Ecological Applications, 18(8), 2008, pp. 1901–1913

MARKOV CHAIN MONTE CARLO METHODS FOR ASSIGNING LARVAE TO NATAL SITES USING NATURAL GEOCHEMICAL TAGS

J. Wilson White,[1,2,4] Julie D. Standish,[1] Simon R. Thorrold,[3] and Robert R. Warner[1]

[1]*Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, California 93106 USA*
[2]*Department of Wildlife, Fish, and Conservation Biology, University of California, Davis, One Shields Avenue, Davis, California 95616 USA*
[3]*Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543 USA*

Abstract. Geochemical signatures deposited in otoliths are a potentially powerful means of identifying the origin and dispersal history of fish. However, current analytical methods for assigning natal origins of fish in mixed-stock analyses require knowledge of the number of potential sources and their characteristic geochemical signatures. Such baseline data are difficult or impossible to obtain for many species. A new approach to this problem can be found in iterative Markov Chain Monte Carlo (MCMC) algorithms that simultaneously estimate population parameters and assign individuals to groups. MCMC procedures only require an estimate of the number of source populations, and post hoc model selection based on the deviance information criterion can be used to infer the correct number of chemically distinct sources. We describe the basics of the MCMC approach and outline the specific decisions required when implementing the technique with otolith geochemical data. We also illustrate the use of the MCMC approach on simulated data and empirical geochemical signatures in otoliths from young-of-the-year and adult weakfish, *Cynoscion regalis*, from the U.S. Atlantic coast. While we describe how investigators can use MCMC to complement existing analytical tools for use with otolith geochemical data, the MCMC approach is suitable for any mixed-stock problem with a continuous, multivariate data.

Key words: Cynoscion regalis; *deviance information criterion; Gibbs sampler; Markov Chain Monte Carlo; mixed-stock analysis; mixture model; natal source; otolith geochemistry; population assignment; weakfish.*

Introduction

The precarious state of many exploited marine populations (Botsford et al. 1997, Jackson et al. 2001) has sparked considerable interest in place-based management, including the implementation of marine protected areas (Lubchenco et al. 2003). However, the success of place-based management can be complicated by the long-distance movements undertaken by many marine species. Most benthic organisms have dispersive planktonic larval stages, so juveniles recruiting to one area may have been spawned elsewhere (Mora and Sale 2002). Many other fishes migrate between feeding and spawning grounds as adults, so harvests may consist of a mixture of multiple independently reproducing stocks (e.g., Knutsen et al. 2007). In either case, effective management hinges upon successful determination of the natal origin of individuals at a particular location (Carr and Reed 1993, Warner et al. 2000, Botsford et al. 2003). Specifically, it is important to know how many natal sources contribute to a sample (Palsbøll et al. 2006, Waples and Gaggiotti 2006) and the degree to which larvae are exchanged among subpopulations within a larger metapopulation (Kritzer and Sale 2004). Modeling efforts make it clear that misjudging the number or identity of sources contributing to the harvested population at a particular location can lead to management failures (Crowder et al. 2000, Stockhausen et al. 2000). Recent advances in statistical computing and the development of Markov Chain Monte Carlo (MCMC) techniques offer some potential solutions to the daunting problem of mixed-stock analysis. Here we outline the basics of this analytical approach and illustrate its use with both simulated data and empirical data sets.

Determining natal origins remains a difficult problem in marine population ecology, especially for the case of larval dispersal (Levin 2006). While efforts to tag large numbers of larvae prior to dispersal have met with some success (Jones et al. 1999, 2005, Almany et al. 2007), the logistics of this approach are daunting at large scales. Attention has turned to the use of "natural" tags, including chemical composition of calcified structures in fish and invertebrates, to characterize stock structure in marine species (reviewed by Campana et al. 2000), identify different dispersal histories (Swearer et al. 1999), and track the movement of individuals among habitats over different life stages (Thorrold et al. 2001, Warner et al. 2005, Zacherl 2005, Becker et al. 2007, Standish et al. 2008). In particular, geochemical signatures in fish otoliths are widely used as records of the environmental

Manuscript received 26 October 2007; revised 19 March 2008; accepted 11 April 2008. Corresponding Editor: K. Stokesbury.
[4] E-mail: jwwhite@ucdavis.edu

history experienced by an individual (Campana and Thorrold 2001), and in this paper we use "otolith geochemistry" as a generic shorthand for all natural tag techniques. We describe the use of MCMC in an otolith geochemistry context, but the basic approach is suitable for any multivariate continuous data set describing a mixture of populations.

The assignment of post-dispersal individuals to the habitat or population of origin using otolith geochemistry typically requires two levels of sampling. An important first step is to characterize the elemental or isotopic signatures of potentially contributing populations or habitats by sampling pre-dispersal individuals. When dispersal occurs as larvae, this requires sampling propagules before they enter the plankton; for example by collecting benthic eggs (e.g., Ruttenberg and Warner 2006), pre-paturition larvae still inside the mother (e.g., from rockfishes [*Sebastes* sp.] or other ovoviviparous fishes; Warner et al. 2005), recently spawned pelagic larvae (for broadcast spawners, e.g., DiBacco and Levin 2000), or by culturing larvae in situ (Becker et al. 2007). Older individuals (i.e., post-disperal recruits or adults) are then sampled from locations of interest and geochemical signatures in that portion of the otolith deposited during the pre-dispersal stage are compared to the source signatures obtained in step one, typically using a multivariate technique such as discriminant function analysis (DFA; e.g., Brown 2006) or maximum likelihood algorithms (MLE; e.g., Thorrold et al. 2001). A major limitation of these assignment techniques is the requirement that all potential source populations or habitats must be sampled. The natal signature of an unsampled source is by definition unknown, so individuals originating from those areas will be necessarily misclassified. It is possible to use MLE techniques (but not DFA) to identify individuals that are unlikely to have originated in any of the sampled natal sources (Standish et al. 2008), but even in this case, no additional inference can be made about the identity of those alternative, unsampled sources. These techniques are thus vulnerable to error whenever recruits from distant, unknown, or simply unsampled source populations are present. In certain systems, such as estuarine-dependent species where spawning locations are discrete and well characterized (Thorrold et al. 2001), this assumption may not be limiting. However, for other species, such as open-coast spawners where much less is known about specific spawning locations, it may be difficult to know whether all natal sources were characterized sufficiently. In some cases, sampling all natal sources is simply not feasible (Gillanders and Kingsford 1996, Warner et al. 2005).

The task of assigning fish to stocks or natal sources is a special case of the statistical problem of mixture models, in which the goal is to identify the number of unique groups contributing to a mixed sample and to classify individuals into groups (Titterington et al. 1985). A popular approach to this type of problem is to use iterative MCMC algorithms (Gilks et al. 1996). MCMC techniques have been developed extensively for use in population genetics, where investigators also face a mixture-model problem: the presence of cryptic population structure (Pritchard et al. 2000, Pella and Masuda 2001). There is great potential for the MCMC approach in an otolith geochemistry context, primarily because MCMC can complement existing techniques (e.g., DFA) by simultaneously estimating the number of unique natal sources contributing to a sample of unknowns and assigning individuals to each source without any prior information on the number or identity of the sources. We first describe the basics of the MCMC approach and then demonstrate its use on simulated mixed-stock data sets and published data sets of weakfish (*Cynoscion regalis*) otolith elemental signatures from the U.S. Atlantic coast.

Methods

Natal source assignment as a mixture model problem

For the purposes of assignment using otolith geochemistry, a natal source is not strictly equivalent to a source population in the traditional ecological sense and carries no assumptions about demographic closure or genetic isolation. Rather, a natal source is a geographic locality with a distinctive geochemical signature such that the concentration of a given element or isotope in the otolith of a fish from that source can be considered a random draw from a multivariate normal distribution of concentrations.

To understand the MCMC approach to the mixture problem, consider the classic statistical analogy of urns containing balls of multiple colors in different proportions. If the proportion of each type of ball in each urn is known (the parameters), it is easy to predict the expected composition of a sample of balls drawn from several urns (the data). MCMC performs the reverse operation: given the sample of balls, it obtains estimates of the composition of each urn. This is done by applying Bayes' theorem, which describes the probability of the parameters given the data. In statistical notation, this is $\Pr(\text{parameters} \mid \text{data})$. This method is computationally intensive and requires the careful choice of appropriate probability distributions and resampling algorithms, notably the Gibbs sampler and the Metropolis-Hastings sampler. We outline the basics of these methods in the following paragraphs; more detail can be found in the Appendix and in technical reviews of MCMC techniques by Gilks et al. (1996), Robert and Casella (2004), and Jasra et al. (2005).

