources or addressed the problem in an ad hoc way by combining either related sources prior to analysis or the estimated proportions of related sources following analysis. Neither of these latter approaches explicitly account for uncertainty in source combinations within the likelihood framework. In this paper we incorporate uncertainty in both the number of source groups and group assignment within a formal Bayesian mixing model framework. By dynamically exploring model complexity due to aggregating sources based on shared proportional contributions, we can estimate posterior probabilities of alternative group configurations, and construct posterior dendrograms of group membership. We apply this method to simulated data, and illustrate applications to two consumer datasets (rainbow trout, coastal mink). Our results demonstrate that estimating, rather than fixing, the number of proportional contributions in a mixing model can improve model inference and reduce bias in estimates of source contributions to a mixture.

Key words: Bayesian analysis; diet; mixing model; source partitioning; stable isotope.

INTRODUCTION

Stable isotope mixing model analysis is an increasingly common approach in environmental sciences. Ecologists use mixing models to determine the proportional contribution of sources to a mixture based on their respective isotope signatures (Schwarcz 1991, Phillips 2001); applications include estimation of the relative importance of food sources to animal diets, pollution sources to air or water bodies, carbon sources to soil organic matter, and numerous others (Lajtha and Michener 1994). Initial mixing model methods relied on analytic solutions to mixing equations that are only tractable when the number of sources is equal to or less than 1 + the number of isotope signatures. Phillips and Gregg (2003) introduced an approach (IsoSource) that yields estimates of proportional contribution when the number of sources is too large to permit a unique solution. Subsequently, Moore and Semmens (2008) introduced a Bayesian mixing model framework capable of accounting for any number of sources. Once solutions with
any number of sources became possible, researchers were faced with the questions: how many sources are too many? Should the contribution of different sources to a mixture be aggregated or considered separately?

The decision as to what constitutes a mixing model source can have critical ramifications for the output of mixing models. Phillips et al. (2005) demonstrated that uncertainty in source contribution estimates depends on the grouping of sources (and variance of the aggregated sources). On the other hand, biological systems such as food webs are often complex (e.g., Polis 1991), and mixing models with overly grouped sources may miss important ecological complexity. Furthermore, mixing models yield biased results if known diet components are omitted from such models; in fact, most mixing model formulations assume that all source components are included in the analysis (Phillips and Gregg 2003, Moore and Semmens 2008).

The standard approach for reducing the number of sources in a mixing model is to combine sources. When source isotope signatures are similar and the combined source group has some biological significance, source reduction through combination can improve mixing model tractability and inference (Phillips et al. 2005). Phillips et al. (2005) presented both a priori and a posteriori approaches to combining sources. In the a priori approach, researchers decide on logical source groupings based on isotope equivalence and biological relatedness prior to analyzing the mixing model. In the a posteriori approach, mixing model outputs from the full set of sources are combined through both logical groupings and consideration for the fact that certain groupings will give more constrained and interpretable estimates of proportional source contributions than will others. This latter approach balances the need for informative model outputs with decisions regarding the biological and isotopic prudence of source combinations. It does not, however, explicitly incorporate the number of source groups, or how sources are divided among groups into mixing model inference. Doing so requires a model framework that evaluates the likelihood of combinations by dynamically exploring model complexity due to aggregation/disaggregation of source combinations. Fortunately, recent advances in Bayesian techniques such as reversible jump Markov chain Monte Carlo (MCMC) allow integration over models of different dimensions (King et al. 2009). As such, these techniques offer a natural extension to recently developed Bayesian stable isotope mixing models.

Bayesian mixing models have evolved rapidly in recent years from the original MixSIR model proposed by Moore and Semmens (2008), to the inclusion of residual error (Parnell et al. 2010), hierarchical variation (Semmens et al. 2009), and source uncertainty (Ward et al. 2010). The principal driver behind this evolution is the goal of more accurately accounting for uncertainty and variability in isotope data and the mixing process. Despite their sophistication, to date none of these methods have addressed the issue of estimating source combinations within the modeling framework. In this paper, we present a Bayesian mixing model formulation that evaluates the posterior probabilities associated with source combinations while simultaneously estimating the proportional contribution of each source group to the mixture. In essence, this method produces a similar result to the a posteriori approach of Phillips et al. (2005), but evaluates the likelihood of alternative source groupings within the mixing model. However, rather than simply pooling source data based on groupings, we instead preserve potential differences in source signatures and evaluate support for shared proportional contribution terms across sources. In so doing we capture likely source groupings based both on isotopic similarity and those resulting from non-distinct source contributions due to the geometry of the mixing space. To demonstrate the utility of this modeling approach we present model output based on simulated data and dietary data on two populations: coastal mink (Mustela vision) in Alaska originally published by Ben-David et al. (1997), and steelhead (Oncorhynchus mykiss) in Oregon. This method rigorously addresses the uncertainty of source classification, while simultaneously improving estimates of source contributions and reducing bias. All code is available for free, at http://www.ecologybox.org (“Mixing model with unknown groups”).
METHODS

