Transdimensional approximate Bayesian computation for inference on invasive species models with latent variables of unknown dimension

Oksana A. Chkrebtii, Erin K. Cameron, David A. Campbell, Erin M. Bayne

A R T I C L E   I N F O

Article history:
Received 20 March 2014
Received in revised form 27 December 2014
Accepted 5 January 2015
Available online 14 January 2015

Keywords:
Likelihood-free inference
Markov chain Monte Carlo
Non-native earthworms
Reversible jump

A B S T R A C T

Accurate information on patterns of introduction and spread of non-native species is essential for making predictions and management decisions. In many cases, estimating unknown rates of introduction and spread from observed data requires evaluating intractable variable-dimensional integrals. In general, inference on the large class of models containing latent variables of large or variable dimension precludes the use of exact sampling techniques. Approximate Bayesian computation (ABC) methods provide an alternative to exact sampling but rely on inefficient conditional simulation of the latent variables. To accomplish this task efficiently, a new transdimensional Monte Carlo sampler is developed for approximate Bayesian model inference and used to estimate rates of introduction and spread for the non-native earthworm species *Dendrobaena octaedra* (Savigny) along roads in the boreal forest of northern Alberta. Using low and high estimates of introduction and spread rates, the extent of earthworm invasions in northeastern Alberta is simulated to project the proportion of suitable habitat invaded in the year following data collection.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Biological invasions are occurring at unprecedented rates worldwide (Ricciardi, 2007) but often remain undetected until invading species have spread extensively. Detailed records documenting the time of initial introduction and subsequent changes in distribution are therefore not available for many invasions (Fang, 2005). This information is critical for projection of future expansion and for development of appropriate management strategies (Abbott, 2006). However, in cases where long-term temporal data is unavailable, current distribution patterns of non-native species can be used to infer how the invasion process may have occurred (Fang, 2005; McIntire and Fajardo, 2009).

Invasions that occur below-ground, such as earthworm invasions, can be particularly difficult to track over time and may initially proceed unnoticed. Non-native earthworms are currently spreading in many forests across North America.
(Frelich et al., 2006; Hendrix et al., 2008) but have limited ability to disperse actively (Marinissen and van den Bosch, 1992). Therefore, passive jump dispersal via abandonment of bait by anglers and transport along roads via vehicle traffic is thought to be important in their spread (Gundale et al., 2005; Frelich et al., 2006; Cameron et al., 2007). In northern hardwood and boreal forests, which are devoid of native earthworm species, earthworm invasions are causing significant changes to nutrient cycling and soil structure (Bohlen et al., 2004; Frelich et al., 2006; Hale et al., 2006). These changes have led to cascading effects on songbird (Loss and Blair, 2011) and plant communities (Scheu and Parkinson, 1994; Hale et al., 2006; Nuzzo et al., 2009). Because earthworms can affect other species directly, as well as indirectly via changes in the physical environment, their invasions may cause substantial changes over a large area of the Canadian boreal forest in the future. Accurate estimates of earthworm introduction and spread rates in this region are thus critical for the development of appropriate management strategies.

For many invasive species, spread occurs via a combination of diffusive spread around invaded sites and jump dispersal to new locations (i.e., stratified diffusion; Shigesada et al., 1995). Because even rare long-distance jump dispersal events result in faster spread than would be expected with diffusive spread alone (Shigesada et al., 1995), estimates of both introduction and spread rates are typically needed to predict the future spatial extent of an invasive species. Introductions of invasive species can be described using point process models, while models for diffusive spread following an introduction are application-dependent and can be complex (e.g., Illian et al., 2009). As a result, data likelihoods for such models are not often available in closed form.

Approximate Bayesian computation (ABC, Diggle and Gratton, 1984; Beaumont et al., 2002; Marjoram et al., 2003) has proven to be a useful approach for approximate inference on intractable likelihood problems, including distinguishing among introduction scenarios and invasion routes of non-native species (e.g., Miller et al., 2005; Lombaert et al., 2010). It relies on repeated model simulation in the absence of an explicit likelihood function. Proposed parameter values are accepted or rejected based on the distance between low-dimensional summaries of a conditional model realization and the observed data.

ABC algorithms are generally computationally intensive, and can result in unreliable estimates when too few proposals are accepted (e.g., Fearnhead and Prangle, 2012, Lemma 1). Existing implementations typically construct a Markov chain with a dependent proposal mechanism defined on a fixed probability space (Marjoram et al., 2003). However, in many models, the dimensionality of the parameter space varies (Green, 1995; Richardson and Green, 1997). In this paper we develop a transdimensional ABC algorithm that allows efficient exploration of parameter subspaces of variable dimension. This approach is applied to estimate introduction and spread rates of non-native earthworms in the boreal forest.

The paper is organized as follows. Section 2 describes a general point-process model for introduction of an invasive species and its subsequent spread. As the likelihood function is an intractable integral of variable dimension, a basic ABC algorithm for obtaining a sample from an approximate posterior distribution associated with this model is described. The novel transdimensional ABC approach is then presented as an efficient alternative to existing sampling methods. Section 3 introduces the motivating problem of estimating the rates of introduction and spread of the earthworm species Dendrobaena octaedra (Savigny) in northern Alberta by combining information from two datasets. The hierarchical model of the population dynamics is described, and algorithmic implementation details are provided. Inference results are summarized in Section 4, and are then utilized in a spatio-temporal simulation model to project the extent of invasion in a sample region in Section 5. Concluding remarks are presented in Section 6.

2. Approximate Bayesian computation

Let us consider a model for the invasion of a non-native species over a given region with the vector of unknown model parameters, $\theta$, such as rates of introduction, birth, predation, or spread. A stochastic mechanism generates $k \in \mathbb{N}$ introduction events on the spatio-temporal horizon $\mathcal{H} \subset \mathbb{R}^d$. Denote by $x^{(i)} \in \mathcal{H}$ the location of $i$th, $1 \leq i \leq k$, latent introduction event, and define the vector concatenation,

$$x_k = [x^{(1)}, \ldots, x^{(k)}]^\top \in \mathcal{H}^k \subset \mathbb{R}^{dk}.$$ 

The spatio-temporal spread resulting from the introductions $x_k$ follows a deterministic or stochastic model, generating the data $Y \in \mathcal{Y}$. The dependence structure of model components on the parameters $\theta$ is illustrated in Fig. 1(a).