Describing the mechanics of MCMC is necessarily a notation-intensive exercise, so Table 1 summarizes the symbols used throughout the paper. Consider a sample of n individuals (i.e., recently settled larvae) drawn from κ natal sources. These might be recently settled recruits at a particular coastal location, and the primary question is which recruit originated in which source. Each individual i has a concentration x_l^i for variable l

TABLE 1. List of symbols used in the paper.

Symbol	Type	Sub-element (if applicable)	Definition
Parameter			
b	scalar		no. burn-in iterations
c	scalar		thinning interval
$\mathbf{\Phi}$	vector	ϕ_k	source mixture proportions
κ	scalar		actual no. sources in mixture
K	scalar		total no. clusters in a mixture
L	scalar		total no. elements or isotopes sampled
M	scalar		total no. Markov chain iterations
$\boldsymbol{\mu}$	matrix	$\mu_{k,l}$	actual source means
$\mathbf{P}$	matrix	$P_{k,l}$	source sample means
$\mathbf{Q}$	matrix		source assignment probabilities
$\mathbf{S}$	array	$\mathbf{S}_k$	covariance matrices
$\mathbf{X}$	matrix	$\mathbf{x}^i$	observations (data)
$\mathbf{Z}$	vector	z^i	source assignments
Parameter indices			
i			individual observation
j			individual observation
k			source
l			element or isotope
m			Markov chain step
n			no. observations in a sample

(these variables might be elemental concentrations or isotope ratios); a total of L variables are sampled, giving each individual a "signature" $\mathbf{x}^i$ (a vector of length L). Throughout this paper, subscripted indices refer to the natal source, k, or geochemical variable, l, with which a parameter is associated, while superscripted indices designate particular observations, either an individual sampling unit, i, or, if in parentheses, a step in a Markov chain, (m).

Given the $n \times L$ matrix of signatures $\mathbf{X}$ (the data from the sample of recruits), the goal is to estimate (1) the number of natal sources, κ, contributing to the sample, (2) the relative proportion of individuals from each natal source in the mixture, $\mathbf{\Phi}$, (3) the vectors of signature means, $\mathbf{P}$, and (4) covariance matrices, $\mathbf{S}$, corresponding to each of the κ sources, and (most importantly) (5) the source assignments, $\mathbf{Z}$, an $n \times 1$ vector containing each individual's natal source assignment. This problem is twofold. If the number of sources in the mixture, κ, is known, it is straightforward to estimate source parameters and source assignments using MCMC methods. Estimating κ must be done separately and is most easily treated as a model selection problem.

Estimating natal source parameters and assignments using MCMC

Given the data, $\mathbf{X}$, and assuming, for the moment, a particular number of natal sources, K, the values of $\mathbf{Z}$, $\mathbf{P}$, $\mathbf{S}$, and $\mathbf{\Phi}$ could be estimated using Bayes' theorem. Bayes' theorem gives the probability of a parameter θ conditional on the data, D. The probability of a particular value of θ given D, $\pi(\theta \mid D)$ (called the posterior), will be equal to the likelihood of the data given that parameter value, $f(D \mid \theta)$, multiplied by the prior probability of the parameter taking that value, and scaled by the likelihood integrated over all possible values of θ, $\int \pi(D \mid \theta)\, d\theta$ (Hilborn and Mangel 1997, Clark 2007). For the set of parameters in a mixture model, Bayes' theorem can be written as

$$\pi(\mathbf{Z}, \mathbf{P}, \mathbf{S}, \Phi|\mathbf{X}) = \frac{f(\mathbf{X}|\mathbf{Z}, \mathbf{P}, \mathbf{S}, \Phi) f(\mathbf{Z}) f(\mathbf{P}) f(\mathbf{S}) f(\Phi)}{\int f(\mathbf{X}|\mathbf{Z}, \mathbf{P}, \mathbf{S}, \Phi) f(\mathbf{Z}) f(\mathbf{P}) f(\mathbf{S}) f(\Phi)\, d\mathbf{Z}\, d\mathbf{P}\, d\mathbf{S}\, d\Phi} \quad (1)$$

where f is a generic probability density function and π is the posterior probability density to be estimated. The mode of the multivariate distribution π provides an estimate for each parameter, and the shape of π indicates the variance around those estimates. It is straightforward to write expressions for the likelihood $f(\mathbf{X} \mid \mathbf{Z}, \mathbf{P}, \mathbf{S}, \mathbf{\Phi})$ and priors $f(\mathbf{Z})$, $f(\mathbf{P})$, $f(\mathbf{S})$, and $f(\mathbf{\Phi})$ in the numerator of Eq. 1, but calculating the denominator requires integrating over a high-dimensional parameter space, which is daunting. MCMC methods avoid this difficulty by generating a sequence (a Markov chain) of parameter values for $\mathbf{Z}$, $\mathbf{P}$, $\mathbf{S}$, and $\mathbf{\Phi}$ that are approximate samples from the posterior distribution $\pi(\mathbf{Z}, \mathbf{P}, \mathbf{S}, \mathbf{\Phi} \mid \mathbf{X})$. In this sequence, each step depends only on the value of the previous step, so it is a Markov chain. By generating a long chain of samples, it is possible to approximate the posterior distribution and use it to estimate the correct parameter values (a Monte Carlo technique). This approach is easiest to visualize with univariate data (Fig. 1). The methods we describe here largely follow standard MCMC practices that one could implement in WinBUGS (Spiegelhalter et al. 2004) or R (*available online*),[5] although we also describe several nonstandard MCMC steps that might be more appropriate for otolith

[5] ⟨http://www.r-project.org⟩

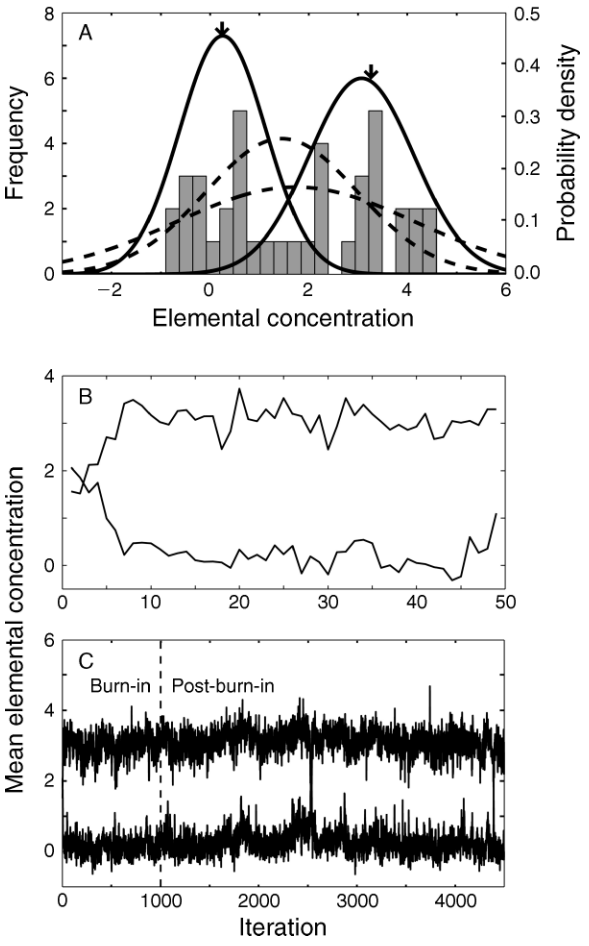

FIG. 1. Application of Markov Chain Monte Carlo (MCMC) methods to a univariate mixture model problem. Histogram (A) shows data sampled from two populations. Black arrows indicate true population means; the dashed lines indicate the posterior distribution estimated by the mean and standard deviation obtained at the first MCMC iteration; solid lines indicate the posterior distributions derived from full MCMC procedure. The progression of the Markov chain is shown for (B) the initial portion of the chain and (C) the entire chain. The initial "burn-in" portion of the chain shown in panel (C) is discarded, and the remainder forms the posterior distributions (solid lines) depicted in panel (A).

geochemical applications but which require independent programming. Consequently, we programmed in Matlab and provide our code as Supplementary Material.

