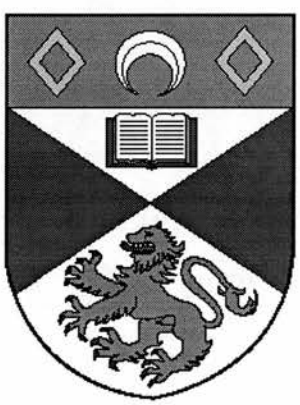
Spatial and temporal models, with applications in ornithology

Rachel M. Fewster



Thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

in the School of Mathematical and
Computational Sciences,

UNIVERSITY OF ST ANDREWS.

October, 1998.

Declarations

1. I, Rachel M. Fewster, hereby certify that this thesis, which is approximately 60,000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

   date 7/10/98 signature of candidate .

2. I was admitted as a research student in October 1995 and as a candidate for the degree of PhD in October 1996; the higher study for which this is a record was carried out in the University of St Andrews between 1995 and 1998.

   date 7/10/98 signature of candidate .

3. I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

   date 7/10/98 signature of supervisor

4. In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any *bona fide* library or research worker.

   date 7/10/98 signature of candidate .

Abstract

Statistical models for ecological populations can provide much information about the effect of factors such as habitat, climate and land management on population range and abundance. Insights gained through application of these models may be used in the formulation of effective management plans for the future conservation of species and their environment. Two types of models are presented, together with applications to ornithological data. The first provides a framework for the analysis of abundance trends in farmland species over the period 1962 to 1995, using data from the Common Birds Census of the British Trust for Ornithology. An approach based on generalized additive models is adopted, which extends previous analysis methods by providing a sound theoretical basis and uniting them under a single umbrella. A procedure is introduced for identifying years in which a significant change occurred in the direction of the population trajectory, and significance tests for drawing inference from the trend curve are described.

The second model examines the process by which a population spreads through its environment in space and time, taking into account spatial heterogeneity and the ease with which individuals can move between sites that are remote from each other. The basic model may be fitted in a straightforward fashion when survey data are available at frequent and regular intervals, but fitting is problematic when this is not the case. It is shown that the difficulties may be overcome by recasting the model as a modified branching process, and using a branching process recurrence relation to approximate the likelihood function. Simulation-based methods for parameter estimation are developed in addition to this approach, and results from the various techniques are found to square closely with one another. The methods are illustrated using survey data for the woodlark, *Lullula arborea*, from Thetford Forest, East Anglia, UK.

Acknowledgements

I would like to thank the following people. Firstly my supervisor, Steve Buckland, for his help and patience. My friends and colleagues in the Statistical Ecology Group and in the wider Maths department at St Andrews, for useful discussions. Guy Warner, for invaluable computing assistance. My parents and friends, and those people acknowledged elsewhere in the thesis for specific help.

This work was funded by a research grant from the EPSRC.

Contents

Chapter 1

Introduction

The impact of human activity on the environment has been such that management of ecological populations is now essential for the preservation of species and habitat. Many species in the UK and elsewhere are in danger of local or global extinction. The corncrake *Crex crex* is a prime example of a bird which has declined dramatically throughout its range over the past few decades, following changes in farming practices. Populations of other species are reaching unsustainable levels due to extermination of their natural predators: the red deer *Cervus elaphus* is one such species, and has been responsible for considerable habitat damage in Scotland through overgrazing. Still more problems are caused by invasion of a non-native species into an ecosystem, for instance the grey squirrel *Sciurus carolinensis* in Britain, which might be endangering the native red squirrel *Sciurus vulgaris*, or the brushtail possum *Trichosurus vulpecula* which has become the subject of a major and costly eradication programme in New Zealand. A number of projects are currently underway in the UK to reintroduce species to areas from which they have become extinct, such as the red kite *Milvus milvus*, and the white-tailed eagle *Haliaeetus albicilla.*

These examples all serve to illustrate the fact that mistakes in land management are ultimately expensive both in terms of public money and in richness of the environment, and it is important to ensure that environmental policy decisions are based on the best scientific advice available. An understanding of the ways in which an ecological population changes with its environment over space and time is central to the design of a sound management policy. Statistical models that capture the relationship between a population and its spatial environment contribute a great deal to this understanding, and can provide much insight into the potential consequences of management decisions.

The aim of a spatio-temporal model in ecology is to determine the influences that govern population spread or decline in a region. The models may be divided into two broad categories: descriptive, or empirical; and mechanistic, or process models. Descriptive models aim to describe or summarize population patterns, as functions of predictors such as habitat and climate. They identify those factors that are most highly correlated with population incidence, and quantify their effects. Likely causes of sudden or gradual changes in the fortunes of a population can be brought to light, and this information is useful for subsequent management strategy.

In occasional circumstances a fitted model of the descriptive type might be used to predict patterns of occurrence in new regions or at future times, but for the most part it is unwise to extrapolate beyond the scope of the data to which the model was fitted. A model might suggest that a certain crop is beneficial or detrimental to a population, or even provide an impression of the optimal mix of habitat cover for a species; and this information is likely to be applicable to other populations of the same species in different regions. However, it is unlikely that the model can predict the full pattern of occurrence in a different region, as there will be many aspects of the new situation about which the original model has no information. Baseline population levels will be different in the new region, and topographical features will affect the ease with which the population can expand or sustain itself in fragmented habitat.

For this reason it is useful to look more deeply into the mechanisms by which population changes occur, and produce models that aid our understanding of these mechanisms. Although no model will ever be completely transferable from one set of circumstances to another, the procedure becomes more justifiable as the model becomes more fundamental. In the ideal situation, a complete understanding of the life processes of a species — birth, death, and movement — and their relation to external influences such as habitat and climate, would enable sound judgements to be made about any population of the species. Possible impacts of habitat loss or new management schemes, or the likely success of a relocation or reintroduction programme, could then be evaluated. This is the philosophy behind mechanistic spatio-temporal models: a direct approach to the modelling of the *processes* of population change. Descriptive models, by contrast, are aimed at investigation of the *outcome* of these processes.

A wide literature is available on descriptive statistical models, and methodology in common use includes multiple regression, generalized linear or additive modelling, and generalized linear mixed modelling. Mechanistic models, since they are more detailed, also tend to be

more specific; and it is difficult to conceive of a general framework for their analysis. The present thesis provides an example of both a descriptive and a mechanistic spatio-temporal model, applied in each case to ornithological data. Both types of models are useful and necessary. The descriptive model, in Part I of the thesis, aims to determine trends in population abundance over time for several species of farmland birds. Account is taken of the spatial nature of the population data, although the model is primarily temporal. An approach based on generalized additive models is adopted.

In Part II of the thesis a mechanistic spatio-temporal model is introduced, in which the spread of a population through a survey region is modelled by recreating the process of colonization from one site in the survey region to another. The probability that individuals move from an occupied site to a target site is formulated as a parametrized function of the distance between the two sites and the habitat quality at the target site, and the aim of the material in Part II is to provide methodology for estimating the parameters of this formulation.

History of population modelling

Although the mathematical study of population models has a long and distinguished history, it seems that relatively little attention has been paid to the fitting of spatio-temporal mechanistic models to ecological data in the past. Ecological population models have always borrowed largely from similar models in epidemiology, and although the many parallels between the two are worthy of exploitation, there are still sufficiently many differences to warrant separate treatment.

Early research in the field of population modelling focused on deterministic formulations (e.g. Kendall 1948; Kendall 1965; Daniels 1995). These were perceived to be more tractable than their stochastic counterparts, while capturing the important features of asymptotic behaviour. Advocates of this approach were nonetheless aware of its limitations, justifying it only on the basis of mathematical convenience (Maynard Smith 1974). Recent authors have considered deterministic models in parallel with their stochastic analogues (e.g. Renshaw 1991; Hengeveld 1989).

Over the last two decades there has been increasing interest in stochastic spatial modelling. Theoretical research has built on the work of Hammersley and Welsh (1965), McKean (1975) and Mollison (1977) for the study of asymptotic model behaviour such as velocity of spread (Metz & van den Bosch 1995), limiting shape (Cox & Durrett 1988) and critical

values (Buttell *et al.* 1993). Applied research has centred mainly on descriptive models for trends analysis and abundance estimation (e.g. Borchers *et al.* 1997; Bak & Nieuwland 1995). Spatio-temporal mechanistic models have been fitted to epidemiological data by Besag (1977) and Gibson (1997). Full theoretical treatments of stochastic spatial processes with applications in ecology have been attempted, but tend to involve oversimplifying assumptions (e.g. Bramson *et al.* 1998).

The aim in this thesis is to strike the middle ground between theory and application. Where a model is amenable to analytic treatment, that approach will be favoured. In recognition of the dangers of oversimplification for the purposes of mathematical convenience, however, methods will also be provided for cases where a direct analytic approach is not possible. Although there is much potential for methodological research in the field of stochastic spatial processes, the present work is driven by an equally urgent need in the field of application. Our wildlife populations are monitored and managed more extensively than ever before, and there is a need to develop statistical methods to keep pace with the increasing volume of survey data. Large sums of public money and charitable donations are channelled into conservation and habitat revitalization projects, and in order for this money to be put to best possible effect it is important to ensure that the response of the ecosystem to management proposals is well understood.

Part I

Trend model

Chapter 2

Analysis of population trends from annual census data

1 Introduction

Monitoring of trend in wildlife populations is essential for the development of long-term conservation strategy. The impact of past change in land management must be understood in order to gain insight into likely outcomes of future change. For this reason much effort has been devoted to collection of wildlife abundance data, especially for bird populations. In the UK, the British Trust for Ornithology (BTO) has been coordinating the Common Birds Census (Marchant *et al.* 1990) since 1962, and has recently initiated the UK Breeding Bird Survey (Gregory *et al.* in press). The North American Breeding Bird Survey (Droege 1990) began in 1966, and covers much of Canada and the US. Over 40 000 participants are involved annually in the North American Christmas Bird Count (Butcher 1990), which has been running since 1900.

All of these surveys are carried out by volunteers, and this has important consequences for analysis of the data. There is an inevitable lack of continuity as observers are unavailable or leave the scheme, and as new observers join. This leaves large quantities of missing data. Observers differ in their ability to detect and count birds, so an apparent decline in abundance could be due in reality to a change of observer. Some surveys, such as the UK Common Birds Census (CBC), allow observers to select their own sites for coverage: an understandable tendency to prefer the more interesting sites means that the selected sites are not necessarily representative of national habitat.

These problems have led to the development of many different analysis methods for large-scale monitoring data (ter Braak *et al.* 1994), each with certain weaknesses. Moreover, techniques suitable for one survey might not be feasible for another, even when the survey designs are similar. For example, the North American Breeding Bird Survey (BBS) is commonly analysed using route regression, in which a separate temporal abundance trend is fitted to data from every site. This method would be infeasible for the British CBC, since it is rare for a single site to contain data from sufficiently many years to allow a meaningful trend to be fitted. Conversely, the BBS involves data from some ten times more sites than the CBC; ergo a method that works well for the CBC might be infeasible for the BBS, simply due to the practicalities of model-fitting.

In this chapter a new approach to the analysis of annual census data is introduced, based on generalized additive models. Existing methods are briefly reviewed, and the formulation of the generalized additive model follows naturally from these. Expressions are derived for the calculation of annual indices of abundance, and significance tests for drawing inference from the index series are described. A method is presented that allows identification of years in which a significant change occurred in the direction of the population trajectory: the curve of second derivatives of the abundance index curve is estimated, and where the second derivative is significantly different from zero, a change in the abundance trajectory is estimated to have taken place. Finally some discussion is provided of covariate models, which enable factors such as geographical location or climate to be taken into account. The methods are illustrated using data from the UK Common Birds Census for 13 species of farmland bird.

Throughout the chapter, attention is focused on the study of trend or pattern in abundance, rather than on absolute levels of abundance. Methods of testing the timing and statistical significance of apparent change are provided, but percentage declines are not emphasized. At present, there is no long-running survey that is sufficiently extensive and representative to allow valid conclusions to be drawn about the absolute abundance of common species on a national scale. Further, the relationship between the number of birds detected and the true number present in a site is often uncertain. In the North American BBS, for example, only 3 minutes are allowed at each sampling point to assess the abundance of every species present (Droege 1990); it is unreasonable to expect abundance records to be accurate in these circumstances.

Inference about percentage change can also be distorted by systematic variation in the proportion of individuals that are detected. For some species there is likely to be a

systematic increase in detectability with local abundance: where abundance is high there is more need to defend territory and compete for resources, both of which cause an increase in conspicuity. In this case, the probability of detecting each individual rises according to the number of individuals present, so that the expected number of detections does not increase linearly with abundance. This situation is illustrated by a simulated example in Figure 1, where detections are considered to be independent and to occur with the same probability for all individuals. When probability of detection for each individual increases with abundance, as in curve (iii), the apparent percentage declines and increases are steeper than the true changes (curve (i)). In curves (ii) and (iv) there is a linear relationship between true abundance and number of detections, and in these curves the true percentage changes are preserved.

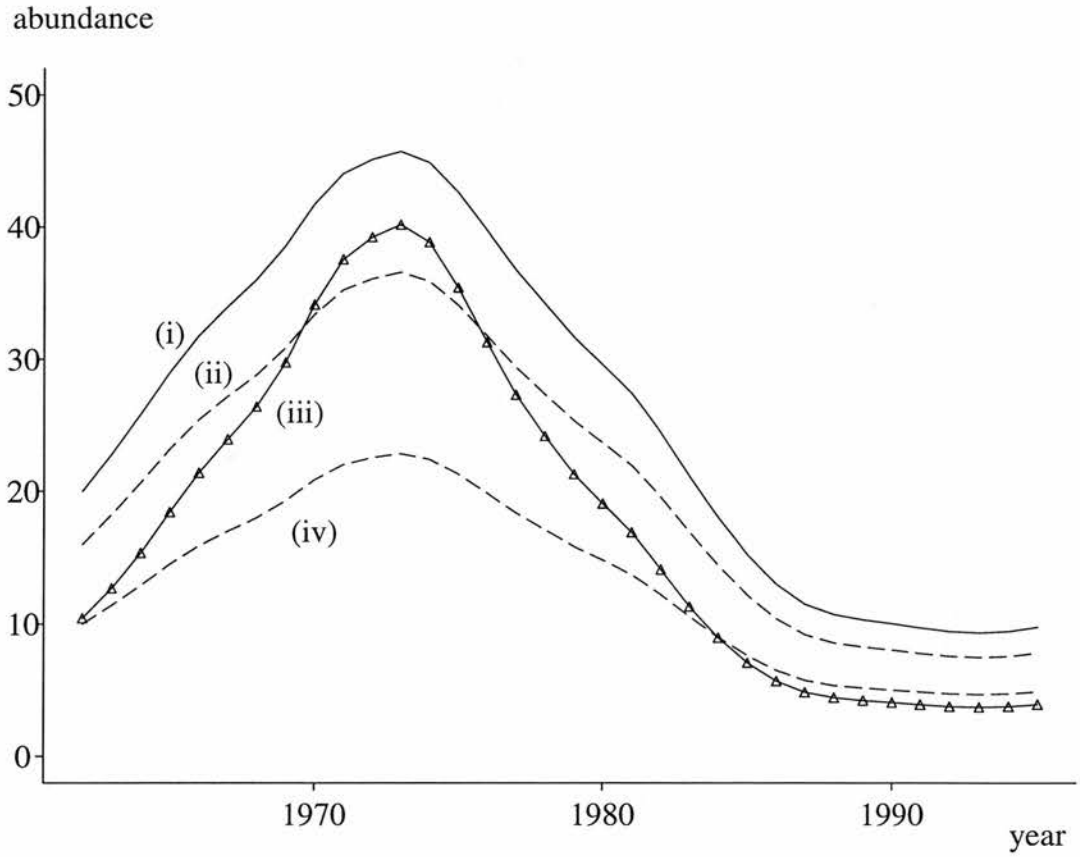


Figure 1: Simulated curves showing non-linear variation of detected abundance with true abundance. Curve (i) is the true abundance curve. Curves (ii) and (iv) vary linearly with (i) and preserve the abundance ratio between any two years. Curve (iii) varies non-linearly with (i) and abundance ratios are distorted.

Another source of systematic variation in detectability might be habitat succession. It is often more difficult to detect individuals in areas of dense vegetation, for example, than where vegetation is sparse. Consequently, a monitoring survey that covers regions where

there is large-scale habitat succession might be subject to systematic change over time in the ratio of detected to true abundance. Once again, this would cause the true percentage changes to be misrepresented.

Use of specialized abundance estimation techniques such as distance sampling (Buckland *et al.* 1993) would greatly improve the potential of large-scale monitoring schemes to provide reliable abundance information. This analysis will be possible with data from the new UK Breeding Bird Survey (Gregory *et al.* in press), which has been in operation since 1994. Further improvements could be accomplished using the mark-recapture line transect methods of Borchers *et al.* (in press), that would allow uncertain trackline detection and covariates to be incorporated into abundance estimation at the site level. However, limited resources and a need to make the survey work accessible to as many volunteers as possible exclude some of the more sophisticated options.

2 Review of previous analysis methods

2.1 Survey designs

In general, survey data from annual monitoring schemes consist of a set of N sites, monitored over T years, with counts y_{it} of the number of individuals of a species observed in site i in year t. These data are to be used to deduce information about temporal abundance trends across the survey region. The definition of a site differs between surveys: in the UK Common Birds Census (CBC) the sites are plots of farmland or woodland, while for the North American BBS the sites, or *routes*, are 24.5 mile stretches along secondary roads, censused at 50 points spaced at intervals of half a mile along the route.

The way in which the counts y_{it} are derived also differs between surveys. In the CBC, frequent visits are made to each plot throughout the breeding season. Numbers and locations of individuals of each species are recorded on every visit, and the records from the whole breeding season are combined to provide an estimate of the number of territory-holding males in the plot, using the methods of North (1977; 1979). For the North American BBS, a 3-minute point count is conducted at each of the 50 stops along a route, and the number of individuals of every species seen or heard during the 3-minute interval is recorded. In some cases, the count y_{it} for a species along route i in year t is taken to be the sum of the 50 counts along the route. However, since it is recognized that a realistic impression of abundance cannot be obtained in only 3 minutes, several analysts

have chosen to define the count y_{it} as the *number* of the 50 stops in which the species was detected (James *et al.* 1990; Droege 1990). This measure is regarded as less sensitive to observer differences, and more reliable, since presence or absence is easier to determine than abundance. As Droege (1990) points out, though, the measure is one of extent rather than abundance.

A common feature of large-scale survey data is the high proportion of missing counts: typically no site will contain data from all T years. The datasets analysed in this chapter from the CBC were found to have between 79% and 82% of possible site-year counts missing, for each of 13 species. Plots that were covered by more than one observer over the survey period were recorded as separate sites, one for each observer; this accounts partly for the very high proportion of missing entries.

Five of the main approaches to analysis of data of this type are reviewed here. These are the chain or ratio method; the Mountford method; route regression; log-linear Poisson regression using the software package TRIM (*Trends and Indices for Monitoring Data:* Pannekoek & van Strien 1996); and generalized linear models with parametric trend estimation (Link & Sauer 1997a). The basic models used for each of these methods are similar, but there are important differences in implementation.

2.2 Chain method

In the chain or ratio method (Marchant *et al.* 1990; ter Braak *et al.* 1994; Mountford 1982), indices of annual abundance are calculated as simple ratios of observed counts between successive years. In order for the counts to be comparable, only those sites that are surveyed in both of the two years are included in the ratio. Thus if S_t is the set of sites that are surveyed in both of years t and $t+1$, the ratio $r_{t+1,t}$ of abundance in year $t+1$ to abundance in year t is calculated as

$$r_{t+1,t} = \frac{\sum_{i \in S_t} y_{i,t+1}}{\sum_{i \in S_t} y_{it}} .$$

The change in abundance between two non-consecutive years is calculated as a product of the ratios for intermediate years: for example between years t and $t+n$ the change is

$$r_{t+n,t+n-1} \times r_{t+n-1,t+n-2} \times \ldots \times r_{t+1,t} . \qquad (1)$$

Unless the sets of sites $S_t, \ldots, S_{t+n-1}$ coincide, this is not the same as the ratio $r_{t+n,t}$. Consequently the method does not use all of the available data, and performs only indirect comparisons between non-consecutive years. In order to calculate annual indices of abundance, the estimates of annual change are chained round a base year.

The chain method is now generally regarded as inadequate, due to inefficient use of the data and tendency to generate spurious trends through random drift (Mountford 1982; Mountford 1985; Peach & Baillie 1994; ter Braak *et al.* 1994). Random variability in sites which enter the analysis only for a short time can cause spurious trends in the overall indices (ter Braak *et al.* 1994).

2.3 Mountford method

The Mountford method (Mountford 1982; 1985) was proposed as an extension to the chain method that generates more precise indices and makes more efficient use of the data. A formal model is proposed for the counts y_{it}, namely

$$Y_{it} = a_i \times b_t + \epsilon_{it}\,, \tag{2}$$

where Y_{it} is the random variable underlying the count y_{it}, and a_i, b_t represent the effects particular to site i and year t respectively. The error ϵ_{it} follows a distribution with zero mean, and variance proportional to $a_i b_t$: for example, Y_{it} might be a Poisson random variable with mean $a_i b_t$. In fact, the distribution of ϵ_{it} is not used in the analysis, since estimation of the parameters proceeds via an extended ratio method rather than by likelihood-based techniques, and the variance-covariance matrix is estimated using sample variances.

The parameters of interest in (2) are the year effects $b_t : t = 1, \ldots, T$. Instead of employing only the ratios $r_{t+1,t}$ of counts between consecutive years, as for the chain method, any ratio r_{t_1,t_2} is admissible in calculation of the Mountford index. As with the chain method, the ratios r_{t_1,t_2} involve only those sites common to both year t_1 and year t_2, and proportional changes in abundance are assumed uniform across all sites over any time period.

The calculation of the Mountford indices is based on the observation that, if all counts were complete in every year, the ratio r_{t_1,t_2} would be equal to any product of ratios of the

form

$$r_{t_1,u_1} \times r_{u_1,u_2} \times \ldots \times r_{u_{k-1},u_k} \times r_{u_k,t_2}\,, \tag{3}$$

for any years $u_1, \ldots, u_k$ and integer k. Since the counts are not complete, this product will take different values for each different selection of $\{u_1, \ldots, u_k\}$, although each one should be close to the ratio r_{t_1,t_2}. These ratio products therefore provide numerous estimates of the ratio of total count in year t_1 to that in year t_2, the underlying value of which is b_{t_1}/b_{t_2} in terms of the parameters of (2). The Mountford estimate of b_{t_1}/b_{t_2} is selected to be the product in (3) whose logarithm has minimum variance. Indices of annual abundance are obtained as deviations from the mean year effect: that is,

$$\log b_t - \frac{1}{T}\sum_{u=1}^{T} \log b_u$$

for year t. Alternatively they may be calculated as deviations from a base reference year, as for the chain method.

Mountford (1982) commented that maximum likelihood methods of parameter estimation would be impracticable for the model (2), since the manipulation of prohibitively large matrices would be required. Developments in computer technology in recent years mean that such maximum likelihood analysis is now routine, and it forms the basis of the generalized additive model approach described in Section 3.

An advantage of the Mountford method is that variance estimation is possible analytically using large-sample theory. However, the Mountford method is somewhat rough and ready: the indices are selected to give greatest possible precision, but this might be at the expense of some bias. The assumption that population change is uniform across all sites over any time period is central to the success of the approach, but is likely to break down after a few years. For this reason, the Mountford method has been implemented by the BTO using a 'moving-windows' technique (Peach & Baillie 1994) which applies the method to 'windows' of only a few years at a time.

Both the log-linear Poisson regression method in section 2.5 and the generalized additive model approach developed in section 3 make a similar assumption of homogeneity of population change across sites, but mitigation is afforded in these cases by two things. Firstly, in both cases it is possible to divide the survey area into smaller regions or to add covariates, so that there is potential for non-homogeneous trends to be fitted. This could not be easily accommodated within the Mountford framework. Secondly, in the Mountford method the assumption of homogeneity across sites is not only explicit in the

model, but also implicit in the estimation of the model parameters. For example, if the estimate of b_1/b_2 is given by $r_{15} \times r_{52}$, then the assumption must hold between years 1 and 5 and between years 5 and 2 in order for the estimate to be reasonable. For the Poisson regression and generalized additive models, on the other hand, the assumption enters explicitly through the model but is not implicit in the fitting of the model. Although a breakdown in the assumption would mean that the fitted model did not provide an accurate description of the population at large, the parameter estimates would nonetheless still be valid within the context of the model. In view of the fact that very few fitted models provide accurate descriptions of complex systems, it would be wrong to dismiss these models on this basis alone.

Both the Mountford and chain methods concentrate on production of annual indices of abundance, which reflect annual fluctuations in both population levels and data quality. For analysis of smooth population trends, it is necessary to find some pattern beneath the 'noise' of annual fluctuation. This may be achieved to some extent by the application of a smoothing algorithm to the index curves. For example, Siriwardena *et al.* (1998) applied a compound running-median algorithm to Mountford index series derived from CBC data. However, smoothing the output from one model amounts to application of a second model, and a model which incorporates trend estimation directly should be preferred.

2.4 Route regression

Route regression (Geissler & Noon 1981) differs from the previous two methods in that it focuses on the explicit modelling of trend, rather than on the production of annual abundance indices. Trend is modelled separately for every site, or route, and the resulting trend estimates are aggregated into an overall estimate at the region or state level using a variety of weighting factors.

A popular version of the basic model is as follows (James *et al.* 1990; ter Braak *et al.* 1994):

$$\log(Y_{it} + 0.5) = \log a_i + t \log b_i + \epsilon_{it} \,. \tag{4}$$

The constant 0.5 added to the count Y_{it} is intended to correct for zero counts. A linear relationship is assumed for each site i between corrected log-count and time, with site intercept $\log a_i$ and site slope $\log b_i$. The error term ϵ_{it} is assumed to follow a normal distribution with mean zero.

The model (4) is often extended to allow for observer effects (Geissler & Sauer 1990; Sauer & Geissler 1990), becoming

$$\log(Y_{it} + 0.5) = \log a_i + t \log b_i + \log c_r + \epsilon_{it}\,, \tag{5}$$

where the route i in year t was surveyed by observer r, and $\log c_r$ is the corresponding observer effect.

The route trend is defined as b_i, and is obtained by back-transforming the regression estimate of $\log b_i$. The back-transformation is approximately

$$\exp\left(\log(b_i) - 0.5\hat{\mathrm{V}}\mathrm{ar}(\log(b_i))\right)\,,$$

and is designed to compensate for the skewness of log-normal counts (Geissler & Sauer 1990; Bradu & Mundlak 1970).

Regional trends are calculated as a weighted mean of route trends. Weights are allotted according to a number of factors, notably the area covered by the route, abundance along the route, and precision of the trend estimate for the route (Geissler & Sauer 1990; James *et al.* 1990). For example, a route in which the population abundance diminishes over the survey period from two individuals to one might have a trend estimate of 0.5, while a second route in which the population increases from 100 to 200 individuals might have an estimate of 2. It would clearly be wrong to take the estimate of trend across both routes as an unweighted mean of the two estimates. However, it is not clear what the weighting system should be in order to give an accurate impression of overall trend.

Alternative implementations of the route-regression method differ in details such as the additive constant used to correct the counts, the method of back-transforming the route trend estimates, and the weighting scheme for aggregation of trend estimates (Thomas 1996). Indices of annual abundance are sometimes obtained via an *ad hoc* residual method (Sauer & Geissler 1990). Variance estimation is generally performed via the bootstrap (Geissler & Sauer 1990; ter Braak *et al.* 1994).

There are many criticisms of the route-regression model as formulated above. The first and most important is the impossibility of finding any objective means of allotting weights to the route trends so that they may be aggregated into an overall trend: the choice of weights can have a dramatic effect on the final results (James *et al.* 1990). Secondly, the model formulation is sloppy: the counts Y_{it}, which are discrete, cannot follow a continuous log-normal distribution. This is of particular consequence when counts are small. The choice

of additive constant used to correct for zero counts is arbitrary and can have an effect on trend estimates (ter Braak *et al.* 1994). Back-transformation of the route trends can lead to bias (Geissler & Sauer 1990; ter Braak *et al.* 1994).

Most of these criticisms could easily be addressed by re-casting the models (4) and (5) as generalized linear models (GLMs). In this case, the *expected* counts are modelled as linear functions on the log scale, rather than the raw counts: so a raw count of zero immediately ceases to pose a problem since its expectation, however small, is always greater than zero. In the GLM framework the distribution of the counts can also be chosen more appropriately, for example the Poisson or negative binomial distributions. Methods of overall trend estimation are easily devised that render back-transformation and weighting of route trends unnecessary (Section 3).

A further criticism of route regression that has now been addressed by some authors is the constraint of linearity in the log-trend. James *et al.* (1990; 1996), and Taub (1990) have all used locally weighted regression smoothers to incorporate non-linear trends. While representing a major step forward, these methods are still lacking in rigorous model formulation. Taub (1990) does not provide a model for the counts at all, but smoothes the raw data; this restricts potential for further inference such as prediction or tests of population change. James *et al.* (1990; 1996) use similar formulations to (4) and (5), subject to the criticisms above. None of these authors use the smooth trend curves for direct inference about trend: Taub (1990) regards the smooth curves as of visual or descriptive purpose only, while James *et al.* (1996) use the non-linear curves purely for estimation of the significance of linear changes between particular years. Again, the methods would benefit from re-casting into a more rigorous framework, this time as generalized additive models (GAMs). Use of the smooth trend curve obtained from a generalized additive model to draw direct inference about population change is described in Section 3.

In summary, the route regression method has all the ingredients of a successful modelling strategy, but implementation to date has left much room for improvement. The direct approach to the modelling of trend, and the potential for inclusion of observer effects and other covariates, are definite bonuses that will be emulated in the GAM approach developed in this chapter. A final problem with the method, however, is the high parametrization required. Data from the UK Common Birds Census, for example, are simply not extensive enough to allow a separate trend to be fitted in each of the surveyed sites. Pooling of the data across many or all sites is necessary, and this is easily incorporated into the GAM framework.

2.5 Log-linear Poisson regression

Log-linear Poisson regression methods for analysis of survey data have been implemented in a specialized software package known as TRIM (*Trends and Indices for Monitoring Data*: Pannekoek & van Strien 1996). These methods are cast as generalized linear models (McCullagh & Nelder 1989), with Poisson error distribution and logarithmic link function; consequently the basic models may be fitted using many general-purpose statistical packages such as S-PLUS (Statistical Sciences Inc. 1993), GLIM (Aitkin *et al.* 1989) or SAS (SAS Institute Inc. 1996).

Two primary models are included in the TRIM package: a linear trend model and an 'annual model'. In both cases the count Y_{it} is assumed to follow a Poisson distribution with mean μ_{it}, and all counts are assumed independent. In the linear trend model, the mean μ_{it} is modelled as follows:

$$\log(\mu_{it}) = \alpha_i + \gamma t\,. \tag{6}$$

Although similar to the route regression model (4), in this case there is a single parameter γ for all sites, whereas the route regression model used a different parameter $\log(b_i)$ for every site. Of course, neither option is precluded in either route-regression or log-linear Poisson regression. The most important differences between the formulations (6) and (4) lie in the particulars of the model: firstly, it is the expected count μ_{it} that is modelled as a linear function on the log-scale, rather than the corrected raw counts $Y_{it}+0.5$; and secondly the counts Y_{it} are assumed to be Poisson random variables rather than log-normal. The log-linear Poisson formulation is thereby much more satisfying in theoretical terms than the route regression model, and avoids the associated difficulties. The principal criticism of the linear trend model is the constraint that counts should vary linearly with time on the log-scale. For any detailed analysis of timing and causes of change in population levels, a simple linear trend is not satisfactory.

The TRIM annual model concentrates on annual fluctuations: the year variable t enters as a factor (categorical variable) instead of a continuous variable as in the linear trend model (6). The expected counts are therefore modelled as

$$\log(\mu_{it}) = \alpha_i + \beta_t\,. \tag{7}$$

Here α_i and β_t are referred to as the *site effect for site* i and the *year effect for year* t

respectively. The model is fitted by finding maximum likelihood estimates $\hat{\alpha}_i$ and $\hat{\beta}_t$ for the N parameters α_i and the T parameters β_t. Once these estimates are obtained, the predicted count for site i in year t is given by

$$\hat{\mu}_{it} = \exp\left(\hat{\alpha}_i + \hat{\beta}_t\right), \tag{8}$$

and the total predicted count for year t is

$$\sum_{i=1}^{N} \hat{\mu}_{it} = \exp(\hat{\beta}_t) \sum_{i=1}^{N} \exp(\hat{\alpha}_i). \tag{9}$$

These predicted counts are used in place of the missing data in the original sample to obtain annual estimates of relative abundance. The abundance index for year t is defined as

$$\mathcal{I}_t = \frac{\text{total predicted count for year } t}{\text{total predicted count for year 1}} = \frac{\exp(\hat{\beta}_t)}{\exp(\hat{\beta}_1)}. \tag{10}$$

There is no difficulty with back-transformation of parameter estimates, as with the route-regression method, or with unbalanced data-sets as with the Mountford and chain methods. The index is a measure of relative abundance and has no units.

A useful feature of log-linear Poisson regression, or indeed any GLM fitted with canonical link function and including an intercept or factor, is the fact that the sum (9) of predicted counts for year t is equal to the sum of observed counts y_{it}, where they exist, added to the predicted counts for the sites where no observations exist (Nelder & Wedderburn 1972; ter Braak *et al.* 1994). This lends justification to the definition used for the annual indices. The index for year t is also a ratio of exponentially transformed year effects, making for easy calculation.

The disadvantage of the annual model (7) is the lack of an explicit trend function. A separate parameter is allotted to each year, with no consideration of the sequence of years. This leads to unconstrained annual estimates, rather than a smooth trend curve. The linear trend model and the annual model of TRIM therefore lie at opposite ends of the spectrum: in the first, trend is explicit but too simplistic, while in the second trend is not explicit and secondary smoothing is required before conclusions about long-term change may be drawn. It is the middle ground between these extremes that is exploited in the generalized additive modelling framework to be developed: trend curves are produced which are smooth but need not be simplistic.

The TRIM program is equipped with further features, such as inclusion of categorical covariates into the linear predictor functions (6) and (7), and corrections for overdispersion and serial correlation in the counts when estimating the standard errors of the parameter estimates. Of these, only the serial correlation correction is not commonly available in general-purpose statistical packages such as S-PLUS, and most general-purpose packages allow inclusion of continuous covariates in addition to categorical variables. The TRIM program does, however, provide significant improvements in speed over the general software for larger data-sets. The TRIM corrections for serial correlation and overdispersion are obtained by providing models for the variance and correlation structure of the counts, in addition to modelling the counts themselves; the resulting standard error estimates should not be taken as definitive since the models might not be appropriate. (Standard error estimates obtained from S-PLUS analyses, however, should be treated with even more caution since they are obtained from a series of rough approximations.) Inclusion of the TRIM correlation and overdispersion features slows the model-fitting process enormously: for example, using CBC data for the skylark, the basic model took 2.5 minutes to fit on a Pentium PC with 16MB RAM running Windows95; with serial correlation and overdispersion corrections the time taken rose to 9.5 hours.

2.6 Generalized linear models with parametric trend estimation

Link & Sauer (1997a; 1997b) propose an extension of the log-linear Poisson models of the previous section. Their approach is designed to produce a smooth trend function intermediate between the linear trend and annual models of TRIM, and of all the methods reviewed it is closest to the generalized additive modelling approach of Section 3. Counts Y_{it} are assumed independent, and follow either Poisson or negative binomial distributions. The negative binomial distribution is often used in place of the Poisson when data are thought to be overdispersed (e.g. Gotway & Stroup 1997; Augustin *et al.* in review).

The basic model for the mean counts $\{\mu_{it}\}$ may be written as

$$\log(\mu_{it}) = \eta_r + \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3 + \dots \,, \tag{11}$$

where site i was surveyed by observer r in year t, and observer effects η_r are used instead of the site effects α_i of section 2.5. Observer effects are considered important because the model is applied to data from the North American BBS, in which the same route is commonly surveyed by many different observers over time. The issue of observer effects

does not arise to the same extent in some other surveys such as the British CBC, in which each site is typically surveyed by only one observer. Link & Sauer (1997a) also include parameters to adjust for inexperienced observers, but these are omitted for clarity.

The model (11) is a GLM, although Link & Sauer choose an alternative fitting procedure in which a conditional likelihood based on sufficient or approximately sufficient statistics for η is maximized with respect to the remaining parameters. It is not clear what advantage this procedure has over a straightforward GLM fit; however, the particulars of model-fitting are not of primary interest in the present discussion.

By adding higher powers of t to the expression (11), a smooth trend curve of any complexity may be accommodated, thereby fulfilling one of the principal objectives of analysis of monitoring data: the direct modelling of smooth trend. Conceptually, the generalized additive models of Section 3 are very similar to this model — indeed, experiment indicates that for polynomials of high order, the trend function obtained from (11) is much the same as that obtained from a GAM with equivalent parametrization. The residual deviance values from both types of models are also comparable. In the GAM approach, however, the trend function is estimated non-parametrically from the data: no parametric specification such as (11) is possible. In this way the trend function need not conform to any rigid parametric dependence on time. Although the parametric curve (11) is subject to fewer constraints as the degree of the polynomial is raised, dealing with polynomials of very large order becomes cumbersome — and those options providing most flexibility are unlikely even to be considered. It will become clear in Section 3 that restriction of attention to models with low flexibility is likely to hide many interesting features of trend.

The global nature of the polynomial regression fit is also somewhat unappealing for trend analysis. A parametric specification has no advantage in predictive capability, since even one or two years beyond the end of the time series the best-fit polynomial invariably produces predictions that are wildly infeasible. Hence it is unrealistic to believe that the data really follow a polynomial pattern. A local non-parametric fit seems more suitable conceptually, even though there may be little difference in the results.

Link & Sauer (1997a) do not describe means of explicit inference about population change from the trend curve obtained through their model. The methodology developed in sections 3.5 and 3.6 for inference from GAM trend curves would, however, be equally applicable to parametric trend curves.

3 Analysis of population trends using generalized additive models

3.1 Description of the data

A new approach to analysis of large-scale monitoring data is now developed, based on generalized additive models (GAMs). The methods are illustrated using data from the UK Common Birds Census (CBC), described in section 2.1. Data from farmland plots between 1962 and 1995 are analysed for 13 species:

bullfinch *Pyrrhula pyrrhula* (349),
chaffinch *Fringilla coelebs* (519),
corn bunting *Miliaria calandra* (185),
goldfinch *Carduelis carduelis* (412),
greenfinch *Carduelis chloris* (456),
grey partridge *Perdix perdix* (384),
linnet *Carduelis cannabina* (463),
reed bunting *Emberiza schoeniclus* (350),
skylark *Alauda arvensis* (499),
stock dove *Columba oenas* (262),
tree sparrow *Passer montanus* (328),
turtle dove *Streptopelia turtur* (219),
yellowhammer *Emberiza citrinella* (446).

The numbers in brackets denote the number of sites involved in the analysis for each species.

There is much concern about the conservation status of many of these species, following declines across Europe in both abundance and range (Marchant & Gregory 1994; Tucker & Heath 1994; Fuller *et al.* 1995; Siriwardena *et al.* 1998; Gibbons *et al.* 1993). Knowledge of the timing of downturns in population numbers might indicate which of the numerous changes in farming practice have had adverse effects for each species.

3.2 Formulation of the model

The generalized additive model (Hastie & Tibshirani 1990) is a flexible extension of the generalized linear model, and the methods used in this section follow naturally from the log-linear Poisson GLMs detailed in section 2.5. The counts y_{it} are again assumed to follow independent Poisson distributions with mean μ_{it} for the count in site i in year t. The linear predictors, which in the log-linear Poisson GLMs were given by equations (6) and (7), are however replaced by a single generic *additive predictor*, which allows log mean abundance to vary as any smooth function of time rather than as a linear function alone. The form of the predictor function is the principal difference between the GLM and the GAM.

The additive predictor may be written as

$$\log(\mu_{it}) = \alpha_i + s(t)\,. \tag{12}$$

Here, the smooth function of time in the additive predictor is represented by $s(t)$, so that the expected count μ_{it} in site i in year t is dependent on the site effect α_i, and on any number of other smoothly-varying quantities which are summarized by the value $s(t)$ in year t. The GAM is fitted by estimating the parameters α_i and the smooth function s.

From (12) it is clear that the log-linear Poisson models in section 2.5 are special cases of the GAM formulation: for the linear trend model (6), the function s is parametrized as $s(t) = \gamma t$ for the single parameter γ, while for the annual model (7) s is parametrized as $s(t) = \beta_t$ for the T parameters $\beta_1, \ldots, \beta_T$. In the second case the function s is no longer smooth, but is obtained by joining the estimates β_t with straight lines.

These two cases lie at the extremes of the GAM framework: the first with maximum smoothness in the function s (a single straight line), and the second with minimum smoothness (a sequence of unconstrained estimates joined by linear segments). They are also the only cases for which a parametric specification is possible for the function s. Between the two extremes lie functions with greater flexibility to fit the data than a straight line, but which provide smooth trends through the data rather than disconnected annual estimates. These are the functions that are obtained through generalized additive modelling.

The output from the GAM is easily visualized. The fitted year effect curve $\hat{s}(t)$ is common to all sites, so that for any two sites i_1 and i_2 the curves $\log(\mu_{i_1 t})$ and $\log(\mu_{i_2 t})$ are parallel. The intercepts of these curves are determined by the site effects, respectively α_{i_1} and α_{i_2}.

Every site, therefore, is subject to the same trend in the logarithm of expected count over time; although the absolute values differ between sites. In order to overcome the restriction that the same log-trend must be followed in every site, the survey region may be split up into as many subsets of sites as required, and a separate GAM fitted for each subset. The extreme at which a separate GAM is fitted in every site is equivalent to the re-cast route regression framework mentioned in section 2.4: the additive predictor could then be written as $\log(\mu_{it}) = \alpha_i + s_i(t)$. More discussion of this is provided in section 3.7.

Once an estimate $\hat{s}$ of the smooth function s has been obtained, the annual abundance index curve $\mathcal{I}(t)$ is calculated as with the GLMs in section 2.5:

$$\mathcal{I}(t) = \frac{\text{total predicted count for year } t}{\text{total predicted count for year 1}} = \frac{\exp\left(\hat{s}(t)\right)}{\exp\left(\hat{s}(1)\right)}. \tag{13}$$

$\mathcal{I}$ is now written as a smooth function of t, rather than as a set of point estimates as in (10).

3.3 Estimation of the smooth function s and specification of the degree of smoothing

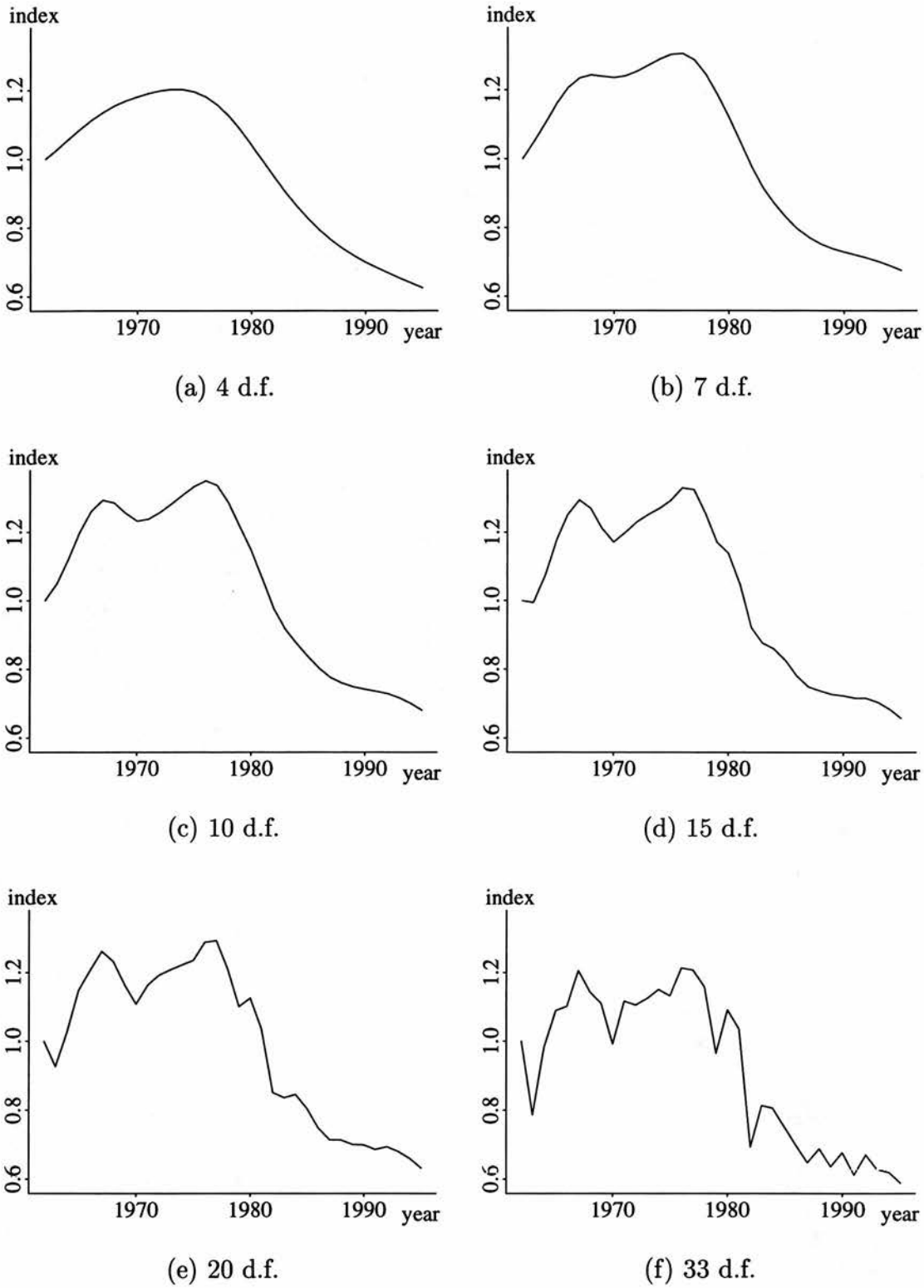
The function s is estimated non-parametrically from the data, using scatterplot smoothers. The shape of the function is therefore defined by the data rather than being constrained to a parametric form. The level of smoothing to be applied is decided upon before the model is fitted. It should be emphasized that the smoothing process in a GAM is part of the model-fitting procedure, and is not merely a smoothing of fitted values previously obtained from the model. The fits were accomplished using the S-PLUS function `gam`.

There is considerable choice over the scatterplot smoother to be used in estimating s, although experiment shows that the fitted curve varies little between choices. All the analyses presented here employ smoothing splines; other choices include locally weighted regression smoothers, kernel smoothers and running-median smoothers. Hastie & Tibshirani (1990) give a comprehensive treatment of scatterplot smoothers and their properties. Smoothing splines are piecewise cubic polynomial fits to the data, motivated by a penalized least squares criterion which is designed to optimize the fit while penalizing roughness to some pre-determined extent. The extent to which roughness is penalized, or equivalently the level of smoothing that is applied, is calibrated by a quantity known as the *degrees of freedom* (d.f.). As the degrees of freedom are increased, the function s gains in flexibility: more turning points and gradient changes are accommodated. A straight line $s(t) = \gamma t$

corresponds to a fit with 1 d.f. — the curve follows a single direction only, with no gradient changes or turning points. Conversely, a piecewise linear fit $s(t) = \beta_t$ $(t = 1, \ldots, T)$ corresponds to a fit with $T-1$ d.f.; a separate gradient is allowed between each successive pair of points. The degrees of freedom associated with the curve s may loosely be interpreted as the number of parameters used in fitting s.

The choice of the value for d.f. is an important part of the modelling process. For clarity, s_d shall be used to refer to the curve s that is fitted using a smoothing spline on the variable t with d degrees of freedom, and the associated model is described as a 'GAM with d degrees of freedom'. Broadly speaking, d indicates the extent to which the smoothed curve may include changes of direction. The choice of d depends largely on the objectives of the analysis. For inference about long-term trends a fairly smooth index curve is required, corresponding to low d.f., while information about annual fluctuations requires unconstrained annual estimates and the maximum value of d. Consideration of the length of the time series is also important: the larger the number of years T, the higher the value of d must be in order to maintain a particular level of flexibility in the trend curve.

For trend estimation with CBC data, a value of d was selected that was high enough to ensure that the trend was not constrained to an unduly simplistic pattern, but low enough to remove all roughness from the output. Abundance index curves obtained from GAMs fitted to CBC data for the skylark with a range of d.f. are shown in Figure 2; the curve for $d = 10$ was selected as most appropriate for intermediate trend estimation, allowing enough flexibility to capture all important features of the population trajectory while remaining sufficiently smooth to ensure the underlying trend is clearly displayed. Experiments with truncation of CBC data sets suggested that a choice of d that is roughly 0.3 times the length of the time series produces a trend curve with suitable complexity and smoothness, although it is stressed that advice will vary according to precise objectives and data. In practice, the fitted trend curve changes little for small changes in d. It is always instructive to plot indices from GAMs with a range of d.f. to ensure that important features of the trend are not lost through over-smoothing.



(a) 4 d.f.

(b) 7 d.f.

(c) 10 d.f.

(d) 15 d.f.

(e) 20 d.f.

(f) 33 d.f.

Figure 2: Abundance indices from GAMs with various degrees of freedom fitted to CBC data for the skylark.

3.4 Diagnostics

A full set of diagnostic tools for generalized additive models has yet to be developed; however, there are informal ways of verifying the goodness of fit which serve to draw attention to likely problems. Plots of residuals and standard error bands for the fitted year effect curve provide a visual test of the fit and are readily provided by packages such as S-PLUS, which was used for all the analyses presented here. Figure 3 shows the year effect curve $s_{10}(t)$ from a GAM with 10 d.f. fitted to CBC data for the corn bunting, together with partial deviance residuals and twice standard error bands. Details of the calculation of these are given in Chambers & Hastie (1993). If the fit is satisfactory, residuals should be distributed evenly above and below the fitted curve, and standard error bands should be consistently narrow. Regions where the residuals 'track' the fitted curve on one side only, or where the standard error bands are particularly wide, serve as warning signals and may indicate that the data are too sparse in these places. Along the bottom of Figure 3 is a rug-plot, which shows the density of observations at each point in time: solid black indicates a high density of observations. The diagnostic plot in Figure 3 shows a healthy fit, although standard errors are somewhat higher at the beginning and end of the time period; this is partly due to data sparsity at these points, and also to the fact that the GAM fit is always less reliable at the endpoints of the range.

For data sets as extensive as those from the CBC, and given the flexibility of the GAM, it is unlikely that the fit will be noticeably poor unless the data are exceptionally sparse. If the choice of degrees of freedom is too low, there might not be sufficient flexibility to represent the true trend accurately, and patterns in the residuals might alert to this: the value of d.f. should then be raised accordingly. If the data are sparse, then a value of d.f. that is too high might result in overfit and a lower value could be used. In reality, however, the parametrization used for the smooth term in the model (12) is negligible compared with the number of parameters required to fit the site effects; reformulation of the model in a more parsimonious covariate framework (section 3.7) would therefore be a better means of coping with sparse data.

A number of approaches to automatic selection of the value of d.f. are outlined in Hastie & Tibshirani (1990), such as generalized cross-validation and Akaike's Information Criterion (AIC). Similar methods may be used in models with several covariates in the additive predictor, to determine which to include. However, the methods can be cumbersome and are not always effective, and for the model presented here the choice of d.f. is more usefully based on the objectives of the analysis than on an automatic selection criterion.

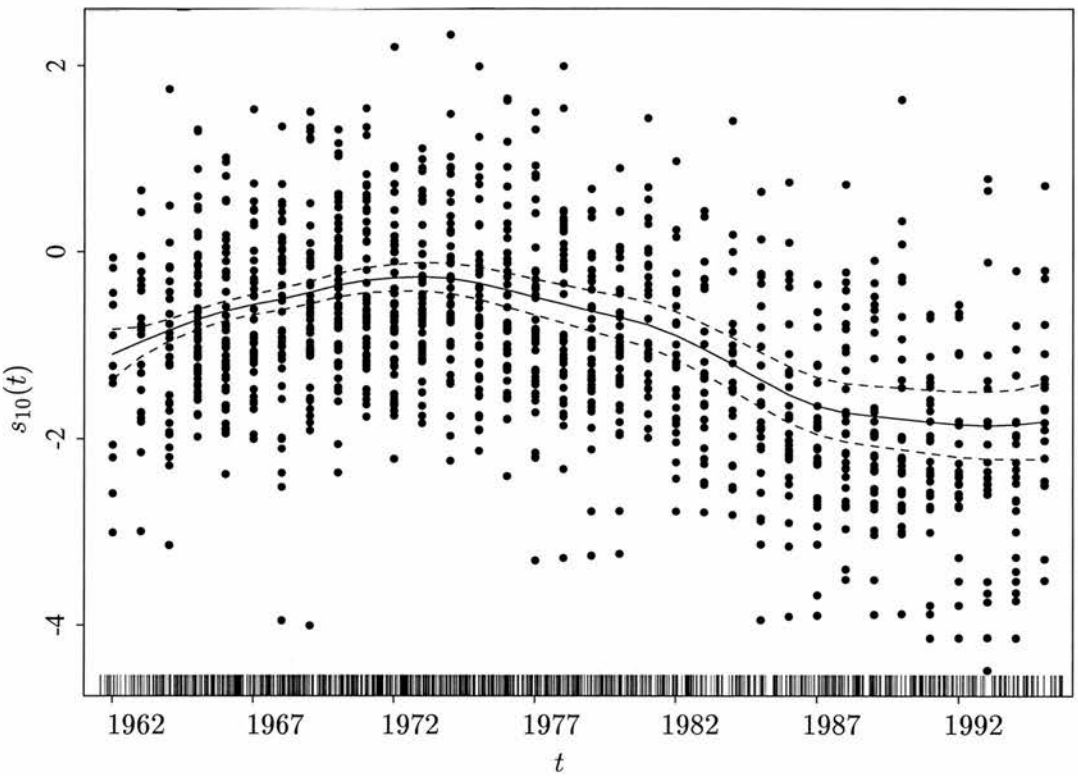


Figure 3: Year effect curve $s_{10}(t)$ from a GAM with 10 d.f. fitted to CBC data for the corn bunting (solid line), with twice standard error bands (dashed lines), partial deviance residuals (black dots), and rug-plot. The residuals are well distributed above and below the curve; the standard error bands are mostly narrow, widening at the endpoints where the observations are more sparse.

Some feeling for the suitability of the model (12) may also be obtained graphically (Mountford 1982). According to (12), all sites in the survey are subject to the same temporal trend on a log scale. For any site i in two different years t_1 and t_2, the following relationship should therefore be satisfied:

$$\left.\begin{array}{rcl} \log \mu_{it_1} & = & \alpha_i + s(t_1) \\ \log \mu_{it_2} & = & \alpha_i + s(t_2) \end{array}\right\} \quad \Rightarrow \quad \frac{\mu_{it_1}}{\mu_{it_2}} = \frac{\exp s(t_1)}{\exp s(t_2)}.$$

This indicates that the ratio of expected counts in any two years should be constant for all sites, and equal to the ratio of abundance indices for the two years. It is instructive to plot the observed counts for years t_1 and t_2, for all sites that were surveyed in both years, to see whether the data support this assumption. Plots for two species over four pairs of years are shown in Figures 4 and 5. Each dot represents a single site i that was surveyed in both year t_1 and year t_2. The straight line gives the ratio of abundance indices obtained from a GAM with 10 d.f. fitted to the full data for each species. If the model assumption is reasonable, the dots should follow a linear pattern along the plotted line: this would

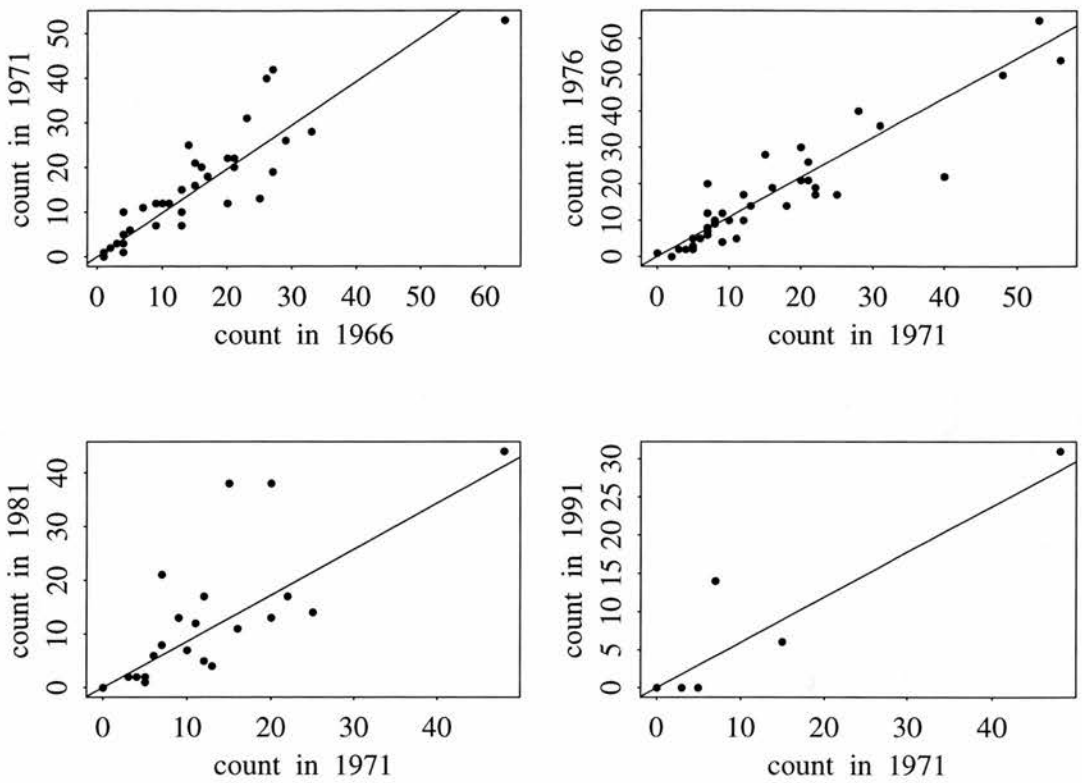


Figure 4: Scatterplots of skylark counts in selected pairs of years for those sites surveyed in both years. Each dot represents a single site. The lines give the estimated proportional change between the years, from a GAM with 10 d.f.

indicate that population abundance is changing uniformly across all sites, and that the proportional change has been correctly estimated by the GAM.