For consistency, we retain the notation used in previous Bayesian isotope mixing model papers (Moore and Semmens 2008, Semmens et al. 2009, Ward et al. 2010). Given the isotopic signatures of mixtures ($x$) and source parameters ($u, \sigma$), the goal of isotopic mixing models is to estimate the vector of relative source contributions ($f$). The likelihood of diet contributions, $f$, from multiple sources to a mixture’s isotopic signature is partly dependent on whether and how sources are grouped (i.e., share elements of $f$), and which mixing model is used in the analysis (e.g., whether or not residual error is included; Parnell et al. 2010). Previously, we used information theoretic approaches to evaluate the data support for alternative mixing models (Semmens et al. 2010). Previously, we used information theoretic approaches to evaluate the data support for alternative mixing models (Semmens et al. 2010). Previously, we used information theoretic approaches to evaluate the data support for alternative mixing models (Semmens et al. 2010). Previously, we used information theoretic approaches to evaluate the data support for alternative mixing models (Semmens et al. 2010).

In order to account for uncertainty in source groupings, we integrate over both the number and assignment of source groups. This procedure is outlined in the following pseudo-code:

1. Draw a random number of groups, $G$ (assigned a Dirichlet prior). We use 2 groups as a lower bound and the total number of sources ($N$) as an upper bound.

2. Randomly assign each source (1:$N$) to a unique group (1:$G$) with a single shared diet contribution term.

3. Given the group membership, calculate the mean and variance of the groups. There are several approaches for doing this. If sources are assumed to be drawn from the same resource pool (e.g., Chinook salmon and sockeye salmon being combined into a ‘salmon’ group), the group means and variances are calculated using formulas for weighted means and pooled variances. These same formulas are typically used to combine sources a priori. The isotopic mean for each group becomes

$$u_k = \frac{\sum_{i=1}^{k} (n_i \cdot u_i)}{\sum_{i=1}^{k} n_i}$$

and the pooled sample variance for each group is

$$s_g^2 = \frac{\sum_{i=1}^{k} (n_i - 1) \cdot s_i^2}{\sum_{i=1}^{k} (n_i - 1)}.$$

In most situations, making the assumption of a common resource pool may not be appropriate given that ecologists typically identify source groups based on biological uniqueness. Under scenarios where resources are not drawn from a common pool, the group means and variances are summed rather than averaged, and independent of sample sizes,

$$u_g = \sum_{i=1}^{k} u_i$$

and

$$s_g^2 = \sum_{i=1}^{k} s_i^2.$$

Although this may seem counterintuitive, the mathematical justification can be demonstrated by examining the formula for the mixture mean and variance for an arbitrary isotope. Using a 4-source example, the mixture parameters are $u_{mix} = f_1u_1 + f_2u_2 + f_3u_3 + f_4u_4$ and $s_{mix}^2 = f_1^2s_1^2 + f_2^2s_2^2 + f_3^2s_3^2 + f_4^2s_4^2$. If we proposed combining sources 3 and 4, these equations could be written as $u_{mix} = f_1u_1 + f_2u_2 + f_3u_3 + u_4$ and $s_{mix}^2 = f_1^2s_1^2 + f_2^2s_2^2 + f_3^2(s_3^2 + s_4^2)$.
any other model parameters (such as residual variances). Using all model parameters, then calculate the likelihood (and posterior) conditional on the group means and variances.

It is important to emphasize that this approach evaluates the probabilities that the various sources being grouped in the mixing model share relative source contribution terms \((f)\). Two sources might have the same contribution if they occupy the same isotope space, or because they contribute similarly to a mixture despite occupying distinct regions of isotope space.

Within the Bayesian mixing model framework, parameter estimation can be done either using the Sampling-Importance-Resampling (SIR) algorithm (Moore and Semmens 2008) or MCMC techniques (Semmens et al. 2009, Parnell et al. 2010). We have implemented both approaches for this problem using the R statistical environment; because the code is more transparent, we focus on the SIR approach (http://www.EcologyBox.org).