Exact inference for model parameters $\theta$ is based on the posterior probability density,

$$\pi(\theta \mid Y) \propto p(Y \mid \theta) \pi(\theta), \quad (1)$$

where the first factor on the right hand side is the likelihood of the data given $\theta$, and the second is its prior. The model parameters impact the data indirectly through the number and configuration of introduction locations, $x_k$, which are not themselves of interest. Therefore the likelihood in expression (1) is obtained by integration with respect to the number and location of the latent introduction events:

$$p(Y \mid \theta) = \sum_{k=0}^{\infty} \int_{\mathcal{H}^k} p(Y \mid x_k, \theta) p(x_k \mid k, \theta) p(k \mid \theta) \, dx_k. \quad (2)$$
Estimation for this class of models poses analytical and computational challenges. In practice, evaluating \( p(Y \mid x_k, \theta) \) is often infeasible, especially under complex models for interaction and spread, or when the data is only partially observed in space. Furthermore, the domain of integration in (2) changes dimension with the dimension of the latent variable, leading to further difficulties in obtaining a closed form likelihood. For the motivating invasive species model in Section 3, even under simple deterministic linear spread in one dimension, evaluating the likelihood (2) presents a geometric and combinatorial challenge whose complexity renders exact inference practically limited. As a result, approximate simulation-based inference, such as ABC, often becomes the only tool available to approach such problems.

2.1. Approximate Bayesian inference for latent variable models

Assume that the model can be forward-simulated to generate a synthetic dataset \( D \in \mathcal{Y} \) from \( p(D \mid x_k, \theta) \). Under deterministic spread, \( p(D \mid x_k, \theta) \) is a point mass density around a data simulation function.

ABC replaces (1) with an approximation of the \( D \)-augmented posterior density marginalized over the synthetic data, \( D \), and latent variables, \((k, x_k)\). The approximation is based on the use of a low-dimensional summary statistic function \( s(\cdot) \), as follows:

\[
\pi(\theta \mid Y) \propto \sum_{k=0}^{\infty} \int_{\mathcal{X}^k} \pi(\theta) \ p(k \mid \theta) \ p(x_k \mid k, \theta) \ p(Y \mid x_k, \theta) \ dx_k \\
\propto \sum_{k=0}^{\infty} \int_Y \int_{\mathcal{X}^k} \pi(\theta) \ p(k \mid \theta) \ p(x_k \mid k, \theta) \ p(D \mid x_k, \theta) \ p(Y \mid s(Y), x_k, \theta) \ p(s(Y) \mid D) \ dx_k \ dD \\
\approx \sum_{k=0}^{\infty} \int_Y \int_{\mathcal{X}^k} \pi(\theta) \ p(k \mid \theta) \ p(x_k \mid k, \theta) \ p(D \mid x_k, \theta) \ p(s(Y) \mid D) \ dx_k \ dD \\
\equiv \pi_{ABC}(\theta \mid Y).
\]

The accuracy of the ABC approximation depends on the degree of sufficiency of the data summary \( s(Y) \) for \((\theta, k, x_k)\), and on the discrepancy term \( p(s(Y) \mid D) \) relating the summarized synthetic data to the observed data by a kernel \( K_{\epsilon} \) with bandwidth \( \epsilon \geq 0 \):

\[
p(s(Y) \mid D) = K_{\epsilon}[s(Y), s(D)].
\]

When \( s(Y) \) is sufficient for \((\theta, k, x_k)\) and \( \epsilon = 0 \), the ABC posterior density is exact. In other words, when \( p(Y \mid s(Y), x_k, \theta) = p(Y \mid s(Y)) \) and \( p(s(Y) \mid D) \) is a point mass function centred at \( s(D) \), then \( \pi(\theta \mid Y) = \pi_{ABC}(\theta \mid Y) \). However, low-dimensional sufficient statistics cannot in general be obtained when the likelihood is unknown, so the data summaries employed for ABC are chosen subjectively, leading to an approximate posterior density (Fearnhead and Prangle, 2012).

The bandwidth \( \epsilon \) controls the tolerance for the discrepancy between the summarized real and synthetic data. Effectively, \( \epsilon \) controls the trade-off between the dimension of the summary statistic and Monte Carlo error from too few data matches (e.g., Fearnhead and Prangle, 2012, Lemma 1).

Under the data simulation model, the ABC posterior for the general invasion model (Fig. 1(a)) is:

\[
\pi_{ABC}(\theta \mid Y) \propto \int_{\mathcal{Y}} \left[ \sum_{k=0}^{\infty} \int_{\mathcal{X}^k} p(D \mid x_k, \theta) \ p(x_k \mid k, \theta) \ p(k \mid \theta) \ dx_k \right] p(s(Y) \mid D) \ dD. \\
\]

Estimates of the mean, mode, or quantiles of \( \pi_{ABC}(\theta \mid Y) \) can now be obtained from a Monte Carlo sample. One simple method to produce such a sample is rejection ABC (Sisson and Fan, 2010), shown in Algorithm 1. More efficient ABC-MCMC sampling strategies rely on dependent proposals for parameters. However, our introduction model involves the parameter \( x_k \) whose dimension changes with \( k \), and therefore we develop a transdimensional sampling approach to avoid further approximation (such as the approximation of Congdon, 2006).
used as bait and therefore transport by vehicles is the key mechanism involved in passive dispersal (Tiunov et al., 2006).

earthworm species in northern Alberta (Cameron et al., 2007; Cameron and Bayne, 2009). This species is not commonly focused on the litter-dwelling earthworm (Appendix A).

produces a Markov chain whose stationary distribution has the desired ABC posterior density ((3); proof is provided in Appendix A).

The resulting transdimensional Metropolis–Hastings random walk ABC sampler is described in Algorithm 2 and generates synthetic data obtained from an area of positive posterior probability (e.g. via rejection ABC).