The plots for the skylark (Figure 4) seem to bear good testimony to this assumption. Variation about the straight line increases as the two years t_1 and t_2 become further apart: this is to be expected as there is greater likelihood that the site habitat will change significantly over a longer time period. With the goldfinch (Figure 5), counts are lower and the sample size is smaller, making the relationship less clear, although a linear pattern does not seem unreasonable. The two plots shown are fairly representative of the plots for the other 11 species examined.

Formal tests of linearity are not included here, since it should be emphasized that the sites included in Figures 4 and 5 are only a subset of the full dataset. Indeed, there may be an association between longevity of a survey site and lack of change in habitat over the survey period, which would make these sites unrepresentative of the survey sample. Furthermore, even when the trend is heterogeneous across sites, curvature in the plots would only emerge if the heterogeneity were correlated with abundance — otherwise the pattern would still be broadly linear but with wide scatter about the line. While a markedly non-linear pattern

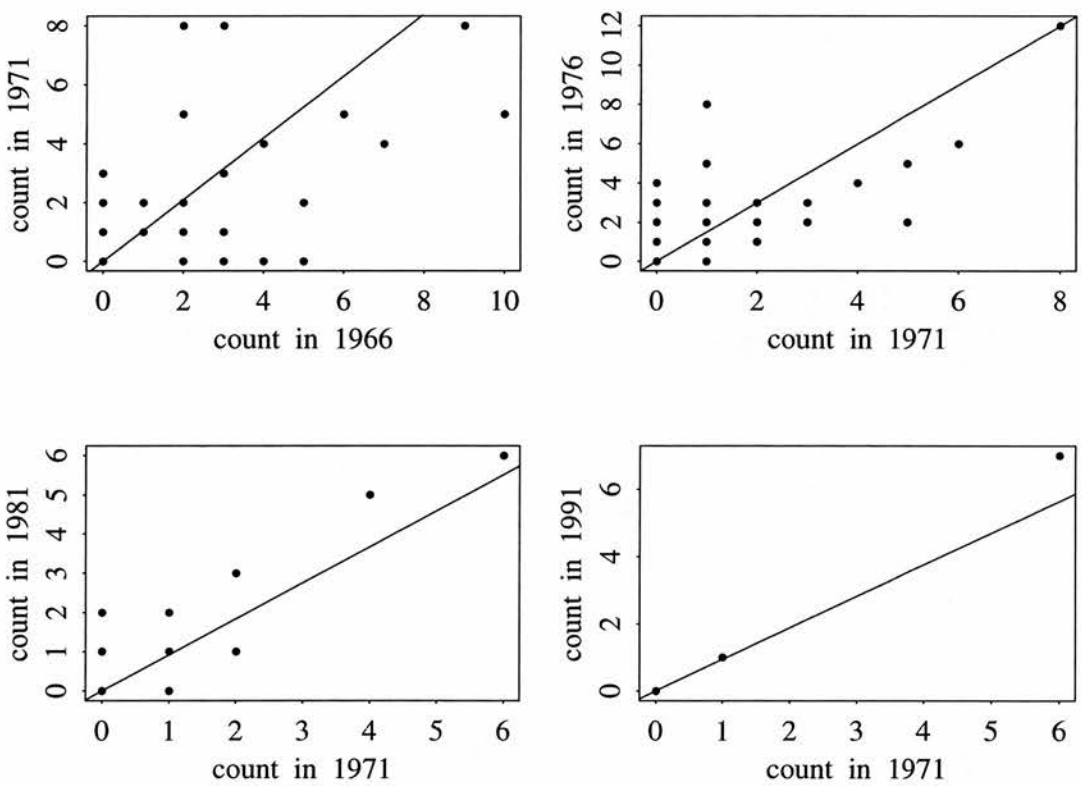


Figure 5: Scatterplots of goldfinch counts in selected pairs of years for those sites surveyed in both years. Each dot represents a single site. The lines give the estimated proportional change between the years, from a GAM with 10 d.f.

in the plots gives definite grounds for concern, therefore, a linear plot does not prove that all is well.

If there is reason to believe that the assumption of uniform trend across all sites is invalid, the regional model described in section 3.7 is recommended. There, a separate trend is accommodated within each of several regions. Interestingly, despite the promising appearance of the plots for the skylark in Figure 4, application of the regional model to these data indicated that the variation in trend between regions was significant (section 3.7).

3.5 Inference from the GAM indices

Inference about population trend may be drawn directly and unambiguously from the fitted abundance indices, since trend is modelled explicitly in the GAM: so periods of decline, increase and stability may be identified by visual inspection of the abundance curve, and are unconfounded by the noise of annual fluctuation. Attention is focused on the detection of patterns in abundance rather than on the magnitude of the indices, for the reasons given in Section 1. Accordingly, while the statistical significance level associated with a decline or increase is estimated, no conclusions are drawn about the magnitude of the change in real terms.

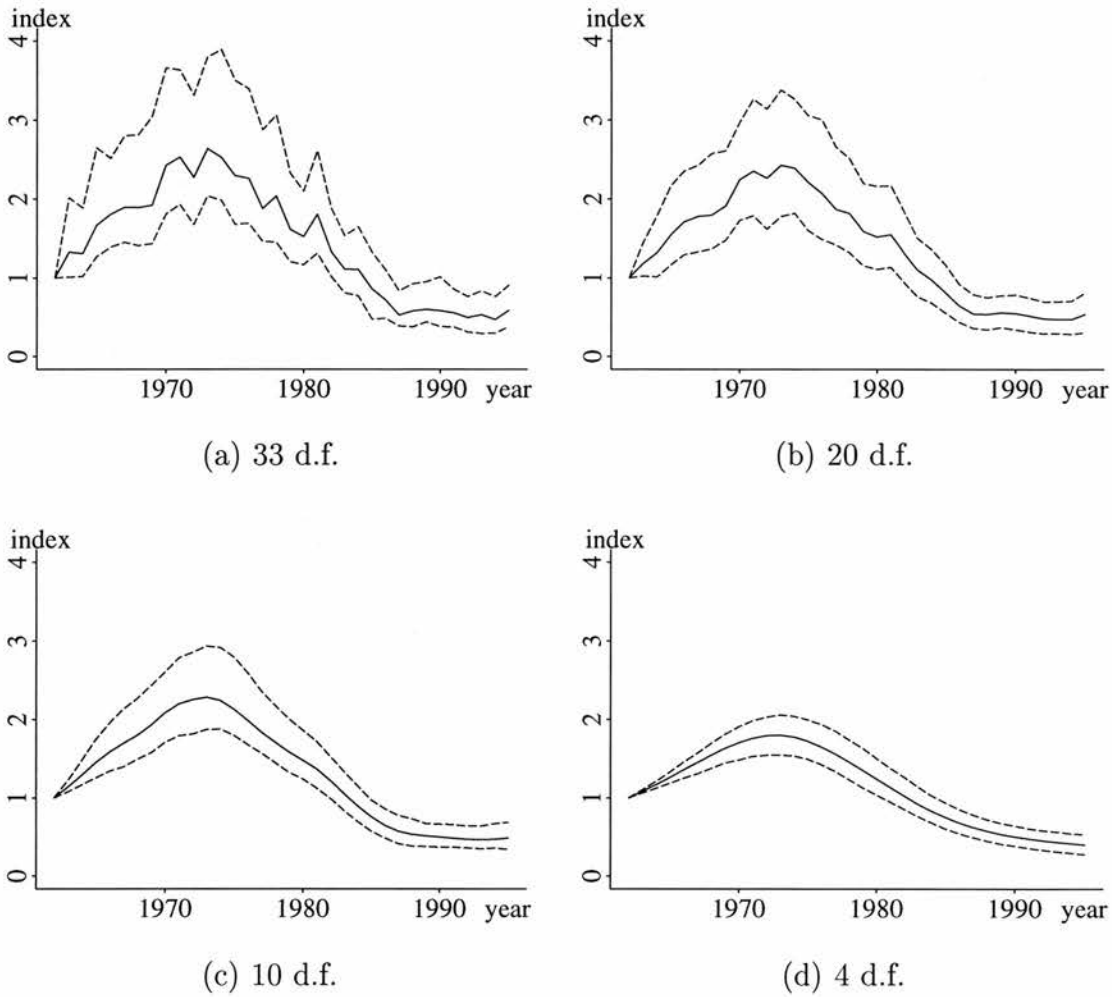
In order to quantify the significance of apparent change, estimates of precision are required for the abundance indices. Approximate confidence intervals for the abundance indices are obtained by means of the bootstrap. This is considered preferable to the use of standard error estimates provided for the fitted year effect curve $\hat{s}_d(t)$ by software packages such as S-PLUS, for a number of reasons. Firstly, use of standard error bands to provide confidence intervals demands that an assumption be made of the statistical distribution of the fitted year effects. Although normality of the fitted year effects will be assumed in the informal tests to follow, it is desirable to make as few such assumptions as possible. Secondly, the calculation of standard error estimates for smooth terms in S-PLUS involves a series of approximations (Chambers & Hastie 1993); the estimates obtained are adequate for detecting spurious features of fit as described in section 3.4, but attempts to use them in computing intervals with a precise confidence level are likely to be misleading. Thirdly, the abundance index curve is obtained through an exponential transformation (13) of the fitted year effect curve; to obtain standard errors of the abundance indices, the standard errors of the fitted year effects would therefore also need to be transformed. In order to be accurate this requires information about the covariances of the fitted year effects, which

is not available in S-PLUS and is extremely difficult to obtain (Chambers & Hastie 1993).

Bootstrapping proceeds by drawing for each of a given number of replicates a random sample of size N, with replacement, from the original N sites. The N sites in the sample are treated as distinct, although in practice there will be some duplicates. A GAM is fitted to the sample from each replicate, and the annual abundance indices are calculated. After B bootstrap replicates, there are B values for the abundance index in each year. These are sorted into ascending order, and approximate $100(1-2\alpha)\%$ confidence limits for this year are provided by the lth lowest and uth highest values, where $l = (B+1)\alpha$ and $u = (B+1)(1-\alpha)$ (Buckland 1984). This procedure provides a percentile interval which may be treated as an approximate confidence interval: the condition under which the approximation is exact is given in Buckland (1984).

Confidence intervals provide an informal test for the significance of population change: if the indices for two given years are assumed independent and normally distributed with standard errors of comparable size, then to a good approximation the difference between the indices is significant at the 5% level if there is no overlap in their 85% confidence intervals (Buckland *et al.* 1992). The test is two-tailed, and is surprisingly robust to differences in standard errors: standard errors differing by a factor of up to about 2 are quite acceptable. The assumption of independence is reasonable as long as the comparison involves indices separated by some years. To enable this test to be performed easily, all index curves are presented with 85% confidence intervals, obtained via the bootstrap from $B = 119$ bootstrap replicates. Bootstrapping has the disadvantage of being a very lengthy process, since a new GAM must be fitted for each replicate: for the 13 species presented here, between 1 and 12 minutes were required per bootstrap replicate in S-PLUS (version 3.4 for UNIX). For practical purposes this limits the number of bootstrap replicates that can be made. Substantial improvements in speed are achieved by use of dedicated software such as the Fortran program GAMFIT (Hastie & Tibshirani 1990), although this does not provide a full interface for subsequent analyses.

Bootstrapped 85% confidence intervals for abundance indices are illustrated in Figure 6 for four different GAMs applied to CBC data for the corn bunting. The most notable feature is the narrowing of the confidence interval as the amount of smoothing is increased. This is to be expected, because a model with a low level of smoothing incorporates annual fluctuations and is sure to have high uncertainty in the output, while a model with high levels of smoothing concentrates on trend and has much less associated uncertainty. This demonstrates a further advantage of incorporating smoothing procedures directly into



(a) 33 d.f.

(b) 20 d.f.

(c) 10 d.f.

(d) 4 d.f.

Figure 6: Bootstrapped 85% confidence limits (dashed lines) for the abundance indices from GAMs with various degrees of freedom applied to CBC data for the corn bunting.

the model fitting: the confidence intervals are much narrower than those that would be obtained when simply smoothing the output from a previous model. For example, if the output from the GAM with 33 d.f. (which corresponds to the log-linear Poisson regression model in (7)) were smoothed, the resulting confidence intervals would inherit all the uncertainty of annual fluctuation and remain as wide as those in Figure 6 (a); the output from a GAM with smoothing incorporated directly into the model, on the other hand, yields narrower confidence intervals as in Figure 6 (b), (c) and (d). Of course, if too much smoothing is incorporated the trend curve does not have sufficient flexibility to represent the true pattern of population change, so a balance must be found.

The bootstrap procedure also allows estimation of the pointwise standard errors of the annual abundance indices, by using the sample standard error of the indices from the bootstrap replicates. This enables quantification of the significance of population change

between any two years t_1 and t_2. The indices $\mathcal{I}(t_1)$ and $\mathcal{I}(t_2)$ are assumed independent and normally distributed, with means ν_1 and ν_2 respectively and variances given by the bootstrap estimates. The significance of the population change between years t_1 and t_2 is taken to be the significance level at which the null hypothesis $H_0 : \nu_1 = \nu_2$ is rejected.

The normal distribution is used as a reference distribution for these significance tests purely out of convenience: there is no evidence that this is the true distribution of the indices. The tests are, therefore, only approximate. However, they are not highly sensitive to the degree of smoothing chosen in the analysis, and they remain valid for detecting change in abundance even when the number of detections provides a distorted estimate of true abundance.

3.6 Analysis of second derivatives

The ability to use techniques such as the second derivative analysis presented in this section is one of the principal advantages of the GAM approach. The GAM provides a smooth differentiable curve for the abundance indices, and the derivatives reveal information about the shape of the curve.

The second derivative of the index curve $\mathcal{I}(t)$ at time t is a measure of the curvature of $\mathcal{I}$ at that time. If the second derivative is greater than zero, the curve is turning upwards ($\smile$), while if it is less than zero the curve is turning downwards ($\frown$). The magnitude of the second derivative indicates the tightness of the curvature: a second derivative of approximately zero signifies a linear region of the index curve, meaning that the population trajectory is changing at a steady rate. Years in which the second derivative is markedly different from zero are those where something is happening to alter the rate of population change. These years are referred to as *change-points*, and it is shown how change-points that are statistically significant may be identified. The timing of the change-points might help to suggest causes of the change.

The second derivatives of $\mathcal{I}$ are not directly available as mathematical expressions since $\mathcal{I}$ is a non-parametric curve; instead numerical gradient estimates are used. The first derivative, or gradient, of $\mathcal{I}$ at time t is approximated by

$$\mathcal{I}^*(t) = \frac{\mathcal{I}(t + \frac{r}{2}) - \mathcal{I}(t - \frac{r}{2})}{r} \tag{14}$$

for some small r. Substituting equation (14) into itself yields an estimate of the second

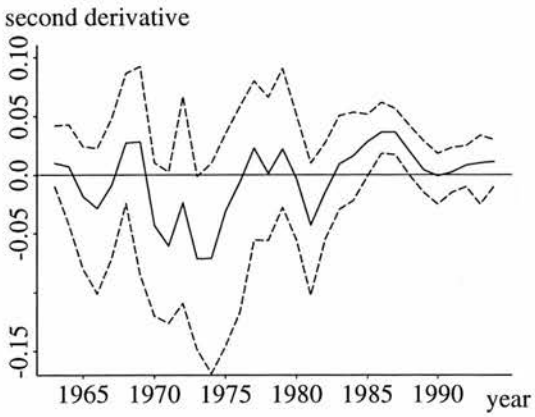
derivative at point t:

$$\begin{aligned} \mathcal{I}^{**}(t) &= \frac{\mathcal{I}^{*}(t+\frac{r}{2}) - \mathcal{I}^{*}(t-\frac{r}{2})}{r} \\ &= \frac{\mathcal{I}(t+r) - 2\mathcal{I}(t) + \mathcal{I}(t-r)}{r^2}. \end{aligned} \tag{15}$$

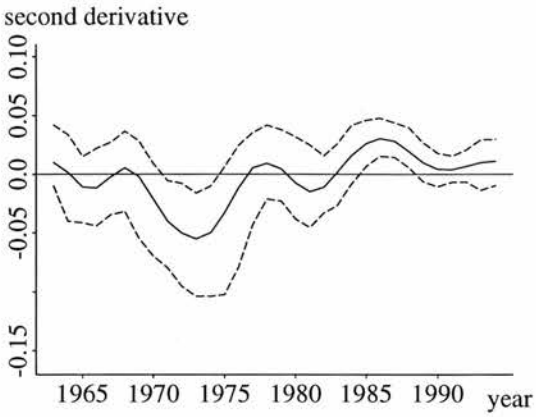
Once a value for r has been chosen, therefore, estimation of the second derivative is straightforward. The quantity r is referred to as the *window size* — the length of the interval used in gradient estimation. Although the smallest values of r give the most accurate second derivative estimates, there is some merit in choosing a higher value of r since this has the effect of smoothing the second derivative curve and making it easier to interpret. Figure 7 shows second derivative estimates obtained from the CBC index curve for the corn bunting, for three different values of r. A satisfactory compromise between small r and smooth output occurs at $r = 3$ years.

In order to detect years where the second derivative is significantly different from zero, bootstrapped approximate confidence intervals are provided for the second derivative curve. For each bootstrap replicate, a GAM is fitted, model indices are obtained, and the associated curve of estimated second derivatives is found. The confidence limits are determined by ordering the second derivative point estimates for each year and selecting the appropriate upper and lower percentiles as described in section 3.5. The approximate confidence intervals may be used to elicit an approximate hypothesis test of the same significance level: the hypothesis that the second derivative estimate for any year derives from a distribution with mean zero is rejected if the confidence interval for the second derivative estimate does not contain the point zero. From a 95% confidence interval, the two-tailed hypothesis may be rejected at the 5% level. If the distribution of the second derivative satisfies certain symmetry conditions then the hypothesis test is exact (Buckland 1984).

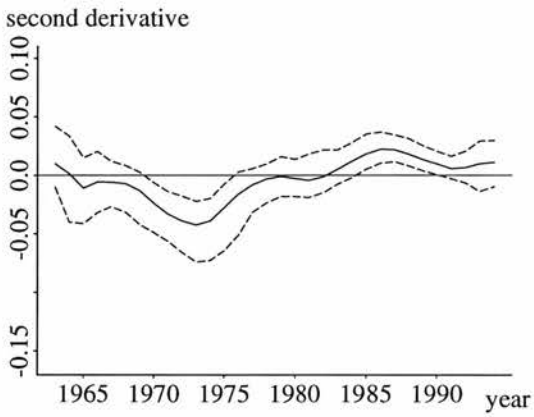
Bootstrapped 95% pointwise confidence intervals for the second derivatives of the corn bunting index curve are shown in Figure 7. The plot from a window size of 1 is too jagged to enable viable inference, so all results in this chapter will use a window size of 3. The significant change-points at the 5% level are those years in Figure 7(b) for which the confidence intervals do not contain the point zero: that is, 1971–1974 and 1985–1988. Between 1971 and 1974 the second derivative curve is significantly less than zero, indicating a change for the worse: the population curve turns downwards. The converse is true for the period 1985–1988: the second derivative is significantly positive, indicating an upturn,



(a) Window size 1.



(b) Window size 3.



(c) Window size 5.

Figure 7: Estimated second derivative curves (solid lines) for the abundance indices obtained from a GAM with 10 d.f. fitted to CBC data for the corn bunting. The dashed lines are bootstrapped 95% confidence intervals.

or lower rate of decline, in the population. These features are verified by examination of the corresponding abundance index curve for the corn bunting in Figure 6(c).

3.7 Covariate models

A full covariate model may be incorporated into the GAM framework by adding extra terms to the additive predictor in (12). The extra terms might contain information on site habitat, in which case the site effects α_i could be replaced by more parsimonious expressions involving habitat covariates such as percentage crop cover or site area. Alternatively the covariates might represent geographical location of the sites, or average daily minimum temperatures if hard winters are thought to have had an impact on abundance.

Each variable in the additive predictor may be regarded as either a categorical, parametric continuous or non-parametric continuous variable. For example, in (12) the site number i is a categorical quantity, while t is a continuous variable treated non-parametrically via the expression $s(t)$. A parametric treatment of the time variable occurs in the GLM formulation $\log(\mu_{it}) = \alpha_i + \gamma t$ (equation (6)) while an alternative categorical treatment is given in the GLM formulation $\log(\mu_{it}) = \alpha_i + \beta_t$ (equation (7)).

The advantages of the covariate approach lie in parsimony and in the accommodation of separate trends for every site. However, if site effect parameters are to be dispensed with altogether, care must be taken to include all covariates of potential importance in the model. Serious bias in prediction could arise if a major source of variation is omitted. If many covariates are to be included, model selection becomes an important consideration; stepwise model selection may be conducted by means of AIC or other criteria (Chambers & Hastie 1993).

The potential of covariate models will be illustrated here by inclusion of a categorical variable representing the geographical region in which each site is located: north-east, north-west, south-east or south-west. The four regions divide the available CBC plots into approximately equal samples around easting 45 and northing 21 of the UK national grid (Marchant & Gregory 1994). Dominant habitats differ between regions: for example the north-east region contains the low-lying, intensively arable land of East Anglia and Lincolnshire, while the other regions are dominated by grazing and mixed farms. The four regions also differ in topography, temperatures and patterns of rainfall.

The purpose of the material in this section is to indicate how extra terms may be added to the model, and how their significance may be tested. The region variable is included as an interaction term: that is, a region-by-time interaction that fits a separate smooth trend $s(t)$ in each of the four regions. Note that this increases the parametrization considerably: a smooth trend with d degrees of freedom in each of four regions corresponds to a total of $4 \times d$ equivalent parameters on the trend terms, as opposed to only d equivalent parameters if the region effect is not included. If a species exhibits marked regional variation in trend, however, the extra parameters are likely to be worthwhile.

The additive predictor for the GAM with region-time interaction becomes

$$\log(\mu_{it}) = \alpha_i + s_d^{(k)}(t) \tag{16}$$

for the mean count μ_{it} in site i at time t, where k denotes the region in which site i is

located ($k = 1, 2, 3, 4$), and $s_d^{(1)}(t)$, $s_d^{(2)}(t)$, $s_d^{(3)}(t)$ and $s_d^{(4)}(t)$ are the four corresponding smooth regional trends, each with d d.f. The index curve $\mathcal{I}(t)$ is once again the ratio of total predicted abundance (across all regions) in year t to that in year 1. At present, there is no facility in S-PLUS for fitting an interaction between a categorical variable and a smooth term. With the simple model here, the fit is accomplished using four separate GAMs: one for each region. With more complicated models, there might be further covariates that must be fitted to data from all regions simultaneously, such as percentage crop cover of site i; in that case it is not possible to fit separate GAMs and an iterative scheme is required. It is hoped that future developments of S-PLUS will incorporate this facility.

The significance of the region-by-time interaction may be tested using the analysis of deviance technique (Hastie & Tibshirani 1990). The statistical distribution of difference in residual deviance between the models with and without regional trend is approximated by a χ^2 distribution, with degrees of freedom equal to the difference in degrees of freedom between the two models —i.e. $4 \times d - d = 3d$. The region-by-time interaction is considered significant if the observed decrease in deviance due to its inclusion is extreme for the χ^2_{3d} distribution. Although it is known that the difference in residual deviance between the two nested GAMs does not have a χ^2 distribution, even asymptotically, the approximation has been shown to provide a useful means of model selection (Hastie & Tibshirani 1990). If the dispersion parameter is estimated from the data, for example when there is overdispersion, the usual F-test approximation may be applied: the F-statistic is obtained by dividing the difference in residual deviance between the two models by $3d$ times the mean deviance for the full model with interaction.

4 Examples

4.1 Results

The methods of Section 3 are now applied to CBC data for the 13 species listed in section 3.1. The generalized additive model given by equation (12) was fitted to the data for each species, using a smoothing spline with 10 d.f. on the year variable. The index curve $\mathcal{I}(t)$ was calculated, bootstrapped 85% confidence intervals and pointwise standard error estimates for $\mathcal{I}(t)$ were obtained, and the second derivative of the index curve was estimated for each bootstrap replicate. Significant change-points were identified using a

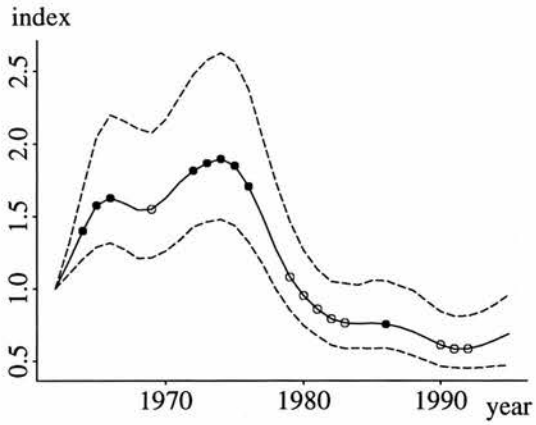
95% bootstrapped confidence interval for the second derivative curve, with a window size of 3 years. Diagnostic plots for the 13 GAMs were all good.

The index curves for 12 species are shown in Figure 8; results for the skylark are shown separately in Figure 9. From examination of the 85% confidence intervals, it appears that significant change at the 5% level has occurred between the beginning and end of the time period for all species except goldfinch and reed bunting; the reed bunting population has seen significant change during the period but an initial population rise was negated by a subsequent decline. Of the remaining species, all but the chaffinch, greenfinch and stock dove have experienced significant decline.

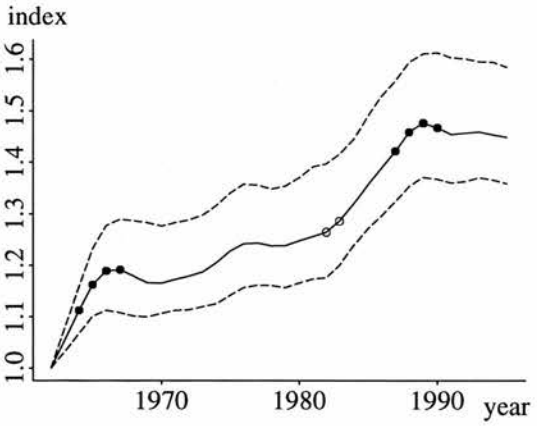
More formal quantification of the significance of long-term changes is presented in Table 1. Three years have been selected for comparison: 1965, 1975 and 1992. Since the GAM fit can be unreliable at the end-points of the time period, use of the results from 1962–1964 at one end and 1993–1995 at the other is avoided; this ensures inference is not biased by spurious end effects. Many of the 13 species have a peak in abundance between 1973 and 1976, so 1975 was selected as a suitable intermediate point for comparison. For years t_1 and t_2, which may be any two of 1965, 1975 and 1992, it is assumed that $\mathcal{I}(t_1) \sim N(\nu_1, \sigma_1^2)$, $\mathcal{I}(t_2) \sim N(\nu_2, \sigma_2^2)$, and $\mathcal{I}(t_1)$ and $\mathcal{I}(t_2)$ are independent. Under the null hypothesis $H_0 : \nu_1 = \nu_2$, the distribution of $\mathcal{I}(t_1) - \mathcal{I}(t_2)$ is $N(0, \sigma_1^2 + \sigma_2^2)$. The significance level given in Table 1 is that at which the null hypothesis H_0 would be rejected in a two-tailed test against a general alternative hypothesis. Estimates of σ_1^2 and σ_2^2 are obtained from the bootstrap replicates.

The results in Table 1 confirm the patterns detected from visual inspection of the indices: bullfinch, corn bunting, grey partridge, linnet, skylark, tree sparrow, turtle dove and yellowhammer populations all experienced highly significant declines between 1975 and 1992; little overall change was recorded for the goldfinch, greenfinch and reed bunting, and only chaffinch and stock dove populations increased significantly between 1965 and 1992. The yellowhammer is unusual in that its decline began in the mid-1980s rather than in the 1970s. The stock dove population has undergone a dramatic increase, although exceptionally high variance in the fit reduces the associated significance level. The high variance is probably due to the fact that about 20% of the recorded count for the stock dove was obtained from only six sites.

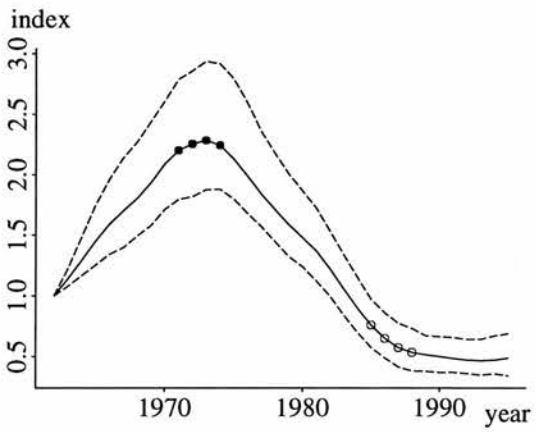
Significant change-points from second derivative analysis are marked on the index curves in Figure 8. These reveal some striking patterns, more easily visualized in Table 2. From



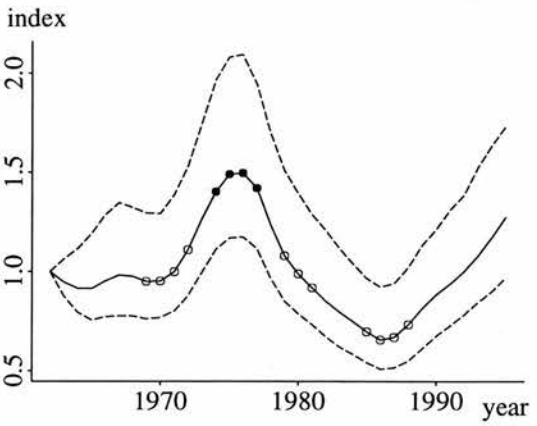
(a) Bullfinch.



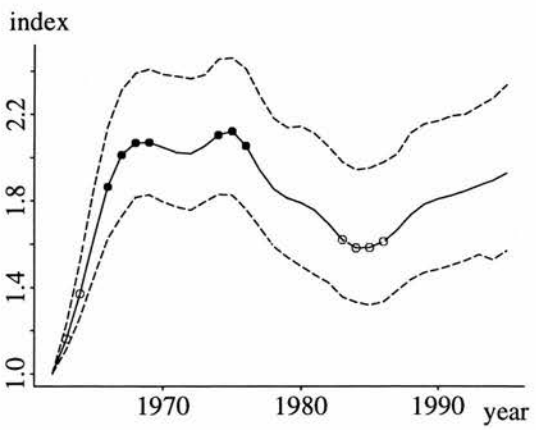
(b) Chaffinch.



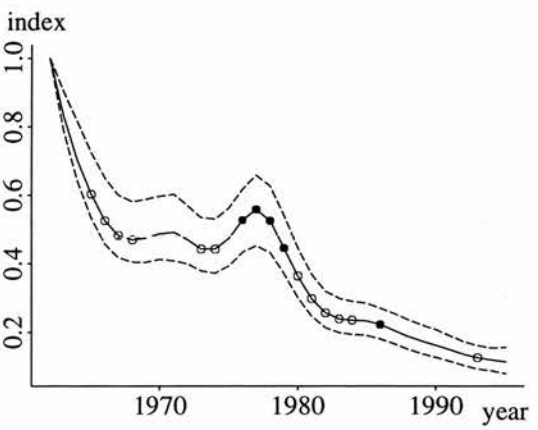
(c) Corn bunting.



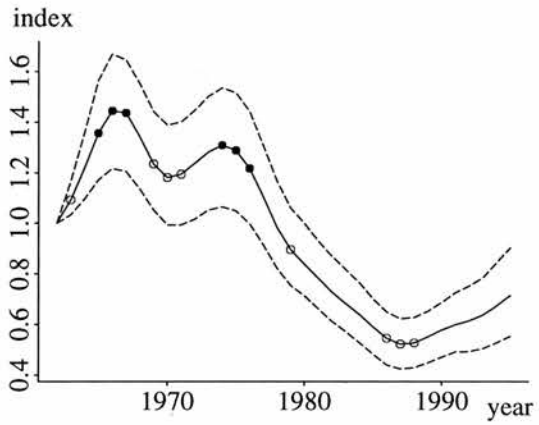
(d) Goldfinch.



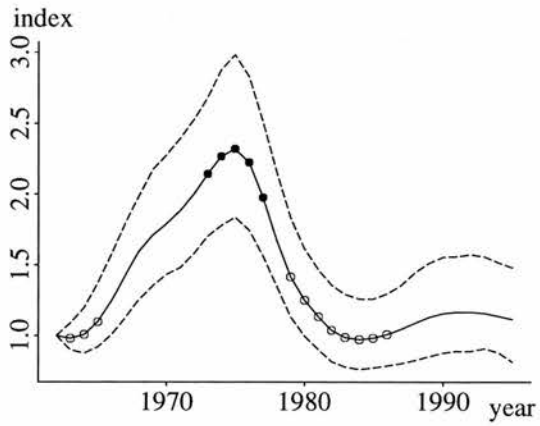
(e) Greenfinch.



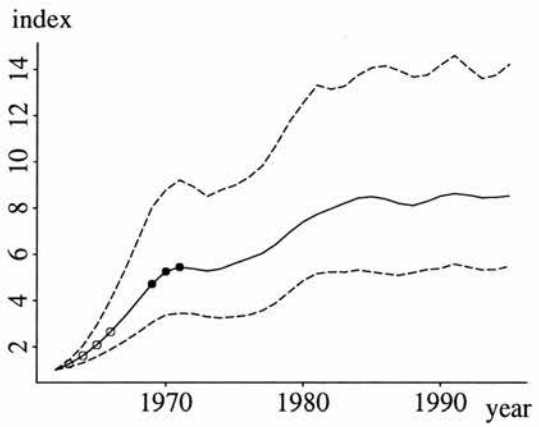
(f) Grey partridge.



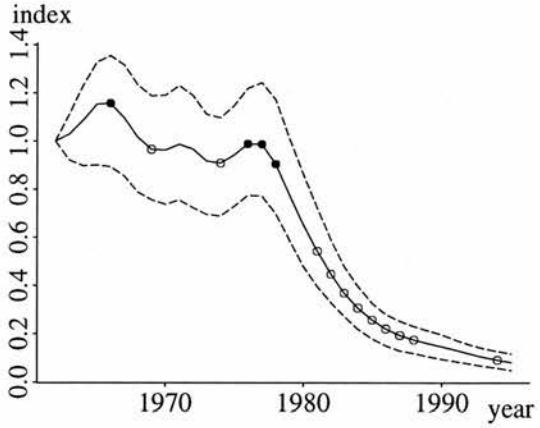
(g) Linnet.



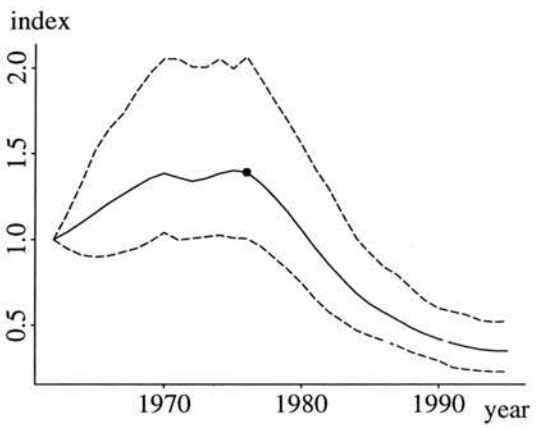
(h) Reed bunting.



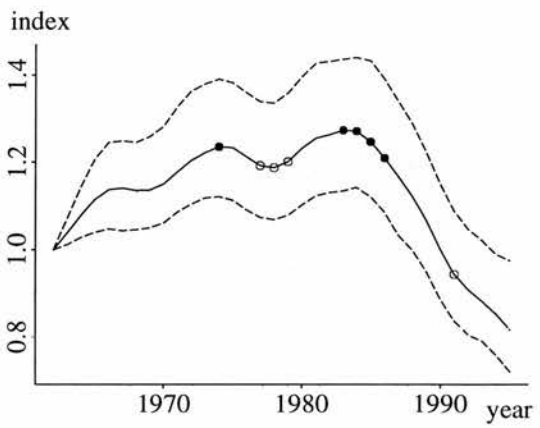
(i) Stock dove.



(j) Tree sparrow.



(k) Turtle dove.



(l) Yellowhammer.

Figure 8: Index curves from CBC data for 12 farmland species. The solid lines give the index curves from a GAM with 10 d.f., and the dashed lines represent upper and lower 85% bootstrapped confidence limits. Significant change-points are marked on the index curves: a filled circle (•) denotes a point at which the second derivative is significantly negative (a downturn in the index curve), and an unfilled circle (○) denotes a point at which the second derivative is significantly positive (an upturn in the index curve).

Species	1965–1975 change : p–value	1975–1992 change : p–value	1965–1992 change : p–value
Bullfinch	up : 0.53	down* : 7.9×10^{-4}	down* : 3.2×10^{-4}
Chaffinch	up : 0.43	up* : 0.032	up* : 0.0018
Corn bunting	up : 0.082	down* : 7.9×10^{-6}	down* : 9.7×10^{-7}
Goldfinch	up : 0.098	down : 0.20	up : 0.75
Greenfinch	up : 0.057	down : 0.40	up : 0.41
Grey partridge	down : 0.14	down* : 5.1×10^{-8}	down* : 9.8×10^{-12}
Linnet	down : 0.73	down* : 9.5×10^{-5}	down* : 1.7×10^{-6}
Reed bunting	up* : 0.0058	down* : 0.016	up : 0.81
Skylark	up : 0.27	down* : 3.0×10^{-7}	down* : 6.2×10^{-7}
Stock dove	up : 0.064	up : 0.38	up* : 0.026
Tree sparrow	down : 0.35	down* : 4.3×10^{-7}	down* : 9.3×10^{-11}
Turtle dove	up : 0.59	down* : 0.013	down* : 0.0026
Yellowhammer	up : 0.27	down* : 0.0077	down* : 0.036

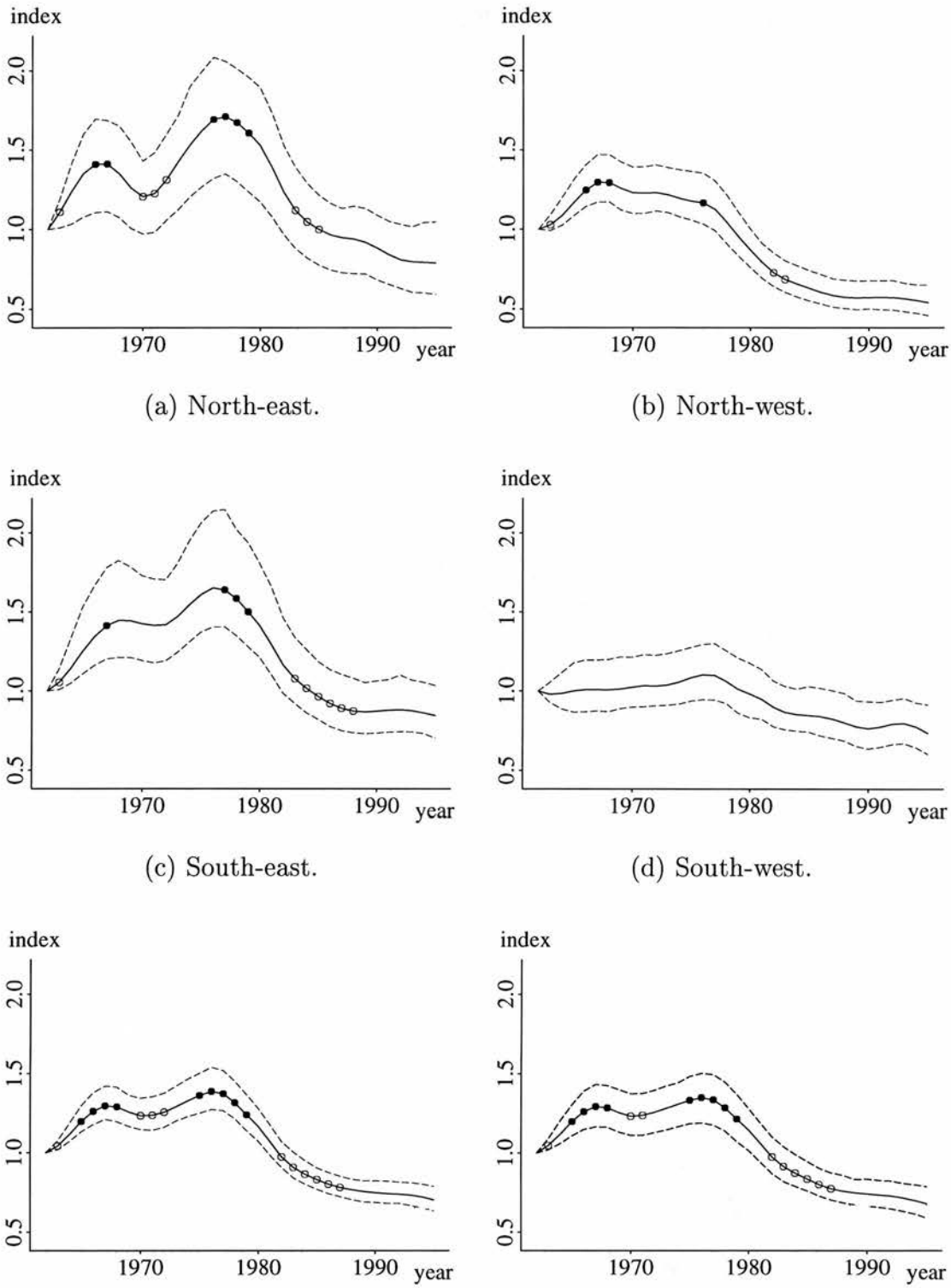
Table 1: Directions and significances of population change for 13 farmland species using output from GAMs with 10 d.f. fitted to CBC data, together with bootstrapped standard error estimates. Changes are considered between 1965 and 1975, between 1975 and 1992, and between 1965 and 1992. p–values are from a two-tailed test of no change between the indices. The smaller the p–value, the more significant is the change. Changes for which the p–value is less than 0.05 are marked with an asterisk (*).

Table 2 it is clear that all species except the chaffinch and stock dove experienced a significant downturn in population trajectory between 1972 and 1978, corresponding in all cases to the beginning of a period of decline. Conversely, between 1980 and 1989, a significant upturn in population trajectory occurred for most species, corresponding in some cases to the beginning of a period of increase (reed bunting, linnet, greenfinch, goldfinch and chaffinch) and in others to a decrease in the rate of decline (tree sparrow, skylark, grey partridge, corn bunting and bullfinch).

	Year																															
Species	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
Bullfinch		●	●	●			○			●	●	●	●	●			○	○	○	○	○			●				○	○	○		
Chaffinch		●	●	●	●															○	○				●	●	●	●				
Corn bunting									●	●	●	●											○	○	○	○						
Goldfinch							○	○	○	○		●	●	●	●		○	○	○				○	○	○	○						
Greenfinch	○	○		●	●	●	●					●	●	●							○	○	○	○								
Grey partridge			○	○	○	○					○	○		●	●	●	●	○	○	○	○	○		●							○	
Linnet	○		●	●	●		○	○	○			●	●	●			○							○	○	○						
Reed bunting	○	○	○								●	●	●	●	●		○	○	○	○	○	○	○	○								
Skylark	○		●	●	●	●		○	○				●	●	●	●	●			○	○	○	○	○	○							
Stock dove	○	○	○	○			●	●	●																							
Tree sparrow				●			○					○		●	●	●			○	○	○	○	○	○	○	○						○
Turtle dove														●																		
Yellowhammer												●			○	○	○				●	●	●	●					○			

Table 2: Significant change-points for the 13 farmland species. Filled circles (●) denote significant downturns in the population trajectory, unfilled circles (○) denote significant upturns.

The region-time interaction model is illustrated using the skylark data. Independent smooth trend functions were fitted for the four separate regions, each one with 10 degrees of freedom. The results are shown in Figure 9: bootstrapped 85% confidence intervals and significant change-points are presented for each region. There are some differences between the four regional trends shown in Figure 9 (a) to (d): in particular the skylark population appears to have remained almost stable in the south-west while in the eastern regions it has undergone noticeable oscillation. In all regions, however, a period of stability or increase lasting until the late 1970s was followed by a period of decline. The overall abundance indices from the regional model are shown in Figure 9 (e): the index for year t is the ratio of total predicted count over all regions in year t to that in year 1. The confidence intervals for the overall indices are much narrower than those from the regional



(a) North-east.

(b) North-west.

(c) South-east.

(d) South-west.

(e) All regions.

(f) Non-regional model.

Figure 9: Results from the region-by-time model for CBC skylark data. Index curves are shown as solid lines, bootstrapped 85% confidence limits as dashed lines and change-points as filled and unfilled circles. Regional index curves are shown in (a)–(d); in (e) the regional results are pooled to yield overall abundance estimates, and in (f) the results of the original model (without the region-by-time interaction term) are shown.

curves; to understand this it should be noted that the regional index curves are all ratios, not absolute values, and in fact the sum across all regions is dominated by the count from the north-west region. The overall indices therefore inherit the low variance of those from the north-west. The indices from the original model without region-time interaction (specified by equation (12)) are shown in Figure 9 (f), and are almost identical to those from the regional model in (e). In this case at least, the regional term makes negligible difference to inference.

Analysis of deviance tests suggest that the regional model provides a significant improvement in fit for the skylark: a drop of 180.3 in residual deviance for an increase of 30 equivalent parameters is highly significant for the χ^2_{30} distribution (for significance at the 5% level, a decrease in residual deviance of 43.8 or more is required). The decision as to whether to use the regional model or the original model therefore hinges on the objectives of the analysis: in the case of the skylark, the regional model highlights discrepant regional trends and gives more accurate predictions at the level of individual sites, but has little bearing on inference about overall trend.

The regional model was also fitted to CBC data for the chaffinch and bullfinch: the regional effect was found to be significant at the 5% level for the chaffinch (drop in deviance = 177.7) but not for the bullfinch (drop in deviance = 39.6). In the case of the chaffinch an overall population increase occurred in all regions, but while the increase was steady in the western regions, there was a fall in abundance between 1972 and 1980 in the eastern regions. For the bullfinch the pattern of change was broadly similar across all regions.

4.2 Interpretation

The trend curves of Figure 8 and the change-points in Table 2 reveal some conspicuous patterns across the 13 farmland species considered. The mixed fortunes of species in the mid-1960s probably reflect the different responses to the harsh winter weather of 1962–63. Depending on ecology and demography, by 1965–66 most affected species were either reaching the end of a period of steep decline, or arriving at a plateau after a steep recovery (Baillie 1990; Marchant *et al.* 1990; Greenwood & Baillie 1991). The 1970s saw a range of changes in agriculture which have been thought detrimental to farmland bird populations (O'Connor & Shrubb 1986; Fuller *et al.* 1995; Baillie *et al.* 1997; Siriwardena *et al.* 1998), borne out by the collection of negative change-points observed in this interval. Some variation in the precise timing of the change-points is expected as a result of differences in

lifespan and reproductive strategy between species. For example, coincident depressions in breeding success for a short-lived and a long-lived species tend to be manifested in the abundance of the short-lived species much earlier, while concurrent depressions in adult survival are apparent first in the abundance of the long-lived species. This is especially true if individuals of the short-lived species mature earlier and produce more offspring than those of the long-lived species.

While downturns dominated the 1970s, the reverse is true of the 1980s. For the most part the upturns of the 1980s have been followed by a period of relative stability, albeit at a lower level than prior to the mid-1970s. The farmland populations of the 13 seed-eating species have therefore failed to recover from their severe declines. In some cases this might be indicative of a national decline in abundance; in others, the birds may have been forced into less preferred habitat.

The regional model (16) applied to data for the skylark reveals interesting differences in regional trend. The steeper declines in the east may reflect the arable farmland that is characteristic of these regions: agricultural changes over the last 30 years have been more dramatic in arable areas than on grazing lands, and the baseline population density in the early 1960s was lower in arable parts. Population increases are seen in the mid-1960s for all regions except the south-west: these increases probably represent population recoveries following the harsh winter weather of 1962-1963, which was less severe in the south-west region. Use of the regional model is encouraged for all species, at least for exploratory data analysis.

The significant change-points identified from the GAM index curves in Figure 8 may be compared against those obtained from secondarily smoothed Mountford indices by Siriwardena *et al.* (1998). The change-points found by the two methods correspond fairly closely, as do the smoothed abundance curves. The GAM approach is to be favoured due to better precision, but the close correspondence with the results from the Mountford method indicates that earlier analyses need not be discarded.

5 Extensions and conclusions

The basic GAM formulation presented in this chapter can be extended in many ways. There are a number of choices for error distribution and link function: an obvious extension would be investigation of the negative binomial distribution in place of the Poisson, with

the logarithmic link function. Addition of covariates into the basic model might be one of the most promising lines of future enquiry and could lead to a parsimonious model able to accommodate a separate trend in each site.

Monitoring schemes should be designed to cover a representative sample of sites; otherwise the trends obtained must strictly be interpreted as trends over selected habitat rather than true national trend. CBC sites were selected by observers, although the plots chosen have been shown to be representative of lowland farmland in southern Britain (Fuller *et al.* 1985). Continuity of survey cover is also important: continuous coverage of a small sample of sites is more valuable for trend analysis than patchy coverage of a larger sample. These issues have been addressed in the UK Breeding Bird Survey (Gregory *et al.* in press) which is in the process of replacing the CBC.

When the survey covers a random sample of sites, generalized additive mixed models (GAMMs) might be worth considering. Site effects in a GAMM are no longer regarded as fixed unknowns, as in equation (12), but as realizations from a parametric distribution whose parameters are to be estimated. Without a random sample of sites, the data might lie predominantly in the tail of this distribution and bias in prediction would result. The mixed model would be more parsimonious than the fixed effects model, although some investigation of a suitable trend index would be required. Although there has been much recent interest in methodology for generalized linear mixed models (e.g. Schall 1991; Engel & Keen 1994), there seems to be little literature on generalized additive mixed models. Development of such models might prove a valuable tool for estimating trends in wildlife populations from annual surveys of randomly selected sites.

In conclusion, the GAM approach presented here provides the most general and flexible framework currently available for trend analysis of census data, surpassing the chain and Mountford methods and incorporating all the models of sections 2.4, 2.5 and 2.6 as special cases. Similarities between GAM indices and those from the Mountford method and log-linear Poisson regression for the CBC data are encouraging, although inference based on more primitive techniques such as the chain method should be re-assessed.

Acknowledgements

This work was partially funded by the UK Ministry for Agriculture, Fisheries and Food, under contract CSA3109. The biological interpretations and references in section 4.2 were provided by Gavin Siriwardena and Stephen Baillie of the British Trust for Ornithology, Thetford, and by Jeremy Wilson of the Royal Society for the Protection of Birds.

Part II

Colonization model

Chapter 3

Introduction to the colonization model and the woodlark data

The following chapters will be concerned with approaches to the fitting of one particular mechanistic spatio-temporal model, when survey data are incomplete. The methods will be applied to a set of survey data for the woodlark *Lullula arborea*, a scarce songbird whose UK population is concentrated in the south of England. The survey data are taken from Thetford Forest, a conifer plantation in the Breckland region of Norfolk and Suffolk, UK.

The present chapter introduces the model and dataset, and describes the difficulties with model-fitting that motivate the subsequent work.

1 The colonization model

1.1 Model description

The model to be studied will be referred to as the *colonization model*, and was proposed by Buckland & Elston (1993) for investigating the change in spatial distribution of a wildlife population over time. A similar but more basic model was applied by Besag (1977) in an epidemiological context, to analyse the spread of footrot in endives.

The colonization model is described as follows. The survey area consists of N sites, which might be chosen arbitrarily, for example as atlas grid squares, or might be determined

ecologically as patches of suitable habitat such as forest clearings or marshes. An initial survey of the species of interest is taken over the N sites, and a record of presence or absence is made for each site. The survey data are assumed to be complete, and the time of the initial survey is defined as time 0. The sites are numbered $1, \ldots, N$.

A second survey is taken over the same N sites some time later, defined as time 1. The population is assumed to be closed in that all individuals present in the survey region at time 1 have ancestry in the individuals present at time 0. If individuals present in site h at time 1 are descendants of individuals present in site i at time 0, the site i is said to be an *ancestor* of the site h. Each site may have several ancestors. The *colonization probability* p_{ih} is defined as the probability that site i is an ancestor of site h: that is, that individuals or their offspring present in site i at time 0 have colonized site h by time 1. The colonization probabilities are the quantities for which a model is proposed.

An occupied site i at time 0 may be the ancestor of many sites or of none; there is no stipulation that the probabilities $\{p_{ih}\}_{h=1}^{N}$ should sum to 1, and they are not Markov chain transition probabilities. Colonizations of and from sites are assumed to be independent over a single time period: that is, the event that site i is an ancestor of site h is independent of the event that site i is an ancestor of site k for all i, h and k; and the event that site i is an ancestor of site h is independent of the event that site j is an ancestor of site h for all i, j and h. This assumption is not unreasonable as long as the time period is short compared with the rate at which colonizations take place — in particular, for a one-year time period it will be reasonable for many species other than the strongly territorial. The size of the sites relative to the territory size of the species of interest will also affect the validity of the independence assumption.

Figure 1 provides an illustration of the colonization mechanism. In most cases, the probability p_{hh} that site h remains occupied over the time period will be less than 1 for all sites h. This represents an important difference between the colonization model and many epidemiological models, in which an 'occupied' unit corresponds to an infected individual, and recovery is not possible.

The model for the colonization probabilities $\{p_{ih}\}$ may take any form, but a simple version of the form suggested by Buckland & Elston (1993) is

$$p_{ih} = p_0 \exp\left(-a\,\delta_{ih} - b\,\varsigma_h\right) \tag{1}$$

where δ_{ih} is the Euclidean distance between sites i and h, ς_h represents a habitat suitability

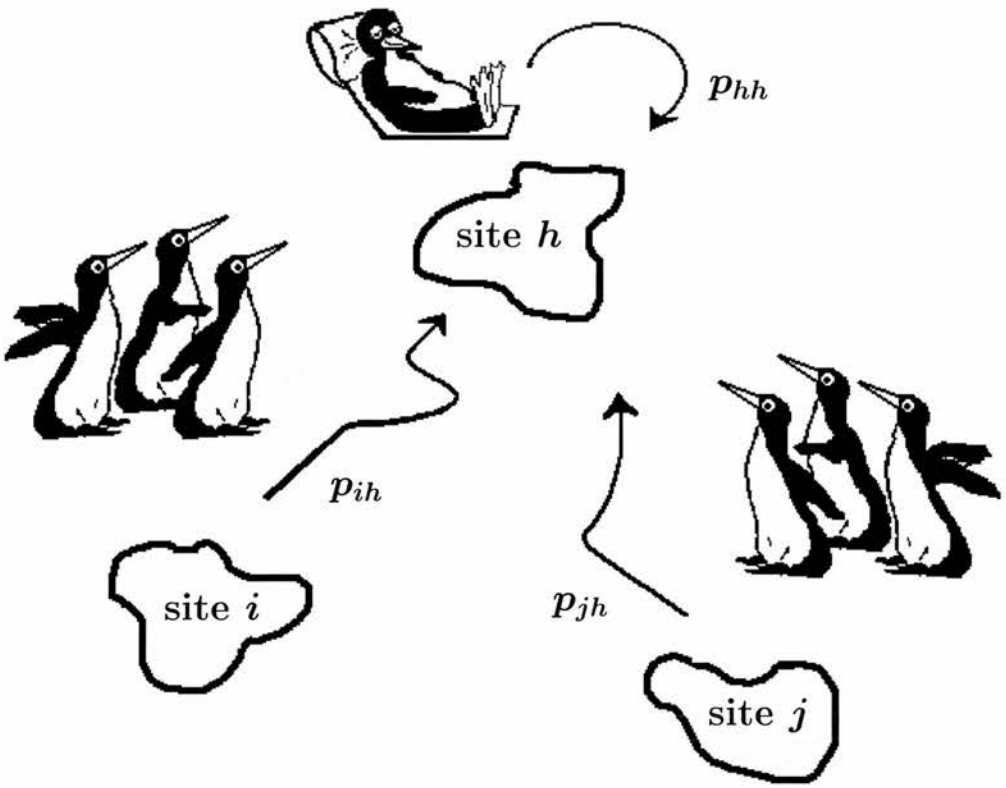
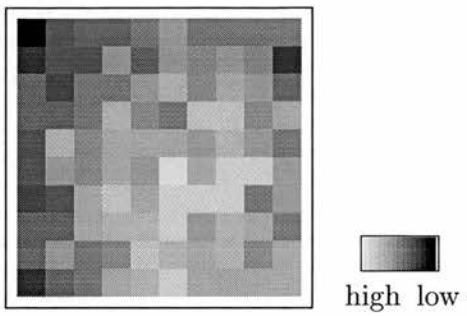


Figure 1: The movement of individuals among sites in the survey region between times 0 and 1 determines the pattern of occupation at time 1.

