Inference From Gene to Population

Propagating Uncertainty In Estimates of Population Characteristics Through Ecological Scales

Joseph Daniel Chipperfield

PhD

Department of Biology
May 2010

Corrections and Response to Examiners

I would like to thank the examiners for providing me with very constructive feedback. Below is a list of the corrections made in response to the examiners comments.

Chapter 1

1. An extra section on the philosophy behind Bayes theorem and its applicability to the problems contained in this thesis is presented
2. Much of the wording relating to the introduction of the parentage analysis techniques has been changed to make it more clear. Terms related to 'null alleles' and 'phenotypes' are now defined in the correct places and an more explanation has been given to the differences between the different marker types
3. Typographical errors have been corrected

Chapter 2

1. Mistakes in mathematical notation have been corrected
2. Misplaced text appearing in results section has now been moved to the correct location and the results section has been expanded
3. The megagametophyte dataset has been released as part of the `ecomodtools` package for the R statistical platform on the R-Forge repository
4. A section has been added to the chapter explaining why it is not appropriate to rerun the analysis for the joint distribution of the allele frequencies and the error model parameters when comparing the outputs to a metric of 'truth' that is also dependent upon the observation model
5. Quantification of the convergence has been added using the multivariate scale reduction factor
6. Extra sections added to the discussion highlighting the fundamental underpinning of allele frequency estimates to many analyses in population genetics. Reiterate the improvements that the described method provides to the field
7. Typographical errors corrected

Chapter 3

1. Errors in mathematical notation has been corrected

2. The zebrafinch dataset currently has an embargo on it until it is published. When it becomes publically available, I will distribute it with the `ecomodtools` package for the R statistical platform on the R-FORGE repository

3. Added analysis of the observation model error estimates and parentage under vague and informative priors

4. Typographical errors corrected

Chapter 4

1. Algorithms rewritten for clarity

2. Corrections made to algorithms that were mis-specified

3. The kernel smoothing particle filter replaced with a Reversible Jump MCMC particle filter that has substantially reduced bias

4. Typographical errors corrected

Chapter 6

1. The last section has been expanded to give a more thorough example of a situation where employing all the methods described in this thesis would bring extra benefit. I have used the example of the cane toads invasion of Australia

2. Typographical errors corrected

Other Corrections

1. Computational code for all chapters has been included as part of the `ecomodtools` package for the R statistical platform on the R-FORGE repository. This can be accessed at `https://r-forge.r-project.org/projects/ecomodtools/`

2. An extra chapter has been added for the scaling up of dispersal from individual-level dispersal kernels to population-level measures of habitat connectivity

3. Upon initial analysis, the Melancholy thistle dataset had some very complex issues to address. Lots of evidence of clonal as well as seed dispersal. Needed to develop a framework to account for these phenomena and this looked likely to be far too complex for used as a simple example of deriving estimates of dispersal kernels from paternity analysis. I have instead included an addendum to chapter 3 that describes the mathematical framework for the estimation of population parameters from parentage data. This should provide the necessary link to chapter 4

General Abstract

A current trend in population biology is the increasing realisation of the effect of individual variability on some of the big patterns of population dynamics. Simultaneously, the field of population genetics continues to develop a sophisticated theoretical basis for the inference of large-scale population dynamics from information derived from the smallest ecological unit, that of the gene.

This thesis aims to contribute to the synthesis of these two fields by outlining a series of novel methods that can be used in the scaling up of genetic information to individual dynamics, and, eventually, to inference of patterns of the population. A critical feature of the methods described here is the preservation and propagation of uncertainty in estimates at each stage of the analysis. The thesis begins by introducing an estimation procedure for the calculation of allele frequencies when observation error means that frequencies cannot be directly observed. Genotyping errors can also prove troublesome in the field of parentage analysis, the basis of many models of inference of population-level processes. Any assignment errors made at this stage can be disastrous for any inference build upon these assignments. I describe a novel method of conducting parentage analysis, extend these methods for a series of common marker types and arbitrary ploidy, and show how uncertainty in parentage allocations can be propagated robustly to further stages of analysis.

I review a set of new methods that may prove useful for the fitting of individual-based models to real data. I describe how these methods can be applied in the context of individual-based modelling and describe an extension of the methods to efficiently handle common data used to parametrise individual-based models. I discuss that individual-based models may provide a key bridging discipline between the field of traditional population ecology and population genetics. Finally, I describe a method to use information on dispersal collected at the individual-level to inform population-level estimate of immigration and emigration rates of spatially-explicit models of population dynamics.

Acknowledgements

If writing science is the birth of ideas then at times this thesis has felt like a difficult labour. Stretching the analogy to its breaking point, I can say that this thesis would not have been possible without a competent and patient team of midwives and doctors.

Firstly, I'd like to say a big thank you to my supervisors, Calvin Dytham, Chris Thomas, Roger Butlin and Jon Bridle. They have been wonderfully supportive and patient with me and I owe them a debt of gratitude that I will never to be able to repay.

The dreaded 'write-up' year has been a particularly harsh year, a perfect storm of finance worries, work difficulties and stresses from looming deadlines. I would like to give a heartfelt thanks to those who have given me shelter during this storm, sometimes in the literal as well as metaphorical sense. In particular I'd like to thank, Pippa Gillingham and Graeme Glover, Alex Dumbrell and Hannah Lewis, Kate Somerwill, and the entire Pond family for their kind and continued support in this regard.

I have had the good fortune to work with two fantastic research groups over the course of my thesis. For this I'd like to thank the staff and students of the Fabrikschleichach field station of the University of Würzburg for their support, fantastic cooked lunches and their acceptance of my antisocial working times. I'd also like to thank the entire theoretical and whole organism ecology lab groups at the University of York. They have provided a wonderful working atmosphere with much friendly debate and humour. Stuart Priest has helped me with a number of technological disasters, averting crisis after crisis.

I would like to show my gratitude to my mother and my grandmother for providing moral support and the odd food parcel throughout my PhD. A big thank you also goes to Kate Pond, whose continued encouragement has kept me going through some of the darkest times.

Finally, I'd like to thank a friend who sadly could not be around for the completion of my PhD. Leonie Hassett has reminded me that I should always continue to strive to follow my dreams and squeeze the most out life in any circumstance. Rest in peace Nones.

Author's Declaration

I hereby declare that this thesis is my own work and effort and that it has not been previously submitted anywhere for any award. Where other sources of information have been used they have been acknowledged.

Signature: .....

Data: .....

Chapter 2 uses raw macrogametophyte data kindly provided by Professor Nathalie Isabel at Natural Resources Canada and published in Isabel *et al.* (1995) and Isabel *et al.* (1999). Sampling, extraction and amplification of the megagametophytes was performed by Professor Nathalie Isabel and her research team.

Chapter 3 uses zebrafinch SNP data kindly provided by Dr Jon Slate and Professor Tim Birkhead at the University of Shefield. All molecular assays were performed by this research group.

Chapter 4 uses molehill location data. This data was kindly provided by Dr Katja Schiffers of the Laboratoire d'Ecologie Alpine, Université Joseph Fourier / CNRS Grenoble and published in Schiffers *et al.* (2008).

A modified version of chapter 5 has been accepted for publication in the journal *Methods in Ecology and Evolution*. DOI:10.1111/j.2041-210X.2011.00117.x

Contents

List of Tables

List of Figures

CHAPTER 1

General Introduction

The application of theories of population genetics is an application of reductionism. The field of population genetics aims to infer the highest level single-species phenomena, the structure and dynamics of populations, from the smallest units of ecological information, that of molecular genetics. This can be interpreted as a two-stage model where the characteristics of individuals are inferred from their genetics and the characteristics of populations are inferred from the characteristics of individuals. Figure 1.1 illustrates how each level in the hierarchy informs the next, allowing the combination of genetic information across loci and individuals to infer properties of the population as a whole.

The study of population genetics does not demand that all population level phenomena need necessarily be reducible to molecular equivalents. Indeed, some authors such as Hull (1974) argue that this is impossible. However, the careful analysis of genetic data can reveal some facets of population structure, both present and historic. For example, population bottlenecks leave genetic fingerprints in the form of reduced genetic diversity, both in terms of reduced heterozygosity (Wright, 1931, 1938; Nei *et al.*, 1975; Chakraborty & Nei, 1977) and fewer unique allele types (Leberg, 1992; Brookes *et al.*, 1997).

Inferences made about the individual from its genetics have some level of uncertainty attached to them however. Scaling up from the individual to the population requires that we

Figure 1.1: Illustration of the direction of inference in studies of population genetics. At each stage it is necessary to propagate any uncertainty in the inference robustly.

account for all uncertainties in the inferred attributes for each individual when we derive properties about the population. For example, parentage analysis could be employed, using the observed genotypes, to assign offspring to parent pairs and calculate the breeding success for each individual (Williams & DeWoody, 2009; Gopurenko *et al.*, 2007). These estimates may be combined to form population-level estimates of growth rates and used in models of population dynamics (Kendall & Wittmann, 2010). However, if there is uncertainty in the estimates of individual reproductive success then, to assess the validity of the composite estimates of population growth rate, this uncertainty must also be incorporated into the estimates of the variation in population growth rates.

Most methods of error propagation focus on the recalculation of a variance after transformation of a random variable with a known variance. In many studies, estimates for a variable, X, are often given as the sample mean with a confidence interval, standard error, standard deviation, or some other simple function of the variance, σ_X^2. If we were to use the estimate of X to calculate another value of interest, Z, say by some sort of linear transform, then it is important to ensure that any uncertainty in the original estimate for X is carried through in calculation of the variance of Z, σ_Z^2. Say that X is known or assumed to have been drawn from a normal distribution. Consequently, the entire distribution of X can be described by the first two moments, the mean and the variance. From the known result that random variables that are linear transformations of a normally distributed random variable are themselves normally distributed, it has been shown that σ_Z^2 is also a linear transform of σ_X^2. Bevington & Robinson (2002) give a number of key results in error propagation for functions that have a number of inputs when each of the input estimates are assumed to be normally distributed.

The standard results of Bevington & Robinson (2002) have limited applicability for the sorts of problems involved in scaling up from genetic data to population parameters however. The types of model that I describe in this thesis are complex and the input variables are combined in non-linear ways. Except for a limited set of circumstances, see Goodman (1960), non-linear error propagation can usually only be achieved by using standard results on linearised approximations, such as Taylor expansions. The necessary truncation of the approximating Taylor series means that any estimates of propagated uncertainty can only be calculated approximately. Moreover, the distributions of input variables described in this project are rarely normal and so the full distribution of the input variable cannot be adequately described by the first two moments alone. Indeed, the only way to describe the uncertainty in these situations is to update the

full probability density or mass function for the variable of interest at each stage of the analysis.

This thesis describes a series of methods to update the uncertainty surrounding key attributes of a population derived from data drawn from molecular techniques. It is shown how Bayesian techniques can be employed to update the distribution of key model parameters at each stage of analysis, and, when analytical methods fail, how Monte Carlo methods can be employed to draw samples from the distribution.

1.1 Bayesian Inference

When confronted with a set of data, D, our first instinct as natural scientists is to find a model, M, that will explain the most variation that we see in the data. Either the investigator can choose from an existing suite of models or one can be constructed for this purpose. Once we have selected a model, we can then ask questions such as: 'if my model is true, what is the probability that it would generate the set of data that I have observed?'. Any given model can also have a set of parameters, θ_M, that determine its behaviour. This time our question of the data can be a little nuanced: 'if my model is true, and its parameters are equal to θ_M, then what is the probability that it would generate the set of data that I have observed?'. More formally, we are interested in deriving an equation for the probability, $\mathbb{P}(D|M,\theta_M)$. This quantity is known as the 'likelihood' (*sensu* Fisher, 1922). It is common practise to find the values of θ_M that give the maximum value for the likelihood, These are the 'maximum-likelihood estimators'.

Unfortunately the likelihood alone is not useful when the objective is to propagate parameter uncertainty. Under these circumstances the investigator is more interested in the quantity $\mathbb{P}(M,\theta_M|D)$: the probability that model M is the true model with set of parameters θ_M given the information gleaned from the data, or in Bayesian parlance, the 'posterior'. This quantity is related to the likelihood by Bayes' Theorem (Bayes, 1763) where:

$$\mathbb{P}(M,\theta_M|D) = \frac{\mathbb{P}(D|M,\theta_M)\,\mathbb{P}(M,\theta_M)}{\mathbb{P}(D)} \tag{1.1}$$

The denominator of the right-hand term of equation 1.1 is a normalising constant that ensures the probability of the posterior sums to one across all possible values for the model parameters. The numerator of the right-hand term of equation 1.1 has two is the product of the likelihood and the prior ($\mathbb{P}(M,\theta_M)$). The prior represents the probability that model M is true with

parametrisation θ_M before the data are taken into account. It can be a subject of controversy how this prior distribution is set (see Suppes, 2007), but for the most part it can be considered an amalgam of knowledge drawn from previous experiments and observations. In this sense it is possible to daisy-chain multiple Bayesian analyses together where the posterior from set of analyses can form the prior for the next set of analysis. This allows us to feed data from multiple data sources to draw inference on the same set of parameter values. We will make substantial use of Bayesian inference in this thesis as this property lends itself very well to the propagation of information from one ecological scale to the next.

Unfortunately the posterior probability density/mass function is often difficult to derive directly. This is because the normalising constant, $\mathbb{P}(D)$, cannot often be described in a closed-form. Whilst, we may not be able to describe the posterior probability density function in a useful form, we can use a number of specialised algorithms such as the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970; Chib & Greenberg, 1995) or the Gibbs sampler (Casella & George, 1992) to sample from it. Chapters 2, 3, and 4 all describe specific sampling algorithms tailored to their particular application.

1.2 Allele Frequencies and Population Structure

1.2.1 Hardy-Weinberg Frequencies

One of the key elements of classical population genetics is the description of allele frequencies in and between populations. Early pioneers in the field, Hardy (1908) and Weinberg (1908, translated in Weinberg 1963), were the first to describe the expected distribution of alleles in a population where allele proportions and genotype proportions are in equilibrium. They describe a diploid system with two alleles, A_1 and A_2, with proportional frequencies, f_1 and f_2 respectively. They show that if alleles are allowed to recombine freely then we would expect the genotype frequencies to follow

$$q_{A_1 A_1} = f_1^2 \tag{1.2}$$

$$q_{A_1 A_2} = f_1 f_2 \tag{1.3}$$

$$q_{A_2 A_2} = f_2^2 \tag{1.4}$$

where $q_{A_1 A_1}$, $q_{A_1 A_2}$, and $q_{A_2 A_2}$ are the relative genotype frequencies of an A_1 homozygote, a heterozygote, and an A_2 homozygote respectively.

By defining the vector $\mathbf{G}$ to be a genotype frequency vector with each element, G_i, representing the quantity of allele type i present in the genotype, it is then possible to generalise the Hardy-Weinberg law to incorporate polyallelic and polyploidal systems. Here $q_{\mathbf{G}}$ is the expected proportion of genotype frequency vector $\mathbf{G}$ found in the population under conditions of Hardy-Weinberg equilibrium. The assumption of random mating and assortment of alleles present in Hardy-Weinberg equilibrium means that the genotype frequency vector can be assumed to be drawn from a multinomial distribution with probability vector parameter equal to the vector of relative frequencies present in the population, $\mathbf{f}$, such that

$$q_{\mathbf{G}} = \begin{cases} \frac{C!}{\prod_i \mathrm{G}_i!} \prod_i \mathrm{f}_i & \sum_i \mathrm{G}_i = C \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

where C is the system ploidy.

The application of Hardy-Weinberg equilibrium concepts makes many assumptions about the population however. Sampling for the genotype frequency vector $\mathbf{G}$ is taken with replacement and so for the multinomial approximation to hold then the population must be large. If the population drops to a small size, sampling alleles for $\mathbf{G}$ from the population with replacement becomes a bad approximation to the process, the genotypes of individuals can no longer be assumed independent, and the distribution of a single individuals genotype, conditional on the set of genotypes present in the rest of the sampled individual, instead follows a multivariate hypergeometric distribution. Alleles must be randomly mixing and so must not exhibit any significant frequency stratification, such as in sub populations (Wahlund, 1928, translated in Wahlund 1975) or sexual segregation (although see extensions detailed in Moree, 1950; Stark, 1976). There are a number of phenomena that could disrupt Hardy-Weinberg equilibrium such as assortative mating, immigration and emigration, mutation, or selection acting on the phenotype associated with the alleles. This sensitivity of the theory to common assumption violations has prompted the development of a number of formal statistical tests for deviation from Hardy-Weinberg equilibrium (Wigginton *et al.*, 2005; Guo & Thompson, 1992; Emigh, 1980).

1.2.2 Wright's F Statistics

Wright (1951) describes a 'fixation index' to describe divergences from Hardy-Weinberg equi-

librium. Here, departures from the theoretical equilibrium are exemplified by differences in the expected homozygote to heterozygote ratio. In the simple biallelic diploid case we can respecify the derivation of $q_{A_1\,A_1}$, $q_{A_1\,A_2}$, and $q_{A_2\,A_2}$:

$$q_{A_1\,A_1} = (1 - F)\, f_1^2 + F f_1 \tag{1.6}$$

$$q_{A_1\,A_2} = (1 - F)\, 2 f_1 f_2 \tag{1.7}$$

$$q_{A_2\,A_2} = (1 - F)\, f_2^2 + F f_2 \tag{1.8}$$

where F is the fixation index. The value of F has an upper value of 1 but a complex lower bound that is dependent upon the allele frequencies (Zhivotovsky, 1999, and chapter 2) but is in all cases less than or equal to zero. As F tends towards 1 the relative frequency of homozygotes increases and eventually results in complete fixation, where no heterozygotes exist in the population, indicative of an inbred population. Conversely, values for F that are less than zero produces an excess of heterozygotes. In this sense it is possible to interpret deviations of F from zero as deviations from Hardy-Weinberg equilibrium (Wright, 1922).

A common tactic in population biology is to use allele frequencies and zygosities to find or show genetic structure within a population. For example, suppose that there exists *a priori* a criterion, or set of criteria, with which a population can be feasibly subdivided. The next step of the analysis would involve testing the putative subdivisions for genetic differentiation. Wright (1951) developed this idea further, defining three related fixation indices for the examination of population structure: F_{IS}, F_{ST}, and F_{IT}. F_{IS} is a measure of the within-sub population divergence from Hardy-Weinberg equilibrium, a measure of the zygosity of individuals relative to the sub population. F_{ST} describes the zygosity of the sub populations relative to the total population. Finally, F_{IT} is defined as the a measure of relative zygosity from individual to total population. Unfortunately multiple definitions of the so-called F-statistics, particularly F_{ST}, has led to some confusion in the literature. For clarity, we use the more widely accepted definition of Holsinger & Weir (2009). Under this definition, the F-statistics of Wright (1951) are linked to the genotype frequencies in sub populations S_1 and S_2 according to the following

relationships:

$$q_{A_1 A_1}^{(S_1)} = \left(1 - F_{IS}^{(S_1)}\right) {f_1^{(S_1)}}^2 + F_{IS}^{(S_1)} f_1^{(S_1)} \tag{1.9}$$

$$q_{A_1 A_2}^{(S_1)} = \left(1 - F_{IS}^{(S_1)}\right) 2 f_1^{(S_1)} f_2^{(S_1)} \tag{1.10}$$

$$q_{A_2 A_2}^{(S_1)} = \left(1 - F_{IS}^{(S_1)}\right) {f_2^{(S_1)}}^2 + F_{IS}^{(S_1)} f_2^{(S_1)} \tag{1.11}$$

$$q_{A_1 A_1}^{(S_2)} = \left(1 - F_{IS}^{(S_2)}\right) {f_1^{(S_2)}}^2 + F_{IS}^{(S_2)} f_1^{(S_2)} \tag{1.12}$$

$$q_{A_1 A_2}^{(S_2)} = \left(1 - F_{IS}^{(S_2)}\right) 2 f_1^{(S_2)} f_2^{(S_2)} \tag{1.13}$$

$$q_{A_2 A_2}^{(S_2)} = \left(1 - F_{IS}^{(S_2)}\right) {f_2^{(S_2)}}^2 + F_{IS}^{(S_2)} f_2^{(S_2)} \tag{1.14}$$

$$q_{A_1 A_1}^{(T)} = \pi^2 + F_{IT} \pi \left(1 - \pi\right) \tag{1.15}$$

$$q_{A_1 A_2}^{(T)} = 2\pi \left(1 - \pi\right) \left(1 - F_{IT}\right) \tag{1.16}$$

$$q_{A_2 A_2}^{(T)} = \left(1 - \pi\right)^2 + F_{IT} \pi \left(1 - \pi\right) \tag{1.17}$$

$$\left(1 - F_{IT}\right) = \left[1 - c F_{IS}^{(S_1)} - \left(1 - c\right) F_{IS}^{(S_2)}\right] \left(1 - F_{ST}\right) \tag{1.18}$$

where $\pi = c f_1^{(S_1)} + \left(1 - c\right) f_1^{(S_2)}$ and c is the proportion of individuals sampled from sub population S_1. $q_{\cdot}^{(S_1)}$, $q_{\cdot}^{(S_2)}$, and $q_{\cdot}^{(T)}$ represent the relevant genotype proportions at sub population S_1, sub population S_2, and the total population (T) respectively. $F_{IS}^{(S_1)}$ and $F_{IS}^{(S_2)}$ are the sub population specific inbreeding coefficients.

A number of alternative test statistics for fixation have also been proposed (Slatkin, 1995; Spitze, 1993; Excoffier *et al.*, 1992; Nei, 1973; Fisher, 1949; Pearl, 1917) but all are closely related to the F-statistics described here (Holsinger & Weir, 2009; Wright, 1951). Moreover, all statistics used to infer population structure require the accurate assessment of allele frequencies and genotypes.

The genotype, the underlying genetic basis of an individual, is never directly available however. The visible expression of the genotype is referred to as its phenotype. The usage of the term phenotype is often considered to be related to the physical attributes of the individual concerned but, in the realm of population genetics, this term is usually used more specifically to relate to the visible alleles in a genetic assay. From a statistical standpoint it is apt to consider

the genotype as a hidden state which expresses itself through an imperfect observation process as the phenotype.

One commonly encountered form of imperfect observation comes in the form of 'null' alleles. Null alleles are alleles that can only be observed in homozygous individuals (which we will henceforth refer to as 'homozygous nulls'). In hetorozygous individuals with at least one null allele preset (henceforth referred to as null heterozygtes), the presence of a null allele can make the individual appear homozygous for the non-null allele when it is in fact cryptically heterozygous. Dominant markers such as Amplified Fragment Length Polymorphisms (AFLPs; Vos *et al.*, 1995) or Random Amplified Polymorphic DNA (RAPDs; Williams *et al.*, 1990) only have two allele types denoting a band presence (positive allele) or a band absence (negative allele) at a given locus. If a positive allele is present then a band appears at the respective place on the assay. It is therefore impossible to distinguish a heterozygote individual from a homozygote individual with multiple copies of the positive allele when employing these marker types. In this sense the negative allele is a 'null' allele although for these marker types the term 'recessive' is more commonly used.

Even markers that are normally codominant, such as microsatellites or Single Nucleotide Polymorphisms (SNPs), can include 'null' alleles (Dakin & Avise, 2004). These alleles usually result from a mutation at a primer binding site resulting in a failure to amplify the respective marker. A number of methods exist for the estimation of null allele frequencies in dominant markers (see Foll *et al.*, 2008; Holsinger *et al.*, 2002; Hill & Weir, 2004; Zhivotovsky, 1999; Lynch & Milligan, 1994; Stewart Jr & Excoffier, 1996). Chapter 2 extends these methods to also incorporate allele frequency estimations for codominant, polyallelic markers that can be used to directly estimate fixation indices. Chapter 2 also includes an option to incorporate other forms of genotyping error into estimates of allele frequency, thus allowing for uncertainty related to both genotyping errors and dominancy to be included in estimates of allele frequencies, and, furthermore, in statistics that rely on these estimates of allele frequencies.

1.3 Parentage Analysis and Population Parameter Estimation

The use of methods of parentage analysis to derive information regarding population-level relevant parameters is not a new field. Many otherwise cryptic facets of a species' breeding

strategy can be diagnosed with the application of paternity analysis, including, but not limited to, the frequency of extra-pair copulations (Zheng *et al.*, 2010; Wojczulanis-Jakubas *et al.*, 2009; McEachern *et al.*, 2009; Uller & Olsson, 2008; Simmons *et al.*, 2007), diagnosis of assortative mating (Bos *et al.*, 2009), and variances in breeding success (Seamons & Quinn, 2010; Doerksen & Herbinger, 2008; Tatarenkov *et al.*, 2008). These distributions of individual breeding success can be used to infer effective population size (Bouteiller & Perrin, 2000; Hill, 1972), or included as a variable in more complex models of breeding strategy (for example Vanpé *et al.*, 2009a).

If information pertaining to the location of samples is also available to the investigator then it is also possible, once parentage is established, to formulate model to describe the geographical spread of genetic information (Saenz-Agudelo *et al.*, 2009). This application of parentage analysis allows assessment of the inter-generational movement between patches in a network of habitats (Botsford *et al.*, 2009; Saenz-Agudelo *et al.*, 2009; Stow & Sunnucks, 2004). If accurate point-to-point distance estimates can be made between the parent individuals and the offspring, then it is possible to construct a probability distribution of dispersal distances from the mother to the offspring (Broquet & Petit, 2009; Robledo-Arnuncio & Garca, 2007), otherwise known as the (maternal) dispersal kernel. It is also possible to calculate the distance over which paternal contributions can be made, such as the distance of pollen dispersal in plants (Robledo-Arnuncio & Gil, 2005).

Paternity analysis methods falls into three broad categories: exclusion methods, likelihood methods, and fractional methods. Exclusion methods exclude potential parent pairs based on genotype incompatibilities between the putative parents and the offspring. Exclusion methods fail when the number of loci are few, or exhibit low polymorphism as to be insufficient to exclude all but one parent pair. Conversely, mutations, genotype errors, or the presence of recessive alleles may result in the erroneous exclusion of the true parent pair under these methods (Cifuentes *et al.*, 2006). Likelihood methods such as those employed by Meagher & Thompson (1986) or Marshall *et al.* (1998) allow the weighting of the non-excluded parent pairs based on the probabilities of the observed parental genotypes resulting in the observed offspring genotype. However, these methods when implemented are often simply used to assign paternity to the most likely parent pair. This ignores the uncertainty associated with the parentage assignment and may bias subsequent analyses for which parentage assignments form the basis.

Fractional methods provide the best mechanism for retaining uncertainties in parentage as-

signments. Here no absolute assignment is made and different parentage pairs are weighted according to their likelihood (Devlin *et al.*, 1988) or the posterior probability output from a Bayesian analysis such as that implemented in the R package MasterBayes (Hadfield *et al.*, 2006). This allows the investigator to weight the conclusions of subsequent analyses according to the probabilities of the set of parentage assignment on which they are based.

Whilst many of the current methods of parentage allow for some degree of genotyping error (see Jones & Ardren, 2003), most employ observation models that are only suitable for codominant markers. Moreover, even if codominant markers are being used, very few methods of parentage analysis allow for the incorporation of null alleles. Chapter 3 describes a series of observation models to link a true genotype to an observed phenotype. This allows the implementation of a new method of fractional parentage analysis that allows for the joint inference of parentage from mixtures of different marker types for systems of arbitrary ploidy.

1.4 Individual-Based Models in Population Modelling

We have already discussed how estimates of indices of population structure are determined by estimates of allele frequencies, for which an example allele frequency estimation framework is described in chapter 2. We have also discussed how parentage analysis techniques, such as those described in chapter 3, can be employed to derive estimates of population parameters. The next step is to produce a method by which this information can be used to model population dynamics.

One method is to simply enter the garnered values of the parameters into models of population dynamics, ensuring that sensitivity of the outputs over the credible range of parameter estimates given their uncertainty is taken into account. Another method gaining popularity is individual-based modelling (Łomnicki, 1999; Grimm *et al.*, 1999; DeAngelis *et al.*, 1994). Here individuals are described separately, either mathematically or, more commonly, in the context of a simulation model, and population-level phenomenon emerge from the interactions of individuals with themselves and the environment. The draw of individual-based modelling comes with the ability to model much more of the complexity of the system. For example, the Soda model of Bennett *et al.* (2009) permits the assessment of human disturbance on relatively complex individual-level behaviour that would be difficult to assess using standard population models.

The added complexity of this so-called individual-based ecology is also its curse however. The data generation mechanism of individual-based models cannot typically be expressed in a closed analytical form and so the probability of observing a data set given a set of parameter values, the likelihood, is difficult to calculate. The combination of the absence of a likelihood function and the fact that individual-based models commonly have a higher parameter load than their simple analytical equivalents means that they are notoriously difficult to fit to real data and analyse (Murdoch *et al.*, 1992; Beissinger & Westphal, 1998). Some authors have suggested a 'pattern-orientated' approach (Grimm *et al.*, 1996; Wiegand *et al.*, 2003; Grimm *et al.*, 2005) where potential parameter values are filtered according to their ability to reproduce patterns of interest in the data. Pattern oriented modelling has yet to adopt a rigorous statistical framework however, and currently there exists no mechanism for the assessment of the relative probability of different parameter values. Some promise has been made on this front with the advent of methods to apply approximate Bayesian techniques to models for which there is no likelihood (see Beaumont *et al.*, 2002; Sisson *et al.*, 2007; Marjoram *et al.*, 2003; Toni *et al.*, 2009). Chapter 4, introduces these methods to the field of individual-based modelling and extends these methods to incorporate scenarios where dynamics models are being fitted to a data time series.

1.5 Scaling up Further to Metapopulation Models

Individual-based models can provide very useful and detailed descriptions of how individuals move and interact with their environment but their application to large scale problems can be limited. At the very large scales it can be computationally prohibitive to simulate enough individuals to truly represent the population of interest. Under these situations it is common for the investigator to rephrase the individual-based model as a metapopulation model or a lattice-based population model. If however you have used the techniques outlined in chapters 2, 3, and 4 to produce a wonderfully parameterised individual-based model, then how can you make use of this parameterisation when stepping up to the next ecological scale?

Some parameters are easier to scale-up than others. Estimates of individual-level fecundity and mortality can be related to the growth parameter in most population growth models (see Law *et al.*, 2003; Murrell *et al.*, 2004). Dispersal can be a very difficult process to scale appropriately however. Firstly a dispersal process that is specified at the level of the individual,

and in continuous space, can be difficult to translate into a dispersal mechanism described in terms of an artificial geometry placed upon the landscape (Holland *et al.*, 2007). Whilst lattice-based models do exist (Chesson & Lee, 2005), these models are not described in terms of individual movements and it may be difficult to see how the parameters of the two models match. Chapter 5 describes four different methods for the approximation of individual-level and continuous-space dispersal on a grid. The chapter details how to use the derived cell-to-cell transition probabilities to estimate connectivity between nodes in a metacommunity, thus providing the last step in scaling from gene to population.

CHAPTER 2

The unknown genotype: estimating recessive allele frequency in the presence of observation error

Summary

1. The non-expression of recessive alleles in the presence of their dominant counterparts results in the presence of cryptic heterozygotes.

2. High incidence of cryptic heterozygotes can result in significant biases in the calculation of indices of population structure. Moreover, zygosity misdiagnosis can result in the erroneous exclusion of true parentage pairs in paternity analysis.

3. Methods exist to estimate the frequencies of the recessive allele type but none take into account the extra complications arising from errors of genotyping.

4. We describe here a method for calculating allele frequencies for alleles with expression hierarchies, even in the presence of genotyping error. The method is applied to the expressed phenotype of RAPD megagametophyte data for the eastern white pine (*Pinus strobus*) and black spruce (*Picea mariana*) and for which individual haploid runs can provide accurate estimates of recessive allele frequency.

5. The methods described in this chapter provide reliable estimates of the null allele frequency over a range of different genotyping error rates. Unlike existing methods, we show

that our estimates are accurate even when the null allele frequency is low.

6. We discuss possible extensions to the algorithm for the joint calculation of the inbreeding coefficient, F_{IS}, and outline some of the difficulties in achieving this.

2.1 Introduction

Molecular methodologies have far expanded the potential field of inference for ecological problems. The otherwise cryptic assessment of differentiation between populations, made possible by the tools developed in the now mature discipline of quantitative population genetics, offers insights into dispersal and historic vicariance events and allows estimation of the level of inbreeding and assortative mating occurring within populations. Where sufficient data are available, such examinations can be supplemented by the use of parentage analysis techniques: the assignment of paternity, and often jointly, maternity, to an offspring can elucidate the mechanisms that drive the geographic diffusion of genes. This in turn allows differentiation of the contributions of gamete transfer and post-natal dispersal to gene flow in addition to the estimation of other population parameters critical to conservation biology (Haig, 1998).

The application of such methods can be problematic when some individuals have genotypes containing alleles that are not expressed using conventional molecular techniques. Biallelic markers, such as Amplified Fragment Length Polymorphism (AFLPs: Vos *et al.*, 1995) or Random Amplified Polymorphic DNA (RAPD: Williams *et al.*, 1990), exhibit one of only two possible phenotypes: either a band is present at a given loci, or it is absent. At the genotypic level, the allele responsible for band presence is dominant and a positive phenotype can, in the absence of observation error, arise from both homozygous positive and heterozygous genotypes. Only a homozygous negative genotype produces an absent band phenotype. Even co-dominant markers such as microsatellites may still exhibit a 'null' allele which only becomes expressed in the homozygous case (Dakin & Avise, 2004).

Null alleles can cause a number of problems in paternity analysis. The simulations of Dakin & Avise (2004) show that, for realistic incidence rates of null alleles, the confusion of null heterozygotes for non-null homozygotes will result in a small under-estimation of exclusion power. More importantly, the misdiagnosis of a heterozygous offspring with one null allele (or more in a polyploidal system) as a non-null homozygote could result in the erroneous exclusion of a parent pair if both were also null heterozygotes but also misdiagnosed as differing non-null

homozygotes. The potential for null allele presence to incorrectly assign parentage is worrying, indeed Dakin & Avise (2004) state probabilities of excluding an actual parent as high as 15% when null allele frequencies were are around 20%: the upper end of null allele frequency reported in the literature they review. Commonly used paternity analysis software such as CERVUS (Marshall *et al.*, 1998), FAMOZ (Gerber *et al.*, 2003), PAPA (Duchesne *et al.*, 2002) and PARENTE (Cercueil *et al.*, 2002), do not treat null alleles as a special case and assume that the frequency is sufficiently low as to not affect assignment, or argue that the genotyping error models allow for enough flexibility to counteract a modest null allele load (see Jones & Ardren, 2003). CERVUS does include a method of assessing loci for the presence of null alleles by calculating departures from Hardy-Weinberg equilibrium (Marshall *et al.*, 1998) but such departures can also arise from many phenomena unrelated to null allele frequency including inbreeding and Wahlund effects (Chakraborty *et al.*, 1992; Dakin & Avise, 2004).

Several metrics for the assessment of population structure from genetic data require approximations of allele frequency. The popular F_{ST} statistic of Wright (1951), for the calculation of population subdivision, requires knowledge of mean allele frequencies across all populations as well as the variance between them (Weir & Cockerham, 1984). Similarly, the genetic distance statistic, D, of Nei (1972), the basis of which was later developed into the statistics D_{ST}, G_{ST} and R_{ST} among others (Nei, 1973), require calculations relating to the probability of picking identical alleles from pools of potential alleles for which estimates of allele frequencies are essential. In the case of null alleles, simply adding up the number of observed homozygotes, ignoring a potentially large number of cryptic alleles present in heterozygotes, and multiplying the total by the ploidy will obviously result in downward biases in allele frequency estimates, and hence, biases in the calculation of statistics of population structure.

An early attempt to estimate null allele frequency, as applied in Stewart Jr & Excoffier (1996), involved simply taking the square root of the null homozygote frequency. This technique, in its most basic form, requires the assumption of Hardy-Weinberg equilibrium, but, if an estimate of the inbreeding coefficient is known, then it is possible to extend this method to populations exhibiting significant inbreeding (or outbreeding in the case of negative inbreeding coefficients). The main problem with this technique is that, especially at low allele frequencies, estimates tend to exhibit a strong downward bias. Lynch & Milligan (1994) present an asymptotically unbiased estimator for null allele frequency, but this requires the exclusion of loci with less than three null homozygotes, creating a sampling bias of loci that exhibit high

null allele frequencies (Isabel *et al.*, 1995, 1999; Szmidt *et al.*, 1996; Zhivotovsky, 1999). More recent estimators fare much better: the popular estimator described by Zhivotovsky (1999) and demonstrated on the dataset of Isabel *et al.* (1995) was shown to exhibit a much reduced bias compared to the estimators of Lynch & Milligan (1994) and Stewart Jr & Excoffier (1996), although the relative performance of all estimators is not always pronounced for all data sets (see Krauss, 2000). Further developments, such as the moment estimation technique of Hill & Weir (2004) and Markov Chain Monte Carlo (MCMC) sampler of Holsinger *et al.* (2002), have incorporated the joint estimation of both the null allele frequency and indices of population structure. In some cases the Holsinger *et al.* (2002) estimator has been shown to perform quite poorly for null allele estimation but a reformulation by Foll *et al.* (2008) may have gone some way to correcting these biases.

The calculation for null allele estimation is further confounded by the incidence of errors in the genotype scoring process. Aside from the incidence of true null homozygotes, where mutations arising at primer annealing sites result in non-amplification (Kwok *et al.*, 1990), there are many ways in which insidious fake null homozygotes can arise. Insufficient quantity or poor quality template can, in some cases, lead to a failed amplification and incorrect diagnosis of a null-homozygote (Dakin & Avise, 2004). None of the null allele frequency estimation techniques described above take into account genotyping error. Although most parentage analysis programs incorporate some form of genotyping error (Jones & Ardren, 2003), those that attempt to identify loci with high null allele incidence do not allow for such error when calculating departure from Hardy-Weinberg equilibrium.

We present here a theoretical extension of the work of Zhivotovsky (1999) to allow the calculation of null allele frequencies in the presence of genotyping error. This extension allows for estimation of not only of biallelic recessive allele frequencies, but also null allele frequencies in codominant markers. We describe how to robustly propagate uncertainty in such estimates to indices of population structure, allowing for departures from Hardy-Weinberg equilibrium, even when uncertainty exists in the coefficient of inbreeding.

2.2 Materials and Methods

2.2.1 Genotype Probabilities

We first define the vector, $\mathbf{f}_j$, as an allele frequency vector with each element, f_{j_a}, equal to the proportional frequency of allele a at locus j of the source population from which the samples have been made. The vector $\mathbf{f}_j$ is a list of exhaustive possible allele frequencies including the null allele such that $\sum_a \mathrm{f}_{j_a} = 1$. From this vector it is then possible to calculate the probability that an individual, i, drawn at random from a large population in Hardy-Weinberg equilibrium (HW) exhibits genotype, G_{ij}, at locus j:

$$\mathbb{P}_{HW}\left(G_{ij}\right) = \prod_{k=1}^{C_j} \sum_a \mathrm{f}_{j_a} \omega_a\left(v_{ijk}\right) \tag{2.1}$$

$\omega_a\left(v_{ijk}\right)$ is an indicator function which equals one when the allele at position k is the same as the value of allele a, and zero at all other times. Here the 'value' of an allele is left to be defined appropriately to the marker system used in the analysis: in dominant, biallelic marker systems the usual nomenclature is to denote the dominant allele with a + and the recessive allele with a −. In codominant, polyallelic marker systems, such as microsatellites and RFLPs, it is usual that allele values be simply represented by the fragment length. C_j denotes the ploidy of locus j. In most analyses C_j will be held constant between loci.

Continuing from both Lynch & Milligan (1994) and Zhivotovsky (1999), we extend the calculation of expected diploid genotype frequencies for populations that exhibit some degree of departure from Hardy-Weinberg equilibrium in the biallelic case to also include polyallelic marker systems. In a biallelic marker with alleles M and m, we start with the assertion that the probabilities of selecting an individual at random in a large inbreeding or outbreeding (IO) source population with the each of the genotypes below are the following:

$$\mathbb{P}_{IO}\left(G_{ij} = MM\right) = \left(1 - F_{IS}\right)\mathbb{P}_{HW}\left(G_{ij} = MM\right) + \mathrm{f}_{j_M} F_{IS} \tag{2.2}$$

$$\mathbb{P}_{IO}\left(G_{ij} = Mm \cup G_{ij} = mM\right) = \left(1 - F_{IS}\right)\mathbb{P}_{HW}\left(G_{ij} = Mm \cup G_{ij} = mM\right) \tag{2.3}$$

$$\mathbb{P}_{IO}\left(G_{ij} = mm\right) = \left(1 - F_{IS}\right)\mathbb{P}_{HW}\left(G_{ij} = mm\right) + \mathrm{f}_{j_m} F_{IS} \tag{2.4}$$

Here, F_{IS} is the coefficient of inbreeding. Values of F_{IS} approaching unity denote populations exhibiting extreme inbreeding and heterozygote deficiency. Values of F_{IS} that are negative may include outbred populations with an excess of heterozygotes. A source population in Hardy-

Weinberg equilibrium is a special case of the inbreeding/outbreeding model with $F_{IS} = 0$. From this basis it is a simple conceptual step to include polyallelic marker systems:

$$\mathbb{P}_{IO}\left(G_{ij}\right) = \begin{cases} \left(1 - F_{IS}\right)\mathbb{P}_{HW}\left(G_{ij}\right) + \mathrm{f}_{j_a} F_{IS} & \text{if homozygous for allele } a \\ \left(1 - F_{IS}\right)\mathbb{P}_{HW}\left(G_{ij}\right) & \text{if heterozygous} \end{cases} \tag{2.5}$$

which, using the result of equation 2.1, equates to the following:

$$\mathbb{P}_{IO}\left(G_{ij}\right) = \begin{cases} \left(1 - F_{IS}\right)\left(\mathrm{f}_{j_a}\right)^{C_j} + F_{IS}\mathrm{f}_{j_a} & \text{if homozygous for allele } a \\ \left(1 - F_{IS}\right)\prod_{k=1}^{C_j}\sum_{a}\mathrm{f}_{j_a}\omega_a\left(v_{ijk}\right) & \text{if heterozygous} \end{cases} \tag{2.6}$$

The lower bound of F_{IS} is related to the allele frequency vector, $\mathbf{f}_j$, of each locus. To avoid negative homozygote probabilities and heterozygote probabilities greater than one when F_{IS} is negative, any potential values for $\mathbf{f}_{ij}$ and F_{IS} must conform to certain restrictions. To ensure that $\mathbb{P}_{IO} \geq 0$ for the homozygous case of allele a at locus j of sample i then

$$F_{IS} \geq -\frac{\left(\mathrm{f}_{j_a}\right)^{C_j - 1}}{1 - \left(\mathrm{f}_{j_a}\right)^{C_j - 1}} \tag{2.7}$$

Equally, to ensure that $\mathbb{P}_{IO} \leq 1$ in the heterozygous case, the following inequality must also hold:

$$F_{IS} \geq 1 - \frac{1}{\prod_{k=1}^{C_j}\sum_{a}\mathrm{f}_{j_a}\omega_a\left(v_{ijk}\right)} \tag{2.8}$$

It is important to note that, under this specification, genotype probabilities are location specific; a heterozygote with genotype Mm is not the same as mM and to calculate the probability of either genotype occurring, it is necessary to sum the relevant probabilities such that:

$$\mathbb{P}.\left(G_{ij} = Mm \cup G_{ij} = mM\right) = \mathbb{P}.\left(G_{ij} = Mm\right) + \mathbb{P}.\left(G_{ij} = mM\right) \tag{2.9}$$

$\mathbb{P}.\left(x\right)$ denotes the probability of the genotype, x, being drawn from a large population in either under Hardy-Weinberg equilibrium (HW), or, from the more general inbreeding/outbreeding source population (IO).

2.2.2 Including Observation Error

Equations 2.1 and 2.5 describe the likelihood of a genotype at a given locus. However, even in the absence of genotyping errors, the existence of null alleles ensures that the observed phenotype is not always an accurate representation of the genotype. Indeed, for any given phenotype there

may be a number of possible genotypes that will provide the same observation. In order to calculate the likelihood of observing a phenotype, O_{ij}, at allele j and individual i given the allele frequencies of the source population and the inbreeding coefficient it is therefore necessary to integrate over all genotype possibilities. If we let H_j denote the set of possible genotypes then:

$$\mathbb{P}\left(O_{ij}|\mathbf{f}_j, F_{IS}, \boldsymbol{\beta}_{ij}\right) = \sum_{G_{ij} \in H_j} \mathbb{P}\left(O_{ij}|G_{ij}, \boldsymbol{\beta}_{ij}\right) \mathbb{P}\left(G_{ij}|\mathbf{f}_j, F_{IS}\right) \tag{2.10}$$

where $\mathbb{P}\left(G_{ij}|\mathbf{f}_j, F_{IS}\right)$ is the likelihood of drawing genotype G_{ij} from a large population with allele frequencies, $\mathbf{f}_j$, and inbreeding coefficient F_{IS}. Crucially, the probability of observing a phenotype, O_{ij}, given the 'true' genotype, G_{ij}, or $\mathbb{P}\left(O_{ij}|G_{ij}\right)$, is the observation model. Here it is possible to specify any mechanism that results in partial observation of the genotype such as the masking of recessive alleles, or genotype scoring errors. The vector, $\boldsymbol{\beta}_{ij}$, is a set of parameters for the observation model for the marker at locus j of sample i. In most analyses the observation model will not be locus or sample specific; the value of $\boldsymbol{\beta}_{ij}$ will not vary between samples and loci. We include the notation here to allow departure from this simple case for when markers of different types are to be jointly analysed or, in the rare case where particular samples are known *a priori* to exhibit different error rates.

For biallelic markers in systems of low ploidy the set of possible genotypes at locus j, H_j, is very small. For example, a diploid AFLP marker has only four genotypic states: $++$, $+-$, $-+$ and $--$. In general, the number of genotype combinations (N_j) possible under any given marker system at locus j is given by $N_j = A_j^{C_j}$, where the allele total, A_j, includes any null alleles. However, it is not inconceivable that the set of genotypes can, for some markers, become very large and theoretically infinite. For marker systems where alleles are defined by their fragment sizes, there is no upper bound to the number of alleles. Although theoretically this may be the case, practically it is only possible to observe fragment sizes defined within the bounds that the experimental protocol defines. Moreover, the alleles present in the sampled population are limited to those present in the source population. Obviously, it is impossible to know the full set of alleles present in the source population, but, for practical purposes, it may be sufficient to assume that all possible alleles are present in the sampled population. An alternative way to address this problem is to include an extra co-dominant allele type that represents all unsampled alleles in the source population and estimate the frequency of the unsampled allele like any other allele.

By assuming that each individual i is an independent sample, and that, within this individual, the genotypes at each locus are independent from the genotypes at the other loci, it is possible to calculate the joint likelihood of observing the entire set of phenotypes across all samples and loci, O. If we let $\mathbf{f}$ represent the set of allele frequencies across all loci, $\{\mathbf{f}_1, \ldots, \mathbf{f}_j, \ldots, \mathbf{f}_L\}$, and $\boldsymbol{\beta}$ represent the set of observation model parameters across all samples and loci, $\{\boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{ij}, \ldots, \boldsymbol{\beta}_{SL}\}$, where L is the number of loci sampled, and S, the number of samples, then:

$$\mathbb{P}\left(O|\mathbf{f}, F_{IS}, \boldsymbol{\beta}\right) = \prod_j \prod_i \mathbb{P}\left(O_{ij}|\mathbf{f}_j, F_{IS}, \boldsymbol{\beta}_{ij}\right) \tag{2.11}$$

By simple application of Bayes theorem:

$$\mathbb{P}\left(\mathbf{f}, F_{IS}, \boldsymbol{\beta}|O\right) = \frac{\mathbb{P}\left(O|\mathbf{f}, F_{IS}, \boldsymbol{\beta}\right)\mathbb{P}\left(\mathbf{f}, F_{IS}, \boldsymbol{\beta}\right)}{\iiint\limits_{V_{\mathbf{f}F_{IS}\boldsymbol{\beta}}} \mathbb{P}\left(O|\mathbf{f}, F_{IS}, \boldsymbol{\beta}\right)\mathbb{P}\left(\mathbf{f}, F_{IS}, \boldsymbol{\beta}\right)\, d\mathbf{f}\, dF_{IS}\, d\boldsymbol{\beta}} \tag{2.12}$$

where $V_{\mathbf{f}F_{IS}\boldsymbol{\beta}}$ is a joint volume of integration for the parameters to be estimated. If some information, derived independently from the data used in the study, already exists on known distributions or values of model parameters then this previous knowledge can be expressed through the prior, $\mathbb{P}\left(\mathbf{f}, F_{IS}, \boldsymbol{\beta}\right)$. For the most part, little or nothing can be inferred about the model parameters before the analysis takes place and so, in most applications, the prior is set to be minimally informative. It is important to note that, whatever prior is used, the allele frequencies and the value for the inbreeding/outbreeding coefficient, F_{IS}, are not independent and must be treated as such in specification for the prior functional form. For all loci, any values for allele frequencies which do not sum to one, or values of F_{IS} that, given the allele frequencies, do not adhere to the conditions specified in equations 2.7 and 2.8, must be given a zero weight for the probability calculations to be valid.

Estimates for allele frequency, F_{IS} and parameters of the phenotype observation model can be achieved using techniques to evaluate equation 2.12. Simple methods of Monte Carlo integration (see Fishman, 1996) to approximate the denominator will not work in this instance. The requirement that allele frequencies for each locus sum to one, alongside the fact that the lower bound of F_{IS} is dependent upon these frequencies, complicates the generation of sample values for these parameters with uniform support over the volume of integration, a necessary step in Monte Carlo integration. Some flavours of Monte Carlo integration, such as those that implement importance sampling (Oh & Berger, 1993), relax the requirement to generate sample values with uniform support. Even on application of these methods, cases where there

are a large number of loci, the volume of integration is likely to be highly dimensional and the convergence of Monte Carlo methods of integration for estimation of the constant on the denominator of equation 2.12 may be slow. Where no finite bounds exist for the parameters of the observation model, it is not possible to use these techniques, regardless of loci dimensionality.

Whilst it may be difficult to compute the probability density function of the posterior distribution directly, it is possible to sample from the posterior distribution using Markov Chain Monte Carlo (MCMC) sampling (Gilks *et al.*, 1996). The Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970; Chib & Greenberg, 1995) generates samples from the target distribution by taking a set of possible parameter values, $\boldsymbol{\theta}$, and proposing new values, $\boldsymbol{\theta}^*$, according to a proposal density, $K(\boldsymbol{\theta}^*|\boldsymbol{\theta})$. Either $\boldsymbol{\theta}$ or $\boldsymbol{\theta}^*$ is chosen randomly as a sample from the posterior distribution with relative acceptance probability weighted according to the relative likelihood of observing the data using the parameters, their prior probability, and the probability that they were proposed. The algorithm is repeated, taking $\boldsymbol{\theta}$ as the most recently accepted sample each time. As the number of repeats becomes large the density of the set of samples generated from the algorithm converges on the probability density function given by the posterior distribution.

The samples given from the Metropolis-Hastings algorithm are guaranteed eventual convergence, regardless of the proposal density chosen, provided that the probability of proposing a vector of parameter values is greater than zero for all values with some posterior support. The choice of proposal density does however effect the rate of convergence and inappropriate choices for the proposal density may mean that converge becomes too slow to be computationally feasible.

2.2.3 An Observation Model for Dominant Markers

The model framework description presented here allows for the fitting of an application and/or marker specific model of observation error. Here observation error can be split into two separate components: firstly, laboratory errors such as those arising from contamination or failed amplification can produce errors in genotype assignment. Secondly, given a particular genotype, misdiagnosed or not, some marker systems exhibit different hierarchies of dominance resulting in the non-expression of recessive alleles.

We describe here a simple model for phenotype expression for dominant, biallelic markers

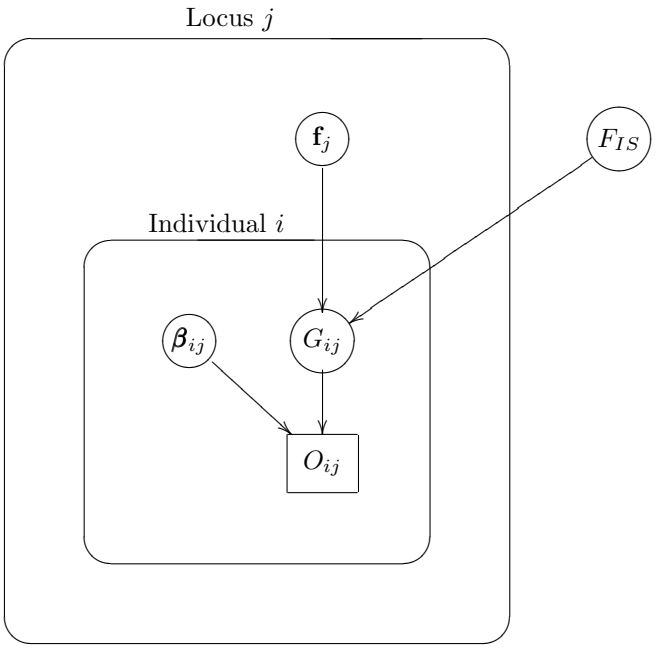


Figure 2.1: Directed acyclic graph of the modelling framework described in this paper. Circular nodes represent parameters to be estimated: F_{IS} is the inbreeding coefficient, $\mathbf{f}_j$ is the vector of relative allele frequencies, and $\boldsymbol{\beta}_{ij}$ is a vector of parameters for the observation model. G_{ij} is the true genotype of individual i at locus j whilst O_{ij} is the observed phenotype, represented as a rectangular node because data are available for the value of this parameter. Variables contained within the outer frame, are, or can be, specified uniquely to each locus whilst variables also contained with the inner frame, are, or can be, specified uniquely to each sample and locus combination.

with dominant allele, denoted by '+', and a recessive allele, denoted by '−'. We define two parameters, ϕ_j and ψ_j, that determine two sources of error. The first source of error, given by ϕ_j, is the probability of misdiagnosis of a dominant gene at locus j as a recessive gene. This can arise from non-amplification due to sample contamination by inhibitory agents (Opel *et al.*, 2010; Wilson, 1997), laboratory errors or low-quality or insufficiently populous template (Watts *et al.*, 2007; Gagneux *et al.*, 1997; Broquet & Petit, 2004; Taberlet *et al.*, 1996). ψ_j, the second source of error, is the probability of misdiagnosing a null allele as a dominant type. This error type, albeit rarer, can arise from sample contamination or errors of allele identification caused by the confusion of background fluorescence in the gels as band presence (Whitlock *et al.*, 2008).

Using this framework we can consider the number of truly positive homologous genes at locus j of sample i that are correctly identified as such, as a random variable, A_{ij+}, drawn from a binomial distribution with parameters n_{ij+}, the number of truly positive homologous genes, and trial success parameter, $1 - \phi_j$, the probability of correctly diagnosing a positive allele. In a similar vein, we define the random variable, A_{ij-}, as the number of recessive homologous genes at locus j of sample i incorrectly identified as the dominant type. A_{ij-} is assumed to be drawn independently from A_{ij+} according to a binomial distribution with parameters ψ_j, the probability of incorrectly diagnosing a recessive allele as its dominant analogue, and n_{ij-}, the number of negative homologous genes such that $C_j = n_{ij+} + n_{ij-}$. From this specification a band absence can only be observed when all truly positive homologous genes are misdiagnosed as negative ($A_{ij+} = 0$) and that all truly negative genes are correctly identified ($A_{ij-} = 0$).

So, if

$$A_{ij+} \sim \text{Bin}\left(n_{ij+}, 1 - \phi_j\right) \tag{2.13}$$

$$A_{ij-} \sim \text{Bin}\left(n_{ij-}, \psi_j\right) \tag{2.14}$$

then

$$\mathbb{P}\left(O_{ij}|G_{ij}, \phi_j, \psi_j\right) = \begin{cases} 1 - \mathbb{P}\left(A_{ij+} = 0\right)\mathbb{P}\left(A_{ij-} = 0\right) & \text{if } O_{ij} = + \\ \mathbb{P}\left(A_{ij+} = 0\right)\mathbb{P}\left(A_{ij-} = 0\right) & \text{if } O_{ij} = - \end{cases} \tag{2.15}$$

$$= \begin{cases} 1 - \phi_j^{n_{ij+}}\left(1 - \psi_j\right)^{n_{ij-}} & \text{if } O_{ij} = + \\ \phi_j^{n_{ij+}}\left(1 - \psi_j\right)^{n_{ij-}} & \text{if } O_{ij} = - \end{cases} \tag{2.16}$$

This observation model can be used in equation 2.10 to calculate the likelihood of the entire set of observation data given a set of allele frequencies, value of F_{IS}, and observation model parameters ($\boldsymbol{\beta}_{ij} = [\phi_j \quad \psi_j]^T$).

2.2.4 An Example Dataset

To assess the accuracy of the methods described here we apply the model to two data sets from which accurate estimates of allele frequencies can be made. The first dataset is that collected from 75 individuals over five natural populations of the black spruce, *Picea mariana*, spread throughout a 1000km part of its range in Québec, Canada. The laboratory protocol, sampling regime and loci selection criteria are published in Isabel *et al.* (1995). The second dataset is that of the eastern white pine, *Pinus strobus*. Again 75 individuals were sampled across five populations according to the methods published in Isabel *et al.* (1999). Further details of the sampled locations can be found in Beaulieu & Simon (1994) (locations sampled are indexed by ANT, BRO, SCH, USB, and ZEP).

In both data sets a series of RAPD markers amplified from haploid sexual tissues (6 to 8 megagametophytes) are taken from each individual sampled. Because the gametic tissue is haploid it is easier to make accurate inferences pertaining to the genotype of the sampled individual because in each megagametophyte sample, recessive alleles are expressed without possible masking from dominant alleles at the same locus. Inferences made from gamete tissues are not entirely free from error however. When the number of gametic tissue samples are few there may be sampling error associated with the inferred homozygotes. For example, in the diploid case, an individual, i, with M_{ij+} positively identified megagametophytes at locus j and no negatively identified megagametophytes ($M_{ij-} = 0$) still has a $\left(\frac{1}{2}\right)^M$ probability of being observed in a truly heterozygous individual for which the analogous allele has not been sampled.

Although in this instance it is possible to isolate and remove the aspect of observation error that relates to hierarchies of allele dominance, gamete tissue is still susceptible to the types of genotyping error discussed in this paper. Assuming the genotyping error process described in the previous section, we can assign the probability of observing a positive allele at locus j from a randomly selected megagametophyte of individual i, d_{ij+}, as the probability of sampling and correctly identifying a megagametophyte with positive genotype plus the probability of sampling a megagametophyte with a negative genotype but erroneously identifying it as positive such

that

$$d_{ij+} = \frac{Z_{ij+}}{C_j}(1-\phi_j) + \frac{C_j - Z_{ij+}}{C_j}\psi_j \tag{2.17}$$

where Z_{ij+} is the number of positive alleles in the true, non-gametic, genotype at locus j of individual i. If we assume that each sample of a megagametophyte is an independent Bernoulli trial with probability of observing a positive result, d_{ij+}, then the random variable, M_{ij+}, describing the number of observed positive megagametophytes is drawn from a binomial distribution. If we let K_{ij} be the number of megagametophytes sampled from individual i at locus j it follows that

$$M_{ij+} \sim \text{Bin}(K_{ij}, d_{ij+}) \tag{2.18}$$

and so

$$\mathbb{P}(M_{ij+}|Z_{ij+}, K_{ij}, C_j, \phi_j, \psi_j) = \begin{cases} 0 & \text{if } M_{ij+} > K_{ij+} \text{ or } M_{ij+} < 0 \\ \binom{K_{ij+}}{M_{ij+}} d_{ij+}^{M_{ij+}} (1-d_{ij+})^{K_{ij+}-M_{ij+}} & \text{otherwise} \end{cases} \tag{2.19}$$

Equation 2.19 describes the likelihood of observing the number of positively identified megagametophytes given a known zygosity. By application of Bayes theorem it is possible to use this likelihood function to infer the number of positive alleles in the genotype at locus j of individual i:

$$\mathbb{P}(Z_{ij+}|M_{ij+}, K_{ij}, C_j, \phi_j, \psi_j) = \frac{\mathbb{P}(M_{ij+}|Z_{ij+}, K_{ij}, C_j, \phi_j, \psi_j)\,\mathbb{P}(Z_{ij+})}{\sum_{k=0}^{C_j} \mathbb{P}(M_{ij+}|k, K_{ij}, C_j, \phi_j, \psi_j)\,\mathbb{P}(k)} \tag{2.20}$$

For most applications no prior information is known on the probabilities of the positive allele count at a locus and so we can use a non-informative formulation that reduces the second terms on both the numerator and the denominator, $\mathbb{P}(Z_{ij+})$ and $\mathbb{P}(k)$ respectively, to the constant $\frac{1}{C_j}$. This reduces equation 2.20 to the following:

$$\mathbb{P}(Z_{ij+}|M_{ij+}, K_{ij}, C_j, \phi_j, \psi_j) = \frac{\mathbb{P}(M_{ij+}|Z_{ij+}, K_{ij}, C_j, \phi_j, \psi_j)}{\sum_{k=0}^{C_j} \mathbb{P}(M_{ij+}|k, K_{ij}, C_j, \phi_j, \psi_j)} \tag{2.21}$$

The next step in the analysis of the megagametophyte data is to pool the information from each of the samples to provide a locus specific estimate of dominant allele frequency. We define the random variable Z_{j+}, where $0 \leq Z_{j+} \leq C_j S$, as the total number of positive alleles

across all S samples at locus j such that $Z_{j+} = Z_{1j+} + \cdots + Z_{ij+} + \cdots + Z_{Sj+}$. If we let $\ell_{ij}(x) = \mathbb{P}(x|M_{ij+}, K_{ij}, C_j, \phi_j, \psi_j)$, the probability mass function of the number of positive alleles in the genotype of the ith individual, then it is possible to define the probability mass function of the random variable Z_{j+} as $\upsilon_j(x)$, where $\upsilon_j(x) = \mathbb{P}(Z_{j+} = x|\mathbf{M}_{j+}, \mathbf{K}_j, C_j, \phi_j, \psi_j)$, $\mathbf{M}_{j+} = \{M_{1j+}, \ldots, M_{ij+}, \ldots, M_{Sj+}\}$, and $\mathbf{K}_j = \{K_{1j}, \ldots, K_{ij}, \ldots, K_{Sj}\}$ as the following:

$$\upsilon_j(x) = (\ldots(\ldots((\ell_{1j} * \ell_{2j}) * \ell_{3j}) \cdots * \ell_{ij}) \cdots * \ell_{Sj})(x) \tag{2.22}$$

where $*$ is a convolution operator such that

$$(\ell_{hj} * \ell_{ij})(x) = \sum_{\tau=0}^{\min\{x, C_j S\}} \ell_{hj}(x - \tau)\, \ell_{ij}(\tau) \tag{2.23}$$

The absence of a simple, closed form, of the probability mass function of Z_{j+}, $\upsilon_j(x)$, for an arbitrary number of loci requires that evaluation be performed numerically. Fortunately we can adapt the algorithm published in Butler & Stephens (1993) to allow the precise evaluation of the probability mass function at all possible values of Z_{j+}. This function was originally developed to calculate the exact probabilities of the resultant probability mass function of a random variable that is sum of a set of random variables independently drawn from binomial distributions with varying trial success probabilities. In this application, we keep the general structure of the algorithm intact but instead change the steps that use the probability of the mass function to use values calculated using equation 2.21 instead.