2.2. Transdimensional approximate Bayesian computation

Addressing the problem of low acceptance rates and unreliable estimates from the approximation in (3), we develop an efficient algorithm for obtaining samples. Instead of relying on conditional simulation on the variable-dimensional model subspace, a transdimensional approach (Green, 1995) is adopted allowing proposals between probability spaces of different dimensions.

In order to construct transitions between all model spaces of different dimension, it is sufficient to define pair-wise transitions between all model index pairs (i, j) ∈ 𝕃 × 𝕃, with associated model-specific parameter vectors 𝜙, ∈ ℜni and 𝜙, ∈ ℜnj. Model-specific vectors are augmented by auxiliary random variables, 𝑢, ∈ ℜni and 𝑢, ∈ ℜnj respectively, chosen by convenience under the constraint 𝑛i + 𝑚i = 𝑛j + 𝑚j, which ensures that both model spaces have the same dimension.

The diffeomorphic transformation 𝜙ij : ℜni × ℜmj → ℜnj × ℜmj is defined corresponding to the mapping (𝜙, 𝑢, ) → (𝜙, 𝑢, ). For a given transformation 𝜙ij with Jacobian 𝑗ij = |∂𝜙ij(𝜙, 𝑢, )/∂(𝜙, 𝑢, )|, a proposed move from model i to j is constructed by first drawing a vector 𝑢, from the density 𝑝, 𝑢, (𝑢, ) and then obtaining (𝜙, 𝑢, ) = 𝜙ij(𝜙, 𝑢, ). The inverse mapping (𝜙, 𝑢, ) → (𝜙, 𝑢, ) can be accomplished by using the transformation 𝜙ij. For simplicity, one generally sets 𝑚j = 0 for all (i, j) such that 𝑛j ≥ 𝑛i.

The resulting transdimensional Metropolis–Hastings random walk ABC sampler is described in Algorithm 2 and produces a Markov chain whose stationary distribution has the desired ABC posterior density ((3); proof is provided in Appendix A).

3. Motivating application

3.1. Field methods

Data on earthworm occurrence were collected in the boreal forest of northern Alberta, Canada between 54.4° N and 57.8° N latitude and 110.1° W and 119.8° W longitude (Fig. 2; see Cameron and Bayne, 2009 for further description). We focused on the litter-dwelling earthworm D. octaedra (Savigny), which is introduced from Europe and is the most common earthworm species in northern Alberta (Cameron et al., 2007; Cameron and Bayne, 2009). This species is not commonly used as bait and therefore transport by vehicles is the key mechanism involved in passive dispersal (Tiunov et al., 2006).

Earthworms were sampled at roads ranging in age from 6 to 56 years old in 2006. At each road, sampling occurred along a 50.25 m transect which ran parallel to the road. Leaf litter was hand-sorted to determine earthworm occurrence in six 0.0625 m² (25 by 25 cm) quadrats with a 9.75 m gap between quadrats on each transect. Transects were 1–2 m into the
Fig. 2. Survey locations (n = 78), represented by black circles, within the boreal forest of northern Alberta. Sites were clustered into 8 groups according to spatial location, represented by ovals.

forest from its edge, with alternate quadrats located ∼5 m farther into the forest interior. We will hereafter refer to this as the spatial data Y.

To examine the spread rate within our study area, additional temporal data on earthworm occurrence were collected along transects perpendicular to roads where earthworms were already established (n = 26). These transects were 500 m long, with sampling quadrats every 50 m. Sampling occurred in 2006 (Cameron and Bayne, 2009, 2015) and again in 2012/13. We subtracted the distance of the farthest quadrate with earthworms present in 2006 from the distance of the farthest occupied quadrate in 2012/13 and divided by 6 years. We will refer to this as the temporal data Z.

3.2. Model inference for earthworm invasions

We now propose a model for the invasion of earthworms along one spatial dimension over time. Our framework is applicable for any model with a latent variable structure, requiring integration with respect to parameters of variable dimension, as long as data can be obtained by forward-simulation from the model.

The 78 observation transects were grouped into eight categories based on their spatial location on the landscape (Fig. 2). Each of these groups contained young, intermediate-aged, and old roads. Observations \( Y_{gr} \in \{0, 1\}^6 \) for road \( r = 1, \ldots, n_g \) in group \( g = 1, \ldots, 8 \) consist of binary error-free measurements of presence or absence at the sampled 6 quadrats of the corresponding transect. For notational clarity, we omit dependence of the introduction locations on \( k \) and instead define \( x_{gr} \) to be the value of \( x_k \) associated with group \( g \) and road \( r \).

For every road in the study, we model earthworm introductions and spread over the spatio-temporal horizon \( \mathcal{H}_{gr} \), defined by the largest possible extent of activity that can affect the observed data under our model, as shown in Fig. 3. We assume that the introduction rate is constant within each of the 8 selected road groups per unit area of \( \mathcal{H}_{gr} \) by modelling the \( k \) introductions according to a homogeneous space–time Poisson process with rate \( \lambda_g \times \text{Area of } \mathcal{H}_{gr} \in \mathbb{R}^+ \), measured in introductions/(m × yr) to account for different road ages.