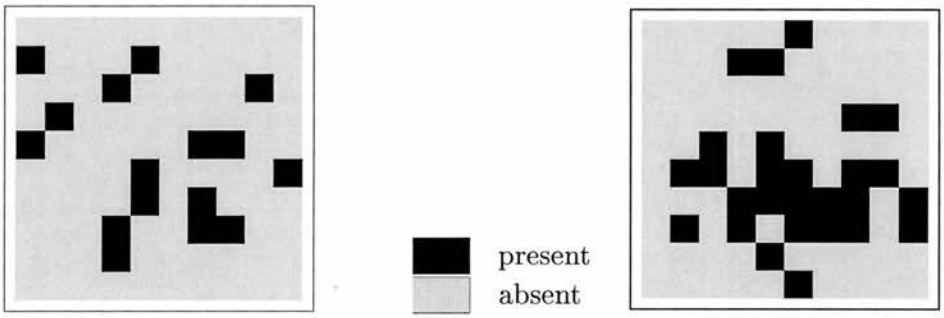
score for site h at time 1, and p_0, a and b are parameters to be estimated, with $a, b \geq 0$, $0 < p_0 \leq 1$. The habitat suitability score ς_h is based on ecological knowledge, and is a measure of the quality of habitat in site h for the species of interest. The measure decreases as habitat quality improves. One way of obtaining the suitability values is to conduct a logistic regression of presence/absence on habitat covariates in a small region where the species is well-established, to determine the dependence of species' occurrence on habitat when the issue of non-occupation due to remoteness from the current distribution may be disregarded. The fitted logistic model may then be used to estimate the fitted probability π_h that any other site h would be occupied based on habitat cover alone. Once the values $\{\pi_h\}$ have been computed for all sites h, a simple transformation may be applied to yield the suitability scores. The transformation recommended by Buckland & Elston (1993) is

$$\varsigma_h = \frac{\max_k\{\pi_k\} - \pi_h}{\max_k\{\pi_k\} - \min_k\{\pi_k\}}, \tag{2}$$

whereby the most suitable site has a habitat suitability score of 0, and the least suitable site has a score of 1. This is the method that will be adopted here. The details of habitat modelling are beyond the scope of the present discussion, and the suitability scores



(a) Site suitabilities.



(b) Pattern at $t = 0$.

(c) Pattern at $t = 1$.

Figure 2: A simulated set of site suitabilities is shown in (a), depicting a region of good habitat in the centre and poorer habitat around the edges. An initial distribution of presence/absence is given in (b), and a second distribution in (c). Between the two surveys the population distribution has consolidated in the region of good habitat but dwindled in the poorer habitat.

are henceforth assumed to have been determined prior to the fitting of the colonization model. Note that the logistic model operates in a descriptive capacity only, by determining occupation probabilities as a function of habitat in the absence of other considerations. The colonization model, on the other hand, aims to emulate the process by which a species can expand into new areas or contract in range, and takes into account the fact that occupation of a site is not determined entirely by habitat quality.

Equation (1) provides a model for the colonization probabilities that renders colonization between two sites i and h more likely if the sites are close together than if they are remote from each other, and more likely if the target site h contains habitat of high suitability for the species of interest. By estimating the parameters a, b and p_0 of the model, an impression can be formed as to the mechanism behind the movement of individuals between sites — for example, which of habitat suitability and remoteness is the dominant factor in determining probability of occupation.

An example of the data required for the application of the colonization model is shown in Figure 2, where site suitabilities and two occupation distributions have been simulated on a grid of 100 site squares. Presented with information such as this, the aim is to estimate the parameters of the colonization probabilities (e.g. a, b and p_0) that have the effect of producing the final distribution, given the initial distribution and the site suitabilities.

1.2 Fitting the one-stage colonization model

If the colonization model extends over a single time-step, the likelihood function with respect to the parameters of the colonization probabilities is readily calculated (Buckland & Elston 1993). Let $\boldsymbol{y^{(0)}} = \left(y_1^{(0)}, \ldots, y_N^{(0)}\right)$ be the observed distribution of presence/absence at time $t = 0$: each $y_i^{(0)}$ is 1 or 0 according to whether site i is occupied or unoccupied respectively. Similarly, let $\boldsymbol{y^{(1)}} = \left(y_1^{(1)}, \ldots, y_N^{(1)}\right)$ be the observed distribution at time $t = 1$, and suppose that the parameters of the colonization probabilities $\{p_{ih}\}$ are summarized by the vector $\boldsymbol{\theta}$; e.g. $\boldsymbol{\theta} = (a, b, p_0)$.

The *occupation probability* p_h is defined for site h as the probability that h is occupied at the time $t = 1$ of the second survey. Since colonizations from all sites i at time 0 to site h at time 1 are independent, the probability that site h is unoccupied at time 1 is given by

$$1 - p_h \;=\; \prod_{i\,:\,y_i^{(0)}=1} (1 - p_{ih})$$

— that is, the product over those sites i occupied at time 0 of the probability that site i does not colonize h over the single time step. The occupation probabilities are thus expressed in terms of the parameters $\boldsymbol{\theta}$ of the colonization probabilities.

The one-step likelihood $L\left(\boldsymbol{y^{(1)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(0)}}\right)$ now follows immediately from the fact that colonizations to all sites h occur independently of each other:

$$L\left(\boldsymbol{y^{(1)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(0)}}\right) = \prod_{h\,:\,y_h^{(1)}=0} (1 - p_h) \prod_{h\,:\,y_h^{(1)}=1} p_h \,. \tag{3}$$

The likelihood is a function of the parameters $\boldsymbol{\theta}$ through the occupation probabilities $\{p_h\}$, and may be maximized with respect to $\boldsymbol{\theta}$ to yield the maximum likelihood parameter estimates. Note that the likelihood function corresponds to a discrete probability function rather than a probability density.

1.3 Extension to a T–stage model

Suppose now that the survey period extends over more than a single time-step. If survey data is acquired from equally-spaced times $0, 1, 2, \ldots, T$, the overall likelihood of the observations $\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}}, \ldots, \boldsymbol{y^{(T)}}$ conditional on the initial distribution $\boldsymbol{y^{(0)}}$ is

$$L\left(\boldsymbol{y^{(T)}},..,\boldsymbol{y^{(1)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(0)}}\right) = L\left(\boldsymbol{y^{(T)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(T-1)}}\right) L\left(\boldsymbol{y^{(T-1)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(T-2)}}\right)\ldots L\left(\boldsymbol{y^{(1)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(0)}}\right),$$

because the distribution at any time-point depends only on the distribution at the previous time-point, as with a Markov process. The overall likelihood is therefore a product of single-step likelihoods, each of which may be calculated as in equation (3). The parameters $\boldsymbol{\theta}$ are constant over the whole time period.

However, difficulties arise if surveys did not take place at some or all of the intermediate times $1, 2, \ldots, T-1$. Suppose, for example, that surveys were carried out in years numbered 0, 3, 4 and 6. In order to make optimal use of the data in determining the parameter values, a joint likelihood

$$L\left(\boldsymbol{y^{(6)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(4)}}\right) L\left(\boldsymbol{y^{(4)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(3)}}\right) L\left(\boldsymbol{y^{(3)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(0)}}\right) \tag{4}$$

is required. The parameters $\boldsymbol{\theta}$ must be chosen to relate to a time period of a particular length: if $\boldsymbol{\theta}$ relates to a time period of length ℓ, for example, the colonization probability $p_{ih} = p_{ih}(\boldsymbol{\theta})$ denotes the probability that site h is colonized at time $t + \ell$ by individuals or their offspring that were present in site i at time t, for $t = 0, 1, \ldots$.

If the parameters $\boldsymbol{\theta}$ relate to the interval of a single year, the one-step component of the likelihood $L\left(\boldsymbol{y^{(4)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(3)}}\right)$ is calculable, but neither $L\left(\boldsymbol{y^{(6)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(4)}}\right)$ nor $L\left(\boldsymbol{y^{(3)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(0)}}\right)$ may be obtained in the same way. Similarly, parameters relating to either two or three years are suitable for one of the factors in (4), but not for the other two. If the colonization model is to be applied to data stemming from a number of surveys at irregular intervals, therefore, some method of computing a likelihood relating to a longer time period than that of the parameters is required. In the above example the parameters would be chosen to relate to an interval of one year, and some way of overcoming the problem of missing surveys would be required for the calculation of the likelihoods $L\left(\boldsymbol{y^{(6)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(4)}}\right)$ and $L\left(\boldsymbol{y^{(3)}}\middle|\boldsymbol{\theta},\boldsymbol{y^{(0)}}\right)$. Note that the time period spanned by the parameters cannot be greater than the shortest interval between surveys: there is no possibility of accommodating data taken after a one-year interval if the parameters relate to an interval of two years.

The issue of obtaining parameter estimates for the colonization model when some surveys are missing is the principal theme of Chapters 5 and 6. One approach is to sum over all possible intermediate distributions: e.g.

$$L\left(\boldsymbol{y}^{(2)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right) = \sum_{\boldsymbol{y}^{(1)}} L\left(\boldsymbol{y}^{(2)}, \boldsymbol{y}^{(1)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right) .$$

Since the sum ranges over 2^N possible distributions $\boldsymbol{y}^{(1)}$, however, this approach is computationally intractable. The time required for a survey region covering a modest $N = 100$ sites, using a computer capable of 1 billion likelihood evaluations per second, would be some 4×10^{11} centuries. This is thought to be excessive.

1.4 Further reasons for dividing the time period

The problem of fitting the colonization model when there are missing survey data may be summarized as the computation of a likelihood relating to a longer time period than that of the parameters. Although the primary reason for developing methodology for coping with this circumstance is to enable the model to be fitted to data from several surveys taken at irregularly-spaced intervals, there are also other situations where the methodology could prove useful.

Suppose for example that there are two surveys with an elapsed time of T years between them. It would be possible to define the time of the first survey as time 0 and the time of the second survey as time 1, and fit a one-step model for which a single time-step involves T years. However, if the fitted model is to be used to obtain predictions of species' distribution subsequent to the time of the second survey, the fitted parameters will yield predictions only at T-year intervals. It might be important to obtain a prediction for some time that is not an integral multiple of the sampling period T, and for this the original survey period must be divided appropriately and the fit must accommodate missing intermediate surveys. When $T = 10$, for instance, and a predicted species' distribution is required 5 years after the time of the second survey, the model could be fitted with a time-step of 5 years. This would leave a single missing survey at time 1 : 5 years after the first survey, but would ensure that the required prediction could be obtained.

Another reason for dividing the time period is to improve the validity of the independence assumption. Recall that colonizations of and from sites are assumed to be independent over a single time period. This assumption is most defensible when the length of a single

time period is short — otherwise there is time for unobserved sequences of colonizations to take place, and independence does not hold. One such colonization sequence is illustrated in Figure 3, which was obtained by simulating from the colonization model itself. A single occupation in the bottom right-hand corner of the distribution for year 2 gives rise to further occupations in that corner by years 6 and 7. The occupations in year 7 are not conditionally independent given the distribution in year 0, because they have all arisen from the same colonization in year 2. Attempts to apply the colonization model using a single time-step from year 0 to year 7, therefore, might yield inaccurate results as the independence assumption is not valid for an interval of this length. On the other hand, if the period between surveys were divided into smaller time units and the model fitted as if there were missing intermediate surveys, the assumption of independence might be applied far more confidently.

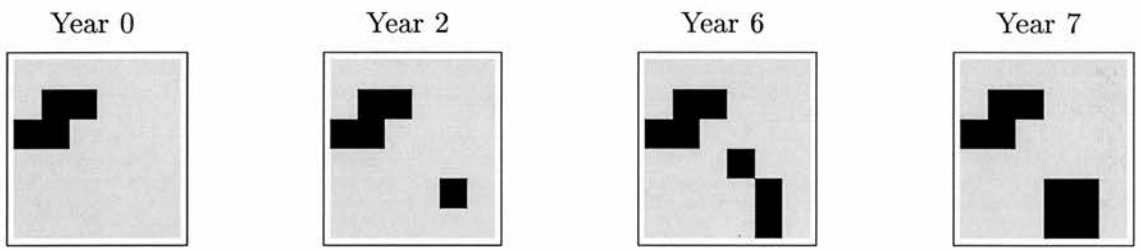


Figure 3: Illustration of a colonization sequence.

2 The woodlark data

The woodlark data were collected between 1986 and 1990 in a comprehensive study covering habitat selection, diet and breeding success of woodlarks in Thetford Forest. The project was funded by the UK Forestry Commission and the Royal Society for the Protection of Birds (RSPB), with a view to determining management strategies that would benefit the woodlark population in the forest. Full details of the study are given in Bowden & Green (1992). Figure 4 shows a map of south-east Britain with the approximate location and extent of Thetford Forest.

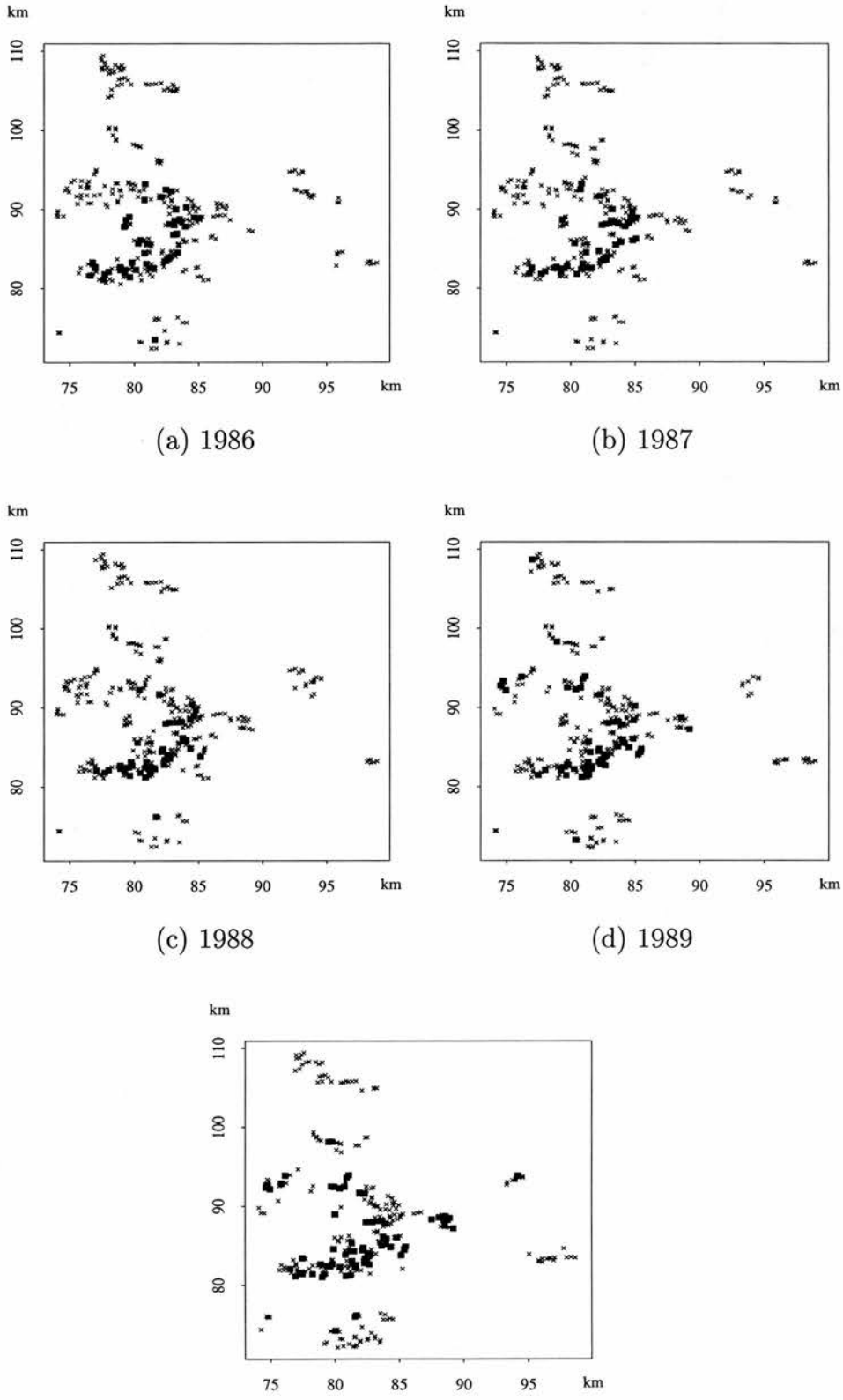
Thetford Forest is divided for management purposes into compartments of about 15 to 25 hectares each. The compartments are further divided into sub-compartments according to the age or species of the trees; these are convenient units to use as sites. The composition of the forest is approximately 37% Scots Pine, 48% Corsican Pine, 5% other pine and 10% broadleaved trees (Bowden & Green 1992).

Figure 4: Map of south-east Britain showing the approximate location of Thetford Forest.

The habitat requirements of woodlarks are quite particular, with the ideal habitat comprising a mixture of bare ground and short grass and herbs, together with an available song-post. Regions of forest with mature trees are unsuitable for woodlarks, probably due to the associated lack of foraging area (Bowden & Green 1992). Regions where trees have been newly felled or recently planted, however, do provide good habitat, with the woodlarks exhibiting a preference for plantations with trees less than 5 years old. The forestry policy therefore has a clear effect on the availability of suitable woodlark habitat, as areas are constantly logged and replanted. In any year, the woodlark survey included only those sites with trees of age up to 7 years; the other sites were discarded as unsuitable.

The woodlark population increased in Thetford Forest over the survey period 1986–1990, notably over the final two years of the period. The numbers of sites occupied by woodlarks from 1986 to 1990 were respectively 45, 40, 43, 57 and 74, out of a total of respectively 338, 320, 358, 333 and 316 suitable sites in the forest. Figure 5 shows the occupation patterns for the five years. The occupied sites almost always corresponded to a single pair of breeding birds, or a territory-holding unpaired male. The song of the male woodlarks makes them much easier to detect than females, and is also a key identification feature. Only on very rare occasions did more than one pair occupy a single site, and on no occasion more than 3 pairs.

The habitat suitability scores for the colonization model were obtained by a logistic regression on habitat covariates in the manner described in the previous section. A stepwise model selection procedure was carried out, and the final model included the following covariates (Bowden & Green 1992): site area; year; percentage bare ground; percentage



(a) 1986

(b) 1987

(c) 1988

(d) 1989

(e) 1990

Figure 5: Distributions of woodlarks in suitable sites in Thetford Forest from 1986 to 1990. A suitable but unoccupied site is marked by a cross, and a suitable occupied site by a black square.

debris; percentage brash; percentage moss; percentage short grass; percentage bracken; percentage shrub cover. All of these except site area, year and percentage bracken were included as quadratic terms. The resulting suitability scores were transformed using the procedure of equation (2), and suitabilities were re-computed for each year of the survey.

The colonization model is well-suited to the woodlark data, since habitat and distance from sites currently colonized are both thought to affect woodlark occupation. Male woodlarks rarely move more than 0.5 km between successive territories from one year to the next (Bowden & Green 1992). Birds in their first breeding season are likely to establish a territory between 1 km and 10 km from their natal site. Furthermore, to a large extent the Thetford community is closed: most of the breeding individuals in the forest are thought to have hatched there. It is known that some emigration occurs out of the forest — ringed individuals have been found holding territories as far away as Lincolnshire — but this is rather unusual, perhaps involving only 3–8% of young birds fledged in any year (Bowden & Green 1992). The extent of immigration into Thetford Forest is not known, but for the purposes of the colonization model it will be assumed negligible. This assumption is not unreasonable given the limited movements of most woodlarks.

The application of the colonization model will provide information that could be useful for management purposes. If the parameter estimates indicate that distance from the current distribution is a crucial factor in determining site occupation, then it would be important for new clear-felled regions to be created close to currently occupied sites: otherwise, management effort could be wasted in regions that are unlikely to be colonized by woodlarks, and intense competition in other areas might lead to starvation or low breeding success. Since the future habitat map of the forest is controlled by the Forestry Commission in their felling strategy, predictions of woodlark distributions under various postulated plans might be useful in order to assess which would bring greatest benefit to the population.

As a complete survey, the woodlark data is ideal for testing the methodology to be developed in Chapters 5 and 6. The model may be fitted with some years of survey data artificially omitted, and the resulting fitted models may be checked against those involving all five years of survey data.

3 Outline of Part II

An overview of the following chapters is now provided. Chapter 4 is concerned with the development of indices for assessing the similarity of two spatial distributions of species' occurrence. Such indices are useful for comparing observed distributions with predictions obtained from a fitted model. They can be used to verify that the model fit is satisfactory, and to compare the outputs of different models with a view to selecting the most reliable. The material is not directly connected with the colonization model, but the techniques developed will be used extensively with the colonization model in the ensuing chapters.

Chapter 5 deals with the analytic fitting of the colonization model when there are missing survey data. A maximum likelihood approach is adopted, and methods of constructing an approximate likelihood function are developed. In Chapter 6 the colonization model is extended to the case where even the one-step likelihood (3) cannot be calculated. Simulation approaches are suggested for dealing with this case.

Acknowledgements

I am grateful to Rhys Green of the RSPB Research Department for making the woodlark data available, converting the format of the data files and providing detailed maps of Thetford Forest.

Chapter 4

Similarity coefficients for spatial ecological distributions

1 Introduction

Measures of the similarity between spatial distributions of species' occurrence are often required in ecological studies. A spatial distribution is a map of designated sites or grid squares, each site containing either a record of presence or absence for the species, or a count of species' abundance. Similarity coefficients are used to rank similarity in range and occurrence between different species, or to quantify distributional change for a single species over a number of years. A further use is in model selection: predictions generated from a number of competing models for species' occurrence are matched against the observed distribution, and the model producing the closest match is selected. It is important in this context that the similarity measure is able to reward similarities in global pattern without placing undue penalty on small local differences.

Various measures have been proposed for the quantification of similarity between spatial distributions, each with certain advantages and disadvantages. Digby and Kempton (1987) summarize the more widely used measures, indicating their suitabilities for a range of applications. The most appropriate similarity coefficient for a particular problem is determined by a number of factors: the size of population sample, whether or not presence and absence are to be treated symmetrically, and in the case of abundance data the relative weights attached to large and small counts. However, all of the measures in common use rely on a few simple summary statistics, and as such are unlikely to convey

an adequate impression of global similarity or trend.

An improved method of quantifying similarity is presented in this chapter, focusing on the global features of distributions and ignoring small local discrepancies. The approach is suitable for both presence-absence and abundance data, and a simple illustration of its use in model selection is provided in each instance. The merits and drawbacks of traditional similarity indices for this purpose are discussed briefly.

2 Matching problems in ecology

2.1 Binary data

A *spatial binary distribution* is defined as an arrangement of N sites with species' presence (1) or absence (0) recorded in each. The 'sites' might for example be atlas grid squares, forest clearings or areas of wetland. Sites are numbered $1, 2, \ldots, N$ and are said to have *status* 1 or 0. The aim is to derive a measure of similarity between two binary distributions over the same N sites.

Traditional similarity coefficients are defined in terms of the quantities $\{n_{xy} : x, y = 0, 1\}$, where n_{xy} is the number of sites with status x in the first distribution and status y in the second. Widely used is the simple matching coefficient (SMC), which is given by

$$\frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} \tag{1}$$

and is simply the proportion of matched sites — those having equal status in both distributions. Presence and absence are treated symmetrically in that conjoint absences are allotted the same weight as co-occurrences.

In some contexts it is preferable to judge similarity only on the basis of co-occurrence, thereby ignoring the n_{00} negative matches; the usual coefficient of this type is attributable to Jaccard (1901) and is written as

$$\frac{n_{11}}{n_{11} + n_{10} + n_{01}}. \tag{2}$$

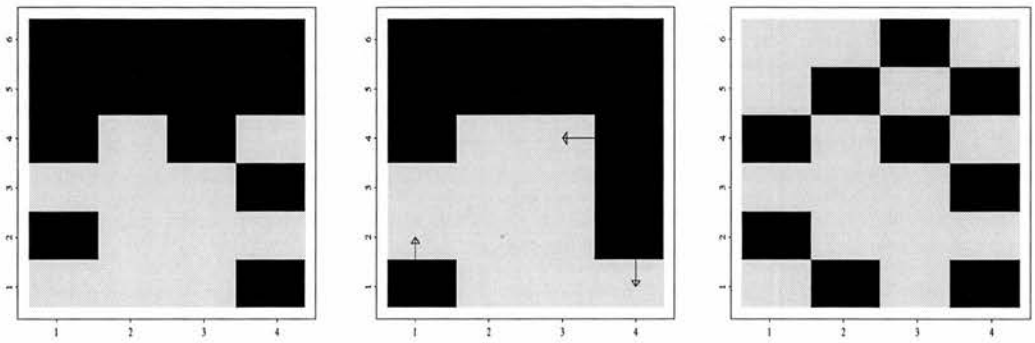
Variants of Jaccard's coefficient have also been proposed but are unlikely to deliver markedly different results; details are given in Digby and Kempton (1987).

The traditional similarity coefficients are easily calculated and make for a quick comparison between distributions, but their values can be misleading. Only the *numbers* of matched and mismatched sites are considered, whereas in fact it is the *positions* of mismatches that have the greatest bearing on overall similarity. A large block of mismatched sites all with the same status is indicative of a global dissimilarity between two distributions, while mismatches that are scattered evenly amongst the sites and are of either status could be attributable to chance alone and need not imply that the distributions are very different. This problem is of particular concern when comparing observed distributions with those generated by statistical models, since a model might yield predictions giving a very good impression of overall trend but attaining only a low similarity coefficient.

A method for improving the similarity coefficients above is developed in this chapter. In order to take account of the positions and status of mismatched sites, the sites are arranged into small groups or cliques. The clique associated with any site i is a set of sites 'close' to i in some sense: two sites are placed in the same clique if it is reasonable to suppose that movement of individuals can occur freely between them. Cliques may be chosen on the basis of common habitat, or according to proximity alone. The clique relation is reflexive and symmetric, but not necessarily transitive — for example sites i and j need not be close enough to each other for movement to occur freely between them despite both being close enough to a third site k.

A 'swap' is permitted between any two sites with opposite status in a clique. Swapping two sites corresponds to the movement of individuals out of one and into the other, and following the swap the status of each site is reversed. A mismatched site, once swapped, may thenceforth be regarded as a matched site. After the maximum number of swaps have been performed, the new similarity of the two distributions may be assessed using any variant of the simple matching coefficient or Jaccard coefficient. The maximum number of matched sites obtainable by performing within-clique swaps will be referred to as the *best attainable match* (BAM). It is only beneficial to swap a pair of sites if both are mismatched, since it is stipulated that each site may be swapped at most once. Every swap therefore increases the number of matched sites by two.

A simple illustration of the technique is provided by Figure 1. The sites are arranged as grid squares, with the clique of any site being the set of its horizontal, vertical and diagonal neighbours. Presence and absence are denoted by dark and light shading. A reference distribution is shown in (a), and two model-predicted distributions are depicted in (b) and (c); these are to be matched against (a). In distribution (b) three swaps may



(a) Observed pattern.

(b) Prediction 1. SMC = 18/24, BAM = 24/24.

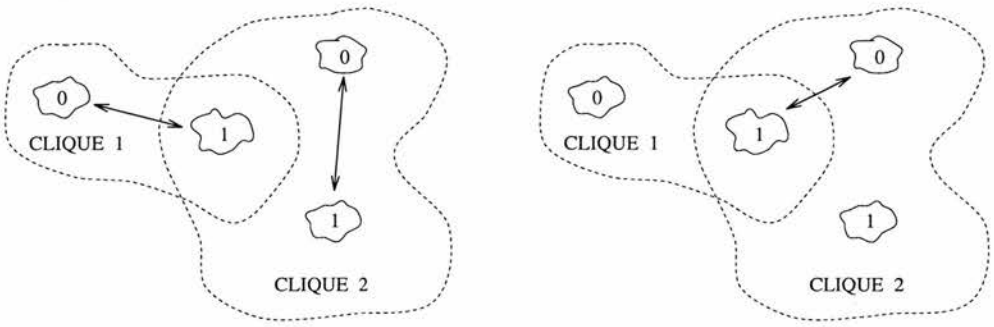
(c) Prediction 2. SMC = 18/24, BAM = 18/24.

Figure 1: The predicted distributions both yield a simple matching coefficient of 0.75, but (b) clearly captures more of the observed pattern than does (c). This discrepancy is apparent in the BAM coefficients.

be made, denoted by arrows, so as to increase the total number of matched sites by six. By contrast, the mismatches in (c) are mostly of the same status and concentrated in the top two rows so that no swaps are possible. The BAM coefficient therefore brings out the greater overall similarity of (b) to (a) where the simple matching coefficient does not.

Evaluation of the best attainable match is equivalent to determination of the maximum number of swaps that can be performed, and is complicated by the fact that optimality must be preserved by every swap. Figure 2 demonstrates the importance of the order in which sites are swapped: optimality is achieved in (a), but an injudicious swap in (b) precludes further swaps and optimal swap is not achieved. In Section 3 it is shown that the best attainable match may be computed using graph theoretic techniques, and an algorithm is developed in Section 4 to perform the calculation.

The swapping approach has the desired effect of allowing similarity to be judged primarily on the basis of global features. Provided that cliques are chosen wisely the method makes sound ecological sense, since the definition of a 'site' is often arbitrary in ecological surveys — atlas grid squares, for example. Individuals present in one site at the time of a survey may easily move subsequently to adjacent sites, so the observed distributions themselves should not be regarded as absolute. Account must of course be taken of species' mobility: it is unusual, for example, for fish to move readily between unconnected lakes. In practice the clique of any site should be chosen to be fairly small, perhaps containing only those sites immediately adjacent to it. For computational purposes large cliques will lead to some loss in efficiency.



(a) Optimal swap achieved. (b) Optimal swap not achieved.

Figure 2: Illustration of different swapping configurations. Four mismatched sites are shown, together with their status (0 or 1). Swaps are denoted by arrows, and cliques are bordered by dotted lines. Two sites may be swapped only if they occur in a common clique. Maximum swap is achieved in (a), but in (b) an injudicious choice for the first swap has precluded further swaps.

2.2 Abundance data

A *spatial abundance distribution* differs from a binary distribution in that each site has associated with it a count reflecting the estimated number of individuals in that site. The count in site i will be written as c_i, and the similarity of two distributions with counts $\{c_{1i}\}$ and $\{c_{2i}\}$ (i=1,...,N) is to be quantified.

With abundance data it is usual practice to measure the 'distance' between distributions rather than their similarity. Digby and Kempton (1987) list a number of possible coefficients, of which the most commonly used are the Euclidean distance

$$\frac{1}{N}\sum_{i=1}^{N}(c_{1i}-c_{2i})^2\,, \tag{3}$$

and the Manhattan distance from Cain and Harrison (1958),

$$\frac{1}{N}\sum_{i=1}^{N}|c_{1i}-c_{2i}|\,. \tag{4}$$

These differ mainly in the weight allotted to large values: the Euclidean distance tends to be dominated by contributions from a small number of sites with large discrepancy between the two distributions, and for this reason the Manhattan distance seems more appropriate for assessing global similarity.

An alternative measure of similarity is provided by the sample correlation coefficient

(Digby & Kempton 1987),

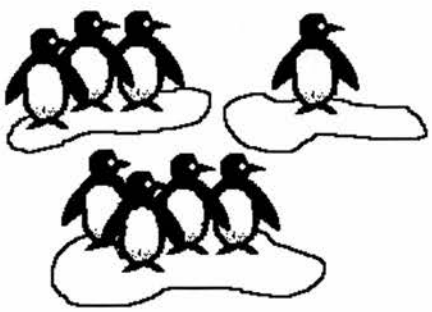
$$\frac{\sum_{i=1}^{N}(c_{1i}-\bar{c}_1)\,(c_{2i}-\bar{c}_2)}{\sqrt{\sum_{i=1}^{N}(c_{1i}-\bar{c}_1)^2\,\sum_{i=1}^{N}(c_{2i}-\bar{c}_2)^2}},$$

where $\bar{c}_t = \frac{1}{N}\sum_{i=1}^{N} c_{ti}$ is the mean of counts in distribution t. This coefficient quantifies similarity in fluctuations about the mean rather than similarity in absolute values, and has been used in a number of contexts including taxonomy and image analysis (for which 'counts' are interpreted as levels of a digitized image and 'sites' as pixels). The appropriateness of adjusting for the mean depends entirely on the application. Hiby and Lovell (1990) use the correlation coefficient to match together digitized photographs of seals for identification purposes; correcting for the mean in this context ensures that discrepancies in average brightness and contrast between photographs do not affect the outcome. In most contexts, however, care must be taken to ensure that the adjustment is valid, and certainly when comparing model predictions against observed data it is inappropriate to use this coefficient since all information on absolute abundance is lost.

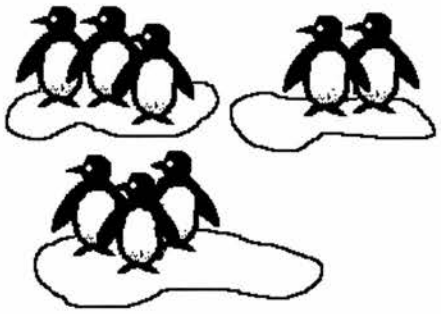
The ideas of section 2.1 may be extended to derive an improved distance coefficient for abundance data. Cliques are defined for each site as before, but swaps are now performed at the level of individuals. The individuals in a site may move into any other sites within the same clique, and all individuals move independently of each other. Accordingly, if a model predicts n animals too few for one site and m animals too many in a neighbouring site, then $\min\{m,n\}$ individuals may be swapped between the sites to even out the imbalance. The idea is illustrated in Figure 3.

For convenience the first distribution is taken as a point of reference, and all swaps are assumed to take place in the second distribution. Sites for which too many animals occur in the second distribution are referred to as *surplus* sites, while those in which too few animals occur are described as *deficit* sites. The surplus in a site can be dispersed between any number of deficit sites in its clique. The final distance coefficient between the first and second distributions is defined as the minimum value of the Manhattan distance (4) obtainable by performing within-clique swaps.

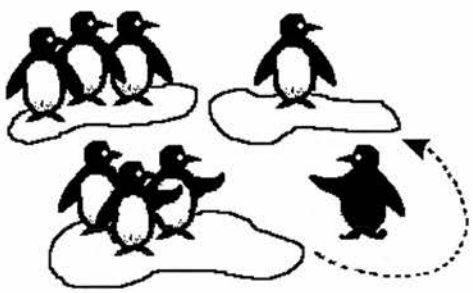
The Manhattan distance has the useful property that every swap decreases the distance by $2/N$, regardless of where the swap is performed. This greatly simplifies the computation of the minimum attainable distance since only the number of swaps need be found, not their whereabouts. The Euclidean distance does not have this property, so computation of the minimum attainable distance with this measure would require considerable extra



(a) Distribution 1.



(b) Distribution 2.



(c) Surplus individuals swapped.

Figure 3: Two abundance distributions over the same three sites are shown in (a) and (b). The three sites are all part of a single clique. An individual is moved between two of the sites in (c) to improve the match.

effort; nor is it guaranteed that the minimum distance will be attained when the number of swaps is at a maximum. If required, a mean Euclidean distance over a number of different maximum swap configurations could be obtained, but this is somewhat heavy-handed for determination of a similarity coefficient and use of the Manhattan distance is recommended.

The maximum number of swaps for an abundance distribution may be found in a similar manner to that for a binary distribution, but by nature the problem is more computationally demanding — especially if the counts involved are large. For very large counts it is recommended that values are rounded to (say) the nearest ten or hundred, so that each swap involves ten or one hundred individuals rather than just one. An algorithm for finding the maximum number of swaps for abundance distributions is developed in Section 4.

3 Evaluation of the best attainable match

3.1 Definitions

In this section it is shown how the evaluation of the best attainable match reduces to the problem of finding the cardinality of a maximum matching in a bipartite graph. This is used to provide an algorithm for the computation.

There are N sites in the survey region, labelled $1, 2, \ldots, N$. Each site i has associated with it a *clique*: a set of sites $\{j_1, \ldots, j_{n_i}\}$ such that movement of individuals can occur freely between site i and site j_k for all $k = 1, \ldots, n_i$. The clique associated with site i is written $clique(i)$.

Given two binary distributions, the first is regarded as fixed and a mismatched site is defined as a *mismatched 0* if it has status 0 in the second distribution, or a *mismatched 1* if it has status 1 in the second distribution. Two mismatched sites i and j are said to be *swappable* if $i \in clique(j)$ (or equivalently $j \in clique(i)$) and i, j have opposite status in the second distribution. There is no interest in swapping mismatched sites that both have the same status; nor in swapping a mismatched site with a matched site.

The *valency* of a mismatched site is the number of sites with which it is swappable. The *valency set*, $Val(i)$, of the mismatched site i is the set of sites swappable with i:

$$Val(i) = \{j : j \text{ is swappable with } i\}.$$

Once two swappable sites are *swapped*, the status of each is changed and the sites are thenceforth regarded as matches; they cannot be swapped further.

A *swapping configuration* of the sites is a set of swaps. The *best possible swap* (BPS) is the maximum number of swaps over all possible swapping configurations. A swapping configuration which attains the BPS is referred to as a *BPS configuration.* Such a configuration will yield the best attainable match.

For any swapping configuration,

$$match\ obtained\ =\ original\ match\ +\ 2 \times (number\ of\ swaps)\,,$$

since each swap increases the match by 2. Thus

$$number\ of\ swaps\ in\ a\ BPS\ configuration\ =\ \frac{BAM - original\ match}{2}\ . \qquad (5)$$

The cliques are not in general equivalence classes since the clique relation is non-transitive. In the event that the cliques do constitute equivalence classes, the BAM problem becomes trivial and the best attainable match is given by

$$N\ -\ \sum_{cliques\ \mathcal{C}} \left| \left(\#\ mismatched\ 0s\ in\ \mathcal{C} \right) - \left(\#\ mismatched\ 1s\ in\ \mathcal{C} \right) \right| .$$

3.2 Formulation as a bipartite graph

A graph $\mathcal{G}$ consists of a finite set $V = \{a, b, c, ...\}$ of elements called *vertices*, together with an *edge set* E of unordered pairs of distinct vertices of V, e.g. $E = \{\{a, b\}, \{a, e\}, \{b, f\}, ...\}$. The graph $\mathcal{G}$ is said to be *bipartite* if its vertex set may be partitioned into two sets X and Y such that every edge of $\mathcal{G}$ is of the form $\{x, y\}$ with $x \in X$ and $y \in Y$.

Let $X = \{x_1, x_2, \ldots, x_m\}$, and $Y = \{y_1, y_2, \ldots, y_n\}$.

The bipartite graph $\mathcal{G}$ is characterized by the matrix $B = (b_{ij})$, where

$$b_{ij} = \begin{cases} 1 & \text{if } \{x_i, y_j\} \text{ forms an edge of } \mathcal{G}\,, \\ 0 & \text{otherwise.} \end{cases}$$

The matrix B summarizes all information about $\mathcal{G}$.

A *matching* M in the bipartite graph $\mathcal{G}$ is a subset of its edges such that no two share a common vertex. Thus a matching is a 1-1 correspondence between two subsets $X' \subseteq X$ and $Y' \subseteq Y$. The *cardinality* of the matching M is the common cardinality of X' and Y'. In the matrix B, a matching of cardinality t corresponds to a set of t 1-entries, with no two on the same line: here, a *line* denotes row or column.

The BAM problem may be formulated as follows:

1. the set of mismatches forms the vertex set V of a bipartite graph $\mathcal{G}$:
   - V is partitioned as (X, Y), where $X = \{\text{mismatched 1s}\}$, $Y = \{\text{mismatched 0s}\}$;

- the edge set of $\mathcal{G}$ is $E = \{ \{x, y\} : x \in X,\ y \in Y$ and x, y are swappable$\}$;
- the valency matrix of $\mathcal{G}$ is the 0–1 matrix B such that $b_{ij} = 1$ iff x_i and y_j are swappable.

2. any set of swaps corresponds to a matching M on $\mathcal{G}$.

3. the maximum cardinality of a matching in B is equal to the BPS for the two distributions: the maximum number of swaps that can be made.

From this and equation (5) it follows that the problem of finding the best attainable match is equivalent to that of finding the maximum cardinality of a matching in a bipartite graph $\mathcal{G}$.

3.3 Matching algorithms

The evaluation of the maximum cardinality of a matching in a bipartite graph has been the focus of much attention in the graph theory and operations research literature. Matching algorithms are applied to scheduling and assignment problems (Hartsfield & Ringel 1990; Dekel & Sahni 1984), and sometimes arise in the context of transportation through a network. There are many special-case algorithms, applied for example to convex, complete, circular convex and regular bipartite graphs (Dekel & Sahni 1984; Karzanov 1987; Liang & Blum 1995; Csima & Lovasz 1992), and the matching problem also finds application in combinatorics (Jockusch 1994) and group theory (Aharoni *et al.* 1995).

A number of algorithms have been proposed for the determination of the maximum matching, predominantly variants of the Hungarian algorithm developed by Kuhn (1955) and based on the work of Hungarian mathematicians König (1936) and Egerváry (1955). An interesting account of the early development of the algorithm is given in Lenstra *et al.* (1991). The Hungarian algorithm begins by constructing an arbitrary matching on the bipartite graph and proceeds by searching for augmenting alternating paths through the graph, whereby new vertices may be added to the matching. An alternating path is one whose edges are alternately in and out of the matching, and an augmenting path is an alternating path joining two unmatched vertices (Hartsfield & Ringel 1990).

When no more augmenting paths can be found, the matching is maximal. The most efficient implementation of the Hungarian algorithm is due to Hopcroft & Karp (1973), and has complexity $O(n^{2.5})$ where n is the number of vertices in the bipartite graph.

Recently the power of parallel computing has been exploited to reduce further the time demands of matching algorithms (e.g. Karp *et al.* 1985; Kim & Chwa 1987; Shiloach & Vishkin 1982).

The Ford–Fulkerson algorithm (Ford & Fulkerson 1956) for maximizing flow in a network provides a rather different approach to the matching problem. Here the bipartite graph is regarded as a network with one set of vertices connected to a source, the other set to a sink. An initial flow (perhaps zero) is proposed, and paths through the network along which flow can be increased are determined by a node-labelling procedure. The Ford–Fulkerson algorithm is a general method for finding optimal network flow, and although there exist implementations of polynomial complexity (Goldfarb & Hao 1993) the Hungarian method is generally preferred for the matching problem.

A new algorithm is presented here, specifically designed for the ecological problem outlined in Section 2. Graphs arising from the ecological application are characterized by their sparseness, and the algorithm exploits this heavily. The algorithm generalizes naturally from the case of binary data to that of abundance data, for which the Hungarian algorithm becomes cumbersome, and it makes better use of the sparsity in the graph than the Ford–Fulkerson method. Both the Ford–Fulkerson and Hungarian algorithms construct an explicit matching on the graph: this is unnecessary within the context of the ecological problem, so the new algorithm does not expend time on it.

The algorithm is not proposed as a replacement of other implementations, which are likely to cope far more efficiently with the matching problem for a general graph. However, in Section 5 it will be demonstrated that the algorithm runs very efficiently within the context of the ecological application.

4 The best attainable match algorithm

4.1 The König Theorem

A fundamental result in Graph Theory is the König Theorem (König 1936), which states:

Theorem 1 *The maximum cardinality of a matching in the bipartite graph $\mathcal{G}$ is equal to the minimum number of vertices of $\mathcal{G}$ required to cover $\mathcal{G}$.* □

A *cover* of $\mathcal{G}$ is a set of vertices S such that every edge in $\mathcal{G}$ has at least one endpoint in S.

The proof of Theorem 1 proceeds by straightforward induction and may be found, for example, in Brualdi & Ryser (1991), chapter 1. When applied to a bipartite graph $\mathcal{G}$ with valency matrix B, the König Theorem states that the best possible swap on $\mathcal{G}$ is equal to the minimum number of lines required to cover the matrix B —that is, the least number of lines of B such that every 1-element of B lies on one of the lines.

The König theorem is used in the following sections to construct an algorithm for finding the best attainable match on $\mathcal{G}$.

4.2 Blocks

Let $\mathcal{G}$ be the bipartite graph arising from the ecological distributions, defined on page 67. For the present, attention is restricted to the case of binary data. Let $X = \{\text{mismatched 1s}\}$ and $Y = \{\text{mismatched 0s}\}$ as before, and recall that the valency set of any element is defined as the set of elements swappable with it.

Suppose r, s are integers with $s \geq r$ and $r,\ s \geq 1$. An *(r,s)–block* is defined to be a set $X' \cup Y'$ for some $X' \subseteq X$, $Y' \subseteq Y$ such that
either

> $X' = \{x_1, \ldots, x_r\}$ has r elements, $Y' = \{y_1, \ldots, y_s\}$ has $s \geq r$ elements, and $\bigcup_{i=1}^{s} Val(y_i) = X'$,

or

> $Y' = \{y_1, \ldots, y_r\}$ has r elements, $X' = \{x_1, \ldots, x_s\}$ has $s \geq r$ elements, and $\bigcup_{i=1}^{s} Val(x_i) = Y'$.

As a matter of convenience it is prescribed that every site included in a block must be valent with at least one other site in the block.

Intuitively speaking, an (r, s)–block is a set of $s \geq r$ mismatches, each of which can be swapped only within (some subset of) a smaller set of r opposite mismatches. It is not necessary that these r opposite mismatches should be swappable only within the s–set.

The (r, s)–block is said to be *swappable* if the BPS within the block is equal to r — that is, if maximum swap is achievable within the block. When a swappable (r, s)–block is *fully*

swapped, r of the sites in the s–set are swapped with the r sites in the r–set, the swap of $\mathcal{G}$ is increased by r, and no site in the block is eligible to be swapped further. If fewer than r of the sites in the s–set are swapped, the block is said to be *partially swapped.*

Suppose $r' \leq r$, $s' \leq s$. An (r', s')–*subblock* of the (r, s)–block $X' \cup Y'$ is a set $X'' \cup Y''$, with $X'' \subseteq X'$, $Y'' \subseteq Y'$ such that $X'' \cup Y''$ constitutes an (r', s')–block of $\mathcal{G}$. Such a subblock is said to be *proper* if $r' < r$.

An (r, s)–block is referred to as *minimal* if it contains no proper subblocks.

4.3 Clusters

Let $X' \subseteq X$, $Y' \subseteq Y$. The set $X' \cup Y'$ is described as a *cluster* if

(i) $Val(y) \subseteq X' \quad \forall y \in Y'$;

(ii) $Val(x) \subseteq Y' \quad \forall x \in X'$;

(iii) if $x \in X'$ and $y \in Y'$ then $\exists$ a sequence $x = x_0,\ y_0,\ x_1,\ y_1, \ldots, x_n,\ y_n = y$ in $X' \cup Y'$ such that all possible pairs (x_i, y_i) and (y_i, x_{i+1}) are swappable. Such a sequence will be referred to as a *path.*

A cluster $\mathcal{C}$ is therefore a set of sites such that any two sites with opposite status in $\mathcal{C}$ can be joined by a path in $\mathcal{C}$ through swappable sites only, and if site i is a member of $\mathcal{C}$ then all sites swappable with i are also members of $\mathcal{C}$. In graph-theoretic terminology, each cluster corresponds to a *connected component* of the graph $\mathcal{G}$.

The cluster $\mathcal{C} = X' \cup Y'$ is referred to as *perfect* if all sites in $\mathcal{C}$ have equal valency. It is immediate from this definition that $|X'| = |Y'|$, since if all sites in $\mathcal{C}$ have valency k then the number of edges in $\mathcal{C}$ is given by both $k|X'|$ and $k|Y'|$.

A cluster $X' \cup Y'$ is described as *swappable* if the BPS within the cluster is equal to $\min\{|X'|, |Y'|\}$. When the cluster is *swapped*, the BPS of $\mathcal{G}$ is increased by $\min\{|X'|, |Y'|\}$ and no sites in the cluster can be swapped further.

The clusters of the graph $\mathcal{G}$ form a partition of $\mathcal{G}$ into a number of sub-graphs. Swapping a site in one cluster can have no effect on possible swaps within a different cluster, since there is no path between any element of the first cluster and an element of the second. The clusters may therefore be treated independently of one another.

4.4 Swappability

Theorem 1 is used to prove the following propositions.

Proposition 1 *Every minimal (r, s)–block is swappable.*

Proof Let $\mathcal{B} = X' \cup Y'$ be a minimal (r, s)–block. Without loss of generality, assume that $|Y'| = r$, $|X'| = s$.

$\mathcal{B}$ has associated with it the $s \times r$ matrix $B' = (b'_{ij})$, where

$$b'_{ij} = \begin{cases} 1 & \text{if } x'_i \text{ and } y'_j \text{ are swappable,} \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that $\mathcal{B}$ is not swappable. From Theorem 1 it follows that there exists a cover of B' in $q < r$ lines.

Now B' has at least one 1-entry in every row and every column, since every element in a block is valent with at least one other element in the block. Thus the q-cover must involve a mixture of rows and columns, since if it involved either alone then $\geq r$ lines would be required.

Suppose the q-cover involves l columns and m rows, where $l + m = q$ and $l,\ m > 0$. Then there are $s - m$ rows not in the cover, whose 1-entries must be covered by the l columns (Figure 4). Further, $l + m < r \leq s$ (so $l < s - m$), and hence this constitutes a proper $(l, s - m)$–subblock of $\mathcal{B}$, thereby contradicting the minimality of $\mathcal{B}$.

If $\mathcal{B}$ has no proper subblocks, it must follow that $\mathcal{B}$ is swappable. □

It is always optimal to swap a swappable (r, s)–block, since the sites in the s-set cannot be swapped elsewhere, and those in the r-set cannot contribute more than r swaps to the total no matter where the swap takes place. The proposition therefore demonstrates that it is always optimal to swap a minimal (r, s)–block.

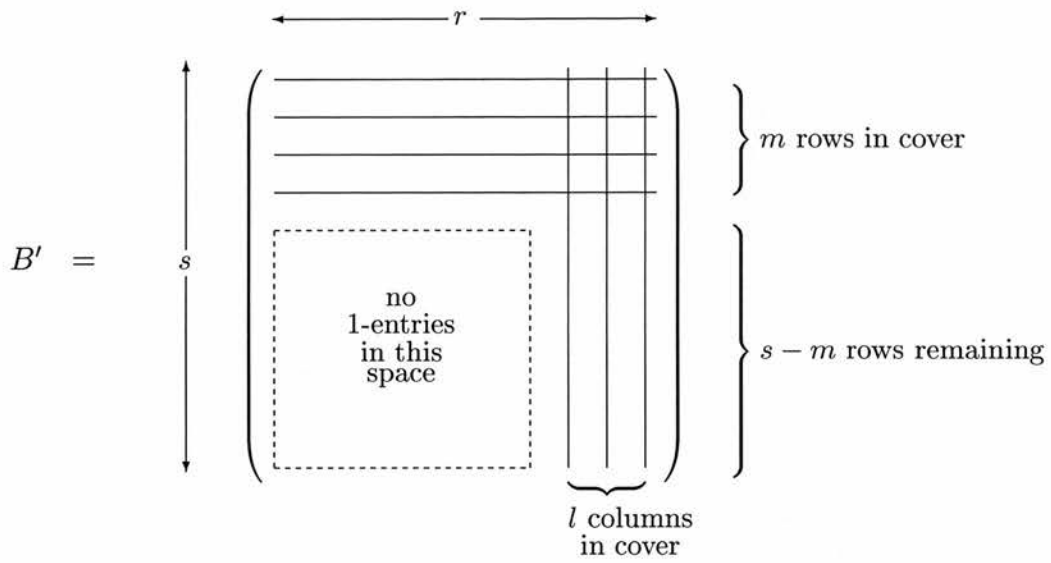


Figure 4: Diagram of the matrix B' of Proposition 1 with rows and columns permuted as necessary. Each horizontal line corresponds to an element of X' and each vertical line to an element of Y', and together these elements cover all 1-entries of the matrix B'. If there are $l + m = q < r$ elements in the cover then a proper $(l, s - m)$–subblock must exist in the bottom right-hand corner of the matrix.

Proposition 2 *Every perfect cluster is swappable.*

Proof Let $\mathcal{C} = X' \cup Y'$ be a perfect cluster, such that every element in $\mathcal{C}$ has valency k. Let the associated matrix be $B' = (b'_{ij})$, where

$$b'_{ij} = \begin{cases} 1 & \text{if } x'_i \text{ and } y'_j \text{ are swappable}, \\ 0 & \text{otherwise.} \end{cases}$$

It was shown in section 4.3 that $|X'| = |Y'|$, so B' is an $r \times r$ matrix with precisely k entries in each column and each row. Thus any line of B' covers at most k new 1-entries of B'. Since B' has $r \times k$ 1-entries in total, at least r lines are required to cover all 1-entries of B' (using the cover of either r rows or r columns). By Theorem 1 it follows that the BPS of $\mathcal{C}$ is r and the perfect cluster $\mathcal{C}$ is swappable. □

Clusters are self-contained and independent of each other, so it is clearly optimal to swap any cluster known to be swappable. The results of this section therefore indicate that it is always optimal to swap minimal blocks and perfect clusters.

4.5 The Algorithm

The BAM algorithm for binary data is now outlined.

The graph $\mathcal{G}$, matrix B and partition (X, Y) of $\mathcal{G}$ are defined as in section 3.2. The algorithm searches $\mathcal{G}$ for perfect clusters and minimal blocks, swapping these as found. Since $\mathcal{G}$ itself constitutes either a $(|X|, |Y|)$–block or a $(|Y|, |X|)$–block, all sites in $\mathcal{G}$ must occur in a block of some size.

Blocks are found by a recursive procedure described below. The search is conducted in increasing order, so that all $(2, \cdot)$–blocks are eliminated before searching for $(3, \cdot)$–blocks, and so on. In this way all blocks detected are certain to be minimal. The simplest type of (r, s)–block occurs when $r = 1$: such blocks are referred to as *singletons* and consist of one or more sites all reliant on a single opposite site for their swap. All singletons are minimal and therefore swappable, and provide a convenient starting point for the algorithm.

Once the singletons have been eliminated the algorithm searches for $(2, s)$–blocks, and so on. In many cases the elimination of an (r, s)–block leaves smaller blocks elsewhere in the graph that must be dealt with before any larger blocks are sought.

The perfection or otherwise of a cluster is quickly determined by checking whether or not all sites in the cluster have equal valency: this procedure demands little computational effort. In practice the graph is divided into its clusters (connected components) before scanning for blocks of size ≥ 2, in order to reduce the computational effort in block-hunting. At this stage perfect clusters are identified and swapped.

The algorithm is summarized in Figure 5. In addition various shortcuts are incorporated, notably checks to determine whether or not a swapped block has left behind it any smaller blocks. Once a site is swapped the valencies of all other sites swappable with it must be decreased accordingly. The algorithm terminates as soon as all unswapped sites have valency 0 — i.e. no more swaps are possible. Termination must occur since all sites with valency > 0 are contained in a block of some size.

The recursive procedure for finding blocks is as follows. Suppose without loss of generality that the algorithm is searching for (r, s)–blocks where the r-set is a subset of Y. An initial site in X with valency $v \leq r$ is selected for inclusion in the block, and this site together with its valent elements are temporarily deleted from the graph. It remains to find $s - 1$ further elements of X whose valency sets lie within an $(r - v)$-subset of Y, which problem

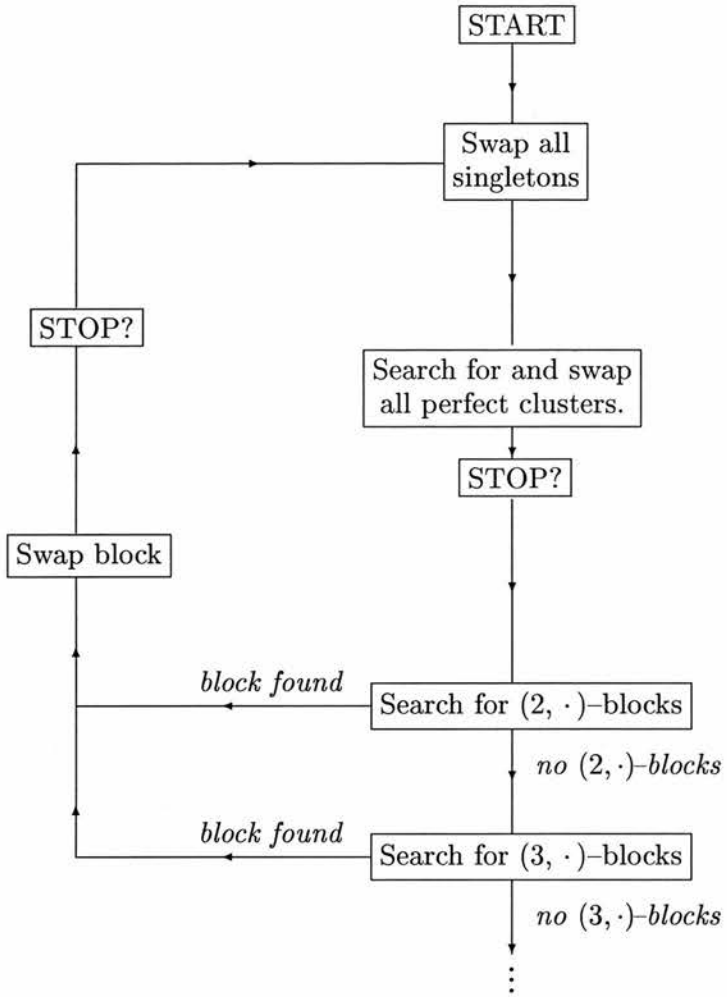


Figure 5: The BAM algorithm for binary data.

is precisely that of finding an $(r - v, s - 1)$–block. The initial element is chosen so as to minimize $r - v$ and thus the recursion is seldom very deep, terminating if $r - v = 1$. In Section 5 it is demonstrated that for the application in hand almost all swaps occur as singletons, perfect clusters and $(2, \cdot)$–blocks, and these the algorithm deals with very quickly.