To illustrate our modeling approach, we followed the approach of Phillips and Gregg (2003) and generated simulated data from a 4-source 2-isotope model. We then applied the mixing model algorithm incorporating grouping uncertainty to each model configuration. Hierarchical cluster analysis (‘hclust’ in R) was then used to construct posterior dendrograms between sources based on the dissimilarity matrices from each matrix of posterior group assignments. We chose to calculate the dissimilarity matrices using all saved SIR draws (integrating over the numbers of groups), however similar dendrograms may be generated for each level of grouping \((2, 3, \ldots, N)\).

To illustrate an ecological application of our procedure, we applied the model to two different stable isotope data sets: 1) stable isotope signatures of juvenile steelhead and their prey from the Middle Fork John Day river in Oregon, USA and 2) stable isotope signatures of coastal mink and their prey from SE Alaska, USA (Ben-David et al. 1997). The steelhead data come from the Integrated Status and Effectiveness Monitoring Program (ISEMP), coordinated through the Northwest Fisheries Science Center, NOAA. The ISEMP program is tasked with developing effective monitoring tools and techniques associated with restoration actions for Endangered Species Act (ESA) listed pacific salmonids. As part of this mission, ISEMP personnel collected stable isotope data in the Middle Fork John Day river in order to identify the importance of invertebrate prey sources to the summer diets of steelhead. At each of 8 sites, terrestrial invertebrates samples were collected and stable isotope analysis was done for ants and spiders (common terrestrial taxa). Additionally, benthic invertebrate samples were collected at each site. Invertebrates were identified to family, and stable isotope analysis was conducted on the 9 most abundant (estimated by weight) families. These families were subsequently assigned to 3 functional feeding groups: predators, detritivores and herbivores. Thus, the 5 prey groups considered in this example are spiders, ants, benthic predators, benthic herbivores and benthic detritivores (Fig. 1). An important consideration in developing mixing models for this system was whether prey groups should be further consolidated. For instance, should ants and spiders be grouped into a single terrestrial invertebrate category? Or, should benthic detritivores and herbivores be grouped into a single category given the high trophic overlap of their taxonomic members? In our analysis, we made the assumption that sources being combined could potentially be drawn from the same resource pool. Using our modeling approach with formulas for pooled means and variances, we demonstrate how to address these questions.

As an example of a situation where sources may represent distinct resource pools, we reanalyzed a classic dataset for coastal mink (Ben-David et al. 1997). This dataset is particularly appropriate, because the number of potential sources is large \((n = 7)\), and because it is well known, having been re-analyzed on multiple occasions (e.g., Phillips 2001, Phillips and Gregg 2003). In both the original analysis (Ben-David et al. 1997) and subsequent re-analyses (Phillips and Gregg 2003), each prey item was assumed to be a unique source group (Fig. 2). Because the 7 sources are thought to represent fundamentally different trophic pools, we assumed that sources were not being drawn from the same resource pool (but still may share proportional contribution terms).

In all our findings, model posterior results include independent draws from all likely source
combinations and associated proportional source contributions. For instance, in a mixing model with a maximum of 4 sources, one draw might include two source groups that combine 1, 2 and 3, 4; the next draw might include 3 source groups with 1 and 2 separate and 3, 4 combined. Using all posterior draws for a given model run, we can construct dendrograms that depict how often candidate sources are considered in the same group, and histograms that depict the frequency of different numbers of sources in the posterior chain. Researchers may choose the most likely number of sources and assign source membership based strictly on the frequency of posterior co-occurrence in order to identify the most parsimonious model formulation. However, other groupings with only slightly lower posterior probabilities may be preferred from a standpoint of biological relevance and interpretability. The computation of posterior probabilities for various competing models can alternatively allow the researcher to choose models that temper statistical likelihood with biological meaningfulness, weighing the tradeoffs between them.

RESULTS

We simulated four sets of source geometry, following Phillips and Gregg (2003), to illuminate uncertainty in both the number of groups and similarity between sources. In scenarios with diffuse solutions due to uninformative source geometry (Fig. 3A, B), our approach for including uncertainty in source groupings doesn’t provide a “magic bullet” for generating better estimates of relative contributions; systems with poor geometry will likely continue to have highly confounded estimates of relative contributions. In this first uninformative scenario, model posteriors favor grouping sources with disparate isotopic signatures because those combinations allow the group means to be identical to the consumer signature (Fig. 3A, B).