For a \( T_{gr} \)-year-old road, earthworms are assumed to spread linearly in time from point-source locations, \( x_{gr} \), at an unknown constant rate, \( v \in \mathbb{R}^+ \) (measured in m/yr), defining the geometry of the horizon \( \mathcal{H}_{gr} = H(v, T_{gr}) \in \mathbb{R}^2 \). If earthworms spread far enough from \( x_{gr} \) to overlap with a sampling quadrat, a presence indicator is recorded as shown in Fig. 3. We denote the resulting presence or absence of worms at the measurement locations by \( m(x_{gr}, v, T_{gr}) \). Since we assume that presence or absence of worms in a quadrat is measured without error, the data generating mechanism \( D = m(x_{gr}, v, T_{gr}) \) is conditionally deterministic. For expositional clarity we use the point mass density, \( \delta m(x_{gr}, v, T_{gr}) \), to highlight our use of a conditionally deterministic data generating mechanism. The resulting hierarchical model, illustrated
Fig. 3. Simulation horizon, \( H_{gr} \), consisting of all spatio-temporal allocations where an introduction event will affect the data according to our spread model. Two introductions and subsequent spread are shown shaded in grey. The data generated by the illustrated invasion is 
\[
\begin{aligned}
& m(x_{gr}, \nu, T_{gr}) = [1, 0, 0, 1, 1, 0], \\
& \text{in Fig. 1(b), is:}
\end{aligned}
\]

\[
Y_{gr} \mid x_{gr}, \nu \sim \delta_m(x_{gr}, \nu, T_{gr}),
\]

\[
x_{gr} \mid k_{gr}, \nu \sim \text{Uniform}(H(\nu, T_{gr})),
\]

\[
k_{gr} \mid \lambda_g, \nu \sim \text{Poisson}(\lambda_g \times \text{Area of } H(\nu, T_{gr})),
\]

\[
Z \mid \nu \sim \text{Normal}(\nu, \sigma^2),
\]

\[
\nu \sim \text{Normal}(10, 10^2),
\]

\[
\lambda_g \sim \text{Exponential}(1),
\]

\[
\sigma^2 \sim \pi(\sigma^2) \propto 1/\sigma^2.
\]

\[\text{(4)}\]

in Fig. 1(b), is:

\[
Y_{gr} \mid x_{gr}, \nu \sim \delta_m(x_{gr}, \nu, T_{gr}),
\]

\[
x_{gr} \mid k_{gr}, \nu \sim \text{Uniform}(H(\nu, T_{gr})),
\]

\[
k_{gr} \mid \lambda_g, \nu \sim \text{Poisson}(\lambda_g \times \text{Area of } H(\nu, T_{gr})),
\]

\[
Z \mid \nu \sim \text{Normal}(\nu, \sigma^2),
\]

\[
\nu \sim \text{Normal}(10, 10^2),
\]

\[
\lambda_g \sim \text{Exponential}(1),
\]

\[
\sigma^2 \sim \pi(\sigma^2) \propto 1/\sigma^2.
\]

The prior on \( \nu \) is based on related studies of earthworm spread rates (Hale, 2008) but has an inflated variance to account for differences in soil characteristics and climate between studies, while the prior on \( \lambda_g \) reflects the expectation that introduction events should be relatively rare in the boreal forest.

Under this simple scenario, the space–time horizon \( H_{gr} \) can be subdivided into distinct regions whose shape depends on \( \nu \) and road age, and the effect on the data of an introduction event occurring within each region can be computed exactly. Therefore, we can obtain \( p(Y_{gr} \mid x_{gr}, \nu) \) in closed form using a geometric and combinatorial argument, and subsequently compute the likelihood \( p(Y_{gr} \mid x_{gr}, \nu) \) through integration over these distinct regions, as shown in the Supplementary materials (see Appendix B). In practice, such lengthy likelihood calculations are not feasible for general practitioners and their complexity introduces the potential for errors in implementation. However, forward simulation is often fast and computationally accessible for practitioners. As the difficulties with exact implementation increase with all but the simplest models, we illustrate transdimensional ABC as a practical alternative for this problem.

Our inference about introduction and spread is based on the marginal ABC posterior density:

\[
\pi_{ABC}(\lambda_1, \ldots, \lambda_8, \nu, \sigma^2 \mid Y_{11, \ldots, Y_{8, ng}, Z}) \propto p(\sigma^2 \mid \nu, Z) p(\nu \mid Z)
\]

\[
\times \prod_{g=1}^{8} \prod_{r=1}^{n_g} \left[ \pi(\lambda_g) p(s(Y_{gr}) \mid D_{gr}) \sum_{k_{gr}=0}^{\infty} \int_{H(\nu, T_{gr})} p(x_{gr} \mid k_{gr}, \nu) p(k_{gr} \mid \lambda_g, \nu) \, dx_{gr} \right].
\]

\[\text{(5)}\]

We use the transdimensional ABC sampler described in Algorithm 3, which is a variant of Algorithm 2, implemented with a birth–death proposal for the number of introductions for each road. A birth–death proposal consists of either increasing by one (birth), decreasing by one (death), or maintaining unchanged the number of introduction events with some probability. This sampling algorithm mimics the data-generating process to produce synthetic data efficiently, which leads to the fast mixing desired in an ABC algorithm. The birth–death proposal is a special case of the general transdimensional proposal with Jacobian of the transformation given by \(|J| = 1\), because \( \phi_{ij} \) is taken to be the identity function of the auxiliary variables, which are sampled uniformly on \( H_{gr} \).
Due to the efficiency of this algorithm, we are able to use an error tolerance of $\epsilon = 0$, with a point mass kernel distance metric to define the term,

$$p \left( s(Y_{gr}) \mid D_{gr} \right) = \mathbb{I} \{ s(Y_{gr}) = s(D_{gr}) \} .$$

This corresponds to accepting proposed parameters only when the resulting summarized simulated data matches the observed simulated data exactly. For the summary statistic, $s$, we chose a 2-dimensional vector consisting of: (i) the number of consecutive occupied quadrats (strings of 1s in $Y_{gr}$), and (ii) the total number of occupied quadrats:

$$s(Y_{gr}) = \left[ \text{number of strings of consecutive 1s}, \sum_{i=1}^{g} Y_{gr}^{(i)} \right]^\top .$$

Each consecutive sequence of occupied quadrats (strings of 1s) indicates that at least one introduction must have occurred. The length of each consecutive string of occupied quadrats can help distinguish between a recent introduction (e.g., 1 or 2 consecutive occurrences) and one or more old introductions (e.g., 6 consecutive occurrences).