The BAM algorithm has been made freely available on the world-wide web as a `.tar` file, at http://www-ruwpa.mcs.st-and.ac.uk/~rachel/software/index.html. The program is written in Sun PASCAL, as a widely-available language with elegant set-handling facilities. The programs have been tested on a UNIX platform with operating systems SunOS 4 and 5 (Solaris 1 and Solaris 2) although not on other systems.

4.6 Abundance data

Given two distributions with counts $\{c_{1i}\}_{i=1}^{N}$ and $\{c_{2i}\}_{i=1}^{N}$ as in section 2.2, the difference $d_i = c_{2i} - c_{1i}$ is computed for each site i. If $d_i > 0$ then site i is said to contain $|d_i|$ *surplus* individuals, while if $d_i < 0$ site i is said to contain $|d_i|$ *deficit* individuals. Sites with $d_i = 0$ are matched exactly and need not be considered further. Surplus and deficit individuals then form the vertices of a bipartite graph $\mathcal{G}$ with partition $X = \{\text{surplus individuals}\}$, $Y = \{\text{deficit individuals}\}$, and the edges of $\mathcal{G}$ extend between surplus and deficit individuals that reside in sites in the same clique.

In order to calculate the minimum attainable value of the Manhattan distance (4), the maximum number of surplus individuals that may be swapped with deficit individuals must be found; each swap decreases the Manhattan distance by $2/N$. Once again the problem reduces to evaluation of the maximum cardinality of a matching in the bipartite graph $\mathcal{G}$. However, in the case of abundance data, the calculation is made unwieldy by the fact that two sites in a clique containing respectively m surplus and n deficit individuals give rise to $m \times n$ identical edges in the graph: any one of the m surplus individuals could be swapped with any one of the n deficit individuals. This would cause a vast increase in the time and memory required for application of the normal matching algorithms, including the algorithm developed in section 4.5.

Instead, a formulation is preferred that acknowledges the duplication of edges in the graph $\mathcal{G}$. Indeed, the problem is perhaps more akin to the flow-max paradigm mentioned in section 3.3, traditionally tackled using the Ford–Fulkerson algorithm. A site containing m surplus individuals would correspond in that formulation to a node connected to a source by an arc of capacity m, and a site containing n deficit individuals would correspond to a node connected to a sink by an arc of capacity n. The most efficient implementations of the Ford–Fulkerson algorithm proceed by establishing an initial flow through the resulting network and determining paths along which the flow can be augmented by a maximal amount. The original flow is then replaced by the augmented flow and the process is repeated until no more augmenting paths exist. In terms of the ecological problem here, arbitrary choices would be made as to which individuals should be swapped, and if it later transpired that those choices precluded optimal swap then considerable back-tracking might be necessary.

The notions of blocks, clusters and singletons outlined in sections 4.2, 4.3 and 4.5 may be extended easily to the case of abundance data, and provide an alternative to the Ford–

Fulkerson algorithm. First some new terminology is defined.

Let Z be a set of surplus or deficit individuals: that is, $Z \subseteq X$ or $Z \subseteq Y$. The set of sites that contain individuals in Z is written as $\mathcal{S}(Z)$. For a surplus or deficit individual x, the set of individuals swappable with x is given by $Val(x)$ as before. In addition, the notion of *site valency* is introduced: the site i is *site-valent* with the site j if the individuals in i are swappable with the individuals in j.

The set $Siteval(i)$ is defined as the set of sites that are site-valent with site i. The *site-valency* of site i is the number of elements in $Siteval(i)$. If y is an individual in site i, the notions of individual valency and site-valency are linked by the expression $Siteval(i) = \mathcal{S}(Val(y))$.

Recall from section 4.2 that sets $X' \subseteq X$ and $Y' \subseteq Y$ form a $(|X'|, |Y'|)$–block iff $X' = \bigcup_{y \in Y'} Val(y)$ and $|Y'| \geq |X'|$. For convenience it is assumed that Y' contains all individuals whose valency sets lie inside X'. Note that all relationships are simply reversed for a $(|Y'|, |X'|)$–block. Then

$$|Y'| \geq |X'| \iff \sum_{j \in \mathcal{S}(Y')} |d_j| \geq \sum_{i \in \mathcal{S}(X')} |d_i|, \tag{6}$$

and

$$X' = \bigcup_{y \in Y'} Val(y) \iff \mathcal{S}(X') = \bigcup_{j \in \mathcal{S}(Y')} Siteval\,(j). \tag{7}$$

Using (6) and (7) the problem may be transformed from one that involves individuals within sites to one involving sites alone. A new type of block is now defined for graphs arising from abundance distributions.

Let t, q be integers with $t > 0$. A *(t, q)–siteblock* is a set $\mathcal{S}(X') \cup \mathcal{S}(Y')$ for some $X' \subseteq X$, $Y' \subseteq Y$ such that
either

$\mathcal{S}(Y')$ has t elements, $\mathcal{S}(X') = \bigcup_{j \in \mathcal{S}(Y')} Siteval\,(j)$,
and $\sum_{j \in \mathcal{S}(Y')} |d_j| - \sum_{i \in \mathcal{S}(X')} |d_i| \geq q$,

or

$\mathcal{S}(X')$ has t elements, $\mathcal{S}(Y') = \bigcup_{i \in \mathcal{S}(X')} Siteval\,(i)$,
and $\sum_{i \in \mathcal{S}(X')} |d_i| - \sum_{j \in \mathcal{S}(Y')} |d_j| \geq q$.

The siteblock definition is motivated by the same idea as the block definition of section 4.2: namely, a set of t sites all competing amongst the same set of opposite sites for their swap. Because of the differing numbers of individuals in each site, however, it is not sufficient to consider only the number of sites in each set as for the (r, s)–blocks of section 4.2. This is why the siteblocks are defined in terms of the weight difference q.

The link between siteblocks and blocks is nonetheless apparent. It is immediate from (6) and (7) that every (r, s)–block in the abundance graph is a $(t, 0)$–siteblock for some t, and vice versa. Furthermore, if a $(t, 0)$–siteblock contains no $(w, 0)$–siteblocks for $w < t$ then the associated (r, s)–block is minimal and full swap is possible within the siteblock. Thus instead of finding blocks of individuals, which are likely to be of unmanageable size, attention is restricted to finding $(t, 0)$–siteblocks, where t is typically small.

The (t, q)–siteblocks are found by a recursive procedure similar to that described for blocks in section 4.5. An initial site i_1 in $\mathcal{S}(X)$ (say) is selected that maximizes $q_1 = |d_{i_1}| - \sum_{j \in \mathrm{A}_1} |d_j|$, where $\mathrm{A}_1 = \mathit{Siteval}\,(i_1)$. If $q_1 \geq 0$ then $\{i_1\} \cup A_1$ constitutes a $(1, 0)$–siteblock which can be swapped. If $t = 1$ the procedure is then complete. Suppose now that $t > 1$: all $(1, 0)$–siteblocks have been eliminated previously, so $q_1 < 0$. The procedure continues by temporarily deleting $\{i_1\} \cup A_1$ from the graph, whereupon it remains to find a set $\{i_2, \ldots, i_t\}$ of $t-1$ further elements of $\mathcal{S}(X)$ such that $\sum_{k=2}^{t} |d_{i_k}| - \sum_{j \in \mathrm{A}_2} |d_j| + q_1 \geq 0$, where $\mathrm{A}_2 = \bigcup_{k=2}^{t} \mathit{Siteval}\,(i_k) \backslash \mathrm{A}_1$. This is precisely the problem of finding a $(t-1,\ -q_1)$–siteblock in the depleted graph, whence the recursion proceeds.

Perfect clusters are handled exactly as for binary data, and arise when all individuals in a cluster have equal valency. There is also a further structure, namely *site-singletons*: these are sites which have a site-valency of 1, and they do not always occur as siteblocks. Site-singletons may be swapped immediately since no decision need be made about which sites they are best swapped with, and the total swap on $\mathcal{G}$ can be no greater if they are unswapped than its value when they are swapped.

The calculation of the BPS for abundance data is computationally more demanding than that for binary data. In most applications, exact matches in counts are rare and consequently cluster sizes tend to be large. Nonetheless, observation suggests that minimal $(t, 0)$–siteblocks seldom occur for $t > 4$, and the majority of swaps are derived from site-singletons and $(1, 0)$–siteblocks. The outline of the algorithm is given in Figure 6, and a PASCAL implementation is available at the WWW address given on page 75.

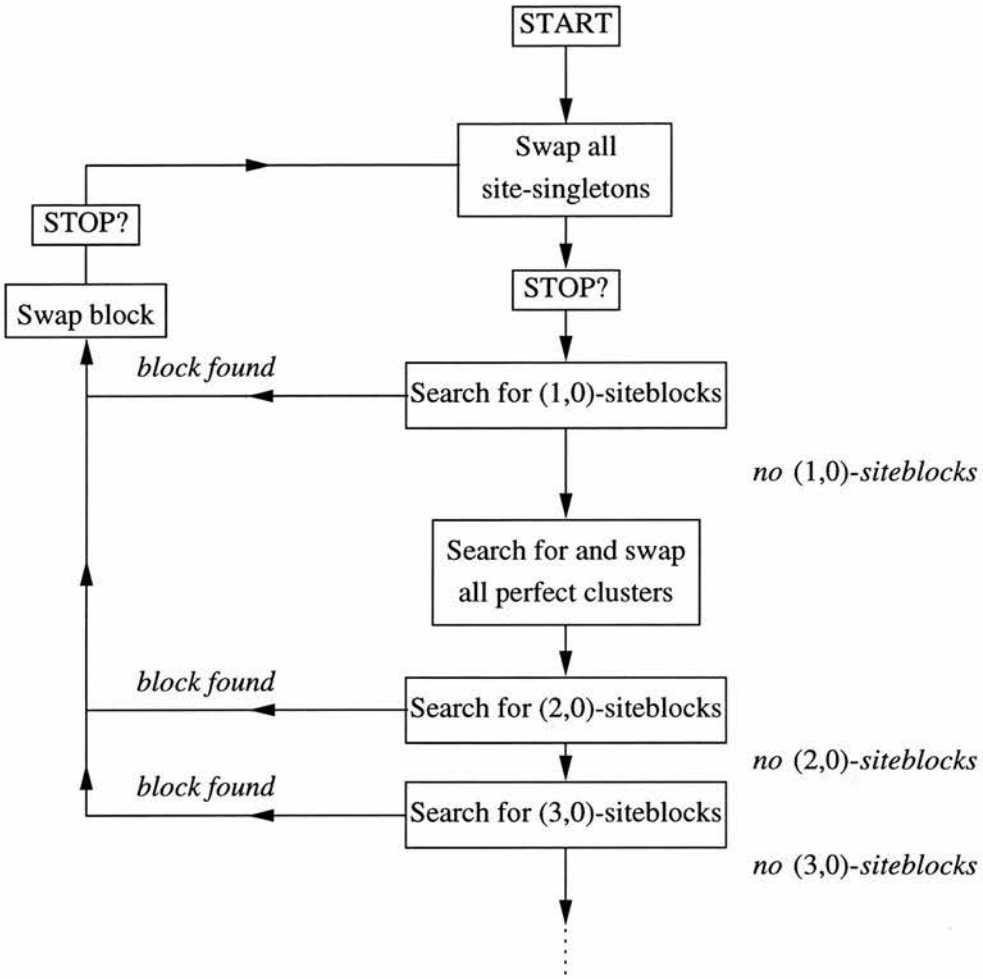


Figure 6: The BAM algorithm for abundance data.

Once the BPS is determined for two abundance distributions, the new value of the Manhattan distance can be calculated as

$$\frac{1}{N}\left\{\sum_{i=1}^{N}|c_{1i}-c_{2i}|-2\times\mathrm{BPS}\right\},$$

since every swapped individual reduces the sum by 2.

5 Simulation studies

5.1 Binary data

The algorithm proposed in Section 4 is intended for relatively sparse graphs that arise from the ecological application. Theoretical analysis of algorithmic complexity is not worthwhile, since performance is only of interest within the scope of the ecological application: worst-case results for a general graph are of no importance. Instead, simulated spatial distributions are used to demonstrate the practical performance of the algorithm.

It is anticipated that the clique of any site will be chosen so as to contain only a small number of sites — perhaps only those immediately adjacent to it — thereby restricting the number of edges stemming from the associated vertex in the graph. The algorithm will be most efficient if the graph $\mathcal{G}$ contains a sizeable proportion of singletons (vertices with only one incident edge) and small blocks. The results of this section demonstrate that this is the case in practice.

For illustrative purposes simulations are performed on an $n \times n$ grid of squares with the clique of any site defined as the set of its eight nearest neighbours (horizontal, diagonal and vertical), or as many of these as exist. A pair of spatial distributions is generated by allotting a probability of presence p to each square in the first grid and q to each square in the second, and drawing stochastic realizations from these probabilities. It is readily shown that the probability that two sites in the same clique are both mismatched and have opposite status is maximized when $p = q = 0.5$. This corresponds to maximizing the expected number of edges in the graph, and therefore provides the most demanding test of the algorithm. Nonetheless the graph remains quite sparse: a mismatched site with eight neighbours is a singleton with probability 0.267 and has an expected number of swappable sites equal to 2.

The primary interest in this section is the behaviour of the algorithm as the number of sites $N = n^2$ grows large. Results are presented from the 'worst case' $p = q = 0.5$, for which the expected number of mismatched sites is $N/2$. This is also the expected number of vertices in the graph. The algorithm is implemented in Sun PASCAL on a Sun Ultra Enterprise 150 server with 256MB RAM. Figure 7 shows the increase in running time with N for selected values of N up to 14400. The algorithm is untested beyond here, and some systems will be unable to cope with the memory demands for distributions as large as these. Ecological surveys rarely contain data from more than a few hundred sites, so the

upper limit of 14400 is considered more than adequate.

The figure indicates that even for very large distributions the algorithm is executed in a matter of seconds on the specified system. In addition, the least squares regression line from the model $\log(time) = a + b\log(N)$ is plotted on Figure 7. The model coefficients are $a = -17.1$, $b = 2.06$, and the fitted curve clearly captures the shape of the relationship. It follows that at least in this range the running time of the algorithm is approximately $O(N^2)$. The most efficient implementation of the Hungarian algorithm for a general graph is of complexity $O(N^{2.5})$ (Hopcroft & Karp 1973), although the performance of that algorithm on this application has not been tested. For all methods, extra time is needed to construct the graph before applying the matching algorithm.

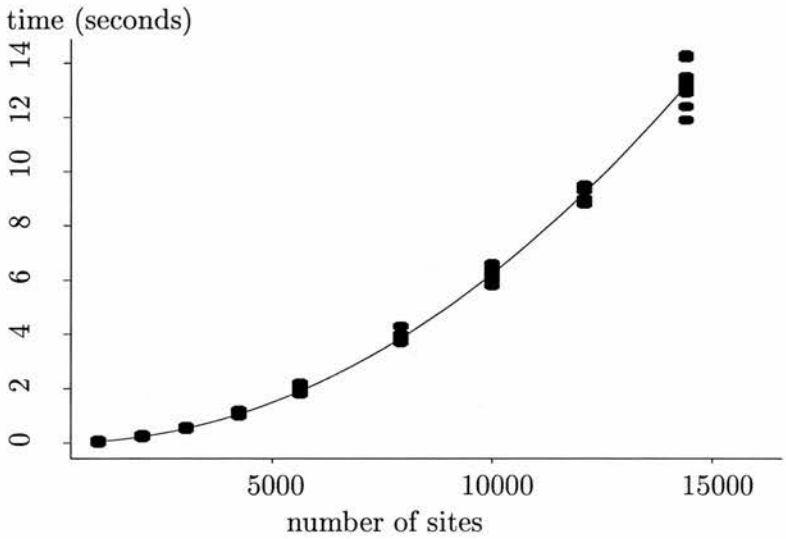


Figure 7: Running times in seconds for the algorithm when $p = q = 0.5$ and $N = 1024$, 2025, 3025, 4225, 5625, 7921, 10000, 12100, 14400. The points show the results of 10 simulations at each value of N and the line represents the least squares fit from the linear model $\log(time) = a + b\log(N)$. The coefficients are $a = -17.1$, $b = 2.06$.

To some extent the running time will depend on the number of $(2, \cdot)$–blocks and above found in the graph, since each such block involves at least one call of the recursive procedure described in section 4.5. In practice there is often no appreciable difference in the time taken to find blocks of different sizes, and blocks are seldom large: the largest block size observed out of many simulations for $N = 10000$, $p = q = 0.5$ was 13, while the mean block size was 2.52. The division of the graph into clusters prior to searching for blocks ensures that the algorithm is handling only a small number of vertices at once: over 20 simulations with $N = 10000$ involving 862 clusters the median cluster size was found to be 8. Figure 8 shows the number of blocks found for a range of N: the relationship is roughly linear.

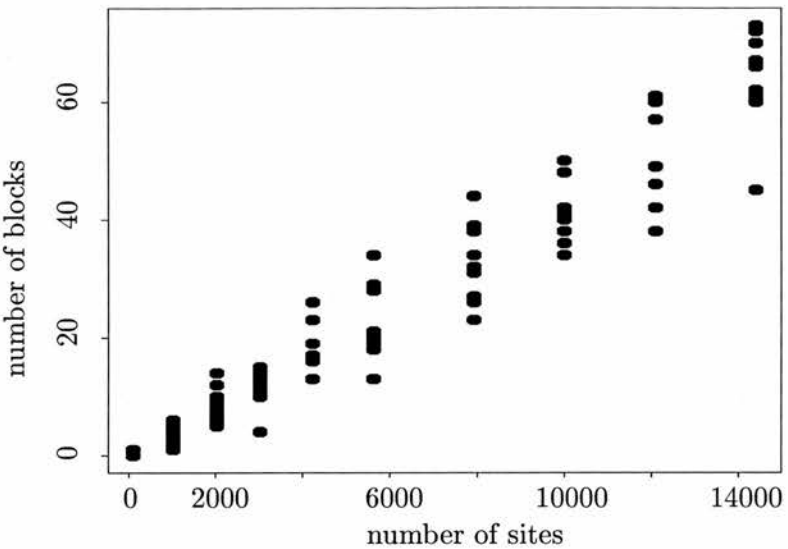


Figure 8: Number of blocks found in each of 10 simulations for $N = 100$, 1024, 2025, 3025, 4225, 5625, 7921, 10000, 12100, 14400.

By far the greatest amount of effort in the BAM algorithm, however, is devoted to singletons. As singletons are swapped, more are created and it can be seen from Figure 9 that singletons comprise about 90% of the total BPS. Informal trials suggest that swapping the initial set of singletons takes up approximately half of the running time, with the remainder being divided between perfect clusters, blocks and subsequent singletons. The sparsity in the graph is therefore heavily exploited.

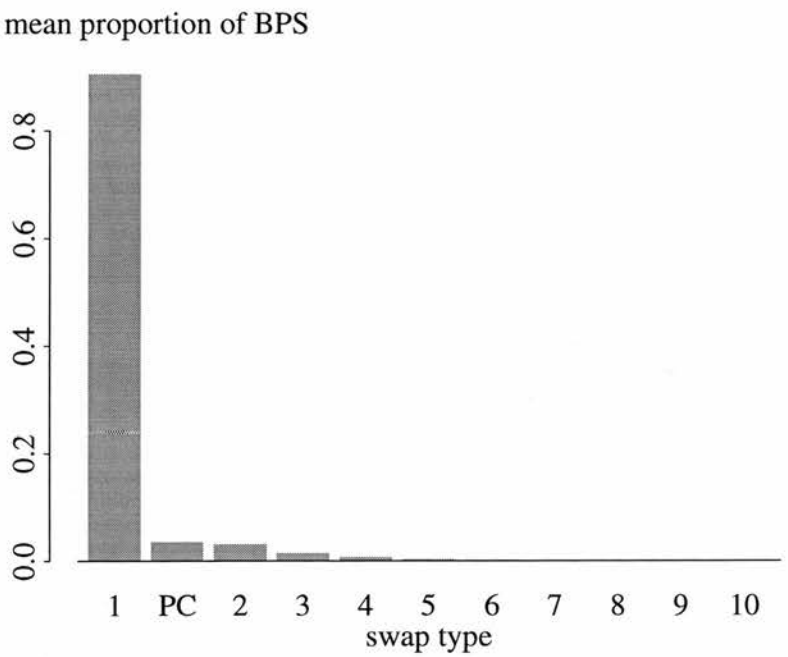


Figure 9: Mean proportion of BPS attributable to singletons (1), perfect clusters (PC), $(2, \cdot)$–blocks (2), ..., and $(10, \cdot)$–blocks (10) from 10 simulations with $N = 10000$.

5.2 Abundance data

Performance results for abundance data are more difficult to obtain than those for binary data, because running time is heavily dependent on the abundance distributions to be matched. Site for site, the abundance algorithm is much slower than the binary algorithm, and it is less predictable. Two pairs of distributions of the same size and generated by the same process (e.g. Poisson counts with constant mean) may differ considerably in the running time required. Large discrepancies in counts between the two distributions cause loss in efficiency, so it may be necessary to bin counts into classes of 10, 100 and so forth before applying the algorithm. Informal trials indicate that distributions with a few hundred sites are handled fairly efficiently, but distributions of over a thousand sites are likely to cause frequent problems with deep recursion. Algorithms such as the Ford-Fulkerson (Ford & Fulkerson 1956) which do not rely on recursion might perform better in these cases, although a direct comparison has not been attempted.

Results presented in this section are from abundance distributions on a square grid of $N = 400$ sites with nearest-neighbour cliques as before. Counts were generated from a discrete uniform distribution on $[0, 10]$, and 60 trials were conducted on the system specified in section 5.1. The running time varied considerably between trials: for 47 of the 60 trials less than one second was required, for a further 10 trials between 1 and 6 seconds were required, and three other trials took respectively 30, 100 and 333 seconds. The pattern was similarly unpredictable for other grid sizes and count distributions.

Figure 10 shows the mean occurrence of perfect clusters, site-singletons and siteblocks over the 60 trials. Perfect clusters are unusual for abundance distributions, and only 16 instances were noted in the 60 trials. Site-singletons were the most abundant of all entities, at 5689 instances, and (1, 0)–siteblocks were also abundant at 2464 instances. Site-singletons and (1, 0)–siteblocks require very little computer time. There was one instance of a (14, 0)–siteblock and one (11, 0)–siteblock, but all other blocks were of size less than 10. The amount of time spent by the algorithm in detecting the largest blocks can be disproportionate to the swap obtained, and some workaround is probably necessary before the abundance algorithm can be used reliably.

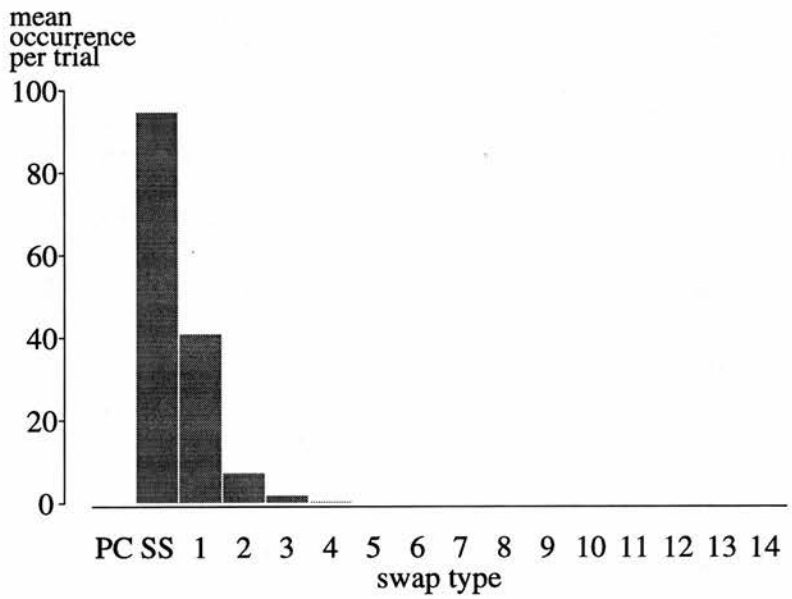


Figure 10: Mean occurrence per trial of perfect clusters (PC), site-singletons (SS), $(1,0)$–siteblocks (1), $(2,0)$–siteblocks (2), ..., and $(14,0)$–siteblocks (14) from 60 simulations with $N = 400$ and counts $\sim$ Uniform$[0,\ 10]$.

6 Examples

6.1 Abundance algorithm: simulated data

For an illustration of the abundance algorithm, counts were simulated on a 20×20 grid of squares, with the clique of any site defined as the set of its eight nearest neighbours (horizontal, vertical and diagonal) or as many of these as exist. The count in square (x, y) was generated according to a Poisson distribution with rate $\exp(0.2 + 0.05\,x)$ for $x, y = 1, 2, \ldots, 20$; the data therefore adhered to a generalized linear model (GLM) formulation with Poisson error, logarithmic link function and linear predictor $0.2 + 0.05\,x$. These data were thenceforth regarded as 'observed', and exhibited a clear trend for count increasing with x (Figure 11 (a)).

Two models were fitted to the observed data:

Model A

GLM with Poisson error, log link and linear predictor $\alpha + \beta x$
(correct model);

Model B

GLM with Poisson error, log link and linear predictor $\gamma + \delta y$
(incorrect model).

The coefficients from the fitted models were

Model A

$\hat{\alpha} = 0.19, \qquad \hat{\beta} = 0.053$
(close to the true values $\alpha = 0.2,\ \beta = 0.05$);

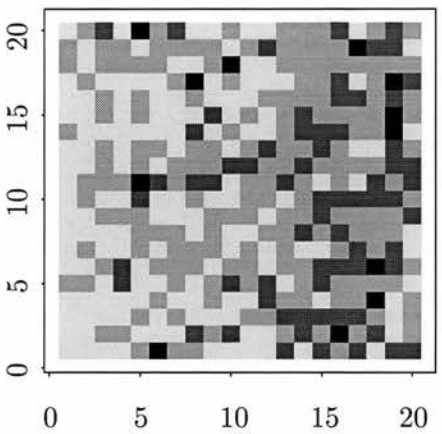
Model B

$\hat{\gamma} = 0.74, \qquad \hat{\delta} = 0.0049$
(meaningless in the context of the true model),

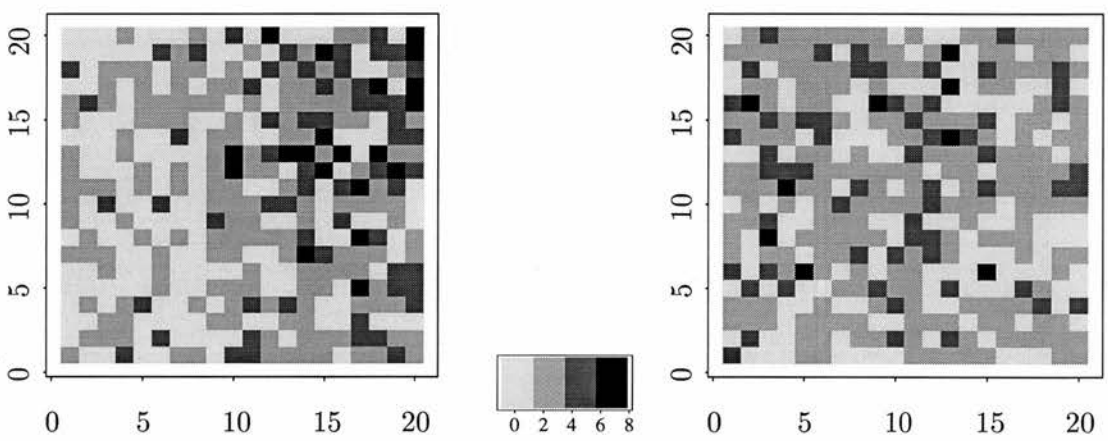
and the models yielded residual deviances of 409.1 and 488.1 respectively.

Model selection for a simple example such as this would generally be based on deviance values, and indeed the deviance from the correct model is appreciably lower than that

from the incorrect model. However, ecological models are frequently not amenable to such straightforward analysis and therefore it is instructive to test the ability of similarity coefficients to distinguish between predictions from good and bad models.



(a) Observed data.



(b) Prediction from Model A. Distance before swaps = 1.61, after swaps = 0.39.

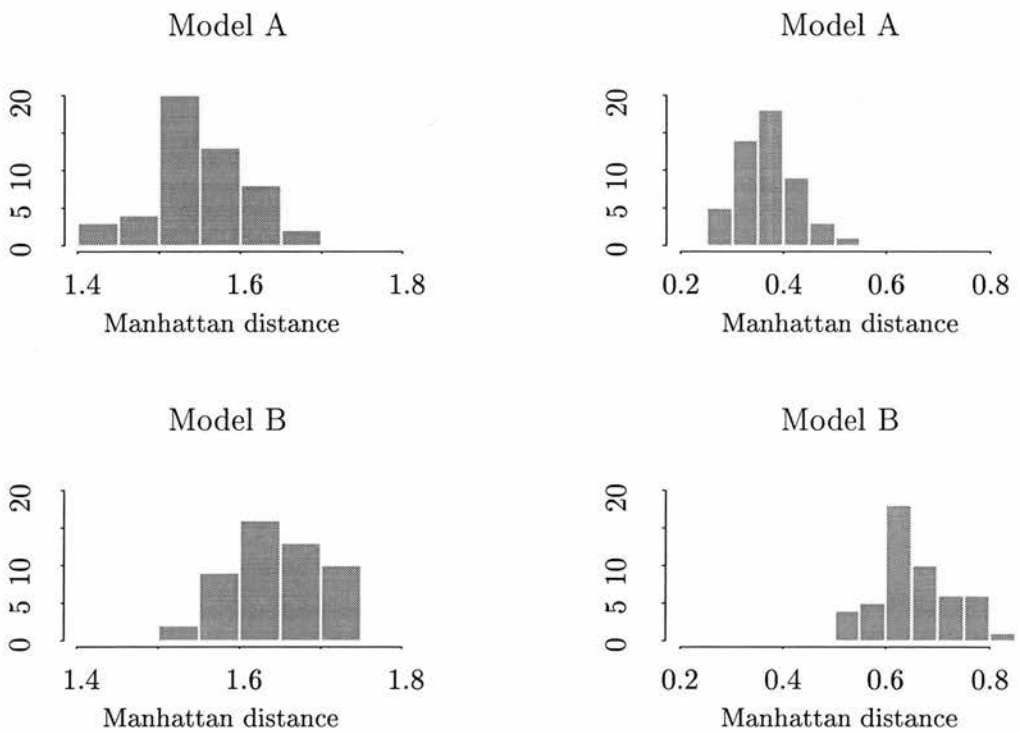
(c) Prediction from Model B. Distance before swaps = 1.60, after swaps = 0.75.

Figure 11: Comparison between observed data and predictions from Models A and B. Manhattan distances between the distribution in (a) and those in (b) and (c) are given before and after swaps are performed.

Figure 11 shows the observed data together with typical predictions from Models A and B. The prediction from Model A follows the same trend as the observed data, while that from Model B displays no clear trend at all. The Manhattan distances between the observed data and the two predictions are however almost identical, and indeed that for Model B is slightly lower in this instance. The distances after swaps have been performed, on the other hand, discriminate much better between good and bad prediction, with the distance

for Model B almost double that for Model A.

Histograms of Manhattan distances between the observed data and 50 predictions from each of Models A and B are given in Figure 12 (a), from which it is evident that no clear conclusions can be drawn as to which is the better model. By contrast, the histograms in Figure 12 (b) of the Manhattan distances from these same predictions after swaps have been performed indicate unequivocally that Model A yields the better results: the maximum post-swap distance obtained from Model A after 50 trials was less than the minimum post-swap distance obtained from Model B. The swapping approach therefore assists with identification of the better model, while the original distance measure is of little help.



(a) Before swaps.

(b) After swaps.

Figure 12: Histograms of Manhattan distances between observed data and predictions from Model A (top), and between observed data and predictions from Model B (bottom), before and after swaps.

6.2 Binary algorithm: distribution of red deer

Augustin *et al.* (1996) propose four modelling strategies for prediction of the spatial binary distribution of red deer *Cervus elaphus* in the Grampian Region of Scotland. The data, illustrated in Figure 13, consist of recorded presence or absence of red deer for each of 1277 1 km grid squares in the region. Each model is fitted to a 20% sample of the data, and produces as output an estimated probability of occupation for every site. Stochastic realizations of presence-absence are generated from these probabilities, and the models are compared on their ability to predict the observed pattern correctly. Comparisons are conducted using the simple matching coefficient.

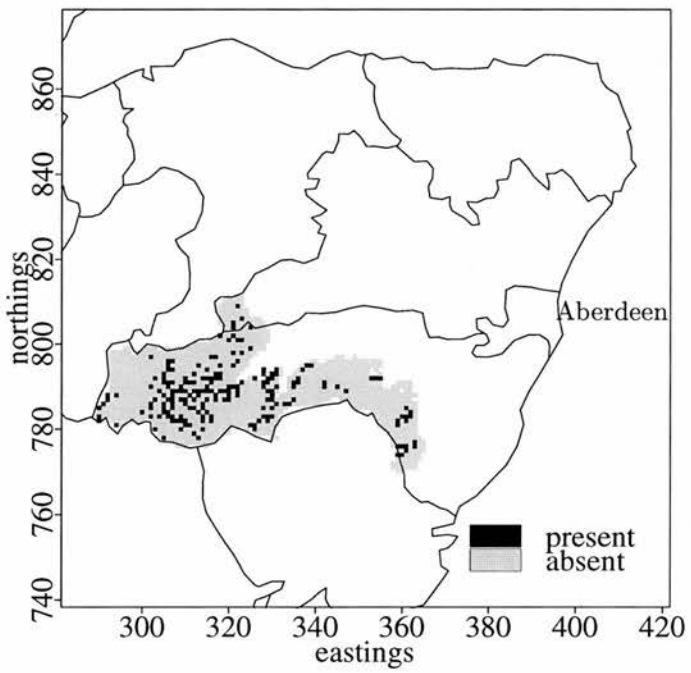
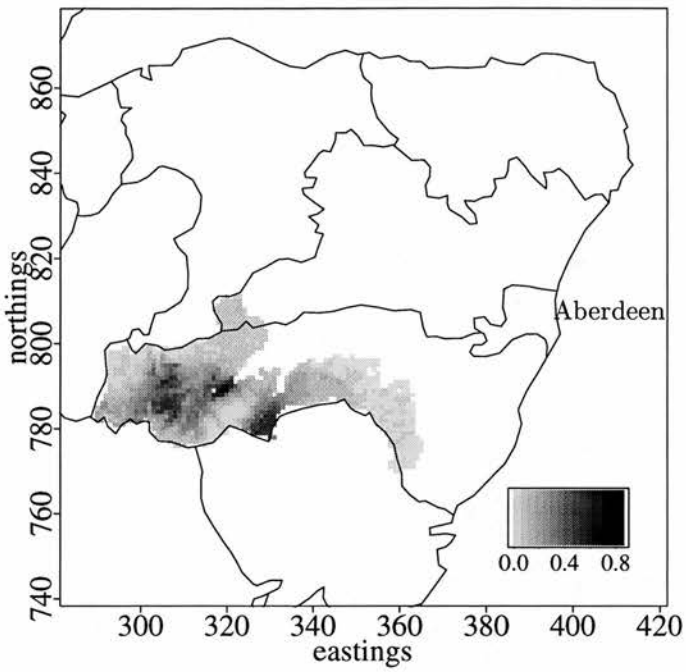


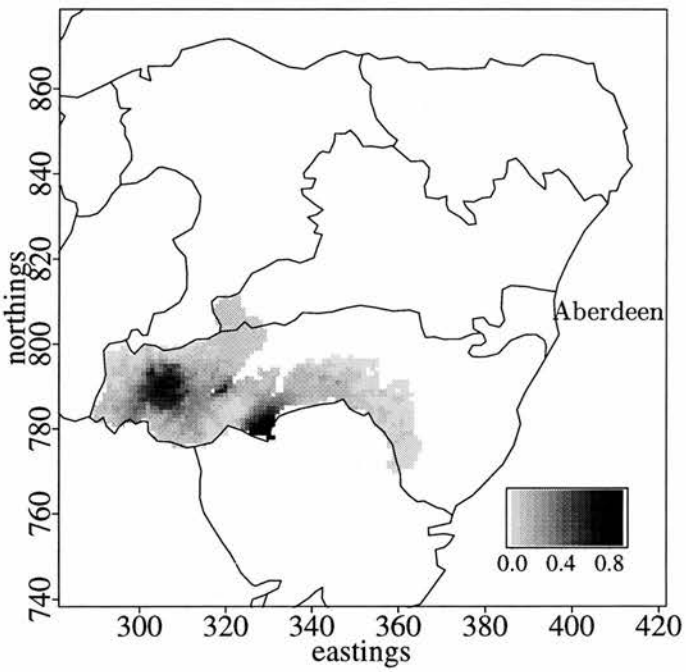
Figure 13: Map of the observed binary distribution of red deer in the Grampian Region of Scotland.

The models used by Augustin *et al.* are subject to stochastic variation throughout the fitting process, so the final probabilities differ between fits. Figure 14 shows probabilities from models 3 and 4 of that paper, averaged over 120 runs. It is considered that these average values provide an adequate summary of the output from each model and may be used for the purpose of model selection.

By contrast with the previous example, both of the models illustrated here suggest a similar pattern of occupation throughout the region. Model 3 is an autologistic regression model, combined with the Gibbs sampler. The Gibbs sampler is used to generate records of presence or absence from those squares where the true observation is missing, allowing



(a) Model 3.



(b) Model 4.

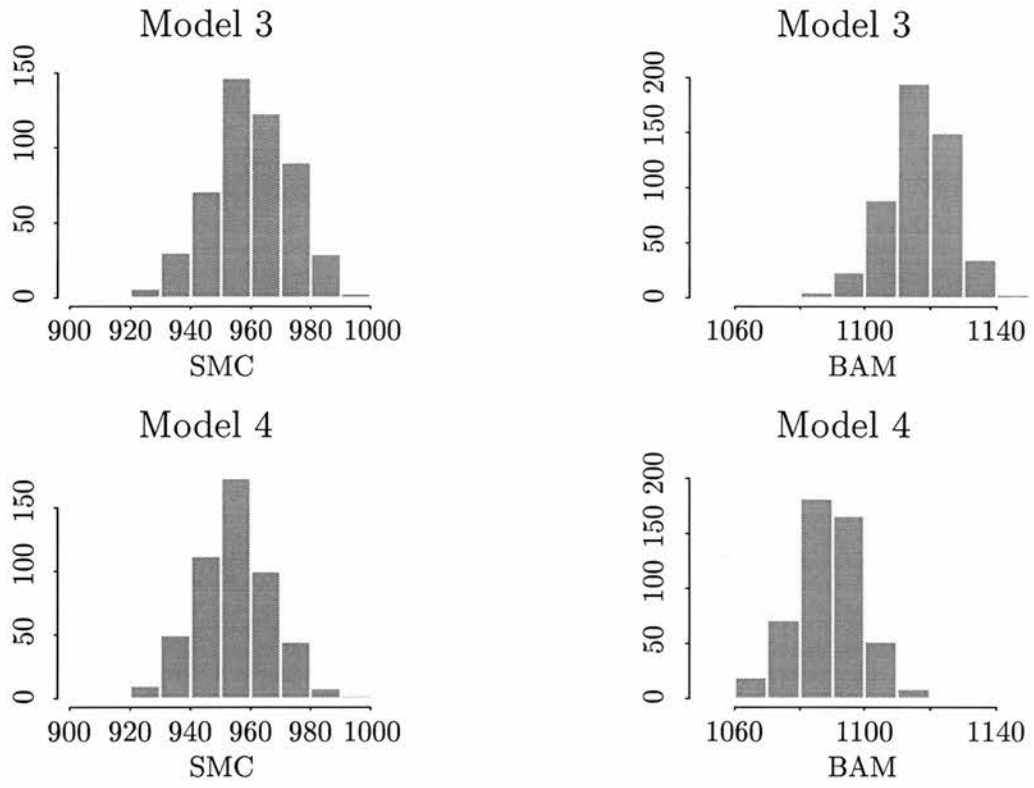
Figure 14: Estimated probabilities of occupation obtained from Models 3 and 4 of Augustin *et al.* (1996), averaged over 120 runs.

the autocovariate value to be calculated from these squares. In Model 4, a modified version of the Gibbs sampler is employed: instead of generating a record of presence or absence for each unsurveyed square, the autocovariate values for those squares are calculated directly from the underlying probabilities of presence-absence, as obtained from the Gibbs sampler. The modification has the effects of reducing stochastic variability in the model-fitting process, improving the convergence of the algorithm and reducing the computer time required to fit the model.

The performance of the two modelling approaches may be assessed by comparing predictions from the fitted models against the observed distribution. Predictions from Models 3 and 4 correspond to stochastic realizations of presence-absence from the probabilities in Figure 14 (a) and (b).

Histograms of the simple matching coefficient and best attainable match obtained by comparing the observed data in Figure 13 against 500 predictions from each of Models 3 and 4 are shown in Figure 15 (a) and (b). The clique of each site was taken as its eight nearest neighbours as in previous examples, and for the purpose of illustration the coefficients have not been scaled by the number of sites. The histograms of simple matching coefficient in (a) suggest that there is little difference between the models. However, the BAM histograms in (b) suggest that the match from Model 3 is better than that from Model 4.

The 500 predictions from the two models may be regarded as paired trials, and percentile intervals may be calculated for the quantities $SMC_3 - SMC_4$ and $BAM_3 - BAM_4$, where the subscripts indicate the model from which the corresponding coefficients were derived. For the simple matching coefficient, the central 95% percentile interval from 500 trials was $[-31, 41]$, providing no clear evidence that either one of the models yields better predictions. From the BAM coefficients, on the other hand, the central 95% percentile interval was $[1, 56]$, indicating that within the paired trials the BAM coefficients from Model 3 were almost always greater than those from Model 4. The percentile intervals may be used as an approximate test of the null hypothesis that the similarity coefficients from the two models derive from distributions with equal mean. The null hypothesis is rejected at the 5% level for the BAM coefficients, since the percentile interval does not encompass the point zero. The null hypothesis is not rejected at the 5% level for the SMC. The BAM coefficients have therefore uncovered evidence of a difference between the predictions of Models 3 and 4 that was not apparent from the simple matching coefficients, and the advantages of Model 4 appear to have occurred at some expense in performance.



(a) Simple matching coefficient.

(b) Best attainable match.

Figure 15: Histograms of simple matching coefficient and best attainable match from 500 stochastic realizations of presence-absence using average probabilities from each of Models 3 and 4.

The material in this section is intended only for illustrative purposes. If a detailed investigation of the performance of the above models is required, it might be advisable to select a clique size larger than that presented here since a single herd of deer can roam freely over a fairly large area. The binary algorithm is used for much more extensive analyses in the following chapters.

7 Concluding remarks

The best attainable match approach has been shown to produce improved similarity measures with respect to the most important features of a spatial distribution. Within the ecological context it allows for population mobility and provides more reliable means of discriminating between different population models than traditional coefficients. Models that yield good predictions of overall pattern are not penalized for minor departures from the observed data, and there is no need to place unrealistic trust in the accuracy of

the observed distribution. The flexibility in choice of clique structure allows a range of environmental considerations to be taken into account.

Algorithms have been presented for the computation of the maximum swap for both binary and abundance data. The binary algorithm runs efficiently and can be used with distributions of several thousand sites. The abundance algorithm is less efficient and can be unpredictable. A possible alternative is to use the abundance algorithm to eliminate all site-singletons and siteblocks of size (say) 4 or less, and to use the Ford-Fulkerson algorithm for those sites that form larger siteblocks. Elimination of the site-singletons and smaller blocks would break up the graph considerably and make it more tractable under the Ford-Fulkerson algorithm.

The measures of similarity presented in this chapter should be interpreted as comparative rather than absolute. Given a single observed distribution a similarity measure can be used to determine which of several predictions is most similar to it, but is not intended to indicate *how* similar they are in any absolute sense. For this reason, the values of best attainable match or minimum attainable distance arising from one problem should not be compared against those from another, and doubling the similarity index between two distributions in no way makes them 'twice as similar'. Notions of what is meant by similarity vary greatly between contexts, so attempts to produce an absolute index of similarity are probably misplaced.

Reference has been made to the fact that the BAM algorithms, both for presence-absence and for abundance data, do not construct explicit matchings on the graph but simply find the maximum number of elements in such a matching. For the most part this is an advantage, since the algorithms are used in conjunction with similarity measures for which the precise locations of swaps are immaterial — namely the simple matching coefficient and the Manhattan distance. These similarity measures will be suitable for most purposes, but there may be cases where a different measure is preferred for which the locations of swaps do affect the final score. Examples are the correlation coefficient and Euclidean distance measure mentioned in section 2.2. In cases such as these, a tremendous amount of extra effort would generally be required to determine which of the many possible BPS configurations yields the maximum similarity coefficient or minimum distance. In some instances the optimal configuration might not even attain the BPS.

One way in which this problem can be overcome is by computation of a mean similarity or distance over a number of possible BPS configurations. It would be reasonably straight-

forward to include in the BAM algorithms a procedure for the explicit construction of a matching. Since the algorithms pass through vertices in numerical order, a different numbering of the same vertices would produce a different matching. In this way several values of similarity or distance could be obtained by running the algorithm a number of times with vertex numbers for each run determined by a random permutation of $\{1, \ldots, N\}$. From these a mean similarity or distance could be calculated. Use of location-sensitive distance measures is, however, somewhat contrary to the spirit of the ecological application: the movements of animals are not usually encumbered by such considerations.

Topics for further work include investigation of the performance of the method in other applications. In particular there are problems in image analysis for which the best attainable match formulation could prove useful.

Acknowledgements

I am grateful to Nicole Augustin for providing the red deer data in section 6.2, and to Gavin Gibson for providing references to the Hungarian and Ford-Fulkerson algorithms.

Chapter 5

Analytic approach to parameter estimation in the colonization model

This chapter deals with likelihood calculations for the colonization model introduced in Chapter 3, when there are missing survey data. The problem described in Chapter 3 was the calculation of a likelihood function relating to a longer time period than that of the parameters: it may be expressed simply as the calculation of the quantity

$$L\left(\boldsymbol{y}^{(T)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right),$$

where $\boldsymbol{y}^{(T)}$ is the observed presence/absence distribution at time T, $\boldsymbol{y}^{(0)}$ is the initial presence/absence distribution at time 0, $\boldsymbol{\theta}$ is the vector of parameters and L is the likelihood function. The survey data $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(T-1)}$ are missing, and the parameters $\boldsymbol{\theta}$ relate to a single time-step. The parameters enter the likelihood through the colonization probabilities, $\{p_{ih}\}$, and their relation to a single time-step should be understood to mean that $p_{ih} = p_{ih}(\boldsymbol{\theta})$ is the probability that individuals or their offspring present in site i at time t have colonized site h by time $t+1$, one time-step later, for $t = 0, \ldots, T-1$. Note that p_{ih} might be a function of time, for example if $p_{ih} = p_0 \exp(-a\,\delta_{ih} - b\,\varsigma_h)$ as given in Chapter 3, and the site suitabilities ς_h change from one time-point to the next as for the woodlark data. In that case, p_{ih} will be written as $p_{ih}^{(t)}$. There is generally assumed to be no change in the parameters $\boldsymbol{\theta}$ over the time period $t = 0, \ldots, T$.

1 Formulation

1.1 Analogy with a branching process

Figure 1 shows in diagrammatic form the hidden processes of the colonization model for a system of 12 sites over 3 time-steps. The twelve sites are pictured separately for each of the four time-points, and colonizations are marked by arrows. The observed part of the process consists only of the pattern of presence and absence at two time-points: in this case at times $t = 0$ and $t = 3$, shown in Figure 2.

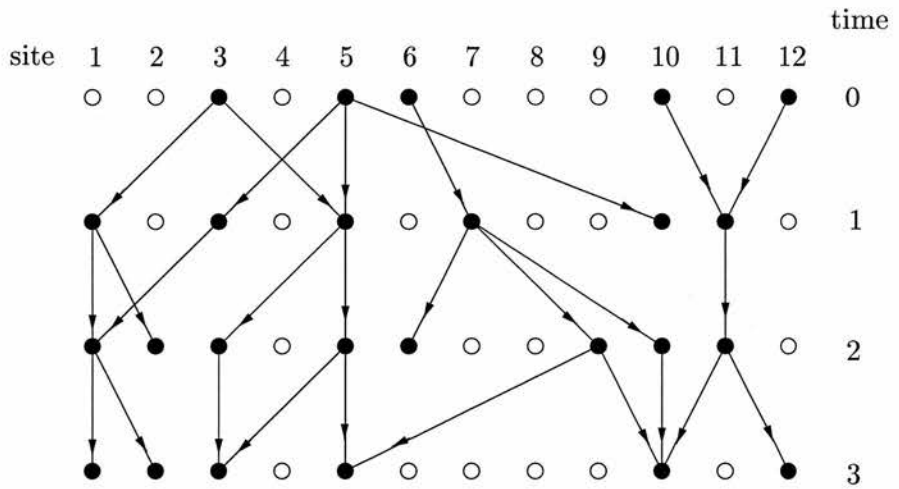


Figure 1: Diagrammatic representation of the hidden colonization process over 3 time-steps for 12 sites. Black circles represent occupied sites, and white circles represent unoccupied sites. Colonizations are denoted by arrows, and may occur only between an occupied site at time t and a site at time $t+1$ ($t = 0, 1, 2$). Any site that is colonized becomes occupied.

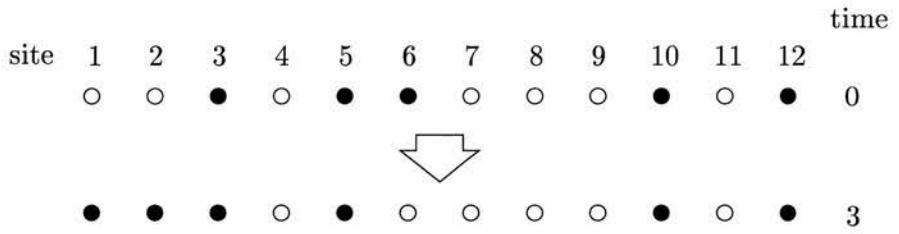


Figure 2: Observed part of the colonization process, consisting only of the pattern of presence and absence in all sites at times $t = 0$ and $t = 3$.

The hidden processes of Figure 1 bear some resemblance to a branching process. Branching processes are described in Volume I of Feller (1968), and are simple reproductive processes whereby individuals have a lifetime of a single time-step, at the end of which they produce offspring independently of each other. The offspring repeat the process in the next time-step, and so on. A branching process might involve individuals of several different *types*, where an individual of any type may produce offspring of any other type, each with a

certain probability.

In the current analogy, individuals of the branching process correspond to occupied sites, and reproductions correspond to colonizations of new sites. Individuals occur as N different types, where N is the total number of sites, since any site can colonize any other site but with a different probability for each. The colonization probability p_{ih} is therefore analogous to the probability that an individual of type i produces an offspring of type h in the next generation.

The parallel between the branching process and the colonization model is not complete, however, since the independence condition in the branching process does not hold for the colonization model. Suppose for instance that some site is colonized at time 1 by two of the occupied sites at time 0. This situation arises in Figure 1, as site 11 is colonized by both of sites 10 and 12. According to the branching process model, this would place two individuals of type 11 in the population at time 1, and these two individuals would then proceed to reproduce independently of each other. In the colonization model, however, there is no record of the number of ancestors associated with a newly-occupied site: site 11 is simply classified as occupied, regardless of the number of sites at time 0 from which it has been colonized. The two colonies in site 11 at time 1 are therefore considered to amalgamate in order to behave as a single occupied site.

It might seem at first that the branching process model is more realistic than the colonization model, since separate colonies in large sites might be expected to behave independently of each other in their colonizations. In fact, there are a number of problems with the raw branching process model in the ecological setting. Firstly, stationarity is lost. The time of the first survey is arbitrary, and at that time each site is observed to be either occupied or unoccupied. There is no knowledge of the ancestry of any site, and consequently no concept of a sub-colony in the site. According to the branching process model, however, ancestry becomes a crucial factor for all times $t > 0$: subcolonies are created within sites, each retaining their ancestral identity; these subcolonies give rise independently to new colonies in the next generation, and so on. Such a process has the potential to form large numbers of colonies in every site within a few years, each behaving independently. This would vastly augment the site-wise probabilities of new colonizations in the following generation. At best, the assumption of independence of colonizations of and from sites would become invalid; at worst, the process could go completely out of control. It is far more realistic to assume that some sort of stationarity has been reached, whereby the probability of colonization of one site from another is dependent on factors such as habitat suitability

but independent of the arbitrary starting point of the survey.

The second problem with the raw branching process is the lack of relevant ancestry data. Just as ancestry is unknown at the time of the first survey, so is it unknown at the time of the second — thus the number of subcolonies within any site is also unknown. Most of the structure of the colonization process would therefore have to be ignored in the fitting of the model.

Although the branching process is deemed unsuitable as a model in its raw form, there are considerable advantages to pursuing the similarities between that process and the process of the colonization model. The independence of reproductions in the branching process makes for easy calculation of certain important quantities, most notably the probability of non-occupation of any site at any time-point. If a site is unoccupied at a given time, colonization of the site must have failed along all possible ancestry routes. Since ancestry is preserved by the branching process model, and all individuals with different ancestry behave independently of each other, this probability is easy to compute.

Specifically, let $G_h^{(T)}(i,t)$ be the probability that site h is unoccupied at time T, given a single ancestor colony in site i at time t. In the analogy with a branching process, this is the probability that there is no individual of type h at time T, given a single ancestor of type i at time t. For the branching process,

$$G_h^{(T)}(i,t) = \prod_{j=1}^{N} \left(1 - p_{ij} + p_{ij}\, G_h^{(T)}(j,t+1)\right) , \tag{1}$$

because any ancestry route from site i at time t to site h at time T must pass through some site j at time $t+1$, so if colonization has failed from site i to site h it must have failed along each route i to j to h. The probability of a single failed route is the probability $1 - p_{ij}$ of failure in the initial step from i to j, plus the probability $p_{ij}\, G_h^{(T)}(j,t+1)$ of success in this initial step followed by failure in reaching h from j. Under the branching process model, the probabilities of failure from i to h via site j are independent for all j. The probability of non-occupation of h at time T may therefore be solved using the recurrence relation (1).

A recurrence of the form (1) is not possible for the colonization model, because ancestry is not preserved and consequently the probabilities of colonization failure from i to h via j are not independent over j. For example, consider the two routes $i \to j_1 \to k \to \cdots \to h$ and $i \to j_2 \to k \to \cdots \to h$. These routes both pass through the same site k, at which

point they become indistinguishable under the colonization model — despite having passed through different sites j_1 and j_2 at the first step. Failure of colonization along these routes cannot therefore be regarded as independent events, and this inhibits the expression of non-occupation probabilities in terms of a recursion.

Without a model formulation under which ancestry is preserved, it is not clear how probabilities of non-occupation might be calculable. In the next section it is shown that the branching process model can be modified so that the advantages of the recurrence relation (1) are retained, while emulating the non-occupation probabilities of the colonization model exactly.

1.2 Modification of the branching process

The essential difference between the colonization model and the branching process model is the behaviour when a single site is colonized from several sites in the previous generation. Under the colonization model, the separate colonies amalgamate to form a single occupied site. Under the branching process model, they remain separate and proceed to undertake new colonizations independently of each other. The only practical difference that this makes, however, is the net probability that a colonization occurs from that site to new sites in the following generation.

Figure 3 shows an example of the behaviour under the colonization model. Site i is colonized from three sites, j, k and l, at time 1, putting three colonies in i. These colonies constitute one occupied site, and the probability that site i colonizes any site h in the following generation is p_{ih}. Figure 4 shows the analogous situation for the branching process model. The three colonies in site i all have an independent opportunity to colonize site h in the next generation. In the raw branching process model, this probability would be p_{ih}; however, on Figure 4 it is written as q_{ih}. The overall probability that site i colonizes site h in the next generation depends on the colony-scale colonization probability q_{ih} and the number of colonies in site i; in this example with 3 colonies it is given by $1-(1-q_{ih})^3$.

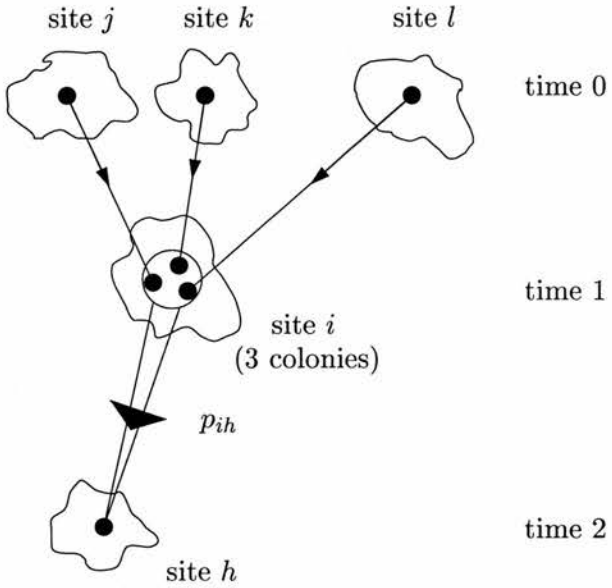


Figure 3: Colonization model behaviour when a site is colonized from more than one site in the previous generation. Site i has three colonies at time 1, which combine to act as a single occupied site. The probability that i colonizes any other site h in the next generation is the colonization probability p_{ih}, regardless of the number of colonies present in site i.