Perhaps the most efficient MCMC technique for sampling from the posterior distribution $\pi(\mathbf{Z}, \mathbf{P}, \mathbf{S}, \mathbf{\Phi} \mid \mathbf{X})$ is the Gibbs sampler, which obtains a new value for a parameter by taking random samples from the probability distribution of that parameter conditional on the other parameters and the data. This requires the full conditional distributions for each parameter (e.g., $f[\mathbf{P} \mid \mathbf{X}, \mathbf{Z}, \mathbf{S}, \mathbf{\Phi}]$) but initial values only for $\mathbf{Z}$. Furthermore, a well-mixing Gibbs sampler will quickly move away from the initial state, so with an adequate burn-in (the discarded initial portion of the Markov chain, prior to convergence on π; Fig. 1B, C) the final result will be insensitive to the values chosen for $\mathbf{Z}^{(0)}$. With K sources and no other prior information, it is reasonable to assume a uniform distribution for $\mathbf{Z}^{(0)}$; that is, each individual in the sample has an equal probability of originating in any of the sources:

$$\Pr(z^i = k) = 1/K \tag{2}$$

where z^i is the ith element of $\mathbf{Z}$, i.e., the population assignment of individual i.

After individuals are initially assigned to sources using Eq. 2, the Gibbs sampler generates a Markov chain of parameter values ($\mathbf{S}^{(1)}$, $\mathbf{P}^{(1)}$, $\mathbf{Z}^{(1)}$, $\mathbf{\Phi}^{(1)}$), ($\mathbf{S}^{(2)}$, $\mathbf{P}^{(2)}$, $\mathbf{Z}^{(2)}$, $\mathbf{\Phi}^{(2)}$), an so on, by iterating the following four steps:

Step 1: Sample $\mathbf{S}^{(m)}$ from $f(\mathbf{S} \mid \mathbf{X}, \mathbf{Z}^{(m-1)})$
Step 2: Sample $\mathbf{P}^{(m)}$ from $f(\mathbf{P} \mid \mathbf{X}, \mathbf{S}^{(m)}, \mathbf{Z}^{(m-1)})$
Step 3: Sample $\mathbf{Z}^{(m)}$ from $f(\mathbf{Z} \mid \mathbf{X}, \mathbf{P}^{(m)}, \mathbf{S}^{(m)}, \mathbf{\Phi}^{(m-1)})$
Step 4: Sample $\mathbf{\Phi}^{(m)}$ from $f(\mathbf{Z} \mid \mathbf{X}, \mathbf{P}^{(m)}, \mathbf{S}^{(m)}, {}^{(m)})$.

Here (m) indicates the current step in the Markov chain, and f indicates a generic conditional distribution. For example, the Gibbs sampler will begin by drawing at random a value for $\mathbf{S}^{(1)}$ from the distribution of $\mathbf{S}$ conditional on $\mathbf{X}$ and the initial value of $\mathbf{Z}$, $\mathbf{Z}^{(0)}$. As mentioned above, the Gibbs sampler only requires an initial value for the parameter $\mathbf{Z}$. The initial values of $\mathbf{S}$ and $\mathbf{P}$ are conditional on $\mathbf{Z}^{(0)}$, and one generally places relatively flat, noninformative priors on these conditional distributions. If actual prior information is known about any of the parameters, this could be incorporated into the MCMC framework, but we assume that such information does not exist in the examples given here.

We provide a formal description of the various distributions needed for the Gibbs sampler in the Appendix. Briefly, the likelihood of an individual belonging to natal source i is multivariate normal, given means $\mathbf{P}_i$, covariance $\mathbf{S}_i$, and the proportion ϕ_i of the sample drawn from source i. Assignments, $\mathbf{Z}$, are then drawn (Step 3) from a multinomial distribution defined by those assignment likelihoods (normalized so that they sum to unity). Geochemical signature means, $\mathbf{P}$, are randomly generated (Step 2) from a multivariate normal distribution defined by the data, $\mathbf{X}$, (assigned to sources according to $\mathbf{Z}$) and covariances, $\mathbf{S}$. The use of these distributions require that the data be multivariate normal or be transformed to approximate normality; alternatively a different, more appropriate distribution could be utilized. In practice, we have found that most otolith geochemical data are either normal or lognormal, and that the MCMC techniques we use here are rather robust to deviations from normality.

Ideally, $\mathbf{\Phi}$ could be generated (Step 4) using the Gibbs sampler by using a Dirichlet distribution, which is the multivariate generalization of the beta distribution and the standard prior for use with multinomial distributions like a mixture of stocks (cf. Pritchard et al. 2000). However, we have found that such algorithms frequently encounter "trapping states" early in the Markov chain

in which $\phi_k = 0$ for all but one population. In such cases the Gibbs sampler ceases to mix and produces the uninteresting and often incorrect result that $\mathbf{X}$ is drawn from a single source. An alternative approach is to use a Metropolis-Hastings sampler to estimate $\mathbf{\Phi}$ independently, an approach that we outline in the Appendix. A Metropolis-Hastings sampler works by generating a candidate value for a parameter, then using a probabilistic rule (based on the likelihood of the data given that parameter value) to determine whether or not the candidate value is chosen as the next step in the Markov chain. Like the Gibbs sampler it produces draws from the desired posterior distribution, but it is slower and less efficient than Gibbs and is used when full conditional distributions are unavailable or unwieldy.

The proper assumptions to make regarding the covariances, $\mathbf{S}$, (Step 1) are an area of some debate. Generating random samples of covariance matrices is complicated by the requirement that they be positive definite (Zhang et al. 2004). The criteria defining positive definiteness require an understanding of matrix algebra, but this constraint is analogous to the requirement that a univariate variance be a positive real number (Horn and Johnson 1985). Generating acceptable matrices becomes easier if one assumes that the covariances of each source population are equal (e.g., Pella and Masuda 2005) or have some structural features in common (e.g., a common dominant eigenvector; Zhang et al. 2004). However, it may be desirable to assume that different sources have completely different covariance matrices, so we describe a technique in the Appendix that permits this assumption.

The sampler repeats Steps 1–4 M times. The initial values in the chain tend to explore parameter space and are discarded as the "burn-in" (Fig. 1B), but the chain eventually converges on a stable distribution (Fig. 1C). For sufficiently large burn-in, b, and thinning interval, c, the values ($\mathbf{S}^{(b)}$, $\mathbf{P}^{(b)}$, $\mathbf{Z}^{(b)}$, $\mathbf{\Phi}^{(b)}$), ($\mathbf{S}^{(b+c)}$, $\mathbf{P}^{(b+c)}$, $\mathbf{Z}^{(b+c)}$, $\mathbf{\Phi}^{(b+c)}$), ($\mathbf{S}^{(b+2c)}$, $\mathbf{P}^{(b+2c)}$, $\mathbf{Z}^{(b+2c)}$, $\mathbf{\Phi}^{(b+2c)}$), and so on, will be independent samples from the stable distribution of ($\mathbf{P}$, $\mathbf{S}$, $\mathbf{Z}$, $\mathbf{\Phi}$) and the expected values of that distribution can be estimated as the mean of those independent samples (Fig. 1A). From the stable distribution of $\mathbf{P}$ one can calculate the mean element concentrations or isotope ratios for each population; from the distribution of $\mathbf{Z}$ one can generate an $n \times K$ matrix, $\mathbf{Q}$, containing the probability of each individual being assigned to each source.

We should note that MCMC is an asymptotic technique: the Markov chain is only certain to describe the posterior distribution π if both it and the burn-in are of nearly infinite length. In other words, there is always concern that the Markov chain has not actually converged on the intended distribution. In recent years, there has been discussion of "perfect" MCMC samplers, which avoid this difficulty (Casella et al. 2001), but this technique is computationally infeasible (at present) for data sets of the size and dimensionality common in otolith geochemistry. A less elegant but more practical (and more widely used) approach is twofold: (1) monitor the output of a chain to determine when it convergences on a stable distribution of values for each parameter and (2) run multiple, independent chains with different initial values to confirm that they converge on the same posterior distribution. For the data sets used here, we found that burn-in of 5000 iterations followed by an additional 10 000 iterations was generally sufficient to attain convergence (we also found that a thinning interval of $c = 1$ was adequate). Running longer chains is usually desirable, but this dramatically increases the computational time required for the relabeling step (described in the Appendix) so, for our examples, we used the shortest possible chains.