**Algorithm 3** Birth-death ABC

at step $\ell = 0$, initialize $(\lambda, v, k, x)^{(0)}$ from an area of positive posterior probability and generate $D$ from $m(x, v, T)$

for $\ell = 1$ to $l$ do

sample $\sigma^{2(\ell)} \sim p(\sigma^{2} \mid \nu^{(\ell - 1)}, x^{(\ell - 1)}, k^{(\ell - 1)}, Y, Z)$

sample $\nu^{(\ell)} \sim p(\nu \mid \lambda^{(\ell - 1)}, x^{(\ell - 1)}, k^{(\ell - 1)}, \sigma^{2(\ell)}, Y, Z)$

for $g = 1$ to 8 do

sample $\lambda_{g}^{(\ell)} \sim p(\lambda \mid v^{(\ell)}, x^{(\ell - 1)}, k^{(\ell - 1)}, \sigma^{2(\ell)}, Y, Z)$

for $r = 1$ to $n_g$ do

sample $x_{gr}^{(\ell)}$ from $p(x_{gr} \mid \nu^{(\ell)}, k^{(\ell - 1)} - 1, \sigma^{2(\ell)}, Y^{(\ell)}, Z^{(\ell)})$

generate synthetic data $D_{gr}$ from $m(x_{gr}, \nu^{(\ell)}, T_{gr})$

propose $b \sim \text{Discrete Uniform}\{-1, 0, 1\}$ and set $k_{gr}^{*} = k_{gr}^{(\ell - 1)} + b$

if $b = 1$ then

sample $x \sim U \left[ H(\nu^{(\ell)}, T_{gr}) \right]$

else if $b = -1$ then

generate $d \sim \text{Uniform}\{1, \ldots, k_{gr}^{(\ell - 1)}\}$

$x_{gr}^{*} = (x_{gr}^{(\ell)})_{-d}$

else

$x_{gr}^{*} = x_{gr}^{(\ell)}$

end if

generate synthetic data $D_{gr}^{*}$ from $m(x_{gr}^{*}, \nu^{(\ell)}, T_{gr})$ and calculate:

$$\alpha = \frac{p(s(Y) \mid D^{*})}{p(s(Y) \mid D)} \times \frac{p(x_{gr}^{*} \mid k_{gr}^{*}) p(k_{gr}^{*})}{p(x_{gr}^{(\ell)} \mid k_{gr}^{(\ell)}) p(k_{gr}^{(\ell)})} \times \frac{q(k_{gr}^{*} \mid k_{gr}^{(\ell)})}{q(k_{gr}^{(\ell)} \mid k_{gr}^{*})}$$

if $\min\{1, \alpha\} \geq \text{Uniform}(0, 1)$ then

set $(k_{gr}^{(\ell)}, x_{gr}^{(\ell)}, D_{gr}^{(\ell)}) \leftarrow (k_{gr}^{*}, x_{gr}^{*}, D_{gr}^{*})$

else

set $(k_{gr}^{(\ell)}, x_{gr}^{(\ell)}, D_{gr}^{(\ell)}) \leftarrow (k_{gr}^{(\ell - 1)}, x_{gr}^{(\ell - 1)}, D_{gr}^{(\ell - 1)})$

end if

end for

end for

4. Inference results and simulation study

4.1. Inference on introduction and spread rates

For this analysis we performed 250,000 iterations of Algorithm 3 and discarded the first 25,000 iterations as burn-in. A Geweke convergence diagnostic was performed comparing the first 50,000 iterations (post burn-in) to the last 50,000 iterations for each of $\lambda_g, g = 1, \ldots, 8$ and $v$ assuming a 4% autocorrelation. The test statistic returned p-values in the range of 0.17 to 0.65 for each parameter. Bivariate posterior density heat maps for the introduction and spread rates, $v$ and $\lambda_g$, are shown in Fig. 4. The marginal posterior densities of $\lambda_g, g = 1, \ldots, 8$ can be classified into three categories as shown in the top panel of Fig. 5. Groups 4 and 2 have the lowest posterior median rates of introduction ($1.31 \times 10^{-5}$ and $2.69 \times 10^{-5}$ introductions/yr) respectively. Groups 1, 3, 5, and 8 have posterior median introduction rates in the
mid-range \( (4.03 \times 10^{-5}, 4.12 \times 10^{-5}, 3.50 \times 10^{-5}, \text{and} 4.09 \times 10^{-5} \) introductions/(m x yr) respectively). Finally, groups 6 and 7 have the highest posterior median introduction rates \( (6.53 \times 10^{-5}, 7.38 \times 10^{-5} \) introductions/(m x yr) respectively). The marginal posterior of \( \nu \) is shown in the middle panel of Fig. 5, where the posterior median was estimated at 13.93 m/yr with a 95% credible interval between 10.60 and 16.99 m/yr. The bottom panel of Fig. 5 shows the marginal density of spatio-temporal introduction locations, \( \pi_{ACB} \{ x_{1,1}, \ldots, x_{8,n_g}, Z \} \), from a 26 year old road in group 1. The associated presence or absence data, \( Y_{1,11} = [1, 1, 0, 0, 1, 0] \), indicates an absence between two consecutive “split” observed presences. The split between subsequent presence measurements induces the distinct marginal density of introductions over the spatio-temporal horizon. This posterior non-uniformity suggests that employing any ABC sampling scheme that generates all introduction events, \( x_{gr} \), from the uniform prior will be extremely inefficient, and supports the use of a transdimensional approach.

4.2. Simulation study

A simulation study was conducted to examine the degree of posterior approximation resulting from summarizing the data by the statistic \( s \), as well as the appropriateness of the model for describing the invasion process. Although these effects are difficult to distinguish from a simulation study, we are able to attribute low coverage for two of the road groups studied to model specification.