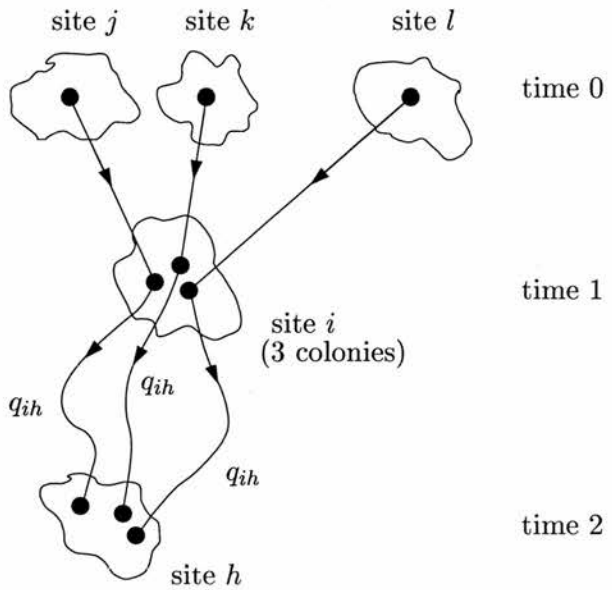


Figure 4: Branching process model behaviour when a site is colonized from more than one site in the previous generation. Site i has three colonies at time 1, which act separately and independently. The probability that i colonizes any other site h in the next generation depends on the number of colonies present in i, and for 3 colonies is given by $1-(1-q_{ih})^3$. Each colony will establish a new separate colony in site h if the colonization is successful.

The required modification to the branching process is now evident. The colony-scale colonization probabilities, which are marked on Figure 4 as q_{ih}, may be decreased so that the overall probability that site i colonizes site h attains the required value under the colonization model, which is p_{ih}. In the raw branching process, the colony-scale probabilities q_{ih} would simply be equal to p_{ih}. Reducing the colonization probabilities at the colony scale, and assuming independence of colonizations at this scale, has the effect of mimicking the non-independence of colonization paths at the site scale.

The modification of the branching process will ensure that all colonization probabilities at the site scale are identical to those of the colonization model, while still allowing the probability of non-occupation of any site at any time-point to be obtained by means of a recurrence relation as in (1). From the non-occupation probabilities, the probabilities of occupation are also immediate; for any site h the single-site likelihood $L\left(y_h^{(T)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(0)}}\right)$ is then either a probability of non-occupation at time T $(y_h^{(T)} = 0)$ or of occupation at time T: $y_h^{(T)} = 1$. Discussion of the calculation of the full likelihood $L\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(0)}}\right)$ is deferred until later sections, but it should be clear at this stage that the branching process analogy will greatly facilitate its computation.

The outstanding issue in applying the modification of the branching process is the calculation of the colony-scale colonization probabilities q_{ih}. This is not straightforward, since the required value q_{ih} depends upon the unknown number of colonies in site i. In the context of Figure 4, for example, there are 3 colonies in site i at time 1; the quantity q_{ih} must therefore be chosen to satisfy the expression

$$1 - (1 - q_{ih})^3 = p_{ih} \,.$$

In general, however, the distribution of colonies at time 1 is unobserved: it is not even known whether site i is occupied at that time. The probability q_{ih} must therefore be chosen by taking into account all possible numbers of colonies in the site i at time 1, and their probabilities. Since the distribution of the number of colonies in any site i will differ from one time-point to the next, the colony-scale colonization probabilities $\{q_{ih}\}_{h=1}^{N}$ will change accordingly. For this reason they will henceforth be written as $\{q_{ih}^{(t)}\}$ for sites $i, h = 1, \ldots, N$ and times $t = 0, \ldots, T-1$. The colony-scale probability $q_{ih}^{(t)}$ is the probability that a single colony in site i at time t will colonize site h at time $t+1$.

Most of this chapter will be devoted to the derivation of expressions for the colony-scale probabilities $\{q_{ih}^{(t)}\}$ and their approximate solutions. First of all, it is useful to establish

precise definitions of terminology.

A *colonization* from a site i to a site h is the process of individuals or their offspring that were present in site i at some time t moving to site h at time $t+1$.

A *colony* is established as a result of every colonization. Each occupied site at time 0 is defined arbitrarily as containing a single colony. At time 1, the number of colonies in a site h is precisely equal to the number of sites at time 0 from which colonizations to site h have occurred. All colonies in site h at time 1 are then able to produce new colonizations independently, and each colonization establishes a new colony in the target site (Figure 5). The essential feature is that the full ancestry of any colony is uniquely defined.

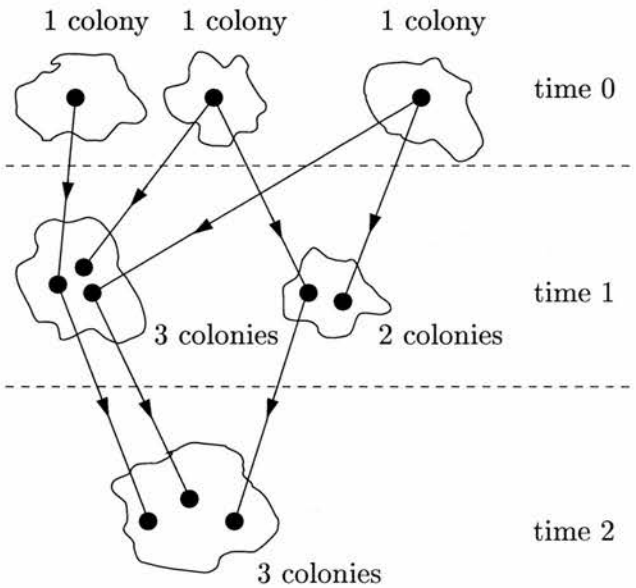


Figure 5: Definition of a colony. At time 0, all occupied sites contain precisely one colony. At time 1, several colonies may be established in a site, and these colonies can each establish new colonies at time 2.

The *raw branching process* model is the conventional branching process with individuals of N types. There is a constant probability p_{ih} that an individual of type i in any generation produces offspring of type h. No individual may produce more than one offspring of any type, but there is no restriction on the number of offspring of different types that an individual may produce.

The *modified branching process* is the raw branching process modified so that the probability $q_{ih}^{(t)}$ of an individual of type i at time t producing an offspring of type h at time $t+1$ is reduced according to the statistical distribution of the number of colonies in site i at time t. The modification is chosen so that the occupation probabilities of the colonization model are reproduced exactly.

The *colony-scale* colonization probability $q_{ih}^{(t)}$ is the probability that a single colony in site i at time t colonizes site h at time $t+1$. It corresponds to the appropriate probability of the modified branching process. The *site-scale* colonization probability p_{ih} is the probability that site i colonizes site h over any time-step, given that site i is occupied at the beginning of the time-step. It may be thought of as the probability that *some* colony in i at time t colonizes site h at time $t+1$, given that there is at least one colony in site i at time t. For the most part p_{ih} is assumed to be constant over time, but if there is time-dependence the probability should be written as $p_{ih}^{(t)}$.

The aim of the work in forthcoming sections is the approximate calculation of the colony-scale probabilities $q_{ih}^{(t)}$. These will eventually enable the likelihood function to be approximately determined; the parameter estimates may be finally obtained by maximizing the likelihood with respect to the parameters.

1.3 Overview

The plan for the remainder of this chapter is as follows. In Section 2 an expression will be derived for the colony-scale colonization probability $q_{ih}^{(t)}$, which will be shown to be the solution to a polynomial of prohibitively large order. Close exponential approximations to the polynomial function are derived in Section 3, which enable the approximate solution for $q_{ih}^{(t)}$ to be found relatively easily. In Section 4 the closeness and reliability of the exponential approximations are examined, by means of exact error bounds and simulation tests. Methods of likelihood estimation from the modified branching process model are presented in Section 5, at the end of which a full summary of the analytic approach to parameter estimation is provided. Section 6 gives some practical details of model implementation, together with two examples: application to the woodlark data, and application to a simulated dataset. A few final points are presented in Section 8.

A glossary of the notation used in this chapter is provided in Appendix B.

2 Derivation of the colony-scale colonization probabilities

2.1 Expression for the colony-scale colonization probabilities

Let $N_i^{(t)}$ be the random variable denoting the number of colonies in site i at time t, and let $p_i^{(t)}$ be the probability that site i is occupied at time t:

$$p_i^{(t)} = \mathbb{P}(N_i^{(t)} > 0)\,.$$

Recall that the colony-scale colonization probability $q_{ih}^{(t)}$ is the probability that a single given colony in site i at time t colonizes the site h at time $t+1$. The conditional probability p_{ih} that site h is colonized at time $t+1$ by *some* colony in site i at time t, given that i is occupied at time t, is referred to as the site-scale colonization probability. To avoid excessive cluttering of notation the site-scale colonization probabilities are assumed constant over time, although the extension to the time-varying case is straightforward.

Each of the $N_i^{(t)}$ colonies in site i at time t has an independent probability $q_{ih}^{(t)}$ of colonizing site h at time $t+1$. Taking into account all possible values of $N_i^{(t)}$, and the probability of each value, the colony-scale colonization probability $q_{ih}^{(t)}$ is chosen to ensure that the overall probability of h being colonized from i during the interval $(t, t+1]$, conditional on i being occupied at the beginning of the interval, is the site-scale probability p_{ih} in accordance with the original model. The aim of this and following sections is to calculate the colony-scale colonization probabilities.

The notation $i \to^{(t)} h$ is used to denote colonization at the site scale from site i to site h in the interval $(t, t+1]$, and the notation $i \not\to^{(t)} h$ to indicate no colonization between i and h in this interval. The relationship between the colony-scale probabilities $\left\{q_{ih}^{(t)}\right\}$ and the site-scale probabilities $\{\, p_{ih}\}$ is established by considering $\mathbb{P}(i \not\to^{(t)} h)$ from the two different perspectives. At the site scale,

$$\begin{aligned}
\mathbb{P}(i \not\to^{(t)} h) &= \mathbb{P}(i \not\to^{(t)} h \,|\, N_i^{(t)} > 0)\,\mathbb{P}(N_i^{(t)} > 0) + \mathbb{P}(i \not\to^{(t)} h \,|\, N_i^{(t)} = 0)\,\mathbb{P}(N_i^{(t)} = 0) \\
&= (1 - p_{ih})\, p_i^{(t)} + (1 - p_i^{(t)}) \\
&= 1 - p_{ih}\, p_i^{(t)}\,; \qquad (2)
\end{aligned}$$

while at the colony scale, in terms of $q_{ih}^{(t)}$,

$$\begin{aligned}
\mathbb{P}(i \not\to^{(t)} h) &= \mathbb{P}(i \not\to^{(t)} h \,|\, N_i^{(t)} = 0)\, \mathbb{P}(N_i^{(t)} = 0) \quad + \\
&\qquad \mathbb{P}(i \not\to^{(t)} h \,|\, N_i^{(t)} = 1)\, \mathbb{P}(N_i^{(t)} = 1) \quad + \\
&\qquad\qquad \mathbb{P}(i \not\to^{(t)} h \,|\, N_i^{(t)} = 2)\, \mathbb{P}(N_i^{(t)} = 2) \quad + \quad \ldots \\
&= \mathbb{P}(N_i^{(t)} = 0) \quad + \quad \left(1 - q_{ih}^{(t)}\right) \mathbb{P}(N_i^{(t)} = 1) \quad + \\
&\qquad\qquad \left(1 - q_{ih}^{(t)}\right)^2 \mathbb{P}(N_i^{(t)} = 2) \quad + \quad \ldots \\
&= \sum_{r \geq 0} \left(1 - q_{ih}^{(t)}\right)^r \mathbb{P}(N_i^{(t)} = r) \\
&= \mathbb{E}\left\{ \left(1 - q_{ih}^{(t)}\right)^{N_i^{(t)}} \right\} . \qquad (3)
\end{aligned}$$

Putting (2) and (3) together, $q_{ih}^{(t)}$ must be chosen so that

$$\mathbb{E}\left\{ \left(1 - q_{ih}^{(t)}\right)^{N_i^{(t)}} \right\} = 1 - p_{ih}\, p_i^{(t)} . \qquad (4)$$

For ease of notation the function $s_i^{(t)}$ is introduced, defined as

$$s_i^{(t)}(x) = \mathbb{E}\left\{ (1-x)^{N_i^{(t)}} \right\} . \qquad (5)$$

The equation to be solved for $q_{ih}^{(t)}$ may therefore be written

$$s_i^{(t)}\left(q_{ih}^{(t)}\right) - 1 + p_{ih}\, p_i^{(t)} = 0 . \qquad (6)$$

2.2 Form of $s_i^{(t)}(x)$

In order to investigate the form of the function $s_i^{(t)}$, some further notation is introduced. The random variable $V_i^{(t)}(k,u)$ is defined as the number of colonies established in site i at time t, starting from a single colony in site k at time u. By definition, there is at most one colony in any site in year 0; it follows that

$$N_i^{(t)} = \sum_{k_0 : N_{k_0}^{(0)} = 1} V_i^{(t)}(k_0, 0) . \qquad (7)$$

For convenience no notational distinction is made between the underlying random variable $V_i^{(t)}(k,u)$ and a specific instance of the random variable distributed as $V_i^{(t)}(k,u)$.

From (7),

$$\begin{aligned} s_i^{(t)}(x) &= \mathbb{E}\left\{(1-x)^{N_i^{(t)}}\right\} \\ &= \mathbb{E}\left\{\prod_{k_0:N_{k_0}^{(0)}=1}(1-x)^{V_i^{(t)}(k_0,0)}\right\} \\ &= \prod_{k_0:N_{k_0}^{(0)}=1}\mathbb{E}\left\{(1-x)^{V_i^{(t)}(k_0,0)}\right\} \end{aligned} \tag{8}$$

since the random variables $\left\{V_i^{(t)}(k_0,\,0)\right\}_{k_0:N_{k_0}^{(0)}=1}$ are independent.

Proposition 3 *For any sites k_u and i, and times u and t with $u < t-1$,*

$$\mathbb{E}\left((1-x)^{V_i^{(t)}(k_u,u)}\right) = \prod_{k_{u+1}}\left\{1 - q_{k_u k_{u+1}}^{(u)} + q_{k_u k_{u+1}}^{(u)}\,\mathbb{E}\left((1-x)^{V_i^{(t)}(k_{u+1},u+1)}\right)\right\}. \tag{9}$$

When $u = t-1$,

$$\mathbb{E}\left((1-x)^{V_i^{(t)}(k_{t-1},t-1)}\right) = 1 - q_{k_{t-1}i}^{(t-1)}\,x\,. \tag{10}$$

(Note that the dummy variable k_u is indexed by time u purely to distinguish it from the dummy variables used for different times; the variable k_u can assume any value in $1,\ldots,N$, and the index u does not represent any changing properties of the sites over time.)

Proof For a specific colony in site k_u at time u, let

$$I\{k_u \to^{(u)} k_{u+1}\} = \begin{cases} 1 & \text{if the colony in site } k_u \text{ colonizes site } k_{u+1} \text{ at time } u+1, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$V_i^{(t)}(k_u,u) = \sum_{k_{u+1}} I\{k_u \to^{(u)} k_{u+1}\}\; V_i^{(t)}(k_{u+1},u+1)\,,$$

so

$$\mathbb{E}(1-x)^{V_i^{(t)}(k_u,u)} = \mathbb{E}\left\{\prod_{k_{u+1}}(1-x)^{I\{k_u\to^{(u)}k_{u+1}\}\;V_i^{(t)}(k_{u+1},u+1)}\right\}$$

$$= \prod_{k_{u+1}} \mathbb{E}\left((1-x)^{I\{k_u \to^{(u)} k_{u+1}\}\, V_i^{(t)}(k_{u+1},u+1)}\right)$$

(by independence of colonizations at the colony scale to and from different sites)

$$= \prod_{k_{u+1}} \mathbb{E}\left\{\mathbb{E}\left((1-x)^{I\{k_u \to^{(u)} k_{u+1}\}\, V_i^{(t)}(k_{u+1},u+1)} \;\middle|\; I\{k_u \to^{(u)} k_{u+1}\}\right)\right\}$$

$$= \prod_{k_{u+1}} \Big\{\mathbb{E}\left((1-x)^0\right)\mathbb{P}\left(I\{k_u \to^{(u)} k_{u+1}\}=0\right) \quad + \quad \mathbb{E}\left((1-x)^{V_i^{(t)}(k_{u+1},u+1)}\right)\mathbb{P}\left(I\{k_u \to^{(u)} k_{u+1}\}=1\right)\Big\}$$

$$= \prod_{k_{u+1}} \left\{1 - q^{(u)}_{k_u k_{u+1}} + q^{(u)}_{k_u k_{u+1}}\mathbb{E}\left((1-x)^{V_i^{(t)}(k_{u+1},u+1)}\right)\right\} \quad \text{as required.}$$

When $u = t-1$, the random variable $V_i^{(t)}(k_{t-1},\, t-1)$ is equal to $I\{k_{t-1} \to^{(t-1)} i\}$ and can take only the values 0 and 1, whence

$$\mathbb{E}\left((1-x)^{V_i^{(t)}(k_{t-1},t-1)}\right) = (1-x)\,\mathbb{P}\left(I\{k_{t-1} \to^{(t-1)} i\}=1\right) + (1-x)^0\,\mathbb{P}\left(I\{k_{t-1} \to^{(t-1)} i\}=0\right)$$

$$= (1-x)\,q^{(t-1)}_{k_{t-1}i} + 1 - q^{(t-1)}_{k_{t-1}i}$$

$$= 1 - q^{(t-1)}_{k_{t-1}i}\,x \qquad \text{as required.} \qquad \square$$

Proposition 4 *$s_i^{(t)}(x)$ is given by the following polynomial for $t \geq 2$:*

$$s_i^{(t)}(x) = \prod_{k_0 : N^{(0)}_{k_0}=1} \prod_{k_1=1}^{N} \left\{1 - q^{(0)}_{k_0k_1} + q^{(0)}_{k_0k_1} \prod_{k_2=1}^{N} \left\{1 - q^{(1)}_{k_1k_2} + q^{(1)}_{k_1k_2} \prod_{k_3=1}^{N} \left\{1 - q^{(2)}_{k_2k_3} + \dots \right.\right.\right.$$
$$\left.\left.\left. \dots + q^{(t-3)}_{k_{t-3}k_{t-2}} \prod_{k_{t-1}=1}^{N} \left\{1 - q^{(t-2)}_{k_{t-2}k_{t-1}}\, q^{(t-1)}_{k_{t-1}i}\, x\right\} \dots\right\}\right\}\right\}. \qquad (11)$$

(Note that $q^{(0)}_{k_0k_1} = p_{k_0k_1}$ and that the first product is over those sites k_0 occupied at time 0.)

Proof Immediate from (8), (9) and (10). □

Equations (6) and (11) indicate that the colony-scale colonization probability $q_{ih}^{(t)}$ is the solution to a polynomial given by $s_i^{(t)}(x) - 1 + p_{ih}\, p_i^{(t)} = 0$, involving the quantities $\left\{q_{vw}^{(u)}\right\}$ for all sites v and w and all times $u < t$. Since the polynomial has large degree, it is necessary to establish existence and uniqueness of roots in the interval of interest.

Proposition 5 *Let* $f(x) = s_i^{(t)}(x) - 1 + p_{ih}\, p_i^{(t)}$. *The equation* $f(x) = 0$ *has precisely one root in the interval* $[0,\, p_{ih}]$, *provided that* $p_i^{(t)} > 0$. *If* $p_i^{(t)} = 0$ *then* $f(x) = 0 \;\; \forall x \in [0,1]$.

Proof First consider the case with $p_i^{(t)} = 0$. In this case site i is unoccupied with certainty at time t, so $N_i^{(t)} = 0$ with probability 1, and by the definition in (5), $s_i^{(t)}(x) = \mathbb{E}(1-x)^{N_i^{(t)}} = 1$. Thus for all $x \in [0,1]$,

$$f(x) \;=\; 1 - 1 + 0 \;=\; 0 \qquad \text{as stated.}$$

When $p_i^{(t)} = 0$, the infinite number of solutions is of no consequence, since if i is unoccupied at time t there is no need to calculate $q_{ih}^{(t)}$.

Now consider the case where $p_i^{(t)} > 0$. It is shown that:

1. $f(0) \geq 0$,
2. $f(1) \leq 0$, and
3. f is monotonically decreasing on the interval $[0,1]$,

proving that there is a single root between 0 and 1. Finally it is shown that this root occurs between 0 and p_{ih}.

1. $f(0) = 1 - 1 + p_{ih}\, p_i^{(t)} = p_{ih}\, p_i^{(t)} \quad \begin{cases} > 0 & \text{if } p_{ih} > 0. \\ = 0 & \text{if } p_{ih} = 0. \end{cases}$

2. $f(1)$:

$$\begin{aligned} f(1) &= \mathbb{E}\left\{(1-1)^{N_i^{(t)}}\right\} - 1 + p_{ih}\, p_i^{(t)} \\ &= \mathbb{P}(N_i^{(t)} = 0) - 1 + p_{ih}\, p_i^{(t)} \\ &= \left(1 - p_i^{(t)}\right) - 1 + p_{ih}\, p_i^{(t)} \\ &= -p_i^{(t)}(1 - p_{ih}) \quad \begin{cases} < 0 & \text{if } p_{ih} < 1. \\ = 0 & \text{if } p_{ih} = 1. \end{cases} \end{aligned}$$

3. Consider

$$f(x) = \mathbb{E}(1-x)^{N_i^{(t)}} - 1 + p_{ih}\, p_i^{(t)}$$

$$= \sum_{r=0}^{\infty} (1-x)^r \mathbb{P}(N_i^{(t)} = r) - 1 + p_{ih}\, p_i^{(t)} .$$

Now if $0 \leq x_1 < x_2 \leq 1$, then $(1-x_1)^r > (1-x_2)^r \ \forall r$, so that $f(x_1) > f(x_2)$. Thus f is monotonically decreasing on $[0,1]$.

Hence if $p_{ih} = 0$ there is a single root $x = 0$, if $p_{ih} = 1$ there is a single root $x = 1$, and if $0 < p_{ih} < 1$ there is a single root x in the interval (0,1).

It remains to show that the single root lies in the interval $[0, p_{ih}]$ when $p_{ih} < 1$ and $p_i^{(t)} > 0$. Let x be the root of f in $[0,1]$. Then

$$f(x) = 0 \quad \Rightarrow \quad \mathbb{E}\left\{(1-x)^{N_i^{(t)}}\right\} = 1 - p_{ih}\, p_i^{(t)} ,$$

$$\mathbb{P}(N_i^{(t)}=0) + (1-x)\mathbb{P}(N_i^{(t)}=1) + (1-x)^2\mathbb{P}(N_i^{(t)}=2) + \ldots = 1 - p_{ih}\left(1 - \mathbb{P}(N_i^{(t)}=0)\right) .$$

Subtracting $\mathbb{P}(N_i^{(t)}=0)$ from both sides gives

$$(1-x)\mathbb{P}(N_i^{(t)}=1) + (1-x)^2\mathbb{P}(N_i^{(t)}=2) + \ldots = (1-p_{ih})\left(1 - \mathbb{P}(N_i^{(t)}=0)\right) ,$$

$$(1-x)\left\{\mathbb{P}(N_i^{(t)}=1) + (1-x)\mathbb{P}(N_i^{(t)}=2) + \ldots\right\} = (1-p_{ih})\left\{\mathbb{P}(N_i^{(t)}=1) + \mathbb{P}(N_i^{(t)}=2) + \ldots\right\} ,$$

whence

$$\frac{1-x}{1-p_{ih}} = \frac{\mathbb{P}(N_i^{(t)}=1) + \mathbb{P}(N_i^{(t)}=2) + \ldots}{\mathbb{P}(N_i^{(t)}=1) + (1-x)\mathbb{P}(N_i^{(t)}=2) + \ldots} . \tag{12}$$

Since $x \geq 0$, each term in the numerator of (12) is $\geq$ the corresponding term in the denominator. Thus

$$\frac{1-x}{1-p_{ih}} \geq 1$$

$$\Rightarrow \quad x \leq p_{ih} ,$$

and the unique root $x \in [0,1]$ lies in the interval $[0, p_{ih}]$. □

Note From equations (6) and (11), the colony-scale colonization probability $q_{ih}^{(t)}$ satisfies

$$1 - p_{ih}\, p_i^{(t)} = \prod_{k_0 : N_{k_0}^{(0)} = 1} \prod_{k_1=1}^{N} \left\{ 1 - q_{k_0 k_1}^{(0)} + q_{k_0 k_1}^{(0)} \prod_{k_2=1}^{N} \left\{ 1 - q_{k_1 k_2}^{(1)} + q_{k_1 k_2}^{(1)} \prod_{k_3=1}^{N} \left\{ 1 - q_{k_2 k_3}^{(2)} + \dots \right.\right.\right.$$

$$\left.\left.\left. \dots + q_{k_{t-3} k_{t-2}}^{(t-3)} \prod_{k_{t-1}=1}^{N} \left\{ 1 - q_{k_{t-2} k_{t-1}}^{(t-2)}\, q_{k_{t-1} i}^{(t-1)}\, q_{ih}^{(t)} \right\} \dots \right\} \right\} \right\} . \qquad (13)$$

The intuition behind this result is as follows. Let A, B be the events

$$A = \left\{ i \not\to^{(t)} h \right\},$$

$$B = \left\{ \exists \text{ no colonization path } k_0 \to^{(0)} k_1 \to^{(1)} \dots \to^{(t-1)} i \to^{(t)} h \text{ for any sites } k_0, \dots, k_{t-1} \right\}.$$

Clearly, $A \subseteq B$ and $B \subseteq A$: thus $A = B$ and $\mathbb{P}(A) = \mathbb{P}(B)$. Now

$$\mathbb{P}(A) = 1 - p_{ih} p_i^{(t)}$$

from (2), while

$$\mathbb{P}(B) = \prod_{k_0 : N_{k_0}^{(0)} = 1} \mathbb{P} \left(\begin{array}{c} \text{no colonization path } k_0 \to^{(0)} k_1 \to^{(1)} \dots \\ \dots k_{t-1} \to^{(t-1)} i \to^{(t)} h \text{ for any } k_1, \dots, k_{t-1} \end{array} \right)$$

$$= \prod_{k_0} \prod_{k_1=1}^{N} \left\{ \mathbb{P}(k_0 \not\to^{(0)} k_1) + \mathbb{P} \left(\begin{array}{c} \text{no path } k_1 \to^{(1)} \dots \to^{(t-1)} \\ i \to^{(t)} h \text{ for any } k_2, \dots, k_{t-1} \end{array} \middle| \, k_0 \to^{(0)} k_1 \right) \mathbb{P}(k_0 \to^{(0)} k_1) \right\}$$

$$= \prod_{k_0 : N_{k_0}^{(0)} = 1} \prod_{k_1=1}^{N} \left\{ 1 - q_{k_0 k_1}^{(0)} + q_{k_0 k_1}^{(0)} \prod_{k_2=1}^{N} \mathbb{P} \left(\begin{array}{c} \text{no path } k_1 \to^{(1)} k_2 \to^{(2)} \dots k_{t-1} \\ \to^{(t-1)} i \to^{(t)} h \text{ for any } k_3, \dots, k_{t-1} \end{array} \right) \right\}$$

$$= \dots ,$$

as in the branching process recurrence relation (1). This expression eventually becomes the right-hand side of (11), and equating $\mathbb{P}(A) = \mathbb{P}(B)$ gives (13).

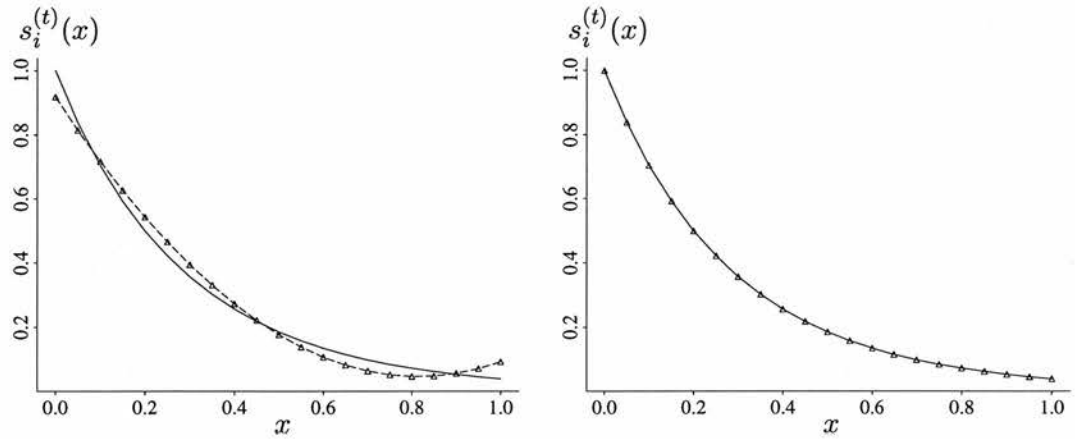
The function $s_i^{(t)}(x) - 1 + p_{ih}\, p_i^{(t)}$ will henceforth be referred to as the *colony-scale polynomial.*

3 Calculation of the colony-scale colonization probabilities

3.1 Motivation

The colony-scale polynomial $s_i^{(t)}(x) - 1 + p_{ih}\, p_i^{(t)}$ to be solved for the colony-scale colonization probability $x = q_{ih}^{(t)}$ is a polynomial in x of degree up to N^t, with precisely one root in the interval $[0,\, p_{ih}]$. It would be extremely time-consuming to find the root of such a polynomial by numerical techniques; furthermore a separate polynomial would have to be solved for every combination of i, h and t. If the modified branching process is to provide a viable means of calculating the likelihood, therefore, the polynomial must be approximated in some way.

In this section a close approximation to the colony-scale polynomial is derived. An exponential approximation is used, motivated by simulation results illustrated in Figure 6. The function $s_i^{(t)}(x)$ was calculated over the interval [0,1] for various values of t and sets of simulated probabilities $\left\{q_{ih}^{(t)}\right\}$, and simple functions were fitted. Exponential approximations of the form $s_i^{(t)}(x) \approx \exp(\alpha + \beta x + \gamma x^2)$ were found to fit the observed functions very well, while the fits from polynomial approximations such as $s_i^{(t)}(x) \approx \alpha + \beta x + \gamma x^2$ were relatively poor. The results in Figure 6 are typical of those obtained from many simulations.



(a) Quadratic approximation.

(b) Exponential approximation.

Figure 6: Best-fit lines from different approximations to $s_i^{(t)}(x)$ for $t = 5$. The solid lines give $s_i^{(5)}(x)$ using simulated values of $\left\{q_{ih}^{(t)}\right\}_{t=0}^{4}$ for $i, h = 1, \ldots, 100$. The dashed lines marked with triangles give the least-squares fits using the models (a) $s_i^{(t)}(x) = \alpha + \beta x + \gamma x^2$ and (b) $s_i^{(t)}(x) = \exp(\alpha + \beta x + \gamma x^2)$. The fit in (b) is much closer than that in (a).

Once the polynomial $s_i^{(t)}(x)$ is approximated as $\exp(\alpha + \beta x + \gamma x^2)$ for α, β and γ to be determined, the solution to the colony-scale polynomial is straightforward: $q_{ih}^{(t)}$ is that value of x satisfying

$$s_i^{(t)}(x) = 1 - p_{ih} p_i^{(t)} ,$$

so the approximate solution occurs when

$$\alpha + \beta x + \gamma x^2 = \log(1 - p_{ih} p_i^{(t)}) \tag{14}$$

which can be solved for x as a simple quadratic equation. Furthermore, since the function $s_i^{(t)}$ does not involve h, the exponential coefficients α, β and γ need only be calculated once to obtain $q_{ih}^{(t)}$ for all h. Again, this represents a huge saving of time over numerical root-finding for the full colony-scale polynomial $s_i^{(t)}(x) - 1 + p_{ih}\, p_i^{(t)}$, which depends on i, t and h. Overall, the exponential approximations will therefore reduce the problem of solving the colony-scale polynomials from one which is computationally impracticable to one with an easily-calculable solution.

From simulation tests and some exact results to be discussed later, the second-order approximations used throughout this section are considered highly adequate. Taking higher-order approximations has two disadvantages. Firstly, the exponential coefficients become much more time-consuming to compute. Secondly, expressions of the form (14) become more difficult to solve when the left-hand side is a cubic or higher-order polynomial.

3.2 Taylor expansion of $\log s_i^{(t)}(x)$

Let

$$g(x) = \log s_i^{(t)}(x) = \log \mathbb{E}\left\{(1-x)^{N_i^{(t)}}\right\}$$

for given i, t. The function g is not indexed by i and t in order to avoid cluttering the notation. It is required to approximate g as

$$g(x) \approx \alpha + \beta x + \gamma x^2 \tag{15}$$

for x in the range [0,1].

Let $g'(x)$ be the first derivative of g with respect to its argument, evaluated at the point x, and let $g''(x)$ be the second derivative. In order to find the coefficients α, β and γ, g

may be expanded as a Taylor series about 0:

$$g(x) = g(0) + x\,g'(0) + \frac{x^2}{2}\,g''(0) + \dots \,. \tag{16}$$

This suggests that a close approximation to g is likely to be provided by assigning $\alpha = g(0)$, $\beta = g'(0)$ and $\gamma = g''(0)/2$.

Now

$$g(0) = \log \mathbb{E}\left\{(1-0)^{N_i^{(t)}}\right\} = \log(1) = 0\,,$$

so $\alpha = 0$. It remains to calculate $g'(0)$ and $g''(0)$.

Calculation of $g'(0)$

Let R be any random variable on the non-negative integers, and let $h(x) = \mathbb{E}(1-x)^R$. Then

$$h(x) = \sum_{r=0}^{\infty} (1-x)^r\, \mathbb{P}(R=r)\,,$$

so

$$h'(x) \;=\; \sum_{r=0}^{\infty} -r\,(1-x)^{r-1}\,\mathbb{P}(R=r) \;=\; -\mathbb{E}\left(R(1-x)^{R-1}\right) \tag{17}$$

and

$$h''(x) \;=\; \sum_{r=0}^{\infty} r\,(r-1)\,(1-x)^{r-2}\,\mathbb{P}(R=r) \;=\; \mathbb{E}\left(R\,(R-1)\,(1-x)^{R-2}\right). \tag{18}$$

Putting $x = 0$ in (17) and (18) gives

$$h'(0) = -\mathbb{E}\,R, \qquad h''(0) = \mathbb{E}\,R^2 - \mathbb{E}\,R\,, \tag{19}$$

as long as these quantities exist.

Now consider $g(x) = \log s_i^{(t)}(x)$. From (8),

$$s_i^{(t)}(x) = \prod_{k_0 : N_{k_0}^{(0)} = 1} \mathbb{E}\left\{(1-x)^{V_i^{(t)}(k_0\,,\,0)}\right\},$$

so that

$$g(x) \;=\; \sum_{k_0 : N_{k_0}^{(0)} = 1} \log \mathbb{E}(1-x)^{V_i^{(t)}(k_0\,,\,0)}\,.$$

Using (17), with $V_i^{(t)}(k_0\,,\,0)$ substituted for R, it follows that

$$g'(x) \;=\; \sum_{k_0:N_{k_0}^{(0)}=1} \frac{-\mathbb{E}\left(V_i^{(t)}(k_0\,,\,0)\,(1-x)^{V_i^{(t)}(k_0\,,0)-1}\right)}{\mathbb{E}\left((1-x)^{V_i^{(t)}(k_0\,,0)}\right)}\,. \tag{20}$$

Thus from (19),

$$g'(0) = \;-\sum_{k_0:N_{k_0}^{(0)}=1} \mathbb{E}\left(V_i^{(t)}(k_0\,,\,0)\right)\,. \tag{21}$$

Proposition 6 *For any sites k_u and i, and times u and t with $u < t-1$,*

$$\mathbb{E}\left(V_i^{(t)}(k_u, u)\right) = \sum_{k_{u+1}} q_{k_u k_{u+1}}^{(u)}\, \mathbb{E}\left(V_i^{(t)}(k_{u+1}, u+1)\right)\,. \tag{22}$$

When $u = t-1$,

$$\mathbb{E}\left(V_i^{(t)}(k_{t-1}, t-1)\right) \;=\; q_{k_{t-1}i}^{(t-1)}\,. \tag{23}$$

Proof Using the notation $I\{k_u \to^{(u)} k_{u+1}\}$ defined on page 105 for a single colony in site k_u at time u,

$$V_i^{(t)}(k_u, u) = \sum_{k_{u+1}} I\{k_u \to^{(u)} k_{u+1}\}\; V_i^{(t)}(k_{u+1}, u+1)\,,$$

so

$$\begin{aligned}
\mathbb{E}\left(V_i^{(t)}(k_u, u)\right) \;&=\; \sum_{k_{u+1}} \mathbb{E}\left(I\{k_u \to^{(u)} k_{u+1}\}\; V_i^{(t)}(k_{u+1}, u+1)\right) \\
&=\; \sum_{k_{u+1}} \mathbb{E}\left\{\mathbb{E}\left(I\{k_u \to^{(u)} k_{u+1}\}\; V_i^{(t)}(k_{u+1}, u+1)\;\middle|\; I\{k_u \to^{(u)} k_{u+1}\}\right)\right\} \\
&=\; \sum_{k_{u+1}} \Big\{\mathbb{E}(0)\,\mathbb{P}\left(I\{k_u \to^{(u)} k_{u+1}\} = 0\right) \;\;+ \\
&\qquad\qquad \mathbb{E}\left(V_i^{(t)}(k_{u+1}, u+1)\right)\,\mathbb{P}\left(I\{k_u \to^{(u)} k_{u+1}\} = 1\right)\Big\} \\
&=\; \sum_{k_{u+1}} q_{k_u k_{u+1}}^{(u)}\, \mathbb{E}\left(V_i^{(t)}(k_{u+1}, u+1)\right) \qquad \text{as required.}
\end{aligned}$$

When $u = t-1$, the random variable $V_i^{(t)}(k_{t-1},\, t-1)$ is equal to the indicator variable

$I\{k_{t-1} \to^{(t-1)} i\}$, whence

$$\begin{aligned} \mathbb{E}\left(V_i^{(t)}(k_{t-1}, t-1)\right) &= \mathbb{P}\left(I\{k_{t-1} \to^{(t-1)} i\} = 1\right) \\ &= q_{k_{t-1}i}^{(t-1)} \qquad \text{as required.} \end{aligned} \qquad \square$$

Proposition 7 *$g'(0)$ is given by the following expression:*

$$g'(0) = -\sum_{k_0 : N_{k_0}^{(0)}=1} \sum_{k_1=1}^{N} \cdots \sum_{k_{t-1}=1}^{N} q_{k_0 k_1}^{(0)} q_{k_1 k_2}^{(1)} \cdots q_{k_{t-1} i}^{(t-1)} \,. \tag{24}$$

Proof Immediate from (21), (22) and (23). $\square$

This gives the coefficient β from (15).

At this stage it is useful to introduce some matrix notation. Let the matrix $Q^{(w)}$ have components $Q^{(w)}[u, v] = q_{uv}^{(w)}$ for $u, v = 1, \ldots N$. The expressions $Q^{(w)}[u, v]$ and $q_{uv}^{(w)}$ will be used interchangeably throughout the rest of the chapter. The notation $Q^{(0)}Q^{(1)} \ldots Q^{(w)}[u, v]$ is understood to represent the (u, v)–th element of the matrix $Q^{(0)}Q^{(1)} \ldots Q^{(w)}$. Further, recall that the vector $\boldsymbol{y^{(0)}}$ denotes the spatial distribution of the population at time $t = 0$: i.e. the ith component of $\boldsymbol{y^{(0)}}$ is

$$y_i^{(0)} = \begin{cases} 1 & \text{if site } i \text{ is occupied at time } t = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The coefficient β in (15) is then given by

$$g'(0) = -\sum_{k_0 : N_{k_0}^{(0)}=1} Q^{(0)}Q^{(1)} \ldots Q^{(t-1)}[k_0, i] = -\boldsymbol{y^{(0)\prime}} Q^{(0)}Q^{(1)} \ldots Q^{(t-1)}[i], \tag{25}$$

where $\boldsymbol{y^{(0)\prime}}$ is used to indicate the vector transpose of $\boldsymbol{y^{(0)}}$.

Calculation of $g''(0)$

Let $w_{k_0}(x) = \mathbb{E}\left((1-x)^{V_i^{(t)}(k_0, 0)}\right)$, and let w'_{k_0} be the first derivative of w_{k_0}. The expression for $g'(x)$ in (20) may be written

$$g'(x) = \sum_{k_0 : N_{k_0}^{(0)}=1} \frac{w'_{k_0}(x)}{w_{k_0}(x)} \,,$$

so that

$$g''(0) = \sum_{k_0 : N_{k_0}^{(0)}=1} \left(\frac{w_{k_0}(0)\, w''_{k_0}(0) - \left(w'_{k_0}(0)\right)^2}{\left(w_{k_0}(0)\right)^2} \right) .$$

Since $w_{k_0}(0) = 1$, this gives

$$\begin{aligned} g''(0) &= \sum_{k_0 : N_{k_0}^{(0)}=1} \left(w''_{k_0}(0) - \left(w'_{k_0}(0)\right)^2 \right) \\ &= \sum_{k_0 : N_{k_0}^{(0)}=1} \left\{ \mathbb{E}\left\{ V_i^{(t)}(k_0\,,\,0)^2 \right\} - \mathbb{E}\left\{ V_i^{(t)}(k_0\,,\,0) \right\} - \left\{ \mathbb{E} V_i^{(t)}(k_0\,,\,0) \right\}^2 \right\} \end{aligned} \quad (26)$$

using (19). From (21) and (25), the second and third terms of (26) are known:

$$\mathbb{E}\left\{ V_i^{(t)}(k_0\,,\,0) \right\} = Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}\,[k_0,\, i]\,; \quad (27)$$

$$\left\{ \mathbb{E} V_i^{(t)}(k_0\,,\,0) \right\}^2 = \left\{ Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}\,[k_0,\, i] \right\}^2 . \quad (28)$$

To calculate $g''(0)$ it remains only to find the first term of (26).

In the light of (28) it is convenient to define a new notation:

Definition: the matrix $M^{(t_1,\ldots,t_r)}$, with elements $M^{(t_1,\ldots,t_r)}[u,\, v] = m_{uv}^{(t_1,\ldots,t_r)}$, is defined for $r \geq 1$ as the element-wise square of the product $Q^{(t_1)} \ldots Q^{(t_r)}$. That is,

$$M^{(t_1,\ldots,t_r)}[u,\, v] = m_{uv}^{(t_1,\ldots,t_r)} = \left\{ Q^{(t_1)} \ldots Q^{(t_r)}\,[u,\, v] \right\}^2 . \quad (29)$$

Note that $m_{uv}^{(t)} = \left(q_{uv}^{(t)}\right)^2$ for any t.

Proposition 8 *For any sites k_u and i, and times u and t with $u < t-1$,*

$$\begin{aligned} \mathbb{E}\left\{ V_i^{(t)}(k_u, u)^2 \right\} &= M^{(u,\ldots,t-1)}[k_u,\, i] - M^{(u)} M^{(u+1,\ldots,t-1)}[k_u,\, i] \\ &\quad + \sum_{k_{u+1}} q_{k_u k_{u+1}}^{(u)} \mathbb{E}\left\{ V_i^{(t)}(k_{u+1}, u+1)^2 \right\} . \end{aligned} \quad (30)$$

When $u = t-1$,

$$\mathbb{E}\left\{ V_i^{(t)}(k_{t-1}, t-1)^2 \right\} = q_{k_{t-1} i}^{(t-1)} . \quad (31)$$

Proof

$$\mathbb{E}\left\{V_i^{(t)}(k_u,u)^2\right\} = \mathbb{E}\left\{\sum_{k_{u+1}} I\{k_u \to^{(u)} k_{u+1}\} V_i^{(t)}(k_{u+1},u+1)\right\}^2$$

$$= \mathbb{E}\left\{\sum_{k_{u+1}} I\{k_u \to^{(u)} k_{u+1}\} V_i^{(t)}(k_{u+1},u+1) \sum_{j_{u+1}} I\{k_u \to^{(u)} j_{u+1}\} V_i^{(t)}(j_{u+1},u+1)\right\}$$

$$= \sum_{k_{u+1}}\sum_{j_{u+1}} \mathbb{E}\left\{I\{k_u \to^{(u)} k_{u+1}\} V_i^{(t)}(k_{u+1},u+1)\right\} \mathbb{E}\left\{I\{k_u \to^{(u)} j_{u+1}\} V_i^{(t)}(j_{u+1},u+1)\right\}$$
$$- \sum_{k_{u+1}} \left(\mathbb{E}\left\{I\{k_u \to^{(u)} k_{u+1}\} V_i^{(t)}(k_{u+1},u+1)\right\}\right)^2$$
$$+ \sum_{k_{u+1}} \mathbb{E}\left\{I\{k_u \to^{(u)} k_{u+1}\}^2 V_i^{(t)}(k_{u+1},u+1)^2\right\}, \qquad (32)$$

where the first term follows from the independence of $I\{k_u \to^{(u)} k_{u+1}\} V_i^{(t)}(k_{u+1},u+1)$ and $I\{k_u \to^{(u)} j_{u+1}\} V_i^{(t)}(j_{u+1},u+1)$ for $k_{u+1} \neq j_{u+1}$, the second term subtracts the incorrect part of the first term in which $k_{u+1} = j_{u+1}$, and the third term is the correct formulation for the case $k_{u+1} = j_{u+1}$.

For clarity, let $I = I\{k_u \to^{(u)} k_{u+1}\}$ and $V = V_i^{(t)}(k_{u+1},u+1)$. Since I and V may be regarded as independent random variables,

$$\mathbb{E}\{I\,V\} = \mathbb{E}I\,\mathbb{E}V = q^{(u)}_{k_u k_{u+1}} \mathbb{E}V, \qquad \mathbb{E}\{I^2\,V^2\} = \mathbb{E}I^2\,\mathbb{E}V^2 = q^{(u)}_{k_u k_{u+1}} \mathbb{E}V^2\,.$$

Continuing from (32) gives

$$\mathbb{E}\left\{V_i^{(t)}(k_u,u)^2\right\} = \left\{\sum_{k_{u+1}} \mathbb{E}(IV)\right\}^2 - \sum_{k_{u+1}} \left\{\mathbb{E}(IV)\right\}^2 + \sum_{k_{u+1}} \mathbb{E}(IV^2)$$

$$= \left\{\sum_{k_{u+1}} q^{(u)}_{k_u k_{u+1}} \mathbb{E}(V)\right\}^2 - \sum_{k_{u\,|\,1}} \left\{q^{(u)}_{k_u k_{u+1}} \mathbb{E}V\right\}^2 + \sum_{k_{u+1}} q^{(u)}_{k_u k_{u+1}} \mathbb{E}V^2\,.$$

But immediate from (22) and (23) is the result:

$$\mathbb{E}V = \mathbb{E}V_i^{(t)}(k_{u+1},u+1) = Q^{(u+1)} \ldots Q^{(t-1)}\,[k_{u+1},\, i]\,,$$

which gives finally

$$\begin{aligned}\mathbb{E}\left\{V_i^{(t)}(k_u,u)^2\right\} &= \left\{Q^{(u)}\ldots Q^{(t-1)}\,[k_u,\,i]\right\}^2 - \sum_{k_{u+1}}\left\{m^{(u)}_{k_uk_{u+1}}m^{(u+1,\ldots,t-1)}_{k_{u+1}i}\right\} \\ &\qquad + \sum_{k_{u+1}} q^{(u)}_{k_uk_{u+1}}\mathbb{E}\left\{V_i^{(t)}(k_{u+1},u+1)^2\right\} \\ &= M^{(u,\ldots,t-1)}\,[k_u,\,i] \;-\; M^{(u)}M^{(u+1,\ldots,t-1)}\,[k_u,\,i] \\ &\qquad + \sum_{k_{u+1}} q^{(u)}_{k_uk_{u+1}}\mathbb{E}\left\{V_i^{(t)}(k_{u+1},u+1)^2\right\},\end{aligned}$$

as required.

When $u = t-1$,

$$\begin{aligned}\mathbb{E}\left\{V_i^{(t)}(k_{t-1},t-1)^2\right\} &= 1^2\,\mathbb{P}\left(I\{k_{t-1}\to^{(t-1)} i\}=1\right) + 0^2\,\mathbb{P}\left(I\{k_{t-1}\to^{(t-1)} i\}=0\right) \\ &= q^{(t-1)}_{k_{t-1}i} \qquad \text{as stated.} \qquad \square\end{aligned}$$

Proposition 9 *For any sites k_0 and i and time t,*

$$\begin{aligned}\mathbb{E}\left\{V_i^{(t)}(k_0,0)^2\right\} &= M^{(0,\ldots,t-1)}\,[k_0,\,i] \;-\; M^{(0)}M^{(1,\ldots,t-1)}\,[k_0,\,i] \\ &+ Q^{(0)}M^{(1,\ldots,t-1)}\,[k_0,\,i] \;-\; Q^{(0)}M^{(1)}M^{(2,\ldots,t-1)}\,[k_0,\,i] \\ &+ Q^{(0)}Q^{(1)}M^{(2,\ldots,t-1)}\,[k_0,\,i] \;-\; Q^{(0)}Q^{(1)}M^{(2)}M^{(3,\ldots,t-1)}\,[k_0,\,i] \\ &\;\;\vdots \\ &+ Q^{(0)}\ldots Q^{(t-3)}M^{(t-2,t-1)}\,[k_0,\,i] \;-\; Q^{(0)}\ldots Q^{(t-3)}M^{(t-2)}M^{(t-1)}\,[k_0,\,i] \\ &+ Q^{(0)}\ldots Q^{(t-1)}\,[k_0,\,i]\,. \qquad (33)\end{aligned}$$

Proof Follows from (30) and (31). $\square$

The final form of $g''(0)$ may now be computed.

Proposition 10 *For given site i and time t, $g''(0)$ is given by*

$$\begin{aligned}g''(0) = \sum_{k_0:N^{(0)}_{k_0}=1} &\left\{Q^{(0)}M^{(1,\ldots,t-1)} + \ldots + Q^{(0)}\ldots Q^{(t-3)}M^{(t-2,t-1)}\right. \\ &- Q^{(0)}M^{(1)}M^{(2,\ldots,t-1)} - \ldots - Q^{(0)}\ldots Q^{(t-3)}M^{(t-2)}M^{(t-1)} \\ &\left.- M^{(0)}M^{(1,\ldots,t-1)}\right\}\,[k_0,\,i]\,. \qquad (34)\end{aligned}$$

Proof From (26), (27), (28) and (29),

$$g''(0) = \sum_{k_0 : N_{k_0}^{(0)} = 1} \left\{ \mathbb{E}\left\{ V_i^{(t)}(k_0\,,\,0)^2 \right\} - Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}\,[k_0,\,i] - M^{(0,\ldots,t-1)}\,[k_0,\,i] \right\}. \tag{35}$$

For each value of k_0, the first term in (35) is given by equation (33), while the second and third terms of (35) cancel out respectively the last and first terms of (33). The remainder is as given in (34). □

Expression (34) becomes $\sum_{k_0 : N_{k_0}^{(0)} = 1} -M^{(0)}\,[k_0,\,i]$ when $t = 1$, and $\sum_{k_0 : N_{k_0}^{(0)} = 1} -M^{(0)} M^{(1)}\,[k_0,\,i]$ when $t = 2$.

Final form of Taylor approximation to $g(x) = \log s_i^{(t)}(x)$

The final form of the Taylor approximation to $g(x) = \log s_i^{(t)}(x)$ is obtained from (16), (25) and (34) as

$$\begin{aligned}
\log s_i^{(t)}(x) \;=\; & x \left(-\boldsymbol{y}^{(0)\prime}\; Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}\,[i] \right) \\
& + \frac{x^2}{2} \left(\boldsymbol{y}^{(0)\prime}\; \left\{ Q^{(0)} M^{(1,\ldots,t-1)} + \ldots + Q^{(0)} \ldots Q^{(t-3)} M^{(t-2,\,t-1)} \right. \right. \\
& - Q^{(0)} M^{(1)} M^{(2,\ldots,t-1)} - \ldots - Q^{(0)} \ldots Q^{(t-3)} M^{(t-2)} M^{(t-1)} \\
& \left. \left. - M^{(0)} M^{(1,\ldots,t-1)} \right\} [i] \right).
\end{aligned} \tag{36}$$

The exponential of the right-hand side of (36) will be referred to as the *second-order* or *quadratic exponential* approximation to $s_i^{(t)}(x)$, and the exponential of the first-order term only will be referred to as the *first order* or *linear exponential* approximation.

3.3 Approximate roots of the colony-scale polynomial

Linear exponential approximation

Recall that $g(x) = \log s_i^{(t)}(x)$ for given site i and time t. The first-order version of (14) is

$$\alpha + \beta x = \log(1 - p_{ih} p_i^{(t)}),$$

and by (16), $\alpha = g(0) = 0$ and $\beta = g'(0)$. The approximate root of the colony-scale polynomial obtained from the linear exponential approximation is therefore

$$q_{ih}^{(t)} \simeq x = \frac{\log(1 - p_{ih} p_i^{(t)})}{g'(0)}. \tag{37}$$

By (25), this becomes

$$q_{ih}^{(t)} \simeq x = -\left(\frac{\log(1 - p_{ih} p_i^{(t)})}{\boldsymbol{y}^{(0)\prime} Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}[\,i\,]} \right). \tag{38}$$

Since $1 - p_{ih} p_i^{(t)} \leq 1$ for all i, h and t, the numerator of (38) is ≤ 0 and the approximate root x is always non-negative.

If the denominator of (37) is zero, then by (24) all possible products $q_{k_0 k_1}^{(0)} \ldots q_{k_{t-1} i}^{(t-1)}$ must be zero. In that case there is zero probability that site i is colonized along any colonization path from the occupied sites at time 0, and consequently $p_i^{(t)} = 0$. The numerator of (37) is also zero in this instance. Since site i is unoccupied with certainty at time t, there is no purpose in calculating $q_{ih}^{(t)}$ for any site h; thus the non-existence of solutions does not pose a problem.

Quadratic exponential approximation

From (14) and (16), the approximate roots to the colony-scale polynomial $s_i^{(t)}(x) - 1 + p_{ih} p_i^{(t)}$ are as follows:

$$x = \frac{-g'(0) \pm \sqrt{g'(0)^2 + 2g''(0) \log(1 - p_{ih} p_i^{(t)})}}{g''(0)}. \tag{39}$$

Since $1 - p_{ih} p_i^{(t)} \leq 1$ for all i, h and t, it follows that $\log(1 - p_{ih} p_i^{(t)}) \leq 0 \ \forall i, h, t$. Also, $-g'(0) \geq 0$ from (25). In order to determine whether it is the positive or negative root of

(39) that is required, consider the two cases $g''(0) < 0$ and $g''(0) > 0$.

(i) $g''(0) < 0$: the second term under the square root in (39) is non-negative, so

$$\sqrt{g'(0)^2 + 2g''(0)\log(1 - p_{ih}p_i^{(t)})} \geq |g'(0)| = -g'(0).$$

It follows that the numerator of (39) is non-positive only if the negative square root is taken, and since the denominator of (39) is also non-positive this gives the sole non-negative solution to (39). Note also that if $g''(0) < 0$, the roots of (39) are always real.

(ii) $g''(0) > 0$: then

$$\sqrt{g'(0)^2 + 2g''(0)\log(1 - p_{ih}p_i^{(t)})} \leq |g'(0)| = -g'(0),$$

so that if real roots to (39) exist, they are both non-negative. The general shape of the quadratic exponential approximation to $s_i^{(t)}(x)$ when $g''(0) > 0$ is given in Figure 7. No scale is shown, since it will vary from case to case, but the general pattern is the same for all cases. The single turning-point in the curve $y = \exp(g'(0)x + g''(0)x^2/2)$, which occurs at $x = -g'(0)/g''(0)$, is also marked on the figure. Now since $s_i^{(t)}(x) = \mathbb{E}(1-x)^{N_i^{(t)}}$, it is known from (17) that $s_i^{(t)}(x)$ decreases monotonically on the interval $[0, 1]$. The turning point in the quadratic exponential approximation to $s_i^{(t)}(x)$ therefore represents a departure from the true curve if it occurs in the interval $[0, 1]$, and consequently it is clear that the required solution to the polynomial (39) is the root that occurs before the turning point — that is, the negative root.

For all values of $g''(0)$, therefore, the negative root to (39) is used as the approximate value of $q_{ih}^{(t)}$. Note that if $g''(0) = 0$, the quadratic exponential approximation reduces to the linear exponential approximation, and the expression (38) should be used in place of (39).

It is not clear that the equation (39) will always have real roots when $g''(0) > 0$, nor that the turning point at $x = -\left(\frac{g'(0)}{g''(0)}\right)$ will always occur outwith the interval [0, 1]. In practice, if the turning point occurs within the interval [0, 1], the linear exponential approximation $s_i^{(t)}(x) \approx \exp(g'(0)x)$ will be used instead of the quadratic exponential approximation, and likewise if there are no real roots to (39). The linear exponential approximation is always monotonically decreasing on the whole interval [0, 1], and a real root for x always exists.