As with any Markov chain of this type, all permutations of $\mathbf{Z}$ have equal likelihood; that is, the labeling of the mixture components (i.e., the sources) is arbitrary. For example, in a two-element mixture of individuals from sources A and B, individuals from A will tend to be assigned to the same group (i.e., share the same value of z) but the identity of this group will change as the Markov chain progresses; at one point individuals in A will have $z = 1$ and individuals from B will have $z = 2$, but if at any point enough individuals from A happen to be assigned $z = 2$ (recall that each iteration of the Gibbs sampler involves a random draw from the conditional distribution of z), a label switch might occur, such that individuals from A begin to be assigned $z = 2$, and those from B are assigned $z = 1$. This so-called label-switching problem makes it impossible to estimate source means and individual assignments from the raw MCMC output, because over the course of the Markov chain, each individual may have been assigned every possible value of z, even if it is associated with a well-defined cluster of observations. To counter this problem, Stephens (2000) proposed a post-hoc relabeling algorithm (see Appendix for details) which we use in all examples presented here.

Estimating κ

The accuracy of the output from the Gibbs sampler described above is contingent upon the number of clusters, K, being equal to the actual number of sources, κ, contributing to the sample. Of course, κ is rarely known with certainty; if it were, we could sample each of the κ sources to obtain a training data set and then use DFA or MLE techniques to assign individuals of unknown origin. The MCMC approach is more useful in cases of unknown κ, and the problem of determining κ is usually solved in one of two ways. Richardson and Green (1997) used reversible-jump MCMC that permitted sources to combine (reducing K) or split (increasing K) in the middle of the Markov chain. This method is problematic in the multivariate normal case because of the difficulty in generating new covariance matrices that are positive definite, requiring the eigenvectors of $\mathbf{S}$ to be kept constant among populations (Zhang et al. 2004).

TABLE 2. Results of Markov Chain Monte Carlo (MCMC) analysis of simulated "unknown" data sets.

Data set	κ	n_i	Δμ	r	ΔDIC for $K = 1$				
					1	2	3	4	5
Zero covariance, two elements	2	20	1	0	0	29†	40	51	61
	2	20	2	0	0	15†	21	31	19
	2	20	3	0	16	0†	9	25	31
Zero covariance, four elements	2	20	1	0	0	11†	14	23	16
	2	20	2	0	17	14†	0	11	24
	2	20	3	0	59	0†	8	18	26
Nonzero covariance, two elements	2	20	3	0.25	12	0†	1	8	21
	2	20	3	0.75	3	0†	5	6	16
Three populations	3	15, 15, 15	3	0	76	10	0†	7	15
	3	20, 20,5	3	0	41	0	12†	25	30
	3	35, 5, 5	3	0	79	22	0†	40	43

Notes: Each data set consisted of n individuals drawn from κ populations (each contributing n_i individuals) with a multivariate normal distribution of elemental signature means (**μ**) with variance 1 and correlation r. The difference between population means for each variable is given by Δμ. MCMC analysis was used to assign individuals to clusters assuming κ was unknown; MCMC runs were performed using $K = 1, 2, \ldots 5$. The ΔDIC model selection criterion is given for each value of K; zeros indicate the number of clusters identified by deviance information criterion (DIC) as the best model, while daggers (†) indicate the model corresponding to the correct number of sources ($K = \kappa$). The ΔDIC model selection criterion is the difference between the DIC for MCMC algorithm runs for each value of K and the lowest observed DIC. The most parsimonious model has ΔDIC = 0.

The use of a mixture parameter (**Φ** in the algorithm we have described) is somewhat similar to reversible-jump MCMC in that sources are allowed to have a zero probability of contributing to the mixture, effectively reducing K. Thus some inference on κ can be made from the posterior distribution of **Φ**. An alternative strategy is to take a model selection approach for comparing MCMC output using different values of K. The deviance information criterion (DIC; Spiegelhalter et al. 2002, Celeux et al. 2006), which is simple to calculate from MCMC output, is emerging as the MCMC equivalent of the Akaike Information Criterion (AIC; Burnham and Anderson 1998), and produces values that can be interpreted in a similar manner. Like AIC, DIC penalizes the adequacy of a model (how well it fits the data) by the number of parameters. In this case, adequacy is measured by Bayesian deviance, D, which is −2 times the log-likelihood: $D = -2\log \Pr(\text{data} \mid \text{parameters})$. DIC is calculated as the mean deviance at each step in the Markov chain, $\overline{D(\mathbf{Z}, \mathbf{P}, \mathbf{S})}$, minus the effective number of parameters in the model, P_d. This latter value is estimated as the difference between the mean deviance and the deviance of the mean values of each parameter: $P_d = \overline{D(\mathbf{Z}, \mathbf{P}, \mathbf{S})} - D(\bar{\mathbf{Z}}, \bar{\mathbf{P}}, \bar{\mathbf{S}})$. The value of K associated with the lowest DIC is preferred (Spiegelhalter et al. 2002).

There is some controversy regarding the use of DIC in mixture models (see discussions accompanying Spiegelhalter et al. 2002, Celeux et al. 2006), and WinBUGS will not allow it (Spiegelhalter et al. 2004). The same problem of symmetrical, nonidentifiable modes that necessitates the relabeling algorithm (Stephens 2000; also see Appendix) also tends to cause P_d to take on illegal, negative values (Celeux et al. 2006). If a relabeling algorithm is not used, the mean parameter values $\bar{\mathbf{Z}}$, $\bar{\mathbf{P}}$, and $\bar{\mathbf{S}}$ are not meaningful because they represent means taken across multiple modes of the mixture. This causes $D(\bar{\mathbf{Z}}, \bar{\mathbf{P}}, \bar{\mathbf{S}})$ to be larger than it should be, producing improperly small or negative values of P_d. Celeux et al. (2006) have proposed a number of alternative formulations for DIC in an attempt to resolve this issue. By applying a relabeling algorithm one can recover appropriate values of $\bar{\mathbf{Z}}$, $\bar{\mathbf{P}}$, and $\bar{\mathbf{S}}$, and for the simulations presented here we found that the original DIC formulation performed well and produced sensible values of P_d, provided the relabeling algorithm had converged to a solution. However, our experience with other data sets suggests that the alternative DIC_3 metric developed by Celeux et al. (2006) is an excellent choice if the original DIC formula yields consistently negative P_d values even after successful relabeling.

EMPIRICAL EXAMPLES

To demonstrate the strengths and limitations of the MCMC approach, we applied it to simulated data generated from known distributions and to actual otolith data taken from fish of known origin. In both cases, the actual natal source assignment of each individual was known, so the method can be judged by its ability to pick the correct number of populations and to assign individuals correctly. Because the value of K in a particular simulation will not always equal the actual number of source contributing to the mixture, it becomes convenient to refer to the latter as "sources" and to use the term "cluster" to refer to the groups identified by MCMC.

Simulated data

We examined the ability of an MCMC algorithm to correctly assign natal origin, estimate source parameters, and select the correct K using 11 simulated data sets spanning a range of mixture scenarios (Table 2). Each individual in a data set (a "recruit") represented an independent draw from a multivariate normal distribution corresponding to the signature of one of κ natal

TABLE 3. Comparison between MCMC and traditional multivariate statistical methods for analysis of the data sets described in Table 2.

Data set	κ	n_i	Δμ	r	Estimated K		Assignment accuracy			
					MCMC (best DIC)	k-means	MCMC (best DIC)	MCMC ($K = \kappa$)	k-means	DFA†
Zero covariance, two elements	2	20	1	0	1	2	0.50	0.60	0.83	0.75
	2	20	2	0	1	2	0.50	0.80	0.90	0.90
	2	20	3	0	2	2	1.00	1.00‡	1.00	1.00
Zero covariance, four elements	2	20	1	0	1	2	0.50	0.50	0.75	0.73
	2	20	2	0	3	2	0.88	0.50	0.93	0.93
	2	20	3	0	2	2	1.00	1.00‡	1.00	1.00
Nonzero covariance, two elements	2	20	3	0.25	2	2	1.00	1.00‡	0.90	0.95
	2	20	3	0.75	2	2	1.00	1.00‡	1.00	1.00
Three populations	3	15, 15, 15	3	0	3	2	1.00	1.00‡	0.67	1.00
	3	20, 20,5	3	0	2	2	0.89	0.96	0.50	0.98
	3	35, 5, 5	3	0	3	2	1.00	1.00‡	0.88	1.00

Notes: Each data set consisted of n individuals drawn from κ populations (each contributing n_i individuals) with a multivariate normal distribution of elemental signature means ($\boldsymbol{\mu}$) with variance 1 and correlation r. The difference between population means for each variable is given by Δμ. We compared MCMC and k-means clustering based on their estimates of K ($K = \kappa$ is the correct result) and compared MCMC to discriminant function analysis (DFA) and k-means clustering based on the accuracy with which individuals were assigned to sources. The assignment accuracy for MCMC is given for the value of K selected by DIC and the correct model for which $K = \kappa$. For DFA, we report the jackknife reclassification success assuming κ is known; for k-means clustering we give the classification success for the value of K identified by the Schwarz criterion.