Algorithm 1: Exact calculation of the probability mass function for total frequencies of the positive allele in biallelic markers taken from haploid gamete tissues

1. Initialise an array a with $C_j S + 1$ elements and index starting at zero. For all elements of a set
$$a_k = \begin{cases} \ell_{1j}(k) & \text{if } k \leq C_j \\ 0 & \text{if } k > C_j \end{cases} \tag{2.24}$$
2. Set the sample iterator $i = 2$.
3. Initialise an array b with $C_j S + 1$ elements and index starting at zero. Initialise each element of b to zero.
4. Initialise an array c with $C_j + 1$ elements and index starting at zero. For all elements of

c set $c_k = \ell_{ij}(k)$.

5. Set the array iterator $k = 0$.
6. Set the probability iterator $h = 0$.
7. Increment the value of b_{k+h} by $a_k c_h$.
8. Increment h. If $h \leq C_j$ then go to step 7.
9. Increment k. If $k \leq C_j(i-1)$ then go to step 6.
10. Set each element of a to the value contained in corresponding element of b.
11. Increment i. If $i \leq S$ then go to step 3.
12. Take the values of element of array a as completed calculations of the probability mass function of variable Z_{j+} evaluated at the relevant index such that $\upsilon_j(k) = a_k$.

The resultant array a_k can be used to calculate the 95% credible interval of positive allele counts at the relevant locus by taking the index values of the array at which the cumulative probabilities first exceed the values of 0.025 and 0.975. These indices, $k_{0.025}$ and $k_{0.975}$ respectively, can be converted into credible intervals of positive allele frequencies by simply dividing the index by the maximum number of positive alleles, $C_j S$.

2.2.5 Fitting the Model

In all analyses it is important to note that the allele frequencies at each locus and the value of the inbreeding/outbreeding coefficient are not independent. Firstly, the frequencies of the total set of possible allele values must sum to one at each locus ,and secondly, that the value for F_{IS} adheres to the conditions laid out in inequalities 2.7 and 2.8. For these conditions to be correctly addressed it is important to give nil prior weight to values that do not conform to these conditions. A simple density function that provides uniform support over the values of $\mathbf{f}_j$ and that satisfy the unit sum requirement is a special parameterisation of the A_j dimensional Dirichlet distribution, where A_j is the number of allele types at locus j, with all shape parameters set to one. This reduces to

$$\mathbb{P}(\mathbf{f}_j) = \begin{cases} \Gamma(A_j) & \text{if } \sum_a \mathrm{f}_{j_a} = 1 \text{ and } 0 \leq \mathrm{f}_{j_a} \leq 1 \text{ for all } a \\ 0 & \text{otherwise} \end{cases} \tag{2.25}$$

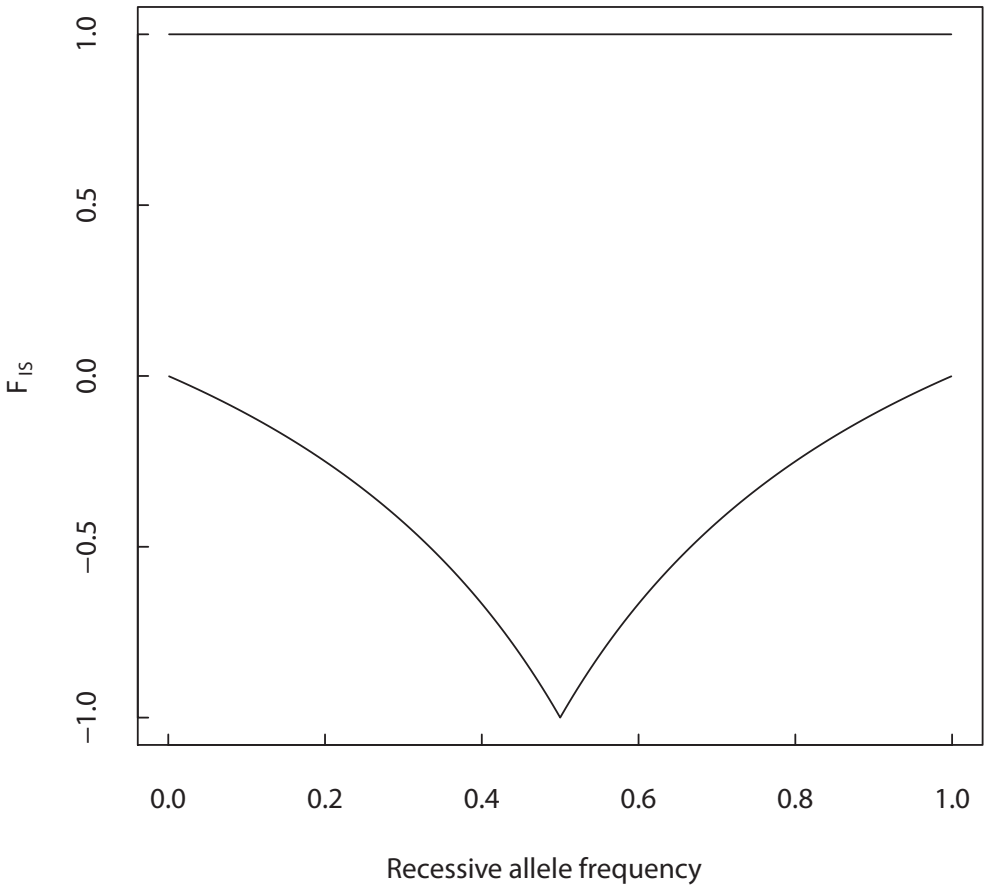


Figure 2.2: Figure illustrating the maximum and minimum values for F_{IS}, denoted by the black lines, at different allele frequencies for a given locus in the biallelic case. When multiple loci are considered, values of F_{IS} are restricted to values that satisfy the boundary conditions at the allele frequencies of each individual locus.

where $\Gamma(x)$ is the gamma function.

The minimum for the range of possible values of F_{IS} is dependent upon the values of the allele frequency vector across all loci, $\mathbf{f}$. For scenarios where the value of F_{IS} is not fixed, we propose the use of a distribution for F_{IS} that is conditionally uniform given a set of allele frequencies so that the minimum possible value of F_{IS} satisfies the conditions in inequalities 2.7 and 2.22 for all possible genotypes at all loci so

$$\mathbb{P}(\mathbf{f}, F_{IS}) = \mathbb{P}(F_{IS}|\mathbf{f}) \prod_j \mathbb{P}(\mathbf{f}_j) \tag{2.26}$$

where

$$\mathbb{P}(F_{IS}|\mathbf{f}) = \begin{cases} \frac{1}{1-z(\mathbf{f})} & \text{if } z(\mathbf{f}) \leq F_{IS} \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.27}$$

and $z(\mathbf{f})$ is a function that calculates the minimum value of F_{IS} given a set of allele frequencies such that

$$z(\mathbf{f}) = \max\{z_1, z_2, \ldots, z_j, \ldots, z_L\} \tag{2.28}$$

$$z_j = \max\left\{1 - \frac{1}{\prod_{k=1}^{C_j} \sum_a \mathrm{f}_{j_a} \omega_a(v_{ijk})}, -\frac{\left(\mathrm{f}_{j_1}\right)^{C_j-1}}{1-\left(\mathrm{f}_{j_1}\right)^{C_j-1}}, -\frac{\left(\mathrm{f}_{j_2}\right)^{C_j-1}}{1-\left(\mathrm{f}_{j_2}\right)^{C_j-1}}, \ldots, -\frac{\left(\mathrm{f}_{j_{A_j}}\right)^{C_j-1}}{1-\left(\mathrm{f}_{j_{A_j}}\right)^{C_j-1}}\right\} \tag{2.29}$$

Figure 2.2 shows the possible boundaries of F_{IS} in the simple, one locus, biallelic case.

The analyses contained in this paper used fixed values for the parameters, ϕ and ψ, in the biallelic observation model but other applications may require an estimate of genotyping error rates from the data. In most implementations, the prior support for the parameters for the observation model can be considered independent of the vector of allele frequencies and inbreeding/outbreeding coefficient. If we consider the parameter values of the observation model at each locus to be also independent then the full joint prior for all inferred values is simply

$$\mathbb{P}(\mathbf{f}, F_{IS}, \boldsymbol{\beta}) = \mathbb{P}(F_{IS}|\mathbf{f}) \prod_j \mathbb{P}(\mathbf{f}_j) \mathbb{P}(\boldsymbol{\beta}_j) \tag{2.30}$$

In order to successfully implement the Metropolis-Hastings algorithm and generate samples

from a distribution with density described by equation 2.12, we require the ability to evaluate three density functions: a prior function (equation 2.30), a likelihood function (equation 2.11), and a proposal function. For the sake of efficiency it is important that the proposal function only propose new values for $\boldsymbol{\beta}$, $\mathbf{f}$, and F_{IS} that have some prior support, which means proposed values of F_{IS} must fall within the acceptable range and that the unit sum requirement of $\mathbf{f}$ is upheld. One way to ensure that these conditions are met is to generate proposed values for allele frequencies, $\mathbf{f}^*$, using a truncated normal distribution for each allele frequency in turn except for the last with a location parameter (corresponding to the mean parameter in a non-truncated normal distribution) set to the previous frequency of the relevant allele type and a shape parameter (corresponding to the standard deviation parameter in a non-truncated normal distribution) denoted by ϵ_1. The truncated normal proposal distribution for each allele type would have a zero lower bound and an upper bound equal to the greatest possible value that could satisfy the unit sum requirement: one minus the sum of the frequencies of all alleles that have been set. The last allele frequency to be set is not drawn from any distribution but simply taken to be the remainder of the available frequency. The probability density function for the proposal of each allele frequency is therefore

$$\mathbb{P}\left(\mathrm{f}^*_{j_1}|\mathrm{f}_{j_1},\epsilon_1\right) = \begin{cases} \dfrac{e^{\frac{-\left(\mathrm{f}^*_{j_1}-\mathrm{f}_{j_1}\right)^2}{2\epsilon_1^2}}}{\epsilon_1\sqrt{2\pi}\left[\Phi\left(\frac{1-\mathrm{f}_{j_1}}{\epsilon_1}\right)-\Phi\left(\frac{-\mathrm{f}_{j_1}}{\epsilon_1}\right)\right]} & \text{if } 0 \le \mathrm{f}_{j_1} \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.31}$$

$$\mathbb{P}\left(\mathrm{f}^*_{j_a}|\mathrm{f}_{j_a},\epsilon_1,\mathrm{f}^*_{j_{a-1}},\mathrm{f}^*_{j_{a-2}},\ldots,\mathrm{f}^*_{j_1}\right) = \begin{cases} \dfrac{e^{\frac{-\left(\mathrm{f}^*_{j_a}-\mathrm{f}_{j_a}\right)^2}{2\epsilon_1^2}}}{\epsilon_1\sqrt{2\pi}\left[\Phi\left(\frac{-\mathrm{f}_{j_a}+\sum_{c=1}^{a-1}\mathrm{f}^*_{j_c}}{\epsilon_1}\right)-\Phi\left(\frac{-\mathrm{f}_{j_a}}{\epsilon_1}\right)\right]} & \text{if } 0 \le \mathrm{f}_{j_a} \le \sum_{c=1}^{a-1}\mathrm{f}^*_{j_c} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } 1 < a < A_j \tag{2.32}$$

$$\mathbb{P}\left(\mathrm{f}^*_{j_{A_j}}|\mathrm{f}^*_{j_{A_j-1}},\mathrm{f}^*_{j_{A_j-2}},\ldots,\mathrm{f}^*_{j_1}\right) = \begin{cases} 1 & \text{if } \mathrm{f}^*_{j_{A_j}} = 1-\sum_{c=1}^{A_j-1}\mathrm{f}^*_{j_c} \\ 0 & \text{otherwise} \end{cases} \tag{2.33}$$

Gathering the probability density functions together gives the joint distribution for a proposed

frequency vector

$$\mathbb{P}\left(\mathbf{f}_j^*|\mathbf{f}_j,\epsilon_1\right) = \mathbb{P}\left(\mathrm{f}_{j_1}^*|\mathrm{f}_{j_1},\epsilon_1\right)\mathbb{P}\left(\mathrm{f}_{j_{A_j}}^*|\mathrm{f}_{j_{A_j-1}}^*,\mathrm{f}_{j_{A_j-2}}^*,\ldots,\mathrm{f}_{j_1}^*\right) \prod_{a=2}^{A_j-1}\mathbb{P}\left(\mathrm{f}_{j_a}^*|\mathrm{f}_{j_a},\epsilon_1,\mathrm{f}_{j_{a-1}}^*,\mathrm{f}_{j_{a-2}}^*,\ldots,\mathrm{f}_{j_1}^*\right) \tag{2.34}$$

In the biallelic case, equation 2.34 simplifies further, removing the product term. Treating the proposal of allele frequency vectors independently across loci means that the joint probability of all proposed allele frequency vectors is simply

$$\mathbb{P}\left(\mathbf{f}^*|\mathbf{f},\epsilon_1\right) = \prod_j \mathbb{P}\left(\mathbf{f}_j^*|\mathbf{f}_j,\epsilon_1\right) \tag{2.35}$$

Finally, once a set of allele frequencies are proposed, a new value for F_{IS}, F_{IS}^*, can be proposed given the restrictions set out in inequalities 2.7 and 2.8. Like the proposition of allele frequencies, we propose values for F_{IS}^* by drawing from a truncated normal distribution defined between an upper value of one and a lower bound given by the lowest possible value of F_{IS} for which inequalities 2.7 and 2.8 hold, $z\left(\mathbf{f}\right)$. This distribution has location parameter given by F_{IS} and shape parameter, ϵ_2, such that

$$\mathbb{P}\left(F_{IS}^*|F_{IS},\epsilon_2,\mathbf{f}^*\right) = \begin{cases} \dfrac{e^{\frac{-(F_{IS}^*-F_{IS})^2}{2\epsilon_2^2}}}{\epsilon_2\sqrt{2\pi}\left[\Phi\left(\frac{1-F_{IS}}{\epsilon_2}\right)-\Phi\left(\frac{z(\mathbf{f}^*)-F_{IS}}{\epsilon_2}\right)\right]} & \text{if } z\left(\mathbf{f}^*\right) \leq F_{IS}^* \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.36}$$

Thus the joint proposal distribution is as follows:

$$\mathbb{P}\left(F_{IS}^*,\mathbf{f}^*,\boldsymbol{\beta}^*|F_{IS},\mathbf{f},\boldsymbol{\beta},\epsilon_1,\epsilon_2\right) = \mathbb{P}\left(F_{IS}^*|F_{IS},\epsilon_2,\mathbf{f}^*\right)\mathbb{P}\left(\mathbf{f}^*|\mathbf{f},\epsilon_1\right)\mathbb{P}\left(\boldsymbol{\beta}^*|\boldsymbol{\beta}\right) \tag{2.37}$$

where $\mathbb{P}\left(\boldsymbol{\beta}^*|\boldsymbol{\beta}\right)$ is a proposal density function for the parameters of the observation model. Here ϵ_1 and ϵ_2 are tunable parameters determining proposal step length in the allele frequencies and inbreeding/outbreeding coefficient respectively. The values used for ϵ_1 and ϵ_2 do not affect the eventual convergence of samples drawn using the Metropolis-Hastings algorithm to the target density but they can affect the efficiency of convergence. Larger steps result in a faster exploration of possible parameter combinations, but steps that are too large result in the proposal of many more low probability combinations, and hence, higher rejection rate (Chib & Greenberg, 1995). Note that it is an easy step to calculate the inverse-step probability of generating the initial set of parameter values from the proposed parameter values $\mathbb{P}\left(F_{IS},\mathbf{f},\boldsymbol{\beta}|F_{IS}^*,\mathbf{f}^*,\boldsymbol{\beta}^*,\epsilon_1,\epsilon_2\right)$:

simply reverse the standard and stared forms of the parameters in equation 2.37 and the relevant sub-formulae.

Algorithm 2: Implementation of the Metropolis-Hastings algorithm for joint estimation of allele frequencies, F_{IS} and observation model parameters

1. Initialise the allele frequency vectors ($\mathbf{f}$), F_{IS} and observation model parameters ($\boldsymbol{\beta}$) using arbitrary values (although these values must be within the support of the prior).

2. For each jth locus:

   (a) Create a new vector of observation parameters at locus j, $\boldsymbol{\beta}_j^*$ from a supplied proposal density. In this study the values of the observation parameters, ϕ_j and ψ_j are fixed and equal across all loci. We include this step here to show the reader where it is possible to allow observation error rates to be estimated from the data (although see discussion).

   (b) Set the remaining frequency counter $c = 1$.

   (c) For each ath allele type at the jth locus except for the last:

      i. Create a new proposal frequency for allele a at locus j, $\mathrm{f}_{j_a}^*$, by drawing a value from a truncated normal distribution with location parameter f_{j_a} and shape parameter ϵ_1 (using algorithms such as those published in Robert, 1995). The possible values for $\mathrm{f}_{j_a}^*$ are truncated between zero and c.

      ii. Decrement c by $\mathrm{f}_{j_a}^*$.

   (d) Set the last element of $\mathbf{f}_j^*$, $\mathrm{f}_{j_{A_j}}^* = c$.

3. Create a new proposal value for the inbreeding/outbreeding coefficient, F_{IS}^*, by drawing a random value from a truncated normal distribution with location parameter F_{IS} and shape parameter ϵ_2. The possible values for F_{IS}^* are truncated between $z(\mathbf{f}^*)$ and 1.

4. Calculate

$$u = \frac{\mathbb{P}(O|\mathbf{f}^*, F_{IS}^*, \boldsymbol{\beta}^*)\,\mathbb{P}(\mathbf{f}^*, F_{IS}^*, \boldsymbol{\beta}^*)\,\mathbb{P}(\mathbf{f}, F_{IS}, \boldsymbol{\beta}|\mathbf{f}^*, F_{IS}^*, \boldsymbol{\beta}^*)}{\mathbb{P}(O|\mathbf{f}, F_{IS}, \boldsymbol{\beta})\,\mathbb{P}(\mathbf{f}, F_{IS}, \boldsymbol{\beta})\,\mathbb{P}(\mathbf{f}^*, F_{IS}^*, \boldsymbol{\beta}^*|\mathbf{f}, F_{IS}, \boldsymbol{\beta})} \tag{2.38}$$

   where the first term on the numerator and the denominator is the likelihood, given by equation 2.11, of the proposed and current parameter values respectively. The second term is the prior expectation of the proposed or current parameter values (see equation 2.30). The final term corresponds the step probabilities given by equation 2.37.

5. Generate a random uniform value, α, between zero and one. If $\alpha < \min\{1, u\}$ then set the current parameter values, $\mathbf{f}$, F_{IS}, and $\boldsymbol{\beta}$, to the proposed values, $\mathbf{f}^*$, F_{IS}^*, and $\boldsymbol{\beta}^*$, respectively.

6. Store the current values of $\mathbf{f}$, F_{IS}, and $\boldsymbol{\beta}$ as a sample from the target distribution.

7. Return to step 2.

2.2.6 Assessing Model Performance

In the previous sections we have shown how to generate credible intervals for allele frequencies at each locus using the raw megagametophyte data of Isabel *et al.* (1995) and Isabel *et al.* (1999). Because this megagametophyte data is much more reliable indicator of the zygosity of each of the individuals tested than phenotype data, we can use the estimates derived using algorithm 1 as a benchmark for the performance of the allele frequency estimates using phenotype data alone. Obviously, in most applications, non-gametic tissue is used and multiple replicates of haploid markers are not available. To simulate the common case, where only a single band presence or absence at each sample and locus represents the total information available to the investigator, we reduce the series of megagametophyte runs at each sample and locus to a single positive or negative value to simulate the corresponding phenotypes of markers taken from non-gametic DNA with the same genotype. Positive phenotypes are generated if any of the megagametophytes of a particular sample exhibit a positive phenotype and, conversely, negative phenotypes are generated only if all the megagametophytes exhibit negative phenotypes. We then fit the hierarchical model to this phenotype data using the methods described in the previous section. This allows the comparison of the allele frequency estimates derived from methods applied to the generated phenotypes to those estimates derived from the full information contained in the megagametophytes.

So far, the description of this allele frequency estimation technique has assumed that the investigator will perform a joint inference of the error parameters, ψ and ϕ, alongside the allele frequencies. Whilst this is encouraged for a standard analysis where there may be some uncertainty over the error rate parameters and where this uncertainty should be incorporated in estimates of the allele frequency estimates, it does not make sense when testing the performance of the model inference of the phenotype against the inference of the allele frequency estimates drawn from the more informative megagametophytes. This is because both the estimate of the

'true' allele frequencies derived from the full megagametophyte data using algorithm 1 and the frequency estimates derived from the hierarchical model described in this chapter and fitted to the generated phenotype data using algorithm 2 are dependant upon these error rate parameters. If we are to use the relatively narrow band of possible allele frequencies from the megagametophyte data as a benchmark for the performance of those allele frequencies generated from a model trained from the phenotype data alone then it is much more sensible to fix the error model parameters and assess the performance of the phenotype-only model in recreating the information present in the megagametophyte allele frequency estimates at given values for the error parameters. We repeat the analysis for each of three possible values for ϕ and ψ: 0.01, 0.05, and 0.1. These values span the range of error rates that are likely in any genetic study. It is important to note that this approach is not a peculiarity to the fact that we are using megagametophyte data as our benchmark estimate for the allele frequencies. Even if we had generated artificial data as our benchmark then we would still require the use of an observation model to convert the simulated genotypes into phenotypes to use as inputs to the phenotype-only model.

The allele estimation procedure described previously was preformed on the observed phenotypes from all available loci of the *Pinus strobus* and *Picea mariana* data sets. Four separate chains of the Metropolis-Hastings algorithm were run for a total of 130000 iterations each and for each locus separately. Starting values for the allele frequencies at each locus were drawn from a uniform distribution between zero and one to initialise each chain. The first 30000 samples were discarded to allow for 'burn-in'. A simple random-walk proposal function truncated between the values of zero and one was used to generate candidate allele frequencies. The standard deviation of the step-length of the random walk proposal function was 0.1. Visual inspection of the trajectory of allele frequency estimates and the mixing of the chains showed ample convergence in each analysis. Moreover, the values for the multivariate scale reduction factor convergence metric proposed by Brooks & Gelman (1998) calculated for each of the analyses ranged from 1.033 to 1.211. Under this convergence criterion, values close to one are indicative of sets of chains for which an increase in run time will not significantly alter the estimation of the parameter values.

2.3 Results

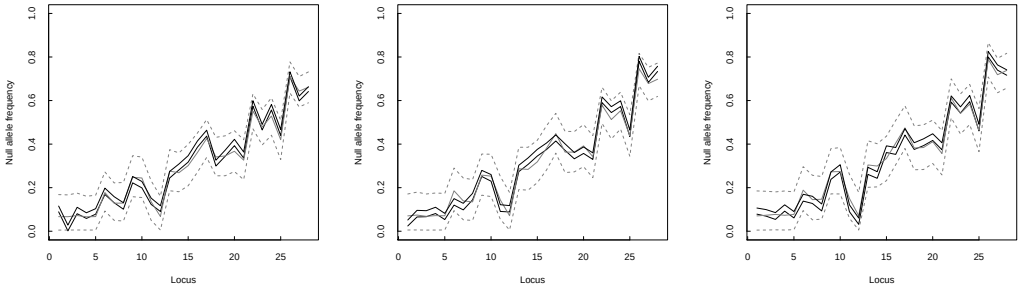
Figures 2.4 and 2.3 show the null allele frequencies estimated for each locus tested from populations of *Pinus strobus* and *Picea mariana* respectively, using the megagametophyte data directly and inferred from using the phenotype data only. The figures illustrate a clear general agreement between the allele frequencies estimated using the phenotype-only model and the more direct inference from the raw megagametophyte data. This relationship appears to hold across the range of potential error values tested. Moreover, the range of the credible interval generated from the phenotype-only model is reasonably narrow, with most allele frequency estimates falling with a 5% band for most loci, even under high rates of observation error. Given that the allele frequency estimates from the megagametophyte data rarely fall outside this band, this suggests that even though information is lost when only phenotype data exists, allele frequency estimates made using the phenotype-only model can still provide reasonable estimates of allele frequency.

2.4 Discussion

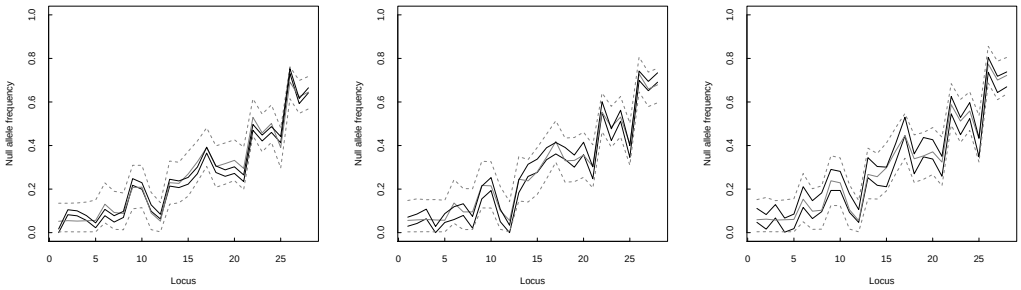
This paper describes a novel method to estimate recessive allele frequencies. Unlike existing techniques, the methods described here allow for the inclusion of uncertainty arising from genotyping errors, and, as we have shown, the allele frequency estimates are robust even when the error rates are high. The quality of the estimation holds even for loci with extreme null allele frequencies. This contrasts with the allele frequency estimator of Zhivotovsky (1999), which can over-predict null allele frequencies when they occur in low frequencies, and the estimator of Lynch & Milligan (1994) which requires that at least three recessive homozygotes are present at a locus in order to calculate the recessive allele frequency.

Allele frequencies lie at the core of a many number of metrics and statistics of population genetics. The F_{ST} metric of Wright (1951), the D index of Nei (1972) and the D_{ST}, G_{ST}, and R_{ST} statistics of Nei (1973) are all common measures of population subdivision and are applied regularly to draw inference about the genetical structure of populations from molecular data. Moreover, formal tests of population subdivision, such as the G log-likelihood ratio test of Goudet *et al.* (1996) also require accurate estimates of allele frequency in order for the test to be statistically robust. The presence of null alleles distorts the perception of the zygosity and allele frequencies in the population, causing difficulty in interpreting the outputs of such methods.

(a) $\phi = 0.01$ and $\psi = 0.01$ (b) $\phi = 0.01$ and $\psi = 0.05$ (c) $\phi = 0.01$ and $\psi = 0.1$



(d) $\phi = 0.05$ and $\psi = 0.01$ (e) $\phi = 0.05$ and $\psi = 0.05$ (f) $\phi = 0.05$ and $\psi = 0.1$



(g) $\phi = 0.1$ and $\psi = 0.01$ (h) $\phi = 0.1$ and $\psi = 0.05$ (i) $\phi = 0.1$ and $\psi = 0.1$

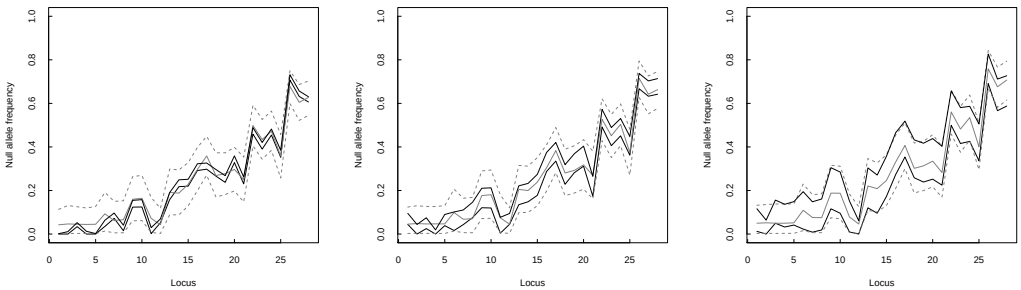
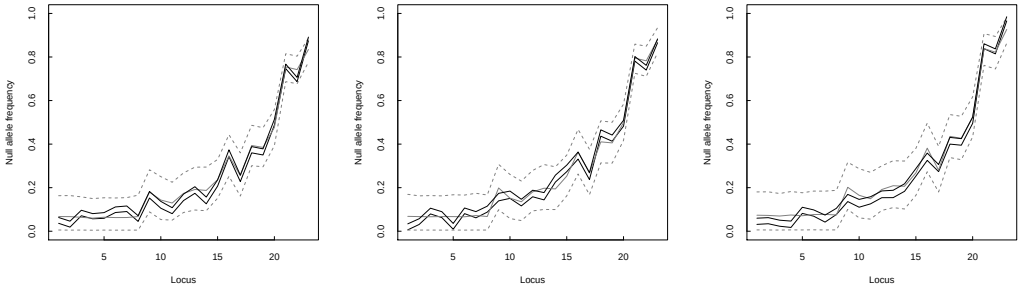
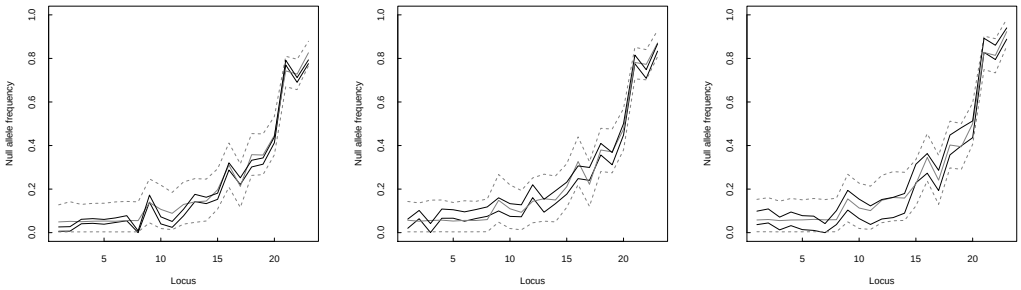


Figure 2.3: Recessive allele frequency estimates for a selection of RAPD loci from *Picea mariana*. Solid black lines represent the 95% credible interval derived from the raw megagametophyte data. The dotted grey lines represent the 95% credible interval of the posterior allele frequency estimates derived using the methods described in this chapter and taking for input the observed phenotypes. The solid grey line is the median value of the allele frequencies sampled using the MCMC sampler.

(a) $\phi = 0.01$ and $\psi = 0.01$ (b) $\phi = 0.01$ and $\psi = 0.05$ (c) $\phi = 0.01$ and $\psi = 0.1$



(d) $\phi = 0.05$ and $\psi = 0.01$ (e) $\phi = 0.05$ and $\psi = 0.05$ (f) $\phi = 0.05$ and $\psi = 0.1$



(g) $\phi = 0.1$ and $\psi = 0.01$ (h) $\phi = 0.1$ and $\psi = 0.05$ (i) $\phi = 0.1$ and $\psi = 0.1$

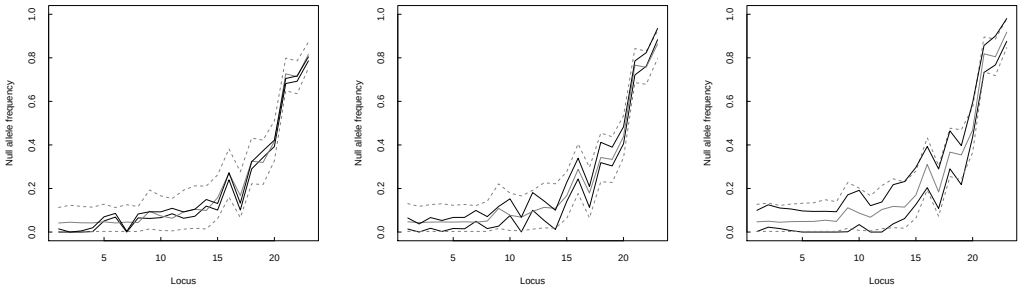


Figure 2.4: Recessive allele frequency estimates for a selection of RAPD loci from *Pinus strobus*. Solid black lines represent the 95% credible interval derived from the raw megagametophyte data. The dotted grey lines represent the 95% credible interval of the posterior allele frequency estimates derived using the methods described in this chapter and taking for input the observed phenotypes. The solid grey line is the median value of the allele frequencies sampled using the MCMC sampler.

The prevailing approach to address such issues is to avoid the use of dominant marker types or codominant marker loci with high null allele frequencies for these types of analyses. This presents a number of added problems however. Much information is wasted in simply removing certain loci from the analysis. Throwing away loci that contain null heterozygotes may artificially bias the analysis as it is those loci exhibiting high levels of homozygosity that are most likely to be removed by this protocol. Null heterozygotes may still be present in the remaining loci and so it is not clear that the allele frequencies derived from the phenotypes from the remaining loci will neccessarily be a good reflection of their genotype distribution even after this removal. A much better approach is to estimate the frequency of null alleles in these loci rather than remove them. The methods provided in this chapter have been shown to provide reliable estimates for the null allele frequency even if these frequencies are extremely high or extremely low and so they can be instrumental in avoiding the wasteful removal of data to perform basic molecular analyses of population structure.

In order to accommodate a wide range of situations, the modelling framework described here allows for the joint estimation of allele frequencies, F_{IS}, and parameters of the observation model. Foll *et al.* (2008) point out that many different combinations of F_{IS} and allele frequency can result in the same expectation of allele frequencies expressed in the phenotype. Except when the number of loci is large, methods which attempt a joint estimation of these parameters can perform poorly (Bonin *et al.*, 2007). However, only part of the problems associated with the joint estimation of allele frequencies and F_{IS} are due to colinearity of the parameters on the likelihood surface. The latest manual of the software package HICKORY (Holsinger *et al.*, 2002) describes an additional ascertainment bias in the joint calculation of F_{IS} and allele frequency: loci are chosen for their polymorphic properties as it is those loci that are the most informative in the inference of population structure (Meudt & Clarke, 2007), not because they are indicative of the total genomic difference between individuals or populations. Excluding non-polymorphic loci results in a bias in the distribution of phenotypes which is used to inform estimates of F_{IS} (Foll *et al.*, 2008). It is not always possible to correct this bias by retaining non-polymorphic loci: RAPD and AFLP loci that express recessive phenotypes across the entire population are impossible to identify.

Ascertainment bias can be partially addressed through the use of suitable priors for allele frequencies or F_{IS}. Zhivotovsky (1999), Foll *et al.* (2008), and Wright (1951) all talk about the use of various parametrisation of a beta distribution, in the biallelic case, to describe the prior

distribution of allele frequencies. Parameters for this prior can be fixed according to a theoretical expectation of the status of the population (Wright, 1951) or themselves estimated as part of the fitting process (Zhivotovsky, 1999; Foll *et al.*, 2008). Both cases can be implemented within this framework as an extra layer in the hierarchical model by adding an extra step where, if we let $\boldsymbol{\alpha}_{ij}$ be the vector of parameters for the distribution describing allele frequencies at locus j of sample i and $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_{11}, \ldots, \boldsymbol{\alpha}_{ij}, \ldots, \boldsymbol{\alpha}_{SL}\}$, the likelihood equation (equation 2.11) is replaced by

$$\mathbb{P}\left(O|\mathbf{f}, F_{IS}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \prod_{j}\prod_{i}\sum_{G_{ij}\in H_j} \mathbb{P}\left(O_{ij}|G_{ij}, \boldsymbol{\beta}_{ij}\right)\mathbb{P}\left(G_{ij}|\mathbf{f}_j, F_{IS}\right)\mathbb{P}\left(\mathbf{f}_j|\boldsymbol{\alpha}_{ij}\right) \tag{2.39}$$

Here $\mathbb{P}\left(\mathbf{f}_j|\boldsymbol{\alpha}_{ij}\right)$ represents the probability of a vector of allele frequencies given a vector of parameters of the distribution controlling the allele frequencies. Samples are then drawn from the following distribution instead of that of equation 2.12 using the Metropolis-Hastings algorithm:

$$\mathbb{P}\left(\mathbf{f}, F_{IS}, \boldsymbol{\beta}, \boldsymbol{\alpha}|O\right) = \frac{\mathbb{P}\left(O|\mathbf{f}, F_{IS}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\mathbb{P}\left(\mathbf{f}, F_{IS}, \boldsymbol{\beta}\right)\mathbb{P}\left(\boldsymbol{\alpha}\right)}{\iiiint_{V_{\mathbf{f}F_{IS}\boldsymbol{\beta}\boldsymbol{\alpha}}} \mathbb{P}\left(O|\mathbf{f}, F_{IS}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\mathbb{P}\left(\mathbf{f}, F_{IS}, \boldsymbol{\beta}\right)\mathbb{P}\left(\boldsymbol{\alpha}\right)\, d\mathbf{f}\, dF_{IS}\, d\boldsymbol{\beta}\, d\boldsymbol{\alpha}} \tag{2.40}$$

$\mathbb{P}\left(\boldsymbol{\alpha}\right)$ is a hyperprior and describes the *a priori* distribution of $\boldsymbol{\alpha}$. In the biallelic case with beta distributed prior described above, the vector $\boldsymbol{\alpha}$ contains the two shape parameters of the beta distribution, λ_1 and λ_2. The beta distribution is a very flexible distribution but when $\lambda_1 > 1$ and $\lambda_2 > 1$ the resultant unimodal distribution gives a zero probability weight to the extremes of the allele frequencies. Given that it is loci with allele frequency extremes that are likely to exhibit low polymorphism, and hence more likely to be excluded from the analysis, then by restricting λ_1 and λ_2 to values that meet these conditions we are provided with a mechanism by which ascertainment bias can be incorporated into the analysis. For markers with more than two allele types, the ascertainment bias can be similarly addressed by specifying parameters with values greater than one for a A_j dimensional Dirichlet distribution.

The addition of a genotyping error model further complicates this inference and has the capacity to extend the parameter likelihood colinearity into higher dimensions, creating problems with parameter identifiability, and resulting in flat marginal posterior distributions for these parameters. It is unlikely that one dataset alone will contain enough information to jointly estimate the entire parameter space and so the authors stress the importance of either fixing some parameters, hence making assumptions about the genotyping error rate or the level of inbreeding or outbreeding, or assigning narrow priors for these parameters. In some cases it may be possible to independently ascertain prior support for the parameters from information

contained in other data sets, for example, as we have shown in this study, allele frequencies at loci can be inferred from haploid gametic tissue.

It is important to note here that whilst we have treated null alleles specially, we have made the strident assumption that they are of all the same type. A null allele can arise from a number of different mutations, and it can be argued that two null alleles, whilst phenotypically indistinct, may have a very different genetic basis. For parentage analysis, the presumed heritability of 'null' without further division results in the non-exclusion of potential parentage pairs of a given offspring where incompatible null genotypes are present. Metrics aimed at estimating inter-population genetic dissimilarity may also exhibit downward biases as diversity in null allele types cannot be incorporated into the analysis. Of course, for this issue to be addressed correctly it is necessary to have some information on the null allele types. In the unlikely scenario that this information exists for the loci of interest and that it is possible to express this expectation as a prior, it is a small step of theoretical development to incorporate this within the framework presented here. For all other occasions, the most parsimonious stance is to treat all null alleles similarly. Adding weight to potentially erroneously allocated parents by creating incompatible matches, when the assertion that null alleles types are differentiable is groundless, garners only questionable results from the application of parentage analysis techniques. Moreover, in the realm of population genetics, the type I error of inferring that populations are genetically distinct when in fact this is not the case is a much more grievous misdemeanour than its type II equivalent: a situation that is encouraged if excessive, but fallacious, variability from different possible null allele types is enforced on the data.

In order to analyse the RAPD data used to illustrate the methods described in this paper, we have only described an observation model relevant to the case of biallelic dominant markers. However, given the open nature for specification of the observation model, it is a simple extension to include observation models appropriate for the calculation of null allele rates amongst codominant markers such as microsatellites or RFLPs. The extension to the codominant case can be achieved most simply by using the bialleleic, dominant marker model described in this paper but treating the null allele as the recessive '$-$' allele and all other allele types as the dominant '$+$' allele. This method is only suitable if the only frequency of interest is that of the null allele. Uncertainty in the frequency of the null allele necessarily means uncertainty in the frequencies of all other allele types, and, in order to achieve a joint estimation of allele frequencies, it is necessary to define a polyallelic observation model.

Allele values in the codominant case are commonly defined by the fragment size and so the component of the observation model that determines genotyping error must contain a mechanism for the misdiagnosis of fragment length sizes. The error model described by Wang (2004) and implemented in the parentage analysis programs MasterBayes (Hadfield *et al.*, 2006) and Colony (Jones & Wang, 2010) describes one such model. Here error rates are apportioned into two types: class I types, those which involve the non-expression real allele types and equivalent to the specification of the ϕ parameter of the model for bialleleic dominant markers described in this paper, and class II types, where non-null allele types are replaced with a randomly selected other non-null allele type.

Parentage analysis programs such as Cervus (Marshall *et al.*, 1998; Slate *et al.*, 2000) and Parente (Cercueil *et al.*, 2002) allow the incorporation of an error model whereby genotyping errors occur with a tunable rate parameter: where errors occur, the genotype at a locus is replaced with a new genotype chosen either according to the expected frequencies under Hardy-Weinberg equilibrium (using the observed allele frequencies), or according to the observed phenotype frequencies of the sampled population (thus preserving population zygosity). In the case of the 'random relabelling' error model which selects a replacement genotype based on expected Hardy-Weinberg frequencies, the implicit assumption is made that allele frequencies can be adequately estimated from the observed phenotype. This is patently not true for dominant markers, but also, more insidiously, untrue for codominant markers containing null alleles. Even in the absence of null alleles, errors related to scoring may introduce some uncertainty into estimates of allele frequency. In addition, there is a logical inconsistency in the error model specification of these programs: for any one genotype there is a possibility of error, and given that an error occurs, a replacement is drawn from a population of genotypes or, in the Hardy-Weinberg assumed case of Cervus, combination of alleles, with frequencies determined by the sampled phenotypes which themselves are assumed to be free of error. Essentially each genotype is treated as error prone but the population from which replacements are drawn is inferred from the same genotypic information but with an error free assumption.

Aside from the logical inconsistency of these error models, it is difficult to envisage how errors of the type implemented in Cervus would actually arise. None of the common genotyping errors described in the methods section of this paper would result in the replacement of a single-locus genotype. Marshall *et al.* (1998) maintain that such an obfuscation in observation

might arise from laboratory labelling errors, but in such an instance, we would expect that the entire multilocus genotype would be replaced for the erroneous sample and not just the genotype at a single locus. Single locus genotype replacements may arise from data entry errors, but, even in these cases, it would be more sensible to draw the replacement genotype from the sample population, as implemented in PARENTE, and not from a theoretical population in Hardy-Weinberg equilibrium.

The allowance for loci specific error models in equation 2.11 permits the joint estimation of error parameters, F_{IS} and allele frequencies from multiple marker types simultaneously. Many studies of population genetics use multiple marker types in their inference. Whilst these markers may not share error parameters they do share information pertaining to the inbreeding/outbreeding of the population. The model described in this paper, alongside those of Holsinger *et al.* (2002) and Foll *et al.* (2008) assume that estimates of allele frequencies are not independent of the deviation of zygosity from Hardy-Weinberg equilibrium and so, under these models, combined multilocus inference of F_{IS} will result in better informed estimates of allele frequencies.

In summary, errors in the genotyping process combined with the observation of recessive alleles only in the homozygote case can result in significant biases in the estimation of allele frequencies. We have described here a flexible method of allele frequency estimation, and, through the example dataset used here, shown its efficacy in the biallelic case. We advocate the use of an observation model that emulates the main sources of genotyping error and preserves the hierarchy of allele dominance. Only then can true level of uncertainties relating to these processes be expressed in estimates of population structure.

CHAPTER 3

A generalised parentage assignment method for mixtures of DNA marker types and arbitrary ploidy levels

Summary

1. Genotyping errors and the presence of unobservable 'null' alleles can significantly bias parentage assignment.

2. There have been a number of papers describing methods for incorporating genotyping error into parentage analysis, but most place significant restrictions on the type of data that can be analysed. Very few available programs can analyse information derived from dominant markers, codominant markers with null alleles, or from markers that do not exhibit standard Mendelian inheritance dynamics.

3. We present here a flexible Bayesian method to allow fractional parentage assignment from a variety of molecular markers with many different modes of inheritance. We present a set of marker-specific observation models that link the underlying true genotypes of samples to the observed phenotypes for arbitrary ploidy.

4. We test the modelling framework in a population of Canary Island zebrafinch for which the parentage is known.

5. We show that the marker specific observation models have the capability to better distin-

guish parentage than generic random relabelling counterparts. We postulate that this may be because of structural deficiencies in the random relabelling model which are unable to account for errors of allelic dropout.

6. Parentage assignments can become poor when there is little prior information about the genotyping error rates. However, even wide but reasonably bounded prior distributions can markedly improve the performance of parentage assignment.

7. The framework presented here provides a novel method for the joint inference of paternity from multiple marker types. The possible inheritance of recessive alleles for an apparently homozygous parent pair is also explicitly calculated, avoiding fallaciously diagnosed incompatibilities caused by the presence of null alleles that are unaccounted for in the parental genotypes. This presents a substantial improvement over previous parentage assignment methodologies.

3.1 Introduction

Many aspects of species' ecology are influenced by patterns of parentage. The sexual behaviour of species, including their degree of polyandry and polygamy and levels of inbreeding, are determined directly by parentage relationships (Marker *et al.*, 2008; Worthington Wilmer *et al.*, 1999; McEachern *et al.*, 2009; Rourke *et al.*, 2009; McLean *et al.*, 2008; Efombagn *et al.*, 2009; Fernandes *et al.*, 2008). When parentage patterns are combined with spatial information it is possible to infer gene flow (Saenz-Agudelo *et al.*, 2009), the movement of individuals and gametes in space. This step is crucial, as it makes accessible the study of a broad range of phenomena and elucidates the mechanisms that drive and constrain it; parentage analysis has been applied to understanding the effects of home range (Martin *et al.*, 2007), dispersal (Piotti *et al.*, 2009; Zeyl *et al.*, 2009) and gamete transfer (Bacles & Ennos, 2008; Krauss *et al.*, 2009; Nakanishi *et al.*, 2009) on gene flow and reproductive success. These processes, founded on parentage patterns, in turn form the basis of population models. Developing an understanding of the reproductive potential of a species, along with the inter-individual variation around it, is an important component of population modelling (Williams & DeWoody, 2009) from which follows estimates of population viability, growth and stability. At the metapopulation level, parentage analysis can be instrumental in fitting models of dispersal and informing prediction of patch occupancy dynamics (Botsford *et al.*, 2009; Planes *et al.*, 2009).

The field of parentage analysis has seen a recent proliferation of techniques (see Jones & Ardren, 2003). Early methods focused on the exclusion of potential mother and father combinations through the observation of an offspring genotype at a locus that could not possibly arise from the recombination of the candidate parent genotypes at said locus (Chakraborty *et al.*, 1974). However, in situations where the number of loci used in the analysis are less numerous, or where each locus exhibits low polymorphism, it may be impossible to exclude all but one parentage pair. Moreover, inter-generational mutations and genotype observation errors in may result in the opposite, and potentially critical, problem of excluding the true parentage pair (Cifuentes *et al.*, 2006). Such situations have driven the theoretical development of parentage analysis techniques.

For situations where simple exclusion analysis is unable to exclude all but one viable parent pair, methods have been developed to weight the resultant possibilities: Meagher & Thompson (1986) develop a likelihood-based method of weighting the probability of the offspring exhibiting a genotype at a given locus given the parental genotypes. This weight is expressed as a ratio relative to the probability that both parents are unrelated to the offspring. In their analysis of a natural population of the perennial herb, *Chamaelirium luteum*, Meagher & Thompson (1987) assigned parentage to the pair that provided the highest unique likelihood ratio. Whilst this method allows for the diagnosis of parentage pairs when exclusion methods alone fail, the output of such an analysis for any parentage pair is still a dichotomous variable where parentage allocation is reduced to possible and not possible outcomes. Indeed Meagher & Thompson (1987) exclude all offspring from further analysis for which a unique parentage pair could not be allocated, amounting to approximately 63 percent of the sampled seedlings. Furthermore, in situations where a number of parentage likelihoods are similar, there does not exist enough support to assume that the most likely pedigree is the correct one (Thomas, 2005). This uncertainty in parentage allocation needs to represented in later analyses.

The alternative tactic of Devlin *et al.* (1988) instead relies on weighting a non-excluded parent pair by its likelihood as a proportion of the likelihood of all non-excluded potential parent pairs. This so-called 'fractional' assignment of parentage pairs can still result in situations where definitive allocation of parentage is lacking, but here ambiguity in the data can be expressed in quantitative terms. For cases where maternity is known, Nielsen *et al.* (2001) respecify the likelihood equations derived in Devlin *et al.* (1988) in a Bayesian context to generate posterior probabilities of paternity. These probabilities are used as fractional weights of paternity in the

program PATRI (Signorovitch & Nielsen, 2002).

Whilst fractional methods provide a rigorous method of ranking non-excluded parentage pairs, PATRI, as currently implemented, does not provide a mechanism for incorporating genotyping error or mutation. Whilst inter-generational mutation is rare, with rates for microsatellites varying between 10^{-3} and 10^{-4} per locus (Ellegren, 2000; Weber & Wong, 1993), cases with sufficiently large samples or where analysed loci are numerous may encounter a small number of mutation errors. Observation errors involving the mis-classification of genotypes during the scoring process are expected to be a much more widespread problem for parentage analyses. Marshall *et al.* (1998), and later revised in Kalinowski *et al.* (2007), extended the likelihood equations of Meagher (1986) to allow for observation error. Under these methods the number of possible parentage pairs for any given offspring is increased as some matches that would otherwise be excluded are retained, albeit with often diminishing likelihood, on the grounds that genetic mismatches may be the result of genotyping error. Unfortunately, the dichotomous assignment of Marshall *et al.* (1998), and that implemented in the software package CERVUS, where one parentage pair must be exclusively assigned to each offspring, can potentially undermine the benefits gained through the inclusion of genotype error (Hadfield *et al.*, 2006). This assertion can be exemplified in the case where a scoring error has occurred at a single locus in the genotype of the true father and results in the observation of a genotype apparently incompatible with the offspring genotype given a maternal genotype. CERVUS, whilst performing better than exclusion methods in that it will assign some probability to the event rather than none (although exclusion methods implemented in the software packages PROBMAX of Danzmann 1997, and NEWPAT of Worthington Wilmer *et al.* 1999, do allow some degree of flexibility by ranking compatibilities or specifying the degree of allowable mismatch respectively), will only correctly diagnose the true father if the likelihood is a arbitrarily set magnitude higher than the erroneous alternatives. Fractional methods, extended to include genotype error, have the potential to fare much better in this regard. Although the ranking of the posterior probability estimate of the true parentage versus the erroneous one would be likely to be the same under this methodology, the fractional allocation of paternity would accurately represent the added uncertainty due to genotyping error.

The MASTERBAYES package for the R statistical platform developed following the work of Hadfield *et al.* (2006) satisfies many of the desirable criteria above. Like PATRI, MASTERBAYES uses fractional paternity assignment, but, unlike PATRI, MASTERBAYES also allows for

the inclusion of genotyping errors in codominant markers, separating those resulting from allelic dropout from stochastic scoring misdiagnoses (Wang, 2004). The strength of the MasterBayes approach lies in the Bayesian calculation of the probability that a candidate parentage pair is the true parentage of a given offspring. The ability to link together Bayesian models to form hierarchical structures allows the parentage analysis to be easily embedded as part of a larger modelling effort. In the case of MasterBayes, this allows non-genetic data such as spatial location to inform parentage relationships. Given that parentage analysis is often conducted as a means to determine other ecological parameters of interest (Haig, 1998) the extensibility of the method used is of key importance. The probabilistic outputs of Bayesian parentage analysis provide a robust technique of propagating uncertainty, such as that associated with observation error or input factors, allowing assessment of the confidence in the derived values of the parameters of interest.

Here, we describe an alternative framework for Bayesian parentage analysis that allows the analysis of genotypes from a range of marker types (including various codominant and dominant markers) to be incorporated into a single analysis using marker-specific observation models. Our method is able to cope with arbitrary levels of ploidy, enabling its application across a wide range of non-model taxa for which parentage analysis in currently problematic. We show how the outputs of the parentage analysis can be used as part of a wider study and how supplementary data can inform not just parameters in linked models that describe the data, but also parentage.

3.2 Materials and Methods

3.2.1 Calculation of Parentage Likelihoods

When assessing a potential parent pair, the likelihood definition of greatest interest for the purposes of parentage assignment is the likelihood of observing multilocus phenotype of offspring i, $\mathbf{O}_i$, if the potential parent pair, m and f (mother and father respectively) were the true parent pair and had observed phenotypes, $\mathbf{O}_m$ and $\mathbf{O}_f$. Specified using another terminology, we need to define the probability of observing offspring genotype, $\mathbf{O}_i$, given that the true mother, $♀_i$, and true father, $♂_i$, for individual i are m and f respectively with observed genotypes, $\mathbf{O}_m$ and $\mathbf{O}_f$, otherwise written as $\mathbf{P}\left(\mathbf{O}_i|♀_i = m, ♂_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right)$. Genotype observations are not perfect, meaning that the observed genotype for any given locus may differ from the true underlying genotypic state of the locus. Therefore in order to evaluate $\mathbb{P}\left(\mathbf{O}_i|♀_i = m, ♂_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right)$,

where $\boldsymbol{\beta}$ is a vector of parameters controlling the observation process, we need to sum over the full set of possible true, but unknown and unobservable, genotypes. We let $\mathbf{G}_i$, $\mathbf{G}_m$, and $\mathbf{G}_f$ represent candidates for the true multilocus genotype across L loci of the offspring, mother, and father respectively such that $\mathbf{G}_{\cdot} = \{\mathbf{G}_{\cdot 1}, \mathbf{G}_{\cdot 2}, \ldots, \mathbf{G}_{\cdot j}, \ldots, \mathbf{G}_{\cdot L}\}$. $\mathbf{G}_{\cdot j}$ is the genotype of the relevant individual at locus j, a vector of length A_j, the number of allele types recorded at locus j. Each element of $\mathbf{G}_{\cdot j}$ represents the quantity of the appropriate allele present at locus j in the genotype of the individual of interest. Consequently, the sum of the elements of the genotype vectors at locus j, $\sum_a \mathrm{G}_{\cdot j_a}$, is equal to the ploidy at that locus C_j. Because the genotype is only recorded as a quantity vector of alleles then it is impossible to distinguish between different combinations of alleles that result in the same overall frequency; the vector $[1\,1]^T$, describing the allele quantities of a biallelic marker, could represent a heterozygote diploid organism with alleles on either of the two possible locations on the homologous chromosomes.

In a similar vein to the definition of the genotype allele frequency vector, $\mathbf{O}_{\cdot} = \{\mathbf{O}_{\cdot 1}, \mathbf{O}_{\cdot 2}, \ldots, \mathbf{O}_{\cdot j}, \ldots, \mathbf{O}_{\cdot L}\}$ where each element of $\mathbf{O}_{\cdot j}$ is the frequency of the observation of the relevant allele. Note that the observation allele frequency vector does not necessarily have to sum to the ploidy of the system like the genotype allele frequency vector. Multiple incidences of the same allele value will, in many marker systems, be hidden from the observer, resulting in a single observation where many alleles of the relevant type exist in the true genotype. Similarly, dominance hierarchies in the marker used may also obstruct observation of the recessive allele. We later describe a series of marker-specific observation models that link the vector of genotype allele frequencies to a vector of observed phenotypes.

It is important to note here that no diagnosis of the true genotype is attempted, or even desired. The probability of an observation depends upon the real state of the system that is being observed. If the real state of the system is unknown, then it is necessary to integrate the probability of our observation given a particular candidate state across all possible real states. Assuming that the probabilities of inheritance and observation for each locus is independent, that is, loci do not exhibit any form of linkage and that the observation process interrogates each locus independently, then

$$\mathbb{P}(\mathbf{O}_i|\mathrm{♀}_i = m, \mathrm{♂}_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}) =$$
$$\prod_j \sum_{\mathbf{G}_{mj}} \sum_{\mathbf{G}_{fj}} \sum_{\mathbf{G}_{ij}} \mathbb{P}\left(\mathbf{G}_{mj}|\mathbf{O}_{mj}, \boldsymbol{\beta}_{mj}\right) \mathbb{P}\left(\mathbf{G}_{fj}|\mathbf{O}_{fj}, \boldsymbol{\beta}_{fj}\right) \mathbb{P}\left(\mathbf{G}_{ij}|\mathbf{G}_{mj}, \mathbf{G}_{fj}\right) \mathbb{P}\left(\mathbf{O}_{ij}|\mathbf{G}_{ij}, \boldsymbol{\beta}_{ij}\right) \quad (3.1)$$

Here, the term $\mathbb{P}(\mathbf{G}_{ij}|\mathbf{G}_{mj}, \mathbf{G}_{fj})$ is the genotype transition probability of a mother with true genotype $\mathbf{G}_{mj}$ and a father with true genotype $\mathbf{G}_{fj}$ producing an offspring with true genotype $\mathbf{G}_{ij}$ at locus j.

In this study, we use a simple Mendelian transition/segregation model with no inter-generational mutation. In the diploid case, Mendelian transition probabilities are simple to calculate but a generalisation to higher ploidy requires extensive calculation for the many different combinations of gametic segregation and sexual recombination. We define $\mathbf{X}_{mj}$ and $\mathbf{X}_{fj}$ as random allele frequency vectors at locus j of a randomly selected gamete of the putative mother and father respectively. Like the genotype and phenotype vectors described previously, each of the A_j elements of $\mathbf{X}_{\cdot j}$ denote the frequency of the respective allele. Mendelian gamete segregation can be considered a form of sampling without replacement from the parental genotype where, given a ploidy, C_j, $\frac{C_j}{2}$ alleles are chosen sequentially from the pool of remaining allele types. In this sense, values for $\mathbf{X}_{mj}$ and $\mathbf{X}_{fj}$ are drawn from a multivariate hypergeometric distribution with probability mass functions $g_m(\mathbf{X}_{mj})$ and $g_f(\mathbf{X}_{fj})$ respectively where

$$g_{\cdot}(\mathbf{X}_{\cdot j}) = \begin{cases} \dfrac{\prod\limits_{a} \binom{\mathrm{G}_{\cdot j_a}}{\mathrm{X}_{\cdot j_a}}}{\binom{C_j}{\frac{C_j}{2}}} & \text{if } \sum\limits_{a} \mathrm{X}_{\cdot j_a} = \dfrac{C_j}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

The offspring genotype at a locus, $\mathbf{G}_{ij}$, is then simply the sum of the two random vectors, $\mathbf{X}_{mj}$ and $\mathbf{X}_{fj}$. The resultant probability mass function of the vector $\mathbf{G}_{ij}$ is therefore the multidimensional convolution of the probability mass functions of the parental gamete genotype vectors, that is, the sum of the probabilities of all possible combinations of the gamete allele frequency vectors that could combine to create the resultant genotype:

$$\begin{aligned} \mathbb{P}(\mathbf{G}_{ij}|\mathbf{G}_{mj}, \mathbf{G}_{fj}) &= (g_m * g_f)(\mathbf{G}_{ij}) \\ &= \sum_{\ell_1=1}^{\frac{C_j}{2}} \sum_{\ell_2=1}^{\frac{C_j}{2}} \cdots \sum_{l_{A_j}=1}^{\frac{C_j}{2}} g_m(\boldsymbol{\ell})\, g_f(\mathbf{G}_{ij} - \boldsymbol{\ell}) \end{aligned} \tag{3.3}$$

where $\boldsymbol{\ell} = \begin{bmatrix} \ell_1 \, \ell_2 \, \ldots \, \ell_{A_j} \end{bmatrix}^T$.