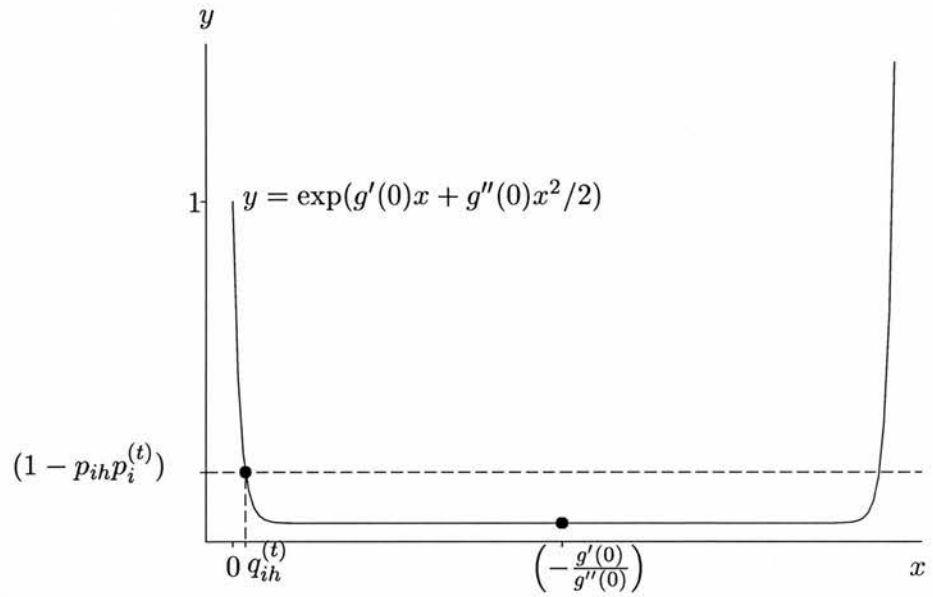


Figure 7: The general shape of the quadratic exponential approximation $y = \exp(g'(0)x + g''(0)x^2/2)$ to the curve $y = s_i^{(t)}(x)$ when $g''(0) > 0$ for x on an unspecified scale. The turning point at $x = -\left(\frac{g'(0)}{g''(0)}\right)$ is marked with a black dot, and the solution for $x = q_{ih}^{(t)}$ is marked at the intersection of the curve with the dotted line $y = 1 - p_{ih}p_i^{(t)}$.

Theoretical conditions for good behaviour of the quadratic exponential approximation have not been found, although simulation results suggest that the use of the linear exponential approximation is unlikely to cause serious problems. The behaviour of the quadratic and linear exponential approximations is examined more closely in the following section.

Calculation of $p_i^{(t)}$

The approximate roots to the colony-scale polynomial obtained from both the linear exponential approximation (37) and the quadratic exponential approximation (39) involve the quantity $p_i^{(t)} = \mathbb{P}(N_i^{(t)} > 0)$. Now

$$s_i^{(t)}(x) \;=\; \mathbb{E}\,(1-x)^{N_i^{(t)}} \;=\; \mathbb{P}(N_i^{(t)} = 0) + (1-x)\mathbb{P}(N_i^{(t)} = 1) + \ldots$$

Putting $x = 1$ gives

$$s_i^{(t)}(1) = \mathbb{P}(N_i^{(t)} = 0) = 1 - p_i^{(t)}\,,$$

and thus $p_i^{(t)} = 1 - s_i^{(t)}(1)$ may be calculated using either the linear or quadratic exponential approximations to $s_i^{(t)}(1)$:

$$p_i^{(t)} \simeq 1 - \exp\left(g'(0)\right) \qquad \text{or} \qquad p_i^{(t)} \simeq 1 - \exp\left(g'(0) + \frac{1}{2}g''(0)\right)\,. \tag{40}$$

It should be noted that when $t = T$, the probability $p_i^{(T)}$ provides the site-wise likelihood $L\left(y_h^{(T)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right)$. According to whether $y_h^{(T)} = 0$ or $y_h^{(T)} = 1$, the likelihood $L\left(y_h^{(T)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right)$ is simply $1 - p_i^{(T)}$ or $p_i^{(T)}$ respectively. The modified branching process formulation has therefore enabled the calculation of the site-wise occupation probabilities at time T, and the branching process recursion (1) has been utilized implicitly through the expression (13) and the observations on page 109. It is not clear how the occupation probabilities $\{p_i^{(T)}\}$ could be calculated without the assistance of the modified branching process.

4 Analysis of the exponential approximations to $s_i^{(t)}(x)$

Before proceeding to use the exponential approximations to $s_i^{(t)}(x)$ in likelihood calculations, it is necessary to evaluate the accuracy and reliability of the approximations. This section presents some exact results for the special cases of $t = 1$ and $t = 2$, and a series of simulation trials to demonstrate the performance of the approximate method.

4.1 Exact analysis for $t = 1, 2$

When $t = 1$ or $t = 2$, the expression (11) for $s_i^{(t)}(x)$ reduces to the simple form

$$s_i^{(t)}(x) = \begin{cases} \prod_{k_0 : N_{k_0}^{(0)} = 1} \left\{1 - q_{k_0 i}^{(0)} x\right\} & (t = 1)\,, \\ \prod_{k_0 : N_{k_0}^{(0)} = 1} \prod_{k_1 = 1}^{N} \left\{1 - q_{k_0 k_1}^{(0)} q_{k_1 i}^{(1)} x\right\} & (t = 2)\,, \end{cases}$$

which may be written generically as

$$\prod_{k=1}^{M} (1 - \alpha_k x) \tag{41}$$

for some integer M and real numbers $\alpha_1, \ldots, \alpha_M$, where $0 \leq \alpha_k \leq 1$ for all k. The linear and quadratic exponential approximations to $s_i^{(t)}(x)$ may be obtained from equation (36) as respectively

$$\exp\left(-S_1\, x\right) \quad \text{and} \quad \exp\left(-S_1\, x - S_2 \frac{x^2}{2}\right),$$

where $S_1 = \sum_{k=1}^{M} \alpha_k$ and $S_2 = \sum_{k=1}^{M} \alpha_k^2$. The values of the individual components α_k are known for $t = 1$, since they are given by the colonization probabilities $\{p_{ih} : i,\ h = 1, \ldots, N\}$. However, for $t = 2$ the components α_k are not known, as the coefficients S_1

and S_2 may be obtained from the corresponding coefficients for $t = 1$ without calculating the products $q^{(0)}_{k_0 k_1} q^{(1)}_{k_1 i}$ individually: viz. $S_1 = \sum_{k_0} \sum_{k_1} q^{(0)}_{k_0 k_1} q^{(1)}_{k_1 i} = \sum_{k_1} q^{(1)}_{k_1 i} \left(\sum_{k_0} q^{(0)}_{k_0 k_1} \right)$ and similarly for S_2.

It is required to derive bounds for the linear and quadratic exponential approximations to $\prod_{k=1}^{M}(1 - \alpha_k x)$, where the components $\{\alpha_k\}$ are unknown. Since the solution x to the colony-scale polynomial $\prod_{k=1}^{M}(1 - \alpha_k x) = (1 - p_{ih} p_i^{(t)})$ might occur at any $x \in [0, 1]$, it is useful to obtain bounds that apply throughout the range $[0, 1]$ rather than global extrema with respect to x. It is also reasonable to produce bounds subject to fixed S_1 and S_2, since most of the structure of the problem would be lost by extremizing over these quantities. For example, putting $\alpha_1 = 1$, $\alpha_2 = \ldots = \alpha_M = 0$ gives $S_1 = 1$, and the first-order error at $x = 1$ is $\exp(-S_1 x) - \prod_k (1 - \alpha_k x) = \exp(-1) = 0.368$, which is unacceptably high. These values, however, are never likely to occur in practice, so this bound is not useful.

The problem is therefore formulated as follows. The linear and quadratic error functions are defined as

$$\begin{aligned} \epsilon_1(x) &= \exp\{-S_1 x\} - \prod_{k=1}^{M}(1 - \alpha_k x) && \textit{linear error,} \\ \epsilon_2(x) &= \exp\left\{-S_1 x - S_2 \frac{x^2}{2}\right\} - \prod_{k=1}^{M}(1 - \alpha_k x) && \textit{quadratic error.} \end{aligned}$$

It is shown in Appendix A that these functions are always non-negative. The linear approximation bounds are obtained by extremizing $\epsilon_1(x)$ over $\alpha_1, \ldots, \alpha_M$, subject to the following constraints: (a) $\sum_{k=1}^{M} \alpha_k = S_1$, and (c) $0 \leq \alpha_k \leq 1 \quad \forall k$. The quadratic approximation bounds are obtained by extremizing $\epsilon_2(x)$ over $\alpha_1, \ldots, \alpha_M$, subject to the constraints (a) $\sum_{k=1}^{M} \alpha_k = S_1$, (b) $\sum_{k=1}^{M} \alpha_k^2 = S_2$, and (c) $0 \leq \alpha_k \leq 1 \ \forall k$. The advantage of this formulation is that it allows the maximum and minimum errors to be determined quickly for a range of plausible $\boldsymbol{\alpha}$, simply by selecting credible values for S_1 and S_2. This is more informative than the calculation of a single error under one particular value of $\boldsymbol{\alpha}$.

A detailed derivation of the linear and quadratic approximation bounds is given in Appendix A, where the following results are obtained.

Linear approximation bounds

The maximum over $\boldsymbol{\alpha} \in \mathbb{R}^M$ of the linear error $\epsilon_1(x)$, subject to the constraints (a) and (c), occurs at $\boldsymbol{\alpha} = (\underbrace{1, \ldots, 1}_{\lfloor S_1 \rfloor}, S_1 - \lfloor S_1 \rfloor, 0, \ldots, 0)$ and is given by

$$\max_{\boldsymbol{\alpha}} (\epsilon_1(x)) = \exp\{-S_1 x\} - \left(1 - x\right)^{\lfloor S_1 \rfloor} \left(1 - (S_1 - \lfloor S_1 \rfloor)x\right) . \tag{42}$$

The minimum over $\boldsymbol{\alpha} \in \mathbb{R}^M$ of $\epsilon_1(x)$, subject to (a) and (c), occurs at $\alpha_1 = \ldots = \alpha_M = S_1/M$ and is given by

$$\min_{\boldsymbol{\alpha}} (\epsilon_1(x)) = \exp\{-S_1 x\} - \left(1 - \frac{S_1}{M} x\right)^M . \tag{43}$$

Equation (43) actually provides a new approximation to $s_i^{(t)}(x)$ with tighter bounds than the linear exponential approximation. Since $\epsilon_1(x) \geq 0$ everywhere, (43) gives

$$0 \;\leq\; \exp\{-S_1 x\} - \left(1 - \frac{S_1}{M} x\right)^M \;\leq\; \exp\{-S_1 x\} - \prod_{k=1}^{M} (1 - \alpha_k x) ,$$

thus

$$\prod_{k=1}^{M} (1 - \alpha_k x) \;\leq\; \left(1 - \frac{S_1}{M} x\right)^M \;\leq\; \exp\{-S_1 x\} .$$

Approximation of $\prod_{k=1}^{M}(1 - \alpha_k x)$ by $\left(1 - \frac{S_1}{M} x\right)^M$ is therefore closer than the linear exponential approximation, and moreover provides an approximate colony-scale polynomial $\left(1 - \frac{S_1}{M} x\right)^M = 1 - p_{ih} p_i^{(t)}$ that is easily solved for x:

$$x = \frac{M}{S_1} \left(1 - \left(1 - p_{ih} p_i^{(t)}\right)^{1/M}\right) .$$

The positive $(1/M)$th root is taken if M is even; this yields the smallest positive solution for x, which is necessarily the correct root.

The new approximation will be used in place of the linear exponential approximation for $t = 1$ and $t = 2$, although in practice the difference between the two is negligible for the large M generally encountered with $t > 1$. Occasionally the revised approximation has a detectable effect for $t = 1$, under a combination of large S_1 and small M.

Quadratic approximation bounds

The maximum over $\boldsymbol{\alpha} \in \mathbb{R}^M$ of the quadratic error $\epsilon_2(x)$, subject to the constraints (a), (b) and (c), occurs at

$$\boldsymbol{\alpha} = (\underbrace{c, \ldots, c}_{n-1}, d, \underbrace{1, \ldots, 1}_{M-n}),$$

where

$$c = \frac{S_1 - M + n}{n} - \frac{1}{n}\sqrt{\frac{n(S_2 - M + n) - (S_1 - M + n)^2}{n-1}},$$

$$d = \frac{S_1 - M + n}{n} + \frac{n-1}{n}\sqrt{\frac{n(S_2 - M + n) - (S_1 - M + n)^2}{n-1}},$$

and

$$n = \left\lceil \frac{(M - S_1)^2}{(S_2 - 2S_1 + M)} \right\rceil .$$

The maximum quadratic error is therefore given by

$$\max \boldsymbol{\alpha} \left(\epsilon_2(x) \right) = \exp\left\{ -S_1 x - S_2 \frac{x^2}{2} \right\} - \left(1 - cx\right)^{n-1} \left(1 - dx\right) \left(1 - x\right)^{M-n} . \quad (44)$$

The minimum over $\boldsymbol{\alpha} \in \mathbb{R}^M$ of the quadratic error $\epsilon_2(x)$, subject to the constraints (a), (b) and (c), occurs at

$$\boldsymbol{\alpha} = (a, \underbrace{b, \ldots, b}_{n-1}, \underbrace{0, \ldots, 0}_{M-n}),$$

where

$$a = \frac{S_1}{n} - \frac{n-1}{n}\sqrt{\frac{nS_2 - S_1^2}{n-1}},$$

$$b = \frac{S_1}{n} + \frac{1}{n}\sqrt{\frac{nS_2 - S_1^2}{n-1}},$$

and

$$n = \left\lceil \frac{S_1^2}{S_2} \right\rceil .$$

The minimum quadratic error is therefore given by

$$\min \boldsymbol{\alpha} \left(\epsilon_2(x) \right) = \exp\left\{ -S_1 x - S_2 \frac{x^2}{2} \right\} - \left(1 - ax\right) \left(1 - bx\right)^{n-1} . \quad (45)$$

As with the linear case, the result (45) suggests an alternative to the quadratic approximation to $s_i^{(t)}(x)$. However, this time there is no simple solution to the resulting colony-scale

polynomial $(1-ax)(1-bx)^{n-1} = 1 - p_{ih}p_i^{(t)}$, so the revised approximation is not of practical benefit.

Linear error subject to quadratic bounds

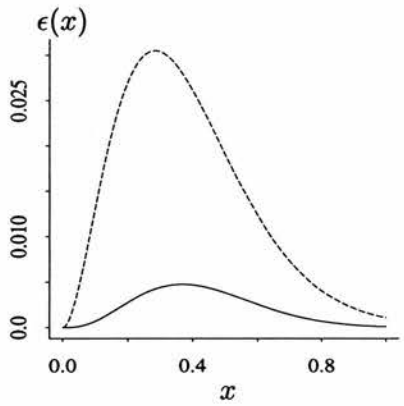
Bounds for the linear error function $\epsilon_1(x)$ subject to all three constraints (a), (b) and (c) may also be derived, even though the constraint (b) does not enter into the error function. The vector $\boldsymbol{\alpha} = (\underbrace{c, \ldots, c}_{n-1}, d, \underbrace{1, \ldots, 1}_{M-n})$, with c, d and n defined as in the previous section, simply minimizes the expression $\prod_{k=1}^{M}(1-\alpha_k x)$ over $\boldsymbol{\alpha}$, subject to the constraints (a), (b) and (c). The same vector $\boldsymbol{\alpha}$ must therefore maximize the linear exponential error, $\exp(-S_1 x) - \prod_{k=1}^{M}(1-\alpha_k x)$, subject to the constraints (a), (b) and (c). The maximum linear error subject to these constraints is consequently

$$\exp\{-S_1 x\} - \Big(1-cx\Big)^{n-1}\Big(1-dx\Big)\Big(1-x\Big)^{M-n}. \tag{46}$$

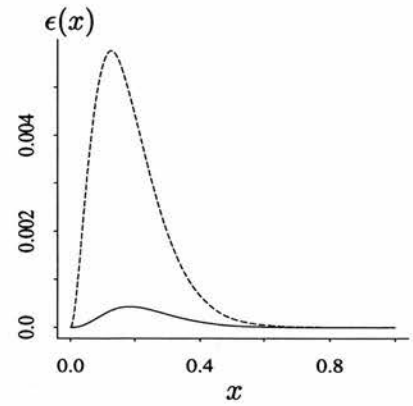
Similarly, the minimum linear error subject to constraints (a), (b) and (c) occurs at $\boldsymbol{\alpha} = (a, \underbrace{b, \ldots, b}_{n-1}, \underbrace{0, \ldots, 0}_{M-n})$, with a, b and the new value of n as defined in the previous section. The minimum linear error subject to these constraints is

$$\exp\{-S_1 x\} - \Big(1-ax\Big)\Big(1-bx\Big)^{n-1}. \tag{47}$$

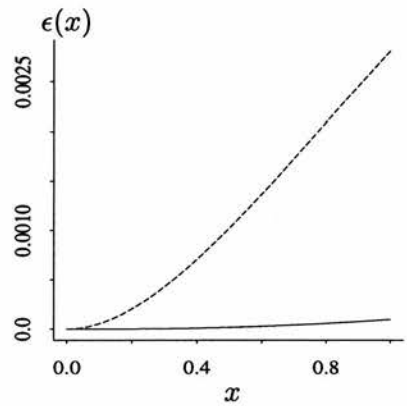
Figure 8 shows the linear and quadratic error bounds, each subject to all three quadratic constraints (a), (b) and (c), for four simulated values of S_1, S_2 and M. The quadratic errors are noticeably smaller than the linear errors, but even the maximum linear errors subject to the given constraints are sufficiently small to be regarded as negligible in the present context. Repeated simulations indicate that the errors remain small throughout the sphere of practical application.



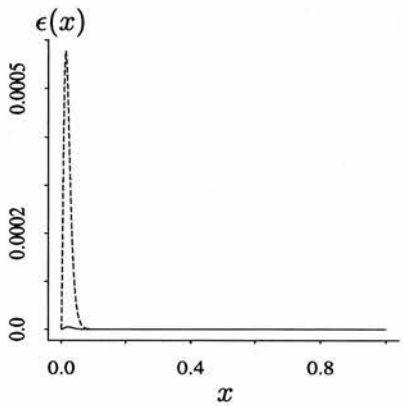
(a) $M = 10,\ S_1 = 6.84,\ S_2 = 5.01.$



(b) $M = 100,\ S_1 = 16.1,\ S_2 = 5.30.$



(c) $M = 100,\ S_1 = 0.788,\ S_2 = 0.012.$



(d) $M = 1000,\ S_1 = 145,\ S_2 = 44.8.$

Figure 8: Error functions for selected M, S_1 and S_2. The solid lines show the maximum quadratic error subject to constraints (a), (b) and (c), while the dashed lines show the maximum linear error, subject to the same (quadratic) constraints.

Remark The approximations to the functions $\left\{s_i^{(t)}\right\}$ are used for two purposes: firstly for the evaluation of the function at some point (for example, at the point $x = 1$ when calculating $p_i^{(t)} = 1 - s_i^{(t)}(1)$), and secondly to obtain solutions for x to the equation $s_i^{(t)}(x) = 1 - p_{ih}p_i^{(t)}$. In general it is more important to obtain accuracy in the second case than in the first, due to the fact that a small error in $s_i^{(t)}(x)$ can correspond to a substantially larger error in the value of x at which $s_i^{(t)}$ attains a particular value.

This point is demonstrated in Figure 9. In the simulation trials to follow the difference between the linear and quadratic exponential approximations is usually slight, but the extra accuracy afforded by the quadratic approximations might nonetheless be useful when solving the colony-scale polynomials.

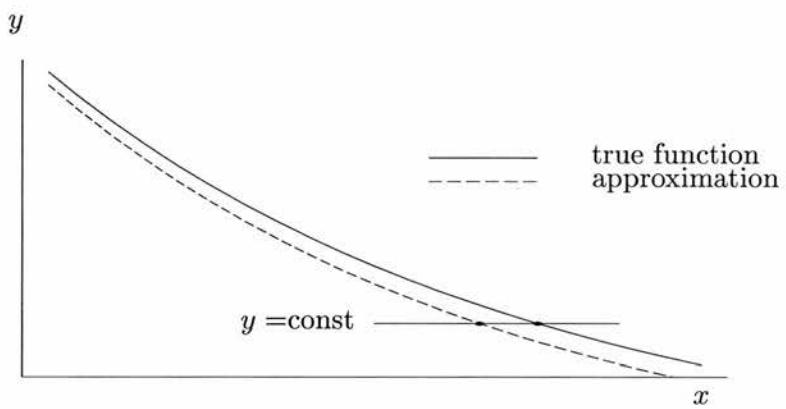


Figure 9: Illustration of the importance of a close approximation to $y = s_i^{(t)}(x)$ when solving the equation $s_i^{(t)}(x) =$const for x. A small error in the y-direction can be magnified considerably in the x-direction.

4.2 Simulation results

Illustrations of the exponential approximations to $s_i^{(t)}(x)$ are now provided using spatial distributions simulated from the colonization model. Four sets of parameters a, b and p_0 were selected, and trials were carried out on square grids of size $N = 100$ and $N = 400$ sites. For each trial, habitat suitability scores were randomly generated for every site, subject to arbitrarily-chosen trends over the grid. The colonization probabilities $\left\{p_{ih} = q_{ih}^{(0)}\right\}$ for each pair (i, h) of sites were thereby calculated, using the expression $p_{ih} = p_0 \exp(-a\,\delta_{ih} - b\,\varsigma_h)$ where δ_{ih} is the distance between sites i and h, and ς_h is the transformed habitat suitability score of site h. An initial distribution was generated, and the exact functions $\left\{s_i^{(t)}\right\}$, together with their approximations, were calculated for $t = 1, \ldots, 6$. Calculation of the

exact functions $\left\{s_i^{(t)}\right\}$ is extremely time-consuming and is not feasible to perform except in the context of simulation studies such as these, where only a few results need be collected.

For each of the four trials, the approximations to the functions $\left\{s_i^{(t)}\right\}$ were calculated using both the linear and the quadratic exponential approximations. The quantities $\left\{p_i^{(t)}\right\}$ were calculated using the same level of approximation as the functions $\left\{s_i^{(t)}\right\}$ — that is, linear or quadratic. The differences in the linear and quadratic approximations at time $t = 1$ propagate forwards to times $t = 2$ and beyond, explaining the small discrepancies between $S_1 = \sum_{k_0} \sum_{k_1} q_{k_0 k_1}^{(0)} q_{k_1 i}^{(1)}$ at time $t = 2$ under the two different approximation levels.

Simulation 1: $\boldsymbol{a = 1.0, \quad b = 2.0, \quad p_0 = 0.40, \quad N = 100.}$

Figure 10 shows a typical prediction over 6 time periods using these parameters on a 10×10 grid of site squares, starting from an initial distribution with occupation in all squares at time 0. Occupied sites are shaded dark, unoccupied sites light, and the numbering of the sites (not shown) increases along the rows starting at the bottom left-hand corner. The population tends to diminish over time. Table 1 shows observed and expected frequencies of occupation over 1000 trials for a single site (site 28) in the distribution. Predictions such as that in Figure 10 were obtained 1000 times, and the occupation status of site 28 at each of times $t = 1, \ldots, 6$ was recorded. The observed frequency of occupation is simply the number of the 1000 trials in which site 28 was occupied. The first-order expected result for time t is $1000 \times p_{28}^{(t)}$, where $p_{28}^{(t)} = 1 - s_{28}^{(t)}(1)$ and all functions $\left\{s_i^{(t)}\right\}$ were approximated using the linear exponential approximation. The second-order expected result was obtained in the same fashion when all functions $\left\{s_i^{(t)}\right\}$ were approximated using the quadratic exponential approximations. The final two columns give the difference between observed and expected occupation frequencies.

The expected frequencies of occupation using both the linear and quadratic exponential approximations to the functions $\left\{s_i^{(t)}\right\}$ are close to the observed frequencies, and none of the observations O_t would be unusual as realizations from either of the binomial distributions $\text{Bin}(1000, E_{1t})$ or $\text{Bin}(1000, E_{2t})$. An expected frequency E_{kt} is marked with a star if the observation O_t is significant at the 5% level for the $\text{Bin}(1000, E_{kt})$ distribution under a two-tailed test. Note that the results are correlated over the time period, since the prediction at time t follows from the prediction at time $t - 1$ for all t. A slight tendency in the results of Table 1 to over-estimate frequency of occupation for site 28 at times $t = 5$ and 6, for example, is not indicative of any sort of drift in this direction, and indeed repetitions of the experiment produce both over- and under-estimates of occupation.

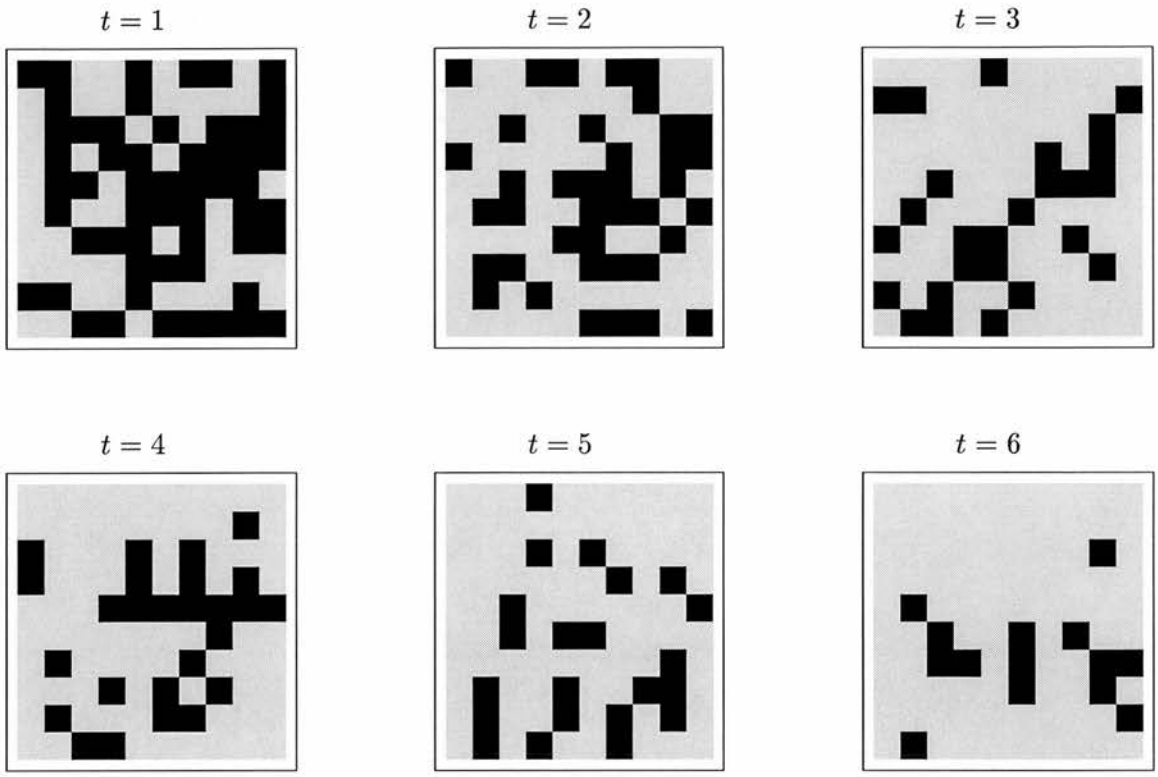
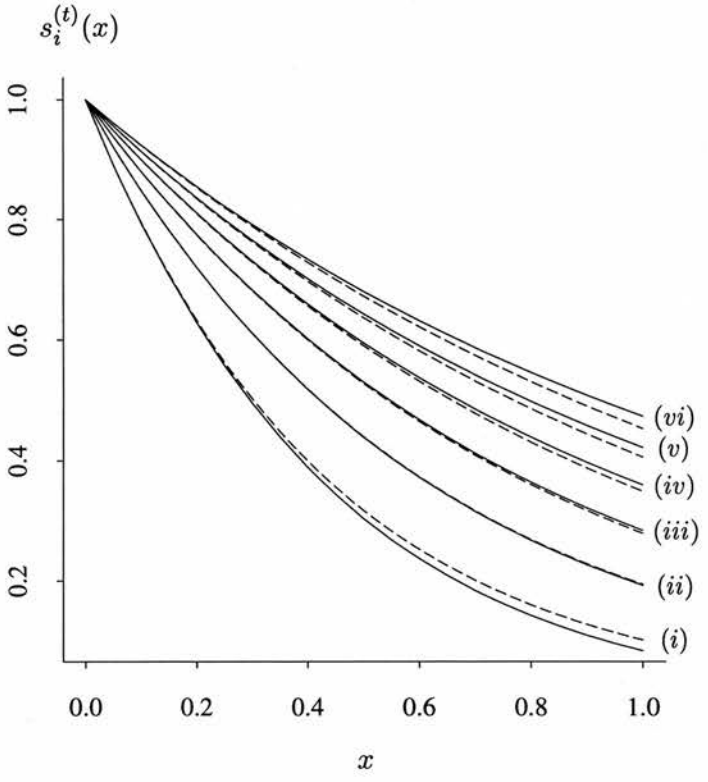


Figure 10: Predicted distributions of occupation using $a = 1.0$, $b = 2.0$, $p_0 = 0.40$. All sites were occupied in the initial distribution at time $t = 0$.

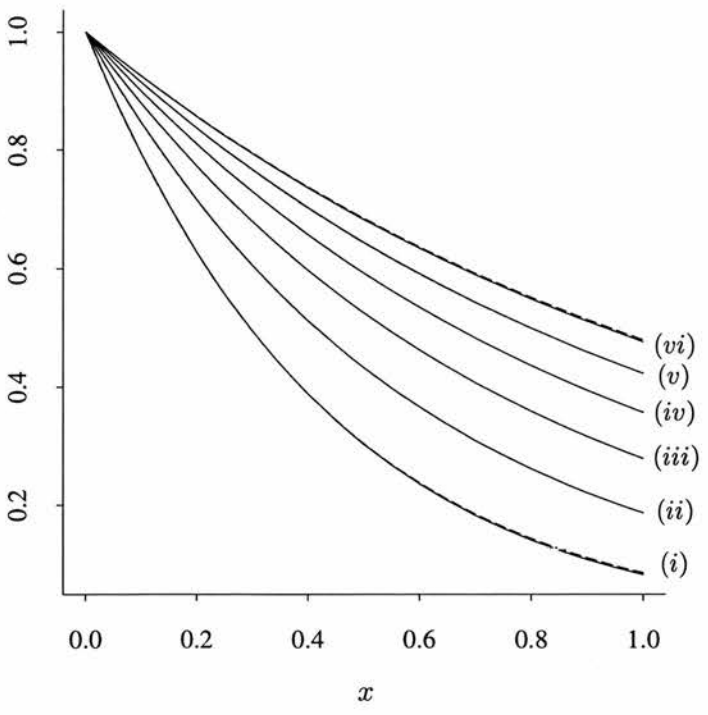
	Number of occupations out of 1000 trials			Prediction error	
t	Obtained O_t	Expected: 1st-order E_{1t}	Expected: 2nd-order E_{2t}	$E_{1t} - O_t$	$E_{2t} - O_t$
1	925	899*	914	-26	-11
2	819	807	813	-12	-6
3	738	721	721	-17	-17
4	666	652	643	-14	-23
5	571	595	577	24	6
6	517	548	520	31	3

Table 1: Observed and expected occupation frequencies from 1000 trials for site 28 under Simulation 1. The single expectation under which O_t would be significant at the 5% level is marked with a star.

Figure 11 shows explicitly the linear and quadratic exponential approximations to the functions $\left\{s_{28}^{(t)}(x) : t = 1, \ldots, 6\right\}$. Estimates of the true functions, given by (11), are shown in solid lines while the exponential approximations are shown in dashed lines. The solid lines represent estimates of the true functions rather than the exact values of the true functions because (11) relies on calculation of the matrices $Q^{(1)}, \ldots, Q^{(5)}$, which must be estimated from equations (38), (39) and (40).



(a) First-order approximations.



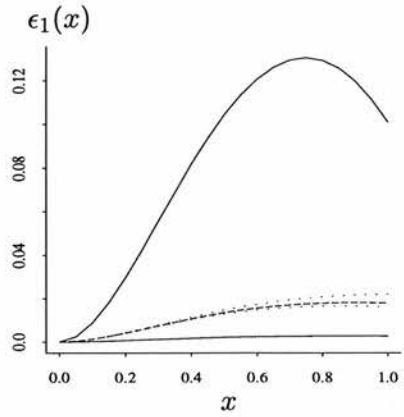
(b) Second-order approximations.

Figure 11: First and second order approximations to $s_i^{(t)}(x)$ from Simulation 1 for $0 \leq x \leq 1$, $i = 28$ and t ranging from 1 (curve (i)) to 6 (curve (vi)). The solid lines give the true functions $s_{28}^{(t)}(x)$ and the dashed lines are obtained from (a) linear and (b) quadratic exponential approximations.

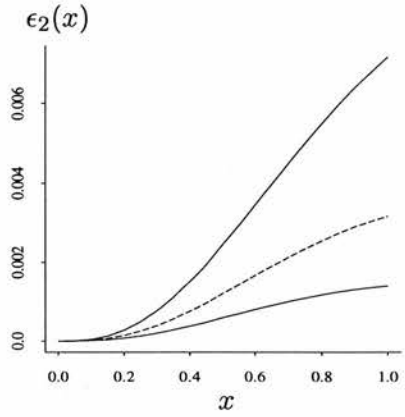
Both the linear and quadratic exponential approximations to $s_i^{(t)}(x)$ in Figure 11 are fairly close to the true curves, although the quadratic approximations are noticeably better. The first-order approximate curves for $t = 1$ and $t = 2$ lie slightly above the true curves, while for $t = 5$ and $t = 6$ this trend is reversed. Nonetheless, the approximations are adequate representations of reality.

Exact results for $t = 1$ and $t = 2$ are shown in Figure 12, again for site 28. Recall that the true functions $s_i^{(t)}(x)$ may be written as $\prod_{k=1}^{M}(1 - \alpha_k x)$ for $t = 1, 2$. Upper and lower bounds with respect to $\boldsymbol{\alpha}$ for the linear error functions $\epsilon_1(x)$, subject to the constraints (a) and (c) of section 4.1, are given in Figure 12 (a) and (c), while bounds for the quadratic errors $\epsilon_2(x)$ for $t = 1$ and 2, subject to the constraints (a), (b) and (c) of section 4.1, are given in Figure 12 (b) and (d). On each diagram the true error, at the true values of $\boldsymbol{\alpha}$, is marked. The quadratic errors, even at the upper bounds, are small enough to be regarded as negligible. The linear errors are small in practice, but not sufficiently so at the upper bounds; however, this is partly an artefact of the constraints used. Figure 12 (a) shows in dotted lines the upper and lower bounds obtained from equations (46) and (47) for the linear error $\epsilon_1(x)$ subject to all three constraints (a), (b) and (c). These bounds are much tighter, and the linear error would be acceptable even at the upper bound.

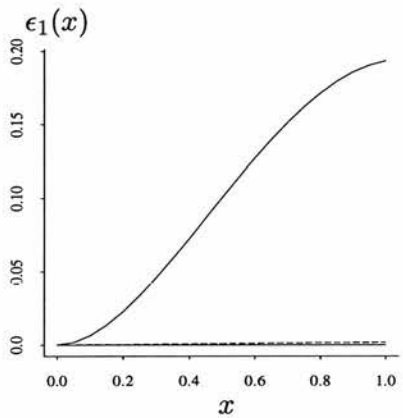
In essence, the plots in Figure 12 therefore demonstrate two things. Firstly, the quadratic exponential approximation is noticeably better than the linear approximation, although both are probably adequate. Secondly, much information about the worth of the exponential approximations is contained in the constraint (b): $\sum_{k=1}^{M} \alpha_k^2 = S_2$. If the linear exponential approximation is to be used, there is likely to be little guarantee that it will be effective unless S_2 is also known. The quadratic approximation is therefore recommended for $t = 1$ and $t = 2$, since if S_2 is known, then so is the quadratic approximation. Naturally, the situation becomes more complicated for $t > 2$, and the extra computational effort required to calculate the quadratic approximation for large t might not be justified.



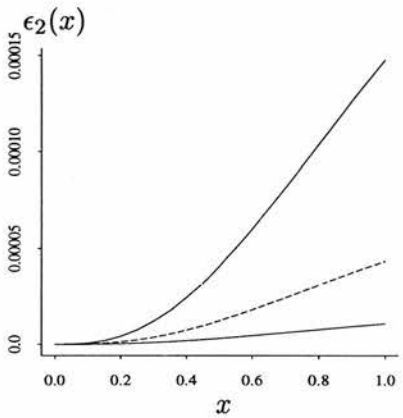
(a) Linear approximation, $t = 1$. $M = 100$, $S_1 = 2.29$.



(b) Quadratic approximation, $t=1$. $M = 100$, $S_1 = 2.29$, $S_2 = 0.314$.



(c) Linear approximation, $t = 2$. $M = 10000$, $S_1 = 1.64$.



(d) Quadratic approximation, $t=2$. $M=10000$, $S_1=1.67$, $S_2=0.017$.

Figure 12: Error functions for $t = 1$ and $t = 2$ under the linear and quadratic exponential approximations, for site 28, Simulation 1. The upper and lower bounds for $\epsilon_k(x)$ $(k = 1, 2)$ are shown for $0 \leq x \leq 1$ in solid lines, and the observed errors are shown in dashed lines. The dotted lines in (a) show the bounds for linear error subject to the extra constraint $\sum_{k=1}^{M} \alpha_k^2 = S_2$. Note the different scales on the linear and quadratic plots.

Simulation 2: $a = 1.0, \quad b = 1.0, \quad p_0 = 0.60, \quad N = 100.$

The initial distribution over the $N = 100$ sites at time $t = 0$ was generated randomly for Simulation 2 using N Bernoulli trials in which each site was allotted a probability of occupation equal to 0.2. This resulted in an initial occupation of 19 sites. Figure 13 shows a typical prediction using parameters $a = 1.0$, $b = 1.0$, $p_0 = 0.60$ starting from this distribution; the population range increases over time. Observed and expected occupation frequencies for an arbitrarily chosen site (site 47) are shown in Table 2. Once again, the observed results match closely with the expected results under both linear and quadratic approximations, with the exception of that for the linear approximation at time $t = 1$. The results from the quadratic approximations at times $t = 5$ and 6 are also somewhat unusual for the appropriate binomial distribution.

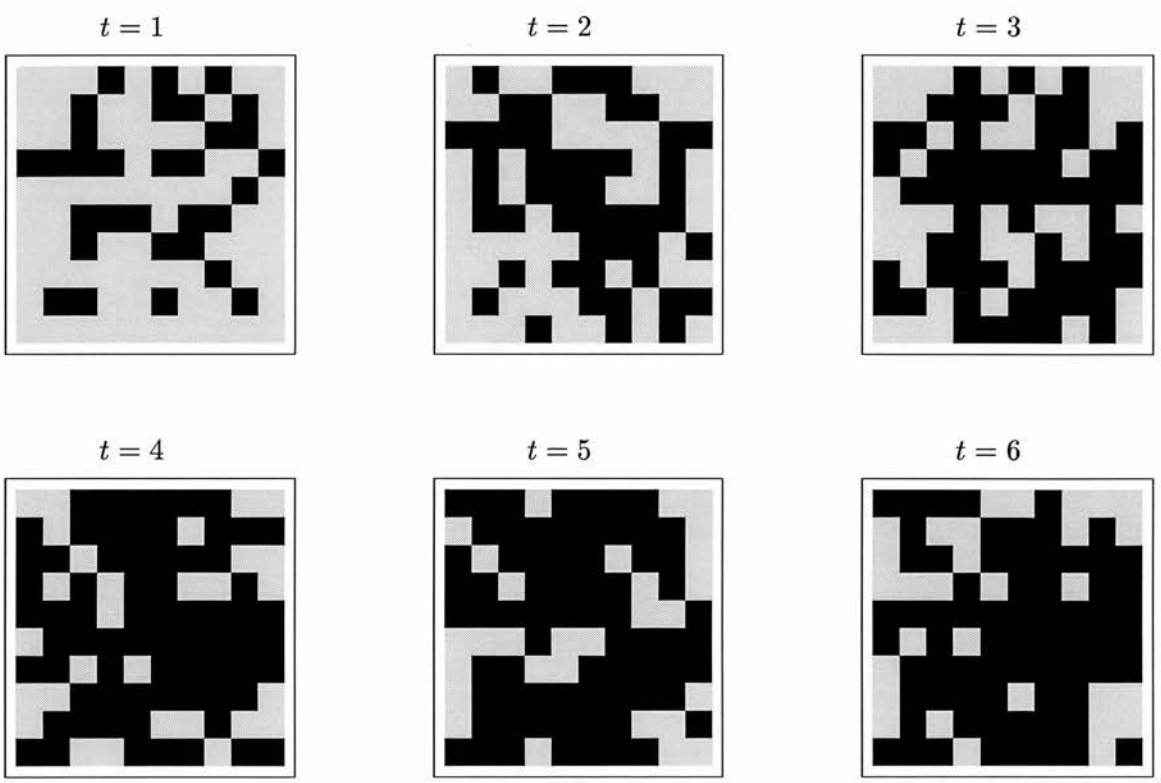


Figure 13: Predicted distributions using $a = 1.0$, $b = 1.0$, $p_0 = 0.60$.

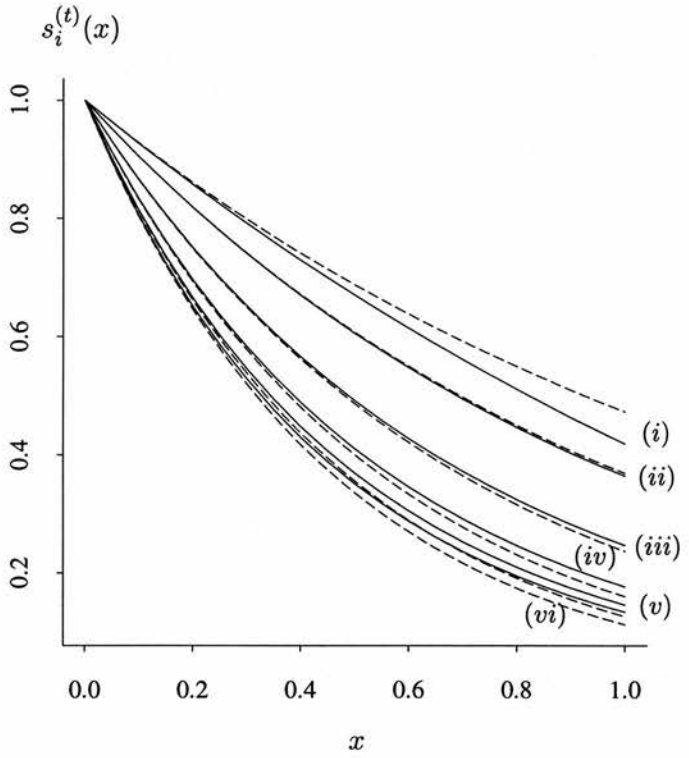
	Number of occupations out of 1000 trials			Prediction error	
t	Obtained O_t	Expected: 1st-order E_{1t}	Expected: 2nd-order E_{2t}	$E_{1t} - O_t$	$E_{2t} - O_t$
1	583	527*	568	-56	-15
2	648	631	650	-17	2
3	776	763	763	-13	-13
4	832	840	823	8	-9
5	885	873	848*	-12	-37
6	894	887	857*	-7	-37

Table 2: Observed and expected occupation frequencies from 1000 trials for site 47 under Simulation 2. Three expectations under which O_t would be significant at the 5% level are marked by stars.

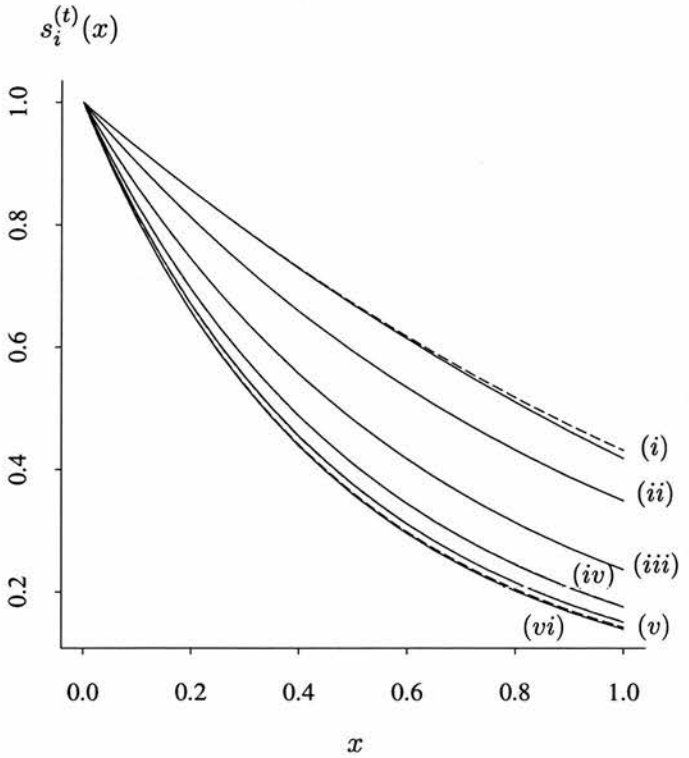
Figure 14 shows the linear and quadratic exponential approximations to the functions $\left\{s_{47}^{(t)}(x)\right\}$ for $t = 1, \ldots, 6$. By contrast with Simulation 1, the probability of occupation in the selected site increases with time, so $s_{47}^{(t)}(1) = 1 - p_{47}^{(t)}$ decreases as t increases.

The quadratic approximations in Figure 14 display a somewhat more dramatic improvement over the linear approximations than in Simulation 1, especially at time $t = 1$ where the linear approximation suffers considerably from the small number $M = 19$ of occupied sites. This accounts for the poor prediction E_{11} in Table 2: the over-estimation of $s_{47}^{(1)}(1)$ in Figure 14 corresponds to the under-estimation of $E_{11} = 1000\left(1 - s_{47}^{(1)}(1)\right)$ in the table. The exact error functions for $t = 1$, 2, site 47, are shown in Figure 15, from which the conclusions are similar to those from Simulation 1.

Note that in the practical application of the colonization model, the approximations to the functions $\left\{s_i^{(t)}(x)\right\}$ will form the basis of likelihood calculations, and the likelihood will be maximized over the parameter space to obtain parameter estimates. The true functions $\left\{s_i^{(t)}(x)\right\}$, represented by solid lines in Figure 14, will not be available for diagnostics; nor will the true error functions given by the dashed lines in Figure 15 (c) and (d). The upper and lower error bounds in Figure 15 therefore comprise the principal remaining diagnostic tool, and it is recommended that a selection are plotted for various parameter vectors and sites once the model has been fitted.

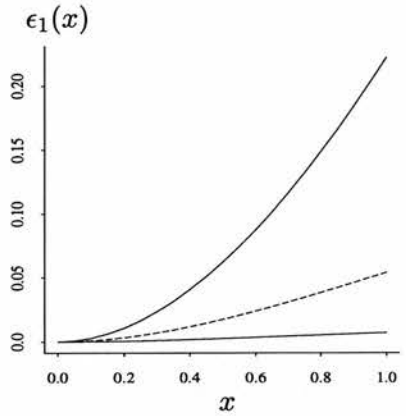


(a) First-order approximations.

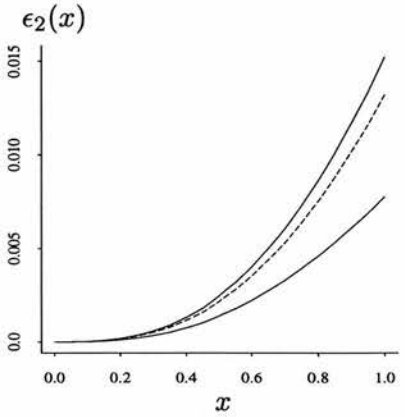


(b) Second-order approximations.

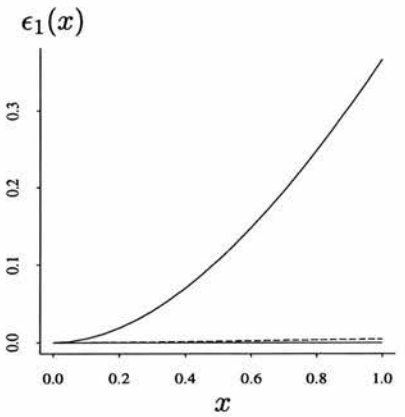
Figure 14: First and second order approximations to $s_i^{(t)}(x)$ from Simulation 2 for $0 \leq x \leq 1$, $i = 47$ and t ranging from 1 (curve (i)) to 6 (curve (vi)). The solid lines give the true functions $s_{47}^{(t)}(x)$ and the dashed lines are obtained from (a) linear and (b) quadratic exponential approximations.



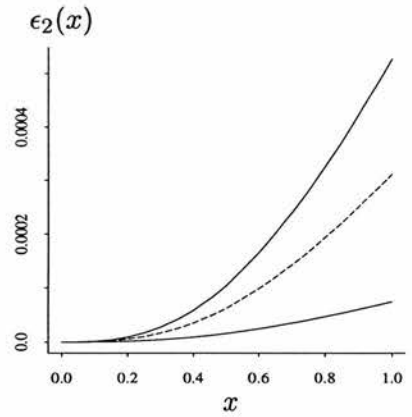
(a) Linear approximation, $t = 1$. $M = 19$, $S_1 = 0.749$.



(b) Quadratic approximation, $t=1$. $M = 19$, $S_1 = 0.749$, $S_2 = 0.181$.



(c) Linear approximation, $t = 2$. $M = 1900$, $S_1 = 0.998$.



(d) Quadratic approximation, $t=2$. $M=1900$, $S_1=1.04$, $S_2=0.026$.

Figure 15: Error functions for $t = 1$ and $t = 2$ under the linear and quadratic exponential approximations, for site 47, Simulation 2. The upper and lower bounds for $\epsilon_k(x)$ $(k = 1, 2)$ are shown for $0 \leq x \leq 1$ in solid lines, and the observed errors are shown in dashed lines.

Simulation 3: $a = 0.10, \quad b = 0.10, \quad p_0 = 0.90, \quad N = 100.$

The parameters in Simulation 3 generate presence in all sites from time $t = 1$ onwards, even when the starting distribution at time $t = 0$ is sparse. The distribution at time $t = 0$ consisted in this case of 22 occupied sites, chosen at random. The expected occupation frequencies E_{1t} and E_{2t} for the selected site 61 were equal to 1000 for all $t = 1, \ldots, 6$, and all of the observed occupation frequencies for this site were also 1000. The approximations for site 61 are extremely close to the true functions (Figure 16); the errors are quantified for $t = 1, 2$ with the appropriate upper and lower bounds in Figure 17.

The first-order approximations in Figure 16(a) raise an interesting point in that the approximate curves for $t = 2, \ldots, t = 6$ are identical. If the estimated values of $p_j^{(t)}$ are 1 for all sites j at time t, it may be shown as follows that the estimated function $s_i^{(t+1)}$ adopts a constant form independent of time. For each j, substituting $p_j^{(t)} = 1$ in equation (38) gives

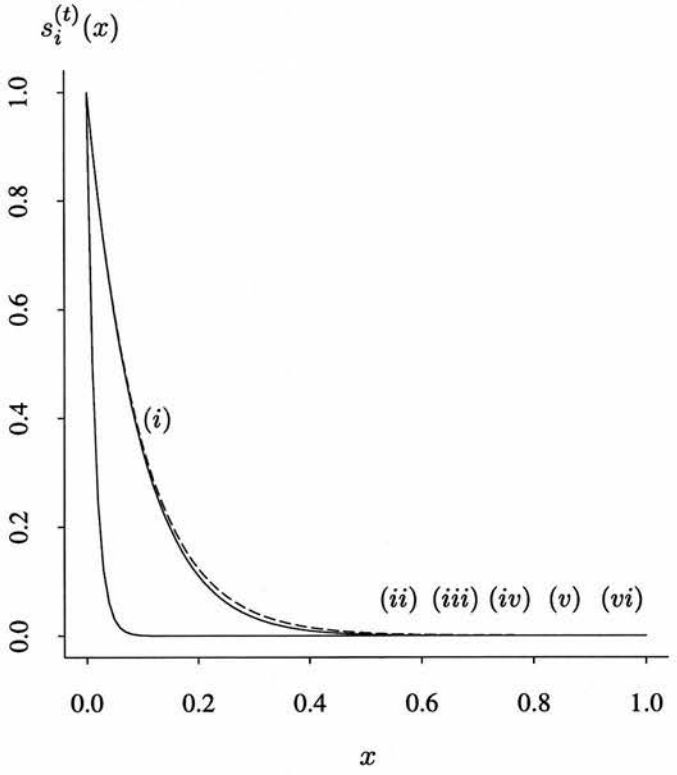
$$q_{ji}^{(t)} = Q^{(t)}[j, i] \simeq -\left(\frac{\log(1 - p_{ji})}{\boldsymbol{y}^{(0)\prime} Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}[j]}\right) \quad \forall j,$$

whence from (25) and the linear truncation of (16),

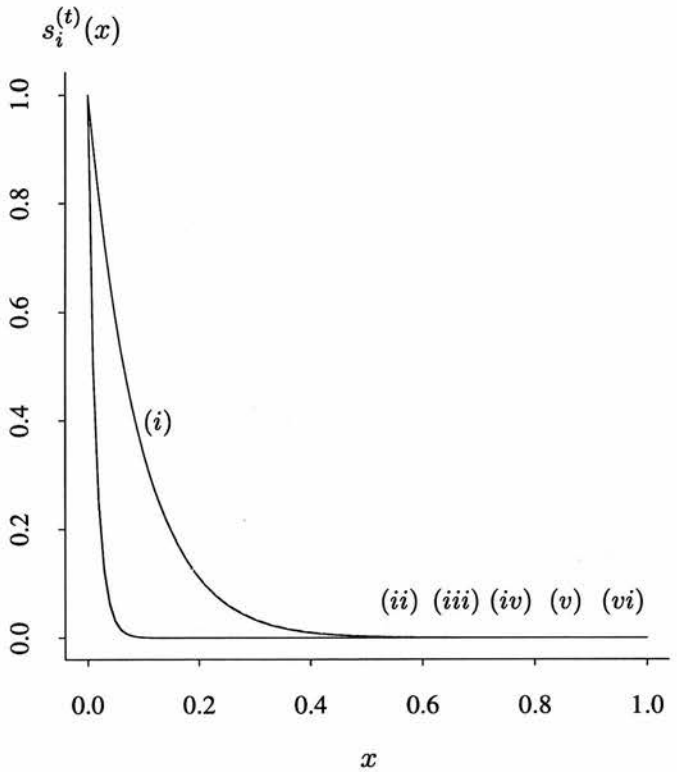
$$\begin{aligned} \log s_i^{(t+1)}(x) &= -x\, \boldsymbol{y}^{(0)\prime} Q^{(0)} Q^{(1)} \ldots Q^{(t-1)} Q^{(t)}[i] \\ &= -x \sum_j \left(\boldsymbol{y}^{(0)\prime} Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}[j]\right) Q^{(t)}[j, i] \\ &= x \sum_j \log(1 - p_{ji})\,. \end{aligned}$$

If the estimated values of $p_j^{(t+1)}$ are also 1 for all j (that is, if $s_j^{(t+1)}(1) = 0 \ \forall j$), the estimated function $s_i^{(t+2)}$ will be identical to $s_i^{(t+1)}$, and so on. The quadratic exponential approximations to $s_i^{(t)}$, however, do not share this property: the dashed curves in Figure 16(b) appear indistinguishable, but are not identical in reality.

This simulation also illustrates the rather unusual situation where the approximation $s_i^{(t)}(x) \simeq (1 - \frac{S_1}{M}x)^M$ is an appreciable improvement over the linear exponential approximation at time $t = 1$. The combination of small M and large S_1 in Figure 17(a) yields a maximum observed error of 0.0132 for the linear exponential approximation, whereas the alternative approximation $s_i^{(t)}(x) \simeq (1 - \frac{S_1}{M}x)^M$ produces a maximum error of 0.000745. The alternative error function is shown in the dotted line on Figure 17(a).

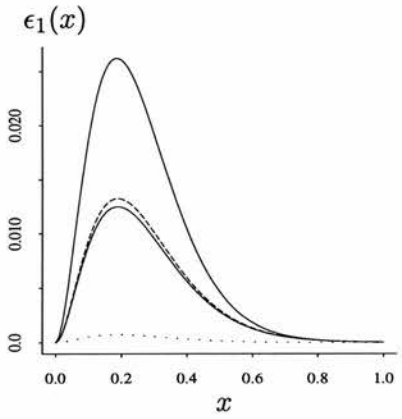


(a) First-order approximations.

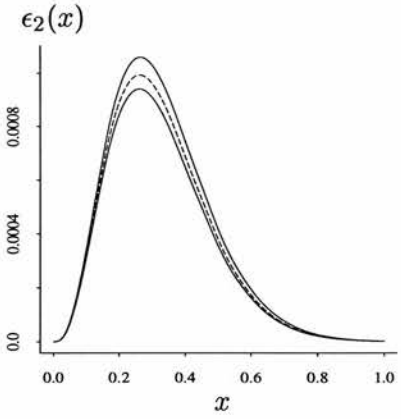


(b) Second-order approximations.