We considered several options for simulation study designs. Generating data according to Model (4) given model parameters drawn from the full ABC posterior, \( \pi_{ACB}(\lambda_1, \ldots, \lambda_8, \nu, k_{1,1}, \ldots, k_{8,n_g}, x_{1,1}, \ldots, x_{8,n_g} \mid Y_{1,1}, \ldots, Y_{8,n_g}, Z) \), would yield data summaries which exactly match the real observed data summaries. Because this would amount to repeating the real data analysis multiple times, we deemed this strategy to be unsuitable. An alternative strategy consists of sampling introductions and spread rates from the marginal ABC posterior (5) and then independently sampling the number and locations of introductions from \( \pi_{ACB}(k_{1,1}, \ldots, k_{8,n_g}, x_{1,1}, \ldots, x_{8,n_g} \mid Y_{1,1}, \ldots, Y_{8,n_g}, Z) \). This strategy was also found unsuitable because of the downward bias on introduction rates resulting from the possibility of generating introduction events, \( x_{gr} \), outside of the horizon, \( H(\nu, T_{gr}) \). In order to avoid the above issues we generated data from Model (4) using fixed \( \lambda_g, g = 1, \ldots, 8 \) and \( \nu \) chosen to be the posterior median values of group specific introduction rates and overall spread rate, as reported in the previous section.

The data generating process and inference were repeated 150 times using the transdimensional birth–death ABC method described in Algorithm 3 implemented with 250 000 iterations, and 95% credible intervals were computed. A Geweke diagnostic was performed component-wise for \( \lambda_g, g = 1, \ldots, 8 \) and \( \nu \), comparing the final 100 000 iterations with the initial 100 000 iterations after discarding 25 000 iterations for burn-in. In all cases convergence was attained. Table 1 shows the median of posterior medians, median of posterior standard deviations (however the skew in the posterior densities implies caution should be used when interpreting this value), and the proportion of times that the 95% credible intervals covered the introduction rate used to generate the data. The 95% credible regions for the simulated datasets covered the

---

**Fig. 4.** Marginal posterior densities of spread rate (vertical axis) and introduction rate (horizontal axis) by group of roads. Red dots indicate the values of \((\lambda, \nu)\) used for the predictive simulation described in Section 5.
Fig. 5. Marginal posterior densities for the rate of introduction, $\lambda_g$ (top). Marginal posterior density for the worm spread rate, $\nu$, (middle). Heatmap showing the marginal posterior of spatio-temporal introductions for the 26 year old road in group 1 (bottom).

true spread rate 100% of the time, and covered the true group specific introduction rates with reasonable frequency for all road groups, except for road groups 6 and 7.

The large values of the spread rates $\lambda_6$ and $\lambda_7$ that were chosen for the simulation study make multiple introduction events likely, especially for older roads. However, uniform sampling from $p(x_{gr} | k_{gr}, \nu)$ over $H(\nu, T_{gr})$ given $k_{gr} \geq 1$ introductions, generates presence in all quadrats with high probability. This is in contrast to the sets of split consecutive presence measurements separated by absences which were observed in the field data. Whereas conditioning on multiple consecutive presence measurements in the field data leads to high estimated introduction rates and marginal posterior densities of introduction events such as that illustrated in Fig. 5, a systematic lack of such multiple consecutive presence measurements in the simulated data leads to an underestimate of the number and rate of introductions. From this we conclude that for groups 6 and 7, the model assumption that introductions occur according to a homogeneous space–time Poisson process may not be reasonable. A more likely scenario for these two groups might consist of increased frequency of introductions in recent years leading to multiple consecutive separated strings of presences. Another possible culprit may be the assumption of a constant spread rate, which in reality may be hindered by obstacles to spread along roads. These possibilities present interesting subjects for further exploration and study of introduction and spread patterns for these individual regions.
5. Discussion and prediction

The highest estimated introduction rates were for groups 6 and 7, which are located in the southwestern part of Alberta’s boreal forest, near the Peace River Region. In this area, human settlement and agricultural conversion occurred earlier and more extensively than in the rest of northern Alberta (Schneider, 2002). The higher earthworm introduction rates may thus be related to the greater intensity and longer history of human activity. However, it is not clear why groups 2 and 4 had substantially lower introduction rates than other locations, as levels of anthropogenic disturbance are relatively similar to levels for groups 1, 3, 5, and 8. More intensive sampling at sites along a gradient of human activity would be needed to examine effects of anthropogenic disturbances on spatial variability of earthworm introduction rates. In all cases, the introduction rates estimated using our approach were lower than previously estimated from the same dataset (Cameron and Bayne, 2009). The analysis in Cameron and Bayne (2009) yielded an estimated introduction rate of $1.03 \times 10^{-3}$ introductions/(m × yr), under the assumption of no active dispersal between sampling sites and a spread rate of 10 m/yr as obtained from literature from other regions. Attributing all occurrences to passive dispersal is likely to have produced an over-estimate of the introduction rate in Cameron and Bayne (2009).

To illustrate how the introduction and spread rate estimates produced by this analysis can be used to model invasive species distributions, we modelled earthworm distribution within the Alberta Pacific Forest Industries Forest Management Area (Al-Pac FMA), a 59 054 km$^2$ area in north-eastern Alberta. We obtained road network and Alberta Vegetation Inventory (AVI) data from Alberta-Pacific Forest Industries and additional data on road ages from the Mistakiis Institute. In ArcGIS 10.1, we generated points every 10 m along the road network, randomly invaded these points using the introduction rate from our analysis, and created buffers around invaded points using the spread rate, as described in Cameron and Bayne (2009). We produced a map of the predicted areal extent of earthworms for 2006 (the year for which we were able to obtain road age data), which was intersected with a GIS layer containing forest habitat suitable for invasion. All forest types were considered to be suitable except stands in which black spruce $Larix laricina$ were dominant, as such forests have highly acidic soils and are thus less likely to be colonized by earthworms (Bouché, 1977; Edwards and Bohlen, 1996). Because we did not have a large number of sites sampled representatively across the Al-Pac FMA, we created the maps in Fig. 6 based on three sets of introduction and spread rates, corresponding to high, middle, and low estimates. The extreme values (frequent introduction with fast spread vs rare introduction with slow spread) were used to heuristically illustrate the magnitude of the prediction uncertainty for the areal extent of earthworms in 2006, while the middle values were used to illustrate predictions of the extent after 50 years following data collection. Since the Al-Pac FMA is located between groups 1–4, a low value of $(\lambda, \nu)$ was selected from the lower end of the bivariate posterior distributions for groups 2 and 4, while a high value of $(\lambda, \nu)$ was selected from the higher end of the distributions for groups 1 and 3. The middle value was taken to be the median of the four group posteriors. All three points are shown in Fig. 4.