In equation 3.1 we make reference to the quantity, $\mathbb{P}(\mathbf{O}_{ij}|\mathbf{G}_{ij}, \boldsymbol{\beta}_{ij})$, the probability of observing offspring phenotype, $\mathbf{O}_{ij}$, at locus j given the true offspring genotype, $\mathbf{G}_{ij}$, or, in order

words, the observation model. $\boldsymbol{\beta}$ in equation 3.1 is a set of vectors of parameters for each combination of locus and sample, where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}, \ldots, \boldsymbol{\beta}_{1L}, \boldsymbol{\beta}_{21}, \boldsymbol{\beta}_{22}, \ldots, \boldsymbol{\beta}_{2L}, \ldots, \boldsymbol{\beta}_{SL}\}$, that control the observation model. For most applications there will be insufficient information *a priori* or garnered from the data to support loci and/or sample specific parameters for the observation model, and so it is likely that each vector in the set $\boldsymbol{\beta}$ will be equal. We however include the relevant notation in this model description for completeness and to provide a point of departure from the ideas contained within this paper for application to more complex scenarios where different loci may have vastly differing error rates. The two parental probabilities of the form $\mathbb{P}(\mathbf{G}_{\cdot j}|\mathbf{O}_{\cdot j}, \boldsymbol{\beta}_{\cdot j})$ can be written in terms of their observation model by application of Bayes Theorem:

$$\mathbb{P}(\mathbf{G}_{\cdot j}|\mathbf{O}_{\cdot j}, \boldsymbol{\beta}_{\cdot j}) = \frac{\mathbb{P}(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\beta}_{\cdot j})\,\mathbb{P}(\mathbf{G}_{\cdot j})}{\sum_{\boldsymbol{\ell}} \mathbb{P}(\mathbf{O}_{\cdot j}|\boldsymbol{\ell}, \boldsymbol{\beta}_{\cdot j})\,\mathbb{P}(\boldsymbol{\ell})} \tag{3.4}$$

Usually, prior support for the true genotype allele frequency vector at locus j, element $\mathbb{P}(\mathbf{G}_{\cdot j})$ in equation 3.4, is set to be uniform over the possible genotype combinations. For any genotype allele frequency vector $\mathbf{G}_{\cdot_j}$ at locus j, there may be a many number of different combinations of genotypes that may give rise to a particular allele frequency. For example, in the bialleleic diploid case, the genotype allele frequency vector $[1 \quad 1]^T$ for alleles a_1 and a_2, could arise from either an a_1 allele on the first homologous chromosome and an a_2 allele on the second homologous chromosome, or, *vice versa*. By treating the genotype allele frequency vector as a description of a multiset, it follows that the number of potential combinations of alleles across the homologous set for any genotype allele frequency vector corresponds to the multinomial coefficient. Assuming uniform support across all combinations of all possible genotype allele frequency vectors, the resultant prior distribution for any genotype allele frequency vector becomes the proportional number of genotype combinations that can arise from the given genotype allele frequency vector relative to the total number of combinations arising from all possible genotype allele frequency vectors:

$$\mathbb{P}\left(\mathbf{G}_{\cdot_j}\right) = \frac{\left(\dfrac{C_j!}{\mathrm{G}_{\cdot j_1}!\mathrm{G}_{\cdot j_2}!\ldots\mathrm{G}_{\cdot j_{A_j}}!}\right)}{\left(\displaystyle\sum_{\boldsymbol{\ell}} \frac{C_j!}{\ell_1!\ell_2!\ldots\ell_{A_j}!}\right)} \tag{3.5}$$

The number of potential true genotype combinations must be finite in order to calculate both the normalising constant in the denominator of equation 3.4 and the summations of equation 3.1. The total number of potential true genotypes at a locus is equal to ${A_j}^{C_j}$. Related as it is to both the total number of allele types identified at the locus and the ploidy, the total number

of potential true genotypes is only finite if both of these quantities are also finite.

Biallelic markers such as amplified fragment length polymorphism markers (AFLP; Vos *et al.*, 1995) only have two observable states at a given locus: positive and negative. This is commonly referred to as the phenotype although it can also be considered to be an imperfect observation of the true genotype. The pool of potential true genotypes is larger for all non-haploids however, for example if we denote + as a presence of the marker on a particular chromosomal copy and − as the absence of the marker, in the diploid case, the true genotype at a given AFLP locus could be $--$, $-+$, $+-$ or $++$. The number of potential true genotypes can potentially be very high in polyploid systems but the fact that AFLP exhibit only two potential allele values ensures that the number of potential genotype combinations is still finite.

Whilst the case for a finite number of potential true genotype combinations may be accurate in the case of genetic markers that have a restricted allelic set, such as amplified fragment length polymorphism or random amplified polymorphic DNA (RAPD; Williams *et al.*, 1990), it is more debatable for polyallelic markers such as microsatellites. The fragment length at a microsatellite locus is theoretically unbounded but experimentally it is only possible to observe fragment lengths between set limits. Only restricting the pool of potential alleles to values that lie within experimentally defined limits can still result in a large, and computationally infeasible set of combinations however. If some information is known about the schema of the repeat unit then it is possible to further reduce the number of potential allele values to those that are multiples of the repeat length. For the microsatellite loci used in this study, we assume that the pool of potential allele values at a particular locus is restricted only to those lengths observed in the genetic data across all individuals at the relevant locus.

3.2.2 Incorporating Genotyping Error

Random Replacement Methods

Early likelihood-based parentage methodologies discriminated between potential parent pairs based solely on genotype recombination probabilities. This technique in its purest form assumes that the observation of the genotypes is perfect: that no genotyping error occurs. The developments of Marshall *et al.* (1998), further revised in Kalinowski *et al.* (2007) and employed in the parentage software CERVUS, allow for some degree of observation error by assuming that all errors are the product of a random relabelling of genotypes at each locus. Under this

specification, the observation model at the j^{th} locus, $\mathbb{P}\left(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\beta}_{\cdot j}\right)$, takes the form:

$$\mathbb{P}\left(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\beta}_{\cdot j}\right) = \begin{cases} 1 - \epsilon_{\cdot j}\left(1 + p_{\mathbf{O}_{\cdot j}}\right) & \text{if } \mathbf{O}_{\cdot j} = \mathbf{G}_{\cdot j} \\ \epsilon_{\cdot j}\left(p_{\mathbf{O}_{\cdot j}}\right) & \text{otherwise} \end{cases} \tag{3.6}$$

where $\epsilon_{\cdot j}$ denotes the probability of a genotype substitution, and $p_{\mathbf{O}_{\cdot j}}$, the proportional frequency of phenotype $\mathbf{O}_{\cdot j}$, at locus j. Here, $\boldsymbol{\beta}_{\cdot j} = [\epsilon_{\cdot j}]$. If substitution occurs, the phenotype at that locus is replaced by another randomly selected phenotype from the pool of observed phenotypes. The case $\mathbf{O}_{\cdot j} = \mathbf{G}_{\cdot j}$ can transpire in one of two possible ways: either no substitution takes place, with probability $1 - \epsilon_{\cdot j}$, or, substitution takes place but it is a silent replacement, the phenotype is substituted for exactly the same phenotype with probability $\epsilon_{\cdot j}\left(p_{\mathbf{O}_{\cdot j}}\right)$. The case where $\mathbf{O}_{\cdot j} \neq \mathbf{G}_{\cdot j}$ can only result from a substitution for a non-equal phenotype in this model.

An alternative specification of the model involves the replacement phenotype not being drawn from the pool of observed phenotypes, but from the pool of genotypes expected under Hardy-Weinberg assumptions with the same allelic frequencies observed in the sampled population. This implementation, whilst increasing the pool of potential genotype combinations arising from a relabelling error, does so at the expense of assumptions of equilibrium. Sampled individuals can now be the receivers of genotypes arising from a theoretical population, and the scenario of alleles combining at a locus to generate new genotypes not represented in the sample is now possible. This can be very problematic if the sampled individuals do not appear to conform to this assumption of equilibrium; substitutions, when they happen, may be for genotype frequencies exhibiting significant differences in zygosity.

The phenotype is only a partial expression of the underlying phenotype however. Random relabelling of observations without regard to the genotypes that underlie the observation can complicate, rather than elucidate, the mechanism that generates sampling errors. Random relabelling models using substitution from a Hardy-Weinberg population will not work correctly if the true allele frequencies cannot be ascertained from the observations. In a dominant marker system such as AFLPs, a positive result only purports to a positive value occurring at least once across the homologous chromosomal copies, and not, to the frequency of that allele. Further methods, such as those of Zhivotovsky (1999), must be employed to infer the frequency of the positive allele. Implemented in their most basic form, random relabelling models treat all

dominant markers as haploid regardless of the actual ploidy. Even in codominant markers, the existence of null alleles can make it difficult to differentiate between a homozygote and a heterozygote with one or more copies of the null allele, creating bias in allele frequency estimation.

Biallelic Dominant Markers

AFLPs and RAPDs No over-arching observation model will be suitable for all marker types and so the first step for joint mixed-marker parentage analysis is to develop marker-specific error models. In the case of dominant markers, it is important to address the two sources of observation error: the possibility that an allele may be incorrectly diagnosed (a positive allele may be observed as a negative and, more rarely, a negative allele may be observed as a positive) and that dominance hierarchies will obscure the expression of recessive alleles. We define $\phi_{\cdot j}$ as the probability that a positive allele at locus j is misdiagnosed as a negative. An error that could arise through inability to extract enough high quality product (Watts *et al.*, 2007; Broquet & Petit, 2004; Gagneux *et al.*, 1997; Taberlet *et al.*, 1996) or through amplification failure from contamination by inhibitory agents (Opel *et al.*, 2010; Wilson, 1997). Similarly, we define $\psi_{\cdot j}$ as the probability that a negative allele at locus j is misdiagnosed as a positive allele. Although much less common, this error could arise through sample contamination or through the confusion of background fluorescence in the gels as band presence (Whitlock *et al.*, 2008). We define the random variables $A_{\cdot j+}$ and $A_{\cdot j-}$ as the number of alleles diagnosed as the dominant allele from the set of truly positive and negative alleles respectively. If we assume that allelic states are independent then:

$$A_{\cdot j+} \sim \text{Bin}\left(n_{\cdot j+}, 1 - \phi_{\cdot j}\right) \tag{3.7}$$

and:

$$A_{\cdot j-} \sim \text{Bin}\left(n_{\cdot j-}, \psi_{\cdot j}\right) \tag{3.8}$$

where $n_{\cdot j+}$ and $n_{\cdot j-}$ are the numbers of truly positive and negative alleles respectively at locus j. It then follows that a positive phenotype can be observed ($\mathbf{O}_{\cdot_{ij}} = \mathbf{+}$ where $\mathbf{+} = [1 \quad 0]^T$) either when at least one of the truly positive alleles are diagnosed as positive ($A_{\cdot j+} \geq 1$) or if

at least one of the truly negative alleles are misdiagnosed as a positive ($A_{\cdot j-} \geq 1$):

$$\begin{aligned}\mathbb{P}\left(\mathbf{O}_{\cdot j} = \mathbf{+}|\mathbf{G}_{\cdot j}, \boldsymbol{\beta}_{\cdot j}\right) &= \mathbb{P}\left(A_{\cdot j+} \geq 1 \cup A_{\cdot j-} \geq 1|\boldsymbol{\beta}_{\cdot j}\right) \\ &= 1 - \mathbb{P}\left(A_{\cdot j+} = 0|\phi_{\cdot j}\right)\mathbb{P}\left(A_{\cdot j-} = 0|\psi_{\cdot j}\right) \\ &= 1 - \phi_{\cdot j}^{n_{\cdot j+}}\left(1 - \psi_{\cdot j}\right)^{n_{\cdot j-}} \end{aligned} \tag{3.9}$$

Similarly, the instance of negative observation ($\mathbf{O}_{\cdot j} = \mathbf{-}$ where $\mathbf{-} = [0 \quad 1]^T$) can occur if all truly positive alleles are misdiagnosed as negative alleles ($A_{\cdot j+} = 0$) and all truly negative alleles are correctly diagnosed ($A_{\cdot j-} = 0$):

$$\begin{aligned}\mathbb{P}\left(\mathbf{O}_{\cdot j} = \mathbf{-}|\mathbf{G}_{\cdot j}, \boldsymbol{\beta}_{\cdot j}\right) &= \mathbb{P}\left(A_{\cdot j+} = 0 \cap A_{\cdot j-} = 0|\boldsymbol{\beta}_{\cdot j}\right) \\ &= \phi_{\cdot j}^{n_{\cdot j+}}\left(1 - \psi_{\cdot j}\right)^{n_{\cdot j-}} \end{aligned} \tag{3.10}$$

where $\boldsymbol{\beta}_{\cdot j} = [\phi_{\cdot j} \quad \psi_{\cdot j}]^T$. Combining the results of equations 3.9 and 3.10, and substituting $n_{\cdot j-} = C_j - n_{\cdot j+}$, we obtain

$$\mathbb{P}\left(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\beta}_{\cdot j}\right) = \begin{cases} 1 - \phi_{\cdot j}^{n_{\cdot j+}}\left(1 - \psi_{\cdot j}\right)^{C_j - n_{\cdot j+}} & \text{if } \mathbf{O}_{\cdot j} = \mathbf{+} \\ \phi_{\cdot j}^{n_{\cdot j+}}\left(1 - \psi_{\cdot j}\right)^{C_j - n_{\cdot j+}} & \text{if } \mathbf{O}_{\cdot j} = \mathbf{-} \end{cases} \tag{3.11}$$

Polyallelic Codominant Markers

A very different tactic must be employed to model codominant markers, although there are some parallels. More information pertaining to the genotype may be exposed in the phenotype of codominant markers but there still may be a number of different mechanisms by which the phenotype is observed given any true genotype. For example, the observation of one band in a diploid organism could represent a homozygote of the relevant allele, or, it could represent a heterozygote with one null allele. There are therefore a many number of possible vectors of genotype allele frequencies that may produce an observed phenotype even in the absence of errors of allele diagnosis. Here we define the potential observed genotype at locus j after allele diagnosis errors are taken into account as $\mathbf{M}_{\cdot j}$. This model formulation separates the observation process into two parts. The first process connects the true genotype allele frequency vector, $\mathbf{G}_{\cdot j}$, and the allele frequency vector after allele diagnosis errors have been made, $\mathbf{M}_{\cdot j}$. The second process describes the obscuring of multiple copies of the same allele type and the non-expression of null alleles in all but homozygotes, linking the allele frequency vector $\mathbf{M}_{\cdot j}$ to the phenotype frequency vector $\mathbf{O}_{\cdot j}$. Figure 3.2 illustrates this two-stage process for a number

of examples.

For the incorporation of errors of allele diagnosis, we define a series of vectors $\boldsymbol{\kappa}_{\cdot j_a}$, for each of the A_j allele types, including any null alleles, at a locus. Each of the vectors are of length A_j with each element of the vector $\kappa_{\cdot j_{a_b}}$ containing the probability of diagnosing allele a as allele b. Where $a = b$, the element of the relevant vector contains the probability of a correct diagnosis. Where $a \neq b$, a misdiagnosis has been made with probability $\kappa_{\cdot j_{a_b}}$. This formulation allows flexibility in not only misdiagnoses rates but also allows the weighting of certain types of misdiagnoses over others. Each copy of a given allele, a, present in the true genotype can be interpreted as any of the A_j allele types. The resultant vector of allele frequencies, $\mathbf{D}_{\cdot j_a}$, after allele diagnosis errors of the population of allele a in the true genotype ($\mathrm{G}_{\cdot j_a}$) have been taken into account, can be considered to follow a multinomial distribution with vector of diagnosis probabilities, $\boldsymbol{\kappa}_{\cdot j_a}$, with probability mass function

$$f_a\left(\mathbf{D}_{\cdot j_a}\right) = \begin{cases} \dfrac{\mathrm{G}_{\cdot j_a}!}{\mathrm{D}_{\cdot j_{a_1}}!\mathrm{D}_{\cdot j_{a_2}}!\ldots\mathrm{D}_{\cdot j_{a_{A_j}}}!}\kappa_{\cdot j_{a_1}}\kappa_{\cdot j_{a_2}}\ldots\kappa_{\cdot j_{a_{A_j}}} & \text{if } \sum_b \mathrm{D}_{\cdot j_{a_b}} = \mathrm{G}_{\cdot j_a} \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

Under this specification, the total vector of allele frequencies after diagnosis errors, $\mathbf{M}_{\cdot j}$, is the sum of the random diagnosis vectors for each allele type in the true genotype, $\sum_a \mathbf{D}_{\cdot j_a}$. The probability mass function of $\mathbf{M}_{\cdot j}$ is therefore the multivariate convolution of the probability mass function of the separate allele diagnosis vectors:

$$\mathbb{P}\left(\mathbf{M}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\kappa}_{\cdot j_1}, \boldsymbol{\kappa}_{\cdot j_2}, \ldots, \boldsymbol{\kappa}_{\cdot j_{A_j}}\right) = \left(\ldots((f_1 * f_2) * f_3)\cdots * f_{A_j}\right)(\mathbf{M}_{\cdot j}) \tag{3.13}$$

where $*$ is a convolution operator such that

$$\left(f_a * f_b\right)(\mathbf{x}) = \sum_{\ell_1=1}^{C_j}\sum_{\ell_2=1}^{C_j}\cdots\sum_{\ell_{A_j}=1}^{C_j} f_a(\mathbf{x})\, f_b(\mathbf{x} - \boldsymbol{\ell}) \tag{3.14}$$

and $\boldsymbol{\ell} = \begin{bmatrix}\ell_1\, \ell_2\, \ldots\, \ell_{A_j}\end{bmatrix}^T$.

The absence of a closed form for equation 3.13 means that numerical techniques must be employed in order to calculate values for the probability mass function of $\mathbf{M}_{\cdot j}$. Butler & Stephens (1993) describe a numerical method to exactly calculate the values from the resulting probability mass function of a random variable that is the sum of a series of binomially distributed

random variables with differing parameters controlling the probability of trial success. What follows is a multivariate extension of the algorithm of Butler & Stephens (1993) suitable for the calculation of random variable that is itself the sum of a series of multinomially distributed random variables.

Algorithm 1: Exact calculation of the probability mass function of a vector of allele frequencies of a potential phenotype after allele diagnosis error.

1. Initialise an A_j dimensional array λ with a length of $C_j + 1$ elements in each dimension and array indices starting at zero in each dimension. Set all elements of λ to zero.
2. Initialise a vector of iterators $\boldsymbol{\ell}$ of length A_j with all elements set to zero.
3. Initialise the iterator t_1 to A_j.
4. If the sum of the vector of iterators $\boldsymbol{\ell}$ equals $\mathrm{G}_{\cdot j_1}$ then set the element of λ at dimensional coordinates $\boldsymbol{\ell}$, $\lambda_{[\boldsymbol{\ell}]}$, equal to $f_1(\boldsymbol{\ell})$.
5. Increment ℓ_{t_1} by one.
6. If $t_1 > 0$ and $\ell_{t_1} > \mathrm{G}_{\cdot j_1}$ then set ℓ_{t_1} to zero and decrement t_1 by one before returning to step 5.
7. If $t_1 > 0$ then return to step 4.
8. Repeat for each value of a between 2 and A_j:
   (a) Initialise an A_j dimensional array θ with a length of C_j+1 elements in each dimension and array indices starting at zero in each dimension. Set all elements of θ to zero.
   (b) Initialise a vector of iterators $\boldsymbol{\tau}$ of length A_j with all elements set to zero.
   (c) Initialise the iterator t_2 to A_j.
   (d) If the sum of the vector of iterators $\boldsymbol{\tau}$ equals $\mathrm{G}_{\cdot j_a}$ then
      i. Set t_1 equal to zero.
      ii. Set all the elements of $\boldsymbol{\ell}$ equal to zero.
      iii. Increment the element $\theta_{[\boldsymbol{\ell}+\boldsymbol{\tau}]}$ by $f_a(\boldsymbol{\tau})\,\lambda_{[\boldsymbol{\ell}]}$.
      iv. Increment ℓ_{t_1} by one.
      v. If $t_1 > 0$ and $\ell_{t_1} > \sum_{b=1}^{a-1} \mathrm{G}_{\cdot j_b}$ then set ℓ_{t_1} to zero and decrement t_1 by one before returning to step 8(d)iv.

vi. If $t_1 > 0$ then return to step 8(d)iii.

(e) Increment τ_{t_2} by one.

(f) If $t_2 > 0$ and $\tau_{t_2} > \mathrm{G}_{\cdot j_a}$ then set τ_{t_2} to zero and decrement t_2 by one before returning to step 8e.

(g) If $t_2 > 0$ then return to step 8d.

(h) Set the values of the elements of λ equal to the values of the respective elements of θ.

9. Return the multidimensional array λ as the probability mass function of the random vector $\mathbf{M}_{\cdot j}$ reading

$$\mathbb{P}\left(\mathbf{M}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\kappa}_{\cdot j_1}, \boldsymbol{\kappa}_{\cdot j_2}, \ldots, \boldsymbol{\kappa}_{\cdot j_{A_j}}\right) = \lambda_{[\mathbf{M}_{\cdot j}]} \tag{3.15}$$

The final stage of the model is to link the possible genotypes after allele diagnosis errors, $\mathbf{M}_{\cdot j}$, to the observed phenotype $\mathbf{O}_{\cdot j}$. To do this, it is necessary to identify those combinations of genotypes that could produce the phenotype of interest: the observation of a diploid species with only one allele observation at a locus could result from either homozygosity of the observed allele, or from a heterozygote with one null allele. More generally, any observations for which fewer alleles are observed than the ploidy requires that there exist more than one possible genotype allele frequency vector from which the phenotype would arise. Any extra unobserved alleles could either be recessive in nature or repeats of those from the set of observed alleles. If we denote the set of possible genotype allele frequency vectors at locus j as $H_{\cdot j}$, then to incorporate this extra uncertainty into the model it is necessary to sum over the set of possible genotypes, weighting each by its probability in light of allele diagnosis errors, to derive the final probability of a given phenotype:

$$\mathbb{P}\left(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\kappa}_{\cdot j_1}, \boldsymbol{\kappa}_{\cdot j_2}, \ldots, \boldsymbol{\kappa}_{\cdot j_{A_j}}\right) = \sum_{\mathbf{M}_{\cdot j} \in H_{\cdot j}} \mathbb{P}\left(\mathbf{M}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\kappa}_{\cdot j_1}, \boldsymbol{\kappa}_{\cdot j_2}, \ldots, \boldsymbol{\kappa}_{\cdot j_{A_j}}\right) \tag{3.16}$$

It is from these basic building blocks, using the algorithm described above to compute the solutions to equation 3.13 and hence solve equation 3.16, that form the basis for a series of models suitable for different kinds of codominant markers. Assuming particular functional forms for the allele diagnosis probabilities leads to a series of specialisations suitable for use in various marker-specific implementations.

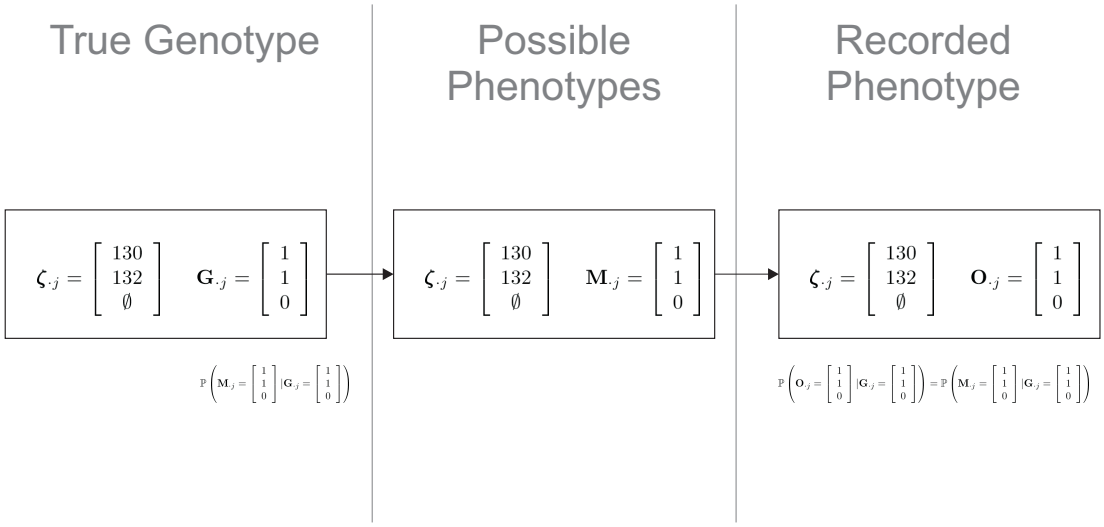
VNTRs and RFLPs In the case of alleles with quantitative traits such as fragment size, as is the case when using microsatellite, ministaellites, or RFLP markers, there exists extra information contained within the scoring process that can inform probabilities related to diagnostic error rates. It is worth noting that if, for example, a heterozygote with fragments of length 130 and 140 are observed, neither necessarily have to be correctly diagnosed. Sources of error here arise from confusion over the quantitative trait of an allele, simple misreading of fragment location or the presence of disturbances such as 'stutter bands' present in some samples (Hoffman & Amos, 2005; Ginot *et al.*, 1996), resulting in the mis-classification of alleles. Despite the introduction of many automated technologies for band analysis, these error rates remain high (Ewen *et al.*, 2000; Ginot *et al.*, 1996). Unlike the random relabelling model, an allele that has been erroneously classified is likely have very similar quantitative traits to the allele that has been falsely substituted for the true allele. A putative observation model for microsatellite allele fragment length, might, for example, use a discrete implementation of the normal distribution defined only over the possible fragment length values, so that the probability of confusing allele a with allele b, $p_{\cdot j_{ab}}$, given that they have lengths $\zeta_{\cdot j_a}$ and $\zeta_{\cdot j_b}$ respectively is given by

$$p_{\cdot j_{ab}} = \frac{\Phi\left[\frac{1}{\sigma_{\cdot j}}\left(\zeta_{\cdot j_b} - \zeta_{\cdot j_a} + \frac{1}{2}\right)\right] - \Phi\left[\frac{1}{\sigma_{\cdot j}}\left(\zeta_{\cdot j_b} - \zeta_{\cdot j_a} - \frac{1}{2}\right)\right]}{\sum_{k\neq\emptyset}\Phi\left[\frac{1}{\sigma_{\cdot j}}\left(\zeta_{\cdot j_k} - \zeta_{\cdot j_a} + \frac{1}{2}\right)\right] - \Phi\left[\frac{1}{\sigma_{\cdot j}}\left(\zeta_{\cdot j_k} - \zeta_{\cdot j_a} - \frac{1}{2}\right)\right]} \tag{3.17}$$

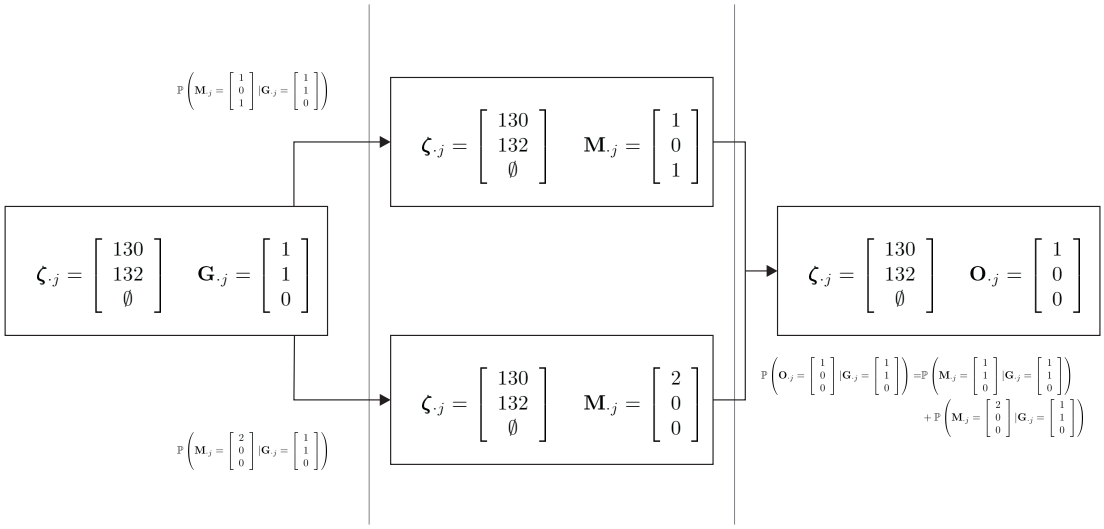
$\Phi(x)$ is the normal distribution function. Here we use the notation $\emptyset$ to denote the null allele so that the denominator of equation 3.17 is the summation over all non-null alleles. The pa-

Figure 3.2 *(on the next page)*: Illustration of the possible combinations of codominant error model outputs, ($\mathbf{M}_{\cdot j}$), that can generate an observed allele frequency vector, ($\mathbf{O}_{\cdot j}$), given a true genotype allele frequency vector, ($\mathbf{G}_{\cdot j}$) with three example scenarios from the diploid case given a vector of fragment lengths that define the respective alleles ($\boldsymbol{\zeta}_{\cdot j} = [130 \quad 132 \quad \emptyset]^T$ where $\emptyset$ is a null allele). The first scenario (a) describes the simple case of the observation of a heterozygous phenotype, with one observed fragment of 132 base units and another at 130 base units. The number of fragments observed is equal to the ploidy and so there is only one possible set of alleles which would result in this observation ($\mathbf{M}_{\cdot j} = [1 \quad 1 \quad 0]^T$). The probability of observing the phenotype given the true genotype is therefore the possibility that the alleles that make up the genotype are diagnosed with frequency given by the one possible set of alleles ($\mathbf{M}_{\cdot j}$). The second scenario, (b), where a homozygous phenotype with allele length 130 is observed, could result from either a genuine homozygous error model output ($\mathbf{M}_{\cdot j} = [2 \quad 0 \quad 0]^T$), or from a heterozygote with one null allele ($\mathbf{M}_{\cdot j} = [1 \quad 0 \quad 1]$). The probability of observing allele frequency vector, $\mathbf{O}_{\cdot j}$, given a genotype, $\mathbf{G}_{\cdot j}$, is therefore the sum of the probabilities of diagnosing alleles with frequencies $[2 \quad 0 \quad 0]^T$ or $[1 \quad 0 \quad 1]^T$. Where one null allele is present in the true genotype such as the third scenario (c), outcomes which require the misidentification of the null allele are impossible (denoted by the dotted lines) in the observation models described for codominant markers in this paper, leaving only one error model outcome that could produce the observed phenotype.

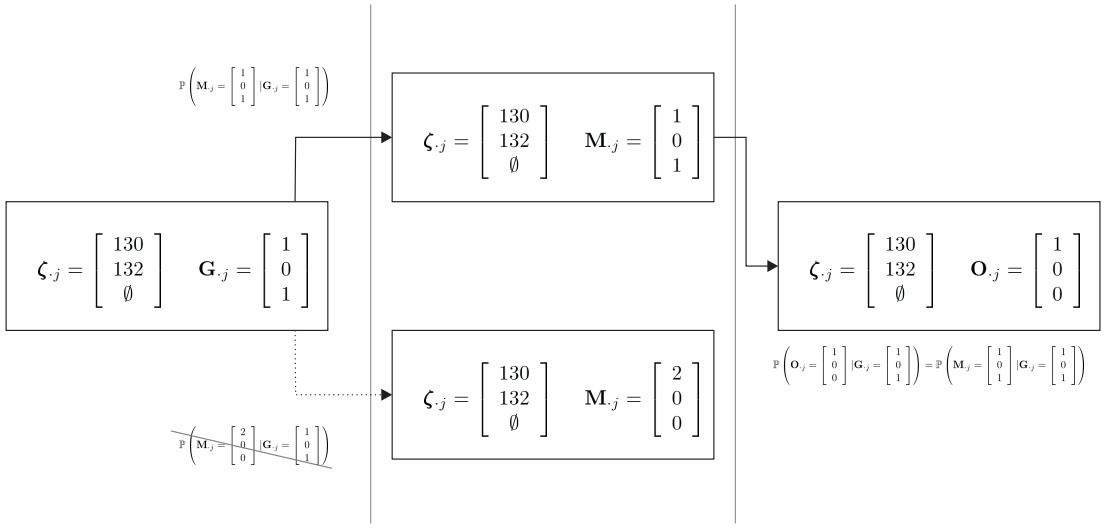
(a)


True Genotype
Possible Phenotypes
Recorded Phenotype
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{O}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$
$\mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\right)$
$\mathbb{P}\left(\mathbf{O}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\right) = \mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\right)$

(b)


$\mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\right)$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{O}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{M}_{\cdot j} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$
$\mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\right)$
$\mathbb{P}\left(\mathbf{O}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\right) = \mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\right)$
$+ \mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\right)$

(c)


$\mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right)$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{O}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$
$\boldsymbol{\zeta}_{\cdot j} = \begin{bmatrix} 130 \\ 132 \\ \emptyset \end{bmatrix}$ $\mathbf{M}_{\cdot j} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$
$\mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right)$
$\mathbb{P}\left(\mathbf{O}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right) = \mathbb{P}\left(\mathbf{M}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} | \mathbf{G}_{\cdot j} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right)$

rameter $\sigma_{.j}$ controls the possible variation in fragment length observation: as $\sigma_{.j} \to 0$ the range of fragment size errors diminishes until all probability weight is given to the correct diagnosis where $a = b$. The distribution tends to a uniform distribution across the set of possible allele values as $\sigma_{.j} \to \infty$.

Errors arising from fragment binning does not tell the whole story however. Like the model for dominant markers, fragments can fail to be detected, either through amplification or extraction error, or through inhibition caused by contaminants. This random dropout rate can be incorporated into the model by the parameter $v_{.j}$. This model therefore distinguishes between the scenario where random dropouts give the appearance of a null homozygote and the situation where genetic mutations at the primer binding site result in a genuine, genotype-driven, band absence. Assuming that null alleles can only be correctly diagnosed as null alleles, the final transition probabilities take the form

$$\kappa_{.j_{a_b}} = \omega_{\emptyset}(a)\,\omega_{\emptyset}(b) + (1 - \omega_{\emptyset}(a))\left[\omega_{\emptyset}(b)\,v_{.j} + (1 - \omega_{\emptyset}(b))\,(1 - v_{.j})\,p_{.j_{ab}}\right] \tag{3.18}$$

where $\omega_{\emptyset}(a)$ is an indicator function:

$$\omega_{\emptyset}(a) = \begin{cases} 1 & \text{if } a \text{ is a null allele} \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

SNPs Single Nucleotide Polymorphisms (SNPs) are markers that use polymorphisms at a particular nucleotide to derive useful information of the system of interest. For the most part, SNPs have two codominant alleles, representing the two nucleotide states and a recessive null allele (Carlson *et al.*, 2006). Recent studies have shown that many commonly used SNP markers in human populations may actually have three codominant states (Hüebner *et al.*, 2007; Hodgkinson & Eyre-Walker, 2010) with the possibility that some may even have the maximum four codominant states (Brookes, 1999).

Regardless of the number of codominant alleles, the observation model for SNP markers uses the same derivation from the general framework for codominant markers described previously. Here, we define the parameter, $\xi_{.j}$, to denote the random dropout rate, or the probability as failing to observe a non-null allele, at locus j. Given that a random dropout error does not occur then another type of error, misclassification of the correct allele type, may occur with probability $\gamma_{.j}$. In the incidence of a classification error we assume that the replacement allele

type is drawn uniformally from the remaining pool of non-null alleles such that the probability of selecting from each of the candidate alleles is $\frac{1}{A_j-2}$. Combining these elements of observation error results in the following formulation for the elements of $\boldsymbol{\kappa}_{\cdot j_a}$:

$$\kappa_{\cdot j_{a_b}} = \omega_{\emptyset}(a)\,\omega_{\emptyset}(b) + (1-\omega_{\emptyset}) \left[\begin{array}{l} \omega_{\emptyset}(b)\,\xi_{\cdot j} + (1-\omega_{\emptyset}(b))\,(1-\xi_{\cdot j}) \\ \left[\omega_a(b)\,(1-\gamma_{\cdot j}) + (1-\omega_a(b))\,\frac{\gamma_{\cdot j}}{A_j-2}\right] \end{array} \right] \tag{3.20}$$

where like $\omega_{\emptyset}(x)$ (see equation 3.19), $\omega_a(b)$ is another indicator function such that

$$\omega_a(b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

Equation	*Parameters*			*Submodel variants*	

Perfect Observation

P1: Observed phenotype is a perfect representation of the genotype

$$\mathbb{P}\left(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}\right) = \begin{cases} 1 & \text{if } \mathbf{O}_{\cdot j} = \mathbf{G}_{\cdot j} \\ 0 & \text{otherwise} \end{cases}$$

Random Relabelling Error

R1: A genotyping error causes a random substitution of another genotype (see equation 3.6)

$$\mathbb{P}\left(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}\right) = \begin{cases} 1 + \epsilon_{\cdot j}\left(p_{\mathbf{O}_{\cdot j}}\right) & \text{if } \mathbf{O}_{\cdot j} = \mathbf{G}_{\cdot j} \\ \epsilon_{\cdot j}\left(p_{\mathbf{O}_{\cdot j}}\right) & \text{otherwise} \end{cases}$$

$\epsilon_{\cdot j}$	Probability that a random substitution is made	$\epsilon_{\cdot j} \in (0,1)$
$p_{\mathbf{O}_{\cdot j}}$	Probability that $\mathbf{O}_{\cdot j}$ is selected as the observed phenotype under a random replacement	$p_{\mathbf{O}_{\cdot j}} \in (0,1)$

R1-1:	$p_{\mathbf{O}_{\cdot j}}$ is sample frequency of $\mathbf{O}_{\cdot j}$
R1-2:	$p_{\mathbf{O}_{\cdot j}}$ is Hardy-Weinberg frequency of $\mathbf{O}_{\cdot j}$

Observation Models for Dominant Markers

D1: Model for biallelic markers such as AFLPs or RAPDs (see equation 3.11)

$$\mathbb{P}\left(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\beta}_{\cdot j}\right) = \begin{cases} 1 - \phi_{\cdot j}^{n_{\cdot j+}}\left(1-\psi_{\cdot j}\right)^{C_j - n_{\cdot j+}} & \text{if } \mathbf{O}_{\cdot j} = \mathbf{+} \\ \phi_{\cdot j}^{n_{\cdot j+}}\left(1-\psi_{\cdot j}\right)^{C_j - n_{\cdot j+}} & \text{if } \mathbf{O}_{\cdot j} = \mathbf{-} \end{cases}$$

$\phi_{\cdot j}$	Probability that a positive allele is misdiagnosed as a negative allele	$\phi_{\cdot j} \in (0,1)$
$\psi_{\cdot j}$	Probability that a negative allele is misdiagnosed as a positive allele	$\psi_{\cdot j} \in (0,1)$

D1-0:	Full model
D1-1:	$\phi_{\cdot j} = 0$
D1-2:	$\psi_{\cdot j} = 0$

Observation Models For Codominant Markers

For all codominant markers

$$\mathbb{P}\left(\mathbf{O}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\kappa}_{\cdot j_1}, \boldsymbol{\kappa}_{\cdot j_2}, \ldots, \boldsymbol{\kappa}_{\cdot j_{A_j}}\right) = \sum_{\mathbf{M}_{\cdot j} \in H_{\cdot j}} \mathbb{P}\left(\mathbf{M}_{\cdot j}|\mathbf{G}_{\cdot j}, \boldsymbol{\kappa}_{\cdot j_1}, \boldsymbol{\kappa}_{\cdot j_2}, \ldots, \boldsymbol{\kappa}_{\cdot j_{A_j}}\right)$$

where $\mathbf{M}_{\cdot j}$ is a random vector of alleles that could be observed as $\mathbf{O}_{\cdot j}$ after diagnostic errors are taken into account. $\mathbf{M}_{\cdot j} = \sum_a \mathbf{D}_{\cdot j_a}$ where

$$\mathbb{P}\left(\mathbf{D}_{\cdot j_a} | \mathrm{G}_{\cdot j_a}, \boldsymbol{\kappa}_{\cdot j_a}\right) = \begin{cases} \dfrac{\mathrm{G}_{\cdot j_a}!}{\mathrm{D}_{\cdot j_{a_1}}! \mathrm{D}_{\cdot j_{a_2}}! \ldots \mathrm{D}_{\cdot j_{a_{A_j}}}!} \kappa_{\cdot j_{a_1}} \kappa_{\cdot j_{a_2}} \ldots \kappa_{\cdot j_{a_{A_j}}} & \text{if } \sum_b \mathrm{D}_{\cdot j_{a_b}} = \mathrm{G}_{\cdot j_a} \\ 0 & \text{otherwise} \end{cases}$$

C1: Model for codominant markers with alleles binned according to fragment length such as VNTRs or RFLPs (see equation 3.18)

$$\kappa_{\cdot j_{a_b}} = \omega_\emptyset(a)\,\omega_\emptyset(b) + (1 - \omega_\emptyset(a))\left[\omega_\emptyset(b)\, v_{\cdot j} + (1 - \omega_\emptyset(b))(1 - v_{\cdot j})\, p_{\cdot j_{ab}}\right]$$

$$p_{\cdot j_{ab}} = \frac{\Phi\left[\frac{1}{\sigma_{\cdot j}}\left(\zeta_{\cdot j_b} - \zeta_{\cdot j_a} + \frac{1}{2}\right)\right] - \Phi\left[\frac{1}{\sigma_{\cdot j}}\left(\zeta_{\cdot j_b} - \zeta_{\cdot j_a} - \frac{1}{2}\right)\right]}{\sum_{k \neq \emptyset} \Phi\left[\frac{1}{\sigma_{\cdot j}}\left(\zeta_{\cdot j_k} - \zeta_{\cdot j_a} + \frac{1}{2}\right)\right] - \Phi\left[\frac{1}{\sigma_{\cdot j}}\left(\zeta_{\cdot j_k} - \zeta_{\cdot j_a} - \frac{1}{2}\right)\right]}$$

$v_{\cdot j}$	The probability of allelic dropout	$v_{\cdot j} \in (0,1)$	C1-0:	Full model
$\sigma_{\cdot j}$	The variance in allele length assignment	$\sigma_{\cdot j} \in \mathbb{R}_+$	C1-1:	$v_{\cdot j} = 0$
			C1-2:	$\sigma_{\cdot j} \to 0$

C2: Model for codominant markers with alleles that are not assigned according to fragment size such as SNPs (see equation 3.20)

$$\kappa_{\cdot j_{a_b}} = \omega_\emptyset(a)\,\omega_\emptyset(b) + (1 - \omega_\emptyset)\left[\begin{array}{l} \omega_\emptyset(b)\,\xi_{\cdot j} + (1 - \omega_\emptyset(b))(1 - \xi_{\cdot j}) \\ \left[\omega_a(b)(1 - \gamma_{\cdot j}) + (1 - \omega_a(b))\frac{\gamma_{\cdot j}}{A_j - 2}\right] \end{array}\right]$$

$\xi_{\cdot j}$	The probability of allelic dropout	$\xi_{\cdot j} \in (0,1)$		
$\gamma_{\cdot j}$	The probability of misclassifying an allele for a another random non-null allele	$\gamma_{\cdot j} \in (0,1)$	C2-0:	Full model
			C2-1:	$\xi_{\cdot j} = 0$
			C2-2:	$\gamma_{\cdot j} = 0$

Table 3.1: A selection of genotyping error models described in this study. All models are used to calculate the probability of observing allele frequency vector, $\mathbf{O}_{\cdot j}$, given a vector of 'true' allele frequencies present in the genotype, $\mathbf{G}_{\cdot j}$ at locus j. The models are categorised into those that are suitable for biallelic dominant markers (such as AFLP or RAPD markers) and those that are more suitable for polyallelic markers with co-dominant alleles (such as microsatellite markers). Each model has a selection of sub-models, each with a reference code, where one or more parameters are fixed or redefined. In addition to the variable parameters in the models, there are number of constants that are defined in table 3.2.

Constant	*Description*
$n_{\cdot j+}$	Number of positive alleles in the true genotype at locus j
C_j	Ploidy at locus j
H_j	The set of possible allele frequency vectors that, after diagnosis errors are taken into account, could result in the observation of frequency vector, $\mathbf{O}_{\cdot j}$, at locus j
$\omega_{\emptyset}(a)$	An identity function that equals one if allele a is a null allele, zero otherwise
$\zeta_{\cdot j_a}$	The fragment length of allele a at locus j
$\omega_a(b)$	An identity function that equals one if allele a is the same allele type as allele b, zero otherwise
A_j	The number of different allele types at locus j

Table 3.2: A list of constants used in the error models described in table 3.1.

Table 3.1 summarises the methods described in this section as well defining some fixed parameter sub-model variants.

3.2.3 Parentage Sampling Algorithm

The derivation of the likelihood equation described in equation 3.1 allows, through application of Bayes theorem, the calculation of the posterior probability of a parentage pair given the phenotype observations and a vector of parameters, $\boldsymbol{\beta}$:

$$\mathbb{P}\left(♀_i = m, ♂_i = f | \mathbf{O}_i, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) = \frac{\mathbb{P}\left(\mathbf{O}_i | ♀_i = m, ♂_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) \mathbb{P}\left(♀_i = m, ♂_i = f\right)}{\sum_a \sum_b \mathbb{P}\left(\mathbf{O}_i | ♀_i = a, ♂_i = b, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) \mathbb{P}\left(♀_i = a, ♂_i = b\right)} \tag{3.22}$$

The quantity, $\mathbb{P}\left(♀_i = m, ♂_i = f\right)$, represents the prior probability that individual m is the mother and individual f is the father of offspring i. In many applications there will not be any prior information regarding probabilities of possible parentage combinations but it is at this stage that it is possible to incorporate known incompatibilities in the mating system by giving such matches a zero weight. For example, many plant species are self-incompatible and so, in the analysis of a species that exhibits such characteristics, $\mathbb{P}\left(♀_i = a, ♂_i = a\right) = 0$ for all a. If we define the term 'mother' to mean 'seed donator' and 'father' to mean pollen donator in the plant sexual system then it is also possible to exclude androecious plants from being the 'mother' and gynoecious plants from being the 'father' at this stage.

When the set of vectors of observation model parameters, $\boldsymbol{\beta}$, are fixed then it is possible to calculate the posterior probabilities for each parentage pair directly using equation 3.22. However, it is unlikely that genotyping error rates are known precisely for any given system. It is possible, however, to extend the analysis to jointly estimate values for $\boldsymbol{\beta}$ from the data as long as previous knowledge of the parameters of the genotyping error rates can be expressed as a prior, $\mathbb{P}\left(\boldsymbol{\beta}\right)$. The posterior density we need to evaluate is now

$$\mathbb{P}\left(♀_i = m, ♂_i = f, \boldsymbol{\beta} | \mathbf{O}_i, \mathbf{O}_m, \mathbf{O}_f\right) = \frac{\mathbb{P}\left(\mathbf{O}_i | ♀_i = m, ♂_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) \mathbb{P}\left(♀_i = m, ♂_i = f\right) \mathbb{P}\left(\boldsymbol{\beta}\right)}{\int_{V_{\boldsymbol{\beta}}} \sum_a \sum_b \mathbb{P}\left(\mathbf{O}_i | ♀_i = a, ♂_i = b, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) \mathbb{P}\left(♀_i = a, ♂_i = b\right) \mathbb{P}\left(\boldsymbol{\beta}\right) \, dV_{\boldsymbol{\beta}}} \tag{3.23}$$

where $V_{\boldsymbol{\beta}}$ is a volume of integration over all possible values of the elements of the comprising vectors of $\boldsymbol{\beta}$.

The new posterior density of equation detailed in equation 3.23 is now sufficiently complex to be difficult to evaluate directly. Instead, possible parentage combinations and values for $\boldsymbol{\beta}$ can be sampled from the posterior distribution using Markov Chain Monte Carlo techniques. Deriving the for the full conditional distributions of $\boldsymbol{\beta}$ for each of the separate observation models is not a trivial task and so this means that it is not possible to use Gibbs sampling to draw these values and so we have instead implemented a single-update Metropolis-Hastings algorithm (see Chib & Greenberg, 1995) for this purpose.

The Metropolis-Hastings algorithm requires the generation of proposal values of the parameters of interest, denoted by $\boldsymbol{\beta}^*$, $\mathrm{♀}_i^*$, and $\mathrm{♂}_i^*$, which are accepted as samples from the target distribution with a probability related to the ratios of their posterior support compared to that of samples generated in the previous iteration of the algorithm. The proposal values are generated from a proposal density. , $\mathbb{P}\left(\mathrm{♀}_i^*, \mathrm{♂}_i^*, \boldsymbol{\beta}^* | \mathrm{♀}_i, \mathrm{♂}_i, \boldsymbol{\beta}\right)$, which describes the probability of generating the candidate values given the last iteration's sample values for $\mathrm{♀}_i$, $\mathrm{♂}_i$, and $\boldsymbol{\beta}$. As long as the proposal probabilities for values that have some posterior support are greater than zero, the exact form of the proposal density does not affect the eventual convergence of the algorithm to the target distribution. The proposal density does however control the efficiency to which samples from the Markov chain convergence towards the posterior distribution. For the generation of proposed values for vectors of observation model parameters we have implemented a multivariate truncated normal distribution (see Horrace, 2005) with probability density function

$$\mathbb{P}\left(\boldsymbol{\beta}^* | \boldsymbol{\beta}, \boldsymbol{\Sigma}\right) = \frac{e^{\left[-\frac{1}{2}\left(\mathbf{K}_{\boldsymbol{\beta}^*}-\mathbf{K}_{\boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{K}_{\boldsymbol{\beta}^*}-\mathbf{K}_{\boldsymbol{\beta}}\right)\right]}}{\int_{L_1^-}^{L_1^+}\int_{L_2^-}^{L_2^+}\cdots\int_{L_N^-}^{L_N^+} e^{\left[-\frac{1}{2}\left(\mathbf{K}_{\boldsymbol{\beta}^*}-\mathbf{K}_{\boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{K}_{\boldsymbol{\beta}^*}-\mathbf{K}_{\boldsymbol{\beta}}\right)\right]} d\left[\mathrm{K}_{\boldsymbol{\beta}^*}\right]_1 \, d\left[\mathrm{K}_{\boldsymbol{\beta}^*}\right]_2 \ldots d\left[\mathrm{K}_{\boldsymbol{\beta}^*}\right]_N} \tag{3.24}$$

where $\mathbf{K}_{\boldsymbol{\beta}}$ and $\mathbf{K}_{\boldsymbol{\beta}^*}$ are respecifications of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ respectively so that the entire set of observation parameters in all samples and at all loci are laid out vertically in one column vector of length N:

$$\mathbf{K}_{\boldsymbol{\beta}} = \begin{bmatrix}\boldsymbol{\beta}_{11}^T & \boldsymbol{\beta}_{12}^T & \ldots & \boldsymbol{\beta}_{1L}^T & \boldsymbol{\beta}_{21}^T & \boldsymbol{\beta}_{22}^T & \ldots & \boldsymbol{\beta}_{2L}^T & \ldots & \boldsymbol{\beta}_{SL}^T\end{bmatrix}^T \tag{3.25}$$

$\left[\mathrm{K}_{\boldsymbol{\beta}^*}\right]_i$ represents the i^{th} element of the vector $\mathbf{K}_{\boldsymbol{\beta}^*}$. The two vectors $\mathbf{L}^- = \begin{bmatrix}L_1^- & L_2^- & \ldots & L_N^-\end{bmatrix}^T$

and $\mathbf{L}^+ = \begin{bmatrix} L_1^+ & L_2^+ & \dots L_N^+ \end{bmatrix}$ contain the limits, upper and lower respectively, for each of the N parameters in $\mathbf{K}_{\boldsymbol{\beta}}$. The limits need not necessarily be finite. Finally, $\boldsymbol{\Sigma}$, is the variance-covariance matrix for the proposal distribution of new parameters. To sample prospective values for each of the parameters independently, only the diagonal elements need be set to non-zero values. This is the best tactic for most applications. However, setting off-diagonal values of the variance-covariance matrix to non-zero values would allow for the efficient sampling of parameters where there is a known colinearity in the likelihood surface for one or more of the parameters. Geweke (1991) describes an algorithm for the efficient sampling from multivariate truncated normal distributions.

One simple method to propose candidate parentage pairs for a given offspring is to select uniformally amongst the parentage pairs with each iteration of the Metropolis-Hastings algorithm. However, except when genotyping errors are set exceptionally high, the weighting of parentage likelihoods will be tightly constrained around the true parent pair. Selecting proposal parentage pairs using a simple uniform distribution may therefore produce a very high rejection rate and inefficient convergence times. However, in equation 3.22 we are presented with the full conditional distribution of the parentage pairs given a set of values for $\boldsymbol{\beta}$. It is also possible to sample values with a probability mass function corresponding to this conditional distribution by sampling from a categorical distribution with categories defined as each of the possible parentage pairs and a probability parameter vector with elements calculated using equation 3.22. This method of parameter selection for possible parentage pairs represents the Gibbs sampling step within a Metropolis-within-Gibbs sampling algorithm (Tierney, 1994).

Algorithm 2: Metropolis-within-Gibbs sampling algorithm for the generation of samples from the joint posterior distribution of parentage and observation model parameters for a given offspring.

1. Initialise the chain with arbitrary values for $♂_i$, $♀_i$, and elements of the vector set $\boldsymbol{\beta}$, ensuring that $\mathbb{P}\left(\mathbf{O}_i | ♀_i = m, ♂_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) > 0$, $\mathbb{P}\left(♀_i = m, ♂_i = f\right) > 0$, and $\mathbb{P}\left(\boldsymbol{\beta}\right) > 0$.

2. *Gibbs Step*: Draw a new parentage combination ($♀_i$ and $♂_i$) from a categorical distribution with probability vector elements set according to equation 3.22 conditional on the current values for $\boldsymbol{\beta}$.

3. *Metropolis-Hastings Step*: Propose a set of new values for the genotype observation model $\boldsymbol{\beta}^*$ from a multivariate truncated normal distribution with probability density function given in equation 3.24.

4. Calculate

$$u = \frac{\mathbb{P}\left(\mathbf{O}_i|\mathbb{Q}_i = m, \mathbb{O}_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}^*\right)\mathbb{P}\left(\mathbb{Q}_i = m, \mathbb{O}_i = f\right)\mathbb{P}\left(\boldsymbol{\beta}^*\right)\mathbb{P}\left(\boldsymbol{\beta}|\boldsymbol{\beta}^*, \boldsymbol{\Sigma}\right)}{\mathbb{P}\left(\mathbf{O}_i|\mathbb{Q}_i = m, \mathbb{O}_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right)\mathbb{P}\left(\mathbb{Q}_i = m, \mathbb{O}_i = f\right)\mathbb{P}\left(\boldsymbol{\beta}\right)\mathbb{P}\left(\boldsymbol{\beta}^*|\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)} \tag{3.26}$$

5. Generate a random uniform value, α, between zero and one. If $\alpha < \min\{1, u\}$ then set the current vectors of parameter values, $\boldsymbol{\beta}$, to the proposed values $\boldsymbol{\beta}^*$.

6. Store the current values of $\mathbb{Q}$, $\mathbb{O}$, and $\boldsymbol{\beta}$ as samples from the target distribution of equation 3.23.

7. Return to step 2.

The implementation of the algorithm above and equations 3.22 and 3.23 assume that the parentage conditions for each offspring are independent and, as such, parentage estimations can be made for each offspring individual separately. If the same pool of individuals are used as potential parents for each of the offspring then this assumption can be violated. For situations where it is likely that the number of offspring to be assigned parentage lie in the upper end or exceed the reproductive potential of an individual then it is unlikely that one individual can be the parent to all offspring in the group. Also, by separating parentage estimation we are also separating the estimation of observation error model parameters. Except in the very rare case where no overlap exists for the pool of potential parents for each offspring and it is expected that the parameters for the observation error model will be significantly different for each of the offspring, it is much more reasonable to include information from the parentage performed for each of the offspring to estimate the parameters of the observation model. In these situations it is necessary to jointly update the entire parentage vectors, $\mathbb{Q} = \begin{bmatrix}\mathbb{Q}_1 & \mathbb{Q}_2 & \dots & \mathbb{Q}_W\end{bmatrix}^T$ and $\mathbb{O} = \begin{bmatrix}\mathbb{O}_1 & \mathbb{O}_2 & \dots & \mathbb{O}_W\end{bmatrix}^T$, where W is the total number of offspring. The new target distribution then becomes

$$\mathbb{P}\left(\mathbb{Q} = \mathbf{m}, \mathbb{O} = \mathbf{f}, \boldsymbol{\beta}|\mathbf{O}\right) = \frac{\mathbb{P}\left(\mathbb{Q} = \mathbf{m}, \mathbb{O} = \mathbf{f}\right)\mathbb{P}\left(\boldsymbol{\beta}\right)\prod_i \mathbb{P}\left(\mathbf{O}_i|\mathbb{Q}_i = m_i, \mathbb{O}_i = f_i, \mathbf{O}_{m_i}, \mathbf{O}_{f_i}, \boldsymbol{\beta}\right)}{\int_{V_{\boldsymbol{\beta}}} \sum_{\mathbf{a}}\sum_{\mathbf{b}} \mathbb{P}\left(\mathbb{Q} = \mathbf{a}, \mathbb{O} = \mathbf{b}\right)\mathbb{P}\left(\boldsymbol{\beta}\right)\prod_i \mathbb{P}\left(\mathbf{O}_i|\mathbb{Q}_i = a_i, \mathbb{O}_i = b_i, \mathbf{O}_{a_i}, \mathbf{O}_{b_i}, \boldsymbol{\beta}\right)\, dV_{\boldsymbol{\beta}}} \tag{3.27}$$

Equation 3.27 assumes that the observation process for each offspring is conditionally independent from the observation process of each of the other offspring given the set of observation model parameter vectors $\boldsymbol{\beta}$:

$$\mathbb{P}\left(\text{♀}_i = m, \text{♂}_i = f | \mathbf{O}_i, \mathbf{O}_m, \mathbf{O}_f, R_i^{\text{♀}}, R_i^{\text{♂}}, \boldsymbol{\beta}\right) = \frac{\mathbb{P}\left(\mathbf{O}_i | \text{♀}_i = m, \text{♂}_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) \mathbb{P}\left(\text{♀}_i = m, \text{♂}_i = f | R_i^{\text{♀}}, R_i^{\text{♂}}\right)}{\sum_a \sum_b \mathbb{P}\left(\mathbf{O}_i | \text{♀}_i = a, \text{♂}_i = b, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) \mathbb{P}\left(\text{♀}_i = a, \text{♂}_i = b | R_i^{\text{♀}}, R_i^{\text{♂}}\right)} \tag{3.28}$$

where $R_i^{\text{♀}}$ and $R_i^{\text{♂}}$ are the sets of elements of the relevant parentage vectors except for element i, $\left\{\text{♀}_1, \text{♀}_2, \ldots, \text{♀}_{i-1}, \text{♀}_{i+1}, \ldots, \text{♀}_W\right\}$ and $\left\{\text{♂}_1, \text{♂}_2, \ldots, \text{♂}_{i-1}, \text{♀}_{i+1}, \ldots, \text{♂}_W\right\}$ respectively. The term $\mathbb{P}\left(\text{♀}_i = m, \text{♂}_i = f | R_i^{\text{♀}}, R_i^{\text{♂}}\right)$ denotes the probability that offspring i has mother m and father f given a set of parentage relationships for all other offspring. It is here that the investigator may, if they so wish, incorporate models of reproductive success to allow for limits on breeding potential. The joint inference of the entire parentage vectors, ♀ and ♂, can be achieved by updating each element of the vector in turn using Gibbs sampling, drawing each element from a categorical distribution with probability vector calculated using equation 3.28.

Algorithm 3: Metropolis-within-Gibbs sampling algorithm for the generation of samples from the joint posterior distribution of observation model parameters and the full parentage vector for W offspring. Allows for dependence of parentage between offspring and the dependency of the observation process on shared parameters between potential parents.