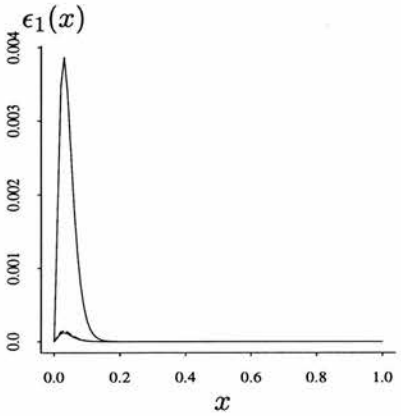
Figure 16: First and second order approximations to $s_i^{(t)}(x)$ from Simulation 3 for $0 \leq x \leq 1$, $i = 61$ and t ranging from 1 (curve (i)) to 6 (curve (vi)). The solid lines give the true functions $s_{61}^{(t)}(x)$ and the dashed lines are obtained from (a) linear and (b) quadratic exponential approximations.



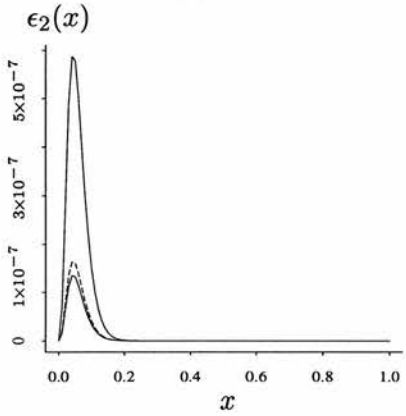
(a) Linear approximation, $t = 1$. $M = 22$, $S_1 = 10.4$.



(b) Quadratic approximation, $t=1$. $M = 22$, $S_1 = 10.4$, $S_2 = 5.22$.



(c) Linear approximation, $t = 2$. $M = 2200$, $S_1 = 70.1$.



(d) Quadratic approximation, $t=2$. $M = 2200$, $S_1 = 68.8$, $S_2 = 2.61$.

Figure 17: Error functions for $t = 1$ and $t = 2$ under the linear and quadratic exponential approximations, for site 61, Simulation 3. The upper and lower bounds for $\epsilon_k(x)$ $(k = 1, 2)$ are shown for $0 \leq x \leq 1$ in solid lines, and the observed errors are shown in dashed lines. The dotted line on (a) gives the error from the alternative approximation $s_i^{(t)}(x) \simeq (1 - \frac{S_1}{M}x)^M$.

Simulation 4: $a = 1.0, \quad b = 2.0, \quad p_0 = 0.60, \quad N = 400.$

The final simulation is conducted on a larger grid of site squares. Figure 18 shows a typical prediction from Simulation 4, beginning from an initial distribution at time $t=0$ composed of 42 occupied sites chosen at random. Observed and expected occupation frequencies for site 203 are given in Table 3; only one of the observed frequencies was unusual for the first-order or second-order expectation. The quadratic approximations to $s_{203}^{(t)}(x)$ in Figure 19 exhibit slight but noticeable improvement over the linear approximations, and the errors for this site at times $t = 1, 2$ (Figure 20) are all small, even at the upper bounds.

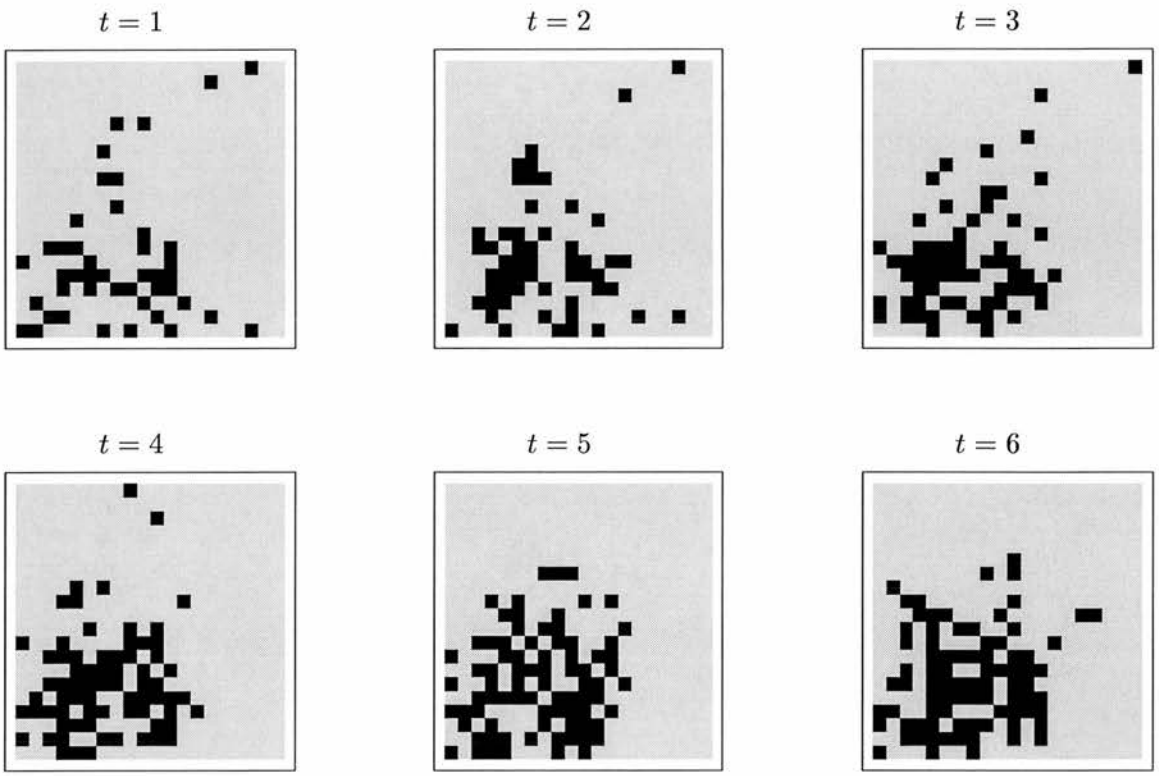
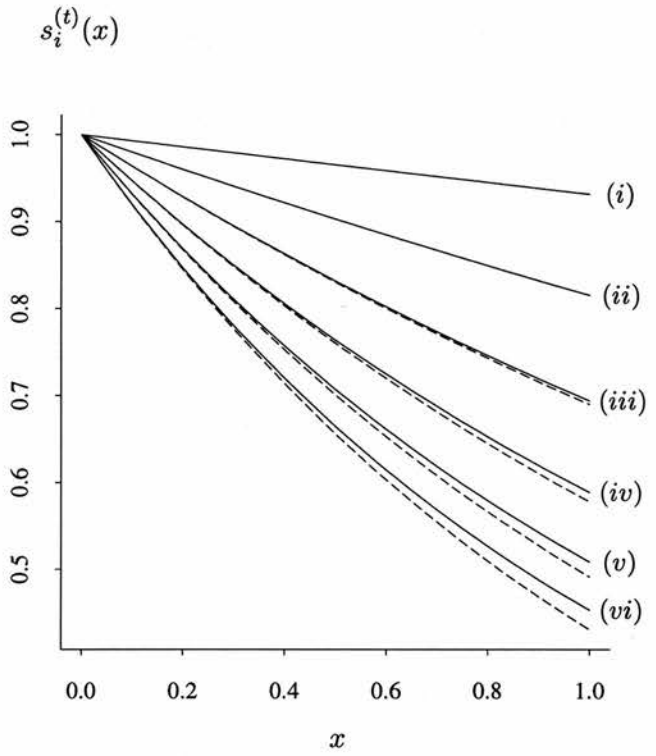


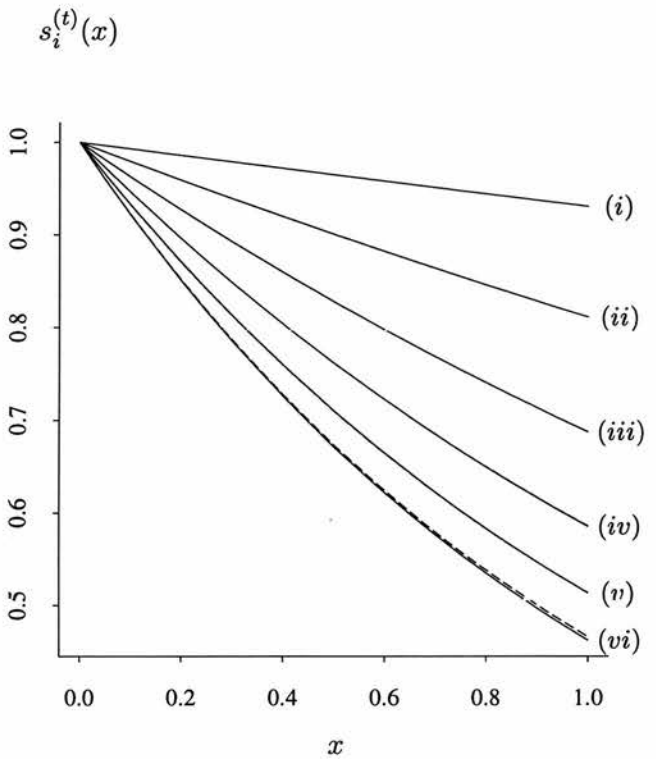
Figure 18: Predicted distributions using $a = 1.0$, $b = 2.0$, $p_0 = 0.60$.

	Number of occupations out of 1000 trials			Prediction error	
t	Obtained O_t	Expected: 1st-order E_{1t}	Expected: 2nd-order E_{2t}	$E_{1t} - O_t$	$E_{2t} - O_t$
1	82	68	69	-14	-13
2	194	185	188	-9	-6
3	294	310	312	16	18
4	429	422	414	-7	-15
5	514	509	486	-5	-28
6	570	570	534*	0	-36

Table 3: Observed and expected occupation frequencies from 1000 trials for site 203 under Simulation 4. One of the results was significant at the 5% level.

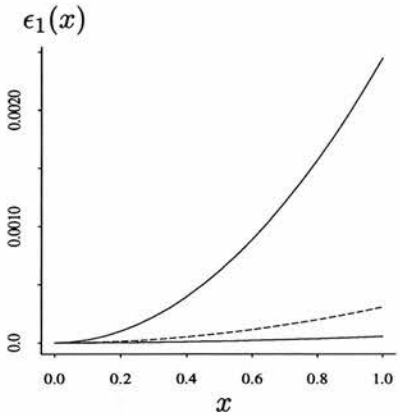


(a) First-order approximations.

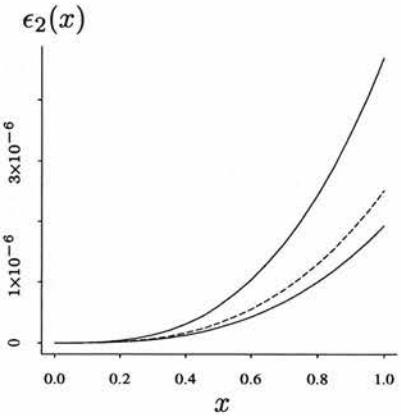


(b) Second-order approximations.

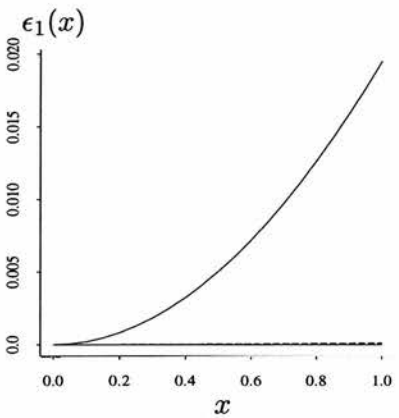
Figure 19: First and second order approximations to $s_i^{(t)}(x)$ from Simulation 4 for $0 \leq x \leq 1$, $i = 203$ and t ranging from 1 (curve (i)) to 6 (curve (vi)). The solid lines give the true functions $s_{203}^{(t)}(x)$ and the dashed lines are obtained from (a) linear and (b) quadratic exponential approximations.



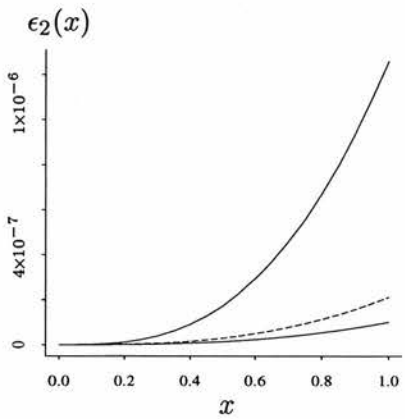
(a) Linear approximation, $t = 1$. $M = 42$, $S_1 = 0.071$.



(b) Quadratic approximation, $t=1$. $M = 42, S_1 = 0.071, S_2 = 0.00065$.



(c) Linear approximation, $t = 2$. $M = 16800$, $S_1 = 0.204$.



(d) Quadratic approximation, $t=2$. $M=16800, S_1=0.208, S_2=0.00028$.

Figure 20: Error functions for $t = 1$ and $t = 2$ under the linear and quadratic exponential approximations, for site 203, Simulation 4. The upper and lower bounds for $\epsilon_k(x)$ $(k = 1, 2)$ are shown for $0 \leq x \leq 1$ in solid lines, and the observed errors are shown in dashed lines.

5 Likelihood estimation

The results of Section 4 suggest that the exponential approximations to the colony-scale polynomials may be used with confidence. In particular the close correspondence between observation and prediction in Tables 1, 2 and 3 indicates that the approximate method may be expected to generate occupation probabilities for each site over a period of several years that tally closely with the true probabilities under the colonization model. It is important also to recognize that there is considerable stochasticity inherent in the colonization model itself: several simulations from the model, each conducted with the same set of parameter values, can lead to a variety of different distribution patterns when run over a number of years. In this context, and in the light of the results from Section 4, the errors introduced in approximating the colony-scale polynomials by exponential functions are of little consequence.

In this section it is shown how the likelihood function may be estimated using the output from the modified branching process.

5.1 Extinction probabilities

The *extinction probability at time* t of a set $A \subseteq \{1, 2, \ldots, N\}$ of sites is defined as the probability that all sites in A are unoccupied at time t. For the purposes of the present discussion there is no stipulation that all sites in A must remain unoccupied at times subsequent to t. Extinction probabilities are readily calculated under the modified branching process formulation, since if extinction has occurred in the set of sites A at time t then colonization must have failed along *all* possible routes between occupied sites at time 0 and sites in A at time t. By contrast, the probability of obtaining any non-zero spatial distribution on A at time t is much more difficult to calculate, because there are large numbers of possible routes along which colonization might have taken place. The ease with which extinction probabilities may be calculated for any set of sites is the principal advantage of formulating the colonization model as a modified branching process.

The approximate extinction probability for a set of sites A at time t may be derived in much the same way as the approximate extinction probability $1 - p_i^{(t)} = s_i^{(t)}(1)$ for the individual site i in Section 3. Let $N_A^{(t)}$ be the random variable denoting the total number of colonies in A at time t:

$$N_A^{(t)} = \sum_{i \in A} N_i^{(t)};$$

and let the function $s_A^{(t)}$ be defined as

$$s_A^{(t)}(x) = \mathbb{E}\left\{(1-x)^{N_A^{(t)}}\right\}.$$

The extinction probability for A at time t is given by

$$\mathbb{P}\left(N_A^{(t)} = 0\right) = s_A^{(t)}(1),$$

and $s_A^{(t)}$ may be approximated by an exponential function in the same way as $s_i^{(t)}$.

Accordingly, let $g_A(x) = \log s_A^{(t)}(x)$, and consider the truncated Taylor series $g_A(x) \approx g_A(0) + x g'_A(0) + x^2 g''_A(0)/2$. Since $g_A(0) = \log(1) = 0$, the first and second order exponential approximations to $s_A^{(t)}(x)$ are respectively $\exp(x g'_A(0))$ and $\exp(x g'_A(0) + x^2 g''_A(0)/2)$.

Let $V_A^{(t)}(k,u)$ be the number of colonies established in the set A at time t, starting from a single ancestor colony in site k at time u. Clearly, $V_A^{(t)}(k,u) = \sum_{i\in A} V_i^{(t)}(k,u)$, where $V_i^{(t)}(k,u)$ is the number of colonies established in the single site i, arising from a single colony in the site k at time u. Using the same reasoning given for equations (21), (22) and (23), the following results are obtained:

$$\begin{aligned}
g'_A(0) &= -\sum_{k_0 : N_{k_0}^{(0)} = 1} \mathbb{E}\left(V_A^{(t)}(k_0\,,\,0)\right); \\
\mathbb{E}\left(V_A^{(t)}(k_u,u)\right) &= \sum_{k_{u+1}} q_{k_u k_{u+1}}^{(u)}\, \mathbb{E}\left(V_A^{(t)}(k_{u+1},u+1)\right) \qquad (u < t-1); \\
\mathbb{E}\left(V_A^{(t)}(k_{t-1},t-1)\right) &= \sum_{i\in A} q_{k_{t-1} i}^{(t-1)};
\end{aligned}$$

whence

$$g'_A(0) = -\sum_{i\in A} \boldsymbol{y}^{(0)\prime} Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}[i]. \tag{48}$$

The first-order approximation to the extinction probability in set A at time t is thus

$$\mathbb{P}\left(N_A^{(t)} = 0\right) = s_A^{(t)}(1) \simeq \exp\left(-\sum_{i\in A} \boldsymbol{y}^{(0)\prime} Q^{(0)} Q^{(1)} \ldots Q^{(t-1)}[i]\right). \tag{49}$$

Note that at the first-order level of approximation, (25) and (49) imply that

$$\mathbb{P}\left(N_A^{(t)} = 0\right) \simeq \prod_{i\in A} \mathbb{P}\left(N_i^{(t)} = 0\right). \tag{50}$$

The expression for $g''_A(0)$ may also be obtained by following the reasoning of section 3.2, although since it is a little more involved than the calculation for $g'_A(0)$ the full derivation will not be given here. Let $\boldsymbol{a}$ be the vector denoting inclusion of the set A: that is, the ith element of $\boldsymbol{a}$ is

$$a_i = \begin{cases} 1 & \text{if } i \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Further let $\boldsymbol{r_A}^{(t_1,\ldots,t_n)}$ be the vector with ith component

$$\boldsymbol{r_A}^{(t_1,\ldots,t_n)}[i] = \left(Q^{(t_1)} \ldots Q^{(t_n)} \boldsymbol{a}[i]\right)^2 .$$

The expression for $g''_A(0)$ is

$$\begin{aligned} g''_A(0) = \; \boldsymbol{y^{(0)}}' \; \Big\{ & Q^{(0)} \boldsymbol{r_A}^{(1,\ldots,t-1)} + \ldots + Q^{(0)} \ldots Q^{(t-2)} \boldsymbol{r_A}^{(t-1)} \\ & - Q^{(0)} M^{(1)} \boldsymbol{r_A}^{(2,\ldots,t-1)} - \ldots - Q^{(0)} \ldots Q^{(t-3)} M^{(t-2)} \boldsymbol{r_A}^{(t-1)} \\ & - Q^{(0)} \ldots Q^{(t-2)} M^{(t-1)} \boldsymbol{a} - M^{(0)} \boldsymbol{r_A}^{(1,\ldots,t-1)} \Big\} \end{aligned} \tag{51}$$

with $g''_A(0) = -\boldsymbol{y^{(0)}}' M^{(0)} \boldsymbol{a}$ when $t = 1$, and (48) and (51) thereby give the second-order approximation to the extinction probability in set A at time t as

$$\mathbb{P}\left(N_A^{(t)} = 0\right) \;=\; s_A^{(t)}(1) \;\simeq\; \exp\left(-\boldsymbol{y^{(0)}}' Q^{(0)} Q^{(1)} \ldots Q^{(t-1)} \boldsymbol{a} + \frac{g''_A(0)}{2}\right) . \tag{52}$$

At the second-order level of approximation it is no longer true that the approximate extinction probability for A is the product of the approximate extinction probabilities of the sites $i \in A$. The product of the approximate second-order extinction probabilities may nonetheless be used to estimate the extinction probability for A, but it should be regarded as a pseudo-estimate of the true extinction probability — just as a product of probabilities of correlated observations might be regarded as a pseudo-likelihood.

Demonstrations of extinction probability estimates are given in Tables 4 and 5, stemming from Simulations 1 and 4 of section 4.2 respectively. For each table, a set A of adjacent sites was selected and 1000 simulations from the colonization model conducted. The observed number of the 1000 trials in which A was occupied was recorded, and compared against the expected number $1000\left(1 - \hat{\mathbb{P}}(N_A^{(t)} = 0)\right)$ using three different estimates $\hat{\mathbb{P}}(N_A^{(t)} = 0)$ of extinction probability: the first-order estimate (49), the second-order pseudo-estimate, and the full second-order estimate (52). For the most part the empirical extinction probabilities are consistent with the estimates, although the full second-order results provide a substantial improvement over the other estimates in Table 5. There is apparently little to

choose between the first-order estimates and the second-order pseudo-estimates. It must be stressed, however, that there is considerable variability in the observations O_t when the experiment is repeated, so it is unwise to draw firm conclusions from the results shown.

	Number of occupations out of 1000 trials				Prediction error		
	Obtained	Expected					
t	O_t	1st-order E_{1t}	2nd-order (pseudo), E_{2t}	2nd-order (full), E_{3t}	$E_{1t} - O_t$	$E_{2t} - O_t$	$E_{2t} - O_t$
1	990	992	993	993	2	3	3
2	862	888*	892*	875	26	30	13
3	695	717	720	686	22	25	-9
4	537	569*	569*	531	32	32	-6
5	443	461	458	421	18	15	-22
6	366	384	379	345	18	13	-21

Table 4: Observed and expected occupation frequencies for a set A of 9 adjacent sites on a grid of size $N = 100$, using the parameters of Simulation 1 in section 4.2. Those expectations under which the observation O_t is significant at the 5% level are marked with a star.

	Number of occupations out of 1000 trials				Prediction error		
	Obtained	Expected					
t	O_t	1st-order E_{1t}	2nd-order (pseudo), E_{2t}	2nd-order (full), E_{3t}	$E_{1t} - O_t$	$E_{2t} - O_t$	$E_{2t} - O_t$
1	693	683	692	692	-10	-1	-1
2	572	594	603*	574	22	31	2
3	525	549	554	506	24	29	-19
4	491	531*	531*	471	40	40	-20
5	471	535*	528*	460	64	57	-11
6	509	554*	540	465*	45	31	-44

Table 5: Observed and expected occupation frequencies for a set A of 21 adjacent sites on a grid of size $N = 400$, using the parameters of Simulation 4 in section 4.2. Those expectations under which the observation O_t is significant at the 5% level are marked with a star.

5.2 Use of extinction probabilities in likelihood calculation

Let $\boldsymbol{y^{(T)}}$ be the observed spatial distribution at the time T of the final survey: that is, $\boldsymbol{y^{(T)}}$ is a vector with ith component

$$y_i^{(T)} = \begin{cases} 1 & \text{if site } i \text{ is occupied at time } T, \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function $L\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y^{(0)}}\right)$ given the observed spatial distribution $\boldsymbol{y^{(0)}}$ at time 0 and the unknown parameters $\boldsymbol{\theta}$ of the colonization probabilities is the discrete joint probability

$$L\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y^{(0)}}\right) = \mathbb{P}\left(y_1^{(T)}, \ldots, y_N^{(T)} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y^{(0)}}\right).$$

As mentioned above, it is straightforward to calculate the joint probability that all sites are unoccupied at time T, but if any of the sites are occupied the calculation becomes much more difficult. However, this difficulty may be overcome to some extent by writing the joint probability of any sequence of 0s and 1s in terms of extinction probabilities only.

Suppose without loss of generality that $y_1^{(T)} = \ldots = y_m^{(T)} = 1$ and $y_{m+1}^{(T)} = \ldots = y_N^{(T)} = 0$. The joint probability of $y_1^{(T)}, \ldots, y_N^{(T)}$ may be obtained using the inclusion/exclusion formula for a probability of unions, using the fact that

$$\begin{aligned}
&\mathbb{P}\left(y_1^{(T)} = 1, \ldots, y_m^{(T)} = 1,\ y_{m+1}^{(T)} = 0, \ldots, y_N^{(T)} = 0\right) \\
&\quad = \mathbb{P}\left(y_1^{(T)} = 1, \ldots, y_m^{(T)} = 1 \,\middle|\, y_{m+1}^{(T)} = 0, \ldots, y_N^{(T)} = 0\right) \mathbb{P}\left(y_{m+1}^{(T)} = 0, \ldots, y_N^{(T)} = 0\right) \\
&\quad = \left\{1 - \mathbb{P}\left(\left(y_1^{(T)} = 0\right) \cup \ldots \cup \left(y_m^{(T)} = 0\right) \,\middle|\, y_{m+1}^{(T)} = 0, \ldots, y_N^{(T)} = 0\right)\right\} \times \\
&\qquad\qquad \mathbb{P}\left(y_{m+1}^{(T)} = 0, \ldots, y_N^{(T)} = 0\right) \quad (53)
\end{aligned}$$

—that is, the joint conditional probability that $y_1^{(T)}, \ldots, y_m^{(T)}$ all take the value 1 is the complement of the conditional probability that at least one of these elements is zero, which may be expressed as the probability of a union.

To simplify notation, let $B = \{m+1, \ldots, N\}$ be the set of sites with status zero at time T, and let $\boldsymbol{y_B^{(T)}}$ denote the vector $(y_{m+1}^{(T)}, \ldots, y_N^{(T)})$. The notation $\mathbf{0}$ is used to indicate the

vector $(0,\ldots,0)$, with length determined by context. Continuing from (53),

$$\mathbb{P}\left(y_1^{(T)}=1,\ldots,y_m^{(T)}=1,\, y_{m+1}^{(T)}=0,\ldots,y_N^{(T)}=0\right)$$

$$= \left\{1-\frac{\mathbb{P}\left\{\left(\left(y_1^{(T)}=0\right)\cup\ldots\cup\left(y_m^{(T)}=0\right)\right)\cap\left(\boldsymbol{y_B^{(T)}=0}\right)\right\}}{\mathbb{P}\left(\boldsymbol{y_B^{(T)}=0}\right)}\right\}\mathbb{P}\left(\boldsymbol{y_B^{(T)}=0}\right)$$

$$= \mathbb{P}\left(\boldsymbol{y_B^{(T)}=0}\right)-\mathbb{P}\left\{\left(\left(y_1^{(T)}=0\right)\cap\left(\boldsymbol{y_B^{(T)}=0}\right)\right)\cup\ldots\cup\left(\left(y_m^{(T)}=0\right)\cap\left(\boldsymbol{y_B^{(T)}=0}\right)\right)\right\},$$

because set intersection $\cap$ is distributive over set union $\cup$. Applying the inclusion/exclusion formula yields

$$\begin{aligned}
&\mathbb{P}\left(y_1^{(T)}=1,\ldots,y_m^{(T)}=1,\, y_{m+1}^{(T)}=0,\ldots,y_N^{(T)}=0\right)\\
&\quad=\mathbb{P}\left(\boldsymbol{y_B^{(T)}=0}\right)-\sum_{\substack{C\subseteq\{1,\ldots,m\}\\ C\neq\emptyset}}(-1)^{|C|-1}\,\mathbb{P}\left(\bigcap_{j\in C}\left\{\left(y_j^{(T)}=0\right)\cap\left(\boldsymbol{y_B^{(T)}=0}\right)\right\}\right)\\
&\quad=\sum_{\substack{C\subseteq\{1,\ldots,m\}\\ \text{including }\emptyset}}(-1)^{|C|}\,\mathbb{P}\left(\bigcap_{j\in C\cup B}\left\{y_j^{(T)}=0\right\}\right), \qquad (54)
\end{aligned}$$

where $\mathbb{P}\left(\bigcap_{j\in S}\left\{y_j^{(T)}=0\right\}\right)$ for any set S denotes the joint probability that $y_j^{(T)}=0$ for all elements j of S, or equivalently the extinction probability for S at time T, $\mathbb{P}(N_S^{(T)}=0)$. The extinction probability for the set $\emptyset$ is defined as 1, so that $\mathbb{P}(N_\emptyset^{(T)}=0)=1$.

Equation (54) therefore demonstrates that the joint probability of any combination of 0s and 1s in the year T distribution may be expressed entirely in terms of extinction probabilities, and therefore the likelihood $L\left(\boldsymbol{y^{(T)}}\,\middle|\,\boldsymbol{\theta},\,\boldsymbol{y^{(0)}}\right)$ may also be expressed in these terms. The practical use of (54) is however restricted by the number of terms involved in the sum: if the distribution at time T includes m occupied sites, a separate extinction probability must be evaluated for every possible subset of $\{1,\ldots,m\}$, entailing 2^m terms in total. In most circumstances a method with this complexity would be out of the question for computational reasons. However, following from the observation of

equation (50) it is shown in the next section that a first-order approximation to the full likelihood is easily calculable, and a suggestion is made as to a possible second-order approximation.

5.3 First-order approximation to the full likelihood

Equation (54) may be rewritten as

$$L\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y^{(0)}}\right) \;=\; \sum_{C \subseteq \{1,\ldots,m\}} (-1)^{|C|}\, \mathbb{P}\left(N^{(T)}_{C \cup B} = 0\right), \tag{55}$$

where it is assumed without loss of generality that the observed spatial distribution at time T is $\boldsymbol{y^{(T)}} = (\underbrace{1,\ldots,1}_{m},\underbrace{0,\ldots,0}_{N-m})$ for some $m \leq N$, and where all probabilities on the right-hand side are implicitly conditional on $\boldsymbol{y^{(0)}}$ and $\boldsymbol{\theta}$. The set B is $\{m+1,\ldots,N\}$ as above. The first-order approximation to the full likelihood is obtained by substituting all extinction probabilities in (55) by their first-order approximations, given by (49).

At the first-order level of approximation, (50) gives $\hat{\mathbb{P}}\left(N^{(t)}_A = 0\right) = \prod_{i \in A} \hat{\mathbb{P}}\left(N^{(t)}_i = 0\right)$ for any set of sites A, where $\hat{\mathbb{P}}$ denotes an estimated probability. Inserting this into (55) gives

$$\begin{aligned}
\hat{L}\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y^{(0)}}\right) \;&=\; \prod_{i \in B} \hat{\mathbb{P}}\left(N^{(T)}_i = 0\right) \sum_{C \subseteq \{1,\ldots,m\}} (-1)^{|C|}\, \hat{\mathbb{P}}\left(N^{(T)}_C = 0\right) \\
&= \prod_{i \in B} \hat{\mathbb{P}}\left(N^{(T)}_i = 0\right) \left\{ \sum_{C \subseteq \{2,\ldots,m\}} (-1)^{|C|}\, \hat{\mathbb{P}}\left(N^{(T)}_C = 0\right) + \sum_{C \subseteq \{2,\ldots,m\}} (-1)^{|C|+1}\, \hat{\mathbb{P}}\left(N^{(T)}_{C \cup \{1\}} = 0\right) \right\} \\
&\qquad\qquad \text{(isolating the terms involving the first site)} \\
&= \prod_{i \in B} \hat{\mathbb{P}}\left(N^{(T)}_i = 0\right) \left\{1 - \hat{\mathbb{P}}\left(N^{(T)}_1 = 0\right)\right\} \sum_{C \subseteq \{2,\ldots,m\}} (-1)^{|C|}\, \hat{\mathbb{P}}\left(N^{(T)}_C = 0\right) \\
&\;\;\vdots \\
&= \prod_{i \in B} \hat{\mathbb{P}}\left(N^{(T)}_i = 0\right) \left\{1 - \hat{\mathbb{P}}\left(N^{(T)}_1 = 0\right)\right\} \ldots \left\{1 - \hat{\mathbb{P}}\left(N^{(T)}_{m-1} = 0\right)\right\} \sum_{C \subseteq \{m\}} (-1)^{|C|}\, \hat{\mathbb{P}}\left(N^{(T)}_C = 0\right) \\
&= \prod_{i \in B} \hat{\mathbb{P}}\left(N^{(T)}_i = 0\right) \prod_{i \in \{1,\ldots,m\}} \left\{1 - \hat{\mathbb{P}}\left(N^{(T)}_i = 0\right)\right\}. \qquad (56)
\end{aligned}$$

The first-order approximation to the full likelihood is therefore precisely the product of the estimated probabilities of the observations for each site, and as such is equivalent to the likelihood that would be obtained if all observations were assumed independent. In many quite general statistical analyses, assumptions of independence — however tenuous — are made in order to justify the multiplication of individual probabilities in this way to yield the likelihood function. Regression models and their generalizations, for instance, are frequently used in these circumstances. In most cases the result should be regarded more accurately as a pseudo-likelihood than a true likelihood. In the case of the colonization model, however, the result (56) lends considerable theoretical justification to the procedure. Throughout the rest of this chapter the term *product-likelihood* will be used to indicate the likelihood obtained through multiplication of non-independent quantities. A number of alternative forms of pseudo-likelihood function have also been widely used in the literature (e.g. Besag 1975; Kalbfleisch 1982).

The first-order approximation to the full likelihood requires that first-order approximations are used for the calculation of $\hat{\mathbb{P}}\left(N_i^{(T)} = 0\right)$ for all sites i at the time T of the final survey; however, second-order approximations may be used for any or all previous times in the calculation of the quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$ for sites i, h and times $t < T$. If the second-order approximations are also used at time T, the product-likelihood function may still be calculated and used for parameter estimation; however, the theoretical justification of the product-likelihood as an approximation to the full likelihood is no longer valid.

5.4 Second-order approximation to the full likelihood

The full second-order estimates of extinction probabilities in (52) are not a product of site-wise extinction probabilities, so the expression (54) does not simplify in the same way as it does for the first-order estimates. Here, an *ad-hoc* method is suggested which might enable the second-order estimates to be used in likelihood calculation.

Consider $L\left(\boldsymbol{y}^{(\boldsymbol{T})} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y}^{(\mathbf{0})}\right) = \mathbb{P}\left(y_1^{(T)}, \ldots, y_N^{(T)}\right)$ where the right-hand side conditions implicitly on the unknown parameters $\boldsymbol{\theta}$ and the distribution $\boldsymbol{y}^{(\mathbf{0})}$ at time 0. Then

$$\mathbb{P}\left(y_1^{(T)}, \ldots, y_N^{(T)}\right) = \mathbb{P}\left(y_1^{(T)} \,\middle|\, y_2^{(T)} \ldots, y_N^{(T)}\right) \mathbb{P}\left(y_2^{(T)} \,\middle|\, y_3^{(T)} \ldots, y_N^{(T)}\right) \ldots \mathbb{P}\left(y_{N-1}^{(T)} \,\middle|\, y_N^{(T)}\right). \tag{57}$$

Each of the conditionals on the right-hand side of (57) may be approximated by a truncated

conditional involving only those sites closest to the target site: for example

$$\mathbb{P}\left(y_1^{(T)} \,\middle|\, y_2^{(T)} \ldots, y_N^{(T)}\right) \simeq \mathbb{P}\left(y_1^{(T)} \,\middle|\, y_2^{(T)}, y_3^{(T)}, y_4^{(T)}\right)$$

if sites 2, 3 and 4 are closest to site 1. The truncated conditionals may then be written as a ratio of joint probabilities, which are calculated using the inclusion/exclusion formula as in equation (54). The complexity of this calculation depends only on the number of 1s involved in the truncated conditional: for example, a conditional involving m 1s demands the calculation of 2^m extinction probabilities. When the distribution $\boldsymbol{y^{(T)}}$ at time T is reasonably sparse and the truncations are chosen appropriately, this does not pose a problem. The second-order likelihood is approximated by the product of the estimated truncated conditionals.

In practice, there are difficulties associated with the implementation of this method. Extinction in a set A at time T generally occurs with low probability, especially if A is large as would be preferable in the circumstances. In such cases, the quadratic exponential approximation to $s_A^{(T)}(x)$ is prone to reach its turning point within the interval [0, 1], and the approximation is not reliable beyond this point (recall Figure 7 on page 121). The first-order estimates of extinction probability must therefore be used instead, and the whole method reduces to the first-order likelihood approximation. Attempts to use the second-order method of likelihood estimation were unsuccessful for this reason. Whether or not this *ad-hoc* means of incorporating the correlation structure of the observations at time T into the likelihood calculation would yield improved parameter estimates remains an open question.

5.5 Summary of the analytic approach to the colonization model

The material of this chapter and the procedure for parameter estimation in the colonization model is now summarized. The parameters $\boldsymbol{\theta}$ enter through the site-scale colonization probabilities, $\{p_{ih}\}_{i,\,h=1,\ldots,N}$, where p_{ih} is the probability that site h is colonized at the end of any time-step by ancestors in site i at the beginning of the time-step. The likelihood $L\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(0)}}\right)$ of the observed distribution at the final time T, conditional on the observed distribution at the initial time 0 and the parameters $\boldsymbol{\theta}$, must be obtained as a function of $\boldsymbol{\theta}$. Once obtained, the likelihood can be maximized with respect to $\boldsymbol{\theta}$ to give the maximum likelihood parameter estimates.

When survey data is not available at every time-point, calculation of the likelihood function

for the colonization model is problematic, since an evaluation of the probabilities of all possible colonization paths from the initial occupied sites to the final occupied sites is computationally infeasible. However, if the ancestry of every occupation is well-defined, it is straightforward to obtain the probability of extinction in any set of sites at any time-point, because extinction demands that colonization should have failed along all possible ancestry paths. In the original colonization model, there is no well-defined notion of ancestry: a site is simply occupied or unoccupied, and there is no framework for recognizing that an occupied site at time t might have ancestry in several sites at time $t-1$.

This motivated the idea of regarding the colonization model as a modified branching process, in which ancestry is clearly preserved over all time-steps. There is no approximation involved at this stage: if the branching process is modified appropriately it reproduces the occupation probabilities of the colonization model exactly. The underlying mechanisms of the two conceptual models, however, are different, and that of the modified branching process facilitates the likelihood calculation through the determination of extinction probabilities.

Associated with the modified branching process formulation is the idea of colonies within sites: each colony has a unique ancestry and acts independently of every other colony in the site. In order to emulate the original colonization model, the colonies in site i at time t, acting together but independently, must have overall probability p_{ih} of colonizing site h at time $t+1$. For this to be possible, each colony in site i at time t must have a reduced probability $q_{ih}^{(t)}$ of colonizing site h at time $t+1$, and the colony-scale colonization probabilities $q_{ih}^{(t)}$ must be calculated in order to obtain the final likelihood. The reduced but independent probabilities at the colony scale mimic the probabilities at the site scale, where colonization paths are not independent but merge and become indistinguishable.

The colony-scale colonization probability $q_{ih}^{(t)}$ was shown in Section 2 to be the unique solution between 0 and p_{ih} of a polynomial with very large order (up to N^t, where N is the number of sites in the survey region). Exact solution of this colony-scale polynomial is infeasible: the coefficients would be extremely time-consuming to compute, and there is no closed-form solution to a general polynomial of such high order. Even if the coefficients of the polynomial were obtained, numerical root-finding would be hampered by the computational expense required for every evaluation of the polynomial, and rounding errors would become significant.

Instead, approximations to the colony-scale polynomial were derived in Section 3, where it

was shown that the required root is approximately the solution x to expressions of the form $\exp(\beta x) = const.$ (first-order or linear exponential approximation), or $\exp(\beta x + \gamma x^2) = const.$ (second-order or quadratic exponential approximation). The coefficients β and γ were derived using Taylor series expansions. In Section 4 it was demonstrated that these approximations are very good within the usual scope of the colonization model.

Finally, the present section has shown via equations (49), (51) and (52) how the colony-scale colonization probabilities $q_{ih}^{(t)}$ are used in calculating extinction probabilities, and thereby how the full likelihood function may be calculated, given particular values of the parameters $\boldsymbol{\theta}$. The approximation to the full likelihood using first-order approximations for all extinction probabilities is precisely the product-likelihood function, while the approximation using second-order extinction probabilities is too expensive to be calculated directly. A possible alternative second-order estimation of the likelihood is unlikely to be successful in implementation.

Once the approximate likelihood function can be calculated for given parameter values, a numerical maximization technique may be used to maximize the likelihood over the parameter space and thereby find the maximum likelihood parameter estimates. In practice, it is preferable to minimize the negative logarithm of the likelihood function. The minimization will be performed in this chapter using the downhill simplex method from Nelder & Mead (1965), as implemented in Numerical Recipes in C (Press *et al.* 1988), although many other algorithms could be used (Press *et al.* 1988). If there are several surveys at irregular spacing, $0, T_1, \ldots, T_n$, the likelihood is obtained for each consecutive pair of surveys and the overall likelihood is the product of the results:

$$L\left(\boldsymbol{y}^{(T_n)}, \ldots, \boldsymbol{y}^{(T_1)} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y}^{(0)}\right) = L\left(\boldsymbol{y}^{(T_n)} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y}^{(T_{n-1})}\right) \times \ldots \times L\left(\boldsymbol{y}^{(T_1)} \,\middle|\, \boldsymbol{\theta},\, \boldsymbol{y}^{(0)}\right).$$

6 Implementation and Examples

6.1 Computational details

There are two aspects to the implementation of the branching-process approach to the fitting of the colonization model, namely the calculation of first and second-order approximations to the colonization probabilities. The first-order approximations are straightforward and quick to compute, but the second-order approximations are more difficult to obtain and involve the multiplication of large matrices.

First-order approximations

In order to calculate the first-order approximation to the likelihood given by equation (56), only the estimated probabilities $p_i^{(T)} = \mathbb{P}\left(N_i^{(T)} > 0\right)$ are required. Let the vector $\boldsymbol{u^{(t)}}$ be defined for $t \geq 1$ as

$$\boldsymbol{u^{(t)\prime}} = \boldsymbol{y^{(0)\prime}} Q^{(0)} \ldots Q^{(t-1)},$$

where a dash indicates vector transpose. By (25) and (40), the first-order approximation to $p_i^{(T)}$ is $\left(1 - \exp(-u_i^{(T)})\right)$, where $u_i^{(T)}$ is the ith element of $\boldsymbol{u^{(T)}}$. To calculate the approximate likelihood, therefore, it is sufficient to find $\boldsymbol{u^{(T)}}$. Now it is clear that $\boldsymbol{u^{(t+1)\prime}} = \boldsymbol{u^{(t)\prime}} Q^{(t)}$ for any t. By (38), the first-order approximation to the elements of $Q^{(t)}$ is given by

$$q_{ih}^{(t)} \simeq -\left(\frac{\log(1 - p_{ih} p_i^{(t)})}{u_i^{(t)}}\right) = -\left(\frac{\log\left(1 - p_{ih}\left(1 - \exp\left(-u_i^{(t)}\right)\right)\right)}{u_i^{(t)}}\right).$$

Thus

$$u_h^{(t+1)} = \sum_{i=1}^{N} u_i^{(t)} q_{ih}^{(t)} = -\sum_{i=1}^{N} \log\left\{1 - p_{ih}\left(1 - \exp\left(-u_i^{(t)}\right)\right)\right\}. \tag{58}$$

The calculation of $\boldsymbol{u^{(T)}}$, and hence the likelihood, therefore proceeds by calculating $\boldsymbol{u^{(1)\prime}} = \boldsymbol{y^{(0)\prime}} Q^{(0)}$ and using (58) to recursively obtain $\boldsymbol{u^{(2)\prime}}, \ldots, \boldsymbol{u^{(T)\prime}}$. The whole procedure is extremely straightforward and does not involve any matrix multiplications. The ease with which the calculation may be performed ensures that there is no practical limit on the length T of survey period that can be accommodated. With very long survey periods, however, the first-order approximation will lose some of its effectiveness.

Second-order approximations

Second-order approximations may be used in the calculation of the quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$ for all sites i, h and times $t < T$. Calculation of the likelihood function can proceed in two different ways. The first possibility is to estimate the final occupation probabilities $p_i^{(T)}$ using the first-order approximations, and calculate the first-order approximation to the likelihood as above. The second possibility is to use the second-order approximations in estimating the final occupation probabilities $p_i^{(T)}$, and to multiply them together as a pseudo-estimate of the likelihood (the product-likelihood function). Both of these approaches will be implemented.

Let the vector $\boldsymbol{z}^{(t)}$ be defined for $t \geq 1$ as

$$\boldsymbol{z}^{(t)\prime} = \boldsymbol{y}^{(0)\prime} Q^{(0)} \ldots Q^{(t-2)} M^{(t-1)},$$

where a dash indicates vector transpose and $\boldsymbol{z}^{(1)\prime} = \boldsymbol{y}^{(0)\prime} M^{(0)}$. Clearly, $\boldsymbol{z}^{(t+1)\prime} = \boldsymbol{u}^{(t)\prime} M^{(t)}$ for all t. Expressions for the quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$ are given in equations (39) and (40), and involve the calculation of $g'(0)$ and $g''(0)$ in the notation of Section 3, where $g(x) = \log s_i^{(t)}(x)$ and is particular to the site i and time t. From (25), $g'(0) = -u_i^{(t)}$. Calculation of $g''(0)$ is more involved, but it is readily shown by rearranging (34) that the following expressions hold for a fixed site i:

$$\begin{aligned}
t=1: \quad g''(0) &= -\boldsymbol{z}^{(1)\prime}[i] \\
t=2: \quad g''(0) &= -\boldsymbol{z}^{(1)\prime} M^{(1)}[i] \\
t=3: \quad g''(0) &= \left\{ \left(\boldsymbol{u}^{(1)\prime} - \boldsymbol{z}^{(1)\prime}\right) M^{(1,2)} - \boldsymbol{z}^{(2)\prime} M^{(2)} \right\}[i] \\
t=4: \quad g''(0) &= \left\{ \left(\boldsymbol{u}^{(1)\prime} - \boldsymbol{z}^{(1)\prime}\right) M^{(1,2,3)} + \left(\boldsymbol{u}^{(2)\prime} - \boldsymbol{z}^{(2)\prime}\right) M^{(2,3)} - \boldsymbol{z}^{(3)\prime} M^{(3)} \right\}[i] \\
&\vdots \\
t=T: \quad g''(0) &= \left\{ \left(\boldsymbol{u}^{(1)\prime} - \boldsymbol{z}^{(1)\prime}\right) M^{(1,\ldots,T-1)} + \ldots + \left(\boldsymbol{u}^{(T-2)\prime} - \boldsymbol{z}^{(T-2)\prime}\right) M^{(T-2,T-1)} \right. \\
&\qquad \left. - \boldsymbol{z}^{(T-1)\prime} M^{(T-1)} \right\}[i].
\end{aligned}$$

Square brackets $[i]$ are used to denote the ith element of any vector. The vectors $\boldsymbol{u}^{(t+1)}$ and $\boldsymbol{z}^{(t+1)}$ are easily determined from $\boldsymbol{u}^{(t)}$ and involve only the multiplication of a vector by a matrix. The expense in the calculation of $g''(0)$ for each time t is due to the computation of the matrices $M^{(1,\ldots,t-1)}, \ldots, M^{(t-2,t-1)}$. There is no means of calculating the matrix $M^{(t_1,\ldots,t_n)}$ other than performing the full matrix multiplication $Q^{(t_1)} \ldots Q^{(t_n)}$. Since each matrix $Q^{(t)}$ has dimensions $N \times N$, where the number of survey sites N is typically a few hundred, these matrix multiplications are expensive both in terms of computer time and storage. Furthermore, for $T \geq 3$, the total number of matrix multiplications required to calculate $g''(0)$ for all $t \leq T$ is $(T-1)(T-2)/2$, which can be prohibitively time-consuming on a large grid. Note that the matrix products need not be calculated again for different sites i, but they do need to be re-calculated for every set of parameter values involved in the maximization process.

6.2 Application to the woodlark data

The woodlark data described in Chapter 3 is now analysed using the techniques developed in this chapter. The dataset consists of five years of complete presence/absence data and habitat suitability information, covering a total of 497 sites. Since there are no years in which data is genuinely missing, the methodology may be tested by omitting some of the surveys and comparing the results against those obtained when all surveys are included. In view of the rapidly changing habitat suitability in each site over the 4-year period, the habitat data will be retained for the purposes of calculating $q_{ih}^{(t)}$, even for those t where the corresponding survey data is omitted. Although it is unlikely in practice that habitat data is available for years in which no survey was conducted, omission of this information would severely bias the model performance in the woodlark case. It is envisaged that the colonization model will more usually be applied in circumstances where habitat suitability changes little between surveys.

The changing habitat suitability for each site over the 4-year period introduces time-dependence to the site-scale colonization probabilities $\{p_{ih}\}$, which should now be written $\left\{p_{ih}^{(t)}\right\}$ to represent the probability that a colonization occurs from site i at time t to site h at time $t+1$. The model for the site-scale colonization probabilities is $p_{ih}^{(t)} = p_0 \exp\left(-a\,\delta_{ih} - b\,\varsigma_h^{(t+1)}\right)$, where $\varsigma_h^{(t+1)}$ is the transformed suitability score for site h at time $t+1$. The theory of the preceding sections may be applied unchanged when the site-scale colonization probabilities are time-varying, with the exception that all occurrences of p_{ih} are replaced by $p_{ih}^{(t)}$. The aim of the model-fitting process is the estimation of the unknown parameters a, b and p_0. Throughout this section, the year 1986 corresponds to time $t = 0$, so that $\boldsymbol{y^{(0)}} = \boldsymbol{y^{(1986)}}$ and so on.

The material of the previous sections suggests a number of different ways of obtaining the final model fit when there is missing survey data, for example using different levels of approximation. In addition, it is instructive to compare the fits obtained by including different subsets of the five available surveys. Eight fits will be provided in total, following the example of Carroll (1876). These are described as follows.

Fit 1: full model. All five years of survey data included. The overall likelihood is

$$L\left(\boldsymbol{y^{(4)}}, \boldsymbol{y^{(3)}}, \boldsymbol{y^{(2)}}, \boldsymbol{y^{(1)}}, \left|\, \boldsymbol{\theta}, \boldsymbol{y^{(0)}}\right.\right) = L\left(\boldsymbol{y^{(4)}} \left|\, \boldsymbol{\theta}, \boldsymbol{y^{(3)}}\right.\right) L\left(\boldsymbol{y^{(3)}} \left|\, \boldsymbol{\theta}, \boldsymbol{y^{(2)}}\right.\right) \times L\left(\boldsymbol{y^{(2)}} \left|\, \boldsymbol{\theta}, \boldsymbol{y^{(1)}}\right.\right) L\left(\boldsymbol{y^{(1)}} \left|\, \boldsymbol{\theta}, \boldsymbol{y^{(0)}}\right.\right),$$

and no approximations are required.

Fit 2: one-step model. Only years 1986 and 1990 included. The four-year period is regarded as a single time-step, so that ostensibly there is no missing data. The single-step site-scale colonization probability from site i at time $t = 0$ (year 1986) to site h at time $t = 4$ (year 1990) is $p_{ih} = p_0 \exp\left(-a\,\delta_{ih} - b\,\varsigma_h^{(1990)}\right)$. No approximations are required.

Fit 3: 1st order. Only years 1986 and 1990 included. A four-step model is applied, with $p_{ih}^{(t)} = p_0 \exp\left(-a\,\delta_{ih} - b\,\varsigma_h^{(t+1)}\right)$ for $t = 0, \ldots, 3$. The modified branching process technique is employed, and first-order (linear exponential) approximations are used to obtain all quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$. The final estimate of the likelihood is the first-order approximation to the full likelihood, obtained by multiplying together the first-order estimates of the probabilities of the observations for 1990 as described in section 5.3.

Fit 4: 1st order (modified). As for Fit 3, with the exception that all approximations for years $t = 1$ and $t = 2$ use the modified approximation $\prod_{k=1}^{M}(1 - \alpha_k x) \simeq \left(1 - \frac{S_1}{M}x\right)^M$ suggested by equation (43) on p. 124, where S_1, S_2, α_k and M are defined in section 4.1.

Fit 5: 2nd order. As for Fit 3, but with second-order (quadratic exponential) approximations used for all quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$ for $t < T = 4$. The first-order approximations are used in the calculation of $p_i^{(T)} = p_i^{(4)}$, and the first-order approximation to the full likelihood is employed as in Fit 3 and Fit 4.

Fit 6: 2nd order (pseudo). As for Fit 5, except that the second-order approximations are used in the calculation of the final-year occupation probabilities $p_i^{(T)} = p_i^{(4)}$ as well as for all the other quantities. The final likelihood estimate is a pseudo-estimate of the true likelihood — that is, the product-likelihood, obtained by multiplying together the second-order estimates of the probabilities of the observations for 1990 as described in section 5.3.

Fit 7: 2-stage. Years 1986, 1989 and 1990 included. The joint likelihood function is given by

$$L\left(\boldsymbol{y}^{(4)}, \boldsymbol{y}^{(3)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right) = L\left(\boldsymbol{y}^{(4)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(3)}\right)\, L\left(\boldsymbol{y}^{(3)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right)$$

and $L\left(\boldsymbol{y}^{(3)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right)$ is estimated using the first-order approximation to the full likelihood and first-order (linear exponential) approximations for all relevant quantities. The single-step likelihood $L\left(\boldsymbol{y}^{(4)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(3)}\right)$ may be calculated directly. This example illustrates the situation where a number of surveys have been taken at irregular spacing.

Fit 8: arbitrary parameters. Only years 1986 and 1990 included. Arbitrary parameter values are used to generate model predictions which may be compared against those

obtained using the maximum likelihood estimates of Fits 1–7. The selected parameters are those from Simulation 1 of section 4.2, namely $a = 1.0$, $b = 2.0$, $p_0 = 0.40$.

The estimated likelihood function was calculated and maximized using software written in the programming language C (Kernighan & Ritchie 1978) and S-PLUS (Statistical Sciences Inc. 1993), version 3.4 for UNIX. The maximization was performed using the function `amoeba` from Press *et al.* (1988). The C routines were dynamically loaded into S-PLUS so as to combine the efficiency of compiled C code with the object permanency, data-manipulation facilities and graphics interface of S-PLUS. However, only slight modification to the code would be required in order to make the programs stand-alone C. Some additional programs were written in PASCAL (Findlay & Watt 1985) to interface with the best attainable match algorithm of Chapter 4.

Approximate timings for a single evaluation of the likelihood on a Sun Ultra 10 server with 192 MB RAM were

	Time (seconds)
Fit 1:	0.2
Fit 2:	0.05
Fit 3:	0.8
Fit 4:	0.8
Fit 5:	40
Fit 6:	111
Fit 7:	0.6

A typical maximization might involve between 60 and 200 evaluations of the likelihood. The use of the second-order approximations in Fit 5 clearly causes a massive increase in computer time, and the time is more than doubled again when second-order approximations are also used in the final year.

The parameter estimates from Fits 1 to 8 are presented in Table 6. The first point to note is that the parameter estimates from Fit 2 (one-step model) are not directly comparable with those from the other fits, because the parameters pertain to a four-year interval rather than a one-year interval. Once Fits 2 and 8 are excluded, the other parameter estimates are extremely close. This is especially true of Fits 3, 4, 5, and 6, which involve the same model and the same years of data but differences in the approximations used in the calculation of the quantities $q_{ih}^{(t)}$, $p_i^{(t)}$ and $L\left(\boldsymbol{y}^{(4)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right)$. However, the results from these fits were also close to those obtained from Fits 1 and 7, which include more of the available survey data. This is encouraging evidence that the methodology for coping with missing years is effective.

Method of fit	Parameter estimates			Mean BAM results for 1990 from 1000 simulations		
	a	b	p_0	match	unmatched 0s	unmatched 1s
1: full model	0.170	1.83	0.0314	254.0	37.3	24.6
2: one-step model	0.122	2.64	0.0800	253.8	31.3	30.9
3: 1st order	0.117	2.55	0.0367	253.6	31.5	30.9
4: 1st order (modified)	0.117	2.56	0.0368	253.7	31.2	31.1
5: 2nd order	0.117	2.57	0.0368	253.9	31.4	30.7
6: 2nd order (pseudo)	0.117	2.57	0.0369	253.6	31.3	31.1
7: 2-stage	0.123	1.71	0.0225	252.5	41.2	22.4
8: arbitrary parameters	1.00	2.00	0.400	234.0	27.8	54.2

Table 6: Parameter estimates and mean BAM scores from the eight different methods of fit. The last three columns give respectively the mean best attainable match, the mean number of mismatched 0s in the predicted distributions after swaps have been performed, and the mean number of mismatched 1s in the predicted distributions after swaps have been performed. Means are taken from 1000 predictions using the parameters shown.

Comparison between the parameter estimates obtained for Fits 3, 4, 5, and 6 is particularly interesting, because it demonstrates the effect on the results of progressively improving the approximations in the hidden stages of model-fitting. In the case of the woodlark data, improvements in the approximations appear to have had a negligible effect on the final outcome. This suggests that further improvements, to third order and deeper, would similarly have little effect and provides considerable assurance that the approximations do not cause bias in the results. At least for this dataset, the basic first-order fit appears to be adequate: the second-order fits clearly do not exhibit sufficient improvement to justify the increased amounts of time and memory required to obtain them. Note, however, that these fits involve only 3 missing years of survey data; improvements may well be more dramatic when there are longer periods of missing years.