† All reclassifications were significantly better than random at $\alpha = 0.05$ (White and Ruttenberg 2007).

‡ The model with $K = \kappa$ was also identified as the best model by DIC.

sources. We measured assignment accuracy by comparing cluster assignments to actual source identity (the actual MCMC cluster assignments [$z = 1$, $z = 2$, and so on] were arbitrary, but it was straightforward to match clusters to sources based on the similarity of signature means) . We considered individuals to be assigned to a cluster if the assignment probability for that cluster was greater than the arbitrary threshold of 0.5.

The results from the first group of data sets in Table 2, with two sources and increasing difference between population means, show that source means must be separated by some minimum difference in order to be resolved into distinct clusters by MCMC. The second group of data sets show that this clustering can be improved by including additional independent variables; for example, increasing the number of elements from two to four improved assignment accuracy from 50% to 88% when source means differed by only two standard deviations. This group of data sets also illustrates a quirk of this procedure: DIC sometimes selected models with too many clusters ($K > \kappa$), but the additional clusters generally had few or no individuals assigned to them. The presence of such empty clusters appears to improve the mixing of the Markov chain. The third group of data sets shows that strong covariance among the variables does not greatly impede the efficacy of the technique, and the fourth group of data sets in Table 2 illustrates the performance of MCMC clustering with multiple sources mixed in unequal proportions. The results were mixed: MCMC was able to resolve and accurately classify individuals from three evenly mixed sources (15:15:15 individuals) or from two minor contributors mixed with a single high-contribution source (35:5:5 individuals), but in the case of one small group mixed with two larger groups (20:20:5), individuals from the smaller group were misclassified as belonging to one of the larger groups.

For comparison, we also applied two traditional statistical techniques to each data set. First, we calculated the jackknife reclassification success statistics for a DFA in which the number and parameters of the actual sources were known, mimicking the best-case scenario of classifying post-dispersal juveniles after exhaustive baseline sampling of pre-dispersal individuals. In all cases, DFA performed the same or better than MCMC at assigning individuals to sources (Table 3). The disparity was especially great in cases where there was little separation between the source means and MCMC was unable to resolve the correct number of clusters. We also applied a k-means cluster analysis to each data set. This technique can be used in an exploratory fashion to discover how well a mixture can be partitioned into a variable number of clusters (Steinley 2006). Following the standard technique, we selected the most parsimonious number of clusters using the Schwarz criterion (an information criterion conceptually similar to AIC and DIC; Pelleg and Moore 2000). In all of the two-source scenarios, the k-means analysis identified the correct number of clusters and had an assignment accuracy that matched or exceeded that of MCMC. As with DFA, the k-means analysis tended to outperform MCMC when there was little difference among the sources in multivariate space. In part this may be because k-means analysis cannot run a $K = 1$ scenario; it is constrained to always find at least two clusters. However, in all of the three-source scenarios, including one with evenly mixed sources, the k-means analysis consistently identified only two clusters, resulting in much lower assignment accuracy than MCMC. Thus in some cases MCMC appears to be superior to the

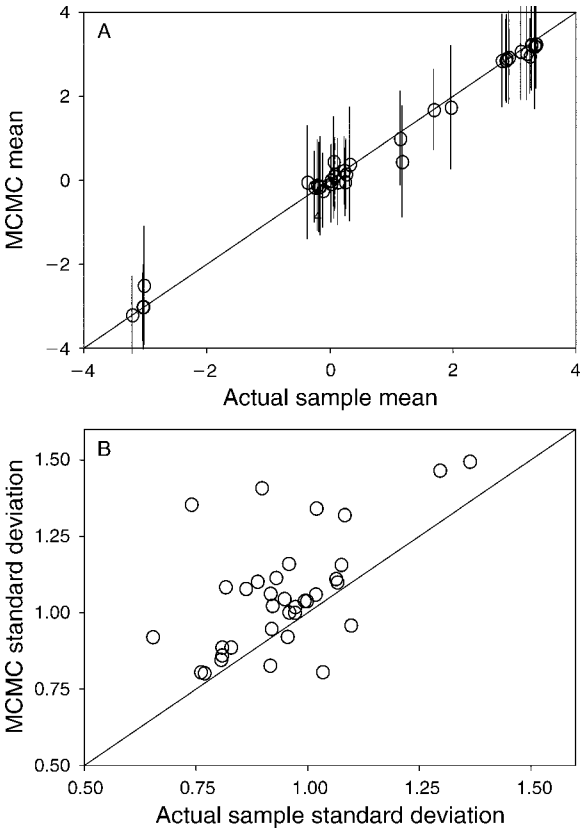

FIG. 2. Comparison of MCMC-estimated population parameters and (A) actual sample means ± SD and (B) standard deviations for all multivariate normal simulated data sets in Table 2. MCMC estimates are taken from the DIC-selected (deviance information criterion) best model.

k-means approach for identifying the correct number of clusters, especially when there are multiple small mixture components. We also note that *k*-means clustering generally assumes that the various clusters have equal covariance matrices (McGarigal et al. 2000), which is not a constraint for the MCMC approach.

The accuracy of MCMC estimation of source parameters can be examined by comparing the actual sample means and variances for each population in the simulated data sets to the MCMC estimates of those parameters for the corresponding clusters in the DIC-selected model. The 95% confidence intervals (CI) for the MCMC-estimated mean overlapped the value of the sample mean in all cases (Fig. 2A). However, the MCMC algorithm consistently overestimated sample variance (Fig. 2B). This discrepancy likely resulted from the occasional misassignment of individuals during the MCMC iterations that inflated the estimates of sample variance used to simulate the covariance matrices at each step.

Weakfish data

To illustrate the merit of the MCMC approach using real otolith data, we chose a data set for which we already possessed reliable estimates of source assignments; that is, the "unknown" data were not completely unknown. The weakfish, *Cynoscion regalis*, presents an ideal case study. Weakfish spawn in estuaries and coastal embayments along the east coast of North America each spring; after remaining in natal estuaries for several months, young-of-the-year (YOY) juveniles join the adult population in the annual autumnal migration to southern overwintering grounds. Thorrold et al. (1998) showed that natal estuaries produce distinctive geochemical signatures in the otoliths of pre-dispersal YOY juvenile weakfish; based on this signature, juveniles could be assigned to their natal estuary with an average accuracy of 93% using an artificial neural network (ANN) method. A later study (Thorrold et al. 2001) used signatures in the natal region of otoliths collected from spawning adults to assign adults to their natal estuary. The authors concluded that weakfish had a relatively high degree (60–81%) of natal homing.

Weakfish have proven to be a useful study species because YOY juveniles are easily collected prior to dispersal, permitting the characterization of natal signatures. In addition, because collection of juveniles and adults span most of the geographical range of the species, the contribution of unknown sources is minimized. We reanalyzed the same juvenile and adult data sets using MCMC as if these advantages were not present: we assumed that the juvenile collection sites were unknown, and we did not use juvenile data as a training data set for assignment of adults.

Two hundred sixty juvenile and 414 adult weakfish were collected at five estuaries: Peconic Bay, New York (NY), Delaware Bay (DE), Chesapeake Bay (CB), Pamlico Sound, North Carolina (PS), and coastal Georgia (GA). Each otolith had a geochemical signature consisting of six variables (see Thorrold et al. 1998, 2001 for details); we did not transform the data prior to analysis.

We performed MCMC analysis independently for the juvenile and adult weakfish data with $K = 1$–6. For the juvenile data, DIC selected a best-fit model with $K = 4$. For this model, all but six individuals were assigned to a cluster with >80% probability, and the clusters largely matched the actual geography of the natal sites (Fig. 3, 4A).