1. Initialise the chain with arbitrary values for the elements of vectors ♀, ♂, and $\boldsymbol{\beta}$ ensuring that $\prod_i \mathbb{P}\left(\mathbf{O}_i | \text{♀}_i = m_i, \text{♂}_i = f_i, \mathbf{O}_{m_i}, \mathbf{O}_{f_i}, \boldsymbol{\beta}\right) > 0$, $\mathbb{P}\left(\boldsymbol{\beta}\right) > 0$, and $\mathbb{P}\left(\text{♀}, \text{♂}\right) > 0$.

2. *Gibbs Step*: For each i^{th} offspring out of the total W:

   (a) Draw values for the parentage pair ♀_i^*, ♂_i^* from a categorical distribution with a vector of probabilities for each of the different parentage combinations calculated using equation 3.28.

   (b) Set $\text{♀}_i = \text{♀}_i^*$ and $\text{♂}_i = \text{♂}_i^*$.

3. *Metropolis-Hastings Step*: Propose a set of new values for the genotype observation model $\boldsymbol{\beta}^*$ from a multivariate truncated normal distribution with probability density function given in equation 3.24.

4. Calculate

$$u = \frac{\mathbb{P}\left(\mathbf{♀} = \mathbf{m}, \mathbf{♂} = \mathbf{f}\right)\mathbb{P}\left(\boldsymbol{\beta}^*\right)\mathbb{P}\left(\boldsymbol{\beta}|\boldsymbol{\beta}^*, \boldsymbol{\Sigma}\right)\prod_i \mathbb{P}\left(\mathbf{O}_i|♀_i = m_i, ♂_i = f_i, \mathbf{O}_{m_i}, \mathbf{O}_{f_i}, \boldsymbol{\beta}^*\right)}{\mathbb{P}\left(\mathbf{♀} = \mathbf{m}, \mathbf{♂} = \mathbf{f}\right)\mathbb{P}\left(\boldsymbol{\beta}\right)\mathbb{P}\left(\boldsymbol{\beta}^*|\boldsymbol{\beta}, \boldsymbol{\Sigma}\right)\prod_i \mathbb{P}\left(\mathbf{O}_i|♀_i = m_i, ♂_i = f_i, \mathbf{O}_{m_i}, \mathbf{O}_{f_i}, \boldsymbol{\beta}\right)} \quad (3.29)$$

5. Generate a random uniform value, α, between zero and one. If $\alpha < \min\{1, u\}$ then set the current vectors of parameter values, $\boldsymbol{\beta}$, to the proposed values $\boldsymbol{\beta}^*$.

6. Store the current values of $♀_i$, $♂_i$, and $\boldsymbol{\beta}$ as samples from the target distribution of equation 3.23.

7. Return to step 2.

3.2.4 An Example Dataset

To demonstrate the methods described in this chapter we apply the techniques using a data set for which the parentage is already known. We use an extensive data set of a Canary Island zebrafinch population with parentage assigned for over 120 offspring from a pool of 340 possible parents. This allows us to compare the results from the parentage analysis to the known parentage and evaluate the performance of the method. After loci with missing values present are removed from the analysis, 157 single-nucleotide polymorphisms (SNPs) remain.

To determine the performance of the models we perform two sets of analyses. The first analysis aims to test the effect of the observation model parameters, from two different phenotype observation models (R1-1 and C2-0 from table 3.1) on the quality of the paternity assignment. To achieve this aim, we run the analysis for a series of parameter values, treating them as fixed parameters in the estimation process. We generate samples from the posterior distribution using the Metropolis-within-Gibbs algorithm described in the previous section. In all cases and in each chain, we run the algorithm for 130000 iterations and discard the first 30000 iterations to allow the chain to 'burn-in'. Each analysis uses a total of five independent chains and each chain was initialised by selecting a parent pair at random as the putative parents for each offspring. Visual inspection of the trajectories of samples generated by each chain showed reasonable mixing between chains and the appearance of convergence in each of the analyses performed. Ranges for one metric of convergence, the multivariate scale reduction factor of Brooks & Gelman (1998), for all of these analyses were between 1.03 and 1.11. These

values suggest that a longer run will not substantially improve the precision of estimates of the parameters.

The quality of the paternity is assessed by using the proportion of the empirically derived posterior samples of parentage that assign the true parent pair as the parents, as an estimate of the posterior probability weight of the true parent pair. Assuming that the parentage of different offspring can be considered independent events, then the probability of correctly assigning the parents of all offspring is simply the product of the posterior probability estimates for the true parents for each of offspring. This index will henceforth be referred to as the 'total probability of correct parentage assignment'.

The second set of analyses involves the joint estimation of parentage and model observation parameters. We test the sensitivity of the estimation process on the prior distribution by considering two different prior density distributions. Our first prior is a vaguely informative prior with the probability of error for all error types given by a normal distribution with a mean of 0.005 and a standard deviation of 0.03 truncated between 0 and 1. Whilst ensuring that posterior estimates for the error values are restricted to an area of feasibility, this prior is still relatively broad and easily contains the values commonly used as fixed error rates in parentage analysis. For example, Marshall *et al.* (1998) use an error rate of 0.01 in their analysis of red deer in their demonstration of the CERVUS parentage analysis package. The second prior tested is an uninformative uniform distribution set over the limits of the error rate parameter (zero and one). This will allow investigation into the extent to which the quality of the parentage assignment is dependent upon knowledge of the observation error rates.

Similarly to the first set of analyses, we run the Metropolis-within-Gibbs algorithm described in the previous section for 130000 iterations and discarded the first 30000 iterations to allow the chain to 'burn-in'. Each analysis uses five independent chains. Initial parentage allocations for each offspring were chosen at random and initial values for the error parameters were drawn from a uniform distribution between zero and one to initialise each chain. New values for the error model parameters were proposed according to a simple random-walk Metropolis-Hastings algorithm except that the distribution was truncated between the limits of the parameter values. The standard deviation of the step-length was 0.1. Visual inspection of the trajectories of the samples suggest that the chains were mixing well and that requirements of convergence were met. In addition, calculation of the multivariate scale reduction factor (Brooks & Gel-

man, 1998), gives values of 1.07 and 1.14 for the fitting of the random relabelling observation model (R1-1 of table 3.1) using the vaguely-informative and non-informative priors respectively. The multivariate scale reduction factor for the analysis using the vaguely-informative and non-informative prior in the fitting of the marker-specific SNP model (C2-0 from table 3.1) was 1.02 and 1.06 respectively. These values suggest that the posterior estimates for the observation model parameters will not be substantially improved by further running of the MCMC algorithm.

3.3 Results

3.3.1 Fixed Error Model Parameters

All observation methods performed well across the entire range of biologically reasonable genotype error rates evaluated in this study, with the total probability of correct parentage assignment consistently higher than 83%. Both the random-relabelling observation model (R1-1 from table 3.1) and the SNP marker-specific observation model (C2-0 from table 3.1) tested show a negative skew of performance with genotyping error rate (see figure 3.3). This is to be expected, when genotyping error rates are set to artificially high levels then the parentage inference becomes blurred and probability weights are spread more evenly amongst the candidate parent pairs, thus lowering the posterior probability attached to the true parent pair. This situation improves gradually as the error rate is lowered, peaking when the fixed error rate matches closely the true genotyping error rate. However, a sharp drop in performance is present once the genotyping error rate is set at artificially low levels. This could be because the diagnosis of erroneous incompatibilities caused by observation errors exclude the true parents, causing a significant negative impact on the total probability of correct parentage assignment.

The SNP marker-specific observation model does however produce higher values for the total probability of correct parentage assignment, for some values of its parameter space. The C2-0 model benefits from an extra free parameter of complexity that the random relabelling model and at least some of this increased performance can attributable to the flexibly of likelihood surface under different parameter combinations. However, the total probability of correct parentage is particularly sensitive to changes in the allelic dropout frequency and relatively insensitive to changes in the allele misdiagnosis parameter (figure 3.3b). This suggests that the majority of errors present in data set are of the allele dropout variety. The only tunable genotyping error rate parameter in the R1-1 observation model controls the rate of genotype substitution, and

the genotype substitution probabilities are derived from the observed frequencies. However, given that null alleles are expressed only in the homozygous case, the substitution probabilities for genotypes containing null alleles in this model are only non-zero for homozygous case. This suggests that the R1-1 exhibits a certain degree of structural inflexibility to deal effectively with recessive alleles.

3.3.2 Variable Error Model Parameters

Figure 3.5 shows the prior probability density plot and the marginal posterior density estimates for the parameters for each of the observation models when the observation model parameters are estimated jointly with the parentage. The quality of the parentage assignment remains high for both observation models when informative priors are specified: the total probability of correct parentage assignment is 87.776% for the random relabelling model and 96.321% for the SNP marker model. The performance of the methods is not maintained when uninformative priors are specified however. The total probability of correct parentage assignment drops to 64.344% and 62.151% when there is no information to constrain the parameters of the observation models to reasonable values.

From figure 3.5 it is apparent that the posterior estimates for the parameters of the observation models are very similar under the specification of an informative prior and all appear very similar in shape to the prior distribution. Additionally, the credible intervals for the random relabelling parameter of model R1-1 and the allele misdiagnosis rate of model C2-0 are very wide when an uninformative prior is specified. This suggests that there is only limited information available in the data to estimate these parameters. Despite this, there is little posterior support for extreme values for the random relabelling parameter and the allele misdiagnosis parameter when the prior is uninformative. This makes sense as low values for the relabelling rates in R1-1 and the allele misdiagnosis rates in C2-0 result in models where the likelihood of any incompatible parent and offspring phenotype combinations is low. In a large data set it is likely that there will be at least one instance where observation error will result in offspring with phenotypes that appear incompatible with any available parent pair and so, in these cases, it is possible to discriminate against these unlikely parameter values even when there is no prior information. Moreover, the allelic dropout rate parameter of model C2-0, whilst wider than under the analysis using the vaguely-informative prior, still has a relatively narrow posterior distribution when a non-informative prior is specified. Allelic dropout provides a very peculiar type of error compared to the other error types as it not only changes the observed allelic

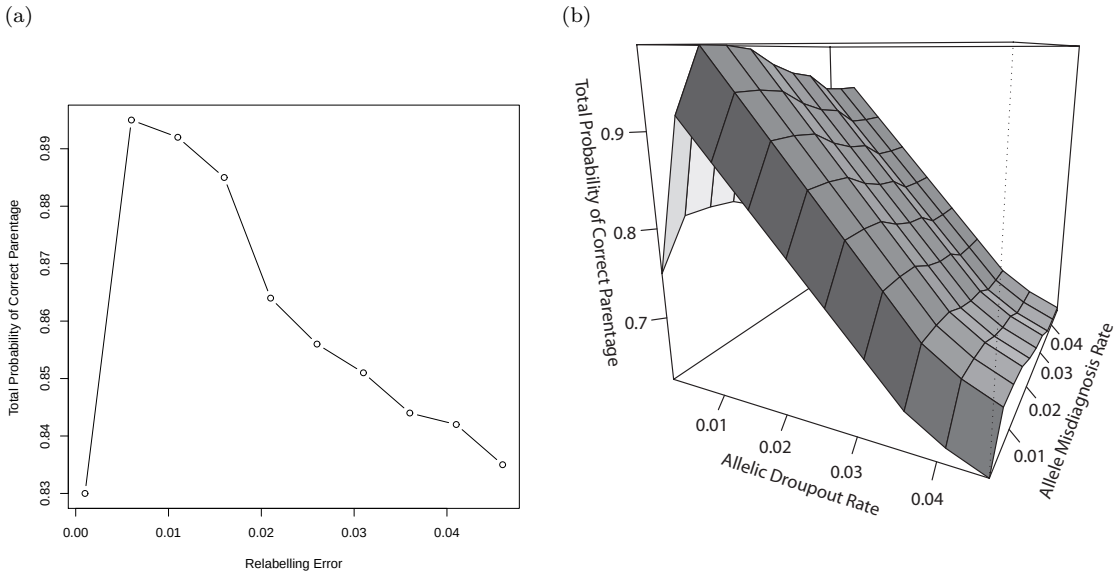


Figure 3.3: The effect of the parameters of the observation models on the total probability of correct parentage assignment. Figure (a) shows the effect of the value of the rate of random relabelling error parameter on the parentage assignment quality for the R1-1 observation model (see table 3.1). Figure (b) illustrates how the parentage assignment quality varies with each of the two parameters for the observation model C2-0 (see table 3.1): the rate of allelic dropout and the rate of allele misdiagnosis.

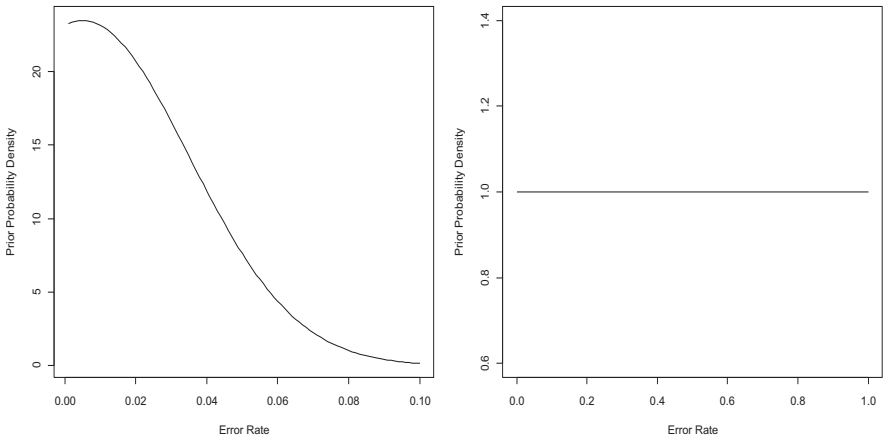
frequencies but also the zygosity of the population. This type of signal may be pretty easy to detect in the data set and so, the frequency of these types of error may be estimated reliably, even when an uninformative prior is used.

3.4 Discussion

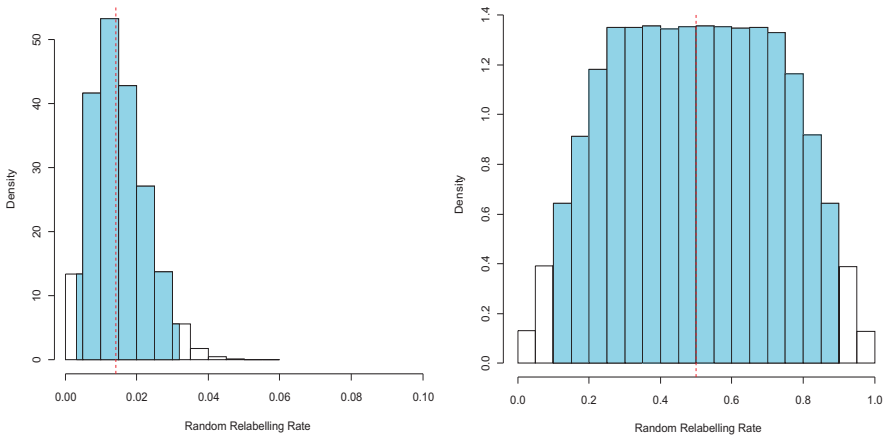
Our results have shown that the application of observation models that describe the error types present in different marker types outperform generic relabelling models on the whole. The

Figure 3.5 *(on the next page)*: Model outputs and priors from an analysis where both observation model parameters and parentage are jointly estimated. Figure (a) displays the prior probability density of observation parameters. The left-hand panel gives a vaguely informative prior with higher densities given to error rates that are likely to exist in standard settings. The right-hand panel is an entirely non-informative prior, where the error rates are equally likely along the entire possible range of values. Figure (b) displays the estimated posterior density for the parameter of the random relabelling model (R1-1) under analyses using the two different prior types. Figure (c) displays the marginal posterior density estimates for the two parameters under the SNP observation model (C2-0) with the leftmost panels displaying results for analyses performed using with the vaguely informative prior and the rightmost panels displaying results for analyses performed using the non-informative prior. The blue shading on the posterior density estimates denotes the 95% credible interval for the parameter estimates and the red dotted line shows the median value of the posterior sample.

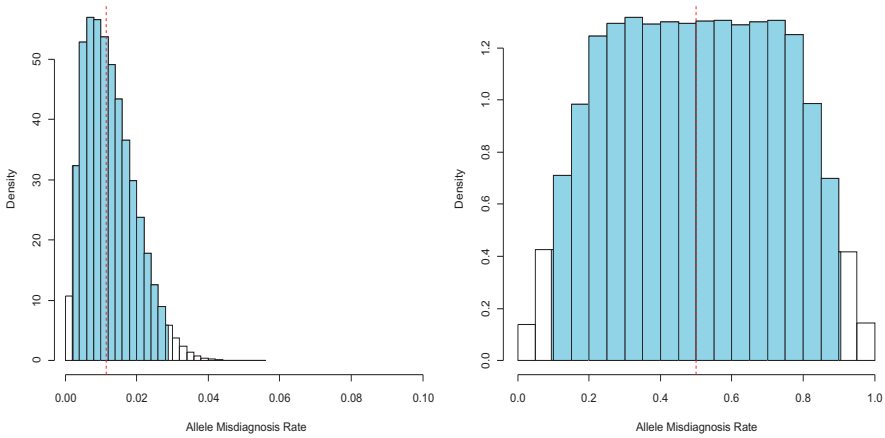
(a) Prior Specification of Error Terms

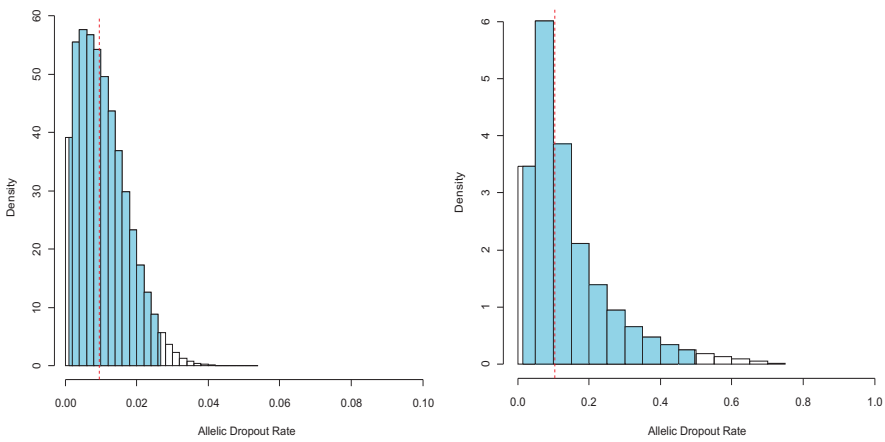


(b) Random Relabelling Model



(c) SNP Observation Model





quality of the parentage assignment is however dependent upon assumptions about the error rate. This is true regardless of the observation model used. The wide credible interval for most observation model parameter estimates when a uninformative prior is specified may explain the poor performance of the parentage assignment in these situations. Parentage methods rely on identifying unlikely or incompatible genotypes to exclude potential parent pairs and, given that our inference of these genotypes are driven by how they are liked to the observations of the phenotype, then the parameters attached to a given observation model will dramatically alter how phenotypes are used to weight combinations of different parents. If we have poor knowledge of the error rate then it becomes difficult to identify the correct parent pair. In short, there is an issue of parameter identifiability in the model and without prior knowledge of one set of parameters then it is difficult use information about the observation error rates to build a picture of parentage. All is not lost however, as we have shown that even vaguely-informative priors can provide the information required for improved parentage assignment.

Different marker systems are observed in different ways: AFLP and RAPD markers allow the simultaneous observation of multiple loci on one gel line so that the position on the gel denotes the locus whilst VNTR and RFLP markers use the positioning on the gel to assign allele types for one locus at a time. Given the very different nature of these observation processes, it seems intuitive that the application of one model for all marker types to describe the transition from genotype to observed phenotype would be insufficient.

One of the criterion for assessing the performance of a model is to ask whether the model, as it is described, adequately mimics the real-world phenomenon for which it is designed to elucidate. The model implemented in Parente (Cercueil *et al.*, 2002) defines a parameter that controls the rate of errors of substitution. When a substitution is made the single locus genotype is replaced with another single locus genotype taken directly from the population of genotypes in the sample. It has been argued that such a model, may adequately describe errors arising from labelling, pipetting or data entry errors (Kalinowski *et al.*, 2007). Relabelling errors are however likely to result in genotype errors across all loci rather than just at a specific locus. Moreover, this list of human errors does not include that of allele binning, an error type which would result in a very different genotypic signal than that occurring from random relabelling, and has the potential to account for much of the observation error (Ewen *et al.*, 2000). Other allele misdiagnosis errors, unrelated to the quantity of the allele in the sampled population, such as allelic dropout from contamination by inhibitory contaminants or preferential ampli-

fication of other allele types are also not included. The error model implemented in CERVUS (Marshall *et al.*, 1998), goes further than that of Cercueil *et al.* (2002) and draws replacements from a theoretical population in Hardy-Weinberg equilibrium with allele frequencies equal to those observed in the sampled population. It is even harder to see how this type of error could occur in the genotyping process as neither labelling, pipetting, or data entry errors involve the investigator finding a replacement for the true sample by going back out into the field and sampling from a population that may not necessarily exhibit the same zygosity as the sampled population. This is particularly true given that the presence of null alleles does not mean that observed allele frequencies of the sampled population represents the true allele frequencies.

Dominant markers, true to their namesake, exhibit a relatively high frequency of recessive markers as it is only the presence of null allele homozygotes in the sample that create the polymorphism from which inference can be drawn. The per locus information content for most dominant markers is much lower than their codominant counterparts, prompting some authors to eschew the use of dominant markers in parentage analysis (Kirst *et al.*, 2005). However, the costs of running dominant markers can be considerably cheaper, and, it may be possible to make up for the lack of power at one locus by inferring parentage relationships from many loci simultaneously (Gerber *et al.*, 2000; Milligan & McMurry, 1993). Currently, very few parentage analysis programs allow for the use of dominant markers in parentage analysis: PROBMAX (Danzmann, 1997), MASTERBAYES (Hadfield *et al.*, 2006), FAMOZ (Gerber *et al.*, 2003) and, more recently, COLONY (Jones & Wang, 2010; Wang, 2004), are notable exceptions to this. However, unlike the observation model for dominant markers described in this study, most of these programs are limited to the analysis of parentage in the diploid case.

Null alleles are much rarer in loci used in codominant markers, not necessarily because they are rarer in the total population of polymorphic loci, but that codominant markers used in parentage analysis are often chosen from loci that exhibit low null allele frequencies (Matson *et al.*, 2008; Castro *et al.*, 2007). This is because null alleles are often considered nuisance alleles in studies using codominant markers (de Sousa *et al.*, 2005; Pemberton *et al.*, 1995); most packages of parentage analysis do not attempt to distinguish between heterozygotes with one or more null alleles and non-null homozygotes, assuming the later, and, therefore, that no recessive alleles exist in the data set. Whilst CERVUS (Marshall *et al.*, 1998; Kalinowski *et al.*, 2007) and NEWPAT (Worthington Wilmer *et al.*, 1999) include a diagnostic tool for the identification of loci with high null allele frequencies, they offer no mechanism to incorporate these loci into the

analysis. PROBMAX (Danzmann, 1997), takes the more conservative approach, and re-codes all non-null homozygotes as heterozygotes with the presence of one null allele for all loci with null alleles present. This method may work satisfactorily for exclusion methods, but, if applied to fractional assignment methods, may bias parentage assignment.

To avoid errors compounded by issues of observability, (de Sousa *et al.*, 2005) advocates removing all loci with suspected null alleles from all analyses. However, excluding all loci from an analysis with suspected null alleles is very wasteful, drains the discrimination power between candidate parent pairs, and may bias the paternity assignment if these loci are informative. Indeed, some studies have shown that the benefits of some of the maximum-likelihood based parentage assignment methods are undermined and perform no better than exclusion methods once the set of sampled loci are trimmed to remove those which are error prone or contain null alleles (Castro *et al.*, 2007). For many species where the number of described microsatellite loci are few, the loss of power from such exclusion practices would be unacceptable.

Dakin & Avise (2004), in their review of 233 articles using data that included null alleles, report that only a small fraction of the 90% of articles the included loci with null alleles in the analysis made any statistical correction for this fact. In contrast, the observation model described in this paper explicitly addresses the different mechanisms by which a phenotype can be observed, including the excess homozygosity arising from null allele presence. Given that some studies show as high as 40% of incompatibilities arising from the presence of null alleles (Bowling *et al.*, 1997), it is becoming ever more important to address null alleles explicitly in models of parentage. Bar some notable exceptions, the two main strategies for dealing with null alleles in parentage analysis is either to remove loci from the analyses where they are suspected to be present or to ignore their presence and to continue the analysis as normal without correction. Both methods have substantial drawbacks. We describe here a method of parentage assignment that explicitly models the inheritance of the null case, allowing the integration of information present in all loci in the inference of paternity without needing to exclude loci with null alleles present.

The special exceptions for null allele transmission used in this paper may go some way to addressing the biases described in other papers. However, the implicit assumption has been made that the 'null' allele is of only one type and our mechanisms of inheritance treat it as such. (Lehmann *et al.*, 1996) show that a 'null' allele may have a basis in many different genetic

characteristics that prevent amplification. If some estimates of the number of null allele types are known, then it may be possible to incorporate these as extra alleles in the observation and genotype vectors. In most cases this information is not available, and even if it were available, it is unlikely that this extra information would play an important role in discriminating between potential parent pairs for a given offspring. Whilst it may be genetically incorrect to aggregate the different null alleles types, it may prove to be the parsimonious stance in the absence of further information and avoids the scenario where incompatibilities between different null allele types are diagnosed but where one of the null allele types does not exist.

The Mendelian model of inheritance described in this paper performs adequately for species with even-valued ploidy and equal genetic investment from both parents. However, more unusual inheritance systems can be incorporated into the analysis by modifying transition equation 3.3. For example, it is possible to describe the more general situation where genetic investment from parents are not equal, respecifying probability mass functions $g_m(\mathbf{X}_{mj})$ and $g_f(\mathbf{X}_{fj})$

$$g_m(\mathbf{X}_{mj}) = \begin{cases} \dfrac{\prod_a \binom{\mathrm{G}_{mj_a}}{\mathrm{X}_{mj_a}}}{\binom{C_{ij}}{M_{ij}}} & \text{if } \sum_a \mathrm{X}_{mj_a} = M_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{3.30}$$

$$g_f(\mathbf{X}_{fj}) = \begin{cases} \dfrac{\prod_a \binom{\mathrm{G}_{fj_a}}{\mathrm{X}_{fj_a}}}{\binom{C_{ij}}{C_{ij}-M_{ij}}} & \text{if } \sum_a \mathrm{X}_{fj_a} = C_{ij} - M_{ij} \\ 0 & \text{otherwise} \end{cases} \tag{3.31}$$

where M_{ij} is the number of copies of locus j maternally inherited by offspring i. C_{ij} is the ploidy of offspring i at locus j. Using this formulation it is possible to describe the inheritance of maternally inherited markers, such as those residing in chloroplasts or mitochondria, by setting $M_{ij} = C_{ij}$. It is also possible to describe sex specific ploidy, for example, haplodiploidy, by setting $C_{ij} = 2$ and $M_{ij} = 1$ if the offspring is female, and setting $C_{ij} = 1$ and $M_{ij} = 1$ if the offspring is male. NEWPAT (Worthington Wilmer *et al.*, 1999) allows the analysis of sex-linked loci, FAMOZ (Gerber *et al.*, 2000) allows the analysis of cytoplasmically inherited loci, and COLONY (Jones & Wang, 2010) allows the analysis of haplodiploid organisms. However, the methodology described in this paper can be extended to cover all of these marker inheritance systems with ease, allowing also the mixing of these inheritance systems so that joint parentage estimation can be made from different markers with different mechanisms of inheritance.

Parentage analysis is often part of a much larger analysis, often used to infer other characteristics such as dispersal abilities (Robledo-Arnuncio & Garca, 2007; Piotti *et al.*, 2009; Zeyl *et al.*, 2009), home range size (Martin *et al.*, 2007), and individual reproductive success (Williams & DeWoody, 2009). These approaches require methods to combine the information from both genetic and non-genetic data to jointly estimate parameters controlling parentage and the phenomenon of interest. Classic approaches have seen the application of a two stage process where parentage is first inferred using one of the many parentage analysis programs available. Once parentage is ascertained the results of the analysis are fed into the second stage of the model for which another set of parameters are estimated. However, it is rare in these analyses that the uncertainty related to the parentage assignment is propagated to the next stage of the model. If parents are incorrectly identified and no information is provided to the strength of the identification then significant biases may arise in estimates of parameters at the next stage of the analysis (Jones, 2003). Fractional parentage methods, such as the one described in this paper, are expected to perform best when propagating uncertainty as they can provide probabilistic estimates for all parent pairs, retaining information when multiple possible crosses are likely.

MasterBayes (Hadfield *et al.*, 2006) goes one step further in this regard and actually incorporates non-genetic information, such as spatial data, into the parentage assignment. Bayesian methodologies can be particularly useful here because they allow easy combination of different models into one hierarchical framework. Unlike methods which simply use the output of one model to become the input of another model, Bayesian hierarchical models use the information collected at any tier of the model to jointly estimate the parameters at all tiers of the model. This 'pulling in' of all available data allows for maximum inference of parameter values and better estimates (Jackson *et al.*, 2009). Whilst exact implementations are beyond the scope of this paper, it is worth noting that the model described here is equally as extensible to the integration of data from other sources, allowing independent data on dispersal ability, breeding success, or home range to influence paternity assignment.

The combination of null allele presence and allelic dropout have accounted for as much as a 53% of false parentage assignments in some studies (Jerry *et al.*, 2004). The mechanisms by which genotyping error and null allele presence exclude parentage pairs are different however, and to adequately model the probabilities associated with observing an offspring phenotype

given the observed phenotypes of putative parents it is necessary to disentangle these effects. The model framework described here does just that, explicitly modelling the inheritance of 'true' genotypes and the observation process that links the phenotype to the 'true' genotypes. We have also shown that this method is easily extensible to include other non-genetic information and allows the drawing of statistical power from multiple marker types and for any ploidy. Only when all available information from all sources are drawn into the analysis can we achieve the best possible estimates for the parameters that govern parentage. The method described here presents one step towards this eventual goal.

CHAPTER 3S

Using Parentage Assignment for the Calculation of Population Parameters

In ecological studies, the parentage assignment resulting from the parentage analysis is, in itself not usually the primary focus of the study. More often, parentage analysis is a precursor to the assessment of other features of interest in the population. We include this section as an addendum to chapter 3 to show how we can use the output from the parentage analysis described there to drive inference about two key parameters of interest: the distribution of offspring for each individual, and the dispersal kernel of offspring from the parent individuals. Inference about these processes can form the basis of an individual-based model as described in chapter 4.

3S.1 Calculating Breeding Success

The breeding success of an individual can be characterised by the probability distribution of offspring that it manages to produce in a given time period. For many species we would expect a difference functional form for breeding success between the sexes, or seed donor or pollen donor when are talking about plants. We denote $K_{\text{♀}}\left(x|\omega_{\text{♀}}\right)$ and $K_{\text{♂}}\left(x|\omega_{\text{♂}}\right)$ as the probabilitiy mass function of breeding success for the maternal (or seed donating plant) and paternal (or pollen donating plant) contributions. Here, $\omega_{\text{♀}}$ and $\omega_{\text{♂}}$ are vectors of parameters controlling the distribution of the breeding success function. In any given iteration of algorithm 3 in chapter

3, we are presented with a set of putative parentage pairs, consisting of two vectors $\mathbf{m}$ and $\mathbf{f}$, denoting the proposed mother (seed donor) and father (pollen donor) respectively for each of the offspring in the population, such that each element of the vector holds the indicator number of the maternally and paternally contributing parent. Under this specification of the model of breeding success, the likelihood of the vectors of putative parentage pairs given the parameters of the breeding success distributions is

$$\mathcal{L}_K\left(\omega_{\female}, \omega_{\male}\right) = \prod_j K_{\female}\left(\zeta_j|\omega_{\female}\right) K_{\male}\left(\xi_j|\omega_{\male}\right) \tag{3S.1}$$

where $\zeta_j = \sum_i \mathbf{1}_j\left(m_i\right)$ and $\xi_j = \sum_i \mathbf{1}_j\left(f_i\right)$ are the number of offspring that potential parent j contributed to maternally and paternally respectively. $\mathbf{1}_j\left(x\right)$ is an indicator function and equals one when individual j is the same individual as individual x and zero at all other times.

3S.2 Calculating Dispersal Distances

If spatial information pertaining to the location of the offspring and parents is available then if is also possible to use the parentage assignment to parameterise estimates of inter-generational dispersal. Animals adhere to a many number of specialised dispersal rules and behavioural tendencies that would be too varied to discuss here. We instead restrict ourselves to the pollination and seed setting dynamics of plants with the hope that the reader may be able to intuit from this simple example to more complex dispersal models. The dispersal capabilities of an individual can be represented by its dispersal kernel, a probability density distribution defined over possible dispersal distances, r, and angles of dispersal θ. We might be inclined to define separate probability density distributions for the different dispersal mechanisms, $D_{\female}\left(r, \theta|\delta_{\female}\right)$ and $D_{\male}\left(r, \theta|\delta_{\male}\right)$, for seed dispersal and pollen dispersal respectively. We define r_{ij} as the distance between individuals i and j, and θ_{ij} as the angle of direction from i to j. For the birth of a new sexually produced individual we can assume that two dispersal events have to have happened: firstly a pollen grain must have dispersed from the pollen donor to the seed donor, and secondly, a seed must have dispersed from the seed donor to the offspring location. The likelihood of the vectors of the vectors of putative parentage pairs given the parameters for the dispersal kernels is therefore

$$\mathcal{L}_D\left(\delta_{\female}, \delta_{\male}\right) = \prod_i D_{\female}\left(r_{m_i i}, \theta_{m_i i}|\delta_{\female}\right) D_{\male}\left(r_{f_i m_i}, \theta_{f_i m_i}|\delta_{\male}\right) \tag{3S.2}$$

3S.3 Adding Population Parameter Inference to the Parentage Analysis

Parameter estimation for the breeding and dispersal parameters can be embedded into the parentage analysis be inserting an extra Metropolis-Hastings sampler between steps 6 and 7 of algorithm 3 of chapter 3. This Metropolis-Hastings sampler takes the standard form:

1. Propose new parameter values for the dispersal and breeding success models from the proposal density $q\left(\omega_{\female}^{*}, \omega_{\male}^{*}, \delta_{\female}^{*}, \delta_{\male}^{*} | \omega_{\female}, \omega_{\male}, \delta_{\female}, \delta_{\male}\right)$

2. Calculate

$$\alpha = \min\left\{1, \frac{\mathcal{L}_K\left(\omega_{\female}^{*}, \omega_{\male}^{*}\right)\mathcal{L}_D\left(\delta_{\female}^{*}, \delta_{\male}^{*}\right)\pi\left(\omega_{\female}^{*}, \omega_{\male}^{*}, \delta_{\female}^{*}, \delta_{\male}^{*}\right)q\left(\omega_{\female}, \omega_{\male}, \delta_{\female}, \delta_{\male} | \omega_{\female}^{*}, \omega_{\male}^{*}, \delta_{\female}^{*}, \delta_{\male}^{*}\right)}{\mathcal{L}_K\left(\omega_{\female}, \omega_{\male}\right)\mathcal{L}_D\left(\delta_{\female}, \delta_{\male}\right)\pi\left(\omega_{\female}, \omega_{\male}, \delta_{\female}, \delta_{\male}\right)q\left(\omega_{\female}^{*}, \omega_{\male}^{*}, \delta_{\female}^{*}, \delta_{\male}^{*} | \omega_{\female}, \omega_{\male}, \delta_{\female}, \delta_{\male}\right)}\right\} \tag{3S.3}$$

3. Draw a random number l from a uniform distribution defined between zero and one. If $l < \alpha$ then accept $\omega_{\female}^{*}$, $\omega_{\male}^{*}$, $\delta_{\female}^{*}$, and $\delta_{\male}^{*}$ as samples from the posterior distribution and set

$$\begin{aligned} \omega_{\female} &= \omega_{\female}^{*} \quad \omega_{\male} = \omega_{\male}^{*} \\ \delta_{\female} &= \delta_{\female}^{*} \quad \delta_{\male} = \delta_{\male}^{*} \end{aligned} \tag{3S.4}$$

   otherwise accept $\omega_{\female}$, $\omega_{\male}$, $\delta_{\female}$, and $\delta_{\male}$ as samples from the posterior distribution.

CHAPTER 4

The application of Approximate Bayesian Computation to Individual-Based Modelling

Summary

1. The application of individual-based models (IBMs) to ecological problems has long been hampered by the inability to derive the likelihood functions used to assess the performance of parameter combinations in describing observed data.

2. Heuristic pattern-matching methods have been suggested by previous authors which require a thorough interrogation of the parameter space. Whilst these methods allow for a reasonable analysis when there are few parameters in the model, they rapidly loose tractability when the dimensionality of the parameter vector increases. The lack of a calculable likelihood means that classical likelihood-driven Bayesian approaches are also impossible to employ.

3. Here we present a number of approximate Bayesian methods that can be used in IBM applications to efficiently and robustly search the available parameter space even when the functional form of the likelihood is not available. We develop a selection of novel algorithms that are more specifically tailored towards the fitting of IBMs to the sort of data that they are likely to be fitted to (such as time series data).

4. We explain how these methods can be extended using an approximate Bayesian equivalent

of reversible-jump Markov chain Monte Carlo (MCMC) to select the best model from a selection of candidate models.

5. A worked example is provided to show the application of these methods to select between three semi-mechanistic models of molehill formation and to assess the posterior support of different parameter combinations based on their ability to recreate the spatial properties of observed molehills.

6. We discuss the how previous examples 'pattern-oriented' modelling fit in within the approximate Bayesian framework and describe the benefits of adopting these more formal methods for the analysis of data.

4.1 Introduction

Modelling is the art of refining complex phenomena into tractable caricatures. Ecological systems are complex, but boiling down the characteristics of that we are interested in into elegant abstractions whilst maintaining realism can be difficult. With increased understanding of this complexity, alongside an invigoured motivation to include it in the formulation of the model, we have witnessed the birth of a many number of methodologies to incorporate complex interactions.

Individual-based modelling is one such methodology. These techniques are often employed by modellers when results arising from individual-level variation cannot be adequately explained by their state-variable, and mostly analytical, counterparts. Here individuals are defined by a set of rules and characteristics that dictate how they are to interact with their abiotic and biotic environment. Population-level phenomena emerge in individual-based models (IBMs) as the cumulative effect of the individuals interaction with the environment and each other rather than hard-coded at a higher level such as those models employed in classic population ecology.

Individual-based models are not without their costs however. The very complexity of these models mean that they are much more difficult to test and analyse (Murdoch *et al.*, 1992; Beissinger & Westphal, 1998). The problems with this complexity also extend to the fitting of these models to data. Except in the most basic of cases, individual-based models are analytically intractable. Some authors have highlighted the ability to estimate certain key features of interest from individual-based models using analytical techniques: Murrell *et al.* (2004) describe how to approximate the dynamics of spatial moments and Ovaskainen & Cornell (2006) illustrate the derivation of asymptotically exact spatial information using perturbation theory. Despite these efforts, for the most part we are forced to rely on Monte Carlo techniques for the analysis and fitting of IBMs.

For our classical model counterparts the process of fitting models to data is a relatively much less painful affair. Once a mathematical description of the underlying process has been described, an error term, if not already implicit in the model, can be assumed about the process. This allows us to describe the probability of observing the data if the structure of the model and a given combination of parameter values were true. This likelihood function can then be maximised either through analytical derivation or the employment of numerical techniques,

such as those described in Nash (1990), to find the combinations of model parameter values that, if true, would have the highest probability of producing the observed data. These optimal parameter values are often referred to as 'maximum-likelihood estimators' and, for most of the regularly applied probability distributions, they can be described in a simple closed-form function of the data. Even in the absence of a full likelihood specification, the mathematical description of the underlying process allows us to maximise or minimise some other relevant metric such as the sum of the squared deviation of the data from the model functional form.

Unfortunately, deriving a likelihood function for an IBM is not a simple affair and, except in the most simple of cases, it is mathematically intractable. An example of this intractability can be illustrated by considering a description of the movement of an individual that disperses to a new location in each time period by selecting a direction (θ) at random according to some known angular probability distribution function, with probability density function $f_1(\theta|\alpha)$, and then independently drawing a dispersal distance (r) from another known probability distribution with probability density function $f_2(r|\alpha)$. Here α is a vector of parameters controlling the shape of the probability density functions. This is a simple setup and is one that has been used frequently as the basis for more complex IBMs (see Dytham & Travis, 2006; Dytham, 2009, for examples). Now, consider that we are presented with some tracking data of an individual's movements in the field and we wish to fit our simple random-walk model to our observed movement data. Under this scenario our likelihood function, $\mathcal{L}(\alpha)$, is the product of the probabilities of dispersing the observed direction and distance in each time period such that

$$\mathcal{L}(\alpha) = \prod_t f_1(\theta_t|\alpha) f_2(r_t|\alpha) \tag{4.1}$$

where r_t and θ_t are the distance and angle of dispersal observed at time t respectively.

So far, as long as the probability density functions f_1 and f_2 are calculable, then our likelihood function is specifiable in a simple and calculable form. However, modelling an individual's movement as a similar process to the Brownian motion of a particle will rarely provide an adequate description of true movement (Turchin, 1998; Codling *et al.*, 2008). One added complication to the model that an ecologist may be keen to include in the model is the interaction of the individual with features in the landscape. For example, we could add an extra movement rule into the model that states that if the path of movement crosses a physical barrier (such as a wall) then it will reflect away from the wall at an angle from the wall equal

to the angle of incidence before continuing its movement. Suddenly, with the addition of this conceptually simple rule, the likelihood of dispersing to a given location becomes much more difficult to calculate: the probability that an individual arrives at a certain location is now the sum of the probability that the individual arrives there directly, without crossing a barrier, and the probability that the individual arrives there indirectly by reflecting off one or more reflective barriers. With the addition of two or more reflecting barriers, there can potentially be an infinite number of ways the individual can arrive at a given location. To calculate the likelihood in these situation would require the evaluation of the convergence properties of an infinite series. We see that in this situation we are left with a model that is relatively easy to simulate from, but for which it is difficult to evaluate the likelihood. Situations like this are not uncommon when dealing with IBMs however.

The inability to specify a likelihood function for most IBMs presents a problem when fitting these models to data. If we are unable to specify how likely the data are given the model structure and parameter values then how do we go about searching the parameter space for values that produce a good fit to the data? In most cases the potential range of values that collected data can take is huge, and for continuous data and data with unbounded ranges, it is infinite. Expecting an IBM to recreate the data for any combination of parameter values is obviously untenable. Grimm *et al.* (1996) argue that although exact data recreation is impossible, and not a useful goal, we can attempt to emulate certain patterns of interest in the data. By defining metrics that measure the difference between the patterns of interest in the data and the emergent properties of the individuals we can systematically search the parameter space for value which, over many simulations, minimise the distance metric (Wiegand *et al.*, 2003). This so-called 'Pattern-Oriented Modelling' or 'POM' (Grimm *et al.*, 1996, 2005) is not dissimilar to what a statistical practitioner might call 'fitting', although typically, when fitting an IBM, a smaller subset of potential parameter values are tested due to the high computational cost associated with Monte Carlo simulation for each parameter combination.

Pattern-oriented modelling as it currently stands faces a number of problems however. The dimensionality of the parameter vector in individual-based modelling is commonly high (DeAngelis & Mooij, 2003). Testing just five values of eight different parameters would result in $5^8 = 390625$ sets of simulations, each of which would require a number of realisations to in order to adequately assess the ability of the parameter combination to emulate the pattern of interest. Even if such a rigorous trial was performed there would be no certainty that one

of the parameter combinations chosen would be the best out of the possible set of parameter combinations, or indeed good, at pattern reconstruction.

In order to elevate pattern-oriented modelling from a heuristic to a methodology it is important to incorporate techniques that can address these problems in a statistically robust manner. This paper aims to flesh out a number of these techniques in this emerging field, describing each in turn, giving an example of their use in fitting a relatively simple mechanistic model of spatio-temporal point pattern dynamics to molehill construction.

4.2 Materials and Methods

4.2.1 Approximate Bayesian Computation

The process of model fitting requires that we search the available parameter space for combinations that are the most probable given the available dataset. We, as investigators, are therefore interested in the quantity $\mathbb{P}(\theta|D)$, the probability of the parameter vector θ given the dataset D. Bayes theorem states that this quantity is given by

$$\mathbb{P}(\theta|D) = \frac{\mathbb{P}(D|\theta)\,\pi(\theta)}{\int \mathbb{P}(D|\theta)\,\pi(\theta)\,d\theta} \tag{4.2}$$

where $\mathbb{P}(D|\theta)$ is the probability of obtaining dataset D with a model parametrised with parameter vector θ, otherwise known as the likelihood. In Bayesian parlance the quantity $\mathbb{P}(\theta|D)$ is often referred to as the 'posterior distribution'. $\pi(\theta)$ is the 'prior probability' of the parameter vector θ and this term represents the probability density of the parameter vector before information has been drawn from the dataset. It is through this quantity that prior knowledge of the system of interest, either through previous study or known biological or physical limits, can be integrated into the fitting process. Except in the most simple of cases, the denominator quantity, a normalising constant, is difficult to calculate analytically. Instead, Bayesian analysis commonly relies on algorithms to sample values of θ from the target distribution. Inference is made in these cases from the empirical distribution of sampled parameter values.

Fitting IBMs using Bayesian techniques comes with added disadvantage that, for the most part, the likelihood function is not known. This quantity is required, not only in the direct calculation of the posterior, $\mathbb{P}(\theta|D)$, but also in most sampling algorithms of it. The demand for fitting complex models to data with analytically intractable likelihoods has resulted in the

development of a number of approximate likelihood-free approaches for Bayesian model fitting.

All methods of approximate Bayesian computation, rely on the calculation of a summary statistic, or set of summary statistics, with which to compare simulated model outputs with the real dataset. We denote the i^{th} summary statistic of the true data, D, as $S_i(D)$ and the same summary statistic calculated for data simulated from the model with parameter vector θ, $\hat{D}$, as $S_i\left(\hat{D}\right)$. The distance between the set of n summary statistics derived from the simulated dataset and the real dataset is calculated using a distance function $\rho\left[\mathbf{S}(D), \mathbf{S}\left(\hat{D}\right)\right]$, where $\mathbf{S}(\cdot) = [S_1(\cdot), \ldots, S_n(\cdot)]$. All algorithms rely on a tolerance parameter, δ, which is used as an acceptance threshold to decide how close the summary statistic calculated from the simulated data has to be to the same statistic calculated from the real data before it is retained as a sample from the posterior distribution. Small values of δ specify a very narrow acceptance criterion, a higher rejection rate, and hence, increased computational time in order to obtain a robust sample from the target distribution. Large values of δ may increase acceptance rate but at the cost of making the approximation to the posterior distribution much more coarse.

Rejection Sampling

The first set of approximate methods considered here are those that simulate parameters from the prior followed by a rejection criterion. Earlier incarnations of rejection algorithms for approximate Bayesian computation do exist (see Tavaré *et al.*, 1997) but they are either designed for a specific application and not easily generalisable to other models, or require an analytic description of the expected value of the comparison statistic which, for most scenarios, is not available. We begin here with a description of the method used in Pritchard *et al.* (1999) for the fitting of human population history models to Y chromosome microsatellite data:

Algorithm 1: Rejection sampling (Pritchard *et al.*, 1999)

1. Draw a random parameter vector θ from a distribution with probability density function $\pi(\theta)$, the prior density.

2. Simulate a dataset, $\hat{D}$, using parameter vector θ.

3. If $\rho\left[\mathbf{S}(D), \mathbf{S}\left(\hat{D}\right)\right] \leq \delta$ (in their original paper Pritchard *et al.*, 1999, required that $|S_i(D) - S_i\left(\hat{D}\right)| \leq \delta$ for all i) then store the parameter vector θ as a sample from the

target distribution.

4. Go to 1.

Beaumont *et al.* (2002) adapt this algorithm further, proposing an ellipsoidal acceptance region at step 3, $\rho\left[\mathbf{S}\left(D\right),\mathbf{S}\left(\hat{D}\right)\right] = \sqrt{\sum_i \left[\frac{S_i(D)-S_i\left(\hat{D}\right)}{\sigma_{S_i}}\right]^2}$ where σ_{S_i} is a scaling constant related to the variance of statistic i. The authors also provide a re-weighting and regression step to attempt to correct for the approximation.

Markov Chain Monte Carlo

Rejection methods have the advantage that they are conceptually simple, easy to code and that the generation of samples from the target distribution can be done in isolation. This latter characteristic permits the parallelisation of sample generation across multiple processors. However, in cases where the prior density is substantially different from that of the posterior density, the rejection rate can become prohibitively high and a thorough sample of the posterior distribution computationally infeasible.

Standard Bayesian analysis makes common use of Markov Chain Monte Carlo methods, such as the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970; Chib & Greenberg, 1995) or the Gibbs sampler (Geman & Geman, 1984), that involve the sampling of posterior parameter estimates from a Markov Chain. The samples drawn from the chain tend towards a sample from the posterior as the chain progresses. Samples may not be drawn independently under this scheme but, at the asymptote, this lack of independence does not affect the eventual converge of the parameter samples to that expected if they were drawn directly and independently from the posterior. Marjoram *et al.* (2003) suggest an approximate analog of the Metropolis-Hasting algorithm commonly employed in traditional Bayesian computation:

Algorithm 2: Markov chain Monte Carlo sampling (Marjoram *et al.*, 2003)

1. Initialise the parameter vector θ with arbitrary values.

2. Propose a new vector of parameter values, θ^*, drawn randomly from a distribution with probability density $q\left(\theta^*|\theta\right)$.

3. Simulate a dataset, $\hat{D}$, using the proposal parameter vector θ^*.

4. If $\rho\left[\mathbf{S}\left(D\right),\mathbf{S}\left(\hat{D}\right)\right] > \delta$ then store θ as a sample from the posterior distribution and go to step 2.

5. Generate a random number, l, from a continuous uniform distribution defined between the limits of 0 and 1. If
$$l > \min\left\{1, \frac{\pi\left(\theta^*\right) q\left(\theta|\theta^*\right)}{\pi\left(\theta\right) q\left(\theta^*|\theta\right)}\right\} \tag{4.3}$$
then store θ as a sample from the posterior distribution and go to step 2.

6. Store θ^* as a sample from the posterior distribution. Set $\theta = \theta^*$ and go to step 2.

The density $q\left(\theta^*|\theta\right)$ in algorithm 2 is a proposal density. As long as the proposal density allows the proposal of parameter combinations that have some posterior support then the exact choice of the proposal distribution does not affect the asymptotic convergence properties of the Markov chain. However, the choice of the proposal density will affect the speed of convergence of the chain: proposal distributions similar to the posterior distribution perform most optimally but, given that in most cases the functional form of the posterior is not known, we often resort to simple proposal distributions with a simple symmetric distribution around the current parameter values.

Markov-Chain Monte Carlo methods in Bayesian analysis need to be applied with care however. The samples drawn from the chain can only truly be considered as samples from the posterior once the chain is run for infinite length of time. In practice, we are happy to accept the samples once the chain has been run for 'long enough' but what constitutes 'long enough' usually depends on the how quickly the chain converges to its stationary distribution. Moreover, when assessing values of θ with very low posterior support the number of rejected proposal values will increase, curtailing the ability of the chain to explore the available parameter space. For this reason, it can be difficult to adequately sample from multimodal target distributions, particularly if there is a large region of low probability separating the modal peaks. It is therefore vitally important when implementing these algorithms that adequate assessment is made of the convergence of the chain, either through visualisation of chain mixing (see Peltonen *et al.*, 2009) or using one or more of the diagnostics described by Mengersen *et al.* (1999).

Sequential Methods

Sisson *et al.* (2007), and later Toni *et al.* (2009), provide the details of a likelihood-free equivalent of a sequential Monte Carlo method using sequential importance sampling (SIS). Importance

sampling requires the use of a number of 'particles', with each particle containing a vector of potential samples of parameter values from the target distribution. The basic premise is that rather then sub sample the particles directly from an approximation to the posterior distribution, such as the algorithm of Pritchard *et al.* (1999), the particles are instead filtered using a number of intermediary distributions. This method attempts to circumvent the potential inefficiency of a high rejection rate when the prior and posterior distributions are very different by filtering the particles through T intermediate stages and replacing particles that are performing badly with new ones. The fact that particles are independently drawn, each with their own trajectory, means that particle filtering methods do not suffer from strong autocorrelation in the posterior sample brought about by slow mixing at local optima and in low likelihood parameter space to the same extent as MCMC methods. Sisson *et al.* (2007) supplement the basic particle filtering algorithm with an extra particle mutation step (see Del Moral *et al.*, 2006) such that the final version (after corrections as published in Toni *et al.*, 2009) is as follows:

Algorithm 3: Sequential importance sampling (Sisson *et al.*, 2007; Toni *et al.*, 2009)

1. Set the particle population iterator $t = 1$.

2. For each particle of the population of size N:

   (a) If $t = 1$, generate a proposed vector of parameter values for particle i, θ_t^*, from an initialisation distribution with density $\mu(\theta_t^*)$. In most implementations this initialisation distribution is set to the prior distribution, $\pi(\theta_t^*)$. For $t > 1$ randomly select a vector of parameter values with replacement from the population of particles, $\theta_{t-1} = \left[\theta_{t-1}^{(1)}, \ldots, \theta_{t-1}^{(N)}\right]$, with the probability of selecting a given particle set equal to its relative contribution to the vector of weights, $\mathbf{W}_{t-1} = \left[W_{t-1}^{(1)}, \ldots, W_{t-1}^{(N)}\right]$. This is the equivalent of drawing a particle from a categorical distribution with probability vector for the categories set equal to the normalised weight vector. Use the randomly selected vector of parameter values, $\theta_{t-1}^{(S)}$, to generate θ_t^* according to a proposal density $q\left(\theta_t^* | \theta_{t-1}^{(S)}\right)$.

   (b) If $\pi(\theta_t^*) = 0$ then go to step 2a.

   (c) Simulate a data set for particle i, $\hat{D}_t^{(i)}$, using the proposed parameter vector θ_t^*.

   (d) If $\rho\left[\mathbf{S}(D), \mathbf{S}\left(\hat{D}_t^{(i)}\right)\right] > \delta_t$ then go to step 2a. δ_t is a population specific acceptance constant that decreases monotonically as $t \to T$ such that $\delta_t > \delta_{t+1}$. Only δ_T affects

the approximation to the target distribution but previous acceptance constants do control the efficiency of the algorithm (see Sisson *et al.*, 2007; Toni *et al.*, 2009).

(e) Set

$$\theta_t^{(i)} = \theta_t^* \tag{4.4}$$

$$W_t^{(i)} = \begin{cases} \frac{\pi\left(\theta_t^{(i)}\right)}{\mu\left(\theta_t^{(i)}\right)} & \text{if } t = 1 \\ \frac{\pi\left(\theta_t^{(i)}\right)}{\sum_j W_{t-1}^{(j)} q\left(\theta_t^{(i)} | \theta_{t-1}^{(j)}\right)} & \text{otherwise} \end{cases} \tag{4.5}$$

3. Normalise $\mathbf{W}_t$ so that $\sum_i W_t^{(i)} = 1$.

4. If $t < T$ then increment the population iterator t and go to step 2.

5. Save the parameter vectors from the Tth population of particles, θ_T, as samples from the posterior distribution.

Individual-based models are, in nearly all cases, dynamic: that is, they do not assume equilibrium and simply generate results using this assumption. Equilibrium may emerge from the fundamental individual-level description of the process, but it is not a requirement of the modelling framework. Most individual-based models, such as the models described in this paper, simulate the emergence of a time series of data. In some cases we have a time series of data with which to compare to the outputs of the model in incremental stages. Given that the main computational cost of fitting these models using the algorithms described above is the simulation of data, it may be useful in these instances to curtail those simulations that are obviously performing badly (not adequately matching the data collected in the early time periods) and use the computational time saved to assess another combination of parameter values.

Here we describe a reformulation of the algorithm of Toni *et al.* (2009), specifically for the fitting of models to time series data. The model to be fit is simulated forward one time step at a time and the simulated outputs compared to the data in that time period only. Unlike the Toni *et al.* (2009) algorithm, where the model is run for the entire duration of the data collection period in each iteration and compared to the entire dataset, the algorithm described below breaks down the dataset into seperate time components, and resampling is done in each time period. This serves to filter out poorly performing particles without the need to simulate such particles through the entire data collection period. Under this specification, each particle

holds not only a set of parameter values but also the current state of the simulation at each time period.

Like the MCMC methods of algoritm 2, the SIS methods of algorithm 3 makes use of a proposal density $q\left(\theta_t^*|\theta_{t-1}^{(S)}\right)$ to generate new possible parameter values to test. This is required because in each iteration of the algorithm there is a selection of a new set of particles drawn at random from the old set. Without enrichment of the diversity of the particles, the compounding of this sampling effect over many iterations will result in a reduction of the diversity of the particles. This 'mutation' of the parameter values is often referred to as 'kernel smoothing' and is commonly applied in the fitting of dynamic models (see Liu & West, 2001; Thomas *et al.*, 2005; Harrison *et al.*, 2006; Newman *et al.*, 2006). However the application of the proposal kernel to the parameters results in a separation between the parameter values and the states that they generate. If the parameter values are greatly perturbed by the proposal kernel then particles may end up holding state information that is unlikely to have been generated using the perturbed parameters, introducing a bias into the analysis (Trenkel *et al.*, 2000; Harrison *et al.*, 2006).

Other authors have postulated that it is better to replace the kernel smoothing step with a sample from an MCMC sampler (Gilks & Berzuini, 2001; Khan *et al.*, 2005; Andrieu *et al.*, 2010). The argument follows that because the sample from the particle filter at any given time step is already a sample from the posterior distribution with respect to all data recorded up until that time step, the MCMC sampler can be already be said to have 'converged'. Samples generated using these methods can therefore be taken to be true samples from the posterior without requiring a formal test of convergence. Moreover, because standard MCMC tests require a re-evaluation of the likelihood, or a re-simulation in the case of approximate methods, the link between the parameters and the states of the model is maintained and the bias present in kernel smoothing methods does not exist in MCMC perturbation. The downside to this approach is that it requires twice as many simulations as the Kernel smoothing method. Below we adapt the MCMC particle filter to the approximate Bayesian framework for the fitting of IBMs to time-series data:

Algorithm 4: Sequential importance sampling with MCMC particle perturbation for time series simulation

1. Set the time iterator $t = 1$.

2. For each particle of the population of size N:

   (a) If $t = 1$ generate a proposed vector of parameters, $\theta_t^{(i)}$, from the prior distribution, $\pi\left(\theta_t^{(i)}\right)$. For $t > 1$, randomly sample, with replacement, a particle from the population of particles, $\theta_{t-1} = \left[\theta_{t-1}^{(1)}, \ldots, \theta_{t-1}^{(N)}\right]$.

   (b) Simulate forward one time step to create a simulated data set (states) for particle i, $\hat{D}_t^{(i)}$, using the parameter vector $\theta_t^{(i)}$ and the previous states of the particle, $\hat{D}_{0:(t-1)}^{(i)}$. For the simulation of models that exhibit Markovian properties, only $\hat{D}_{t-1}^{(i)}$ needs to be considered when generating the set of states for the next time step.

   (c) If $\rho\left[\mathbf{S}\left(D_t\right), \mathbf{S}\left(\hat{D}_t^{(i)}\right)\right] > \delta$ then go to step 2a.

   (d) Perturb the set of parameters according to the proposal distribution, $q\left(\theta_t^* | \theta_t^{(i)}\right)$, to create a candidate set of parameters θ_t^*.

   (e) Simulate from time $t-1$ to time t to create a proposed simulated data set (proposed states), $\hat{D}_t^*$, using the parameter vector θ_t^* and the previous states of particle i, $\hat{D}_{0:(t-1)}^{(i)}$.