Considering our simulation studies, when sources 1 and 4 are isotopically identical (and presumably ecologically equivalent; Fig. 3C), not surprisingly the posterior dendrogram suggests combining sources 1 and 4 in the first group (Fig. 3D). When we rearranged sources 1 and 4 so that
they remain equidistant from sources 2, 3, and the consumer, but source 1 is no longer identical to source 4 (Fig. 3E), the posterior dendrogram (Fig. 3F) remains essentially unchanged. Why does the mixing model suggest combining such disparate signatures? A 4-source mixing model (i.e., no grouping of sources) allows sources 1 and 4 to be included in two ways: a) neither source contributes to the diet, or b) both sources contribute, and their relative contributions are equal. As such, the posterior estimates for the contribution of sources 1 and 4 in the 4-source model are identical ($\mu = 0.11$, $\sigma = 0.06$) and positively correlated. The lack of independence between these sources makes them a logical grouping despite their lack of isotopic equivalence.

Combining sources that are isotopically distinct may be most appropriate when the diet contributions of the sources being combined are deemed to be unimportant in the context of model inference. For instance, using similar source geometry but more sources, some degree of grouping is strongly supported (Fig. 4A, B). It may be reasonable to assume, for instance that sources 3/7/8 are isotopically distinct, but all have a similar estimated contribution. Treating each source as a unique group maximizes the flexibility of the mixing model, and allows each estimated diet proportion to have small variance. This assumption is supported is through the lens of mixture models because of the tradeoff between the bias and variance of the mixture means and variances (Fig. 4C, D). When the contribution of each source is estimated independently, the variance of the mixture is slightly lower than when sources 3/7/8 have the same contribution, however this increased precision is also accompanied by an increase in bias (Fig. 4E).

For the application to coastal mink data, where sources were a priori assumed to be drawn from unique resource pools, the highest posterior probability results from assigning each source to its own group (Fig. 5). If a researcher chooses some amount of grouping, the posterior dendrogram suggests several biologically disparate groups, such as aggregating rodents, ducks, and shrimp. Such a grouping doesn’t reflect the
trophic niche of these species; rather, it reflects the fact that the combined group (through shared contributions) yields a reasonable fit to the consumer data.

In contrast with the mink example, where the most likely solution assigned each source to a unique group, our steelhead example suggests that some of the benthic invertebrate source pools should be grouped. The highest posterior probability is assigned to scenarios with 4 groups.

Fig. 3. Posterior dendrograms for 3 simulated examples of source/consumer isotope geometry. The standard deviation for each source is included as a circle, and consumer signatures are indicated with solid points. Posterior dendrograms should be read from the bottom, with the most similar sources included in the first group, and the most dissimilar source as the last branch. The y-axis on the dendrogram represents the relative distance based on dissimilarity matrices from each matrix of posterior group assignments.
(the most probable source grouping being benthic herbivores and benthic detritivores, Fig. 5). There is considerable uncertainty in whether further aggregation should be done—but all levels of grouping are more strongly supported than no grouping (Fig. 5). This result is somewhat intuitive based both on the source geometry (Fig. 1), and trophic ecology—benthic herbivores and detritivores represent relatively small fractions of steelhead diet (<10%), compared to benthic predators or terrestrial insects, and the degree to which these groups can be readily split into functional feeding groups is uncertain (Mihuc 1997).
DISCUSSION

Despite a rapid evolution in stable isotope mixing models over the past decade, ecologists have continued to struggle with decisions regarding appropriate source combinations when the number of source items is large relative to the number of isotopes in a study (Phillips and Gregg 2003). Previous methods have considered uncertainty in source groupings (e.g., IsoSource; Phillips et al. 2005), however our method is the first to incorporate uncertainty in both the number of source groupings and assignment of sources to groups within the statistical model. Data support for these alternative source configurations are evaluated via the likelihood. By incorporating group uncertainty in a Bayesian statistical framework, we are able to make probabilistic statements about how sources should be clustered.