For the adult data, DIC selected $K = 6$. One of these clusters had an <22% assignment probability for all fish and a second had only 10 fish assigned to it; we restrict our discussion to the remaining four clusters. In analyzing this data set, Thorrold et al. (2001) concentrated on calculating the proportion of adults collected in each estuary that were assigned to that same estuary as their natal site (this was their estimate for site fidelity). We performed a similar analysis to make our results comparable: for each of the four "natal" clusters produced by MCMC, we determined the proportion of adults that had been collected in each of the estuarine

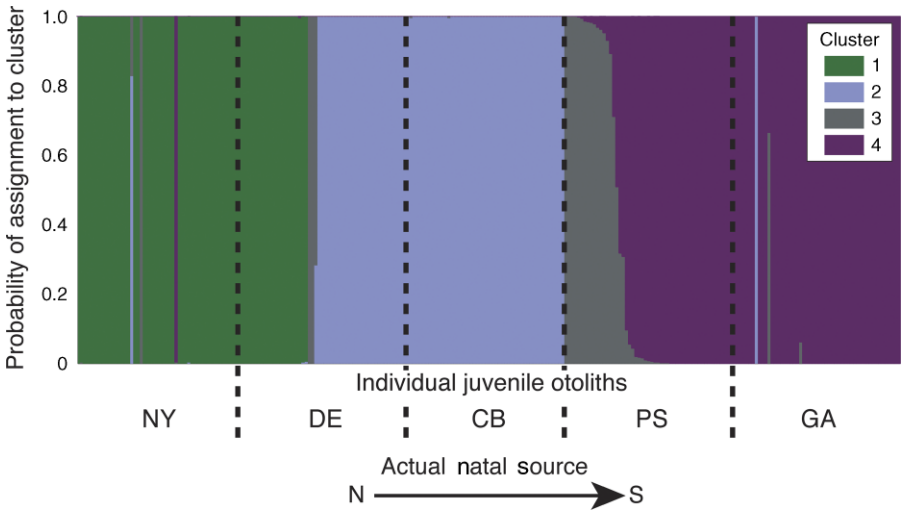

FIG. 3. Assignment probabilities for pre-dispersal juvenile weakfish otoliths. Individuals are grouped along the horizontal axis by the estuary in which they were spawned; estuaries are ordered from north to south as indicated by the arrow. Vertical bars indicate the probability of membership in each of the four clusters in the best-fit model, each indicated by a different color. Each vertical bar along the horizontal axis corresponds to one individual, and the total probability for each individual sums to unity. Estuaries are NY, Peconic Bay, New York; DE, Delaware Bay; CB, Chesapeake Bay; PS, Pamlico Sound, North Carolina; and GA, coastal Georgia.

spawning locations (Fig. 4B). Adults assigned to natal clusters 1 and 2 were predominantly collected in GA and PS, respectively. Natal cluster 4 consisted primarily of fish collected in NY and DE, while adults assigned to natal cluster 3 had been collected in PS, CB, and DE.

DISCUSSION

The examples we provided here illustrate the advantages of the MCMC approach: with no prior information, the algorithm correctly identified the number of sources contributing to mixed samples and assigned individuals to their source populations with a high degree of accuracy. This method is thus an excellent addition to the analytical toolbox of otolith geochemistry investigators. Even in the absence of reliable information on the number or identity of source populations, the number of chemically unique sources contributing to a sample of fish can be identified. The ability of MCMC methods to generate this sort of estimate does not involve any statistical sleight of hand: MCMC is simply a tool for evaluating the multidimensional integral in Bayes' theorem (Eq. 1; Lele et al. 2007).

The MCMC approach is similar in philosophy to that proposed years ago by Smouse et al. (1990) for identifying genetically distinct salmon stocks, but its applicability to multivariate normal data sets in marine systems has been noted only recently (Pella and Masuda 2005). We have described the use of this technique in an otolith geochemistry context, but it could be applied easily to other types of multivariate data used to assign individuals to stocks, such as otolith or scale morphology or profiles of fatty acid composition (Cadrin et al. 2005). While the weakfish case study addressed a case of spatial, not temporal, variability in natal signatures, MCMC could also be used with a sample consisting of multiple recruit cohorts. In such cases, the age data carried in the annuli of each otolith could be used to determine whether clusters resolve spatial or temporal variability (or both). It also may be possible to use MCMC to assign independent samples of fish taken at different times or locations to their respective stocks using aggregate life history parameters calculated for each sample, such as age and size distribution or von Bertalanffy growth parameters (Begg 2005). In such cases, each sample of multiple fish (rather than each individual organism) would be treated as an independent observation by MCMC. Finally, while we have repeatedly referred to otolith geochemistry for simplicity, the techniques outlined here apply equally well to geochemical data collected from mollusk shells, gastropod statoliths, and the hard (or perhaps even soft) parts of other invertebrates (Zacherl 2005, Becker et al. 2007, Carson et al. 2008). The key steps to successful MCMC implementation are determining how the data are distributed and then choosing appropriate conditional probability distributions regardless of the source of the data.

This method is not without limitations, which were highlighted by the results for several of the simulated data sets. First, population means must be sufficiently different for MCMC to resolve natal sources into separate clusters. This limitation is unavoidable and shared by any multivariate classification technique: two statistical populations must be far enough apart in multivariate space, with little overlap in their distributions, to be distinguishable. Second, parsimony-based DIC selection will sometimes select a model with fewer than the correct number of clusters if some sources make

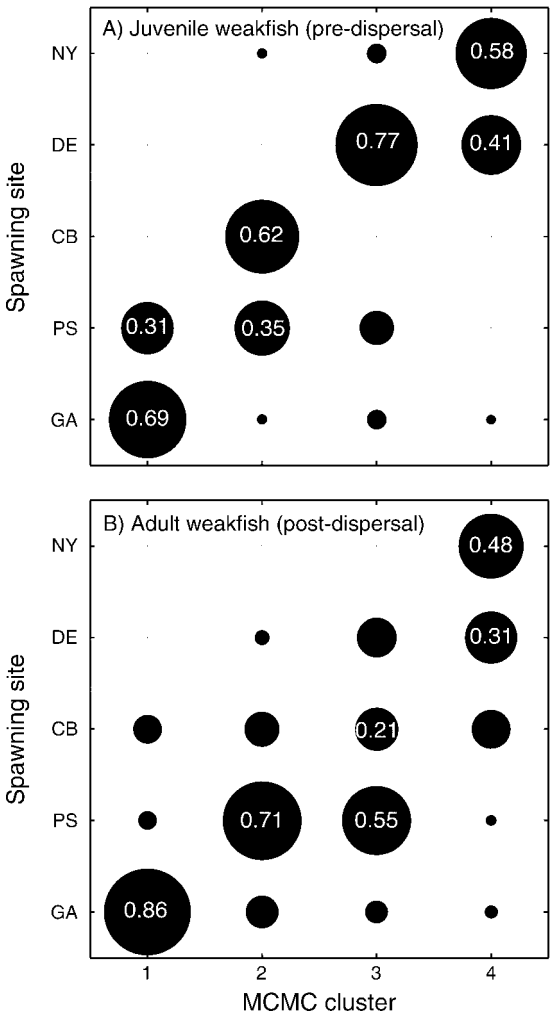

FIG. 4. Proportions of weakfish otoliths collected in each estuarine spawning location that were assigned to each of four clusters by MCMC. (A) Juveniles were collected pre-dispersal, so assignment success reflects accuracy of MCMC clustering. (B) Adults were collected post-dispersal, so assignment probabilities are an estimate of natal homing (cf. Fig. 3 in Thorrold et al. [2001]). Proportions are indicated by the area of the bubble and sum to 1 vertically; values > 0.20 are labeled. Estuaries are as identified in Fig. 3.

only very small contributions to the mixture. In a sense, this is a problem common to traditional forms of data analysis: does an outlying value represent measurement error, process error (i.e., extreme natural variation), or the presence of an additional source? This is a potentially serious problem, because identifying minor contributions from small or distant sources is essential to understanding gene flow (Palumbi 2004) and characterizing the tails of dispersal kernels (Kot et al. 1996). Unfortunately there is no ready statistical solution to this dilemma, but prior knowledge regarding oceanography (Gawarkiewicz et al. 2007) or the distribution of nursery habitats (Beck et al. 2001) could be used to guide inference. Furthermore, if one had access to actual source signatures one might be able to determine whether or not a questionable cluster corresponded to a real natal source. The problem of detecting sources with minor contributions will continue to plague investigators, but it is worth noting that MCMC offers better performance in identifying the correct number of low-contribution sources than the traditional k-means clustering method.