   (f) If $\rho\left[\mathbf{S}\left(D_t\right), \mathbf{S}\left(\hat{D}_t^*\right)\right] <= \delta$ then generate random number, l, from a continuous uniform distribution defined between the limits of 0 and 1. If

$$l > \min\left\{1, \frac{\pi\left(\theta_t^*\right) q\left(\theta_t^{(i)} | \theta_t^*\right)}{\pi\left(\theta_t^{(i)}\right) q\left(\theta_t^* | \theta_t^{(i)}\right)}\right\} \tag{4.6}$$

   then set $\theta_t^{(i)} = \theta_t^*$ and $\hat{D}_t^{(i)} = \hat{D}_t^*$

3. If $t < T$ then increment the time iterator t and go to step 2.

4. Save the parameter vector from the T^{th} population of particles, θ_T, as samples from the posterior distribution.

Sometimes, not only the distribution of parameter values but also the distribution of model outputs used in the fitting process are of interest to the investigator. Inference about unobserved data or predictions outside the realm of the comparison data set is often the aim of the modelling exercise. Even though many locations may be unsampled, the simulation process may propose data values for these locations at the same time as generating data values for sampled locations. In this sense a missing data value can be considered as another parameter, values for which can

be sampled, along with the others, during the fitting process. It is therefore possible to use the distribution of simulated data values at unsampled locations in inferring possible values for these missing data points. Data generated during the parameter sampling process which resulted in a successful sample from the target distribution can be used for this purpose in algorithms 1 and 2. The evolution of the parameter over successive steps in the sequential sampling methods mean that only data generated in the final step of algorithm 3, $\hat{\mathbf{D}}_T = \left[\hat{D}_T^{(1)}, \ldots, \hat{D}_T^{(N)}\right]$, is a suitable approximation for inference purposes. None of the data generated in the fitting process in algorithm 4 is suitable for the inference of missing data. These data must be generated by re-simulating the time series with parameter values drawn from the posterior distribution.

4.2.2 Model Selection

Choosing between a set of possible models or weighting model outputs requires an assessment of performance. In the simplest instance, for comparing models of equal complexity, it may be sufficient to simply compare the fit of the models to the data. Metrics such as the sum of squares or, preferably, a likelihood-based metric would perform adequately in these occasions. As models become more heavily parametrised they are offered greater flexibility and hence the ability to achieve a better fit. Using metrics which only take into account the fit of the model to the data to compare models of differing complexity will result in the favouring of the more complex specifications. In such situations it is also important to balance the fit of the model against its complexity.

In classic maximum-likelihood based approaches to model fitting it is possible to use one of the many indices of information criteria to assess the models in terms of both fit and parsimony (such as those described in Burnham & Anderson, 2001). Bayesian models that have to be fit using Markov Chain Monte Carlo (MCMC) methods often have an analytically intractable maximum-likelihood value, and this quantity, the basis of many information criteria, needs to be approximated numerically. The deviance information criterion of Spiegelhalter *et al.* (2002) can be calculated from the standard MCMC output and thus removes the need for the Monte Carlo evaluation of extra metrics. However, because the methods described in this paper do not use or require the calculation of likelihoods, it is not possible to use standard information theoretic approaches, including DIC, to weight model outputs.

One way to compare model specifications is to include a model indicator, M, to be sampled jointly with the vector of parameters relevant for the model, θ_M, from the target distribution

$\mathbb{P}(M, \theta_M|D)$. The marginal density $\mathbb{P}(M = m|D)$ can be approximated by the proportion of samples taken from the posterior distribution where $M = m$. Using these marginal density estimates it is possible to calculate approximate values for the Bayes factors for each pair of models i and j:

$$B_{ij} = \frac{\mathbb{P}(M = i|D)\,\mathbb{P}(M = j)}{\mathbb{P}(M = j|D)\,\mathbb{P}(M = i)} \tag{4.7}$$

where $\mathbb{P}(M = m)$ is the prior support for model m. The Bayes factor, B_{ij}, summarises the support for model i over model j (see Kass & Raftery, 1995).

A number of joint model and parameter estimation algorithms exist; Grelaud *et al.* (2009) describe a simple extension of the rejection algorithm to allow the estimation of the marginal probability density of model structure. The extension of the sequential importance sampler of Sisson *et al.* (2007) described in Toni *et al.* (2009) is generalised in Toni & Stumpf (2010) to allow for joint estimation of the parameter and the model type. This version of the sequential Monte Carlo sampler uses estimates of the posterior support for each model type to draw initial values for a new model indicator in each iteration of the algorithm. This model indicator is perturbed according to model proposal distribution and a set of candidate parameter values are selected at random from the set of particles of the perturbed model type. Finally, the values for the parameters are perturbed according to a parameter proposal distribution before a simulation is made using the perturbed model type and parameter vector. The results of this simulation are compared to the observed data and the particle is accepted if it meets the required acceptance criteria. However, the estimate for posterior support in each iteration of the algorithm is made by calculating the relative frequency of the model indicator in the population of particles. If the number of models relative to the number of particles is high then the frequency of the relevant model indicators in the population of particles can be low and approximation of the posterior support can be poor. Whilst the algorithm provided by Toni & Stumpf (2010) appears to produce reasonable estimates of model performance, for computational feasibility its application is limited to cases where the number of candidate models are small.

Alternatively, Green (1995) describes a modification of the Metropolis-Hastings algorithm to allow for the movement between models with different numbers and types of parameters. If we define the vector of parameters associated with models m and m^* as θ and θ^* respectively, where model m has r_m parameters and model m^* has r_{m^*} parameters, then the implementation of

the 'reversible-jump' algorithm of Green (1995) requires that we also define a bijection function that can translate the values of the parameters of one model into the parameters of the other (and *vice versa*). For most applications, this function may also require the generation of a vector of random variables, u, such that the bijection describes the recoding of the parameters of model m and random variables u, into θ^* and vector of random variables u^*, where the bijection is given by $(\theta^*, u^*) = g_{mm^*}(\theta, u)$. The vector of random variables u^* is used in the bijection function, $g_{m^*m}(\theta^*, u^*)$, to describe the reverse move from model m^* to model m. There are a number of conditions that restrict the choice of bijection function however. First is the condition of reversibility, where

$$(\theta, u) = g_{mm^*}^{-1}(\theta^*, u^*) = g_{m^*m}(\theta^*, u^*) \tag{4.8}$$

Secondly, the bijection functions must be differentiable or at least partially differentiable with respect to each individual model parameter and random variable present in its list of arguments. Finally, if we define the number of parameters of the random vectors u and u^* as r_u and r_{u^*} respectively, then

$$r_m + r_u = r_{m^*} + r_{u^*} \tag{4.9}$$

This is known as the 'dimension matching condition'.

To account for the change in parameter and model type it is important to redefine the acceptance probability of the proposed transition. Under classic likelihood-based Bayesian analysis, the acceptance probability of a move from model m with parameters θ to model m^* with parameters θ^*, $a_{mm^*}(\theta, \theta^*)$, is given by

$$a_{mm^*}(\theta, \theta^*) = \min\left\{1, \frac{\pi(m^*, \theta^*)\,\mathcal{L}(m^*, \theta^*)\,k_{m^*\to m}q_{m^*m}(u^*)}{\pi(m, \theta)\,\mathcal{L}(m^*, \theta^*)\,k_{m\to m^*}q_{mm^*}(u)}\left|\frac{\partial g_{mm^*}(\theta, u)}{\partial\theta\partial u}\right|\right\} \tag{4.10}$$

where $k_{m\to m^*}$ is the probability of proposing a jump from model m to model m^* and $q_{mm^*}(u)$ is the probability of generating u from the proposal distribution defined when considering moves from model m to model m^*. $\left|\frac{\partial g_{mm^*}(\theta,u)}{\partial\theta\partial u}\right|$ is the absolute value of the determinant of the Jacobian matrix of the bijection $g_{mm^*}(\theta, u)$, and appears in equation 4.10 due to the deterministic transformation applied to parameters when jumping between model types (see Waagepetersen & Sorensen, 2001). $\mathcal{L}(m, \theta)$ is the likelihood of model m and parameter vector θ.

In approximate Bayesian analysis, simulations take the place of likelihood calculation and

so the likelihood terms do not appear in the acceptance probabilities when the reversible jump algorithm is adapted for approximate Bayesian computation. Below we show an adapted version of the reversible-jump MCMC algorithm for application in an approximate Bayesian setting:

Algorithm 5: Reversible-Jump Markov chain Monte Carlo sampling

1. Initialise the parameter vector θ and model indicator m with arbitrary values.
2. Propose a model m^* with probability $k_{m \to m^*}$.
3. Propose a vector of random values, u, drawn from a distribution with probability density $q_{mm^*}(u)$.
4. Apply the bijection, $g_{mm^*}(\theta, u)$, to generate a set of parameters θ^* for model m^*.
5. Simulate a dataset, $\hat{D}$, under model m^* using the proposal parameter vector θ^*.
6. If $\rho\left[\mathbf{S}(D), \mathbf{S}\left(\hat{D}\right)\right] > \delta$ then store θ and m as a sample from the posterior distribution and go to step 2.
7. Generate a random number, l, from a continuous uniform distribution defined between the limits of 0 and 1. If

$$l > \min\left\{1, \frac{\pi(m^*, \theta^*)\, k_{m^* \to m} q_{m^* m}(u^*)}{\pi(m, \theta)\, k_{m \to m^*} q_{mm^*}(u)} \left| \frac{\partial g_{mm^*}(\theta, u)}{\partial \theta \partial u} \right| \right\} \tag{4.11}$$

   then store θ and m as a sample from the target distribution and go to step 2.
8. Store θ^* and m^* as a sample from the target distribution. Set $\theta = \theta^*$ and $m = m^*$, and go to step 2.

In algorithm 4 we have already shown that it possible to embed a Metropolis-Hastings sampler in a particle filter to replenish particles without bias. The reversible-jump algorithm is no exception in this regard: Andrieu *et al.* (2010) and Khan *et al.* (2005) are two examples of studies that have implemented reversible-jump MCMC samplers within a particle filter in the standard likelihood-based Bayesian framework. This allows for joint model and parameter estimation when confronted with time series data. Below we describe an algorithm that integrates these approaches for approximate Bayesian computation:

Algorithm 6: Sequential importance sampling with reversible-jump MCMC particle perturbation for time series simulation

1. Set the time iterator $t = 1$.

2. For each particle of the population of size N:

   (a) If $t = 1$ generate a proposed vector of parameters, $\theta_t^{(i)}$, and a model type, $m_t^{(i)}$ from the prior distribution, $\pi\left(m_t^{(i)}, \theta_t^{(i)}\right)$. For $t > 1$, randomly sample, with replacement, a particle from the population of particles present at the end of the period last time period to provide a candidate vector of parameters, $\theta_t^{(i)}$, and a model type, $m_t^{(i)}$.

   (b) Simulate forward one time step to create a simulated data set (states) for particle i, $\hat{D}_t^{(i)}$, using model $m_t^{(i)}$, the parameter vector $\theta_t^{(i)}$, and the previous states of the particle, $\hat{D}_{0:(t-1)}^{(i)}$. For the simulation of models that exhibit Markovian properties, only $\hat{D}_{t-1}^{(i)}$ needs to be considered when generating the set of states for the next time step.

   (c) If $\rho\left[\mathbf{S}\left(D_t\right), \mathbf{S}\left(\hat{D}_t^{(i)}\right)\right] > \delta$ then go to step 2a.

   (d) Propose a model m_t^* with probability $k_{m_t^{(i)} \to m_t^*}$.

   (e) Propose a vector of random values, u, drawn from a distribution with probability density $q_{m_t^{(i)} m_t^*}(u)$.

   (f) Apply the bijection, $g_{m_t^{(i)} m_t^*}\left(\theta_t^{(i)}, u\right)$, to generate a set of parameters θ_t^* for model m_t^*.

   (g) Simulate from time $t-1$ to time t to create a proposed simulated data set (proposed states), $\hat{D}_t^*$, using model m_t^*, the parameter vector θ_t^*, and the previous states of particle i, $\hat{D}_{0:(t-1)}^{(i)}$.

   (h) If $\rho\left[\mathbf{S}\left(D_t\right), \mathbf{S}\left(\hat{D}_t^*\right)\right] <= \delta$ then generate random number, l, from a continuous uniform distribution defined between the limits of 0 and 1. If

$$l > \min\left\{1, \frac{\pi\left(m_t^*, \theta_t^*\right) k_{m_t^* \to m_t^{(i)}} q_{m_t^* m_t^{(i)}}\left(u^*\right)}{\pi\left(m_t^{(i)}, \theta_t^{(i)}\right) k_{m_t^{(i)} \to m_t^*} q_{m_t^{(i)} m_t^*}(u)} \left| \frac{\partial g_{m_t^{(i)} m_t^*}\left(\theta_t^{(i)}, u\right)}{\partial \theta_t^{(i)} \partial u} \right| \right\} \tag{4.12}$$

   then set $\theta_t^{(i)} = \theta_t^*$, $m_t^{(i)} = m_t^*$ and $\hat{D}_t^{(i)} = \hat{D}_t^*$

3. If $t < T$ then increment the time iterator t and go to step 2.

4. Save the parameter vector from the T^{th} population of particles, θ_T, as samples from the posterior distribution.

4.2.3 An Example Dataset

To illustrate the practical application of the methods described in this paper we use algorithm 6 to fit, and select between, a number of quasi-mechanistic models that could potentially describe the spatial dynamics of molehill production. We compare our model outputs to the dataset described in Schiffers *et al.* (2008) using a suite of assessment metrics. This dataset consists of eight experimental plots located throughout the Untere Havelaue nature reserve in western Brandenberg, Germany. Molehill locations were measured fortnightly at each of the sites according to the sampling regime illustrated in figure 4.1. At each sampling interval, the position of each molehill was recorded for all sites using a tachymeter (Elta-R, Zeiss, Oberkochen).

4.2.4 Models of Molehill Production

One potentially fruitful method of modelling molehill construction would be to look at the spatial properties of the molehill point pattern through time, and use one of the many well-described phenomenological point pattern models, such as those described in Diggle (2003). Whilst these methods may describe, and even predict, molehill appearances accurately, it is difficult to elucidate the mechanisms that drive the spatial properties of molehills from a statistical description of the pattern alone. If the aim of the modelling exercise is to better understand the processes that drive molehill formation, then we need derive a model that can at least emulate these processes.

For the purposes of giving an adequate explanation of the estimation methods employed, we consider here a series of minimally mechanistic models of molehill construction. This allows us to show an application of the methods with some structural realism but avoiding the level of mechanistic detail that would swamp the discussion with details of model implementation. We hope that although the model may be simpler than the individual-based models that are commonly applied to ecological problems, the extensions of the methods to cover IBMs of increased complexity should be intuitive to the reader.

The first model considered is a simple point-process description of molehill 'birth', 'death' and 'dispersal' and takes a similar form to many IBMs in which individuals are represented as a particle in continuous space. In this sense we treat molehills a little like plants, where, in each generation, each individual seeds a random number of 'offspring' which disperse from the parent individual according to a given dispersal kernel before being thinned by a death process. In this

particular example, each individual produces a number of offspring at each generation drawn from a Poisson distribution with mean λ. Offspring disperse from the parent individual at a random distance drawn from an exponential distribution, with mean β, in a random direction that is uniformly drawn between the radian limits of 0 and 2π. Finally, each individual that is not newly-born is removed with a probability τ, representing the removal of molehills from weathering or trampling damage.

The data do not allow for an accurate estimation of mole population densities and so any models derived to describe molehill production need to describe new molehill formation only with respect to the distribution of molehills in the last time period. Modelling molehills as reproducing entities may not make sense on biological grounds but this assumption may provide a sufficient caricature of the relevant point-process dynamics and supply the basis for the more complex departures derived in later models. Indeed, whilst the model lacks an explicit description of the below-ground activity that drives molehill production it may supply some implicit insights.

Under the model described above, the probability that a site will become the location of a new molehill decays exponentially with distance, but equally in all directions, from the parent molehill. However, from the point-patterns of molehill locations published in Schiffers *et al.* (2008) we can see that facets of the below-ground tunnel network are clearly visible on the observed above-ground pattern. It is very unlikely that this pattern could be formed from a spatially isotropic generation process.

Our second model extends the dynamics described in the first model to include an anisotropic dispersal kernel. Instead each molehill 'individual' holds an extra state variable, θ_i, the angle at which the individual dispersed from its parent molehill. Each new molehill is positioned in a direction from the parent molehill drawn at random according to the von Mises distribution (sometimes known as the circular normal distribution, see Fisher, 1993) with a mean equal to the state variable θ_i of the parent molehill and a concentration parameter, analogous to the reciprocal of the variance in unwrapped distributions, κ. In the limit $\kappa \to 0$ the dispersal direction follows a standard uniform distribution between the limits of 0 and 2π; the first model can therefore be thought of a special case of the second model with the value of κ approaching this limit. The state variable θ_i awards the individuals simulated some form of directional 'memory'. Under this specification it is apparent that the locations of any one molehill lineage will

resemble that of a sample of points taken from the path of a correlated random walk (Turchin, 1998), albeit with different functional forms given for step length and orientation.

The creation of a directional memory may allow a more branching point structure, as would be expected for a phenomenon that is sampled from a subterranean network exhibiting such patterns, but with every molehill contributing a statistically equal number of offspring to the next time period it will result in a rather 'bushy' structure. The burrow systems of another subterranean rodent, the silvery mole rat (*Heliophobius argenteocinereus*), as illustrated in Škliba *et al.* (2009), exhibit a small number of long main tunnels with a number of shorter branching side-tunnels. A set of points sampled from such a network would appear quite different from those simulated from the second model, regardless of parameterisation.

Affording model flexibility in order to produce the kind of point patterns that could have been produced from the subterranean networks of Škliba *et al.* (2009) requires a reformulation of the birth process used in the first two models. In this final specification, the total number of new molehills to be generated in any time period is drawn from a Poisson distribution with mean $N\lambda$, where N is the current molehill count. Rather than distribute these offspring amongst the parents for placement these new molehills are instead placed sequentially: each new molehill is assigned the last placed individual as its parent with probability ϕ (for the first offspring to be placed in any time period this is individual placed last in the previous time period), otherwise a parent is allocated at random from those individuals that are not newly-born in the current time period. Once parentage for an offspring is assigned then it is dispersed from the parent individual according to the mechanism described in the second model. In the first time period the first offspring to be placed is assigned a parent at random from the individuals present at model initialisation.

The parameter ϕ has the effect of controlling the 'bushiness' of the underlying subterranean network. In the extreme, a value of 1 for ϕ will result in a series of points lying along a single tunnel with no branching tributaries, with each new offspring automatically assigned parentage of the next offspring in sequence. When $\phi = 0$ every offspring is allocated parentage at random from the set of survivors from the last time period and can be simulated by allocating parentage for the offspring according to a multinomial distribution with a probability vector of identical elements, each with the value $\frac{1}{N'}$, where N' is the number of surviving molehills from the last time period. This is the statistical equivalent of generating a separate independently and

identically distributed Poisson number of offspring for each parent (Johnson *et al.*, 1997) and is the same as the birth process described in the first and second models. In this sense, both the first and second models are special cases of the third model. Table 4.1 summarises the models described here and the parameters that control their behaviour. Figure 4.2 shows an example of one realisation from each of the models.

4.2.5 Implementation

We calculate the parameters and optimal model for the molehill data set using the sequential Monte Carlo algorithm for time series data of algorithm 6. We draw parameter values for the initial population of particles from the independent set of minimally informative priors:

$$\tau \sim \mathrm{U}(0,1) \tag{4.13}$$

$$\lambda \sim \mathrm{TN}\left(\frac{1}{2}, 6, 0, \infty\right) \tag{4.14}$$

$$\beta \sim \mathrm{TN}(3, 10, 0, \infty) \tag{4.15}$$

$$\kappa \sim \mathrm{TN}(10, 50, 0, \infty) \tag{4.16}$$

$$\phi \sim \mathrm{U}(0,1) \tag{4.17}$$

where $\mathrm{U}(a,b)$ is a continuous uniform distribution with density function $f(x|a,b) = \frac{1}{b-a}$. $\mathrm{TN}(\mu,\sigma,a,b)$ is a truncated normal distribution with density function

$$h(x|\mu,\sigma,a,b) = \frac{1}{\sqrt{2\pi}\sigma^2\left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right]} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.18}$$

with $\Phi(x)$ as the normal cumulative density function. Each of three models are given uniform prior weight and the particles are initialised according to this prior distribution.

Each particle is initialised with a set of points taken from the first time slice of data of plot 1. Each individual is initialised with a dispersal bias state variable (θ) drawn randomly from a continuous uniform distribution with limits 0 and 2π, even for particles that are assigned a model indicator for a model that does not allow for dispersal with a directional bias. We iterate through the time steps, simulating data and filtering parameter values according to comparison to the data recorded from plot 1. Once we have finished iterating through the data we are left with a sample of parameter values drawn from the posterior distribution according to the data recorded in plot 1. We then restart the algorithm at the first record in plot 2. Individuals and

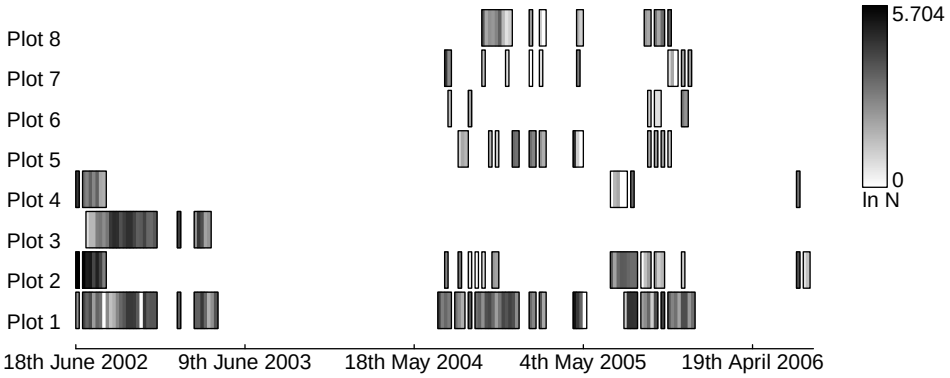
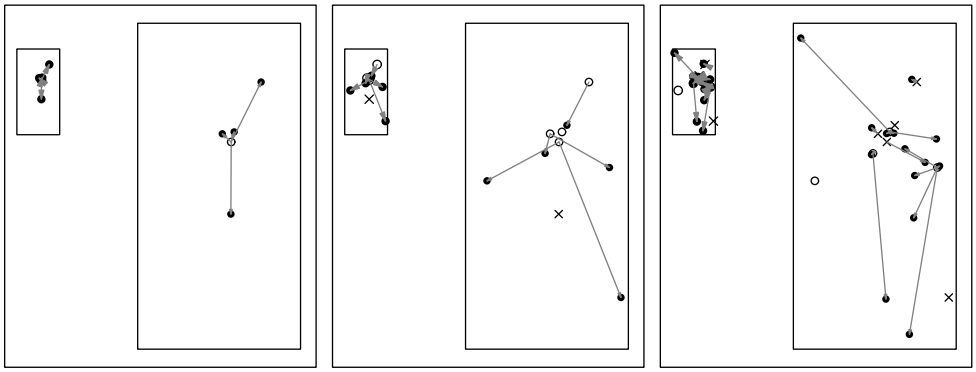


Figure 4.1: Series of sampling schedules for each sampling location in the Untere Havelaue nature reserve in western Brandenburg, Germany. Rectangles indicate that sampling was active during that period. Shading relates to $\ln(N_t)$, the natural logarithm of the number of molehills found at time period t.

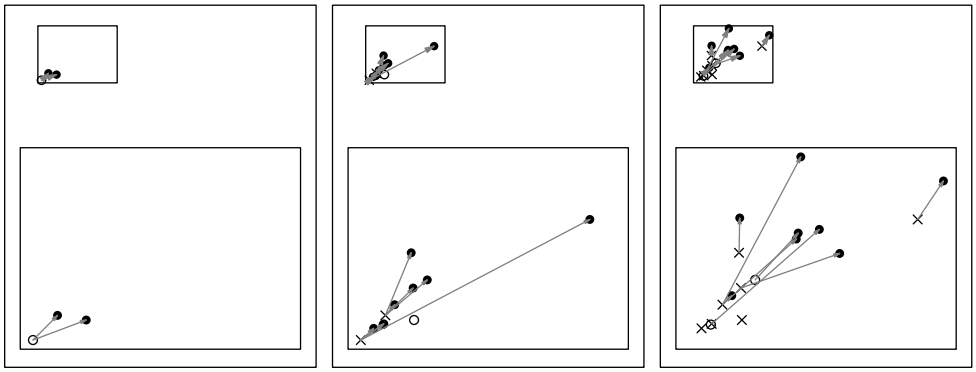
	Parameters	*Model*		
		1	*2*	*3*
$0 \leq \tau \leq 1$	Probability that a molehill is removed in each time period	✓	✓	✓
$\lambda > 0$	The average number of new molehills created per molehill in each time period	✓	✓	✓
$\beta > 0$	The average distance an 'offspring' molehill disperses from the 'parent' molehill in any given dispersal event	✓	✓	✓
$\kappa \geq 0$	Concentration parameter determining the inter-generational correlation in dispersal direction	×	✓	✓
$0 \leq \phi \leq 1$	Parameter determining the proportion of new molehills generated in any time period that lie on new tunnel systems	×	×	✓

Table 4.1: A list of parameters, with a brief explanation of their purpose, used in each of the models described in this study. A tick (✓) represents a free parameter in the model that is to be estimated from the data. A cross (×) represents a parameter that is absent from the model, which in this example, is the same as using the third model but fixing the relevant parameter at zero.

(a) $\tau = \frac{1}{2}$, $\lambda = \frac{3}{2}$, $\beta = \frac{1}{5}$



(b) $\tau = \frac{1}{2}$, $\lambda = \frac{3}{2}$, $\beta = \frac{1}{5}$, $\kappa = 10$



(c) $\tau = \frac{1}{2}$, $\lambda = \frac{3}{2}$, $\beta = \frac{1}{5}$, $\kappa = 10$, $\phi = \frac{9}{10}$

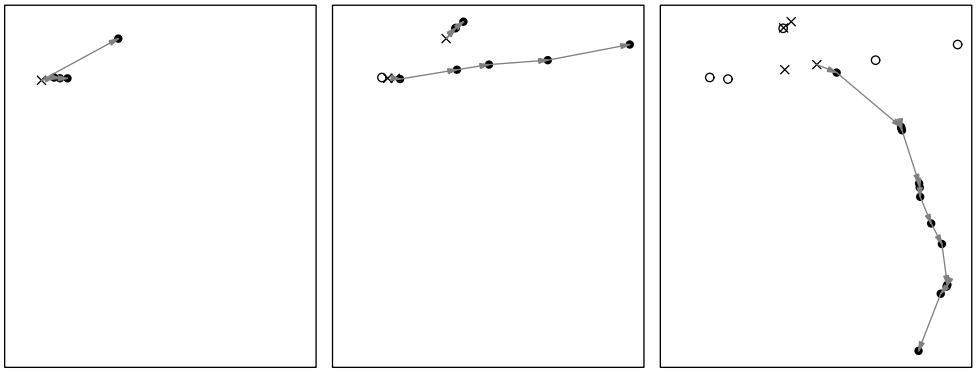


Figure 4.2: Three consecutive generations from a single realisation of each of three models described in this paper. Models 1 to 3 correspond to figures (a) to (b) respectively. In each figure, a filled circle (•) represents the location of a newly-created molehill. Molehills created in previous generations, and which survive the current generation, are denoted by an open circle (∘). Molehills that are removed at the end of the current generation are denoted by crosses (×). A grey arrow connects 'parent' molehills to their 'offspring'. Each realisation is initialised with one molehill (present at the same location for each molehill) using the parameter values shown by the labels. The sub-panels present in series (a) and (b) show an enlargement of the region contained within the area bounded by the smaller rectangle.

state variables for each particle are initialised in exactly the same way as described above but starting values for the parameter vectors and model indicators are drawn from the population of particles that survived the filtering process from iteration through the time steps of the first plot, instead of from the prior distribution. This process is repeated for each plot until the particles have been exposed to the data contained in the time periods of all plots. The final distribution of parameter values and model indicators represents a sample taken from the posterior distribution with respect to the entire data set.

Jumps between model types are proposed with the following probabilities:

$$\begin{aligned} k_{1\to1} &= 0.6 \quad k_{1\to2} = 0.3 \quad k_{1\to3} = 0.1 \\ k_{2\to1} &= 0.2 \quad k_{2\to2} = 0.6 \quad k_{2\to3} = 0.2 \\ k_{3\to1} &= 0.1 \quad k_{3\to2} = 0.3 \quad k_{3\to3} = 0.6 \end{aligned} \tag{4.19}$$

Under all model jumps at least three random numbers, (u_1, u_2, u_3), are generated from a normal distribution with variances σ_1, σ_2, and σ_3, respectively and a mean of zero. The new proposed value for τ, λ, and β (τ^*, λ^*, and β^* respectively) are related to u_1, u_2, and u_3 such that

$$\begin{aligned} \tau^* &= \text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right] \\ \lambda^* &= \lambda e^{u_2} \\ \beta^* &= \beta e^{u_3} \end{aligned} \tag{4.20}$$

Generating proposed values for the κ and σ parameters (κ^* and ϕ^* respectively) for models that contain them requires a little more care however. Jumps to models that have the κ parameter (models 2 and 3) require the generation of a fourth random number, u_4, drawn from a normal distribution with a mean of zero and a variance of σ_4. If the jump is from a model that does not contain the κ parameter (model 1) then $\kappa^* = e^{u_4}$, otherwise $\kappa^* = \kappa e^{u_4}$. Similarly, jumps to models that have the ϕ parameter (model 3) require the generation of a fifth random number, u_5, drawn from a normal distribution with a mean of zero and a variance of σ_5. $\phi^* = \text{logit}^{-1}\, u_5$ if the jump originates from a model does that does not contain a ϕ parameter, otherwise $\phi^* = \text{logit}^{-1}\left[\text{logit}\,\phi + u_5\right]$. The full bijection specification for all jump types is given below:

Jumps from model 1 to model 1

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, u_1^*, u_2^*, u_3^*) &= g_{11}(\tau, \lambda, \beta, u_1, u_2, u_3) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, -u_1, -u_2, -u_3\right)\end{aligned} \tag{4.21}$$

Jumps from model 1 to model 2

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, \kappa^*, u_1^*, u_2^*, u_3^*) &= g_{12}(\tau, \lambda, \beta, u_4, u_1, u_2, u_3) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, e^{u_4}, -u_1, -u_2, -u_3\right)\end{aligned} \tag{4.22}$$

Jumps from model 1 to model 3

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, \kappa^*, \phi^*, u_1^*, u_2^*, u_3^*) &= g_{13}(\tau, \lambda, \beta, u_4, u_5, u_1, u_2, u_3) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, e^{u_4}, \text{logit}^{-1}\,u_5, -u_1, -u_2, -u_3\right)\end{aligned} \tag{4.23}$$

Jumps from model 2 to model 1

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, u_4^*, u_1^*, u_2^*, u_3^*) &= g_{21}(\tau, \lambda, \beta, \kappa, u_1, u_2, u_3) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, \ln\kappa, -u_1, -u_2, -u_3\right)\end{aligned} \tag{4.24}$$

Jumps from model 2 to model 2

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, \kappa^*, u_1^*, u_2^*, u_3^*, u_4^*) &= g_{22}(\tau, \lambda, \beta, \kappa, u_1, u_2, u_3, u_4) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, \kappa e^{u_4}, -u_1, -u_2, -u_3, -u_4\right)\end{aligned} \tag{4.25}$$

Jumps from model 2 to model 3

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, \kappa^*, \phi^*, u_1^*, u_2^*, u_3^*, u_4^*) &= g_{23}(\tau, \lambda, \beta, \kappa, u_5, u_1, u_2, u_3, u_4) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, \kappa e^{u_4}, \text{logit}^{-1}\,u_5,\right. \\ &\qquad \left. -u_1, -u_2, -u_3, -u_4\right)\end{aligned} \tag{4.26}$$

Jumps from model 3 to model 1

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, u_4^*, u_5^*, u_1^*, u_2^*, u_3^*) &= g_{31}(\tau, \lambda, \beta, \kappa, \phi, u_1, u_2, u_3) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, \ln\kappa, \text{logit}\,\phi, -u_1, -u_2, -u_3\right)\end{aligned} \tag{4.27}$$

Jumps from model 3 to model 2

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, \kappa^*, u_5^*, u_1^*, u_2^*, u_3^*, u_4^*) &= g_{32}(\tau, \lambda, \beta, \kappa, \phi, u_1, u_2, u_3, u_4) \\ &= \big(\text{logit}^{-1}[\text{logit}\,\tau + u_1], \lambda e^{u_2}, \beta e^{u_3}, \kappa e^{u_4}, \text{logit}\,\phi, \\ &\qquad -u_1, -u_2, -u_3, -u_4\big)\end{aligned} \tag{4.28}$$

Jumps from model 3 to model 3

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, \kappa^*, \phi^*, u_1^*, u_2^*, u_3^*, u_4^*, u_5^*) &= g_{33}(\tau, \lambda, \beta, \kappa, \phi, u_1, u_2, u_3, u_4, u_5) \\ &= \big(\text{logit}^{-1}[\text{logit}\,\tau + u_1], \lambda e^{u_2}, \beta e^{u_3}, \kappa e^{u_4}, \\ &\qquad \text{logit}^{-1}[\text{logit}\,\phi + u_5], -u_1, -u_2, -u_3, -u_4, -u_5\big)\end{aligned} \tag{4.29}$$

The derivation of the determinant of the Jacobian matrix for each bijection can be found in the appendix of this chapter.

Four statistics are used to assess the performance of the simulated output in recapturing the required facets of the real data. Firstly, we use the population counts of molehills at a given time period. For particle acceptance the population of molehills in the simulated data must be within 14 individuals of the population present in the data set at the relevant time period. This threshold is set at 5% of the observed range of values for population counts in the data taken across all time periods and sites. If we let N_{D_t} and $N_{\hat{D}_t^*}$ be the molehill population sizes in the real and simulated data respectively at time t and C_1 be the criterion acceptance value (here equal to 14), then the first acceptance criterion is met if the following inequality is satisfied:

$$|N_{D_t} - N_{\hat{D}_t^*}| \leq C_1 \tag{4.30}$$

Secondly, we use summary statistics based on the empirical distribution functions of nearest-neighbour distances, $\hat{G}(r)$, and the second-order variances of point-to-point distances (using Ripley's K function), $\hat{K}(r)$ (see pages 17-20 and chapter 4 of Diggle, 2003) on the spatial point pattern of molehills at each time period. Taken collectively, these statistics allow for the assessment of clustering or uniformity in the point pattern. Both statistics are functions of a radial search variable, r. In order to boil the functions down to simple rejection criteria, we calculate the sum of the squared differences between the functions calculated on the real and simulated data, $\hat{G}_{D_t}(r)$ and $\hat{K}_{D_t}(r)$ versus $\hat{G}_{\hat{D}_t^*}(r)$ and $\hat{K}_{\hat{D}_t^*}(r)$ respectively, evaluated at every 10cm interval between 0m and 35m such that the two conditions of proposed particle

acceptance rest on the satisfaction of the two inequalities

$$\sum_r \left[\hat{G}_{D_t}(r) - \hat{G}_{\hat{D}_t^*}(r)\right]^2 \leq C_2 \tag{4.31}$$

$$\sum_r \left[\hat{K}_{D_t}(r) - \hat{K}_{\hat{D}_t^*}(r)\right]^2 \leq C_3 \tag{4.32}$$

where $r \in \left\{x : (\exists k \in \mathbb{N}_0)\left(x = \frac{1}{10}k\right), 0 \leq x \leq 35\right\}$. Note that for computational efficiency, neither statistics are calculated with any edge correction (see pages 5-6 of Diggle, 2003). For the purposes of comparison between two point patterns this should be sufficient as edge effects are treated equally for both point patterns.

Our final acceptance criterion relates to the level of directional bias in the data. A number of measures of anisotropy exist (see Rosenberg, 2004, 2000; Simon, 1997; Mugglestone & Renshaw, 1996, for examples) but all require either sophisticated and computationally expensive calculations, particularly when calculated for every proposed particle value, or require decisions on analysis parameters that make the application of such techniques difficult to automate. Here we propose a simple, *ad hoc* metric to estimate the degree of anisotropy in the point patterns of both the data and the simulation outputs in any time period. We base this metric on the deviation from the null (isotropic) hypothesis that for every point, the distribution of angular directions of all other points represents a sample from a continuous uniform distribution defined between the limits of 0 and 2π. We assess this deviation from circular uniformity by calculating the p-value resulting from a one-sample Kolmogorov-Smirnov test. By using each point present as the reference point in turn, the median of the resultant distribution of p-values is calculated as an estimation of the degree of anisotropy present in the point pattern. If we let V_{D_t} and $V_{\hat{D}_t^*}$ be the median p-values calculated from the real and simulated point patterns respectively then our final criterion correspondence to the validity of the following inequality:

$$|V_{D_t} - V_{\hat{D}_t^*}| \leq C_4 \tag{4.33}$$

We set C_4 similarly to the logic of setting the other acceptance criteria, corresponding to 5% of the range of values of the statistic in the observed data, here 1.1120247×10^{-2}.

In some time periods the number of molehills in the data fall too low to calculate the statistics described above. Time periods where there are fewer than five data points, for which there is only a total of 40 across the 218 time periods for each of the 8 sites, are treated in exactly

the same way as missing data (see below). Simulation output with fewer than three data points are automatically rejected before the spatial statistics are calculated.

In simulating molehill patterns we simulate the model in time blocks corresponding to a weekly period in the real data. Parameter values are therefore scaled to a per weekly basis. From figure 4.1 it is clear that there are gaps in sampling for each of the eight plots. For the purposes of the particle filtering we simply move to the next sampled time period, skipping those time periods with no data attached to them, but to correct for this, we increment the model according to the number of skipped time periods at the relevant data generation steps (steps 2b and 2g of algorithm 6).

4.3 Results

Figure 4.3 displays the marginal density estimates from 100000 particles filtered according to the implementation described previously. Model 1 has very little posterior support, occurring in just 4% of the filtered particles. There is however little to distinguish between the two, more complex models: model 2 occurs in approximately 51% of the filtered particles with the remaining 45% supporting model 3. Indeed, even when model 3 is selected we can see from figure 4.3 that the values for ϕ are generally very low. Given that it is the addition of the ϕ parameter that distinguishes model 3 from model 2, and that low values for ϕ result in sequential molehill placement at very low frequencies, we can see that this placement procedure does not bring about a substantial advantage in the description to the spatial distribution dynamics of molehills.

From figure 4.3 it is clear that the 95% credible interval for parameter τ covers most of the possible range of potential values for the parameter. This suggests that there is very little information present in the data set to discriminate between possible values for τ. To get a better estimate for τ it is necessary to include extra information from other data sources. This information can be expressed in the form of an informative prior, restricting the parameters of τ to values that are likely in the context of a broader study and allowing better estimation of parameters that co-vary with this parameter. The lack of differentiation of τ between model types is not unexpected given its wide distribution in all of the cases.

The dispersal parameter β is highly differentiated between the particles of the different model types. Median molehill 'dispersal' is highest in model 1, followed by model 2, and finally, with the shortest dispersal distances, model 3. This effect could arise from the fact that for any set of parameter values, molehills appear much more aggregated in simulations from model 1 and 2 than a realisation from model 3 with even a low ϕ parameter (see figure 4.2). In order to emulate the spatial dispersion in the data set it is therefore critical that particles of model 1 and model 2 exhibit high dispersal ability to achieve a spacing between molehills that is comparable to the data. Whilst the filtered particles give overwhelming support for the inclusion of an 'inter-generational' correlation of dispersal direction, controlled by the κ parameter in models 2 and 3, there is little differentiation between in the range of credible values between the two models. Moreover the credible interval for κ is broad and the marginal distribution is flat-topped under both model specifications. This suggests that there is strong support for some 'inter-generational' correlation of dispersal direction, as the posterior distribution has little density for values close to zero, but that there is not enough information to discriminate against particular values of κ in the mid-range.

4.4 Discussion

The traditional pattern-oriented approach to the fitting of complex models dictates that we take a selection of parameter values and run multiple realisations of the model we are trying to fit for each parameter combination. We compare facets of the simulated data to the real data using a metric that permits an assessment of the performance of the parameters to match the real data. Parameter combinations that perform well are stored to be investigated further and parameter combinations that performed badly are removed from further analysis.

Whilst rarely specified in such terms, these attempts to sort the likely parameter combinations from the unlikely, are really an attempt to sample from a distribution of parameter values in the same proportion as their probability in the light of the data and what we previously knew about the system: the posterior distribution. The first two steps of pattern-oriented modelling, as defined by Wiegand *et al.* (2003), involves the aggregation of biological information and the estimation of ranges of parameter values. In a Bayesian parlance this sounds very similar to the specification of priors. Step three according to Wiegand *et al.* (2003) is the systematic comparison between the observed patterns and the patterns predicted by the model. This paper details six algorithms that can perform such systematic filtering of parameter values according

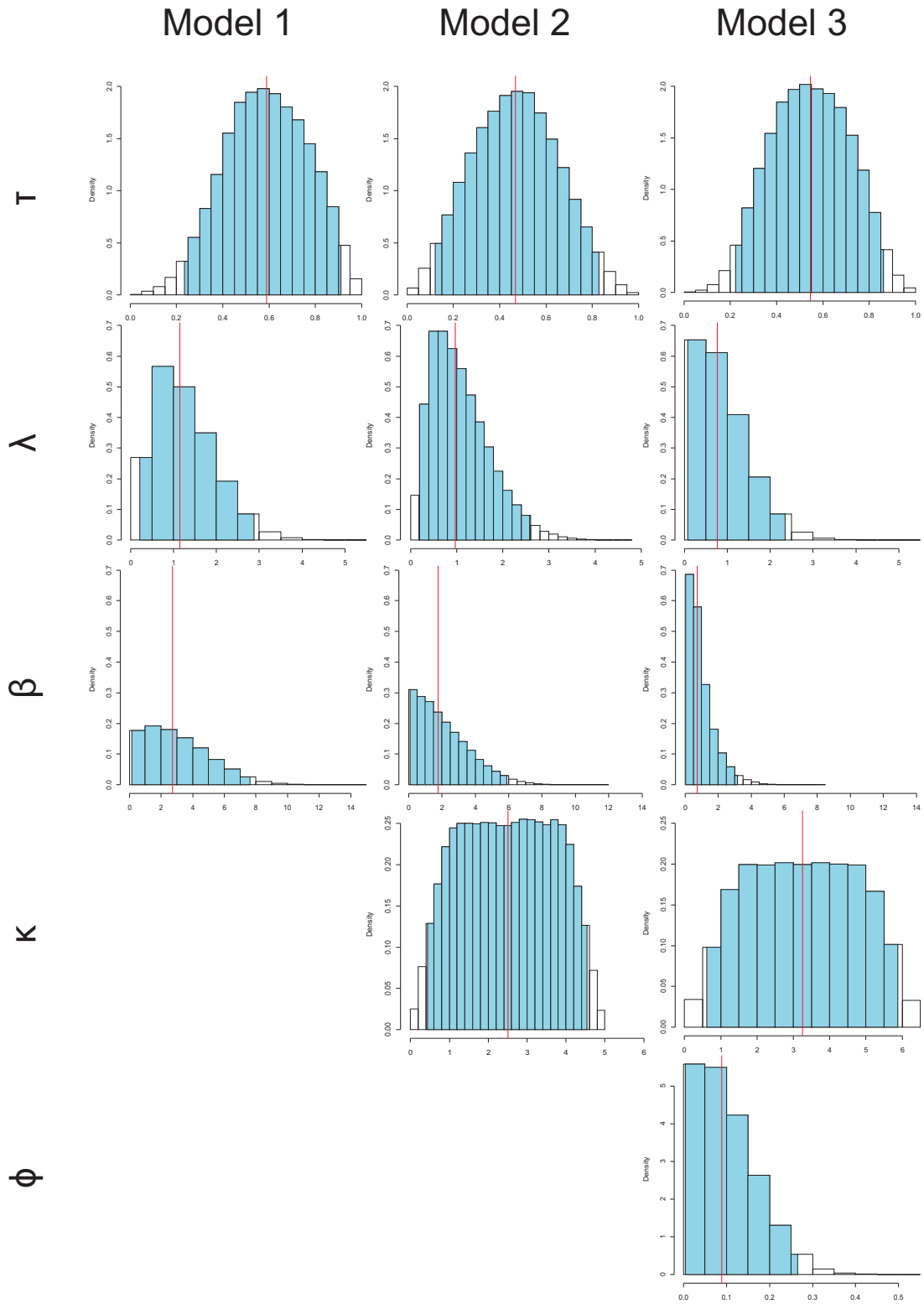


Figure 4.3: Histograms of the marginal densities of the parameter values from 10000 filtered particles for each of the three models of molehill production. The shaded blue region denotes the 95% credible interval for each of the parameter values and the red line denotes the median value of the posterior sample

to the set of comparison statistics that serve to condense the patterns of importance that need to replicated in the model. The parallels appear striking.

Where the methodology outlined in this paper differs from that laid out in Wiegand *et al.* (2003) is that the techniques described here, subject to a few caveats, guarantee the convergence to a sample of parameters from the posterior distribution, albeit an approximation. There are a number of examples of pattern-oriented modelling applied to ecological problems ranging from the implementation of automated parameter selection techniques (Kramer-Schadt *et al.*, 2004; Swanack *et al.*, 2009) to the simulation of a small number of scenarios to test parameter sensitivity (Zinck & Grimm, 2009). The methods employed in these papers may indeed provide estimates of the range of likely parameter values but are unlikely to recreate the relative density of parameter values present in the posterior distribution. Separating 'good' parameter combinations from 'bad' combinations can only be made relative to the selection of parameters tested. If the number of scenarios tested are too few then the risk is run that the best selected parameter set is sufficiently different than the set most likely according to the posterior density. Testing the full range of likely scenarios becomes costly as the dimensionality of the parameterisation increases however. Whilst Bayesian methods do not make the curse of high dimensionality go away, they do provide a systematic and efficient way of searching the parameter space. Providing not only point estimates of 'good' parameter values but also recreating a distribution of parameters with posterior support.

Without adopting the techniques described in this paper, model selection for individual-based models can be a difficult affair. Without an available likelihood it is impossible to apply any of the information theoretic approaches to model selection Burnham & Anderson (2001) and it is difficult to heuristically assess how much extra fit to the desired pattern merits an increase in model complexity. Bayesian methods not only provide a way of selecting appropriate models from a set of candidate models but they can also assign weights to be used in model averaging (Hoeting *et al.*, 1999).

Beyond parameterisation, approximate Bayesian techniques may also be used to synthesise otherwise disparate sources of data. The need to specify priors, often thought a deficiency of the Bayesian analysis, actually benefits the investigator by allowing the input of information derived from other means such as experimental data to form a key part of the model predictions. This particularly important for situations where many data may exist but it is contained in

many small studies with differing objectives and protocols. One example where there exists a plethora of data is the case of the Australian cane toad (*Chaunus [Bufo] marinus*). Its invasion of Eastern Australia has been well documented and it has because a serious pest species. For this species there is much known about the ecophysiological tolerances of this species but the methods of predicting the invasion dynamics using mechanistic models based on environmental biology have been applied independently of correlative methods of range prediction (Phillips *et al.*, 2008). By setting known physiological tolerances as priors on the relevant parameters, it is possible using these methods to describe a mechanistic framework for the species distribution dynamics that is both able recreate dynamic patterns of range changes whilst maintaining biological realism, even if the model description is sufficiently complex to make likelihood calculation intractable.

Individual-based models have been criticised for their typically high parameter load which relies heavily on inference from indirect parameterisation Kramer-Schadt *et al.* (although see 2007). The inability to specify a likelihood function for these parameters makes the task of fitting these models very difficult. This places a high burden on the investigator and can make the application of IBMs appear unnatractive even if the use of these sorts of models to model the study system makes sense from a conceptual point-of-view. This methodology, by drawing in all information that we know about the system of interest, and allowing us to use this information to parameterise the models and choose between competing model architectures, demonstrates one way in which the future of individual-based model research may overcome these hurdles.

Appendix 4.A Derivation of Jacobian determinants for reversible jump MCMC implementation

The function g_{11} defines the transformation of the set of parameters (τ, λ, β) with regard to a vector of randomly generated elements (u_1, u_2, u_3) when proposing moves between parameter values within model 1.

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, u_1^*, u_2^*, u_3^*) &= g_{11}(\tau, \lambda, \beta, u_1, u_2, u_3) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, -u_1, -u_2, -u_3\right)\end{aligned} \tag{4.34}$$

To retain the balance condition, functions which describe proposals of new parameters within a model must be involutary: $g_{11}(g_{11}(\tau, \lambda, \beta, u_1, u_2, u_3)) \equiv (\tau, \lambda, \beta, u_1, u_2, u_3)$. From equation

4.34 we can derive the Jacobian matrix of function g_{11}:

$$\frac{\partial g_{11}\left(\tau,\lambda,\beta,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial u_1\,\partial u_2\,\partial u_3} = \begin{bmatrix} \frac{e^{-u_1}}{(e^{-u_1}\tau-\tau-e^{-u_1})^2} & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{(e^{-u_1}\tau-\tau-e^{-u_1})^2} & 0 & 0 \\ 0 & e^{u_2} & 0 & 0 & \lambda e^{u_2} & 0 \\ 0 & 0 & e^{u_3} & 0 & 0 & \beta e^{u_3} \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.35}$$

Given that the Jacobian matrix of equation 4.35 is triangular, the magnitude of the determinant is simply the absolute value of the product of the diagonal components such that

$$\left|\frac{\partial g_{11}\left(\tau,\lambda,\beta,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial u_1\,\partial u_2\,\partial u_3}\right| = \frac{e^{u_2+u_3-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} \tag{4.36}$$

Function g_{12} describes the bijection for proposals of parameter values of model 2 given the current values for the parameters in model 1, (τ,λ,β), and a vector of randomly generated elements, (u_1,u_2,u_3,u_4), where

$$\begin{aligned}(\tau^*,\lambda^*,\beta^*,\kappa^*,u_1^*,u_2^*,u_3^*) &= g_{12}\left(\tau,\lambda,\beta,u_4,u_1,u_2,u_3\right) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau+u_1\right],\lambda e^{u_2},\beta e^{u_3},e^{u_4},-u_1,-u_2,-u_3\right)\end{aligned} \tag{4.37}$$

with the Jacobian matrix for the bijection given as

$$\frac{\partial g_{12}\left(\tau,\lambda,\beta,u_4,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial u_4\,\partial u_1\,\partial u_2\,\partial u_3} = \begin{bmatrix} \frac{e^{-u_1}}{(e^{-u_1}\tau-\tau-e^{-u_1})^2} & 0 & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{(e^{-u_1}\tau-\tau-e^{-u_1})^2} & 0 & 0 \\ 0 & e^{u_2} & 0 & 0 & 0 & \lambda e^{u_2} & 0 \\ 0 & 0 & e^{u_3} & 0 & 0 & 0 & \beta e^{u_3} \\ 0 & 0 & 0 & e^{u_4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.38}$$

The Jacobian matrix in equation 4.38 is triangular and so the magnitude of the determinant is

simply

$$\left|\frac{\partial g_{12}\left(\tau,\lambda,\beta,u_4,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial u_4\,\partial u_1\,\partial u_2\,\partial u_3}\right| = \frac{e^{u_2+u_3+u_4-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} \tag{4.39}$$

Function g_{13} describes the conversion from a set of parameters in model 1, (τ,λ,β), and a vector of randomly generated numbers, (u_1,u_2,u_3,u_4,u_5), to a set of parameters in model 3, $(\tau^*,\lambda^*,\beta^*,\kappa^*,\phi^*)$:

$$\begin{aligned}\left(\tau^*,\lambda^*,\beta^*,\kappa^*,\phi^*,u_1^*,u_2^*,u_3^*\right) &= g_{13}\left(\tau,\lambda,\beta,u_4,u_5,u_1,u_2,u_3\right)\\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau+u_1\right],\lambda e^{u_2},\beta e^{u_3},e^{u_4},\text{logit}^{-1}\,u_5,-u_1,-u_2,-u_3\right)\end{aligned} \tag{4.40}$$

with Jacobian matrix

$$\frac{\partial g_{13}\left(\tau,\lambda,\beta,u_4,u_5,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial u_4\,\partial u_5\,\partial u_1\,\partial u_2\,\partial u_3} = \begin{bmatrix} \frac{e^{-u_1}}{(e^{-u_1}\tau-\tau-e^{-u_1})^2} & 0 & 0 & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{(e^{-u_1}\tau-\tau-e^{-u_1})^2} & 0 & 0\\ 0 & e^{u_2} & 0 & 0 & 0 & 0 & \lambda e^{u_2} & 0\\ 0 & 0 & e^{u_3} & 0 & 0 & 0 & 0 & \beta e^{u_3}\\ 0 & 0 & 0 & e^{u_4} & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & \frac{e^{-u_5}}{(1+e^{-u_5})^2} & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.41}$$

and, given the triangular properties of the matrix in equation 4.41, the magnitude of the Jacobian determinant is

$$\left|\frac{\partial g_{13}\left(\tau,\lambda,\beta,u_4,u_5,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial u_4\,\partial u_5\,\partial u_1\,\partial u_2\,\partial u_3}\right| = \frac{e^{u_2+u_3+u_4-u_1-u_5}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2\left(1+e^{-u_5}\right)^2} \tag{4.42}$$

Jumps from model 2 to model 1 are described by the function g_{21} where the set of parameters, $(\tau,\lambda,\beta,\kappa)$, and random variables, (u_1,u_2,u_3), are combined to form a set of new parameters, $(\tau^*,\lambda^*,\beta^*)$, where

$$\begin{aligned}\left(\tau^*,\lambda^*,\beta^*,u_4^*,u_1^*,u_2^*,u_3^*\right) &= g_{21}\left(\tau,\lambda,\beta,\kappa,u_1,u_2,u_3\right)\\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau+u_1\right],\lambda e^{u_2},\beta e^{u_3},\ln\kappa,-u_1,-u_2,-u_3\right)\end{aligned} \tag{4.43}$$

The Jacobian matrix for g_{21} is calculated as

$$\frac{\partial g_{21}\left(\tau,\lambda,\beta,\kappa,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial u_1\,\partial u_2\,\partial u_3} = \begin{bmatrix} \frac{e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 \\ 0 & e^{u_2} & 0 & 0 & 0 & \lambda e^{u_2} & 0 \\ 0 & 0 & e^{u_3} & 0 & 0 & 0 & \beta e^{u_3} \\ 0 & 0 & 0 & \frac{1}{\kappa} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.44}$$

where the magnitude of the determinant is given by

$$\left|\frac{\partial g_{21}\left(\tau,\lambda,\beta,\kappa,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial u_1\,\partial u_2\,\partial u_3}\right| = \frac{e^{u_2+u_3-u_1}}{\kappa\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} \tag{4.45}$$

Moves within model 2 are made by generating a random vector, (u_1,u_2,u_3,u_4), and applying the function g_{22} to the vector and the current parameter values $(\tau,\lambda,\beta,\kappa)$. Similarly to g_{11}, g_{22} must be involutary to satisfy the balance condition described in (Green, 1995). Here

$$\begin{aligned}(\tau^*,\lambda^*,\beta^*,\kappa^*,u_1^*,u_2^*,u_3^*,u_4^*) &= g_{22}\left(\tau,\lambda,\beta,\kappa,u_1,u_2,u_3,u_4\right) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau+u_1\right],\lambda e^{u_2},\beta e^{u_3},\kappa e^{u_4},-u_1,-u_2,-u_3,-u_4\right)\end{aligned} \tag{4.46}$$

with Jacobian matrix

$$\frac{\partial g_{22}\left(\tau,\lambda,\beta,\kappa,u_1,u_2,u_3,u_4\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial u_1\,\partial u_2\,\partial u_3\,\partial u_4} = \begin{bmatrix} \frac{e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 & 0 \\ 0 & e^{u_2} & 0 & 0 & 0 & \lambda e^{u_2} & 0 & 0 \\ 0 & 0 & e^{u_3} & 0 & 0 & 0 & \beta e^{u_3} & 0 \\ 0 & 0 & 0 & e^{u_4} & 0 & 0 & 0 & \kappa e^{u_4} \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.47}$$

The magnitude of the Jacobian determinant is as follows:

$$\left|\frac{\partial g_{22}\left(\tau,\lambda,\beta,\kappa,u_1,u_2,u_3,u_4\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial u_1\,\partial u_2\,\partial u_3\,\partial u_4}\right| = \frac{e^{u_2+u_3+u_4-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} \tag{4.48}$$

The jump from model 2 to model 3 is given by the bijection described in function g_{23}, where

$$\begin{aligned}(\tau^*,\lambda^*,\beta^*,\kappa^*,\phi^*,u_1^*,u_2^*,u_3^*,u_4^*) &= g_{23}\left(\tau,\lambda,\beta,\kappa,u_5,u_1,u_2,u_3,u_4\right)\\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau+u_1\right],\lambda e^{u_2},\beta e^{u_3},\kappa e^{u_4},\text{logit}^{-1}\,u_5,-u_1,-u_2,-u_3,-u_4\right)\end{aligned} \tag{4.49}$$

The Jacobian matrix for g_{23} is given by

$$\frac{\partial g_{23}\left(\tau,\lambda,\beta,\kappa,u_5,u_1,u_2,u_3,u_4\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial u_5\,\partial u_1\,\partial u_2\,\partial u_3\,\partial u_4} =$$

$$\begin{bmatrix}
\frac{e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 & 0\\
0 & e^{u_2} & 0 & 0 & 0 & 0 & \lambda e^{u_2} & 0 & 0\\
0 & 0 & e^{u_3} & 0 & 0 & 0 & 0 & \beta e^{u_3} & 0\\
0 & 0 & 0 & e^{u_4} & 0 & 0 & 0 & 0 & \kappa e^{u_4}\\
0 & 0 & 0 & 0 & \frac{e^{-u_5}}{\left(1+e^{-u_5}\right)^2} & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1
\end{bmatrix} \tag{4.50}$$

with determinant of magnitude

$$\left|\frac{\partial g_{23}\left(\tau,\lambda,\beta,\kappa,u_5,u_1,u_2,u_3,u_4\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial u_5\,\partial u_1\,\partial u_2\,\partial u_3\,\partial u_4}\right| = \frac{e^{u_2+u_3+u_4-u_1-u_5}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2\left(1+e^{-u_5}\right)^2} \tag{4.51}$$

Moves from the set of parameters used in model 3 to proposed parameter values in model 1 are described by the bijective function, g_{31}

$$\begin{aligned}(\tau^*,\lambda^*,\beta^*,u_4^*,u_5^*,u_1^*,u_2^*,u_3^*) &= g_{31}\left(\tau,\lambda,\beta,\kappa,\phi,u_1,u_2,u_3\right)\\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau+u_1\right],\lambda e^{u_2},\beta e^{u_3},\ln\kappa,\text{logit}\,\phi,-u_1,-u_2,-u_3\right)\end{aligned} \tag{4.52}$$

with Jacobian matrix

$$\frac{\partial g_{31}\left(\tau,\lambda,\beta,\kappa,\phi,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial\phi\,\partial u_1\,\partial u_2\,\partial u_3} =$$

$$\begin{bmatrix} \frac{e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 \\ 0 & e^{u_2} & 0 & 0 & 0 & 0 & \lambda e^{u_2} & 0 \\ 0 & 0 & e^{u_3} & 0 & 0 & 0 & 0 & \beta e^{u_3} \\ 0 & 0 & 0 & \frac{1}{\kappa} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\phi}+\frac{1}{1-\phi} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.53}$$

for which the magnitude of the determinant is

$$\left|\frac{\partial g_{31}\left(\tau,\lambda,\beta,\kappa,\phi,u_1,u_2,u_3\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial\phi\,\partial u_1\,\partial u_2\,\partial u_3}\right| = \left(\frac{1}{\phi}+\frac{1}{1-\phi}\right)\frac{e^{u_3+u_2-u_1}}{\kappa\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} \tag{4.54}$$

Jumps from model 3 to model 2 are defined by the bijection, g_{32}, where

$$\begin{aligned}\left(\tau^*,\lambda^*,\beta^*,\kappa^*,u_5^*,u_1^*,u_2^*,u_3^*,u_4^*\right) &= g_{32}\left(\tau,\lambda,\beta,\kappa,\phi,u_1,u_2,u_3,u_4\right)\\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau+u_1\right],\lambda e^{u_2},\beta e^{u_3},\kappa e^{u_4},\text{logit}\,\phi,-u_1,-u_2,-u_3,-u_4\right)\end{aligned} \tag{4.55}$$

The function g_{32} has a Jacobian matrix given by

$$\frac{\partial g_{32}\left(\tau,\lambda,\beta,\kappa,\phi,u_1,u_2,u_3,u_4\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial\phi\,\partial u_1\,\partial u_2\,\partial u_3\,\partial u_4} =$$

$$\begin{bmatrix} \frac{e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2} & 0 & 0 & 0 \\ 0 & e^{u_2} & 0 & 0 & 0 & 0 & \lambda e^{u_2} & 0 & 0 \\ 0 & 0 & e^{u_3} & 0 & 0 & 0 & 0 & \beta e^{u_3} & 0 \\ 0 & 0 & 0 & e^{u_4} & 0 & 0 & 0 & 0 & \kappa e^{u_4} \\ 0 & 0 & 0 & 0 & \frac{1}{\phi}+\frac{1}{1-\phi} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.56}$$

with a Jacobian determinant of magnitude

$$\left| \frac{\partial g_{32}\left(\tau, \lambda, \beta, \kappa, \phi, u_1, u_2, u_3, u_4\right)}{\partial \tau \, \partial \lambda \, \partial \beta \, \partial \kappa \, \partial \phi \, \partial u_1 \, \partial u_2 \, \partial u_3 \, \partial u_4} \right| = \frac{e^{u_2+u_3+u_4-u_1}}{\left(e^{-u_1}\tau - \tau - e^{-u_1}\right)^2} \left(\frac{1}{\phi} + \frac{1}{1-\phi} \right) \tag{4.57}$$

Finally, jumps within model 3 are given by the involutary function, g_{33}, where

$$\begin{aligned}(\tau^*, \lambda^*, \beta^*, \kappa^*, \phi^*, u_1^*, u_2^*, u_3^*, u_4^*, u_5^*) &= g_{33}\left(\tau, \lambda, \beta, \kappa, \phi, u_1, u_2, u_3, u_4, u_5\right) \\ &= \left(\text{logit}^{-1}\left[\text{logit}\,\tau + u_1\right], \lambda e^{u_2}, \beta e^{u_3}, \kappa e^{u_4}, \text{logit}^{-1}\left[\text{logit}\,\phi + u_5\right], -u_1, -u_2, -u_3, -u_4, -u_5\right)\end{aligned} \tag{4.58}$$

As before, we derive the Jacobian matrix

$$\frac{\partial g_{33}\left(\tau^*, \lambda^*, \beta^*, \kappa^*, \phi^*, u_1^*, u_2^*, u_3^*, u_4^*, u_5^*\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial\phi\,\partial u_1\,\partial u_2\,\partial u_3\,\partial u_4\,\partial u_5} = \begin{bmatrix} \frac{e^{-u_1}}{(e^{-u_1}\tau-\tau-e^{-u_1})^2} & 0 & 0 & 0 & 0 & \frac{\tau(1-\tau)e^{-u_1}}{(e^{-u_1}\tau-\tau-e^{-u_1})^2} & 0 & 0 & 0 & 0 \\ 0 & e^{u_2} & 0 & 0 & 0 & 0 & \lambda e^{u_2} & 0 & 0 & 0 \\ 0 & 0 & e^{u_3} & 0 & 0 & 0 & 0 & \beta e^{u_3} & 0 & 0 \\ 0 & 0 & 0 & e^{u_4} & 0 & 0 & 0 & 0 & \kappa e^{u_4} & 0 \\ 0 & 0 & 0 & 0 & \frac{e^{-u_5}}{(e^{-u_5}\phi-\phi-e^{-u_5})^2} & 0 & 0 & 0 & 0 & \frac{\phi(1-\phi)e^{-u_5}}{(e^{-u_5}\phi-\phi-e^{-u_5})^2} \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{4.59}$$

and, given the triangular nature of the matrix in equation 4.59, use the diagonal elements to calculate the magnitude of the determinant

$$\left|\frac{\partial g_{33}\left(\tau^*, \lambda^*, \beta^*, \kappa^*, \phi^*, u_1^*, u_2^*, u_3^*, u_4^*, u_5^*\right)}{\partial\tau\,\partial\lambda\,\partial\beta\,\partial\kappa\,\partial\phi\,\partial u_1\,\partial u_2\,\partial u_3\,\partial u_4\,\partial u_5}\right| = \frac{e^{u_2+u_3+u_4-u_1-u_5}}{\left(e^{-u_1}\tau-\tau-e^{-u_1}\right)^2\left(e^{-u_5}\phi-\phi-e^{-u_5}\right)^2} \tag{4.60}$$

CHAPTER 5

On the approximation of continuous dispersal in discrete-space models

Summary

1. Models which represent space as a lattice have a critical function in theoretical and applied ecology. Despite their significance, there is a dearth of appropriate theoretical developments for the description of dispersal across such lattices.