The total area of suitable habitat within the Al-Pac FMA in 2006 was 24 449.3 km$^2$, with a total road length of 22 068 km. Using an introduction rate of $1.5915 \times 10^{-6}$ introductions per metre of road per year and a spread rate of 11.596 m/yr, our model predicts that 8.02 km$^2$ (0.03%) of suitable habitat was invaded by 2006. When using the higher values of $3.7011 \times 10^{-5}$ introductions/(m × yr) and 16.655 m/yr, the model predicts that 905.65 km$^2$ was invaded (3.70% of total suitable habitat). Both maps are shown in Fig. 6. In comparison, Cameron and Bayne (2009) predicted 9.09% of the total area within the Al-Pac FMA would be invaded by earthworms based on their higher estimated rate of introduction of $1.03 \times 10^{-3}$ introductions/(m × yr), and a fixed, non-estimated spread rate of 10 m/yr. However, our current approach to estimating the spread and introduction rates accounts for active dispersal and should produce more realistic estimates (see Fig. 7).

A type of problem that poses challenges for traditional fixed-dimensional formulations such as Metropolis–Hastings ABC is illustrated in Fig. 5. A heat map of the approximate marginal density of introduction locations for one particular road is shown. Because the location vector changes dimension with the number of introductions, a fixed-dimensional ABC algorithm requires a forward-simulation step (rejection ABC) for the conditional simulation of this parameter. Instead, a transdimensional kernel allows us to make use of dependent proposals, thereby reducing Monte Carlo error when the areas of prior and posterior probabilities do not match. In this case, the prior on introduction locations is uniform over the event horizon,

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Simulation study results including the median of posterior medians and coverage probability of the 95% highest posterior credible intervals.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\nu$</td>
<td>$\lambda_1$</td>
</tr>
<tr>
<td>True value</td>
<td>$14.216$</td>
</tr>
<tr>
<td>Median of posterior medians</td>
<td>$15.53$</td>
</tr>
<tr>
<td>Median of posterior standard deviations</td>
<td>$1.23$</td>
</tr>
<tr>
<td>Percent of highest posterior intervals containing true value</td>
<td>$100$</td>
</tr>
</tbody>
</table>

O.A. Chkrebtii et al. / Computational Statistics and Data Analysis 86 (2015) 97–110
but the regions of high posterior density are concentrated in a very thin band around the edges of the horizon, rendering conditional simulation extremely inefficient.

6. Conclusion

Invasions frequently occur through a combination of long-distance jump dispersal events and diffusive spread around invaded sites, resulting in likelihoods that are often intractable. ABC methods provide an approximate inferential framework when the likelihood of the data cannot be evaluated. This paper develops a new efficient ABC sampler for a large class of models with infeasible or fully intractable likelihoods containing variable-dimensional integrals over a set of latent variables.

This class of models is used in a variety of applications. Spatio-temporal dynamic systems often combine stochastic generating mechanisms with complex time-evolution models, so that evaluation of the likelihood requires integration over a variable number of latent point sources (e.g., van Lieshout and van Zwet, 2001). Kingman’s coalescent models (Kingman, 1982) in genetics are another example where observed data is generated via a process that depends on the number and location of latent branching points defining a genealogical tree. Likelihoods of the genetic data are typically unavailable and consist of an integral over the space of all branches (e.g., Tavaré et al., 1997).

Latent variable models with intractable likelihoods pose a challenge to existing ABC–MCMC samplers defined on fixed-dimensional probability spaces. In contrast, we were able to quickly obtain 250 000 MCMC samples per group of roads under each of the conditional values of \( \nu \) using the more efficient transdimensional ABC algorithm proposed in this paper, even when requiring exact matches between summaries (\( \epsilon = 0 \)).

The main structural limitation of our methodology is that the resulting inference is approximate, controlled by the degree of sufficiency of our selected summary statistics and the chosen error tolerance. However, this limitation is inherent in the problem of inferring parameters under intractable likelihoods. All simulation-based methods suffer from the problem of dimensionality, precluding exact likelihood-free inference. Compared to established ABC methods, the sampling efficiency from our methodology allows strict error tolerances to be imposed, thereby improving the approximation. The transdimensional ABC approach proposed in this paper can be applied to problems where the likelihood consists of intractable variable-dimension integrals.
Fig. 7. Al-Pac FMA simulation region. Predicted areal extent of earthworms for 2006 (above) and 2056 (below) are indicated in black under an introduction rate of \(4.0 \times 10^{-5}\) introductions per metre of road per year and a spread rate of \(14.0932\) m/yr. The area invaded is 691.17 km\(^2\) (2.83% of suitable habitat) in 2006 and 9404.36 km\(^2\) (38.46% of suitable habitat) in 2056.

Acknowledgements

We would like to thank Dr. Mark Lewis, Dr. Marty Krkosek, and Stephanie Peacock for their very useful suggestions regarding our model, and to the Bamfield Marine Sciences Centre for providing the opportunity to begin this project as part of a Models in Ecology course. We would also like to acknowledge our many field technicians for assistance with data collection and Charlene Nielsen for her assistance in creating the GIS spread maps. The authors were funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) 2010:389148 and RGPIN04040. The project was also funded by the Alberta Biodiversity Monitoring Institute, Alberta Innovates Bio Solutions BIO-12-006, and the Alberta Livestock and Meat Agency 2012S007S. O.C. gratefully acknowledges the support of the Pacific Institute for Mathematical Sciences International Graduate Training Centre in Mathematical Biology.
Appendix A

A.1. Transdimensional ABC algorithm

Theorem A.1. The Markov chain generated via Algorithm 2 has invariant distribution:

\[
\pi_{ABC}(\theta | Y) \propto \pi(\theta) \sum_{k=0}^{\infty} \int \int_{\mathbb{R}^{nk}} p(D | x_k, \theta) p(x_k | k, \theta) p(k | \theta) d\lambda_k p(s(Y) | D) dD. \tag{A.1}
\]

Proof. Define the augmented spaces \( \mathcal{M}_k = \Theta \times k \times \mathbb{R}^{nk} \times \mathbb{R}^{mk} \times \mathbb{R}^k \), for \( k \in \mathcal{K} \), and the variable-dimensional parameter space \( \mathcal{M} = \bigcup_{k \in \mathcal{K}} \mathcal{M}_k \). Our goal is to construct a Markov chain with invariant distribution, \( \pi(dz) = \pi(z) dz \propto p(s(Y) | D_k) p(D_k | x_k, \theta) p(x_k | k, \theta) p(k | \theta) \pi(\theta) d\lambda, \) where \( \pi(\theta) \) is the Lebesgue measure on \( \sigma(\mathcal{M}) \).