In some cases, DIC can select models with greater than the correct number of clusters, although this appears to be a less serious problem than selecting too few clusters because the "extra" clusters generally have few or no individuals assigned to them. This occurs because the mean Bayesian deviance is minimized when extreme observations are placed in separate clusters, permitting the covariance estimates for the higher-occupancy clusters to shrink. The creation of empty clusters may also be a byproduct of the relabeling procedure: as the initial Markov chain progresses, one cluster is always empty, but the identity of the empty cluster is constantly changing as the chain mixes and the support for each component of the mixture moves from cluster to cluster (this is the essence of the label-switching problem). The existence of an empty cluster may facilitate the mixing of the chain, resulting in lower deviance values and selection of that model. The relabeling algorithm then returns a Markov chain with a single empty cluster. This effect may explain the empty cluster observed in the best-fit model for the simulated data set with four variables and $\Delta\mu = 2$. However, there is the potential for improper inference regarding spurious low-membership clusters: these may lead to errors in estimating long-distance connectivity if the clusters are assumed to represent contributions from distant sources. Here again, ecological knowledge and baseline sampling of sources could be applied to determine whether a particular cluster corresponds to an actual source population or not. In any case, model selection with DIC seems to produce sensible results for the majority of individuals despite the occasional empty or low-membership cluster.

Fortunately, the MCMC approach also produces accurate estimates of source means which could be used to identify the actual natal sources corresponding to each cluster. Such links must be drawn with care, however. The clusters generated by MCMC may be well defined in multivariate space but not in geography. The pre-dispersal weakfish data provide an excellent illustration of this distinction: MCMC did identify a distinct cluster corresponding to Pamlico Sound, but some fish from that estuary were also assigned to the cluster containing fish from Georgia. Inspection of canonical variate plots of these data in the original publication (Thorrold et al. 1998) reveals that the PS and GA data are quite close in multivariate space, explaining the MCMC results. In our simulations, MCMC consistently overestimated sample variances, so that parameter may

be less useful than the mean in relating cluster assignments to actual sources.

MCMC analysis of the weakfish data produced results that were similar, but not identical, to those obtained with traditional methods. For the pre-dispersal juvenile fish, we found a strong geographical pattern of distinct natal signatures. The two geographically extreme estuaries (NY and GA) and the central estuary (CB) were distinct from each other, and fish from the two remaining estuaries were placed in clusters corresponding to their neighbors in multivariate (and geographic) space. For the adults, we found estimates of natal site fidelity similar to those of Thorrold et al. (2001) in the extreme populations (NY and GA). As with the juvenile data, fish collected in CB and DE were not distinguished from each other, and fish collected in PS were assigned to two clusters, although one of the clusters was almost completely restricted to PS fish. While these discrepancies may seem jarring, the advantage of the MCMC approach lies in its non-reliance on prior information. Weakfish spawn at discrete sites and early life stages remain at those sites for months. Imagine instead that we were only able to sample juveniles after they dispersed, when information about sources is unobtainable and MLE analysis is therefore not possible. With an MCMC approach, the sample of 260 juveniles would be resolved into four clusters corresponding to NY/DE, DE/CB, PS, and PS/GA. Similarly, if natal site information were unavailable, adult collections from each spawning site could still be analyzed with MCMC. Such analysis would reveal that adults in NY and GA originated from distinct and unique sources, a third source contributes exclusively to the PS adult population, and that fish from a fourth "source" spawn exclusively in the mid-Atlantic sites. These inferences are not as precise as those made using prior information on natal sources, but would produce similar general conclusions (e.g., weakfish have a high degree of natal homing, making their fishery vulnerable to overexploitation; cf. Thorrold et al. 2001). Of course, had MCMC been used in this way to characterize the adult weakfish population, the next step in the investigation would be an attempt to sample estuaries along the coast to determine which natal source corresponded to which cluster. Without that level of sampling, simply knowing the number of natal sources has limited utility.

Indeed, the primary conceptual difficulty involved in this technique is identifying the biological meaning corresponding to a "source" or "cluster." In an otolith geochemistry context, a statistical source necessarily describes individuals who are sampling the same water mass or source of elements and isotopes (Campana et al. 2000), i.e., it does not necessarily have a particular biological meaning. This statistical source may correspond to any spatial scale, from a single clutch of benthic eggs to fish occupying tens or hundreds of miles of coastline (Thorrold and Hare 2002). Thus simply knowing the number of geochemically unique natal sources may not be helpful without the proper ecological context. While successful application of the MCMC approach does not require exhaustive sampling of every potential natal source, it does require at least some baseline sampling to characterize the spatial scale of variation in natal signatures.

The results from the weakfish case study suggest how MCMC might be incorporated into the existing suite of analytical tools for otolith geochemical investigations. The strength of the approach is that it provides correct (albeit limited) inferences when direct sampling of all sources is infeasible. Thus it could be useful as an exploratory tool to determine the number, geochemical signatures, and relative contributions of natal sources to a sample of post-dispersal individuals. This information could then be used to guide the sampling of pre-dispersal individuals from potential source locations. The eventual goal would be to collect baseline samples of geochemical signatures from the array of natal sources. It would then become possible to apply traditional statistical tools such as DFA, MLE, ANN, or *k*-means clustering to future samples of post-dispersal individuals, because these tools appear to offer more robust and reliable assignment than MCMC (and are simpler to compute) when prior information about the number and identity of sources is available.

In some cases, the use of MCMC would not be limited to exploratory work. In many systems there is temporal variation in natal signatures (Warner et al. 2005, Standish et al. 2008) so a long-term monitoring program might involve repeated use of MCMC, especially if it is difficult to sample pre-dispersal individuals from all sources in each recruitment season. In this sort of system it would be useful to perform baseline sampling to characterize the typical spatial scale of variability in natal signatures, but once such initial sampling has been performed, it could possible to estimate the number of natal sources contributing to subsequent samples of post-dispersal juveniles without sampling all potential sources (and then repeatedly resampling them to account for temporal changes in source signatures; Warner et al. 2005). This would facilitate the identification of sites that consistently receive larval supply from multiple sources (or a single source). MCMC could also used as a check on other analytical methods for systems in which all potential sources are thought to be identified (as in the weakfish example; Thorrold et al. 2001). Unlike DFA and MLE methods, MCMC can identify additional, unsampled sources, and it appears to outperform *k*-means clustering at this task.

Regardless of the species or location of interest, the advantages afforded by MCMC are clear: with some effort, it can be used to identify and assign individuals to previously unsampled sources with a high degree of reliability. If used wisely, it stands to greatly enhance the analytical capabilities of marine and fisheries scientists.

Acknowledgments

Discussions with S. Gaines, J. HilleRisLambers, B. Kinlan, L. W. Botsford, and the Botsford lab group were helpful in developing this work, and the authors are grateful to R. Millar, S. Cadrin, and one anonymous reviewer for insightful comments on the manuscript. J. W. White was supported by NSF Predoctoral, UC Regents, and UCSB Affiliates fellowships. This is contribution number 294 from the Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), funded primarily by the Gordon and Betty Moore Foundation and the David and Lucile Packard Foundation.

Literature Cited

Almany, G. R., M. L. Berumen, S. R. Thorrold, S. Planes, and G. P. Jones. 2007. Local replenishment of coral reef fish populations in a marine reserve. Science 316:742–744.

Beck, M. W., et al. 2001. The identification, conservation, and management of estuarine and marine nurseries for fish and invertebrates. BioScience 51:633–641.

Becker, B. J., L. A. Levin, F. J. Fodrie, and P. A. McMillan. 2007. Complex larval connectivity patterns among marine invertebrate populations. Proceedings of the National Academy of Sciences (USA) 104:3267–3272.

Begg, G. A. 2005. Life history parameters. Pages 119–150 *in* S. X. Cadrin, K. D. Friedland, and J. R. Waldman, editors. Stock identification methods: applications in fishery science. Elsevier Academic Press, Amsterdam, The Netherlands.

Botsford, L. W., J. C. Castilla, and C. W. Peterson. 1997. The management of fisheries and marine ecosystems. Science 277: 509–515.

Botsford, L. W., F. Micheli, and A. Hastings. 2003. Principles for the design of marine reserves. Ecological Applications 13: S25–S31.

Brown, J. A. 2006. Using the chemical composition of otoliths to evaluate the nursery role of estuaries for English sole *Pleuronectes vetulus* populations. Marine Ecology Progress Series 306:269–281.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Cadrin, S. X., K. D. Friedland, and J. R. Waldman. 2005. Stock identification methods: applications in fishery science. Elsevier Academic Press, Amsterdam, The Netherlands.

Campana, S. E., G. A. Chouinard, J. M. Hanson, A. Fréchet, and J. Brattey. 2000. Otolith elemental fingerprints as biological tracers of fish stocks. Fisheries Research 46:343–357.

Campana, S. E., and S. R. Thorrold. 2001. Otoliths, increments, and elements: keys to a comprehensive understanding of fish populations? Canadian Journal of Fisheries and Aquatic Sciences 58:30–38.

Carr, M. H., and D. C. Reed. 1993. Conceptual issues relevant to marine harvest refuges: examples from temperate reef fishes. Canadian Journal of Fisheries and Aquatic Sciences 50:2019–2028.