2. We present a series of methods for approximating continuous dispersal in discrete landscapes (denoted as centroid-to-centroid, centroid-to-area, area-to-centroid and area-to-area dispersal). We describe how these methods can be extended to incorporate different conditions at the boundary of the simulation arena and a framework for approximating continuous dispersal between irregularly shaped patches.

3. Each approximation method was tested against a baseline of continuous Gaussian dispersal in a periodic simulation arena. The residence probabilities for an individual dispersing in each time step according to a Gaussian kernel across grids of three differing resolutions were calculated over a number of dispersal steps. In addition, the steady-state asymptotic properties for the transition matrices for each approximation method and cell resolution were calculated and compared against the uniform expectation under continuous dispersal.

4. All four methods described in this paper provide a reasonable approximation to the con-

tinuous baseline (< 0.03 absolute error in probability calculations) on landscapes with grid cells of length equal to the expected dispersal distance or finer but error increases as grid cells become progressively larger than the expected dispersal distance.

5. Each approximation method exhibits a different spatial pattern of approximation error. Centroid-to-centroid dispersal overestimates residence probabilities near the origin, resulting in decreased invasion rates relative to the baseline diffusion process. All other approximation methods underestimate residence probabilities near the origin and overestimate such probabilities in the peripheries, leading to an overestimation of invasion rates.

6. The asymptotic properties of centroid-to-centroid and area-to-centroid dispersal approximation methods deviate from that which is expected under continuous dispersal. This characteristic renders these methods unsuitable for use in long-term simulation studies where the equilibrium properties of the system are of interest.

7. Centroid-to-area and area-to-area approximation methods exhibit both low approximation error and desirable asymptotic properties however. These methods provide a viable mechanism for linking individual-level dispersal to larger scale characteristics such as metapopulation connectivity.

5.1 Introduction

The extension of ecological models into the spatially explicit realm presents one of the most rewarding but also one of the most challenging aspects of model development. Traditionally, ecological models have focused on describing interactions between individuals in terms of the mean density of individuals in a population. Models derived from this so-called 'mean field' assumption have provided many new insights in ecology but, without the inclusion of local interactions between individuals, the lack of spatial structure in these models can produce very different conclusions on crucial phenomena such as invasion speed and species coexistence than their spatially explicit counterparts (Ovaskainen & Cornell, 2006; Murrell, 2010).

Whilst the spatial element can, in some cases, represent a substantial leap in complexity, it can often elucidate the mechanisms of otherwise confusing observations. For example, the addition of spatial structure in models of predator-prey dynamics in Murrell (2005) and Kondoh (2003) have shown that equilibrium prey densities are negatively linked to the spatial covariance of the antagonists which can increase when prey fecundity is increased. This extension thus provides an alternative spatial explanation for the 'paradox of enrichment' of Rosenzweig (1971). Moreover, some core principles of the theory of competition, such as the assertion that a high ratio of intraspecific to interspecific competition provides community stability (appearing in many text books such as Putman & Wratten, 1984), have been shown to be incomplete when interrogated with models able to explicitly describe and simulate the spatial aggregation of conspecifics (Neuhauser & Pacala, 1999; Murrell, 2010). In applied ecology, spatially explicit models are also commonly used to represent the spatial arrangement of populations and dispersal of individuals, including in the context of reserve selection strategies and responses to climate change (for example Moilanen *et al.*, 2005; Willis *et al.*, 2009).

One of the crucial elements of a spatially explicit model is the specification on how this space is represented. Indeed, Murrell (2005) postulates that the one of reasons why the findings of Wilson *et al.* (1993) appear to contradict the demonstration in Murrell (2005) that increased prey movement reduces the equilibrium population size is that the study of Wilson *et al.* (1993) represents space as a discrete lattice of environments with each patch able to support a maximum of one individual. This type of stochastic cellular automaton is one commonly employed in ecological models (see Silvertown *et al.*, 1992; Jeltsch *et al.*, 1996; Mustin *et al.*, 2009, for more examples), although other variants where populations of more than one individual (as

implemented in Travis & Dytham, 2002), or communities of more than one species (as implemented in Travis *et al.*, 2005), can inhabit a single cell are also used.

Whilst lattice models have the potential to provide many novel ecological insights (Nakamaru, 2006), with some authors exalting these methods as a 'paradigm' (Hogeweg, 1988), their simplification of spatial structure can lead to a number of biases in the interpretation of their output. No more so is this bias shown so prominently than in the methodologies employed to model dispersal through these habitats. The most basic simplification of dispersal, often denoted 'stepping-stone' dispersal or sometimes 'nearest-neighbour' dispersal (Kimura & Weiss, 1964), defines movement as a local process where individuals can only move to adjacent lattice cells with some given probability, usually uniformly selected amongst the neighbourhood of cells (although see Topping *et al.*, 2003; Wiegand *et al.*, 2004, for other weighting methods). For rectangular lattices, different concepts of the neighbourhood are employed (see Milne *et al.*, 1996): 'Moore neighbourhoods' define the eight neighbouring cells in the horizontal, vertical, and diagonal directions as potential destinations for dispersing individuals (Topping *et al.*, 2003; Wiegand *et al.*, 2004, for example), whilst 'von Neumann' neighbourhoods consider only the four cardinally adjacent cells as potential destinations for dispersing individuals (Söndgerath & Schröder, 2002, for example). However, Holland *et al.* (2007) show that both neighbourhood definitions can exhibit unnatural artefacts, both in terms of the spatial densities observed when considering multiple realisations of such defined dispersal events and the maximum traversable distance after a set number of time steps.

In continuous space, the probability density function of dispersal distances of a motile individual (or propagule in sessile organisms) from the point of origin is often referred to as the distance distribution (Nathan & Muller-Landau, 2000), the circular distribution (Wilson, 1993), or the distance pdf (Cousens *et al.*, 2008). These distributions describe the probability of the magnitude of a movement event but not its direction. In a one-dimensional world, the distance distribution is the folded equivalent of a displacement distribution, where displacement also accounts for the direction of movement and can therefore be negative. We can extend these one-dimensional descriptions of displacement into the spatial domain by describing dispersal in terms of its polar coordinates from the point of origin. For models with descriptions of dispersal in continuous space there are a many number of distributions of spatial displacement available to the investigator (see Clark *et al.*, 1999; Cousens *et al.*, 2008). This is not the case for discrete lattice-based dispersal however; outside of simple stepping-stone models of dispersal there is a

dearth of appropriate models for the calculation of cell-to-cell dispersal probabilities. To avoid confusion, the term 'dispersal kernel' will hereafter refer to the probability density function of displacement and not the probability density function of dispersal distance.

To address some of the deficiencies of stepping-stone models of dispersal, Chesson & Lee (2005) describe a number of families of integer-valued displacement distributions for use in lattice models of arbitrary dimensionality. These distributions have the flexibility to allocate non-zero probabilities of dispersal to cells beyond the nearest neighbours, and hence can potentially provide a mechanism of dispersal not too dissimilar to their continuous counterparts. The distributions described in Chesson & Lee (2005) also exhibit a number of desirable qualities that make their development a significant step forward for incorporating more realistic dispersal in cell-based studies. Firstly, most of the distributions described in Chesson & Lee (2005) have functional forms that are closed under convolution. This means that when iterating the dispersal forward a number of time periods, total displacement is simply a re-parametrisation of the one-step displacement distribution. More generally, this means that we are able to parametrise the displacement distribution as a function of time. Secondly, each of the displacement kernels have a parameter controlling the kurtosis of the probability distribution and allowing flexibility in specification of the probability weight of the tails of the distribution. This is particularly useful for helping to include the effects of long distance dispersal that often requires a 'fat-tailed' displacement distribution (Hovestadt *et al.*, 2001; Petrovskii *et al.*, 2008). Finally, the displacement distributions of Chesson & Lee (2005) also exhibit asymptotic radial symmetry, which ameliorates some of the artefacts of lattice-based dispersal described by Holland *et al.* (2007).

Field data such as telemetry or seed shadow data are often used to parametrise continuous models of dispersal (see Greene *et al.*, 2004), but such data are rarely applied so explicitly in the parametrisation of lattice dispersal, nor are such data collected in such a way as to be applicable in these settings. Whilst Chesson & Lee (2005) provide models of lattice dispersal with desirable mathematical properties, the underlying theoretical basis of these models is the mixture of a random quantity of stepping-stone dispersal sub-stages, requiring that individuals disperse cardinally with respect to the artificial geometry placed upon them within each of these dispersal sub-stages. On a two dimensional grid, this means that although an individual can disperse further than the nearest neighbours the final dispersal of the entire time step is comprised of a number of stepping-stone dispersal sub-steps, with each dispersal sub-step lim-

ited to movement within a von Neumann neighbourhood. It is difficult to see the theoretical link between such models and those that are commonly fitted to dispersal data. We adopt here a different approach, and instead describe a general method for the approximation of continuous displacement distributions on lattices of arbitrary resolution. We use this methodology to derive approximate cell-to-cell transition probabilities for commonly employed models of continuous dispersal and describe how this method can be extended to allow for common boundary conditions and irregularly shaped source and destination patches.

For convenience, all notation used in this paper is summarised in table 5.1.

5.2 Materials and Methods

5.2.1 Calculating Transition Probabilities

We first begin by defining the two-dimensional displacement kernel $g_{\cdot}(r, \theta)$, which describes the probability density of a polar displacement of length r (where $r > 0$) at a bearing θ in a single dispersal event. There are a number of different ways to define the direction of dispersal, θ. One common method employed in the mathematical domain is to define θ as the angle of direction measured in an anti-clockwise direction from the x-axis such that $-\pi < \theta \leq \pi$. However, a measurement regime that is more intuitive to field biologists, and one that may be more consistent with the format of collected data, is to define the angle of dispersal as a clockwise bearing from the y-axis, with θ instead defined between the limits $0 \leq \theta < 2\pi$. For the sake of clarity we will adopt the notation θ_1 and θ_2 to refer to the former and latter definitions respectively. It is worth noting that θ_1 and θ_2 are linked by the relationship

$$\theta_1 = \frac{\pi}{2} - \theta_2 + 2\pi \left[1 - \mathbb{H}\left(\frac{3\pi}{2} - \theta_2\right)\right] \tag{5.1}$$

where $\mathbb{H}(x)$ is the step function

$$\mathbb{H}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

Grids are defined on a Cartesian coordinate system and so the polar displacement function must be converted to describe the probability density of displacement to a set of Cartesian destination coordinates, denoted here as $k = (k_x, k_y)$, given a set of source coordinates, $j =$

Table 5.1: Summary of notation used

	Description	*Support*
r	Dispersal distance	$r \in \mathbb{R}^+$
θ_1	Angle of dispersal measured in an anticlockwise direction from the positive x-axis	$\theta_1 \in \mathbb{R} \quad -\pi < \theta_1 \leq \pi$
θ_2	Angle of dispersal measured in a clockwise direction from the positive y-axis	$\theta_2 \in \mathbb{R} \quad 0 \leq \theta_2 < 2\pi$
$\mathbb{H}(x)$	Step function defined in equation 5.2	$\mathbb{H}(x) \in \{0, 1\}$
j	Potential source coordinates (j_x, j_y)	$j_x \in \mathbb{R} \quad j_y \in \mathbb{R}$
k	Potential destination coordinates (k_x, k_y)	$k_x \in \mathbb{R} \quad k_y \in \mathbb{R}$
$g.(r, \theta_2)$	Two-dimensional dispersal kernel described in terms of polar displacement	$g.(r, \theta_2) \in \mathbb{R}^{0+}$
$c.(j_x, j_y, k_x, k_y)$	Two-dimensional dispersal kernel described in terms of the source and destination Cartesian coordinates (see equation 5.6)	$c.(j_x, j_y, k_x, k_y) \in \mathbb{R}^{0+}$
$a_x \quad a_y$	The width and height of the simulation arena respectively	$a_x \in \mathbb{R}^+ \quad a_y \in \mathbb{R}^+$
J	Source cell bounded by j_{x_1} and j_{x_2} on the x-axis and j_{y_1} and j_{y_2} on the y-axis	$j_{x_1} \in \mathbb{R} \quad j_{x_2} \in \mathbb{R} \quad j_{x_1} < j_{x_2}$ $j_{y_1} \in \mathbb{R} \quad j_{y_2} \in \mathbb{R} \quad j_{y_1} < j_{y_2}$
K	Destination cell bounded by k_{x_1} and k_{x_2} on the x-axis and k_{y_1} and k_{y_2} on the y-axis	$k_{x_1} \in \mathbb{R} \quad k_{x_2} \in \mathbb{R} \quad k_{x_1} < k_{x_2}$ $k_{y_1} \in \mathbb{R} \quad k_{y_2} \in \mathbb{R} \quad k_{y_1} < k_{y_2}$
$K^{(i_1,i_2)}$	Translation of the destination cell bounded by $[k_{x_1} + i_1 a_x]$ and $[k_{x_2} + i_1 a_x]$ on the x-axis and by $[k_{y_1} + i_2 a_y]$ and $[k_{y_2} + i_2 a_y]$ on the y-axis	
$\emptyset$	A cell to denote the absorbing state: the area not covered by any of the cells within the simulation arena	
$p_{\cdot JK}^{(\cdot)}$	Probability of moving from cell J to cell K, calculated according to the approximation method in the superscript brackets (CC denotes centroid-to-centroid, AC area-to-centroid, CA centroid-to-area, and AA area-to-area dispersal; equations 5.5, 5.9, 5.7, and 5.8 respectively)	$p_{\cdot JK}^{(\cdot)} \in \mathbb{R} \quad 0 \leq p_{\cdot JK}^{(\cdot)} \leq 1$
$p'^{(\cdot)}_{\cdot JK}$	$p_{\cdot JK}^{(\cdot)}$ corrected for the incorporation of restricting boundary conditions (see equation 5.13)	$p'^{(\cdot)}_{\cdot JK} \in \mathbb{R} \quad 0 \leq p'^{(\cdot)}_{\cdot JK} \leq 1$
$p''^{(\cdot)}_{\cdot JK}$	$p_{\cdot JK}^{(\cdot)}$ corrected for the incorporation of periodic boundary conditions (see equations 5.14, 5.15, and 5.16)	$p''^{(\cdot)}_{\cdot JK} \in \mathbb{R} \quad 0 \leq p''^{(\cdot)}_{\cdot JK} \leq 1$
$\mathbf{J}$	A source patch consisting of $N_{\mathbf{J}}$ cells	
$\mathbf{K}$	A destination patch consisting of $N_{\mathbf{K}}$ cells	
$p_{\cdot \mathbf{JK}}^{(\cdot)}$	Probability of moving from patch $\mathbf{J}$ to patch $\mathbf{K}$ calculated using the underlying cell transition probabilities, $p_{\cdot JK}^{(\cdot)}$, according to equations 5.20 and 5.21	$p_{\cdot \mathbf{JK}}^{(\cdot)} \in \mathbb{R} \quad 0 \leq p_{\cdot \mathbf{JK}}^{(\cdot)} \leq 1$
$\mathbf{P}''^{(\cdot)}$	A transition matrix with each element, $p''^{(\cdot)}_{\cdot JK}$, containing the probability of moving to cell K from cell J with periodic boundary correction applied	
$\mathbf{M}_t^{(\cdot)}$	A vector with each element, $m_{t_J}^{(\cdot)}$, containing the probability that an individual resides within cell J at time t according to the relevant approximation method	$m_{t_J}^{(\cdot)} \in \mathbb{R} \quad 0 \leq m_{t_J}^{(\cdot)} \leq 1$
w_{t_J}	Probability that an individual resides within cell J at time t under a continuous Gaussian diffusion process (see equation 5.26)	$w_{t_J} \in \mathbb{R} \quad 0 \leq w_{t_J} \leq 1$
w''_{t_J}	w_{t_J} with periodic boundary correction applied (see equation 5.27)	$w''_{t_J} \in \mathbb{R} \quad 0 \leq w''_{t_J} \leq 1$

(j_x, j_y). We can rewrite r and θ_2 in terms of these coordinates

$$r = \sqrt{(k_x - j_x)^2 + (k_y - j_y)^2} \tag{5.3}$$

$$\theta_2 = \frac{\pi}{2} - \arctan\left(\frac{k_y - j_y}{k_x - j_x}\right) + \pi\mathbb{H}(j_x - k_x) \tag{5.4}$$

The derivation for r describes the standard magnitude of dispersal distance in a Euclidean two-dimensional coordinate system. However, the formula for θ_2 differs from the standard polar conversion formula as it incorporates both a correction factor to match our definition of θ_2 and also an extra term to make the equation valid regardless of which quadrant the destination coordinate, k, occupies in relation to the source coordinate j.

Centroid-to-Centroid dispersal

The simplest method to approximate a continuous displacement kernel on a lattice is to set the cell-to-cell transition probabilities using the displacement kernel density for the distance from the centroid of the source patch, J, to the centroid of the destination patch, K (one version of the dispersal mechanism implemented in Moilanen, 2004). For this quantity to represent a true probability however, it is necessary to normalise these values by dividing over the sum of the probability densities of the displacement kernel evaluated at the centroids of all candidate dispersal locations. If we denote the centroid-to-centroid transition probability from cell J, bounded between j_{x_1} and j_{x_2} on the x-axis and j_{y_1} and j_{y_2} on the y-axis (where $j_{x_1} < j_{x_2}$ and $j_{y_1} < j_{y_2}$), to cell K, similarly bounded between k_{x_1} and k_{x_2} on the x-axis and k_{y_1} and k_{y_2} on the y-axis, as $p_{JK}^{(CC)}$, then

$$p_{JK}^{(CC)} = \frac{c.\left(\frac{j_{x_2}+j_{x_1}}{2}, \frac{j_{y_2}+j_{y_1}}{2}, \frac{k_{x_2}+k_{x_1}}{2}, \frac{k_{y_2}+k_{y_1}}{2}\right)}{\sum_L c.\left(\frac{j_{x_2}+j_{x_1}}{2}, \frac{j_{y_2}+j_{y_1}}{2}, \frac{l_{x_2}+l_{x_1}}{2}, \frac{l_{y_2}+l_{y_1}}{2}\right)} \tag{5.5}$$

where $c.(j_x, j_y, k_x, k_y)$ is a reparametrisation of the displacement kernel

$$c.(j_x, j_y, k_x, k_y) = g.\begin{pmatrix} r = \sqrt{(k_x - j_x)^2 + (k_y - j_y)^2}, \\ \theta_2 = \frac{\pi}{2} - \arctan\left(\frac{k_y - j_y}{k_x - j_x}\right) + \pi\mathbb{H}(j_x - k_x) \end{pmatrix} \tag{5.6}$$

and L is a candidate destination cell bounded between l_{x_1} and l_{x_2} on the x-axis and l_{y_1} and l_{y_2} on the y-axis.

Centroid-to-Area Dispersal

An alternative derivation of cell transition probabilities is the centroid-to-area definition, with the probability of moving from cell J to cell K denoted here by $p_{\cdot JK}^{(CA)}$. Under this definition, the transition probabilities are defined by the probability that the dispersing individual lands somewhere within the area of the target cell such that

$$p_{\cdot JK}^{(CA)} = \int_{k_{y_1}}^{k_{y_2}} \int_{k_{x_1}}^{k_{x_2}} c_{\cdot} \left(\frac{j_{x_2} + j_{x_1}}{2}, \frac{j_{y_2} + j_{y_1}}{2}, k_x, k_y \right) dk_x dk_y \tag{5.7}$$

Unlike centroid-to-centroid dispersal, centroid-to-area dispersal allows the correct treatment of destination patches that are of different sizes. This is comparable to the models of dispersal implemented in studies such as Hanski *et al.* (2000) and Chapman *et al.* (2007) that weight the dispersal probabilities to destination patches according to area.

Area-to-Area Dispersal

Both centroid-to-centroid dispersal and centroid-to-area dispersal can suffer from severe biases when the cell size is large relative to the expected dispersal distance (Collingham *et al.*, 1996). Under such circumstances, the dispersal distance may need to be improbably large for an individual to move from the centroid to the edge of a source cell, resulting in close to zero probability weights for all possible non-source destination cells. Iterating such models forward a number of time steps can produce a gross underestimation of invasion rates compared to a continuous model counterpart. We can remedy some of these effects by allowing dispersal to originate from alternative points from within the cell. One method, such as that employed in one specification of the SPOMSIM model of Moilanen (2004), describes dispersal in terms of the distance of the nearest edges between patches. Another method, and the one that we will describe here, assumes that dispersal is equally likely from all possible locations from within the cell. Here the locations of individuals are represented as a uniform probability distribution bounded by the spatial boundary coordinates of the cell. The probability of any dispersal event occurring between the source coordinates, j, and destination coordinates, k, is then simply the product of the probability of the origin of the dispersal event, $\mathbb{P}(j) = \frac{1}{(j_{x_2} - j_{x_1})(j_{y_2} - j_{y_1})}$, and the probability of dispersing to the destination given that origin, $\mathbb{P}(k|j) = c_{\cdot}(j_x, j_y, k_x, k_y)$. The transition probability from cell J to cell K, that we denote 'area-to-area' dispersal and by the notation $p_{\cdot JK}^{(AA)}$, requires that we integrate over all possible source coordinates within the boundaries of the source cell and all possible destination coordinates within the boundaries of

the destination cell such that

$$p_{JK}^{(AA)} = \frac{1}{(j_{x_2} - j_{x_1})(j_{y_2} - j_{y_1})} \int_{k_{y_1}}^{k_{y_2}} \int_{k_{x_1}}^{k_{x_2}} \int_{j_{y_1}}^{j_{y_2}} \int_{j_{x_1}}^{j_{x_2}} c.(j_x, j_y, k_x, k_y)\, dj_x dj_y dk_x dk_y \tag{5.8}$$

Area-to-Aentroid Dispersal

One final method for the derivation of transition probabilities on a lattice is area-to-centroid dispersal, $p_{JK}^{(AC)}$. This method is less applicable for use in cell-based dispersal but is included here for the sake of completeness. In a similar manner to the area-to-area dispersal approximation method described above, this definition requires the spatial integration over all possible source coordinates except, that in this case, the destination coordinates are fixed at the centre of the destination cell. However, like centroid-to-centroid dispersal, the final probability requires normalisation such that

$$p_{JK}^{(AC)} = \frac{\int_{j_{y_1}}^{j_{y_2}} \int_{j_{x_1}}^{j_{x_2}} c.\left(j_x, j_y, \frac{k_{x_2} + k_{x_1}}{2}, \frac{k_{y_2} + k_{y_1}}{2}\right) dj_x dj_y}{\sum_L \int_{j_{y_1}}^{j_{y_2}} \int_{j_{x_1}}^{j_{x_2}} c.\left(j_x, j_y, \frac{l_{x_2} + l_{x_1}}{2}, \frac{l_{y_2} + l_{y_1}}{2}\right) dj_x dj_y} \tag{5.9}$$

Figure 5.1 illustrates the four different definitions used in this paper to approximate continuous dispersal when generating lattice-based cell-to-cell transition probabilities.

Appendix A includes a detailed derivation of transition probability estimates under Gaussian dispersal (see Clark *et al.*, 1999) for each of the four approximation methods described in this paper. It may not be easy to derive results analytically for other dispersal kernels however and, in these circumstances, it may be necessary to resort to the application of numerical integration techniques (see Davis & Rabinowitz, 2007) to derive transition probability estimates. Functions to perform any of the approximation methods described in this paper have been provided for the R statistical computing platform as part of the `ecomodtools` package available from RForge (`https://r-forge.r-project.org/projects/ecomodtools/`). The `LatticeTransitionProbs` function of the `ecomodtools` package can be employed to calculate cell-to-cell transition probabilities using the analytic results of commonly employed dispersal kernels, or, by using Monte Carlo integration for an arbitrary, user-defined, dispersal kernel. To install the package and the relevant documentation from R simply type the following at the console whilst connected to the internet: `install.packages("ecomodtools", repos="http://R-Forge.R-project.org")`.

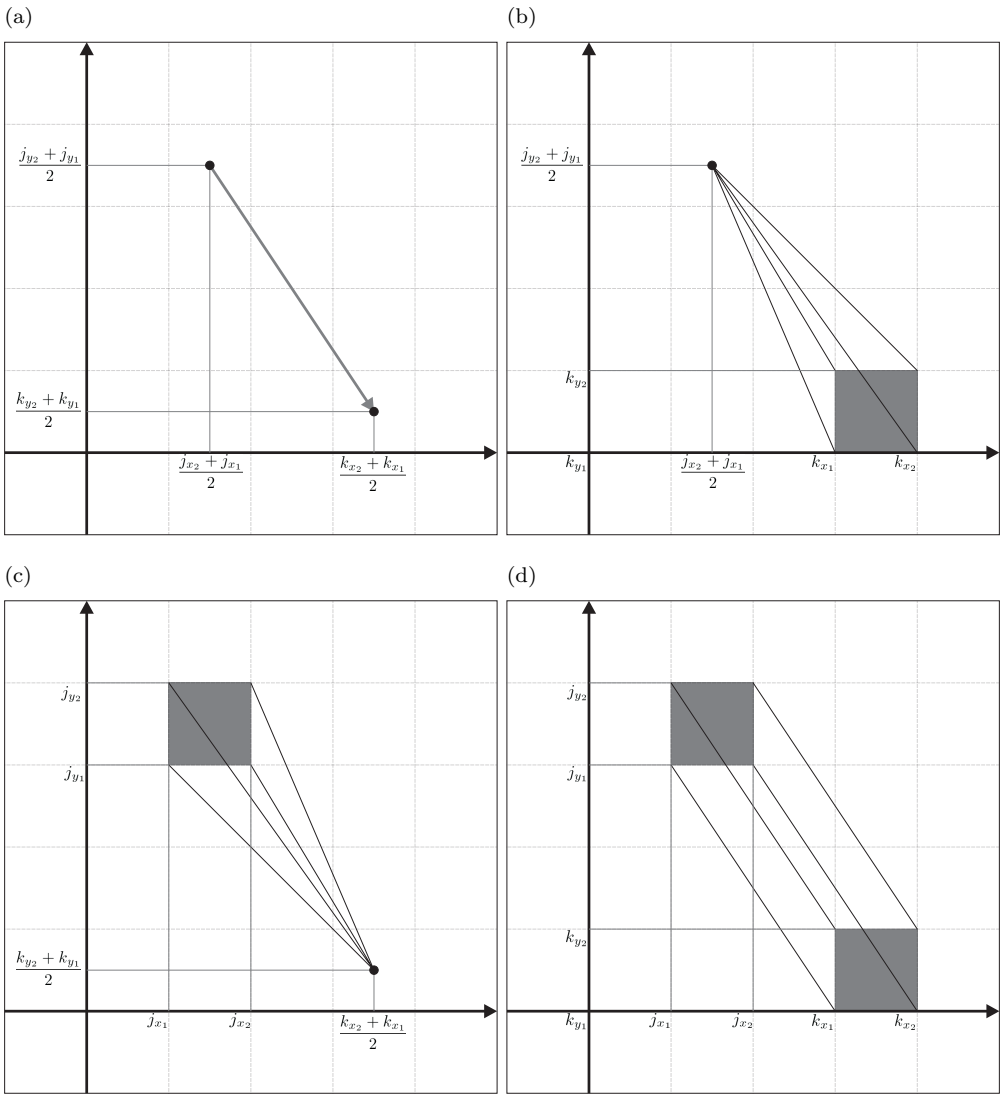


Figure 5.1: Illustration of the four different lattice-based dispersal transition probability definitions described in this paper. Figure (a) illustrates centroid-to-centroid dispersal, $p_{\cdot JK}^{(CC)}$, where dispersal events are assumed to originate from the centre of the cell and dispersing individuals can only disperse to the centroids of the possible destination cells. Centroid-to-area dispersal, $p_{\cdot JK}^{(CA)}$, as depicted in figure (b), shows how the transition probability is defined as the probability of landing anywhere within the boundaries of the destination cell but with all dispersal originating from the centre of the source cell. Area-to-centroid dispersal, figure (c) and $p_{\cdot JK}^{(AC)}$, allows weights the dispersal probabilities of arriving at the centroid of the destination cell given the point of origin of the dispersing individual by the probability that the individual begins its dispersal from that origin. This is assumed to be uniform over the area of the cell. Area-to-area dispersal, $p_{\cdot JK}^{(AA)}$ in figure (d), extends area-to-centroid dispersal by integrating over all possible destination points in the destination cell and relaxing the restriction that individuals can only disperse to centroids. Note that $p_{\cdot JK}^{(CC)}$ and $p_{\cdot JK}^{(AC)}$ require normalisation to represent true transition probabilities.

5.2.2 Composite Dispersal Kernels

Some authors have argued that one dispersal kernel alone does not offer enough flexibility to describe the observed changes in species distributions and that a composite dispersal kernel combining the different modes of dispersal at short and long ranges is preferable (Shigesada *et al.*, 1995; Higgins & Richardson, 1999; Bullock & Clarke, 2000). The commonest form for a composite displacement kernel, denoted here as $g_+(r,\theta)$, usually consists of two sub-kernels, $g_1(r,\theta)$ and $g_2(r,\theta)$, weighted by an extra parameter, ϕ:

$$g_+(r,\theta) = \phi g_1(r,\theta) + (1-\phi)\, g_2(r,\theta) \tag{5.10}$$

Under this specification, each dispersal event involves the drawing of a random distance and direction from a joint distribution described by the probability density function $g_1(r,\theta)$ with probability ϕ (where $0 \leq \phi \leq 1$), otherwise the distance and direction are drawn according to a random number with joint probability density function $g_2(r,\theta)$. This allows the specification of a kernel that describes a common localised dispersal pattern, with a high value of ϕ, but with the possibility of very rare but long distance dispersal events. This formulation of composite dispersal can be included into our cell-to-cell transition probabilities under a lattice-based modelling structure very simply by weighting the transition probabilities corresponding to the composite kernels according to the weighting parameter ϕ such that

$$p^{(\cdot)}_{+_{JK}} = \phi p^{(\cdot)}_{1_{JK}} + (1-\phi)\, p^{(\cdot)}_{2_{JK}} \tag{5.11}$$

where $p^{(\cdot)}_{1_{JK}}$ and $p^{(\cdot)}_{2_{JK}}$ represent the transition probabilities, calculated using any of the four approximation methods described above, of the dispersal described by probability density functions $g_1(r,\theta)$ and $g_2(r,\theta)$ respectively.

5.2.3 Incorporating Boundary Conditions

The methods described here assume that the lattice is infinite in both dimensions. However, for all practical purposes, lattice models must be run over a finite grid. For any model running in a finite space, decisions must be made as to what happens for individuals dispersing outside the boundaries of the model. Research into the effects of boundary conditions is often considered a rather esoteric subject but assumptions regarding individuals at the edge of the simulation arena can exert a dramatic influence on the analysis of the model outputs (Sullivan, 1988; Burton & Travis, 2008). No appraisal of methods for the approximation of continuous dispersal

would therefore be complete without an explanation on how these methods can be adapted to include boundary conditions.

Absorbing Boundary Conditions

One commonly applied boundary condition is the so-called 'absorbing condition'. Here individuals that disperse outside of the simulation arena are removed from the simulation: they are, in effect, killed, although some implementations take care to differentiate between the edge-enforced mortality and standard mortality when describing the results of the model analysis. This boundary condition, whilst easy to implement, has the unfortunate side-effect that individuals in cells close to the border suffer from inflated mortality as it is from these cells that dispersing individuals are more likely to cross the arena threshold. This may not be biologically realistic if, outside the area of study, there exists a means of survival for individuals that disperse outside this region and that it is possible for these individuals to exert some influence over individuals within the study region either through processes such as competition or through the production of offspring immigrating back into the dispersal arena.

Implementing absorbing boundary conditions in the calculation of cell-to-cell dispersal probabilities requires the definition of an extra 'absorbed cell' that contains all individuals that have moved outside of the simulation arena. Once present in the absorbed state, individuals are unable to leave; the probability of moving from the absorbed state to cell K, $p^{(\cdot)}_{\emptyset K} = 0$ and the probability of remaining in the absorbed state, $p^{(\cdot)}_{\emptyset\emptyset} = 1$. By definition, the absorbed state encapsulates any space not defined inside the simulation arena, so the probability of entering the absorbed state from cell J, $p^{(\cdot)}_{J\emptyset}$ is

$$p^{(\cdot)}_{J\emptyset} = 1 - \sum_{L} p^{(\cdot)}_{JL} \tag{5.12}$$

where L is any non-absorbing candidate destination cell inside the simulation arena.

The calculation of centroid-to-centroid and area-to-centroid transition probabilities (equations 5.5 and 5.9 respectively) requires normalisation. This normalisation constant is defined as the sum of the continuous displacement probability density function evaluated at the mid-points of the set of candidate destination cells. However, the absorbing state has no defined mid-point and, as such, it is not easy to implement centroid-to-centroid and area-to-centroid dispersal using absorbing boundary conditions. In principle it may be possible to estimate the weight

of dispersal probability by numerically evaluating an infinite series of cells lying outside of the simulation area and then normalising these probabilities. In practise however, this method may be heavily dependent on the resolution of the extra-arena cells and it is not clear how this method would be applied in situations where the cells inside the simulation arena are not of equal size.

Reflecting and Restricting Boundary Conditions

'Reflecting' and 'restricting' boundary conditions are both methods that ensure that individuals do not leave the simulation arena. Under reflecting boundary conditions, individuals that arrive at the boundaries are reflected back into the simulation area. This process is fairly simple to simulate in continuous space by employing algorithms to test for intersections of the path of movement with the limits of the study region (see O'Rourke, 1994) and correcting the coordinates in these cases according to simple reflection rules. These movement rules translate into quite complex changes in the functional form of the displacement kernel however. These functional changes are specific to the shape and extent of the study region and, for most simulation extents, it is not tractable to describe the net displacement after reflection correction in terms of the original displacement kernel and even less so for a discrete approximation in lattice simulations.

Restricting boundary conditions, where the displacement kernel is truncated so that the probability of movement outside the simulation arena is given a zero weighting, have a much simpler analytical representation. Here the probability of moving from cell J to cell K after a restricting boundary correction factor is applied, $p'^{(\cdot)}_{\cdot JK}$, is simply a normalisation of the standard transition probability

$$p'^{(\cdot)}_{\cdot JK} = \frac{p^{(\cdot)}_{\cdot JK}}{\sum_L p^{(\cdot)}_{\cdot JL}} \tag{5.13}$$

It is worth noting that because the probabilities in centroid-to-centroid and area-to-centroid dispersal (as defined in equations 5.5 and 5.9 respectively) are already normalised, that is $\sum_L p^{(CC)}_{\cdot JL} = 1$ and $\sum_L p^{(AC)}_{\cdot JL} = 1$, then $p'^{(CC)}_{\cdot JK} = p^{(CC)}_{\cdot JK}$ and $p'^{(AC)}_{\cdot JK} = p^{(AC)}_{\cdot JK}$.

Periodic Boundary Conditions

Another commonly applied method considering dynamics at the edge of rectangular simulation arenas is the implementation of periodic boundary conditions. Under these boundary conditions, individuals that exit the simulation arena arrive at the opposite edge. In effect,

the simulation arena becomes a torus and provides a mechanism to approximate an infinite landscape, although at the expense of possible periodicity effects (Hünenberger & McCammon, 1999; Pradeep & Hussain, 2004). In terms of the cell-to-cell transition probabilities described previously, this means that it is now possible to move to a destination cell K by either directly dispersing to it or by dispersing across a boundary and arriving at K indirectly. The updated probability for moving from cell J to cell K for either centroid-to-area or area-to-area dispersal, $p''^{\binom{CA}{AA}}_{\cdot JK}$, can then be defined as

$$\begin{aligned} p''^{\binom{CA}{AA}}_{\cdot JK} &= \sum_{i_1=-\infty}^{\infty} \sum_{i_2=-\infty}^{\infty} p_{\cdot JK(i_1,i_2)}^{\binom{CA}{AA}} \\ &= p_{\cdot JK}^{\binom{CA}{AA}} + \sum_{i_1=1}^{\infty} \left[\begin{array}{c} \left(\sum_{i_2=-i_1}^{i_1} p_{\cdot JK(i_1,i_2)}^{\binom{CA}{AA}} + p_{\cdot JK(-i_1,i_2)}^{\binom{CA}{AA}} + p_{\cdot JK(i_2,i_1)}^{\binom{CA}{AA}} + p_{\cdot JK(i_2,-i_1)}^{\binom{CA}{AA}} \right) \\ -p_{\cdot JK(i_1,i_1)}^{\binom{CA}{AA}} - p_{\cdot JK(-i_1,i_1)}^{\binom{CA}{AA}} - p_{\cdot JK(i_1,-i_1)}^{\binom{CA}{AA}} - p_{\cdot JK(-i_1,-i_1)}^{\binom{CA}{AA}} \end{array} \right] \end{aligned} \tag{5.14}$$

where $K^{(i_1,i_2)}$ is a translation of the cell K such that $K^{(i_1,i_2)}$ is bounded by $[k_{x_1} + a_x i_1]$ and $[k_{x_2} + a_x i_1]$ on the x-axis and by $[k_{y_1} + a_y i_2]$ and $[k_{y_2} + a_y i_2]$ on the y-axis. The values a_x and a_y represent the total width and height of the simulation arena respectively.

Conceptually toroidal boundary conditions can be considered equivalent to an infinite grid of virtual simulation arenas of the same size and shape of our study area arranged in a lattice with the main simulation arena set at the central point (see figure 5.2). The corrected probability of dispersal from cell J to cell K with periodic boundary conditions is, under this conceptual framework, the sum of the probabilities of moving from cell J to each corresponding cell K in each of these virtual simulation arenas. The rearrangement of equation 5.14 allows the expression of the corrected probabilities in terms of one infinite series: dispersal probabilities being successively added from an expanding area of virtual simulation arenas. When $i_1 = 1$, the simulation arenas being evaluated lie in the immediate Moore neighbourhood (diagonal and cardinal neighbours) around our focus simulation arena. As i_1 increases, extra virtual simulation arenas are evaluated in an expanding rectangle providing a better approximation to the periodic boundary corrected probability of dispersal from cell J to cell K.

Whilst an expression for periodic cell-to-cell dispersal probabilities does not exist in a closed form, it is possible to numerically approximate the infinite series that require evaluation in the application of equation 5.14. The requirement for any probability density to integrate to unity over the range of its support ensures that as the distance from the source location gets infinitely

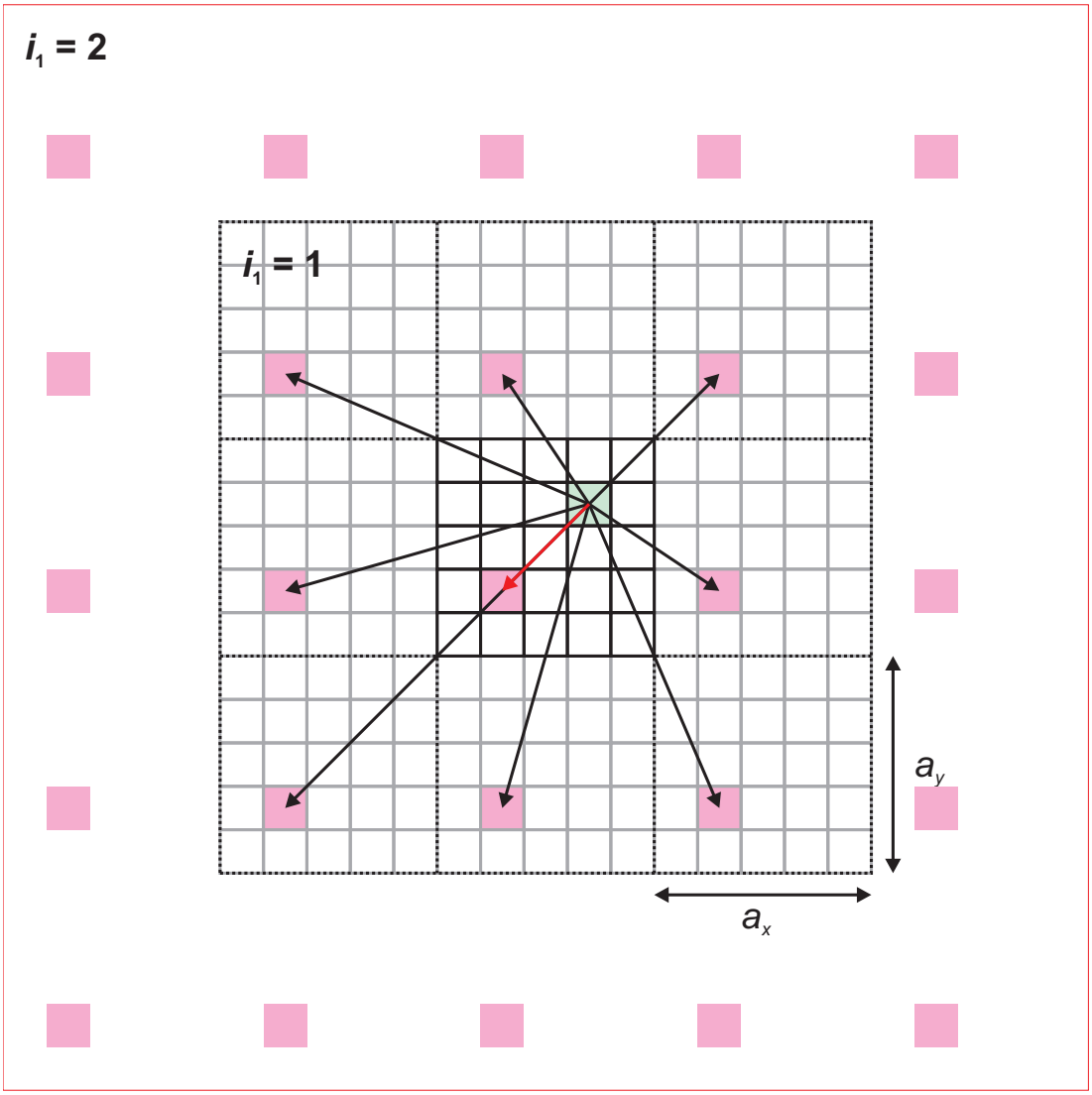


Figure 5.2: Illustration of the implementation of periodic boundary conditions when approximating continuous dispersal on discrete landscapes. The landscape is conceptually arranged as an infinitely sized lattice of simulation arenas centred around the main simulation arena (shown in a solid black lines in the centre of the figure) for which we wish to calculate dispersal. Dispersal from the green cell to the red destination cell within the main simulation arena can occur by direct dispersal (red arrow) or by arriving at the cell by crossing a boundary of the simulation arena and arriving at the cell from the opposite boundary. This is conceptually the same as adding all the probabilities of landing at the corresponding highlighted destination cells in each of virtual arenas (movement shown by black arrows). We approximate this infinite sum by adding the probabilities calculated from the virtual simulation arenas in an expanding rectangular region centred around the main simulation arena. The immediate Moore neighbourhood around the main simulation arena is shown by the dotted black lines (with individual cells shown in grey lines) and corresponds to the set of arenas considered when evaluating the first element of the infinite series ($i_1 = 1$). An outline of the neighbourhood evaluated for the second element of the infinite series ($i_1 = 2$) is also shown.

large, the probability density must tend to zero. As the terms of the infinite series represent ever increasing distances between the source and the translated destination cell then this also ensures the convergence of the infinite series. It is then possible to implement numerical techniques, such as those described in Caliceti *et al.* (2007), for the approximation of the series. So long as the mean dispersal distance is small relative to the size of the simulation arena, the convergence of the series will be rapid and only a handful of terms will need be evaluated to obtain a reasonable approximation.

We know that the sum of dispersal probabilities from cell J to all possible destination cells must equal one under periodic boundary conditions. The sum of direct (uncorrected) dispersal probabilities will not equal one however. This is because these probabilities do not take into account all the possible indirect dispersal events, where an individual travels over a border of the simulation arena and arrives on the destination square from the opposite edge. Upon algorithmic implementation it is useful to calculate the probabilities for all destination cells from a single source concurrently, recalculating all probabilities at every increment of i_1. Implementing the algorithm in this fashion provides a useful metric, equivalent to one minus the sum of the probabilities calculated at the current value of i_1, which is the total probability still 'unaccounted for' in indirect dispersal. A convenient stopping condition for the evaluation of the infinite series is provided when the 'unaccounted' probability is reduced to an acceptably low level. This stopping condition can be set in the `LatticeTransitionProbs` function of the `ecomodtools` R package through the `max.prob` parameter.

Implementing periodic boundary conditions for centroid-to-centroid and area-to-centroid approximation methods is a little more complex. Here, the normalisation step must be performed after the periodic summation step. The probabilities of arriving in cell K after leaving cell J using centroid-to-centroid and area-to-centroid dispersal after periodic boundary conditions have been applied, ${p''}^{(CC)}_{\cdot JK}$ and ${p''}^{(AC)}_{\cdot JK}$ respectively, are defined as

$${p''}^{(CC)}_{\cdot JK} = \frac{\upsilon^{(CC)}(J,K)}{\sum_{L} \upsilon^{(CC)}(J,L)} \tag{5.15}$$

$${p''}^{(AC)}_{\cdot JK} = \frac{\upsilon^{(AC)}(J,L)}{\sum_{L} \upsilon^{(AC)}(J,L)} \tag{5.16}$$

where

$$\upsilon^{\binom{CC}{AC}}(J,K) = \gamma^{\binom{CC}{AC}}_{JK^{(0,0)}} + \sum_{i_1=1}^{\infty}\left[\begin{array}{c}\left(\sum_{i_2=-i_1}^{i_1}\gamma^{\binom{CC}{AC}}_{JK^{(i_1,i_2)}} + \gamma^{\binom{CC}{AC}}_{JK^{(-i_1,i_2)}} + \gamma^{\binom{CC}{AC}}_{JK^{(i_2,i_1)}} + \gamma^{\binom{CC}{AC}}_{JK^{(i_2,-i_1)}}\right) \\ -\gamma^{\binom{CC}{AC}}_{JK^{(i_1,i_1)}} - \gamma^{\binom{CC}{AC}}_{JK^{(-i_1,i_1)}} - \gamma^{\binom{CC}{AC}}_{JK^{(i_1,-i_1)}} - \gamma^{\binom{CC}{AC}}_{JK^{(-i_1,-i_1)}}\end{array}\right] \tag{5.17}$$

and

$$\gamma^{(CC)}_{JK^{(i_1,i_2)}} = c.\left(\frac{j_{x_2}+j_{x_1}}{2}, \frac{j_{y_2}+j_{y_1}}{2}, \frac{k_{x_2}+k_{x_1}}{2} + i_1 a_x, \frac{k_{y_2}+k_{y_1}}{2} + i_2 a_y\right) \tag{5.18}$$

$$\gamma^{(AC)}_{JK^{(i_1,i_2)}} = \int_{j_{x_1}}^{j_{x_2}}\int_{j_{y_1}}^{j_{y_2}} c.\left(j_x, j_y, \frac{k_{x_2}+k_{x_1}}{2} + i_1 a_x, \frac{k_{y_2}+k_{y_1}}{2} + i_2 a_y\right) dj_y dj_x \tag{5.19}$$

Because the corrected transition probabilities require normalisation under centroid-to-centroid and area-to-centroid dispersal, it is not suitable to use the same stopping criteria in the evaluation of the infinite series of equation 5.17 in order to evaluate the infinite series of equation 5.14 for centroid-to-area and area-to-area dispersal. Instead we propose that, for a given value of i_1, further evaluation of the infinite series is stopped if the sum of the added probability for the current iteration across all possible destination cells is below a certain value. When centroid-to-centroid or area-to-centroid dispersal is selected, this stopping value can be controlled by the `max.prob` parameter of the `LatticeTransitionProbs` function of the `ecomodtools` R package. It is important to note that care must be taken when applying this stopping condition. Unlike centroid-to-area and area-to-area dispersal, it is not possible to evaluate the 'unaccounted for' probability. As a result, there is no way to guarantee that a suitable number of virtual simulation arenas have been interrogated to adequately cover the entire range of distances with meaningful non-zero density weights for a given dispersal kernel. This is particularly true for dispersal kernels with modal distances far from the origin of dispersal. These effects are ameliorated somewhat by choosing a large simulation arena relative to the dispersal capabilities of the species being studied.

5.2.4 Extension to Patch-Based Models

The methods described here are not limited to the description of cell-to-cell dispersal. Many metapopulation and metacommunity models use simplified descriptions of the spatial extent of patches. For example, Moilanen (2004) assumes that all patches are circular whilst Hanski (1994) assumes that patch shape is negligible in determining patch connectivity. Both studies assume that inter-patch dispersal probabilities need only be expressed in terms of the shortest distance between patches and the area of the patch. However, if the spatial extent of the

patches can be approximated using a set of cellular pixels then it is possible to bring the methods described here to bear and allow for the description of patch-to-patch dispersal in terms of an underlying continuous displacement kernel. This allows patch connectivity to be described both in terms of the dispersal capabilities of the species of interest and the spatial extent of the individual patches.

Whilst approximating the spatial extent of patches as a series of cells may seem like an abstraction, it rarely represents a loss of information. This is because the representation of habitat areas in spatial data sets, such as the CORINE land cover data set (as described in Brown *et al.*, 2002, for the UK extent), are often stored in a cellular 'raster' format anyway. Even data stored as areal units in 'vector' format can be approximated using fine resolution cellular lattices: the vector LANDCOVER 2000 data set of Fuller *et al.* (2002) is also available in a 25 metre resolution raster version.

If we define a patch, $\mathbf{J}$, as the set of $N_{\mathbf{J}}$ cells that comprise the source patch, with each constituent cell indexed $J_1, J_2, \ldots, J_{N_{\mathbf{J}}}$, and $\mathbf{K}$ as the destination patch of $N_{\mathbf{K}}$ similarly indexed cells, then we can calculate the probability of moving to patch $\mathbf{K}$ given that the source of dispersal originated somewhere within patch $\mathbf{J}$, $p^{(\cdot)}_{\cdot\mathbf{JK}}$, in terms of the component cell-to-cell dispersal probabilities. The event of a dispersing individual relocating to destination cell K_1 and the alternative event of that same individual relocating to any other cell, such as the cell K_2, during a single dispersal event are mutually exclusive. This means that the probability of dispersing to any of the destination cells in patch $\mathbf{K}$ given a specific source cell as the point of origin, $p^{(\cdot)}_{\cdot J_n\mathbf{K}}$, is the sum of the probabilities of dispersal from the source cell to each of the destination cells:

$$p^{(\cdot)}_{\cdot J_n\mathbf{K}} = \sum_{m=1}^{N_{\mathbf{K}}} p^{(\cdot)}_{\cdot J_n K_m} \tag{5.20}$$

If we assume that the probability of the location of the point of origin is uniformly spread across the area of the patch, the final patch-to-patch probability is defined as the sum of the probabilities of dispersing to any of the destination patch cells from each of the source patch cells, but with each probability weighted by the proportional area of the relevant source cell

relative to the total area of the source patch. Therefore,

$$\begin{aligned} p^{(\cdot)}_{\cdot\mathbf{JK}} &= \frac{\sum_{n=1}^{N_\mathbf{J}} \left(j_{n_{x_2}} - j_{n_{x_1}}\right)\left(j_{n_{y_2}} - j_{n_{y_1}}\right) p^{(\cdot)}_{\cdot J_n\mathbf{K}}}{\sum_{l=1}^{N_\mathbf{J}} \left(j_{l_{x_2}} - j_{l_{x_1}}\right)\left(j_{l_{y_2}} - j_{l_{y_1}}\right)} \\ &= \frac{\sum_{n=1}^{N_\mathbf{J}} \left(j_{n_{x_2}} - j_{n_{x_1}}\right)\left(j_{n_{y_2}} - j_{n_{y_1}}\right) \sum_{m=1}^{N_\mathbf{K}} p^{(\cdot)}_{\cdot J_n K_m}}{\sum_{l=1}^{N_\mathbf{J}} \left(j_{l_{x_2}} - j_{l_{x_1}}\right)\left(j_{l_{y_2}} - j_{l_{y_1}}\right)} \end{aligned} \tag{5.21}$$

where $j_{n_{x_1}}$ and $j_{n_{x_2}}$ are the lower and upper boundaries on the x-axis of cell J_n respectively. $j_{n_{y_1}}$ and $j_{n_{y_2}}$ are similarly defined as the lower and upper boundaries of cell J_n on the y-axis. For a regular lattice of cells, where all cells have the same area, equation 5.21 simplifies to $p^{(\cdot)}_{\cdot\mathbf{JK}} = \frac{1}{N_\mathbf{J}} \sum_{n=1}^{N_\mathbf{J}} \sum_{m=1}^{N_\mathbf{K}} p^{(\cdot)}_{\cdot J_n K_m}$.

To satisfy the condition $\sum_\mathbf{K} p_{\cdot\mathbf{JK}} = 1$, a requirement for a properly defined probability mass function, it is necessary that the patches collectively account for all space over which it is possible for individuals to disperse to. While this may be reasonable when deriving movement probabilities for individuals dispersing over landscapes with no gaps, such as the coarse-grained Dirichlet landscapes of Holland *et al.* (2007), this may be unsuitable for application in most metapopulation models where the total area of the patches combined can account for only a very small proportion of the total area of study. In these situations it is important to describe explicitly the fate of individuals that do not disperse successfully to another habitat patch. At one extreme, we can define all patches that are not of suitable habitat as absorbing states and apply the previously described absorbing state correction to the cell-to-cell transition probabilities (and hence to the patch-to-patch transition probabilities). However, for landscapes with patches that are relatively small compared to the total area of study or that appear infrequently, this dispersal-mediated mortality may represent a sizeable mortality risk. For seed dispersal, the displacement to unsuitable soil or environmental conditions may well doom that individual, but in animal dispersal, the description of a 'black hole' effect between patches may present an artificial inflation of dispersal mortality risk. At the other extreme, it is possible to apply restricting boundary conditions to the set of patches so that an individual always successfully disperses to a suitable patch. This in effect truncates the dispersal kernel so that only suitable patches can be dispersed to. Under these conditions, dispersal mortality is always zero, even in very isolated patches. In common application however, it may be most practicable to mix these two extreme scenarios using a method such as the dispersal mixture formula presented in equation 5.11.

5.2.5 Testing the Approximation

To assess the accuracy of the four approximation methods described in this paper, we describe the movement of an individual across the landscape over multiple time periods using each approximation method and compare the probability of the individual residing in each cell over each time period with what would be expected if continuous point to point dispersal was employed.

We define a transition matrix, $\mathbf{P''}_{\cdot}^{(\cdot)}$, as a comprehensive description of cell-to-cell dispersal probabilities and with each element, $p''^{(\cdot)}_{\cdot JK}$, containing the probability of moving to cell K if the dispersal event originated from cell J with periodic boundary correction applied. Here the matrices $\mathbf{P''}_{\cdot}^{(CC)}$, $\mathbf{P''}_{\cdot}^{(CA)}$, $\mathbf{P''}_{\cdot}^{(AC)}$, and $\mathbf{P''}_{\cdot}^{(AA)}$ are defined as dispersal matrices filled with transition probabilities derived using the relevant approximation method. The state-vector, $\mathbf{M}_t^{(\cdot)}$, with elements $m_{t_J}^{(\cdot)}$ contains the probabilities that the individual resides in cell J at time period t. This specification allows the use of $\mathbf{P''}_{\cdot}^{(\cdot)}$ to describe $\mathbf{M}_t^{(\cdot)}$ in terms of a Markov chain recurrence relationship where

$$\mathbf{M}_t^{(\cdot)} = \mathbf{M}_{t-1}^{(\cdot)} \mathbf{P''}_{\cdot}^{(\cdot)} \tag{5.22}$$

For the purposes of this exercise we approximate Gaussian dispersal on a lattice by filling the cell-to-cell transition probabilities in matrix $\mathbf{P''}_{\cdot}^{(\cdot)}$ with those calculated using equations 5.30, 5.35, 5.34, and 5.38 derived in appendix A for centroid-to-centroid, area-to-centroid, centroid-to-area, and area-to-area approximation methods respectively.