We use an argument similar to Green (1995) to define transitions on this parameter space. First, we restrict attention to moves between any two model spaces \( \mathcal{M}_i \) and \( \mathcal{M}_j \). Consider elements \( v = (\theta_i, i, x_i, u_i, D_i) \in \mathcal{M}_i \) and \( w = (\theta_j, j, x_j, u_j, D_j) \in \mathcal{M}_j \), where \( x_k \in \mathbb{R}^{nk}, u_k \in \mathbb{R}^{mk}, k = i, j \). The constraint \( m_i + n_i = m_j + n_j \) ensures that the dimension of \( v \) and \( w \) match. Next, define the following proposal distributions for a move from \( v \) to \( w \) and back:

\[
Q(v, dw) = q(\theta_i | \theta_j) q(j | i) q(x_i, u_j | x_j, u_j) p(D_j | x_j) d\lambda, \\
Q(w, dv) = q(\theta_i | \theta_j) q(i | j) q(x_i, u_i | x_j, u_j) p(D_i | x_i) d\lambda, \\
\]

where \( \lambda \) is the Lebesgue measure on \( \mathcal{B} = \sigma(\mathcal{M}_k), k = 1, 2 \). Now let \( \alpha : \mathcal{M} \times \mathcal{M} \rightarrow [0, 1] \) be an acceptance probability, and define \( \delta_{\alpha} \) to be the Dirac delta measure on \( \mathcal{B} \) centred at \( \alpha \). Then, the corresponding transition kernels are:

\[
P(v, dw) = \alpha(v, w) Q(v, dw) + \delta_{\alpha}(dv) \int_{\mathcal{M}_j} [1 - \alpha(v, w)] Q(w, dm), \\
P(w, dv) = \alpha(w, v) Q(w, dv) + \delta_{\alpha}(dv) \int_{\mathcal{M}_i} [1 - \alpha(w, v)] Q(v, dm),
\]

on the measurable space \( (\mathcal{M}_i \times \mathcal{M}_j, \mathcal{B} \times \mathcal{B}) \), and \( (\mathcal{M}_j \times \mathcal{M}_i, \mathcal{B} \times \mathcal{B}) \), respectively. We assume that the Markov chain associated with this transition kernel is aperiodic and irreducible (true if proposal distribution \( Q(v, dw) \) generates an aperiodic and irreducible chain). The detailed balance condition is satisfied iff \( A, B \in \mathcal{B}, \)

\[
\pi(dv) P(v, dw) = \pi(dw) P(w, dv),
\]

is satisfied for \( A, B \in \mathcal{B}, \)

\[
\pi(dv) Q(v, dw) \alpha(v, w) = \pi(dw) Q(w, dv) \alpha(w, v), \quad v \in A, w \in B.
\]

Next, define a diffeomorphic transformation, \( \phi_\lambda : \mathbb{R}^{nk} \times \mathbb{R}^{mk} \rightarrow \mathbb{R}^{nk} \times \mathbb{R}^{mk} \). Using Lebesgue measure \( \lambda \) on \( (\mathcal{M}_i \times \mathcal{M}_j, \mathcal{B} \times \mathcal{B}) \) define the new measure,

\[
\mu(A \times B) \equiv \lambda\left[ \{ v \in A \cap \mathcal{M}_i, (\theta_i, i, \phi_\lambda(x_i, u_i), D) \in B \cap \mathcal{M}_j \} \cup \{ v \in A \cap \mathcal{M}_j, (\theta_j, j, \phi_\lambda(x_j, u_j), D) \in B \cap \mathcal{M}_i \} \right] \\
= \lambda\left[ \{ v \in A \cap \mathcal{M}_i, (\theta_j, i, \phi_\lambda(x_i, u_i), D) \in B \cap \mathcal{M}_j \} + \{ v \in A \cap \mathcal{M}_j, (\theta_i, j, \phi_\lambda(x_j, u_j), D) \in B \cap \mathcal{M}_i \} \right] \\
= \mu(A \times B)
\]

which is symmetric on \( (\mathcal{M}_i \times \mathcal{M}_j, \mathcal{B} \times \mathcal{B}) \). Then \( \pi(dv) Q(v, dw) \) and \( \pi(dw) Q(w, dv) \) have densities with respect to \( \mu \). These are given by,

\[
f(v, w) = \pi(dv) Q(v, dw) \\
f(w, v) = \pi(dw) Q(w, dv)
\]

respectively. Then reversibility across all moves is guaranteed (Green, 1995; Tierney, 1998) if \( \alpha(v, w) = \min\{1, f(v, w) / f(w, v)\} \) for all \( i, j \in \mathcal{K} \), and so detailed balance is satisfied under

\[
\alpha(v, w) = \min\left\{ \frac{p(s(Y) | D_j) p(x_j | j) p(j | \theta_j) p(\theta_j)}{p(s(Y) | D_i) p(x_i | i) p(i | \theta_i) p(\theta_i)}, \frac{p(s(Y) | D_i) p(x_i | i) p(i | \theta_i) p(\theta_i)}{p(s(Y) | D_j) p(x_j | j) p(j | \theta_j) p(\theta_j)} \right\},
\]

for all move types. As Algorithm 2 does not return the values of the auxiliary vectors, \( u_k \), the resulting marginal invariant distribution is (A.1). □
Appendix B. Supplementary materials

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2015.01.002.

References