Carson, H. S., S. G. Morgan, and P. G. Green. 2008. Fine-scale chemical fingerprinting of an open-coast crustacean for the assessment of population connectivity. Marine Biology 153: 327–335.

Casella, G., M. Lavine, and C. P. Robert. 2001. Explaining the perfect sampler. American Statistician 55:299–305.

Celeux, G., F. Forbes, C. P. Robert, and D. M. Titterington. 2006. Deviance information criteria for missing data models. Bayesian Analysis 1:651–674.

Clark, J. S. 2007. Models for ecological data: an introduction. Princeton University Press, Princeton, New Jersey, USA.

Crowder, L. B., S. J. Lyman, W. F. Figueira, and J. Priddy. 2000. Source-sink population dynamics and the problem of siting marine reserves. Bulletin of Marine Science 66:799–820.

DiBacco, C., and L. A. Levin. 2000. Development and application of elemental fingerprinting to track the dispersal of marine invertebrate larvae. Limnology and Oceanography 45:871–880.

Gawarkiewicz, G., S. Monismith, and J. Largier. 2007. Observing larval transport processes affecting population connectivity: progress and challenges. Oceanography 20:40–53.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov chain Monte Carlo in practice. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Gillanders, B. M., and M. J. Kingsford. 1996. Elements in otoliths may elucidate the contribution of estuarine recruitment to sustaining coastal reef populations of a temperate reef fish. Marine Ecology Progress Series 141:13–20.

Hilborn, R., and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, New Jersey, USA.

Horn, R. A., and C. R. Johnson. 1985. Matrix analysis. Cambridge University Press, Cambridge, UK.

Jackson, J. B. C., et al. 2001. Historical overfishing and the recent collapse of coastal ecosystems. Science 293:629–638.

Jasra, A., C. C. Holmes, and D. A. Stephens. 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science 20:50–67.

Jones, G. P., M. J. Milicich, M. J. Emslie, and C. Lunow. 1999. Self-recruitment in a coral reef fish population. Nature 402: 802–804.

Jones, G. P., S. Planes, and S. R. Thorrold. 2005. Coral reef fish larvae settle close to home. Current Biology 15:1314–1318.

Knutsen, H., P. E. Jorde, O. T. Albert, A. R. Hoelzel, and N. C. Stenseth. 2007. Population genetic structure in the North Atlantic Greenland halibut (*Reinhardtius hippoglossoides*): influenced by oceanic current systems? Canadian Journal of Fisheries and Aquatic Sciences 64:857–866.

Kot, M., M. A. Lewis, and P. van den Driessche. 1996. Dispersal data and the spread of invading organisms. Ecology 77:2027–2042.

Kritzer, J. P., and P. F. Sale. 2004. Metapopulation ecology in the sea: from Levins' model to marine ecology and fisheries science. Fish and Fisheries 5:131–140.

Lele, S. R., B. Dennis, and F. Lutscher. 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecology Letters 10:551–563.

Levin, L. A. 2006. Recent progress in understanding larval dispersal: new directions and digressions. Integrative and Comparative Biology 46:282–297.

Lubchenco, J., S. R. Palumbi, S. D. Gaines, and S. Andelman. 2003. Plugging a hole in the ocean: the emerging science of marine reserves. Ecological Applications 13:S3–S7.

McGarigal, K., S. Cushman, and S. Stafford. 2000. Multivariate statistics for wildlife and ecology research. Springer-Verlag, New York, New York, USA.

Mora, C., and P. F. Sale. 2002. Are coral reef fish populations open or closed? Trends in Ecology and Evolution 17:422–428.

Palsbøll, P. J., M. Bérubé, and F. W. Allendorf. 2006. Identification of management units using population genetic data. Trends in Ecology and Evolution 22:11–16.

Palumbi, S. R. 2004. Marine reserves and ocean neighborhoods: the spatial scale of marine populations and their management. Annual Review of Environment and Resources 29:31–68.

Pella, J., and M. Masuda. 2001. Bayesian methods for analysis of stock mixtures from genetic characters. Fishery Bulletin 99:151–167.

Pella, J., and M. Masuda. 2005. Classical discriminant analysis, classification of individuals, and source population composition of mixtures. Pages 517–570 *in* S. X. Cadrin, K. D. Friedland, and J. R. Waldman, editors. Stock identification

methods: applications in fishery science. Elsevier Academic Press, Amsterdam, The Netherlands.
Pelleg, D., and A. Moore. 2000. X-means: extending *k*-means with efficient estimation of the number of clusters. Pages 737–734 *in* Proceedings of the 17th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, California, USA.
Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
Richardson, S., and P. J. Green. 1997. On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society (B) 59:731–792.
Robert, C. P., and G. Casella. 2004. Monte Carlo statistical methods. Second edition. Springer Science+Business Media, New York, New York, USA.
Ruttenberg, B. I., and R. R. Warner. 2006. Spatial variation in the chemical composition of natal otoliths from a reef fish in the Galápagos Islands. Marine Ecology Progress Series 328: 225–236.
Smouse, P. E., R. S. Waples, and J. A. Tworek. 1990. A genetic mixture analysis for use with incomplete source population data. Canadian Journal of Fisheries and Aquatic Sciences 47: 620–634.
Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society (B) 64:583–639.
Spiegelhalter, D. J., A. Thomas, N. Best, and D. Lunn. 2004. WinBUGS 1.4 User Manual. Technical report. MRC Biostatistics Unit, Cambridge, UK. ⟨http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf⟩
Standish, J. D., M. S. Sheehy, and R. R. Warner. 2008. The use of otolith natal elemental signatures as natural tags to evaluate connectivity among open-coast fish populations. Marine Ecology Progress Series 356:259–268.
Steinley, D. 2006. *K*-means clustering: a half-century synthesis. British Journal of Mathematical and Statistical Psychology 59:1–34.
Stephens, M. 2000. Dealing with label switching in mixture models. Journal of the Royal Statistical Society (B) 62:795–809.
Stockhausen, W. T., R. N. Lipcius, and B. M. Hickey. 2000. Joint effects of larval dispersal, population regulation, marine reserve design, and exploitation on production and recruitment in the Caribbean spiny lobster. Bulletin of Marine Science 66:957–990.
Swearer, S. E., J. E. Caselle, D. W. Lea, and R. R. Warner. 1999. Larval retention and recruitment in an island population of a coral-reef fish. Nature 402:799–802.
Thorrold, S. R., and J. A. Hare. 2002. Otolith applications in reef fish ecology. Pages 243–364 *in* P. F. Sale, editor. Coral reef fishes: dynamics and diversity in a complex ecosystem. Academic Press, San Diego, California, USA.
Thorrold, S., C. M. Jones, P. K. Swart, and T. E. Targett. 1998. Accurate classification of juvenile weakfish *Cynoscion regalis* to estuarine nursery areas based on chemical signatures in otoliths. Marine Ecology Progress Series 173:253–265.
Thorrold, S. R., C. Latkoczy, P. K. Swart, and C. M. Jones. 2001. Natal homing in a marine fish metapopulation. Science 291:297–299.
Titterington, D. M., A. F. Smith, and U. E. Makov. 1985. Statistical analysis of finite mixture distributions. John Wiley and Sons, San Diego, California, USA.
Waples, R. S., and O. Gaggiotti. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology 15:1419–1439.
Warner, R. R., S. E. Swearer, and J. E. Caselle. 2000. Larval accumulation and retention: implications for the design of marine reserves and essential fish habitat. Bulletin of Marine Science 66:821–830.
Warner, R. R., S. E. Swearer, J. E. Caselle, M. Sheehy, and G. Paradis. 2005. Natal trace-elemental signatures in the otoliths of an open-coast fish. Limnology and Oceanography 50: 1529–1542.
White, J. W., and B. I. Ruttenberg. 2007. Discriminant function analysis in marine ecology: some oversights and their solutions. Marine Ecology Progress Series 329:301–305.
Zacherl, D. C. 2005. Spatial and temporal variation in statolith and protoconch trace elements as natural tags to track larval dispersal. Marine Ecology Progress Series 290:145–163.
Zhang, Z., K. L. Chan, Y. Wu, and C. Chen. 2004. Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. Statistical Computing 14:343–355.

APPENDIX A

Details of Markov Chain Monte Carlo methods (*Ecological Archives* A018-068-A1).

SUPPLEMENT

Matlab code containing the MCMC algorithms used in the example (*Ecological Archives* A018-068-S1).