In order to compare the discrete approximations to continuous dispersal it is necessary to derive a cell-based description of residence probability based on continuous dispersal over time. Starting with the Cartesian representation of the Gaussian dispersal kernel as defined in Clark *et al.* (1999), we have shown in appendix B that the total displacement in the Cartesian coordinates, δ_x and δ_y, at time t, arising from Gaussian steps in each time period, is a bivariate-normal random variable with probability density function

$$s_t\left(\delta_x, \delta_y | \alpha\right) = \frac{1}{\pi t \alpha^2} e^{-\frac{\delta_x^2}{t\alpha^2}} e^{-\frac{\delta_y^2}{t\alpha^2}} \tag{5.23}$$

α represents the isotropic standard deviation parameter of displacement in one time step. From equation 5.23 it is possible to derive the probability that an individual resides in cell J at time

t, or w_{t_J}, by integrating the probability density function between the limits of the cell extent:

$$\begin{aligned} w_{t_J} &= \int_{j_{y_1}}^{j_{y_2}} \int_{j_{x_1}}^{j_{x_2}} s_t\left(\delta_x, \delta_y | \alpha\right)\, d\delta_x \delta_y \\ &= \frac{1}{\pi t \alpha^2} \left[\int_{j_{y_1}}^{j_{y_2}} e^{-\frac{\delta_y^2}{t\alpha^2}}\, d\delta_y\right] \left[\int_{j_{x_1}}^{j_{x_2}} e^{-\frac{\delta_x^2}{t\alpha^2}}\, d\delta_x\right] \end{aligned} \tag{5.24}$$

From the identity

$$\begin{aligned} \int_{j_{\cdot_1}}^{j_{\cdot_2}} e^{-\frac{\delta_\cdot^2}{t\alpha^2}}\, d\delta_\cdot &= \sqrt{t}\alpha \int_{\frac{j_{\cdot_1}}{\sqrt{t}\alpha}}^{\frac{j_{\cdot_2}}{\sqrt{t}\alpha}} e^{-\kappa^2}\, d\kappa \\ &= \frac{\sqrt{\pi t}\alpha}{2} \left[\operatorname{erf}\left(\frac{j_{\cdot_2}}{\sqrt{t}\alpha}\right) - \operatorname{erf}\left(\frac{j_{\cdot_1}}{\sqrt{t}\alpha}\right)\right] \end{aligned} \tag{5.25}$$

where κ is a substitution used in integration ($\kappa = \frac{\delta_\cdot}{\sqrt{t}\alpha}$), we can express w_{t_J} in terms of the numerically tractable error function, $\operatorname{erf}(Z)$, as defined in equation 5.32:

$$w_{t_J} = \frac{1}{4} \left[\operatorname{erf}\left(\frac{j_{x_2}}{\sqrt{t}\alpha}\right) - \operatorname{erf}\left(\frac{j_{x_1}}{\sqrt{t}\alpha}\right)\right] \left[\operatorname{erf}\left(\frac{j_{y_2}}{\sqrt{t}\alpha}\right) - \operatorname{erf}\left(\frac{j_{y_1}}{\sqrt{t}\alpha}\right)\right] \tag{5.26}$$

The final element to include in the derivation of continuous Gaussian dispersal in order to make it comparable to the formulation used in the approximations we have applied here is to apply a correction for periodic boundary conditions. Similarly to the derivation for periodic correction derived in equation 5.14, the corrected form of the cell probabilities, ${w''}_{t_J}$ can be defined in terms of the uncorrected probabilities such that

$${w''}_{t_J} = w_{t_J} + \sum_{i_1=1}^{\infty} \left[\begin{array}{c} \sum_{i_2=-i_1}^{i_1} \left(w_{t_{J^{(i_1,i_2)}}} + w_{t_{J^{(-i_1,i_2)}}} + w_{t_{J^{(i_2,i_1)}}} + w_{t_{J^{(i_2,-i_1)}}} \right) \\ -w_{t_{J^{(i_1,i_1)}}} - w_{t_{J^{(-i_1,i_1)}}} - w_{t_{J^{(i_1,-i_1)}}} - w_{t_{J^{(-i_1,-i_1)}}} \end{array} \right] \tag{5.27}$$

Like equations 5.14 and 5.17, the convergent infinite series in equation 5.27 can be evaluated numerically using techniques such as those described in Caliceti *et al.* (2007). Here cell $J^{(i_1,i_2)}$ is a translation of cell J where $J^{(i_1,i_2)}$ is bounded by $[j_{x_1} + a_x i_1]$ and $[j_{x_2} + a_x i_1]$ on the x-axis and by $[j_{y_1} + a_y i_2]$ and $[j_{y_2} + a_y i_2]$ on the y-axis. The vector, ${\mathbf{W}''}_t$, with an element for each cell set to ${w''}_{t_J}$, provides a description of continuous dispersal over time that has a structure allowing comparison to the discrete approximations, $\mathbf{M}_t^{(\cdot)}$, described in this paper.

To test the effect of spatial scale of the lattice on the quality of the approximation, we calculate the relevant probabilities over lattices of three different grid sizes of 1, 3, and 5 units

in width and height. The simulation arena is a total of 45×45 units meaning that, in terms of cell count, the intermediate and coarse grained spatial resolutions comprise of arenas of 15×15 and 9×9 cells respectively. In each calculation we initialise the continuous Gaussian dispersal process with an individual starting at the centre of the grid, which for notational convenience we have designated as the origin of the x and y axes without loss of generality. For the discrete approximations, we initialise the starting probability vector, $\mathbf{M}_0^{(\cdot)}$, so that all elements are zero with the exception of the one cell containing the origin which is given a value of one.

α is set to $\frac{6}{\sqrt{\pi}} \approx 3.385$ for all calculations. Converting the bivariate normal displacement kernel into a probability density function of dispersal distance results in a rescaled Rayleigh distribution (Tufto *et al.*, 1997; Snäll *et al.*, 2007; Cousens *et al.*, 2008) with expected value $\frac{\alpha\sqrt{\pi}}{2}$ (Clark *et al.*, 1998). By setting α to $\frac{6}{\sqrt{\pi}}$, we standardise the expected dispersal distance over one time step to the cell length of the medium resolution grid. This provides a convenient midpoint benchmark to judge the approximation methods at grid resolutions with cell lengths larger than the expected dispersal distance, such as the 5×5 resolution grid, and grids at a finer scale than the scale of dispersal, such as the 1×1 grid.

Residence probabilities were calculated for each cell over the 40 time periods using the transition matrices generated using each of the four approximation methods described in this paper. Each of the resultant vectors of residence probabilities at each time period was compared to those expected under continuous dispersal.

5.3 Results

The absolute range of error values given in table 5.2 show, that for most grid sizes tested here, all four approximation methods provide a reasonable dispersal approximation to what would be expected under continuous dispersal. Here approximation error is defined as the difference between the probability that the individual resides within a cell at a given time period calculated according to the approximation method being tested and the probability that the individual would reside in that cell at the same time period under truly continuous dispersal (equation 5.27). Positive values represent incidences where the residence probabilities calculated by the approximation method exceed those expected under continuous dispersal whilst negative values denote incidences where the 'true' residence probabilities exceed those calculated by the approximation method.

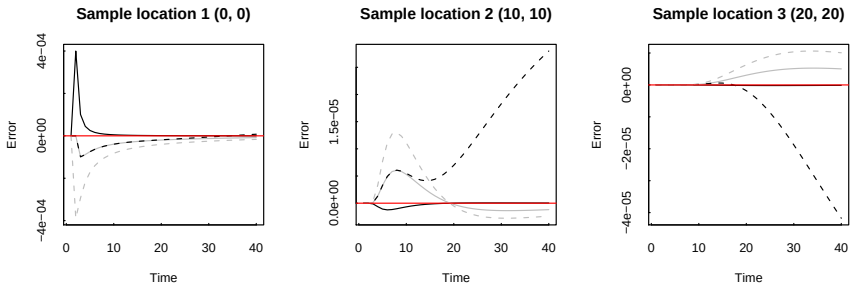
Residence probability estimates are correct to within three decimal places (< 0.0004) of the true probability for calculations made under the fine resolution grid (1×1 cell size) for all estimation methods calculated for all cells over the entire 40 step time period. For medium resolution grids (3×3 cell size), this accuracy reduces to values within 0.03 of the continuous dispersal baseline. At coarse resolutions (5×5 cell size), reasonable approximation to true continuous dispersal is not guaranteed: at the extremes, approximation methods show an inaccuracy in residence probability calculation of up to 0.175.

The time series of approximation error in figure 5.3 shows that the most extreme deviation from continuous dispersal occurs, for all approximation methods and grid resolutions, close to the origin in the earlier time periods. For locations further from the origin, the peak of approximation error occurs later in the time series, and at a much reduced magnitude. As time increases, a wave of increased residence probability spreads out from the centre of the simulation arena; if the timing for the arrival of this probability wave for an approximation method is different than that predicted under continuous dispersal then, during this period of disparity, we observe a peak of approximation method error.

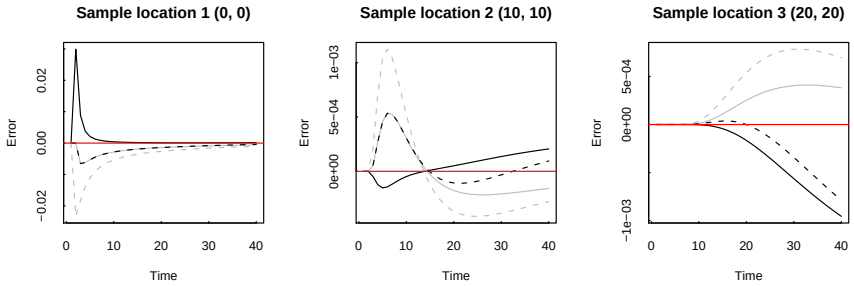
Figures 5.4, 5.5, and 5.6 show the spatial distribution of approximation error on fine, medium, and coarse resolution grids respectively. From these figures we can see that centroid-to-centroid methods tend to over-estimate the probability weights around the origin of dispersal. Conversely, centroid-to-area, area-to-centroid, and area-to-area methods all underestimate the residence probabilities in these areas whilst overestimating residence probabilities in the peripheries.

Under two-dimensional Gaussian diffusion, the variance of the probability mass function of the particle location (equation 5.23) tends to infinity as time increases. The cell residence probabilities, calculated with periodic boundary conditions according to equation 5.27, thus tend towards a uniform distribution bounded by the margins of the simulation arena. Due to the Markovian nature of the calculation mechanism for the residence probabilities for each approximation method, it is possible to calculate the asymptotic probability distribution for such methods. In Markovian models, the distribution of the asymptotic probability of residence is equivalent to the right eigenvector corresponding to the dominant eigenvalue of the transition matrix, rescaled so that all components sum to one. For properly defined transition matrices,

(a) Fine resolution grid with cell size of 1×1 units



(b) Medium resolution grid with cell size of 3×3 units



(c) Coarse resolution grid with cell size of 5×5 units

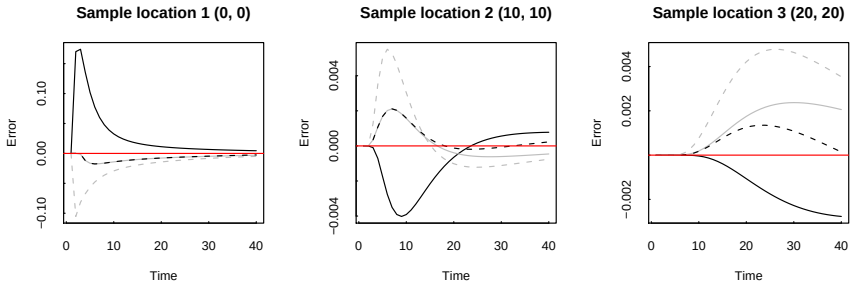
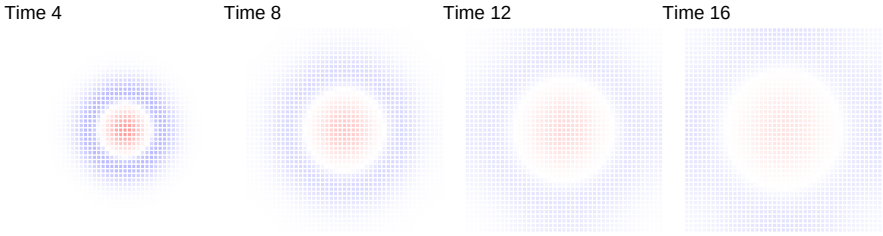


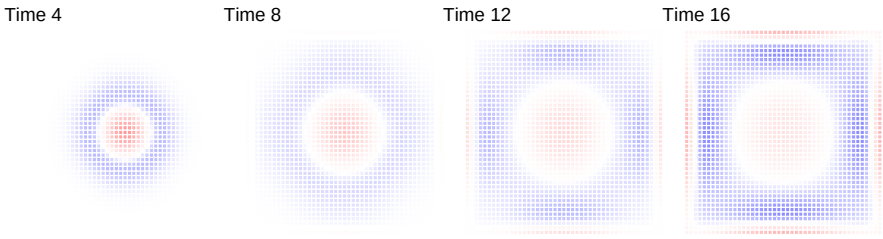
Figure 5.3: Time series of approximation method error at three different sites. Probability error is defined here as the difference of the probability of the individual residing in a cell under continuous dispersal (equation 5.27) and the probability calculated using a discrete approximation method. Positive values represent an 'excess' of probability, where the residence probabilities calculated by the approximation method exceed that expected under continuous dispersal. Conversely, negative values represent residence probabilities calculated by the approximation method below those expected under continuous dispersal. In each panel the solid black line refers to 'centroid-to-centroid' dispersal, the dashed black line to 'area-to-centroid' dispersal, the solid grey line to 'centroid-to-area', and the dashed grey line to 'area-to-area' dispersal. A time series of error is displayed for three cells chosen at successively further distances from the point of origin with sample point one representing the cell containing the origin, sample point two represents the cell containing the point at coordinates $(10, 10)$, and sample point three refers to the cell containing the point at coordinates $(20, 20)$. Figures (a)-(c) show the results at the three different spatial resolutions used in this study, from fine scales to coarse scale. The zero error line is denoted in red.

(a) Centroid-to-centroid dispersal

(b) Centroid-to-area dispersal



(c) Area-to-centroid dispersal



(d) Area-to-area dispersal

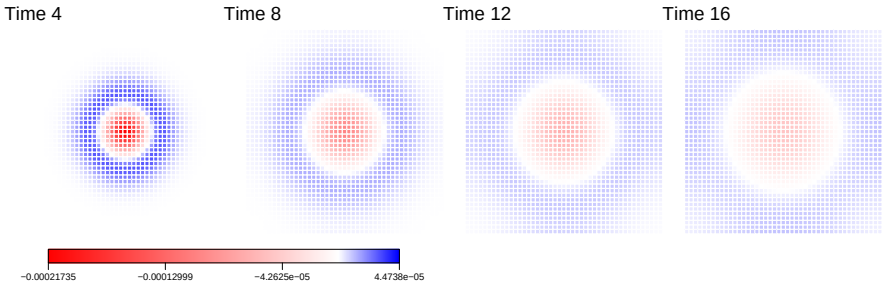
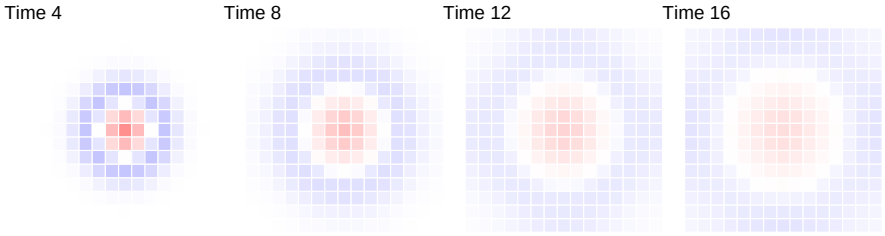


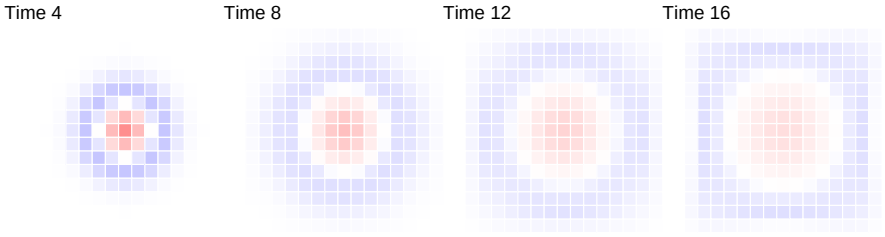
Figure 5.4: Spatial distribution of approximation method error through time on a fine resolution grid (cell size 1×1 units). Probability error is defined here as the difference of the probability of the individual residing in a cell under continuous dispersal (equation 5.27) and the probability calculated using a discrete approximation method. Positive values (blue shading in the panels above) represent an 'excess' of probability, where the residence probabilities calculated by the approximation method exceed that expected under continuous dispersal. Conversely, negative values (red shading in the panels above) represent residence probabilities calculated by the approximation method below those expected under continuous dispersal. Figures (a)-(d) show three snapshots of the spatial error for each of the four approximation methods.

(a) Centroid-to-centroid dispersal

(b) Centroid-to-area dispersal



(c) Area-to-centroid dispersal



(d) Area-to-area dispersal

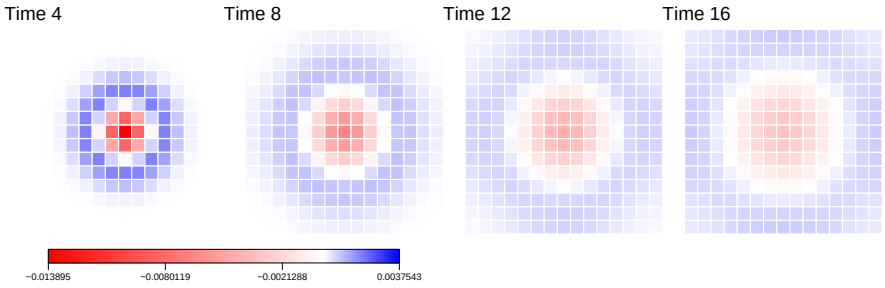
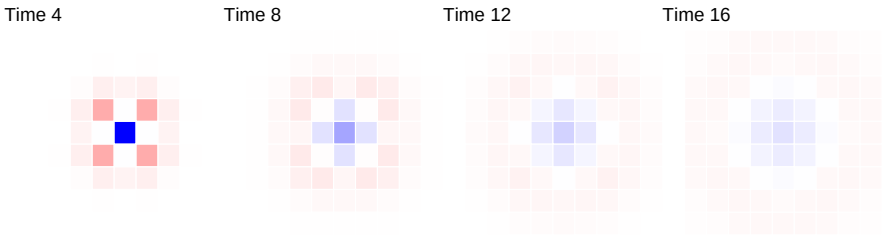
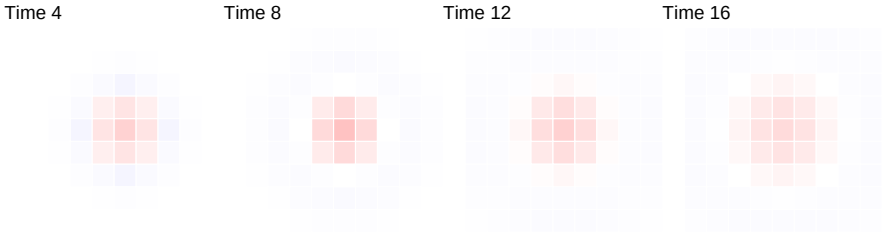


Figure 5.5: Spatial distribution of approximation method error through time on a medium resolution grid (cell size 3×3 units). Probability error is defined here as the difference of the probability of the individual residing in a cell under continuous dispersal (equation 5.27) and the probability calculated using a discrete approximation method. Positive values (blue shading in the panels above) represent an 'excess' of probability, where the residence probabilities calculated by the approximation method exceed that expected under continuous dispersal. Conversely, negative values (red shading in the panels above) represent residence probabilities calculated by the approximation method below those expected under continuous dispersal. Figures (a)-(d) show three snapshots of the spatial error for each of the four approximation methods.

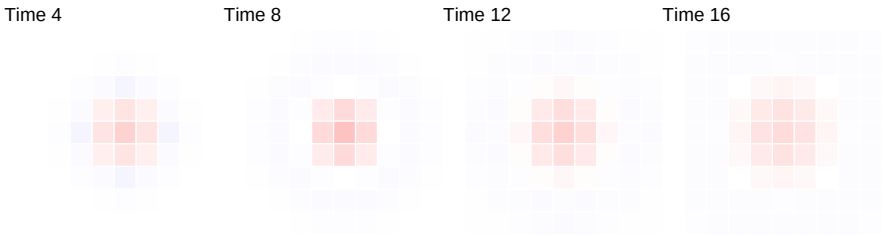
(a) Centroid-to-centroid dispersal



(b) Centroid-to-area dispersal



(c) Area-to-centroid dispersal



(d) Area-to-area dispersal

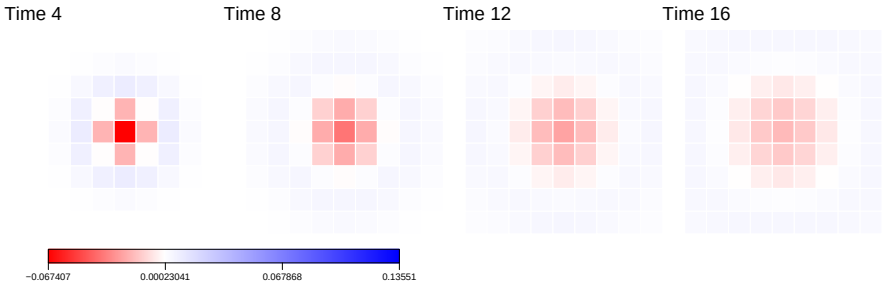


Figure 5.6: Spatial distribution of approximation method error through time on a coarse resolution grid (cell size 5×5 units). Probability error is defined here as the difference of the probability of the individual residing in a cell under continuous dispersal (equation 5.27) and the probability calculated using a discrete approximation method. Positive values (blue shading in the panels above) represent an 'excess' of probability, where the residence probabilities calculated by the approximation method exceed that expected under continuous dispersal. Conversely, negative values (red shading in the panels above) represent residence probabilities calculated by the approximation method below those expected under continuous dispersal. Figures (a)-(d) show three snapshots of the spatial error for each of the four approximation methods.

the dominant eigenvalue is always equal to one as the total probability is conserved between time periods.

Table 5.3 contains the sum of the absolute difference between the asymptotic residence probabilities calculated for each of the approximation methods at each cell resolution and the uniform probabilities expected under continuous dispersal. From table 5.3 it is clear that the centroid-to-area and area-to-area approximations methods exhibit very small deviations from the asymptotic expectation at all cell resolutions ($< 1.8 \times 10^{-8}$). Whilst the total deviation exhibited by the centroid-to-centroid and area-to-centroid methods are still relatively small (< 0.12), they are still many orders of magnitude larger than those exhibited by the areal destination methods and represent a non-negligible departure from the asymptotic optimum.

5.4 Discussion

The methods described in this paper provide a number of different mechanisms to approximate continuous dispersal in lattice-based models. We have shown that, for Gaussian dispersal at least, these approximations hold well at resolutions equivalent to the expected dispersal distance and finer. At coarse resolutions, the approximation methods described in this paper begin to exhibit significant deviations from what would be expected under continuous dispersal. The spatial signal of this error is quite different under the different approximation methods however. The over-estimation of residence probabilities at the core of the range observed under centroid-to-centroid dispersal can be explained by the fact that, under this dispersal regime, the distance between the origin and the destination sites are relatively large compared to the other dispersal approximation methods; centroid-to-area dispersal provides a destination area that has margins closer to the point of dispersal, area-to-centroid dispersal has a margin of the departure area closer to the destination point, and finally, area-to-area dispersal has margins of both destination and origin areas that are yet closer again. This results in centroid-to-centroid approximation methods generating residence probabilities in the nearby and source cells far in excess of what would be expected under continuous dispersal because the probability of spanning the distance between the origin and dispersal centroid for intermediately isolated and distant cells is very low (see Collingham *et al.*, 1996). The unit sum requirement for dispersal probabilities thus requires that the proportional weight be loaded in the nearby cells.

Table 5.2: Table of the range of approximation error for each approximation method and cell resolution over the entire arena and the 40 time steps calculated. Approximation error is defined here as the difference of the probability of the individual residing in a cell under continuous dispersal (equation 5.27) and the probability calculated using a discrete approximation method. Positive values represent an 'excess' of probability, where the residence probabilities calculated by the approximation method exceed that expected under continuous dispersal. Conversely, negative values represent residence probabilities calculated by the approximation method below those expected under continuous dispersal.

	Grid Cell Size		
Approximation Method	1×1	3×3	5×5
Centroid-to-centroid			
Minimum	-5.335×10^{-5}	-3.437×10^{-3}	-2.856×10^{-2}
Maximum	3.999×10^{-4}	2.993×10^{-2}	1.740×10^{-1}
Centroid-to-area			
Minimum	-9.904×10^{-5}	-6.565×10^{-3}	-1.755×10^{-2}
Maximum	1.339×10^{-5}	9.117×10^{-4}	5.381×10^{-3}
Area-to-centroid			
Minimum	-1.758×10^{-4}	-6.565×10^{-3}	-1.755×10^{-2}
Maximum	4.464×10^{-5}	9.177×10^{-4}	5.381×10^{-3}
Area-to-area			
Minimum	-3.885×10^{-4}	-2.333×10^{-2}	-1.053×10^{-1}
Maximum	5.224×10^{-5}	2.702×10^{-3}	1.239×10^{-2}

Table 5.3: Table of sums of asymptotic deviance of approximation methods from continuous dispersal. Elements are calculated from the dominant right eigenvector of the transition matrices used in equation 5.22. The element values are the sum of the absolute difference between the elements of the eigenvector and the uniform probability distribution that represents the asymptotic result of continuous dispersal.

	Grid Cell Size		
Approximation Method	1×1	3×3	5×5
Centroid-to-centroid	2.835×10^{-14}	1.042×10^{-1}	5.111×10^{-2}
Centroid-to-area	8.229×10^{-9}	1.748×10^{-8}	4.751×10^{-13}
Area-to-centroid	1.158×10^{-1}	1.107×10^{-1}	8.365×10^{-2}
Area-to-area	8.902×10^{-10}	4.927×10^{-11}	3.665×10^{-11}

The effect of increased cell size on the spatial distribution of approximation error observed in centroid-to-area, area-to-centroid, and area-to-area dispersal appears to act oppositely to that observed with centroid-to-centroid dispersal. Here, residence probabilities close to the origin of dispersal are underestimated whilst more distant dispersal events are predicted with a greater frequency than that expected under continuous dispersal. For those approximation methods where dispersal originates from an areal unit, this overestimation of residence probability in the peripheral grid cells can be explained from the added dispersal advantage conferred by the assumption that the point of departure is selected uniformly over the originating cell. If the originating cells are large then this first stage in the dispersal process can potentially garner origins of dispersal distant from locations likely to be dispersed to under continuous dispersal in the previous time period. In other words, dispersing individuals are 'pulled' across the interior of cells, effectively accelerating the dispersal rate. This extra process can, once compounded over several time steps, induce considerable increases in the predicted invasion speed.

The point-based origin of dispersal in centroid-to-area dispersal means that it may not be immediately obvious why centroid-to-area dispersal may suffer from the same spatial patterns of approximation error that afflict the area-to-centroid and area-to-area approximation methods. However, this phenomenon can be elucidated by envisioning the scenario where an individual moves from a cell centroid to just inside the margins of a nearby cell in one time period. When the model is iterated to the next time period, the individual is assumed to disperse from the centre of the destination cell of the last time period. Like the areal origin approximation methods, this effect essentially creates an extra intracellular dispersal event in each time period. Compounded over multiple time periods this effect will produce the observed spatial patterning of approximation error and can potentially bias predictions of expansion rates dramatically.

Whilst the absolute approximation error is an area of key consideration when selecting an appropriate approximation method (table 5.2), for simulations run over long-term timescales, particularly those studies that focus on the equilibrium properties of the system, it is also important for the investigator to consider the asymptotic properties of the approximation method applied (table 5.3). Methods that do not create outcomes that tend towards the continuous process that they are supposed to approximate will produce an artefact of approximation and may bias the interpretation of such results. Except at very fine scale resolutions, we have shown here that centroid-to-centroid and area-to-centroid dispersal do not exhibit the requisite asymptotic properties for these purposes. Both of these methods share the characteristic

that they require the evaluation of the dispersal kernel at a point. For a continuous dispersal kernel, the probability of dispersing to a point is infinitesimally small and, in order to express cell-to-cell dispersal probabilities in terms of a true probability that sums to unity across all possible destinations, both approximation methods require normalisation. As a result, both the centroid destination methods cannot characterise a 'true' dispersal process resulting in a long-term deviation from the continuous process, even if the approximation in the short and medium term is accurate.

The sensitivity of the approximation error and asymptotic properties of cell-based dispersal to the resolution of the lattice has resulted in a number of authors suggesting rules for appropriate cell resolution. Martin (1993) expresses such recommendations in terms of a so-called 'm-criterion'. In the context of dispersal approximation, this criterion is only satisfied if the cell length is less than or equal to the expected dispersal distance of the underlying continuous dispersal kernel over one time step. The expected dispersal distance of the underlying dispersal kernel can be calculated by converting the two-dimensional displacement kernel, $g_\cdot(r, \theta)$, into a probability distribution of distances (see Clark *et al.*, 1999; Cousens *et al.*, 2008), and calculating the expected value of this distribution. Rules for lattice-based dispersal have also been documented in Collingham *et al.* (1996), where the authors recommend that the cell lengths should be no longer than one half of the square root of the mean dispersal distance. Both heuristics may be excessively stringent however. The fine resolution grid (cell length 1×1) and dispersal kernel parametrisation evaluated in this study falls slightly outside the maximum cell length criterion of Collingham *et al.* (1996). However, even at the poorest performing locations and time periods within the 40 time periods sampled, all approximation methods described here still give accurate residence probabilities to within three decimal places at this spatial resolution. Moreover, the medium resolution grid (cell length 3×3) falls exactly on the limit of acceptability to satisfy the 'm-criterion' of Martin (1993). Even at this limit, the centroid-to-area and area-to-centroid dispersal methodologies still provide residence probabilities to within two decimal places of the continuous baseline.

For most applications, the methods described here will be employed to generate cell-to-cell transition probabilities for only a small number of parameterisations of the underlying dispersal kernel. In these circumstances, where computational resources are not limiting, it is recommended that fine-scale grid resolutions are used to minimise the approximation error. Small grid sizes also reduce biases in estimates of range expansion and contraction and the

results of such analyses are not so affected by the choice of approximation method. However, in studies where different parameterisations of the dispersal kernel need to be tested, such as in simulations of dispersal evolution (Hovestadt *et al.*, 2001; Travis & Dytham, 2002; Dytham, 2009) or as part of assessing the uncertainty in values of the dispersal parameters on metapopulation connectivity, repeated use of the approximation methods may be required. It is under these conditions that computational requirements may become a real concern. For dispersal kernels that have an analytical result specifying the approximation on discrete landscapes in closed form, or with simple numerically tractable forms (for example, the approximation of Gaussian dispersal derived in appendix A), there may not be much difference between the approximation methods in terms of computational time. For dispersal kernels where the only option is to perform numerical integration to evaluate the integral in the approximations that involve the dispersal to or from areal units (equations 5.7, 5.8, and 5.9), then the number of evaluations of the integrand required to achieve a good approximation will be related to the dimensionality of the integral. As a result, centroid-to-centroid approximation methods can be evaluated in the fastest time as there is no integration involved in their calculation. This is followed by area-to-centroid and centroid-to-area approximation methods which both require integration over one areal unit in their evaluation. Area-to-area methods will be the slowest, as they require integration over both the source and destination areas. It is worth noting however that whilst this relative ordering may be true for calculating the whole matrix of transition probabilities, centroid-to-centroid and area-to-centroid methods both require normalisation and, if only a subset of cell transitions are required, these methods will still require evaluation of the complete matrix in order to ascertain the correct normalisation values. In these instances, centroid-to-area and even area-to-area approximation methods may be able to provide faster results.

In situations where the transition matrix for multiple parameterisations of the underlying dispersal is required, our recommendations for approximation methods are a little more nuanced. Similarly to our recommendation for minimal parameterisation evaluation, we advocate the use of the smallest possible grid that is feasible given the computational resources. At high resolutions, the differences between the approximation methods are negligible and, as such, the method can be selected on the basis of speed alone (favouring centroid-to-centroid methods although see above). However, if the smallest computationally feasible grid size is larger than the expected dispersal distance, and therefore does not satisfy the 'm-criterion' of Martin (1993), then it may become necessary to use the slower areal destination methods (centroid-to-area or

area-to-area) that exhibit strong approximations to continuous dispersal and desirable asymptotic properties, even at coarse resolutions.

All attempts to emulate continuous dispersal on a discrete lattice will suffer from some form of approximation error. Indeed, Chesson & Lee (2005) state that theory relating to the continuous distribution, such as the moments and convolution properties, may not necessarily apply once the distribution has been mapped onto a discrete lattice. Whilst this is undoubtedly true, we argue here that when scaling up from point-to-point continuous dispersal to cell-to-cell dispersal it is useful to maintain a theoretical link between the dispersal as modelled at the smaller scale. In plants, dispersal kernels are most commonly fitted to seed shadow data (Clark *et al.*, 1999) or the outcomes from molecular parentage analysis (Robledo-Arnuncio & Garca, 2007). In animals, mark-release-recapture data (Fujiwara *et al.*, 2006) or telemetry data (Dahl & Willebrand, 2005; Rhoads *et al.*, 2010) methods are most commonly employed. As such, most studies will quote dispersal strategies in terms of point distances. To incorporate the information garnered from these small-scale studies into estimates of cell-to-cell or patch-to-patch connectivity, the calculation of which is an important prerequisite for any form of spatially-explicit metapopulation model (Hanski, 1994; Moilanen, 2004), it is important to define connectivity in terms of parameters derived from data collected at these scales.

Dispersal data collected at the metapopulation level does exist although this often requires considerably more field effort to collect. Hanski *et al.* (2000) describe a likelihood-based approach for the incorporation of observation records of individuals in a patch at a given time period. Whilst methods such as these can be critical in incorporating patch-scale data into model parametrisation and prediction formulation, they can only be expanded to allow inference to be drawn from point-to-point dispersal data if patch connectivity is specified in terms of the parameters relevant to these data. Parametrising metapopulation models in this way provides a mechanism for inference to be drawn from data collected at both levels simultaneously. A core process model, in this case our dispersal kernel, can be linked to a data set via an observation process. For the integration of data collected at different spatial scales, we can define a series of observation models for each data set relating the records to the core process model. For point-level data there exist a number of likelihood methods linking point settlement observations to the underlying dispersal kernel (Ribbens *et al.*, 1994; Clark *et al.*, 1998, 1999; Canham & Uriarte, 2006). For patch-level dispersal observations, the approximation methods described here can be employed to rescale the dispersal kernel to the relevant spatial extent of

the data. This provides a mechanism for the likelihood calculation of observed patch transitions given an underlying dispersal kernel.

For some study species, particularly territorial mammals, the field of dispersal ecology has pursued a much more mechanistic description of movement (for example Will & Tackenburg, 2008; van Moorter *et al.*, 2009). Such models are often described as rule-based because they rely mainly on simulation of individuals that move according to a set of rules rather than through description from a redistribution kernel. Some of the simpler simulation models can still be described in terms of a dispersal kernel and, for the approximation of such models in a discrete landscape, the methods described in the paper remain directly applicable. For the more complex models, where the description of the movement in terms of a dispersal kernel is not tractable, the approximation of transition probabilities must be garnered from direct simulation. Here multiple simulations must be performed. As the number of simulations grows large, the proportion of simulations that reside in each cell at the end of the movement will provide a reasonable approximation to the transition probabilities. If the simulations all start from the centre of the source cell then this corresponds to centroid-to-area dispersal whilst area-to-area dispersal corresponds to a set of simulations that pick a source location at random from within the source cell according to uniform distribution within its borders.

In summary, Holland *et al.* (2007) have shown that nearest neighbour dispersal produces results that are highly dependant upon the geometry of the lattice and the dispersal neighbourhood. For more reasonable implementations of dispersal, we must apply methods that approximate dispersal defined in continuous space to models where space is represented discretely. In most applications, centroid-to-centroid dispersal is used as a default approximation method. Whilst this may represent the least demanding method in terms of computational power, we have demonstrated that such methods can provide a very poor approximation to continuous dispersal: producing biased estimates of invasion speed and asymptotic residence probabilities. Conversely, approximation methods with areal destination spatial units exhibit both desirable asymptotic qualities and high accuracy, even at relatively coarse spatial scales. The adoption of these more complex methods need not be demanding and the use of numerical tools such as the `ecomodtools` package, or through direct derivation (such as that described for Gaussian dispersal in appendix A), can provide the investigator with a much better approximation of continuous dispersal at very little cost in terms of time, either computationally or in implementation. Moreover, we have shown how rows and columns of the transition matrices

generated using these approximation methods can be aggregated to provide estimates of patch connectivity for use in metapopulation and metacommunity models. These methods may provide a valuable part of a suite of techniques to draw inference about ecological processes from data collected at multiple spatial scales.

Appendix 5.A Approximation of Gaussian Dispersal on a Lattice

We start with the two-dimensional displacement kernel for Gaussian dispersal, denoted here by $g_G(r, \theta)$, derived by Clark *et al.* (1999) as a special case of the generalised exponential distribution with probability density function

$$g_G(r, \theta) = \frac{1}{\pi\alpha^2} e^{-\left(\frac{r}{\alpha}\right)^2} \tag{5.28}$$

The Gaussian dispersal as described by Clark *et al.* (1999) is isotropic, so the corresponding probability density function is not dependent upon the direction of travel, θ. Because of this, it is often quoted in the literature as simply $g_G(r)$ but we include it here with the full notation to emphasise the fact that it represents a joint probability density function of distance and direction.

The relevant reparametrisation of $g_G(r, \theta)$ in terms of Cartesian source and destination coordinates according to equation 5.6 is

$$\begin{aligned} c_G(j_x, j_y, k_x, k_y) &= \frac{1}{\pi\alpha^2} e^{-\frac{(k_x - j_x)^2 + (k_y - j_y)^2}{\alpha^2}} \\ &= \frac{1}{\pi\alpha^2} e^{-\left(\frac{k_x - j_x}{\alpha}\right)^2} e^{-\left(\frac{k_y - j_y}{\alpha}\right)^2} \end{aligned} \tag{5.29}$$

This corresponds to the probability density function of a set of destination coordinates drawn from a bivariate normal distribution with mean parameters set according to the source coordinates, with no correlation between the x and y coordinates, and with equal variance in each dimension.

By using the kernel reparametrisation above in equation 5.5, it is possible to describe the Gaussian centroid-to-centroid dispersal probability from source cell J to destination cell K,

$p_{G_{JK}}^{(CC)}$

$$p_{G_{JK}}^{(CC)} = \frac{e^{-\left(\frac{k_{x_2}+k_{x_1}-j_{x_2}-j_{x_1}}{2\alpha}\right)^2} e^{-\left(\frac{k_{y_2}+k_{y_1}-j_{y_2}-j_{y_1}}{2\alpha}\right)^2}}{\sum_L e^{-\left(\frac{l_{x_2}+l_{x_1}-j_{x_2}-j_{x_1}}{2\alpha}\right)^2} e^{-\left(\frac{l_{y_2}+l_{y_1}-j_{y_2}-j_{y_1}}{2\alpha}\right)^2}} \tag{5.30}$$

All of the other forms of lattice dispersal described in this paper require the integration of $c_G\left(j_x, j_y, k_x, k_y\right)$:

$$\begin{aligned}
&\int_{k_{y_1}}^{k_{y_2}} \int_{k_{x_1}}^{k_{x_2}} c_G\left(j_x, j_y, k_x, k_y\right)\, dk_x dk_y \\
&\qquad = \frac{1}{\pi\alpha^2} \int_{k_{y_1}}^{k_{y_2}} e^{-\left(\frac{k_y-j_y}{\alpha}\right)^2} \int_{k_{x_1}}^{k_{x_2}} e^{-\left(\frac{k_x-j_x}{\alpha}\right)^2} dk_x dk_y \\
&\qquad = \frac{1}{\pi\alpha} \int_{k_{y_1}}^{k_{y_2}} e^{-\left(\frac{k_y-j_y}{\alpha}\right)^2} \int_{\frac{k_{x_1}-j_x}{\alpha}}^{\frac{k_{x_2}-j_x}{\alpha}} e^{-\upsilon_x^2} d\upsilon_x dk_y \\
&\qquad = \frac{1}{2\sqrt{\pi}\alpha} \left[\operatorname{erf}\left(\frac{k_{x_2}-j_x}{\alpha}\right) - \operatorname{erf}\left(\frac{k_{x_1}-j_x}{\alpha}\right)\right] \int_{k_{y_1}}^{k_{y_2}} e^{-\left(\frac{k_y-j_y}{\alpha}\right)^2} dk_y \\
&\qquad = \frac{1}{2\sqrt{\pi}} \left[\operatorname{erf}\left(\frac{k_{x_2}-j_x}{\alpha}\right) - \operatorname{erf}\left(\frac{k_{x_1}-j_x}{\alpha}\right)\right] \int_{\frac{k_{y_1}-j_y}{\alpha}}^{\frac{k_{y_2}-j_y}{\alpha}} e^{-\upsilon_y^2} d\upsilon_y \\
&\qquad = \frac{1}{4} \left[\operatorname{erf}\left(\frac{k_{x_2}-j_x}{\alpha}\right) - \operatorname{erf}\left(\frac{k_{x_1}-j_x}{\alpha}\right)\right] \left[\operatorname{erf}\left(\frac{k_{y_2}-j_y}{\alpha}\right) - \operatorname{erf}\left(\frac{k_{y_1}-j_y}{\alpha}\right)\right]
\end{aligned} \tag{5.31}$$

where $\upsilon_x = \frac{k_x-j_x}{\alpha}$ and $\upsilon_y = \frac{k_y-j_y}{\alpha}$ represent substitutions used in the integration and $\operatorname{erf}(x)$ is the error function defined as

$$\operatorname{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \tag{5.32}$$

Whilst the error function does not have a closed analytical form, there exists a number of numerical techniques, such as those described in Cody (1993), for efficient approximate evaluation.

Using a method similar to that used in the derivation of equation 5.31 it can also be shown that

$$\begin{aligned}
&\int_{j_{y_1}}^{j_{y_2}} \int_{j_{x_1}}^{j_{x_2}} c_G\left(j_x, j_y, k_x, k_y\right)\, dj_x dj_y \\
&\qquad = \frac{1}{4} \left[\operatorname{erf}\left(\frac{k_x-j_{x_2}}{\alpha}\right) - \operatorname{erf}\left(\frac{k_x-j_{x_1}}{\alpha}\right)\right] \left[\operatorname{erf}\left(\frac{k_y-j_{y_2}}{\alpha}\right) - \operatorname{erf}\left(\frac{k_y-j_{y_1}}{\alpha}\right)\right]
\end{aligned} \tag{5.33}$$

Substituting the results of equation 5.31 into equation 5.7 and the results of equation 5.33 into equation 5.9 provides a mechanism to specify the Gaussian centroid-to-area and area-to-centroid probabilities, $p_{G_{JK}}^{(CA)}$ and $p_{G_{JK}}^{(AC)}$ respectively, in terms of the numerically tractable error

function so that

$$p_{G_{JK}}^{(CA)} = \frac{1}{4}\left[\operatorname{erf}\left(\frac{2k_{x_2} - j_{x_2} - j_{x_1}}{2\alpha}\right) - \operatorname{erf}\left(\frac{2k_{x_1} - j_{x_2} - j_{x_1}}{2\alpha}\right)\right] \left[\operatorname{erf}\left(\frac{2k_{y_2} - j_{y_2} - j_{y_1}}{2\alpha}\right) - \operatorname{erf}\left(\frac{2k_{y_1} - j_{y_2} - j_{y_1}}{2\alpha}\right)\right] \tag{5.34}$$

and

$$p_{G_{JK}}^{(AC)} = \frac{\left[\operatorname{erf}\left(\frac{k_{x_2} + k_{x_1} - 2j_{x_2}}{2\alpha}\right) - \operatorname{erf}\left(\frac{k_{x_2} + k_{x_1} - 2j_{x_1}}{2\alpha}\right)\right] \left[\operatorname{erf}\left(\frac{k_{y_2} + k_{y_1} - 2j_{y_2}}{2\alpha}\right) - \operatorname{erf}\left(\frac{k_{y_2} + k_{y_1} - 2j_{y_1}}{2\alpha}\right)\right]}{\sum_L \left[\operatorname{erf}\left(\frac{l_{x_2} + l_{x_1} - 2j_{x_2}}{2\alpha}\right) - \operatorname{erf}\left(\frac{l_{x_2} + l_{x_1} - 2j_{x_1}}{2\alpha}\right)\right] \left[\operatorname{erf}\left(\frac{l_{y_2} + l_{y_1} - 2j_{y_2}}{2\alpha}\right) - \operatorname{erf}\left(\frac{l_{y_2} + l_{y_1} - 2j_{y_1}}{2\alpha}\right)\right]} \tag{5.35}$$

Starting from the result of equation 5.33 we show that

$$\begin{aligned} &\int_{k_{y_1}}^{k_{y_2}} \int_{k_{x_1}}^{k_{x_2}} \int_{j_{y_1}}^{j_{y_2}} \int_{j_{x_1}}^{j_{x_2}} c_G\left(j_x, j_y, k_x, k_y\right)\, dj_x dj_y dk_x dk_y \\ &= \frac{1}{4}\left[\int_{k_{y_1}}^{k_{y_2}} \operatorname{erf}\left(\frac{k_y - j_{y_2}}{\alpha}\right) dk_y - \int_{k_{y_1}}^{k_{y_2}} \operatorname{erf}\left(\frac{k_y - j_{y_1}}{\alpha}\right) dk_y\right] \\ &\quad \left[\int_{k_{x_1}}^{k_{x_2}} \operatorname{erf}\left(\frac{k_x - j_{x_2}}{\alpha}\right) dk_x - \int_{k_{x_1}}^{k_{x_2}} \operatorname{erf}\left(\frac{k_x - j_{x_1}}{\alpha}\right) dk_x\right] \end{aligned} \tag{5.36}$$

We can express the indefinite integral of the error function in terms of the error function

$$\begin{aligned} \int \operatorname{erf}\left(\frac{k_\cdot - j_{\cdot_\cdot}}{\alpha}\right) dk_\cdot &= \alpha \int \operatorname{erf}\left(\zeta\right)\, d\zeta \\ &= \left(k_\cdot - j_{\cdot_\cdot}\right) \operatorname{erf}\left(\frac{k_\cdot - j_{\cdot_\cdot}}{\alpha}\right) + \frac{\alpha}{\sqrt{\pi}} e^{-\left(\frac{k_\cdot - j_{\cdot_\cdot}}{\alpha}\right)^2} + C \end{aligned} \tag{5.37}$$

where $\zeta = \frac{k_\cdot - j_{\cdot_\cdot}}{\alpha}$ is a substitution used in the integration and C is a constant resulting from indefinite integration.

Substituting the results of equation 5.37 into equation 5.36 allows us to express the Gaussian

area-to-area transition probabilities, as defined in equation 5.8, in terms of the error function

$$p_{G_{JK}}^{(AA)} = \frac{1}{4\left(j_{x_2}-j_{x_1}\right)\left(j_{y_2}-j_{y_1}\right)} \left[\begin{array}{c} \left(k_{y_2}-j_{y_2}\right)\operatorname{erf}\left(\frac{k_{y_2}-j_{y_2}}{\alpha}\right)-\left(k_{y_1}-j_{y_2}\right)\operatorname{erf}\left(\frac{k_{y_1}-j_{y_2}}{\alpha}\right)- \\ \left(k_{y_2}-j_{y_1}\right)\operatorname{erf}\left(\frac{k_{y_2}-j_{y_1}}{\alpha}\right)+\left(k_{y_1}-j_{y_1}\right)\operatorname{erf}\left(\frac{k_{y_1}-j_{y_1}}{\alpha}\right)+ \\ \frac{\alpha}{\sqrt{\pi}}\left(e^{-\left(\frac{k_{y_2}-j_{y_2}}{\alpha}\right)^2}-e^{-\left(\frac{k_{y_1}-j_{y_2}}{\alpha}\right)^2}-e^{-\left(\frac{k_{y_2}-j_{y_1}}{\alpha}\right)^2}+e^{-\left(\frac{k_{y_1}-j_{y_1}}{\alpha}\right)^2}\right) \end{array}\right] \left[\begin{array}{c} \left(k_{x_2}-j_{x_2}\right)\operatorname{erf}\left(\frac{k_{x_2}-j_{x_2}}{\alpha}\right)-\left(k_{x_1}-j_{x_2}\right)\operatorname{erf}\left(\frac{k_{x_1}-j_{x_2}}{\alpha}\right)- \\ \left(k_{x_2}-j_{x_1}\right)\operatorname{erf}\left(\frac{k_{x_2}-j_{x_1}}{\alpha}\right)+\left(k_{x_1}-j_{x_1}\right)\operatorname{erf}\left(\frac{k_{x_1}-j_{x_1}}{\alpha}\right)+ \\ \frac{\alpha}{\sqrt{\pi}}\left(e^{-\left(\frac{k_{x_2}-j_{x_2}}{\alpha}\right)^2}-e^{-\left(\frac{k_{x_1}-j_{x_2}}{\alpha}\right)^2}-e^{-\left(\frac{k_{x_2}-j_{x_1}}{\alpha}\right)^2}+e^{-\left(\frac{k_{x_1}-j_{x_1}}{\alpha}\right)^2}\right) \end{array}\right] \tag{5.38}$$

Appendix 5.B Derivation of the Probability Density Function of Sums of Uncorrelated Bivariate-Normal Distributed Variables

We begin with the description of two, two-dimensional displacement events such that the change in the x and y dimension, δ_x and δ_y respectively, for each event are described by the bivariate Gaussian distribution. Both displacement events have a mean of zero and separate isotropic standard deviation parameters: α_1 for displacement event one and α_2 for displacement event two. The probability density functions for each event, $s_1\left(\delta_x,\delta_y|\alpha_1\right)$ and $s_2\left(\delta_x,\delta_y|\alpha_2\right)$ respectively, are therefore

$$s_1\left(\delta_x,\delta_y|\alpha_1\right) = \frac{1}{\pi\alpha_1^2}e^{-\left(\frac{\delta_x}{\alpha_1}\right)^2}e^{-\left(\frac{\delta_y}{\alpha_1}\right)^2} \tag{5.39}$$

$$s_2\left(\delta_x,\delta_y|\alpha_2\right) = \frac{1}{\pi\alpha_2^2}e^{-\left(\frac{\delta_x}{\alpha_2}\right)^2}e^{-\left(\frac{\delta_y}{\alpha_2}\right)^2} \tag{5.40}$$

The probability density functions above correspond to the special bivariate-Normal case where there exists no correlation in the variate vector elements, that is, dispersal is not correlated in the x and y spatial dimensions.

The probability density function of the total displacement, $s_3\left(\delta_x,\delta_y|\alpha_1,\alpha_2\right)$, after both displacement events described above are applied in sequence is the two-dimensional convolution

of the two component displacement kernels such that

$$\begin{aligned} s_3\left(\delta_x, \delta_y | \alpha_1, \alpha_2\right) &= \frac{1}{\pi^2 \alpha_1^2 \alpha_2^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\left(\frac{\delta_x - \tau_x}{\alpha_1}\right)^2} e^{-\left(\frac{\delta_y - \tau_y}{\alpha_1}\right)^2} e^{-\left(\frac{\tau_x}{\alpha_2}\right)^2} e^{-\left(\frac{\tau_y}{\alpha_2}\right)^2} d\tau_x d\tau_y \\ &= \frac{1}{\pi^2 \alpha_1^2 \alpha_2^2} \left[\int_{-\infty}^{\infty} e^{-\left(\frac{\delta_x - \tau_x}{\alpha_1}\right)^2} e^{-\left(\frac{\tau_x}{\alpha_2}\right)^2} d\tau_x\right] \left[\int_{-\infty}^{\infty} e^{-\left(\frac{\delta_y - \tau_y}{\alpha_1}\right)^2} e^{-\left(\frac{\tau_y}{\alpha_2}\right)^2} d\tau_y\right] \end{aligned} \tag{5.41}$$

where

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-\left(\frac{\delta_. - \tau_.}{\alpha_1}\right)^2} e^{-\left(\frac{\tau_.}{\alpha_2}\right)^2} d\tau_. &= e^{-\frac{\delta_.^2}{\alpha_2^2 + \alpha_1^2}} \int_{-\infty}^{\infty} e^{-\left(\frac{\alpha_1^2 \tau_. + \alpha_2^2 (\tau_. - \delta_.)}{\alpha_1 \alpha_2 \sqrt{\alpha_2^2 + \alpha_1^2}}\right)^2} d\tau_. \\ &= \frac{\alpha_1 \alpha_2}{\sqrt{\alpha_2^2 + \alpha_1^2}} e^{-\frac{\delta_.^2}{\alpha_2^2 + \alpha_1^2}} \int_{-\infty}^{\infty} e^{-\xi_.^2} d\xi_. \\ &= \sqrt{\frac{\pi}{\alpha_2^2 + \alpha_1^2}} \alpha_1 \alpha_2 e^{-\frac{\delta_.^2}{\alpha_2^2 + \alpha_1^2}} \end{aligned} \tag{5.42}$$

and $\xi_.$ is an integration substitution

$$\xi_. = \frac{\alpha_1^2 \tau_. + \alpha_2^2 \left(\tau_. - \delta_.\right)}{\alpha_1 \alpha_2 \sqrt{\alpha_2^2 + \alpha_1^2}} \tag{5.43}$$

Substituting the results of equation 5.42 into equation 5.41 gives

$$s_3\left(\delta_x, \delta_y | \alpha_1, \alpha_2\right) = \frac{1}{\pi\left(\alpha_2^2 + \alpha_1^2\right)} e^{-\frac{\delta_x^2}{\alpha_2^2 + \alpha_1^2}} e^{-\frac{\delta_y^2}{\alpha_2^2 + \alpha_1^2}} \tag{5.44}$$

From this result it is clear that the probability density function of the total displacement is a reparametrisation of the bivariate-Normal density function with a new standard deviation parameter, α_3, where $\alpha_3^2 = \alpha_1^2 + \alpha_2^2$. More generally, if we describe a discrete time process with bivariate-normal displacement in each time step and for which the standard deviation parameter is set to α isotropically, then the total displacement encountered at a time t is also described by a bivariate-normal process with standard deviation parameter $\alpha_t = \sqrt{t}\alpha$.

CHAPTER 6

Discussion

6.1 Main Conclusions

Many central theories of population genetics, such as the derivation of Hardy-Weinberg equilibrium (Hardy, 1908; Weinberg, 1908, and translated in Weinberg 1963), the description of neutral drift in closed populations (Kimura, 1983, 1986), and the extension of the Wright-Fisher model to describe genealogical coalescence (Kingman, 1982a,b; Hudson, 1983; Tajima, 1983; Tavaré, 1984), all assume, in their basic form at least, that population sizes remain constant. Some progress has been made to extend some of these theories to situations where population size is not constant, such as the extensions of coalescent theory (see Donnelly & Tavaré, 1995), but these attempt are largely restricted to basic, deterministic scenarios of population change.

Conversely, classical population biology typically deals with models of population dynamics that, by their very design, describe changes in population size but ignore individual level variation. The deterministic models of population growth described by Pearl and Verhulst (see Gilpin & Ayala, 1973), Ricker (1954, 1975), Beverton & Holt (1957), and Gompertz (1825), share the characteristic that they do not account for stratification of population-relevant parameters within the population.

Efforts have however been made in the field of population biology to allow for some degree of

within-population differentiation. By defining population structure as a vector it is possible to specify unique fecundity, mortality and growth parameters for each population sub-section such as age class (Leslie, 1945, 1948) or life-cycle stage (Lefkovitch, 1965). Another way to try to incorporate individual differences in demographic parameters is to incorporate those differences implicitly as a stochastic element in the model. This can involve adding demographic and environmental stochasticity terms to existing models describing population dynamics (Brännström & Sumpter, 2006) or by building models of population dynamics from the sum total of sub-models of individual-level reproductive output (Haccou *et al.*, 2007; Lebreton *et al.*, 2007).

Individual-based models provide the next logical step for the investigation of the effect of individual-level variability on population-level phenomena. In these models, the characteristics of the individual can be explicitly incorporated into the model and the interaction between the individuals and the environment lead to the so-called emergent properties of interest (Łomnicki, 1999; Grimm *et al.*, 1999; DeAngelis *et al.*, 1994). Because the model description is already set at the level of the individual, it is with these models that the opportunity to incorporate genetic information is greatest.

The main thrust of this thesis has been to provide methods to allow the scaling up of information found at the level of the gene to investigate patterns at the level of the population. The theoretical basis of both population biology and population genetics has been extended by many authors over the years, to the point that there are many points of synergy between the two. It is not suggested, nor is it the intention, that this thesis is a complete and thorough synthesis of these two related, but separate, fields of ecology. It is intended however that the methods provided in this thesis provide points of departures from which further developments, aimed at linking genetic information to patterns of population dynamics, can be established.

6.1.1 A Flexible and Robust Method of Allele Frequency Estimation

In chapter 1 we have discussed the importance of sound estimates of allele frequencies in inferring population structure. Indeed, the accuracy of estimates of allele frequencies underpins nearly all inference in the field of population genetics. Tests of Hardy-Weinberg equilibrium, such as those described in Engels (2009), Schaid *et al.* (2006), and Troendle & Yu (1994), can be particularly sensitive to biases in allele frequency estimates arsing from genotype observation error (Morin *et al.*, 2009; Mitchell *et al.*, 2003). Allele frequency information can be used to diagnose recent population bottlenecks (Cornuet & Luikart, 1996) as implemented in the computer

program BOTTLENECK (Piry *et al.*, 1999). Tests of population structure, and the estimation of fixation indices, also require accurate estimates of relative allele frequencies (*sensu* Weir & Cockerham, 1984). Imperfect observation, either resulting from allele diagnosis errors (Clarke *et al.*, 2001; Ewen *et al.*, 2000) or the presence of recessive alleles (Dakin & Avise, 2004), all provide mechanisms through which uncertainty in the allele frequency estimates can be introduced.

Both Stewart Jr & Excoffier (1996) and Lynch & Milligan (1994) describe methods to estimate allele frequencies for biallelic dominant markers. Under certain assumptions, Lynch & Milligan (1994) also derives a measure of uncertainty around the point estimate of the allele frequency, but neither method goes so far as to describe the full probability distribution of the recessive allele frequency. The method of Zhivotovsky (1999) is the first method to attempt this. Holsinger *et al.* (2002) and Foll *et al.* (2008) extend the allele estimation procedure to also incorporate the estimation of fixation indices from observed genotype data, but these methods only incorporate the uncertainty arising from observations obscured by allelic dominance. These methods, along with those of Guo & Thompson (1992) for codominant markers, all assume that there is either no genotyping error or that it plays an insignificant part in allele frequency estimation. Chapter 2 represents the first step in incorporating the uncertainties into allele frequency estimates arising from both observation processes: that of recessive allele obfuscation and allele diagnosis. The chapter also continues to describe how the local fixation index, F_{IS}, can also be jointly estimated. By providing estimates of allele frequency and local fixation indices, and, crucially, the uncertainty around these estimates, the methods described in chapter 2 provide a robust basis from which further analysis can be performed.

Whilst the methods described in chapter 2 allows greater inclusion of sources of error in allele frequency estimation, in practice this results in a rather high load of free parameters however. The full list of estimable parameters in this model includes the parameters of the genotyping observation model, the allele frequencies, and the inbreeding / outbreeding coefficient. Some of these parameters may exhibit dependence and so, in the absence of prior information to narrow the probable parameter space, the investigator may suffer problems with parameter identifiability. When the number of loci are few, the data set may contain insufficient information to separate the effects of extreme allele frequency from a population exhibiting severe inbreeding; a population with a large excess of homozygotes (high fixation) is almost indistinguishable from a population where one allele type is at almost full exclusivity. This is not an artefact of the model but a real statistical feature when there is little information in the data set. In these situ-

ations the only solutions are either to increase the number of loci or set more informative priors.

Foll *et al.* (2008) also describe another source of bias in allele frequency estimation based on loci selection criteria. Informative loci are those that exhibit polymorphism. Loci with alleles at extremely low or high proportional frequencies are much less likely to exhibit the relevant level of polymorphism to be selected for use in a study. Because loci with extreme alleleic frequencies are unlikely to be used in a study, Foll *et al.* (2008) argue that a uniform prior of allele frequencies is not suitable. We argue in chapter 2 that this ascertainment bias can be at least partially accounted for by a suitable parameterisation of the beta distribution for biallelic markers and a Dirichlet distribution for polyallelic markers. This allows the investigator to set a prior weight of allele frequencies that tends to zero at the extremes.

6.1.2 Propagating Uncertainty from Parentage Analysis

Patterns of parentage underpin a number of key ideas in the field of population genetics and population ecology. Variation in the breeding success of individuals affects rate of coalescence (Rosenberg & Nordborg, 2002; Nordborg, 2001), demographic stochasticity, and extinction (Haccou *et al.*, 2007). The concept of 'effective population size' introduced by Wright (1931, 1938), and often used as basic parameter in models of population dynamics, can be thought as mechanism to correct for unequal genetic contributions of individuals to the next generation. Where direct observation is too expensive or not possible, one of the many methods of molecular parentage analysis (see Jones & Ardren, 2003) can be applied to fill the gap.

Simple exclusion methods, where potential parent combinations are ruled out based on genetic incompatibility with their offspring, suffer from the problem that if loci number are too few or exhibit insufficient polymorphism then it may not be possible to narrow down potential parents to one pair. However, more importantly, even at low frequencies, genotyping errors and mutations can cause severe problems for exclusion methods of parentage analysis (Cifuentes *et al.*, 2006). Only one error at one locus is needed to create an incompatibility that could exclude the true parent pair.

Likelihood-based methods fare much better in this regard. Here the probability of a putative parent pair is calculated using Mendelian transition / segregation probabilities and an observation model. Likelihood-based methods give a mechanism by which different parent pairs can be weighted and reduces the sensitivity of the assignment to genotyping errors (Jones *et al.*,

2010). However, simple assignment of parentage to pairs which exhibit the highest likelihood, or posterior probability in Bayesian analyses, ignores the uncertainty attached to that assignment. If the parentage assignment is to form the basis of further study then it is imperative that this uncertainty be included in that analysis. To this end, only the so-called 'fractional' methods of parentage analysis (see Devlin *et al.*, 1988) can be used to robustly propagate this uncertainty between hierarchical levels. Fractional methods do not assign paternity, but are used instead to describe the joint probability distribution of the maternity and paternity. This allows the confidence of outputs of further analysis to be weighted according to the probability of their assumptions about the parentage.