Fig. 5. Posterior distributions of the number of source groupings and the posterior dendrograms for our applications to coastal mink in SE Alaska (Ben-David et al. 1997; A, B) and the ISEM steelhead trout example (C, D).
The tradeoff between model bias (too few parameters) and variance has been well studied in ecology (Burnham and Anderson 2002), but not explored within the context of mixing models. Ecologists may expect that increasing the number of sources would increase the variance and reduce the bias, because more complex models generally are better approximations of reality. In the context of mixing models, the complete opposite may be true. Why would a more complex model be more precise yet more biased? The answer lies in the formula for calculating the variance of the mixture. Consider a 1-isotope 2-source model with sources A and B—the variance of the mixture is calculated as \( \sigma_{mix}^2 = p_{A1}^2 \sigma_A^2 + p_{B1}^2 \sigma_B^2 \), where \( p_{A1} \) is the contribution of source A for model 1. Creating a second model (model 2) that introduces a third source C will always reduce at least one of the contributions of sources A and B. This suggests that if the variances of the three sources are equal, adding a third source will force \( (p_{A2}^2 + p_{B2}^2 + p_{C2}^2) < (p_{A1}^2 + p_{B1}^2) \), making the variance of the more complex model smaller. In some cases, adding another source may increase the mixture’s variance, but only when the variance of the third source is large enough to compensate for the change in the variance weights \( (p_i^2) \).

Previous Bayesian stable isotope mixing models have assessed model complexity with information theoretic approaches (Semmens et al. 2009). However, these techniques can’t be used when there is uncertainty in how sources are assigned to groups, because the likelihood surface is multimodal across differing parameter dimensions. Like the information theoretic approach, posterior probabilities balance the tradeoff between model simplicity and variance; more estimated diet proportions increase model complexity, but also increase the bias of the model estimated mixture. Although we examined the sensitivity to the choice of mixing model (with or without residual error terms), model uncertainty was not explicitly included in the mixing model framework. Because we only considered these two relatively simple mixing model formulations, the posterior model probability of each could be estimated by integrating over both models (King et al. 2009; Ward et al., in press). Though our analysis implements Bayesian analyses, ecologists interested in maximum likelihood approaches could use our estimation method to integrate over groupings to find the model with the highest maximum likelihood.

Combining isotopically disparate sources may be counterintuitive, and in opposition to suggestions from previous work (Phillips and Gregg 2003, Phillips et al. 2005). However, sharing contribution terms across sources that represent different ecological roles may reduce uncertainty in some problems, particularly when such sources are not the principal subjects of inference. Further, an ecologist's understanding of biology may ultimately lead to grouping decisions that are sub-optimal from the standpoint of selecting the most parsimonious model. This is not to say that model probabilities trump biological theory; rather, it explicitly acknowledges that researchers should be aware of the potential tradeoff between these competing forms of knowledge. The tradeoff between model variance and bias in the context of model selection is well known (Burnham and Anderson 2002); in stable isotope mixing models, the relationship between variance and bias is complex due to the constraints of proportionality (Fig. 5). Information theoretic model selection tools (Semmens et al. 2009) or the probabilistic Bayesian approach used here seek to minimize the tradeoff between bias and variance.

While we anticipate that the methods developed here will aid ecologists in coping with uncertainty in how sources should be grouped, we acknowledge that they cannot recover informative parameter estimates from poorly formulated mixing models. For instance, the methods outlined here cannot overcome poor source geometry other than to explicitly demonstrate that little information exists to identify distinct diet contributions of prey (Phillips and Gregg 2003). When the consumer signature is within the source polygon, sources that do not define the perimeter of the polygon will add to mixture uncertainty, because they may or may not contribute to the mixture (Phillips and Gregg 2003). Additionally, it is still necessary to avoid discounting sources, particularly when such sources constitute a substantial part of the mixture. Finally, because the posterior parameter/model hypervolume increases geometrically with the number of sources included in the model, too many sources (~15 or more) will
ultimately make the problem intractable given existing computing power. On the other hand, researchers with a large number of potential sources should carefully consider whether it is appropriate to use mixing models to assessing the relative diet proportions of such a generalist. Assuming the data, isotope geometry and trophic ecology are suitable for mixing model analysis, our technique will help researchers refine the mixing model formulation in order to maximize inference. Up until now, the decision of how to group sources has been external to mixing models. Here we provide tools that will allow ecologists to combine biological understanding with model probabilities in order to identify mixing model formulations that maximize the utility of model findings in the context of both ecology and source geometry.

ACKNOWLEDGMENTS

We thank J. Yeakel for criticism that helped improve this manuscript. Nicholas Weber collected, processed, and compiled the ISEMP data used in this manuscript. E. J. Ward and B. X. Semmens contributed equally to this manuscript. Funding for ISEMP was provided by the Bonneville Power Administration. This document has been subjected to the U.S. Environmental Protection 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.

LITERATURE CITED