Chapter 3 discusses the different genotype observation models employed in various parentage analysis computer packages and describes how they may be unsuitable for most marker types. A series of marker-specific observation models are espoused which allow for the presence of recessive alleles and better emulate the observation process for each marker type. A general framework for fractional parentage analysis is described which allows for the application of these marker-specific error models to systems of arbitrary ploidy. Like chapter 2, the framework portrayed in chapter 3 provides another mechanism for the incorporation of information contained within genetic-level data to population-level problems whilst preserving any uncertainty in the data set.

6.1.3 Modelling with Parameter Uncertainty in IBMs

Chapters 2 and 3 both outline methods that can be applied in the calculation of individual-level life history parameters. The main premise of both techniques however, is to preserve the uncertainty that surrounds these parameter estimates for the next stage of the analysis. One broad set of techniques for the investigation of the effect of individual-level dynamics on population-level phenomena include individual-based models (IBMs). This class of models explicitly models the interaction of individuals with each other and with their environment and allows for the assessment of how changes in the behavioural rules of the individuals affects emergent properties of the system being modelled.

One of the major obstacles in adopting an individual-based approach to modelling applied ecological problems is that it is often difficult to fit individual-based models to real data. Moreover, the field of individual-based ecology currently lacks a statistically robust mechanism for the assessment of the effect of parameter uncertainty on model outputs. This feature of IBMs

is primarily driven by the fact that, except in the simplest of cases, there exists no equation expressible in a simple closed form that links the input parameters to the patterns of interest. This analytic intractability means that inference from IBMs rely on Monte Carlo methods. Because extensive simulation from IBMs is costly, the number of parameter combinations that can be tested in any sensitivity analysis is limited. Some heuristics for IBM analysis have been proposed by authors such as Wiegand *et al.* (2003). These allow for a rough estimation of the range of outputs given known valid ranges of inputs, but these methods have do not have a statistical basis and it is unknown how such methods can deal with input parameters that exhibit complex associations such as colinearity.

An alternative tactic for the assessment of parameter uncertainty on model outputs is to view the joint probability distribution of the input parameters as a prior distribution in a Bayesian analysis. The main strength of a Bayesian analysis is that the posterior distribution of parameters from one analysis, like for example the posterior parameter outputs from the methods outlined in chapters 2 and 3, can be used as the prior distribution for a later analysis. In this sense it is possible to 'daisy-chain' together Bayesian models, with each sub model bringing new data to bear on what is known about the set of parameters (Clark & Gelfand, 2006). However, the absence of a tractable likelihood function in most IBMs mean that standard Bayesian techniques cannot be applied in these cases.

Chapter 4 describes the application of an existing set of approximate Bayesian techniques (Pritchard *et al.*, 1999; Beaumont *et al.*, 2002; Marjoram *et al.*, 2003; Sisson *et al.*, 2007; Toni *et al.*, 2009; Toni & Stumpf, 2010) which do not require a known likelihood function, in the context of IBM analysis. Chapter 4 develops the methods described in Sisson *et al.* (2007) and Toni *et al.* (2009); Toni & Stumpf (2010) further, describing an efficient way that IBMs can be fit to time series data. Where chapters 2 and 3 aim to provide mechanisms to scale up from the gene to the individual, the techniques of chapter 4 provide a novel mechanism for linking information at the individual level to inference at the population level.

6.2 Future Work

6.2.1 Combination of Parentage and Allele Frequency Estimation

The methods described for the estimation of allele frequencies and assignment of parentage outlined in chapters 2 and 3 are not technically independent. A key step in the parentage

assignment algorithm of chapter 3 is the summation of observation probabilities over different genotype possibilities (equation 3.1). This requires an application of Bayes theorem to calculate the probability of a genotype allele frequency vector given an observation vector for the two candidate parents (equation 3.4). The prior distribution for the genotype frequency vector in equation 3.5 assumes uniform support over all possible genotypes. This might not always be a sensible prior if allele frequencies are highly uneven; a rare allele homozygote is much less likely than either heterozygotes or homozygotes with common alleles.

However, other aspects of the genetics of the population sampled such as allele frequency and inbreeding / outbreeding estimates can be derived using the methods employed in chapter 2. This can provide an informed prior for parental genotypes in the parentage analysis. Indeed, we can treat the allele frequency vector, $\mathbf{f}$, and the inbreeding coefficient, F_{IS}, as parameters of a hyperprior for the genotypes using the relationship described in equation 2.6, to be estimated as part of the parentage analysis process.

This type of analysis also allows for the assessment of parentage for samples that are missing data at loci. Note here that 'missing data' refers to samples that are known not to been taken and does not cover samples that are homozygous for an allele that does not amplify or for situations where genotyping error results in a band absence. Typically, loci with missing data must be excluded from the analysis which is wasteful, particularly if these loci are informative and have the power to exclude parent pairs for which samples have been taken. It is also not enough to ignore those loci that are missing in the calculation of likelihoods; when comparing two candidate parents as possible parents for an offspring, using a differing number of loci creates a preference bias for the potential parent with less usable loci. This is because, provided that extra loci are not too error prone, an increase in the number of loci results in an increase in the discrimination power of the data set, decreasing likelihoods. However, with extra information pertaining to allele frequencies and zygosity of the population of study, it is possible to include an extra step for the imputation of unobserved genotypes at particular loci (see Hruschka Jr. *et al.*, 2007; Schafer & Graham, 2002). Moreover, this method may provide a mechanism for the calculation of the probability that parentage for an individual lies outside the pool of sampled parents, from a random individual drawn from a population with the same allele frequencies and zygosity as the sampled population.

6.2.2 Dispersal Studies and Parentage Allocation

A critical aspect of a species biology is its ability to disperse. The dispersal ability of a species is often described in terms of its dispersal kernel, a probability density function of dispersal distances from the point of origin (see Levin & Kerster, 1974; Ribbens *et al.*, 1994; Cousens *et al.*, 2008). Genetic information can be brought to bear on studies of dispersal dynamics by providing information to help link offspring individuals to their parents (Pairon *et al.*, 2006). Robledo-Arnuncio & Garca (2007) describe a maximum-likelihood approach to fitting dispersal kernels when the source individual is known. However, exclusion methods of paternity analysis cannot always provide one unambiguous parent pair and, once observation error is taken into account, there will always be some level of uncertainty attached to any parentage assignment.

We can incorporate the fitting of a dispersal kernel by extending the likelihood function of equation 3.1 to include spatial information. If we define x_i, y_i, x_m, and y_m as the x and y coordinates of individual i and putative mother m respectively then the new, full likelihood, becomes

$$\mathbb{P}\left(\mathbf{O}_i, x_i, y_i | \varphi_i = m, \sigma_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}, \boldsymbol{\alpha}_{\varphi}, x_m, y_m\right) = \mathbb{P}\left(\mathbf{O}_i | \varphi_i = m, \sigma_i = f, \mathbf{O}_m, \mathbf{O}_f, \boldsymbol{\beta}\right) \mathbb{P}\left(x_i, y_i | \boldsymbol{\alpha}_{\varphi}, x_m, y_m\right) \quad (6.1)$$

where $\boldsymbol{\alpha}_{\varphi}$ is a vector of parameters for the dispersal kernel to be fitted. The dispersal kernel is described by the probability density function $\mathbb{P}\left(x_i, y_i | \boldsymbol{\alpha}_{\varphi}, x_m, y_m\right)$, which, in the isotropic case, is simply a function of the Euclidean distance between the putative mother and the offspring, $\sqrt{(x_i - x_m)^2 + (y_i - y_m)^2}$. All other notation can be found in the description for equation 3.1 in chapter 3.

Equation 6.1 can be used as the likelihood component of a Bayesian analysis for the calculation of the joint posterior of parentage assignment and dispersal capabilities in the light of genetic observation and spatial location. This integrated analysis allows not only the propagation of uncertainty in parentage assignment to estimates of dispersal kernel parameters but actually allows spatial data to inform paternity assignment also (Hadfield *et al.*, 2006).

6.2.3 Patterns of Parentage and Breeding Success

A common application for methods of parentage analysis is to calculate the individual breeding success for incorporation into models that can link the estimated breeding success to other variables, such as age (Vanpé *et al.*, 2009a) or territory size (Vanpé *et al.*, 2009b). However, most studies do not take into account the uncertainty related to the parentage assignment. Models of breeding success will certainly specify an error structure as part of their definition, and this may go some way to implicitly capturing the variability in the data set, but this is not the same thing as a full-blown observation model that will explicitly address these uncertainties. The Bayesian nature of the analysis method outlined in chapter 3 makes the technique amenable to extension to include the parameterisation and assessment of models of breeding success.

If we define the vectors ♀ and ♂ as random vectors of putative mothers and fathers respectively, with each element, $♀_i$ or $♂_i$, defining a possible parentage combination for offspring i, then it is possible to express a joint-likelihood function for the observation vector of the entire set of offspring genotypes:

$$\mathbb{P}\left(\mathbf{O}|\boldsymbol{\beta},\boldsymbol{\gamma}\right) = \sum_{♀}\sum_{♂}\mathbb{P}\left(♀,♂|\boldsymbol{\gamma}\right)\prod_{i}\mathbb{P}\left(\mathbf{O}_i|♀_i,♂_i,\mathbf{O}_{♀_i},\mathbf{O}_{♂_i},\boldsymbol{\beta}\right) \tag{6.2}$$

Here $\boldsymbol{\gamma}$ is a vector of parameters for the model of reproductive success for which $\mathbb{P}\left(♀,♂|\boldsymbol{\gamma}\right)$ represents the likelihood function of the reproduction sub model. $\mathbb{P}\left(\mathbf{O}_i|♀_i,♂_i,\mathbf{O}_{♀_i},\mathbf{O}_{♂_i},\boldsymbol{\beta}\right)$ is the likelihood function for the paternity assignment of offspring i, as defined in equation 3.1. This likelihood function can form the basis of joint Bayesian analysis of parentage and models of reproductive success. In this sense of a hierarchical model we have specified the parameters of the breeding model, $\boldsymbol{\gamma}$, as parameters for the prior expectation of the vector of joint parentage.

This basic specification provides an extension to the parentage analysis methodology outlined in chapter 3 to provide the joint estimation of parentage and the parameters of genotype observation and breeding success. It is worth noting that under that the extension to the modelling structure described here does not, in its basic form, provide an extra mechanism to add extra data to the inference. This is unlike the extension of parentage model to incorporate dispersal dynamics, described above, where extra spacial information forms a key part of the inference. The extra 'information' here comes in the form of the more detailed structure of the model. This extension does however serve to place biologically reasonable restrictions based on expectations of breeding success on allowable combinations of parentage.

6.2.4 Putting it all Together

We have shown that it is possible to generate estimates of allele frequency and inbreeding level for any real-world population and we have discussed how these estimates can be used to generate indices of population differentiation and structure whilst preserving any uncertainty in these estimates. We have described methods for the assignment of parentage and we have proposed ways in which these methods can be extended to allow estimates of dispersal ability and individual breeding success. Going further, it is possible to link sophisticated models of breeding success to the methods of parentage analysis described in chapter 3, that allow fecundity to vary with extra factors such as the environment, territory conditions, local density, and competition.

One example of where such an approach may be particularly fruitful is in the example of the Australian cane toad. Previous attempts to predict the potential distribution of this invasive species have involved the fitting of climate niche models to occurrence data (see van Beurden, 1981; Sutherst *et al.*, 1996, for early examples). Unfortunately, these approaches are limited by the fact the distribution of the cane toads is not in equilibrium with the environment and they continue to invade into new regions of climate in which they had not previously been observed. This has resulted in very poor predictive success.

More recently, the development of models based on the ecophysiological tolerances of the cane toad have been developed (Kearney *et al.*, 2008), allowing for the generation of surfaces describing the expected fecundity and dispersal of cane toads in different climatic regions. It would be possible to use parentage analysis (using the methods described in chapters 2 and 3) in different climatic regions to refine these estimates and produce a more concrete statistical link between fecundity and the environment. Moreover, this data could be used to supplement the existing telemetry data (see Phillips *et al.*, 2007) in order to derive accurate estimates of local cane toad dispersal (using the methods from chapter 4). Finally, The dispersal and fecundity parameters could form the basis of a metapopulation model (where the models of chapter 5 would need to be employed) to produce a model able to predict cane toad range dynamics at the macroecological scale (Phillips *et al.*, 2008).

Examples such as this demonstrate that low-level descriptions of life history parameters, such as the breeding success and dispersal ability, can form the building blocks for higher-level

individual and population based modelling (Phillips *et al.*, 2008). With estimates for these parameters it is possible to build larger scale models and make predictions of macroecological patterns, drawing inference from data collected at multiple scales using techniques such as those described in chapter 4 and applied in using the techniques in chapter 5. This thesis does not purport to have achieved such an epic synthesis. The methods described in chapters 2, 3, 4 and 5, do however lay the foundations for such scaling by providing mechanisms through which uncertainty and error can be integrated into inference at higher levels of the modelling hierarchy. Robustly propagating uncertainty is an essential part of forming good ecological predictions.

Bibliography

Andrieu, C., Doucet, A. & Holenstein, R. (2010) Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B* **72**, 269–342.

Bacles, C.F.E. & Ennos, R.A. (2008) Paternity analysis of pollen-mediated gene flow for *Fraxinus excelsior* L. in a chronically fragmented landscape. *Heredity* **101**, 368.

Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* **55**, 370–418.

Beaulieu, J. & Simon, J.P. (1994) Genetic structure and variability in *Pinus strobus* in Québec. *Canadian Journal of Forest Research* **24**, 1726–1733.

Beaumont, M.A., Zhang, W. & Balding, D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.

Beissinger, S.R. & Westphal, M.I. (1998) On the use of demographic models of population viability in endangered species management. *Journal of Wildlife Management* **62**, 821–841.

Bennett, V.J., Beard, M., Zollner, P.A., Fernández-Juricic, E., Westphal, L. & LeBlanc, C.L. (2009) Understanding wildlife responses to human disturbance through simulation modelling: a management tool. *Ecological Complexity* **6**, 113–134.

Beverton, R.J.H. & Holt, S.J. (1957) *On the Dynamics of Exploited Fish Populations*. Chapman and Hall, London.

Bevington, P.R. & Robinson, D.K. (2002) *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill.

Bonin, A., Ehrich, D. & Manel, S. (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology* **16**, 3737–3758.

Bos, D.H., Williams, R.N., Gopurenko, D., Bulut, Z. & Dewoody, J.A. (2009) Condition-dependent mate choice and a reproductive disadvantage for MHC-divergent male tiger salamanders. *Molecular Ecology* **18**, 3307–3315.

Botsford, L.W., White, J.W., Coffroth, M.A., Paris, C.B., Planes, S., Shearer, T.L., Thorrold, S.R. & Jones, G.P. (2009) Connectivity and resilience of coral reef metapopulations in marine protected areas: matching empirical efforts to predictive needs. *Coral Reefs* **28**, 327–337.

Bouteiller, C. & Perrin, N. (2000) Individual reproductive success and effective population size in the greater white-toothed shrew *Croeidura russula*. *Proceedings of the Royal Society of London, Series B* **7**, 701–705.

Bowling, A.T., Eggleston-Stott, M.L., Byrns, G., Clark, R.S., Dileanis, S. & Wichim, E. (1997) Validation of microsatellite markers for routine horse parentage testing. *Animal Genetics* **28**, 247–252.

Brännström, Å. & Sumpter, D.J.T. (2006) Stochastic analogues of deterministic single-species population models. *Theoretical Population Biology* **69**, 442–451.

Brookes, A.J. (1999) The essence of SNPs. *Gene* **234**, 177–186.

Brookes, M.I., Graneau, Y.A., King, P., Rose, O.C., Thomas, C.D. & Mallet, J.L.B. (1997) Genetic analysis of founder bottlenecks in the rare British butterfly *Plebejus argus*. *Conservation Biology* **11**, 648–661.

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

Broquet, T. & Petit, E. (2004) Quantifying genotyping erros in non invasive population genetics. *Molecular Ecology* **13**, 3601–3608.

Broquet, T. & Petit, E.J. (2009) Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology, Evolution, and Systematics* **40**, 193–216.

Brown, N., Gerard, F. & Fuller, R. (2002) Mapping of land use classes within the CORINE land cover map of Great Britain. *Cartographic Journal* **39**, 5–14.

Bullock, J.M. & Clarke, R.T. (2000) Long distance seed dispersal by wind: measuring and modelling the tail of the curve. *Oecologia* **124**, 506–521.

Burnham, K.P. & Anderson, D.R. (2001) Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research* **28**, 111–119.

Burton, O.J. & Travis, J.M.J. (2008) Landscape structure and boundary effects determine the fate of mutations occurring during range expansions. *Heredity* **101**, 329–340.

Butler, K. & Stephens, M. (1993) The distribution of a sum of binomial random variables. Technical Report 467, Office of Naval Research.

Caliceti, E., Meyer-Hermann, M., Ribeca, P., Surzhykov, A. & Jentschura, U.D. (2007) From useful algorithms for slowly convergent series to physical predictions based on divergent perturbative expansions. *Physics Reports* **446**, 1–96.

Canham, C.D. & Uriarte, M. (2006) Analysis of neighborhood dynamics of forest ecosystems using likelihood methods and modeling. *Ecological Applications* **16**, 62–73.

Carlson, C.S., Smith, J.D., Stanaway, I.B., Rieder, M.J. & Nickerson, D.A. (2006) Direct detection of null alleles in SNP genotyping data. *Human Molecular Genetics* **15**, 1931–1937.

Casella, G. & George, E.I. (1992) Explaining the Gibbs sampler. *The American Statistician* **46**, 167–174.

Castro, J., Pino, A., Hermida, M., Bouza, C., Chavarrís, D., Merino, P., Sánchez, L. & Martínez, P. (2007) A microsatellite marker tool for parentage assessment in gilthead seabream (*Sparus aurata*). *Aquaculture* **272 S1**, S210–S216.

Cercueil, A., Bellemain, E. & Manel, S. (2002) PARENTE: computer program for parentage analysis. *Journal of Heredity* **93**, 458–459.

Chakraborty, R., De Andrade, M., Daiger, S.P. & Budowle, B. (1992) Apparent heterozygote dificiencies observed in DNA typing data and their implications in forensic applications. *Annals of Human Genetics* **56**, 45–57.

Chakraborty, R. & Nei, M. (1977) Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. *Evolution* **31**, 347–356.

Chakraborty, R., Shaw, M. & Schull, W.J. (1974) Exclusion of paternity: the current state of the art. *American Journal of Human Genetics* **26**, 477–488.

Chapman, D.S., Dytham, C. & Oxford, G.S. (2007) Modelling population redistribution in a leaf beetle: an evaluation of alternative dispersal functions. *Journal of Animal Ecology* **76**, 36–44.

Chesson, P. & Lee, C.T. (2005) Families of discrete kernels for modeling dispersal. *Theoretical Population Biology* **67**, 241–256.

Chib, S. & Greenberg, E. (1995) Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**, 327–335.

Cifuentes, L.O., Martínez, E.H., Acuña, M.P. & Jonquera, H.G. (2006) Probability of exclusion in paternity testing: time to reassess. *Journal of Forensic Science* **51**, 349–350.

Clark, J.S. & Gelfand, A.E. (2006) *Hierarchical Modelling for the Environmental Sciences.* Oxford University Press.

Clark, J.S., Macklin, E. & Wood, L. (1998) Stages and spatial scales of recruitment limitiation in southern Appalachian forests. *Ecological Monographs* **68**, 213–235.

Clark, J.S., Silman, M., Kern, R., Macklin, E. & HilleRisLambers, J. (1999) Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology* **80**, 1475–1494.

Clarke, L.A., Rebelo, C.S., Goncalves, J., Boavida, M.G. & Jordan, P. (2001) PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Molecular Pathology* **54**, 351–353.

Codling, E.A., Plank, M.J. & Benhamou, S. (2008) Random walk models in biology. *Journal of the Royal Society Interface* **5**, 813–834.

Cody, W.J. (1993) Algorithm 715: SPECFUN - a portable FORTRAN package of special function routines and test drivers. *ACM Transactions on Mathematical Software* **19**, 22–32.

Collingham, Y.C., Hill, M.O. & Huntley, B. (1996) The migration of sessile organisms: a simulation model with measurable parameters. *Journal of Vegetation Science* **7**, 831–846.

Cornuet, J.M. & Luikart, G. (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001–2014.

Cousens, R., Dytham, C. & Law, R. (2008) *Dispersal in Plants: A Population Pespective*, chap. 5: Patterns of dispersal from entire plants, pp. 77–110. Oxford University Press, Oxford.

Dahl, F. & Willebrand, T. (2005) Natal dispersal and adult home ranges and site fidelity of mountain hares *Lepus timidus* in the boreal forest of Sweden. *Wildlife Biology* **11**, 309–317.

Dakin, E.E. & Avise, J.C. (2004) Microsatellite null alleles in parentage analysis. *Heredity* **93**, 504–509.

Danzmann, R.G. (1997) PROBMAX: A computer program for assigning unknown parentage in pedigree analysis from known genotypic pools of parents and progeny. *Journal of Heredity* **88**, 333.

Davis, P.J. & Rabinowitz, P. (2007) *Methods of Numerical Integration*. Dover Publications, New York, 2nd edition edn.

de Sousa, S.N., Finkeldey, R. & Gailing, O. (2005) Experimental verification of microsatellite null alleles in norway spruce (*Picea abis* [l.] karst): implications for population genetic studies. *Plant Molecular Biology Reporter* **23**, 113–119.

DeAngelis, D.L. & Mooij, W.M. (2003) *Models in Ecosystem Science*, chap. In praise of mechanistically rich models, pp. 63–82. Princeton University Press, Princeton.

DeAngelis, D.L., Rose, K.A. & Huston, M.A. (1994) *Frontiers in Mathematical Biology*, chap. Individual-oriented approaches to modeling ecological populations and communities, pp. 390–410. Springer, New York.

Del Moral, P., Doucet, A. & Jasra, A. (2006) Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B* **68**, 411–436.

Devlin, B., Roeder, K. & Ellstrand, N.C. (1988) Fractional paternity assignment: theoretical development and comparison to other methods. *Theoretical and Applied Genetics* **76**, 369–380.

Diggle, P.J. (2003) *Statistical Analysis of Spatial Point Patterns*. Arnold.

Doerksen, T.K. & Herbinger, C.M. (2008) Male reproductive success and pedigree error in red spruce open-pollinated and polycross mating systems. *Canadian Journal of Forest Research* **38**, 1742–1749.

Donnelly, P. & Tavaré, S. (1995) Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29**, 401–421.

Duchesne, P., Godbout, M.H. & Bernatchez, L. (2002) PAPA (package for the analysis of parental allocation): a computer program for simulated and real parental allocation. *Molecular Ecology Notes* **2**, 191–193.

Dytham, C. (2009) Evolved dispersal strategies at range margins. *Proceedings of the Royal Society B* **276**, 1407–1413.

Dytham, C. & Travis, J.M.J. (2006) Evolving dispersal and age at death. *Oikos* **113**, 530–538.

Efombagn, M.I.B., Sounigo, O., Eskes, A.B., Motamayor, J.C., Manzanares-Dauleux, M.J., Schnell, R. & Nyassé, S. (2009) Parentage analysis and outcrossing patterns in cacao (*Theobroma cacoa* L.) farms in Cameroon. *Heredity* **103**, 46–53.

Ellegren, H. (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics* **16**, 551–558.

Emigh, T.H. (1980) A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics* **36**, 627–642.

Engels, W.R. (2009) Exact tests for Hardy-Weinberg proportions. *Genetics* **183**, 1431–1441.

Ewen, K.R., Bahlo, M. & Treloar, S.A. (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics* **67**, 727–736.

Excoffier, L., Smouse, P.E. & Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Fernandes, L., Rocheta, M., Cordeiro, J., Pereira, S., Gerber, S., Oliveira, M.M. & Ribeiro, M.M. (2008) Genetic variation, mating patterns and gene flow in a *Pinus pinaster* aiton clonal seed orchard. *Annals of Forest Science* **65**, article 706.

Fisher, N.I. (1993) *Statistical Analysis of Circular Data*. Cambridge University Press.

Fisher, R.A. (1922) On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society of London A* **222**, 309–368.

Fisher, R.A. (1949) *The Theory of Inbreeding*. Oliver and Boyd, Edinburgh.

Fishman, G. (1996) *Monte Carlo: Concepts, Algorithms and Applications*. Springer, New York.

Foll, M., Beaumont, M.A. & Gaggiotti, O. (2008) An approximate bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study population structure. *Genetics* **179**, 927–939.

Fujiwara, M., Anderson, K.E., Neuberg, M.G. & Caswell, H. (2006) On the estimation of dispersal kernels from individual mark-recapture data. *Environmental and Ecological Statistics* **13**, 183–197.

Fuller, R.M., Smith, G.M., Sanderson, J.M., Hill, J.M. & Thompson, A.G. (2002) The UK land cover map 2000: construction of a parcel-based vector map from satellite images. *Cartographic Journal* **39**, 15–25.

Gagneux, P., Boesch, C. & Woodruff, D.S. (1997) Microsatellite errors associated with non invasive genotyping based on nuclear DNA amplified from shed hairs. *Molecular Ecology* **6**, 861–868.

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Gerber, S., Chabrier, P. & Kremer, A. (2003) FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. *Molecular Ecology Notes* **3**, 479–481.

Gerber, S., Mariette, S., Streiff, R., Bodénès, C. & Kremer, A. (2000) Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Molecular Ecology* **9**, 1037–1048.

Geweke, J. (1991) Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. *Computing Science and Statistics: Proceedings of the 23rd Symposium in the Interface* (ed. S.M. Kaufman), Interface Foundation of North America, Fairfax, Virginia.

Gilks, W.R. & Berzuini, C. (2001) Following a moving target: Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society B* **63**, 127–146.

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*, chap. Introducing Markov chain Monte Carlo, pp. 1–19. Chapman and Hall.

Gilpin, M.E. & Ayala, F.J. (1973) Global models of growth and competition. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3590–3593.

Ginot, F., Bordelais, I., Nguyen, S. & Gyapay, G. (1996) Correction of some genotyping errors in automated fluorescent microsatellite analysis by enzymatic removal of one base overhangs. *Nucleic Acids Research* **24**, 540–541.

Gompertz, B. (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London* **115**, 513–585.

Goodman, L.A. (1960) On the exact variance of products. *Journal of the American Statistical Association* **55**, 708–713.

Gopurenko, D., Williams, R.N. & DeWoody, J.A. (2007) Reproductive and mating success in the small-mouthed salamander (*Ambystoma texanum*) estimated via microsatellite parentage analysis. *Evolutionary Biology* **34**, 130–139.

Goudet, J., Raymond, M., de Meeüs, T. & Rousset, F. (1996) Testing differentiation in diploid populations. *Genetics* **144**, 1933–1940.

Green, P.J. (1995) Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Greene, D.F., Canham, C.D., Coates, K.D. & Lepage, P.T. (2004) An evaluation of alternative dispersal functions for trees. *Journal of Ecology* **92**, 758–766.

Grelaud, A., Robert, C.P., Marin, J.M., Rodolphe, F. & Taly, J.F. (2009) ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis* **4**, 317–336.

Grimm, V., Frank, K., Jeltsch, F., Brandl, R., Uchmański, J. & Wissel, C. (1996) Pattern-orientated modelling in population ecology. *Science of the Total Environment* **183**, 151–166.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.H., Weiner, J., Wiegand, T.F. & DeAngelis, D.L. (2005) Pattern-orientated modeling of agent-based complex systems: Lessons from ecology. *Science* **310**, 987–991.

Grimm, V., Wyszomirski, T., Aikman, D. & Uchmański, J. (1999) Individual-based modelling and ecological theory: synthesis of a workshop. *Ecological Modelling* **115**, 275–282.

Guo, S.W. & Thompson, E.A. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372.

Haccou, P., Jagers, P., Vatutin, V.A. & Dieckmann, U. (2007) *Branching Processes: Variation, Growth, and Extinction of Populations*. Cambridge University Press, Cambridge.

Hadfield, J.D., Richardson, D.S. & Burke, T. (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular Ecology* **15**, 3715–3730.

Haig, S. (1998) Molecular contributions to conservation. *Ecology* **79**, 413–425.

Hanski, I. (1994) A practical model of metapopulation dynamics. *Jounal of Animal Ecology* **63**, 151–162.

Hanski, I., Alho, J. & Moilanen, A. (2000) Estimating the parameters of survival and migration of individuals in metapopulations. *Ecology* **81**, 239–251.

Hardy, G.H. (1908) Mendelian proportions in a mixed population. *Science* **28**, 49–50.

Harrison, P.J., Buckland, S.T., Thomas, L., Harris, R., Pomeroy, P.P. & Harwood, J. (2006) Incorperating movement into models of grey seal population dynamics. *Journal of Animal Ecology* **75**, 634–645.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Higgins, S.I. & Richardson, D.M. (1999) Predicting plant migration rates in a changing world: the role of long-distance dispersal. *American Naturalist* **153**, 464–475.

Hill, W.G. (1972) Effective size of populations with overlapping generations. *Theoretical Population Biology* **3**, 278–289.

Hill, W.G. & Weir, B.S. (2004) Moment estimation of population diversity and genetic distance from data on recessive markers. *Molecular Ecology* **13**, 895–908.

Hodgkinson, A. & Eyre-Walker, A. (2010) Human triallalic sites: evidence for a new mutational mechanism? *Genetics* **184**, 233–241.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999) Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382–401.

Hoffman, J.I. & Amos, W. (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology* **14**, 599–612.

Hogeweg, P. (1988) Cellular automata as a paradigm for ecological modeling. *Applied Mathematics and Computation* **27**, 81–100.

Holland, E.P., Aegerter, J.N., Dytham, C. & Smith, G.C. (2007) Landscape as a model: the importance of geometry. *PLOS Computational Biology* **3**, 1979–1992.

Holsinger, K.E., Lewis, P.O. & Dey, D.K. (2002) A Bayesian approach to inferring population structure from dominant markers. *Molecular Ecology* **11**, 1157–1164.

Holsinger, K.E. & Weir, B.S. (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics* **10**, 639–650.

Horrace, W.C. (2005) Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis* **94**, 209–221.

Hovestadt, T., Messner, S. & Poethke, H.J. (2001) Evolution of reduced dispersal mortality and 'fat-tailed' dispersal kernels in autocorrelated landscapes. *Proceedings of the Royal Society, Biological Sciences* **268**, 385–391.

Hruschka Jr., E.R., Hruschka, E.R. & Ebecken, N.F. (2007) Bayesian networks for imputation in classification problems. *Journal of Intelligent Information Systems* **29**, 231–252.

Hudson, R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 183–201.

Hüebner, C., Petermann, I., Browning, B.L., Shelling, A.N. & Ferguson, L.R. (2007) Triallelic Single Nucleotide Polymorphisms and genotyping error in genetic epidemiology studies: MDR1 (ABCB1) G2677/T/A as an example. *Cancer Epidemiology, Biomarkers and Prevention* **16**, 1185–1192.

Hull, D. (1974) *The Philosophy of Biological Science*. Prentice-Hall.

Hünenberger, P.H. & McCammon, J.A. (1999) Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: a continuum electrostatics study. *Biophysical Chemistry* **78**, 69–88.

Isabel, N., Beaulieu, J. & Bousqet, J. (1995) Complete congruence between gene diversity estimates derived from genotypic data at enzyme and random amplified polymorphic DNA loci in black spruce. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 6369–6373.

Isabel, N., Beaulieu, J., Thériault, P. & Bousqet, J. (1999) Direct evidence for biased gene diversity estimates from dominant random amplified polymorphic DNA (RAPD) fingerprints. *Molecular Ecology* **8**, 477–483.

Jackson, C.H., Best, N.G. & Richardson, S. (2009) Bayesian graphical models for regression on multiple data sets with different variables. *Biostatistics* **10**, 335–351.

Jeltsch, F., Milton, S.J., Dean, W.R.J. & van Rooyen, N. (1996) Tree spacing and coexistence in semiarid savannas. *Journal of Ecology* **84**, 583–595.

Jerry, D.R., Preston, N.P., Crocos, P.J., Keys, S., Meadows, J.R.S. & Li, Y. (2004) Parentage determination of Kuruma shrimp *Penaeus (Marsupenaeus) japonicus* using microsatellite markers (bate). *Aquaculture* **235**, 237–247.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1997) *Discrete Multivariate Distributions*, chap. Chapter 35: Multinomial Distributions, pp. 31–92. Wiley-Interscience.

Jones, A.G. & Ardren, W.R. (2003) Methods of parentage analysis in natural populations. *Molecular Ecology* **12**, 2511–2523.

Jones, A.G., Small, C.M., Paczolt, K.A. & Ratterman, N.L. (2010) A practical guide to methods of parentage analysis. *Molecular Ecology Resources* **10**, 6–30.

Jones, B. (2003) Balancing population size and genetic information in parentage analysis studies. *Biometrics* **59**, 694–700.

Jones, O.R. & Wang, J. (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources* **10**, 551–556.

Kalinowski, S.T., Taper, M.L. & Marshall, T.C. (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* **16**, 1099–1106.

Kass, R. & Raftery, A. (1995) Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kearney, M., Phillips, B.J., Tracy, C.R., Christian, K.A., Betts, G. & Porter, W.P. (2008) Modelling species distributions without using species distribution: the cane toad in Australia under current and future climates. *Ecography* **31**, 423–434.

Kendall, B.E. & Wittmann, M.E. (2010) A stochastic model for annual reproductive success. *American Naturalist* **175**, 461–468.

Khan, Z., Balch, T. & Dellaert, F. (2005) MCMC-based particle filtering for tracking a variable number of interacting targets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 1805–1819.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kimura, M. (1986) DNA and the neutral theory. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **312**, 343–354.

Kimura, M. & Weiss, G.H. (1964) The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576.

Kingman, J.F.C. (1982a) The coalescent. *Stochastic Processes and their Applications* **13**, 235–248.

Kingman, J.F.C. (1982b) *Exchangeability in Probability and Statistics*, chap. Exchangeability and the evolution of large populations, pp. 97–112. North-Holland Publishing Company, Amsterdam.

Kirst, M., Cordeiro, C.M., Rezende, D.S.P. & Grattapagla, D. (2005) Power of microsatellite markers for fingerprinting and parentage analysis in *Eucalyptus grandis* breeding populations. *Journal of Heredity* **96**, 161–166.

Kondoh, M. (2003) High reproductive rates result in high predation risks: a mechanism promoting the coexistence of competing prey in spatially structured populations. *American Naturalist* **161**, 299–309.

Kramer-Schadt, S., Revilla, E., Wiegand, T. & Breitenmoser, U. (2004) Fragmented landscapes, road mortality and patch connectivity: modelling influences on the dispersal of eurasian lynx. *Journal of Applied Ecology* **41**, 711–723.

Kramer-Schadt, S., Revilla, E., Wiegand, T. & Grimm, V. (2007) Patterns for parameters in simulation models. *Ecological Modelling* **204**, 553–556.

Krauss, S.L. (2000) Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. *Molecular Ecology* **9**, 1241–1245.

Krauss, S.L., He, T., Barrett, L.G., Lamont, B.B., Enright, N.J., Miller, B.P. & Hanley, M.E. (2009) Contrasting impacts of pollen and seed dispersal on spatial genetic structure in the bird-pollinated *Banksia hookeriana*. *Heredity* **102**, 274–285.

Kwok, S., Kellog, D.E., McKinney, N., Spasic, D., Goda, L., Levenson, C. & Sninsky, J.J. (1990) Effects of primer-template mismatches on the polymerase chain reaction : human immunodefiency virus type 1 model studies. *Nucleic Acids Research* **18**, 999–1005.

Law, R., Murrell, D.J. & Dieckmann, U. (2003) Population growth in space and time: spatial logisitc equations. *Ecology* **84**, 252–262.

Leberg, P.L. (1992) Effects of population bottlenecks on genetic diversity as measured by allozyme electrophoresis. *Evolution* **46**, 477–494.

Lebreton, J.D., Gosselin, F. & Niel, C. (2007) Extinction and viability of populations: paradigms and concepts of extinction models. *Écoscience* **14**, 472–481.

Lefkovitch, L.P. (1965) The study of population growth in organisms grouped by stages. *Biometrics* **21**, 1–18.

Lehmann, T., Hawley, W.A. & Collins, F.H. (1996) An evaluation of evolutionary constraints on microsatellite loci using null alleles. *Genetics* **144**, 1155–1163.

Leslie, P.H. (1945) On the use of matrices in certain population mathematics. *Biometrika* **33**, 183–212.

Leslie, P.H. (1948) Some further nortes on the use of matrices in population mathematics. *Biometrika* **35**, 213–245.

Levin, D.A. & Kerster, H.W. (1974) Gene flow in seed plants. *Evolutionary Biology* **7**, 139–220.

Liu, J. & West, M. (2001) *Sequential Monte Carlo Methods in Practice*, chap. Combined parameter and state estimation in simulation-based filtering, pp. 197–223. Springer-Verlag, Berlin.

Łomnicki, A. (1999) Individual-based models and the individual-based approach to population ecology. *Ecological Modelling* **115**, 191–198.

Lynch, M. & Milligan, B.G. (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology* **3**, 91–99.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. (2003) Markov Chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15324–15328.

Marker, L.L., Wilkerson, A.J.P., Sarno, R.J., Martenson, J., Breitenmoser-Wuersten, C., O'Brien, S.J. & Johnson, W.E. (2008) Molecular genetic insights on cheetah (*Acinonyx jubatus*) ecology and conservation in Namibia. *Journal of Heredity* **99**, 2–13.

Marshall, T.C., Slate, J., Kruuk, L.E.B. & Pemberton, J.M. (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology* **7**, 639–655.

Martin, J.K., Handasyde, K.A. & Taylor, A.C. (2007) Linear roadside remnants: their influence on den-use, home range and mating system in bobucks (*Trichosurus cunninghami*). *Austral Ecology* **32**, 686–696.

Martin, P. (1993) Vegetation responses and feedbacks to climate: a review of models and processes. *Climate Dynamics* **8**, 201–210.

Matson, S.E., Camara, M.D., Eichert, W. & Banks, M.A. (2008) P-LOCI: a computer program for choosing the most efficient set of loci for parentage assignment. *Molecular Ecology Resources* **8**, 765–768.

McEachern, M.B., McElreath, R.L., Van Vuren, D.H. & Eadie, J.M. (2009) Another genetically promiscuous 'polygynous' mammal: mating system variation in *Neotoma fuscipes*. *Animal Behaviour* **77**, 449–455.

McLean, J.E., Seamons, T.R., Dauer, M.B., Bentzen, P. & Quinn, T.P. (2008) Variation in reproductive success and effective number of breeders in a hatchery population of steelhead trout (*Oncorhynchus mykiss*): examination by microsatellite-based parentage analysis. *Conservation Genetics* **9**, 295–304.

Meagher, T.R. (1986) Analysis of paternity with a natural population of *Chamaelirium luteum*. 1. indentification of most-likely male parents. *The American Naturalist* **128**, 199–215.

Meagher, T.R. & Thompson, E. (1986) The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology* **29**, 87–106.

Meagher, T.R. & Thompson, E. (1987) Analysis of parentage for naturally established seedling of *Chamaelirium luteum* (Liliaceae). *Ecology* **68**, 803–812.

Mengersen, K.L., Robert, C.P. & Guihenneuc-Jouyaux, C. (1999) *Bayesian Statistics 6*, chap. MCMC convergence diagnosis: A revieWWW, pp. 415–440. Oxford University Press, New York.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Meudt, H.M. & Clarke, A.C. (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends in Plant Science* **12**, 106–117.

Milligan, B.G. & McMurry, C.K. (1993) Dominant vs codominant genetic markers in the estimation of male mating success. *Molecular Ecology* **2**, 275–283.

Milne, B.T., Johnson, A.R., Keitt, T.H., Hatfield, C.A., David, J. & Hraber, P.T. (1996) Detection of critical densities associated with piñon-juniper ecotones. *Ecology* **77**, 805–821.

Mitchell, A.A., Cutler, D.J. & Chakravorty, A. (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission / disequilibrium test. *American Journal of Human Genetics* **72**, 598–610.

Moilanen, A. (2004) SPOMSIM: software for stochastic patch occupancy models of metapopulation dynamics. *Ecological Modelling* **179**, 533–550.

Moilanen, A., Franco, A.M.A., Early, R.I., Fox, R., Wintle, B. & Thomas, C.D. (2005) Prioritising multiple-use ladscapes for conservation: methods for large multi-species planning problems. *Proceedings of the Royal Society B* **272**, 1885–1891.

Moree, R. (1950) A modification of the Hardy-Weinberg law. *Science* **111**, 691–692.

Morin, P.A., Leduc, R.G., Archer, F.I., Martien, K.K., Huebinger, R., Bickham, S.W. & Taylor, B.L. (2009) Significant deviations from Hardy-Weinberg equilibrium caused by low levels of microsatellite genotyping errors. *Molecular Ecology Resources* **9**, 498–504.

Mugglestone, M.A. & Renshaw, E. (1996) A practical guide to the spectral analysis of spatial point processes. *Computational Statistics and Data Analysis* **21**, 43–65.

Murdoch, W.W., McCauley, E., Nisbet, R.M., Gurney, W.S.C. & de Roos, A.M. (1992) *Individual-based models and approaches in ecology*, chap. Individual-based models: combining testability and generality, pp. 18–35. Chapman and Hall.

Murrell, D.J. (2005) Local spatial structure and predator-prey dynamics: counterintuitive effects of prey enrichment. *American Naturalist* **166**, 354–367.

Murrell, D.J. (2010) When does local spatial structure hinder competitive coexistence and reverse competitive hierarchies? *Ecology* **91**, 1605–1616.

Murrell, D.J., Dieckmann, U. & Law, R. (2004) On moment closures for population dynamics in continuous space. *Journal of Theoretical Biology* **229**, 421–432.

Mustin, K., Benton, T.G., Dytham, C. & Travis, J.M.J. (2009) The dynamics of climate-induced range shifting; perspectives from simulation modelling. *Oikos* **118**, 131–137.

Nakamaru, M. (2006) Lattice models in ecology and social sciences. *Ecological Research* **21**, 364–369.

Nakanishi, A., Tomaru, N., Yoshimaru, H., Manabe, T. & Yamamoto, S. (2009) Effects of seed- and pollen-mediated gene dispersal on genetic structure among *Quercus salicina* saplings. *Heredity* **102**, 182–189.

Nash, J.C. (1990) *Compact Numerical Methods for Computers: Linear Algebra and Function Minimization*. Institute of Physics Publishing.

Nathan, R. & Muller-Landau, H.C. (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology and Evolution* **15**, 278–285.

Nei, M. (1972) Genetic distance between populations. *American Naturalist* **106**, 283–292.

Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 3321–3323.

Nei, M., Maruyama, T. & Chakraborty, R. (1975) The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10.

Neuhauser, C. & Pacala, S.W. (1999) An explicitly spatial version of the Lotka-Volterra model with interspecific competition. *The Annals of Applied Probability* **9**, 1226–1259.

Newman, K.B., Buckland, S.T., Lindley, S.T., Thomas, L. & Fernández, C. (2006) Hidden process models for animal population dynamics. *Ecological Applications* **16**, 74–86.

Nielsen, R., Mattila, D.K., Clapham, P.J. & Palsbøll, P.J. (2001) Statistical approaches to paternity analysis in natural populations and applications to the north Atlantic humpback whale. *Genetics* **157**, 1673–1682.

Nordborg, M. (2001) *Handbook of Statistical Genetics*, chap. Coalescent theory, pp. 179–212. John Wiley & Sons, Chichester.

Oh, M.S. & Berger, J.O. (1993) Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association* **88**, 450–456.

Opel, K.L., Chung, D. & McCord, B.R. (2010) A study of PCR inhibition mechanisms using real time PCR. *Journal of Forensic Sciences* **55**, 25–33.

O'Rourke, J. (1994) *Computational Geometry in C*. Cambridge University Press, Cambridge.

Ovaskainen, O. & Cornell, S.J. (2006) Space and stochasticity in population dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12781–12786.

Pairon, M., Jonard, M. & Jacquemart, A.L. (2006) Modeling seed dispersal of black cherry, an invasive forest tree: how microsatellites may help? *Canadian Journal of Forest Research* **36**, 1385–1394.

Pearl, R. (1917) Studies on inbreeding vii - some further considerations regarding the measurement and numerical expression of degrees of kinship. *American Naturalist* **51**, 545–559.

Peltonen, J., Venna, J. & Kaski, S. (2009) Visualizations for assessing convergence and mixing of Markov chain Monte Carlo simulations. *Computational Statistics and Data Analysis* **53**, 4453–4470.

Pemberton, J.M., Slate, J., Bancroft, D.R. & Barret, J.A. (1995) Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. *Molecular Ecology* **4**, 249–252.

Petrovskii, S., Morozov, A. & Li, B.L. (2008) On a possible origin of the fat-tailed dispersal in population dynamics. *Ecological Complexity* **5**, 146–150.

Phillips, B.J., Brown, G.P., Greenlees, M., Webb, J.K. & Shine, R. (2007) Rapid expansion of the cane toad (*B*ufo marinus) invasion front in tropical australia. *Austral Ecology* **32**, 169–176.

Phillips, B.J., Chipperfield, J.D. & Kearney, M.R. (2008) The toad ahead: challenges of modelling the range and spread of an invasive species. *Wildlife Research* **35**, 222–234.

Piotti, A., Leonardi, S., Piovani, P., Scalfi, M. & Menozzi, P. (2009) Spruce colonization at treeline: where do those seeds come from? *Heredity* **103**, 136–145.

Piry, S., Luikart, G. & Cornuet, J.M. (1999) BOTTLENECK: a computer program for detecting recent reductions in the effective population size using allele frequency data. *Journal of Heredity* **90**, 502–503.

Planes, S., Jones, G.P. & Thorrold, S.R. (2009) Larval dispersal connects fish populations in a network of marine protected areas. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 5693–5697.

Pradeep, D.S. & Hussain, F. (2004) Effects of boundary condition in numerical simulations of vortex dynamics. *Journal of Fluid Mechanics* **516**, 115–124.

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. & Feldman, M.W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791–1798.

Putman, R. & Wratten, S.D. (1984) *Principles of Ecology*, chap. 8: Competition and population stability, pp. 202–218. University of California Press, Berkeley and Los Angeles.

Rhoads, C.L., Bowman, J.L. & Eyler, B. (2010) Home range and movement rates of female exurban white-tailed deer. *Journal of Wildlife Management* **74**, 987–994.

Ribbens, E., Silander Jr., J.A. & Pacala, S.W. (1994) Seedling recruitment in forests: calibrating models to predict patterns of tree seedling dispersion. *Ecology* **75**, 1794–1806.

Ricker, W.E. (1954) Stock and recruitment. *Journal of the Fisheries Research Board of Canada* **11**, 554–623.

Ricker, W.E. (1975) *Computation and interpretation of biological statistics of fish populations.* No. 191 in Bulletin of the Fisheries Resources Board of Canada, Department of Fisheries and the Environment, Ottawa.

Robert, C.P. (1995) Simulation of truncated normal variables. *Statistics and Computing* **5**, 121–125.

Robledo-Arnuncio, J.J. & Garca, C. (2007) Estimation of the seed dispersal kernel from exact identification of source plants. *Molecular Ecology* **16**, 5098–5109.

Robledo-Arnuncio, J.J. & Gil, L. (2005) Patterns of pollen dispersal in a small population of *Pinus sylvestris* L. revealed by total-exclusion paternity analysis. *Heredity* **94**, 13–22.

Rosenberg, M.S. (2000) The bearing correlogram: a new method of analysing directional spatial auticirrelation. *Geographical Analysis* **32**, 267–278.

Rosenberg, M.S. (2004) Wavelet analysis for detecting anisotropy in point patterns. *Journal of Vegetation Science* **15**, 277–284.

Rosenberg, N.A. & Nordborg, M. (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics* **3**, 380–390.

Rosenzweig, M.L. (1971) Paradox of enrichment: destabilization of exploitative ecosystems in ecological time. *Science* **171**, 385–387.

Rourke, M.L., McPartlan, H.C., Ingram, B.A. & Taylor, A.C. (2009) Polygamy and low effective population size in a captive murray cod (*Maccullochella peelii peelii*) population: genetic implications for wild restocking programs. *Marine and Freshwater Research* **60**, 873–883.

Saenz-Agudelo, P., Jones, G.P., Thorrold, S.R. & Planes, S. (2009) Estimating connectivity in marine populations: an empirical evaluation of assignment tests and parentage analysis under different gene flow secenarios. *Molecular Ecology* **18**, 1765–1776.

Schafer, J.L. & Graham, J.W. (2002) Missing data: our view of the state of the art. *Psychological Methods* **7**, 147–177.

Schaid, D.J., Batzler, A.J., Jenkins, G.D. & Hildebrandt, M.A.T. (2006) Exact tests of hardy-weinberg equilibrium and homogeneity of disequilibrium across strata. *American Journal of Human Genetics* **76**, 1071–1080.

Schiffers, K., Schurr, F.M., Tielbrger, K., Urbach, C., Moloney, K. & Jeltsch, F. (2008) Dealing with virtual aggregation - a new index for analysing heterogeneous point patterns. *Ecography* **31**, 545–555.

Seamons, T.R. & Quinn, T.P. (2010) Sex-specific patterns of lifetime reproductive success in single and repeat breeding steelhead trout (*Oncorhynchus mykiss*). *Behavioral Ecology and Sociobiology* **64**, 505–513.

Shigesada, N., Kawasaki, K. & Takeda, Y. (1995) Modeling stratified diffusion in biological invasions. *The American Naturalist* **146**, 229–251.

Signorovitch, J. & Nielsen, R. (2002) PATRI - paternity inference using genetic data. *Bioinformatics* **18**, 341–342.

Silvertown, J., Holtier, S., Johnson, J. & Dale, P. (1992) Cellular automaton models of interspecific competition for space - the effect of pattern on process. *Journal of Ecology* **80**, 527–533.

Simmons, L.W., Beveridge, M. & Kennington, W.J. (2007) Polyandry in the wild: temporal changes in female mating frequency and sperm competition intensity in natural populations of the tittigoniid *Requena verticalis*. *Molecular Ecology* **16**, 4613–4623.

Simon, G. (1997) An angular version of spatial correlations, with exact significance tests. *Geographical Analysis* **29**, 267–278.

Sisson, S.A., Fan, Y. & Tanaka, M.M. (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1760–1765.

Škliba, J., Šumbera, R., Chitaukali, W.N. & Burda, H. (2009) Home-range dynamics in a solitary subterranean rodent. *Ethology* **115**, 217–226.

Slate, J., Marshall, T.C. & Pemberton, J.M. (2000) A retrospective assessment of the accuracy of the paternity inference program cervus. *Molecular Ecology* **9**, 801–808.

Slatkin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462.

Snäll, T., O'Hara, R.B. & Arjas, E. (2007) A mathematical and statistical framework for modelling dispersal. *Oikos* **116**, 1037–1050.

Söndgerath, D. & Schröder, B. (2002) Population dynamics and habitat connectivity affecting the spatial spread of populations - a simulation study. *Landscape Ecology* **17**, 57–70.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A. (2002) Measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* **64**, 583–639.

Spitze, K. (1993) Population structure in *daphnia obtusa*: quantitative genetic and allozymic variation. *Genetics* **135**, 367–374.

Stark, A.E. (1976) Generalization of Hardy-Weinberg law. *Nature* **259**, 44.

Stewart Jr, C.N. & Excoffier, L. (1996) Assessing population genetic structure and variability with RAPD data: Application to *Vaccinium macrocarpon* (American Cranberry). *Journal of Evolutionary Biology* **9**, 153–171.

Stow, A.J. & Sunnucks, P. (2004) High mate and site fidelity in Cunningham's skinks (*egernia cunninghami*) in natural and fragmented habitat. *Molecular Ecology* **13**, 419–430.

Sullivan, P.J. (1988) Effect of boundary conditions, region length, and diffusion rates on a spatially heterogenous predator-prey system. *Ecological Modelling* **43**, 235–249.

Suppes, P. (2007) Where do Bayesian priors come from? *Synthese* **156**, 441–471.

Sutherst, R.W., Floyd, R.B. & Maywald, G.F. (1996) The potential geographical distribution of the cane toad, *Bufo marinus L.* in Australia. *Conservation Biology* **10**, 294–299.

Swanack, T.M., Grant, W.E. & Forstner, M.R.J. (2009) Projecting population trends of endangered amphibian species in the face of uncertainty: A pattern-oriented approach. *Ecological Modelling* **220**, 148–159.

Szmidt, A.E., Wang, X.R. & Z., L.M. (1996) Empirical assessment of allozyme and RAPD variation in *Pinus sylvestris* (L.) using haploid tissue analysis. *Heredity* **76**, 412–420.

Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., Waits, L.P. & Bouvet, J. (1996) Reliable genotyping of samples with very low DNA quantiles using PCR. *Nucleic Acids Research* **24**, 3189–3194.

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

Tatarenkov, A., Healey, C.I.M., Grether, G.F. & Avise, J.C. (2008) Pronounced reproductive skew in a natural population of green swordtails, *Xiphophorus helleri*. *Molecular Ecology* **17**, 4522–4534.

Tavaré, S. (1984) Line-of-descent and genealogical processes, and their applications in population genetic models. *Theoretical Population Biology* **26**, 119–164.

Tavaré, S., Balding, D.J., Griffiths, R.C. & Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518.

Thomas, L., Buckland, S.T., Newman, K.B. & Harwood, J. (2005) A unified framework for modelling wildlife population dynamics. *Australian and New Zealand Journal of Statistics* **47**, 19–34.

Thomas, S.C. (2005) The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **360**, 1457–1467.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Annals of Statistics* **22**, 1701–1728.

Toni, T. & Stumpf, M.P.H. (2010) Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26**, 104–110.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M.P.H. (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* **6**, 187–202.

Topping, C.J., Hansen, T.S., Jensen, T.S., Jepsen, J.U., Nikolajsen, F. & Odderskær, P. (2003) ALMaSS, an agent-based model for animals in temperate European landscapes. *Ecological Modelling* **167**, 65–82.

Travis, J.M.J., Brooker, R.W. & Dytham, C. (2005) The interplay of positive and negative species interactions across an environmental gradient: insights from an individual-based simulation model. *Biology Letters* **1**, 5–8.

Travis, J.M.J. & Dytham, C. (2002) Dispersal evolution during invasions. *Evolutionary Ecology Research* **4**, 1119–1129.

Trenkel, V.M., Elston, D.A. & Buckland, S.T. (2000) Fitting population dynamics models to count and cull data using sequential importance sampling. *Journal of the American Statistical Association* **95**, 363–374.

Troendle, J.F. & Yu, K.F. (1994) A note on testing the Hardy-Weinberg law across strata. *Annals of Human Genetics* **58**, 397–402.

Tufto, J., Engen, S. & Hindar, K. (1997) Stochastic dispersal processes in plant populations. *Theoretical Population Biology* **52**, 16–26.

Turchin, P. (1998) *Quantitative Analysis of Movement*. Sinauer Associates.

Uller, T. & Olsson, M. (2008) Multiple paternity in reptiles: patterns and processes. *Molecular Ecology* **17**, 2566–2580.

van Beurden, E.K. (1981) Bioclimatic limits to the spread of *Bufo marinus* in Australia: a baseline. *Proceedings of the Ecological Society of Australia* **11**, 143–149.

van Moorter, B., Visscher, D., Benhamou, S., Börger, L., Boyce, M.S. & Gaillard, J.M. (2009) Memory keeps you at at home: a mechanistic model for home range emergence. *Oikos* **118**, 641–652.

Vanpé, C., Gaillard, J.M., Morellet, N., Kjellander, P., Liberg, O., Delorme, D. & Hewison, A.J.M. (2009a) Age-specific variation in male breeding success of a territorial ungulate species, the european roe deer. *Journal of Mammalogy* **90**, 661–665.

Vanpé, C., Morellet, N., Kjellander, P., Goulard, M., Liberg, O. & Hewison, A.J.M. (2009b) Access to mates in a territorial ungulate is determined by the size of a male's territory, but not by its habitat quality. *Journal of Animal Ecology* **78**, 42–51.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. & Zabeau, M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**, 4407–4414.

Waagepetersen, R. & Sorensen, D. (2001) A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping. *International Statistical Review* **69**, 49–61.

Wahlund, S. (1928) Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* **11**, 65–106.

Wahlund, S. (1975) *Demographic Genetics*, chap. Zusammensetzung von Population und Korrelationserscheinung vom Standtpunkt der Vererbunglehre aus betrachtet, pp. 224–263. Dowden, Hutchinson and Stroudsburg, Stroudsburg.

Wang, J.L. (2004) Sibship reconstruction from genetic data with typing errors. *Genetics* **166**, 1963–1979.

Watts, P.C., Thompson, D.J., Allen, K.A. & Kemp, S.J. (2007) How useful is DNA extracted from the legs of archived insects for microsatellite-based population genetic analyses? *Journal of Insect Conservation* **11**, 195–198.

Weber, J.L. & Wong, C. (1993) Mutation of human short tandem repeats. *Human Molecular Genetics* **2**, 1123–1128.

Weinberg, W. (1908) Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg* **64**, 368–382.

Weinberg, W. (1963) *Papers on Human Genetics*, chap. Über den Nachweis der Vererbung beim Menschen (On the demonstration of heredity in man), pp. 4–15. Prentice-Hall, New Jersey.

Weir, B.S. & Cockerham, C.C. (1984) Estimating f-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370.

Whitlock, R., Hipperson, H., Mannarelli, M., Butlin, R.K. & Burke, T. (2008) An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error. *Molecular Ecology Resources* **8**, 725–735.

Wiegand, T., Knauer, F., Kaczensky, P. & Naves, J. (2004) Expansion of brown bears (*Ursus arctos*) into the eastern Alps: a spatially explicit population model. *Biodiversity and Conservation* **13**, 79–114.

Wiegand, T.F., Jeltsch, F., Hanski, I. & Grimm, V. (2003) Using pattern-oriented modeling for revealing hidden information: A key for reconciling ecological theory and application. *Oikos* **100**, 209–222.

Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. (2005) A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 887–893.

Will, H. & Tackenburg, O. (2008) A mechanistic simulation model of seed dispersal by animals. *Journal of Ecology* **96**, 1011–1022.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. & Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* **18**, 6531–6535.

Williams, R.N. & DeWoody, J.A. (2009) Reproductive success and sexual selection in wild eastern tiger salamanders (*Ambystoma t. tigrinum*). *Evolutionary Biology* **36**, 201–213.

Willis, S.G., Thomas, C.D., Hill, J.K., Collingham, Y.C., Telfer, M.G., Fox, R. & Huntley, B. (2009) Dynamic distribution modelling: predicting the present from the past. *Ecography* **32**, 5–12.

Wilson, I.G. (1997) Inhibition and facilitation of nucleic acid amplification. *Applied and Environmental Microbiology* **63**, 3741–3751.

Wilson, M.F. (1993) Dispersal mode, seed shadows, and colonization patterns. *Vegetatio* **107/108**, 261–280.

Wilson, W.G., de Roos, A.M. & McCauley, E. (1993) Spatial instabilities within the diffusive Lotka-Volterra system: individual-based simulation results. *Theoretical Population Biology* **43**, 91–127.

Wojczulanis-Jakubas, K., Jakubas, D., Oigarden, T. & Lifjeld, J.T. (2009) Extrapair copulations are frequent but unsuccessful in a highly colonial seabird, the little auk, *Alle alle*. *Animal Behaviorur* **77**, 433–438.

Worthington Wilmer, J., Allen, P.J., Pomeroy, P.P., Twiss, S.D. & Amos, W. (1999) Where have all the fathers gone? An extensive microsatellite analysis of paternity in the grey seal (*Halichoerus grypus*). *Molecular Ecology* **8**, 1417–1429.

Wright, S. (1922) Coefficients of inbreeding and relationship. *American Naturalist* **56**, 330–338.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright, S. (1938) Size of population and breeding structure in relation to evolution. *Science* **87**, 430–431.

Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics* **15**, 323–354.

Zeyl, E., Aars, J., Ehrich, D. & Wiig, Ø. (2009) Families in space: relatedness in the Barents Sea population of polar bears (*Ursus maritimus*). *Molecular Ecology* **18**, 735–749.

Zheng, Y., Deng, D., Li, S. & Fu, J. (2010) Aspects of the breeding biology of the Omei mustache toad (*Leptobrachium boringii*): polygamy and paternal care. *Amphibia-Reptilia* **31**, 183–194.

Zhivotovsky, L.A. (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology* **8**, 907–913.

Zinck, R.D. & Grimm, V. (2009) Unifying wildfire models from ecology and statistical physics. *American Naturalist* **174**, E170–E185.