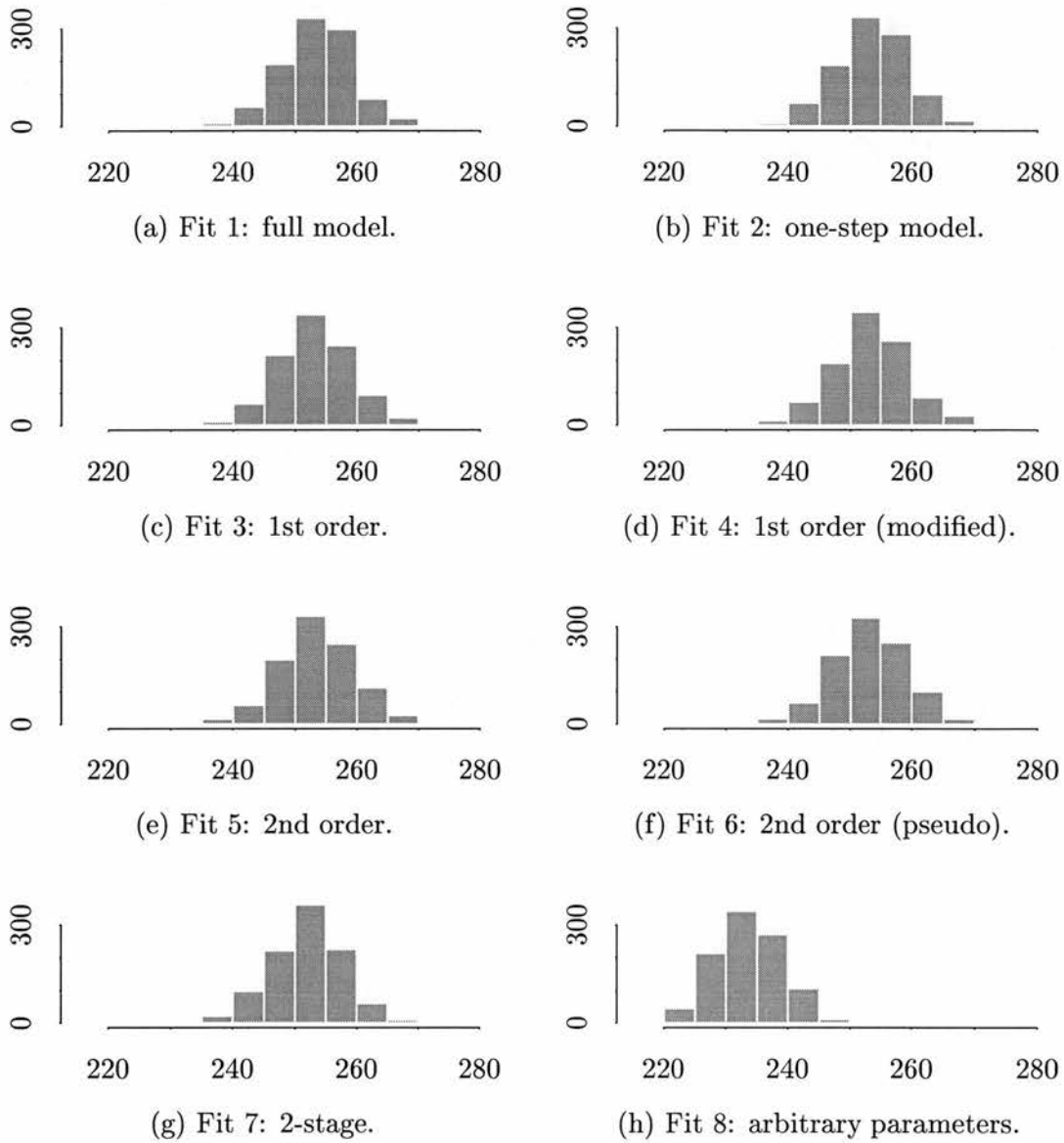
As a point of interest, the values of the log-likelihood obtained at the maximum likelihood

parameter values from Fits 3, 4, 5, 6 and 8 were respectively -131.664, -131.663, -131.664, -131.636 and -223.195. Fits 1 and 7 should be expected to have a lower log-likelihood as they incorporate more years of survey data, and the respective values were -470.880 and -276.523. The one-step log-likelihood was -130.813, which is comparable with the results from Fits 3, 4, 5, 6 and 8 because it involves the same years of survey data; however, since it arises from a different model it is interesting that the value is so close.

Table 6 also contains details of best attainable match results using predictions from the various fitted models. For each of the parameter estimates in Fits 1 to 8, 1000 predicted distributions were simulated for 1990 and compared against the true 1990 distribution. The best attainable match was evaluated, as described in Chapter 4. The clique of any site was chosen to be the set of sites within a 0.5 km radius of that site, as biological evidence suggests that it is uncommon for male woodlarks to move more than 0.5 km between territories in successive years (Bowden & Green 1992). For each simulation the following results were recorded: total number of matched sites after swaps had been performed; number of 0–predictions in the simulated distribution that remained unmatched after swaps had been performed; and number of 1–predictions in the simulated distribution that remained unmatched after swaps had been performed. The mean of each of these quantities over the 1000 simulations is given in Table 6, and histograms of the match results are shown in Figure 21. The results pertain to $N = 316$ sites in 1990: the remaining 181 sites in the distribution were unsuitable for woodlarks in that year and were automatically allotted zero status in the predicted distributions (corresponding to $\varsigma_h^{(1990)} = \infty$ for site h).

Table 6 and Figure 21 indicate that there is negligible difference between the BAM results for the first seven fits. Those fits that involve missing years have therefore attained a standard of prediction that is just as high as that from the fits that use all of the data. Encouragingly, the arbitrary parameters of Fit 8, which have a substantially lower associated likelihood, give appreciably worse BAM results. The BAM values for good parameter estimates should ideally be as high as possible, up to a maximum of 316, while the numbers of mismatched 0s and 1s should be as low as possible; however, it is also important that there should be roughly even numbers of mismatched 0s and 1s. If all the mismatched sites had the same status, for example, the model predictions would clearly be biased.

The parameters from Fits 2, 3, 4, 5, and 6 all yield approximately equal numbers of mismatched 0s and 1s. The imbalances in Fits 1 and 7 may be attributed to the fact that these parameter estimates must accommodate data from years other than 1990, so a compromise fit that is better for the overall data but worse for 1990 alone is obtained.



(a) Fit 1: full model.

(b) Fit 2: one-step model.

(c) Fit 3: 1st order.

(d) Fit 4: 1st order (modified).

(e) Fit 5: 2nd order.

(f) Fit 6: 2nd order (pseudo).

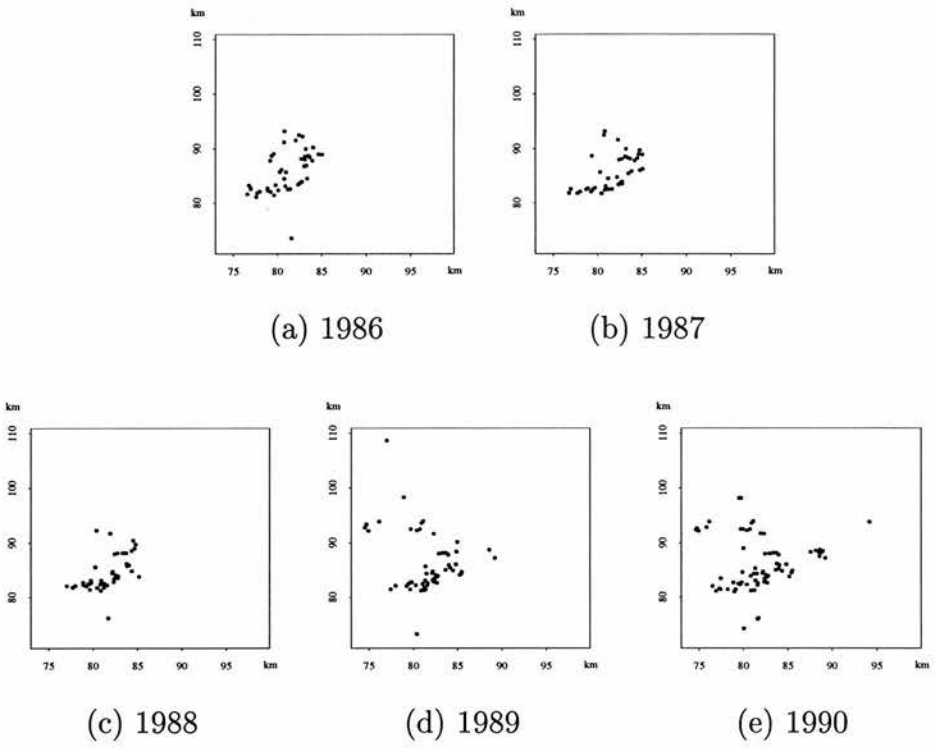
(g) Fit 7: 2-stage.

(h) Fit 8: arbitrary parameters.

Figure 21: Best attainable match results from predicted distributions for 1990, using the parameters from eight different fitting methods. Each histogram shows BAM results from 1000 simulated distributions for 1990, generated from the parameters indicated. The number of sites being compared is $N = 316$.

Unsurprisingly, both of these fits tend to underpredict presence for 1990, due to the fact that the occupation count in 1990 was substantially higher than that in previous years.

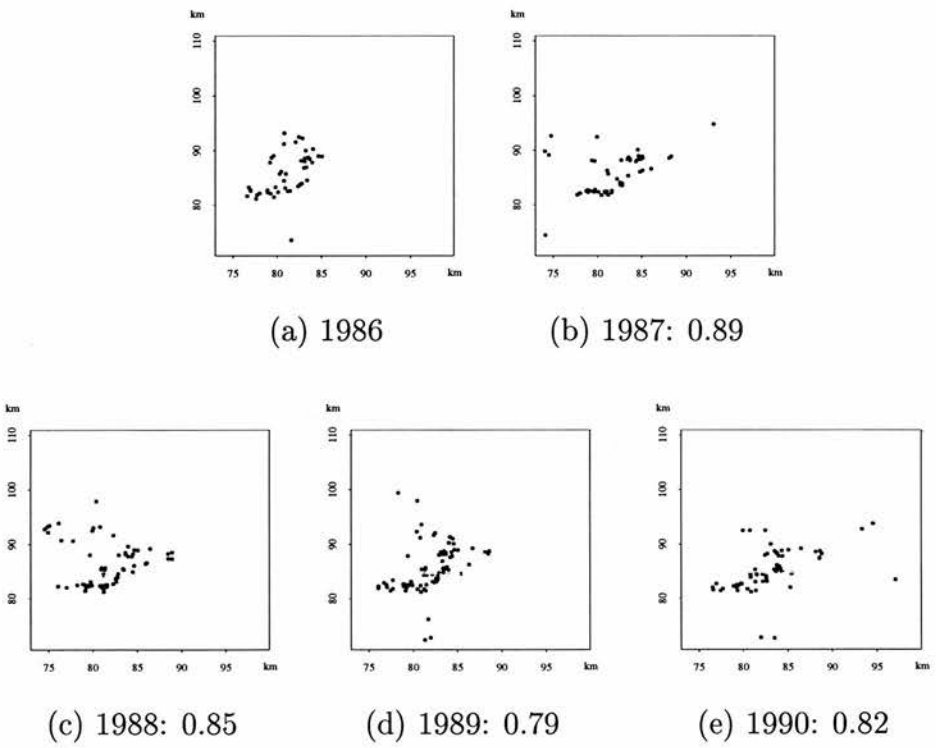
Figure 22 shows the true woodlark distribution in Thetford Forest from 1986 to 1990. Only occupations are shown, so that the full pattern may be seen more clearly. Figures 23, 24 and 25 each show a single prediction from the parameters of Fits 1, 2, and 6 respectively. Each of the three predictions capture much of the observed pattern. BAM coefficients, given by the best attainable match of the prediction against the true distribution, divided by the number of suitable sites in the relevant year, are provided where appropriate.



(a) 1986 (b) 1987

(c) 1988 (d) 1989 (e) 1990

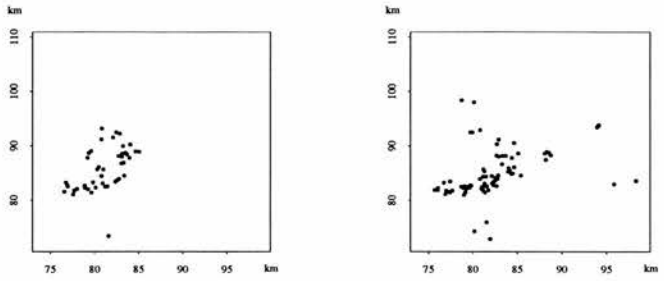
Figure 22: Distribution of woodlark occupations in suitable sites from 1986 to 1990.



(a) 1986 (b) 1987: 0.89

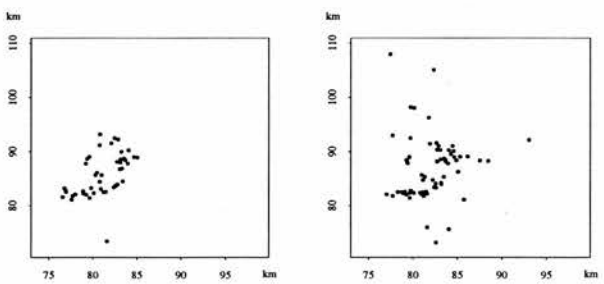
(c) 1988: 0.85 (d) 1989: 0.79 (e) 1990: 0.82

Figure 23: Predicted distribution of woodlark occupations in suitable sites from 1986 to 1990, using the maximum likelihood estimates from Fit 1 (full model). The decimal numbers accompanying the graphs for 1987, 1988, 1989 and 1990 are the BAM coefficients for the associated predictions.

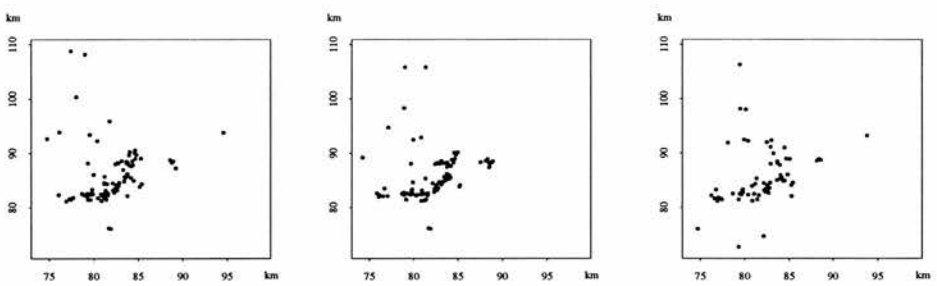


(a) 1986 (b) 1990: 0.82

Figure 24: Predicted distribution of woodlark occupations in suitable sites in 1986 and 1990, using the maximum likelihood estimates from Fit 2 (one-step model). The decimal number accompanying the graph for 1990 is the BAM coefficient for this prediction.



(a) 1986 (b) 1987: 0.86



(c) 1988: 0.84 (d) 1989: 0.77 (e) 1990: 0.82

Figure 25: Predicted distribution of woodlark occupations in suitable sites from 1986 to 1990, using the maximum likelihood estimates from Fit 5 (2nd-order). The decimal numbers accompanying the graphs for 1987, 1988, 1989 and 1990 are the BAM coefficients for the associated predictions.

Diagnostic plots

As a final check on the performance of the fitting procedure when there are missing years of survey data, diagnostic plots are provided as recommended in section 4.2. These plots give the upper bounds on the error functions $\epsilon_1(x)$ and $\epsilon_2(x)$ for the linear and quadratic exponential approximations at times $t = 1$ and $t = 2$, subject to the constraints described in section 4.1. Since the likelihood maximization routine involves evaluations of the likelihood at points throughout the parameter space, it is important to be sure that the approximations are universally good. Quadratic constraints were used for both ϵ_1 and ϵ_2, to give the tightest bounds for the linear error.

Plots were obtained for a range of sites and parameter values using the procedure of Fit 5 (2nd order), and those for site 125 are shown in Figure 26. Fit 5 was chosen in order to provide both linear and quadratic errors, and as an example of a fit with the maximum number of missing years. The selected parameter values are shown in Table 7, and all of them are points at which a likelihood evaluation was observed to take place during the

Graph	t	a	b	p_0	S_1	S_2	M
(a)	1	1.00	1.00	0.500	2.12	0.359	45
(b)	1	1.73	1.18	0.287	0.543	0.0385	45
(c)	1	0.309	4.39	0.261	3.40	0.395	45
(d)	1	0.193	6.98	0.194	3.25	0.294	45
(e)	1	0.191	2.92	0.0674	1.30	0.0467	45
(f)	1	0.117	2.57	0.0368	0.929	0.0212	45
(g)	2	1.00	1.00	0.500	6.87	0.147	14400
(h)	2	1.73	1.18	0.287	0.956	0.00876	14400
(i)	2	0.309	4.39	0.261	8.37	0.0415	14400
(j)	2	0.193	6.98	0.194	6.65	0.0205	14400
(k)	2	0.191	2.92	0.0674	2.14	0.00149	14400
(l)	2	0.117	2.57	0.0368	1.31	0.000369	14400

Table 7: Details of the graphs in Figure 26, giving in each case the time t for which the approximations are examined, the parameter values a, b and p_0, and the values S_1, S_2 and M.

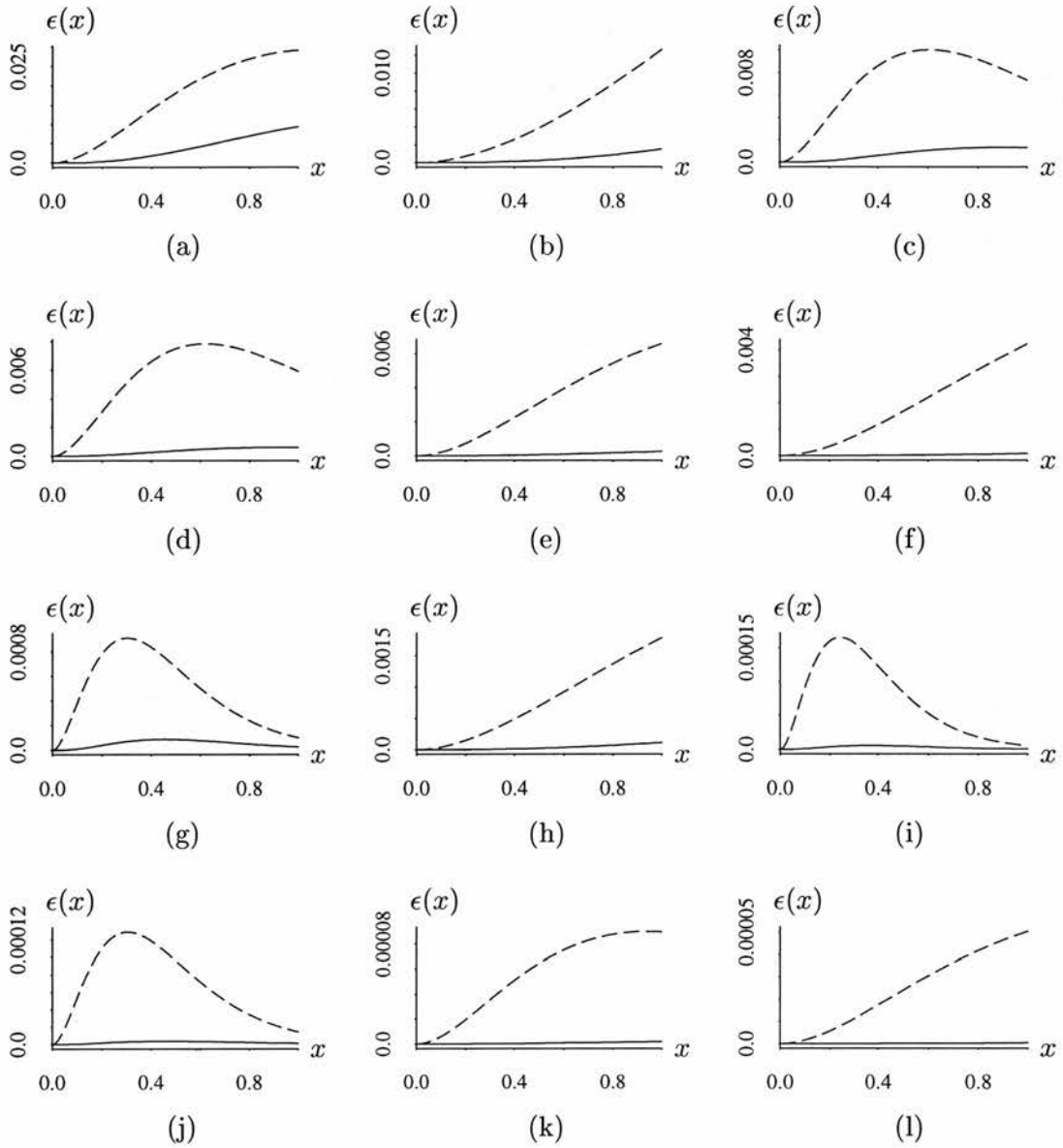


Figure 26: Diagnostic plots for site 125, $t = 1, 2$ and a range of the parameter values encountered during the maximization procedure. The solid lines show the upper bounds for the quadratic error $\epsilon_2(x)$, subject to the three quadratic constraints of Section 4. The dashed lines show the upper bounds for the linear error $\epsilon_1(x)$, subject to the same (quadratic) constraints. The particulars of each curve are given in Table 7.

maximization routine. The first set, corresponding to graphs (a) and (g), was the point chosen for the start of the maximization routine, and the final set (graphs (f) and (l)) is the point of convergence: the maximum likelihood estimates for Fit 5. Also shown in Table 7 for each set of parameter values are the quantities S_1, S_2 and M defined in section 4.1. The error functions ϵ_1 and ϵ_2 are fully determined by these values.

From Figure 26 it is clear that even the upper bounds of the errors are small throughout the parameter space, especially at time $t = 2$ and for parameter values close to the final

maximum likelihood estimate. None of the plots give an unacceptably high upper bound, and there is certainly little cause for concern in a problem involving a few hundred sites when the error in estimating a probability is at most 0.01, as for all plots except (a). Indeed, a large sample would be required to detect an error of 0.01 in any estimated probability. The diagnostic plots obtained for other sites were even better than those for site 125.

Conclusions

The methods developed in this chapter have worked very effectively with the woodlark dataset. The errors in the hidden approximations have been shown to be small, even at the upper bounds. Improvement of the approximations from first-order to second-order has little effect, suggesting that the first-order approximations may be used confidently with these data. This is convenient, since the computations become much more expensive in time and memory when second-order approximations are used. It is likely that the success of the first-order approximations with the woodlark dataset is due in part to the small number of missing surveys.

The parameter estimates when survey data is missing are reasonably close to those incorporating the full data. For example, the log-likelihood of the full model in Fit 1, evaluated at the parameters of Fit 3 (1st order), is -477.370, while the maximized value is -470.880. These likelihoods are obtained as the product of the probabilities for 1327 observations (the total number of suitable sites over the period 1987 to 1990), and the parameters of Fit 3 therefore yield probabilities that are on average 0.995 times those stemming from Fit 1 — a value very close to unity. Similarly, the log-likelihood associated with Fit 3, evaluated at the parameters of Fit 1, is -135.914 — close to the maximized value of -131.664. For the 316 observations in 1990, the parameters of Fit 1 yield probabilities that are on average 0.987 times the optimum values from Fit 3. These results suggest two things: firstly that the methodology for missing years gives plausible results, and secondly that the colonization model itself is reasonably well suited to the woodlark data. If the colonization model were ill-suited to the woodlark data, there would be no reason to expect the fits that incorporate only some of the data to be similar to those incorporating all of it.

Linked to this is the interpretation of the predicted distributions for intermediate years 1987, 1988 and 1989 from Fits 1 and 5 in Figures 23 and 25 respectively. By coincidence, the two predictions shown yield exactly the same BAM coefficient (0.82) for 1990, and it

is to be expected that they should perform equally well for 1990. The parameters from the full model (Figure 23) should also be expected to perform well on the intermediate distributions 1987, 1988 and 1989; but the parameters from Fit 5 would only do so if the colonization model itself were a reasonable representation of reality. The fact that the BAM coefficients from the intermediate distributions of Figure 25 are only marginally less than those of Figure 23 provides some evidence that this is the case. Rigorous determination of the suitability of the colonization model for these data, however, is not the primary concern of this section.

Finally, the maximum likelihood parameters of the one-step model (Fit 2) provide predictions which are surprisingly similar to those from the 4-step model of Fits 1, 3, 4, 5, 6 and 7; the similarities are apparent from Table 6 and Figures 21 and 24. This suggests that there is little problem with the application of the one-step model for these data: the assumption of independence of colonizations of and from sites over the 4-year time step is stronger than the assumption of independence over a one-year time step, but does not appear to have adversely affected the predictive accuracy of the model.

If the colonization model is appropriate for the woodlark data, information about the dependence of colonization probabilities on distance between sites and habitat quality at the target site should be contained in the parameter estimates. Figure 27 summarizes this dependence for the parameter estimates of Fit 3 (1st order). The decrease of colonization probability with deteriorating site suitability is more marked than the decrease of colonization probability with increasing distance. At $\delta_{ih} = 0$, for example, p_{ih} falls by a factor of 2 as ς_h ranges from 0.0 to 0.27: this represents a sharp fall-off since only 35 sites have a suitability score of 0.27 or below, out of the 316 sites that have suitable habitat for woodlarks in 1990. On the other hand, at optimal site suitability ($\varsigma_h = 0$), p_{ih} falls by a factor of 2 as the distance δ_{ih} ranges from 0 km to 6 km. This fall-off is less steep, since on average there are approximately 76 sites within 6 km of any site. The results suggest that habitat is the dominant factor in woodlark occupation: as long as suitable habitat is created, its proximity to the current woodlark occupation is of secondary importance.

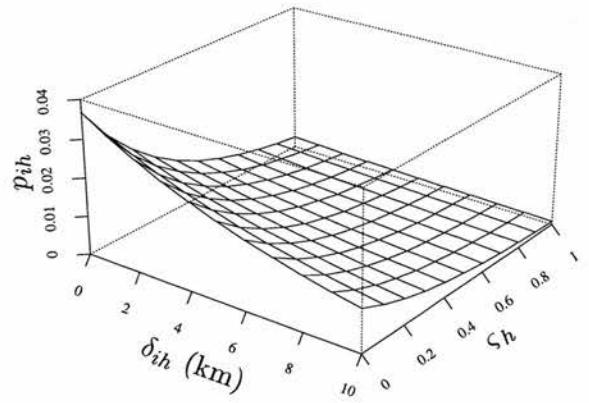


Figure 27: Dependence of colonization probability p_{ih} on distance δ_{ih} between sites i and h and habitat suitability score ς_h of site h, at the maximum likelihood parameter estimates of Fit 3. The relationship is given by $p_{ih} = p_0 \exp(-a\delta_{ih} - b\varsigma_h)$, where $a = 0.117$, $b = 2.55$ and $p_0 = 0.0367$.

The low dependence of colonization probability on distance between sites might seem strange, given that male woodlarks are unlikely to move more than 0.5 km between territories occupied in successive years (Bowden & Green 1992). The longer range colonizations are accounted for, however, by young woodlarks establishing territories for the first time. Figure 28 shows distances between the site of hatching and the site of first known territory for 32 young woodlarks marked with coloured rings, reproduced from Bowden & Green (1992). Colonization distances of 6 km and more are not at all unusual, and the figure helps to give credence to the parameter estimates obtained from the colonization model.

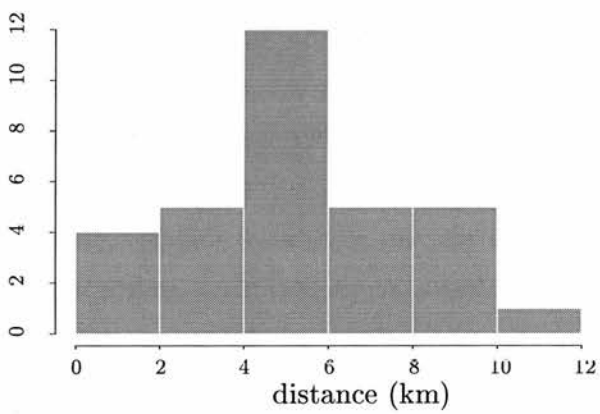


Figure 28: Histogram from Bowden & Green (1992) showing natal dispersal of young woodlarks marked with coloured rings. The histogram shows the distribution of distances moved between the site of hatching and the centre of the first territory on which the bird was sighted in a year subsequent to that of hatching. Reproduced by kind permission of Dr R. E. Green.

6.3 Application to simulated data

The second application uses data that were simulated directly from the colonization model. The parameters, habitat suitability scores and initial distribution were exactly the same as those in Simulation 1 of section 4.2: that is, $a = 1.0$, $b = 2.0$, $p_0 = 0.40$, $N = 100$ and all sites were occupied at time $t = 0$. A twelve-year simulation was obtained from the model, and the resulting distributions are shown in Figure 29. In the subsequent analyses, a single distribution at time T was selected as the 'observed' survey ($T = 3, \ldots, 12$), and all data from the intermediate years $1, 2, \ldots, T-1$ were regarded as missing. The habitat suitabilities, and therefore the colonization probabilities, were taken to be constant over time.

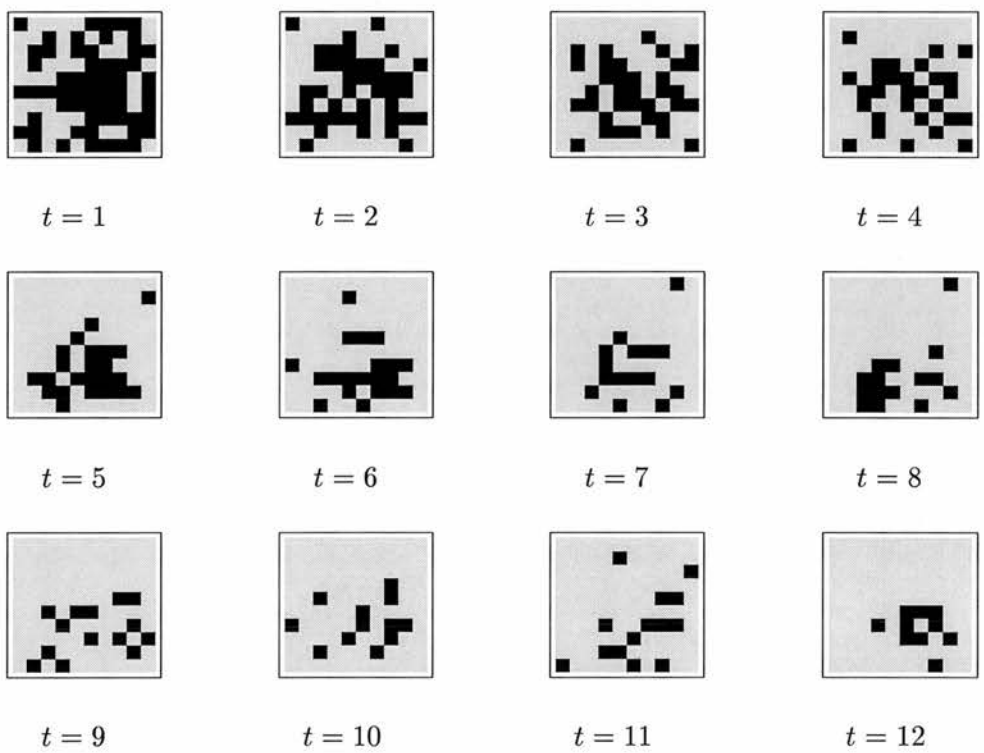


Figure 29: Twelve-year simulation from the colonization model using parameters $a = 1.0$, $b = 2.0$, $p_0 = 0.40$. All sites were occupied at time $t = 0$.

Four fits were obtained for this example:

Fit 1: one-step model. Only years 0 and 12 included. The twelve-year period is regarded as a single time-step, and the single-step colonization probabilities p_{ih} give the probability that site h is colonized at time 12 from ancestors in site i at time 0.

Fit 2: 1st order. Only years 0 and T included; separate fits obtained for $T = 3, \ldots, 12$. A T-step model is applied, and the modified branching process is used with first-order (linear

exponential) approximations for all quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$ $(t \leq T)$. The final likelihood estimate is the first-order approximation to the full likelihood, obtained by multiplying together the first-order estimates of the probabilities of the observations for time T as described in section 5.3.

Fit 3: 2nd order. As for Fit 2, except that second-order (quadratic exponential) approximations are used in the calculation of all quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$ for times $t < T$. First-order approximations are used to calculate the final-year occupation probabilities $p_i^{(T)}$, and the first-order approximation to the full likelihood is used as in Fit 2.

Fit 4: true parameters. The true parameters $a = 1.0$, $b = 2.0$, $p_0 = 0.40$ are used to generate model predictions for comparison against those obtained from the maximum likelihood estimates from Fits 1, 2 and 3.

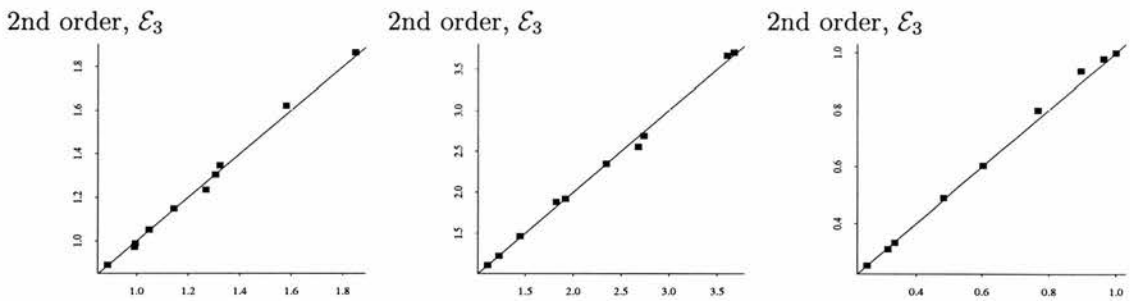
The implementations were carried out using the same software and hardware as in the previous example. Since there are fewer sites involved here, the timings are better and a 12-year fit takes approximately 0.1 seconds per likelihood calculation for Fit 2 (1st order), and 47 seconds for Fit 3 (2nd order). This compares favourably with the 40 seconds required per second-order calculation for a model extending over just four years in the woodlark example.

The purpose of running the procedures from Fits 2 and 3 for $T = 3, \ldots, 12$ is to determine whether the maximum likelihood estimates obtained via the first and second-order approximations exhibit a tendency to drift apart as the number of missing years is increased. Table 8 shows the estimates from Fits 2 and 3 for the ten examples. Although there is greater discrepancy between the two sets of estimates in this example than for the woodlark data, there is no evidence to suggest that this discrepancy widens as the time gap increases. The results are shown graphically for the three parameters in Figure 30, together with the line $y = x$ on which the points would ideally lie. In each case there is little difference between the parameter estimates from Fit 2 and those from Fit 3.

The second salient feature of the results in Table 8 is the variability in parameter estimates between fits. Although the data were generated in a single application of the colonization model, the parameter estimates vary considerably from one value of T to the next and often bear no resemblance to the true values $a = 1.0$, $b = 2.0$, $p_0 = 0.40$. This variability arises because the likelihood function is extremely flat close to the maximum likelihood estimates: there are many parameter values that yield a likelihood value that is almost optimal. For example, under Fit 2 (1st order) with the 12-step model the log-likelihood

	a		b		p_0	
T	1st order	2nd order	1st order	2nd order	1st order	2nd order
3	0.887	0.891	1.11	1.12	0.253	0.254
4	0.995	0.988	1.23	1.23	0.315	0.311
5	1.144	1.149	3.67	3.70	0.962	0.979
6	1.306	1.304	1.92	1.92	0.603	0.604
7	1.048	1.052	2.34	2.35	0.484	0.491
8	1.580	1.620	2.68	2.55	1.000	1.000
9	1.269	1.240	3.60	3.67	1.000	1.000
10	1.849	1.866	1.82	1.88	0.895	0.936
11	0.991	0.972	1.45	1.47	0.335	0.334
12	1.323	1.347	2.74	2.69	0.767	0.798

Table 8: Parameter estimates a, b and p_0 from Fit 2 (1st order) and Fit 3 (2nd order) for T-step models where $T = 3, \ldots, 12$.



(a) Parameter a. (b) Parameter b. (c) Parameter p_0.

Figure 30: Parameter estimates from Fit 2 (1st order) and Fit 3 (2nd order) for T-step models where $T = 3, \ldots, 12$. Each plotted point has coordinates $(\mathcal{E}_2, \mathcal{E}_3)$ where $\mathcal{E}_2$ is the maximum likelihood estimate of the relevant parameter under Fit 2 for one of the times T, and $\mathcal{E}_3$ is the maximum likelihood estimate of the relevant parameter under Fit 3 for the same time T. Also plotted on each graph is the line $y = x$.

is -24.0 at the maximum ($a = 1.323$, $b = 2.74$, $p_0 = 0.767$), and falls only slightly to -25.2 when evaluated at the true parameters $a = 1.0$, $b = 2.0$, $p_0 = 0.40$. Correlations between the parameters a, b and p_0 are also high.

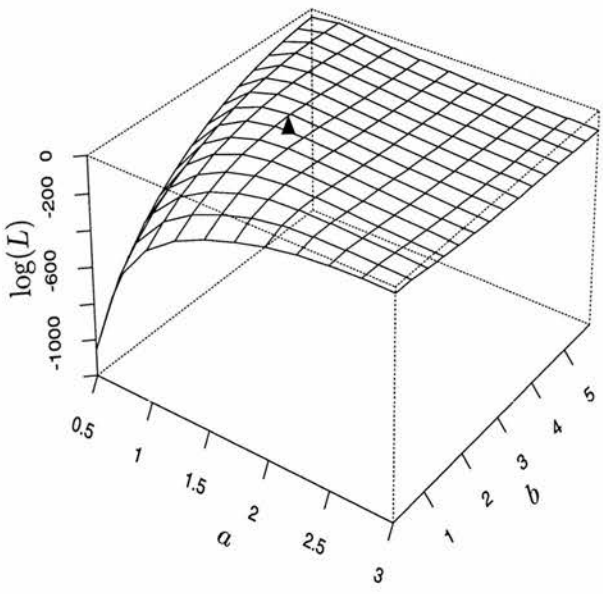
The plots in Figure 31 show the log-likelihood function at the maximum likelihood estimate

of the 12-year model of Fit 2 (1st order). The log-likelihood is particularly flat in the directions of a and b (Figure 31 (a)), when the third parameter p_0 is held constant at its maximum likelihood estimate. There are also marked ridges in the other two plots.

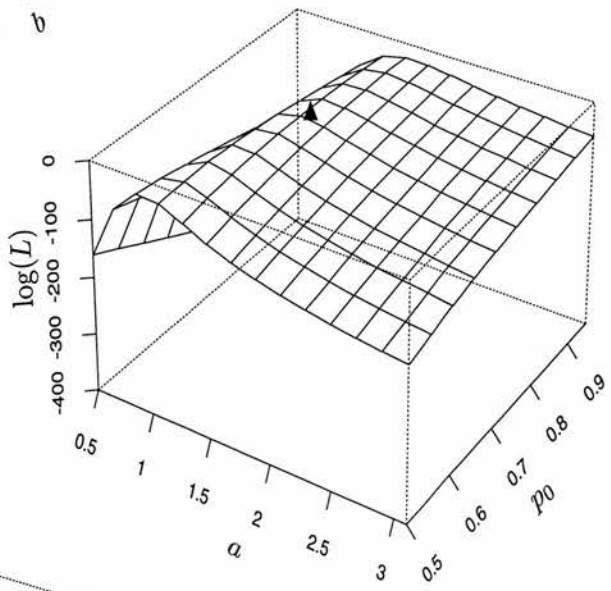
Prediction comparisons from Fits 1–4

The best attainable match algorithm is again used to compare predictions from the various fitting methods. Here, the parameters of Fit 1 (one-step, 12-year) are compared with the 12-step models of Fits 2, 3 and 4, and the results are presented in Table 9 and Figure 32. There is very little to choose between the overall BAM results for Fits 1 to 4; however, the results from Fit 3 (2nd order) are noticeably better balanced in mismatched 1s and 0s than those from Fit 2 (1st order). This suggests that there might be a worthwhile improvement in the second-order approximations over the first-order for these data.

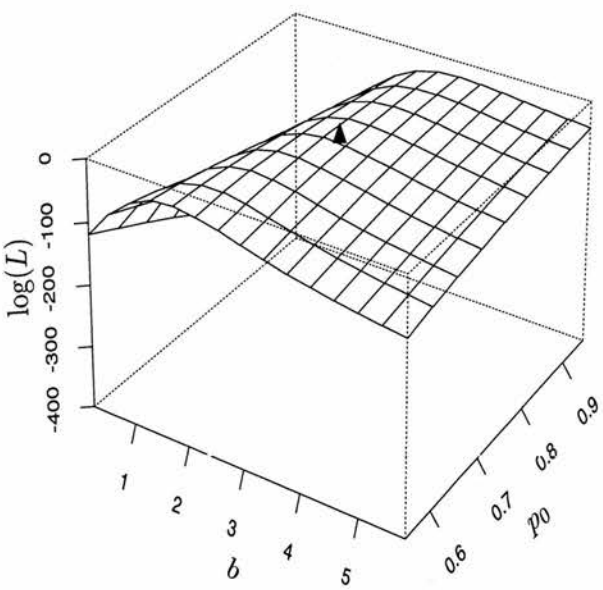
The second-order approximations used in this example did not suffer from the instability described in Figure 7, whereby the turning point of the quadratic exponential approximation to $s_i^{(t)}(x)$ occurs before the point $x = 1$. However, there remains the disadvantage of the time required for the computation of these approximations. The extra effort has brought relatively little reward in this example, although for short time-spans, or when the number of sites N is small, the second-order approximations might provide worthwhile improvements without imposing a disproportionate computational burden. In conclusion, therefore, it is recommended that the second-order approximations be used for the first few hidden years of the fit, and first-order approximations for the remainder. The appropriate cut-off point varies according to the particular details of the application, and requires experimentation to determine.



(a) a, b with p_0 constant.



(b) a, p_0 with b constant.

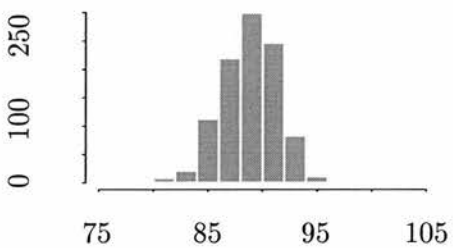


(c) b, p_0 with a constant.

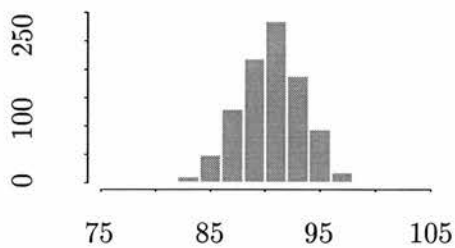
Figure 31: Plots of the log-likelihood function at the maximum likelihood estimates of the 12-step model, Fit 2. For each plot, one parameter is held constant at its maximum likelihood estimate, and the log-likelihood is plotted as the remaining two parameters are varied. The maximum likelihood estimate is marked with a triangle on each plot.

Method of fit	Parameter estimates			Mean BAM results for 1990 from 1000 simulations		
	a	b	p_0	match	unmatched 0s	unmatched 1s
1: one-step model	0.571	4.96	0.0600	89.3	5.4	5.3
2: 1st order	1.32	2.74	0.767	91.0	5.2	3.8
3: 2nd order	1.35	2.69	0.798	90.6	4.6	4.8
4: true parameters	1.00	2.00	0.400	89.9	5.4	4.7

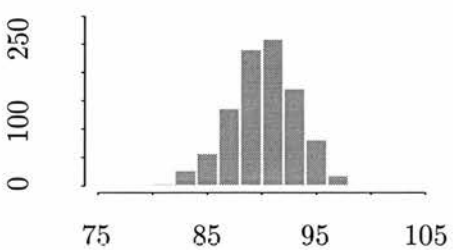
Table 9: Parameter estimates and mean BAM scores from the four different methods of fit for time $T = 12$. The last three columns give respectively the mean best attainable match, the mean number of mismatched 0s in the predicted distributions after swaps have been performed, and the mean number of mismatched 1s in the predicted distributions after swaps have been performed. Means are taken from 1000 predictions using the parameters shown.



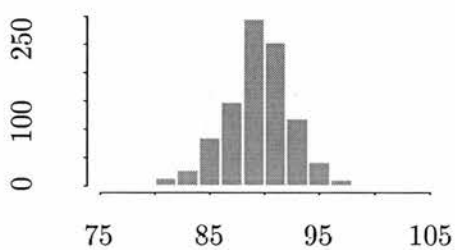
(a) Fit 1: one-step model.



(b) Fit 2: 1st order.



(c) Fit 3: 2nd order.



(d) Fit 4: true parameters.

Figure 32: Best attainable match results from predicted distributions for time $T = 12$, using the parameters from four different fitting methods. Each histogram shows BAM results from 1000 simulated distributions for $T = 12$, generated from the parameters indicated. The number of sites being compared is $N = 100$.

6.4 Variance estimation

The standard errors of the parameter estimates obtained in the previous two sections are estimated using asymptotic theory (Cox & Hinkley 1974). Suppose the observed data are $\boldsymbol{y} = (y_1, \ldots, y_N)$, and suppose that $\boldsymbol{\theta}$ is the vector of parameters and $L(y_r \,|\, \boldsymbol{\theta})$ is the likelihood of a single observation given $\boldsymbol{\theta}$. The information matrix $i_r(\boldsymbol{\theta})$ for a single observation y_r has (j,k)th component

$$[\, i_r(\boldsymbol{\theta}) \,]_{jk} = \mathbb{E}_{\boldsymbol{\theta}} \left(\frac{\partial}{\partial \theta_j} \log L\,(Y_r \,|\boldsymbol{\theta}\,) \, \frac{\partial}{\partial \theta_k} \log L\,(Y_r \,|\boldsymbol{\theta}\,) \right) \tag{59}$$

$$= -\,\mathbb{E}_{\boldsymbol{\theta}} \left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log L\,(Y_r \,|\boldsymbol{\theta}\,) \right), \tag{60}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ is the vector of random variables underlying the observations $\boldsymbol{y}$. The expression for the full-sample information matrix $i_{\cdot}(\boldsymbol{\theta})$ is obtained by substituting $L(\boldsymbol{Y} \,|\, \boldsymbol{\theta})$ for $L(Y_r \,|\, \boldsymbol{\theta})$ in (60).

When the full likelihood is the product of the single-observation likelihoods, i.e. when $L(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \prod_{r=1}^{N} L(y_r \,|\, \boldsymbol{\theta})$, equation (60) gives

$$i_{\cdot}(\boldsymbol{\theta}) = \sum_{r=1}^{N} i_r(\boldsymbol{\theta}) \,.$$

From (59),

$$[\, i_{\cdot}(\boldsymbol{\theta}) \,]_{jk} = \sum_{r=1}^{N} \mathbb{E}_{\boldsymbol{\theta}} \left(\frac{\partial}{\partial \theta_j} \log L\,(Y_r \,|\boldsymbol{\theta}\,) \, \frac{\partial}{\partial \theta_k} \log L\,(Y_r \,|\boldsymbol{\theta}\,) \right)$$

and the information matrix $i_{\cdot}(\boldsymbol{\theta})$ may be estimated for values of $\boldsymbol{\theta}$ close to the true value $\boldsymbol{\theta_0}$ by replacing the expectations in the above expressions by observations:

$$[\, i_{\cdot}(\boldsymbol{\theta}) \,]_{jk} \simeq \sum_{r=1}^{N} \left(\frac{\partial}{\partial \theta_j} \log L\,(y_r \,|\boldsymbol{\theta}\,) \, \frac{\partial}{\partial \theta_k} \log L\,(y_r \,|\boldsymbol{\theta}\,) \right). \tag{61}$$

The asymptotic variance-covariance matrix of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta_0}$ is given in Cox & Hinkley (1974) as the matrix inverse $[\, i_{\cdot}(\boldsymbol{\theta_0}) \,]^{-1}$. This is estimated by the value $\left[\, i_{\cdot}(\hat{\boldsymbol{\theta}}) \,\right]^{-1}$ obtained by replacing $\boldsymbol{\theta_0}$ with $\hat{\boldsymbol{\theta}}$. An estimate of the standard error of any parameter θ_j is then afforded by the square root $\sqrt{\left[\, i_{\cdot}(\hat{\boldsymbol{\theta}}) \,\right]^{-1}_{jj}}$ of the jth diagonal element of $\left[\, i_{\cdot}(\hat{\boldsymbol{\theta}}) \,\right]^{-1}$, with $\left[\, i_{\cdot}(\hat{\boldsymbol{\theta}}) \,\right]^{-1}$ estimated as in (61). The numerical derivatives of (61) were computed using the procedure `dfridr` of Press *et al.* (1988).

Standard error estimates for the parameters a, b and p_0 are given in Table 10 for Fits 1

Method of fit	a		b		p_0	
	MLE	s.e.	MLE	s.e.	MLE	s.e.
1: full model	0.170	0.0335	1.83	0.255	0.0314	0.00476
2: one-step model	0.122	0.0511	2.64	0.379	0.0800	0.0241
3: 1st order	0.117	0.0484	2.55	0.394	0.0367	0.00848
4: 1st order (modified)	0.117	0.0484	2.56	0.394	0.0368	0.00850
5: 2nd order	0.117	0.0491	2.57	0.392	0.0368	0.00880
6: 2nd order (pseudo)	0.117	0.0477	2.57	0.401	0.0369	0.00846
7: 2-stage	0.123	0.0561	1.71	0.477	0.0225	0.00509

Table 10: Maximum likelihood estimates (MLE) and estimated standard errors (s.e.) for the three parameters a, b and p_0 from seven methods of fit for the woodlark data.

to 7 of section 6.2. The observation vector $\boldsymbol{y}$ of equation (61) comprises all survey data: usually $\boldsymbol{y} = \boldsymbol{y^{(4)}}$ but in the case of Fit 1 (full model), $\boldsymbol{y} = \left(\boldsymbol{y^{(1)}}, \boldsymbol{y^{(2)}}, \boldsymbol{y^{(3)}}, \boldsymbol{y^{(4)}}\right)$ and for Fit 7 (2-stage), $\boldsymbol{y} = \left(\boldsymbol{y^{(3)}}, \boldsymbol{y^{(4)}}\right)$. An alternative set of standard error estimates was obtained for Fit 3 (1st order) using the parametric bootstrap. For each of 200 bootstrap replicates a simulated distribution for 1990 was obtained using the parameter estimates from Fit 3, and the likelihood of the simulated distribution was maximized to yield new replicate-specific parameter estimates. The bootstrap estimate of standard error for each parameter is the sample standard error of the 200 replicate-specific estimates. The results for a, b and p_0 were respectively 0.0382, 0.431 and 0.0104, which tally closely with the asymptotic estimates of 0.0484, 0.394 and 0.00848.

Standard error estimates for a, b and p_0 from Fits 1, 2 and 3 of section 6.3 are presented in Table 11. The flatness of the likelihood function for this example causes the variance to be very high. Parametric bootstrap results are unreliable for this case, because with reasonably high frequency a final-year distribution of zeros everywhere is generated. Since the colonization probabilities are given by $p_{ih} = p_0 \exp(-a\delta_{ih} - b\varsigma_h)$, maximization of the likelihood for an all-zero or very sparse distribution might arise from any combination of $a \to \infty$, $b \to \infty$, or $p_0 \to 0$. This clearly has the potential to introduce unrealistically high variance into the estimates of a and b.

Method of fit	a		b		p_0	
	MLE	s.e.	MLE	s.e.	MLE	s.e.
1: one-step model	0.571	5.93	4.96	2.55	0.0600	0.882
2: 1st order	1.32	1.69	2.74	3.68	0.767	0.546
3: 2nd order	1.35	1.69	2.69	3.80	0.798	0.491

Table 11: Maximum likelihood estimates (MLE) and estimated standard errors (s.e.) for the three parameters a, b and p_0 from three methods of fit for the simulated data, $T = 12$.

7 Goodness-of-fit tests for the colonization model

The colonization model has been applied to the woodlark data in this chapter primarily as an illustration of the methodology that has been developed, and there has consequently been little discussion of the fitness of the model for these data. Although not central to the theme of the chapter, the omission will be remedied here with a simple Monte Carlo goodness-of-fit test (Barnard 1963; Morgan 1984). The test involves the selection of a test statistic U, and evaluation of its value u_1 for the observed data $\boldsymbol{y}^{(\boldsymbol{T})}$. A null hypothesis H_0 is devised, under which the distribution of U is specified. Several values $u_2, \ldots, u_n$ of the test statistic are simulated according to the distribution specified by H_0, and these are ranked in increasing order together with the real-data value u_1. Under H_0 all of the values $u_1, \ldots, u_n$ derive from the same distribution, so that all $n!$ permutations are equally likely for the ordered ranking. In particular, the probability that the real-data value has rank m or more is $(n - m + 1)/n$. For a test of size α, the one-tailed null hypothesis is rejected if u_1 has rank $\geq m$, where $(n - m + 1)/n = \alpha$. Marriott (1979) recommends that $n\,\alpha \geq 5$ to ensure that the power of the test is high.

The test was implemented with the test statistic U as the minimized negative log-likelihood, $-\log L\left(\boldsymbol{y}^{(\boldsymbol{T})} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(\boldsymbol{0})}\right)$. The null hypothesis H_0 was designed to test the goodness-of-fit of the fitted model obtained from Fit 3 of section 6.2 (1st order). Under H_0, the distribution of the minimized negative log-likelihood is that characterized by the colonization model with parameters $a = 0.117$, $b = 2.55$ and $p_0 = 0.0367$. Accordingly, 199 simulated realizations of $\boldsymbol{y}^{(\boldsymbol{T})}$ were generated from the colonization model with these parameters, and the model was refitted to the simulated data, whereupon the negative log-likelihood at the minimum was calculated. The real-data value of 131.66 for the minimized negative log-likelihood was ranked in place 91 out of the 200 values in total, providing no evidence

to reject the null hypothesis at the 5% level.

This Monte Carlo test has shown that the observed data for 1990 were not unusually improbable for the fitted colonization model, and as such provides empirical justification for the use of the model with these data. As a caution, it does not provide a test of the wider hypothesis that the model has correctly captured the mechanisms of woodlark colonization. There remains the possibility that attempts to use the fitted colonization model to predict future distributions would be unsuccessful, despite the model providing a good fit to the data observed.

Further Monte Carlo tests using more complete subsets of the woodlark dataset can help to establish this last point — although it is emphasized that the tests provide only an intuitive feel for the outcome, and are not in any sense rigorous proof. The colonization model was fitted to the complete data for 1986 to 1989, yielding parameter estimates $a = 0.204$, $b = 1.84$ and $p_0 = 0.0297$. The test statistic U was chosen to be the negative log-likelihood of the observed data for 1990 under this fitted model, given the observed data for 1989. This provides some indication of the predictive capability of the model for the woodlark data. Under the null hypothesis, U derives from the distribution specified by the colonization model with these parameters.

To test H_0, 199 sets of data were simulated from 1986 to 1990 using the specified parameters. For each, the negative log-likelihood of the simulated data for 1990, given the simulated data for 1989, was calculated and ranked with the real data value. The real data value was ranked in place 200, providing ample evidence on which to reject the null hypothesis. This suggests that the basic colonization model does not provide satisfactory predictions beyond the scope of the data to which it was fitted.

The ability of the colonization model to predict future distributions may be improved by extending the basic model to incorporate time-dependence in the colonization probabilities $\{p_{ih}\}$. The woodlark population in Thetford Forest increased dramatically between 1986 and 1990, with most of the increase occurring in the final year: the number of occupied sites in the five years were respectively 45, 40, 43, 57 and 74. For this reason it is not surprising that the basic model fails to predict the sharp increase that occurred after the 1989 survey. Time-dependence may be incorporated into the colonization probabilities simply by including a time covariate t and an extra time parameter c: for example,

$$p_{ih}^{(t)} = p_0 \exp(-a\,\delta_{ih} - b\,\varsigma_h - c\,t)\,. \tag{62}$$

Using this formulation, the colonization model was again fitted to the data for years 1986 to 1989, giving parameter estimates $a = 0.194$, $b = 1.89$, $p_0 = 0.0236$ and $c = -0.198$. The estimates indicate that the colonization probabilities tend to increase over time ($c < 0$).

The predictive goodness-of-fit test was applied to the time-dependent model, and this time the real data value of U was ranked in place 186 out of 200. With this model, therefore, the null hypothesis should not be rejected at the 5% level. The extra parameter has apparently made an important difference to the predictive ability of the colonization model, despite a relatively small improvement in log-likelihood on the fitted data (-324.27 for the time-dependent model, -326.00 for the original model).

BAM results were also improved under the time-dependent model. Under the basic model, the mean BAM from 1000 predictions for 1990 was 254.7, and the mean numbers of mismatched 0s and 1s were respectively 49.0 and 12.3. Under the time-dependent model, the corresponding results were 255.5, 41.2 and 19.3. A slightly better BAM has been achieved overall, and the balance between the number of mismatched 0s and 1s has been improved. Both the Monte Carlo tests and the BAM results therefore suggest that the time-dependent model should be used for prediction where possible, although the basic model provides an adequate fit within the scope of the original data.

8 Concluding remarks

The examples of Section 6 have demonstrated the successful implementation of the colonization model with missing years of survey data. Only basic implementations were shown, but the full potential of the methodology for fitting the model to data from several surveys taken at irregular spacing and for using the fitted model to obtain predicted distributions for future years is clear. Using the estimated parameters, approximate future extinction probabilities may be obtained for any subset of sites. These might be useful to determine whether there are regions from which the species is in danger of disappearing.

The estimated parameters of the model reveal information about the relative importance of the various covariates — distance and habitat suitability — in determining the probability of colonization between two sites. There is flexibility for the model to be fitted using different functional forms for the colonization probabilities $\{p_{ih}\}$, and different covariates. For example, habitat covariates for each site h could be incorporated directly into the colonization probabilities rather than combined beforehand into a single habitat quality

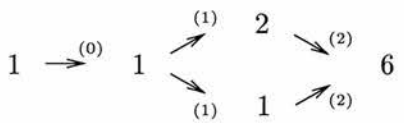
score ς_h, or a time covariate could be included as in Section 7. When only two surveys are available for analysis, however, there is unlikely to be sufficient information to yield a satisfactory estimate of the time parameter.

The principal difficulty in fitting the colonization model to datasets with missing survey data is the choice of approximation level to be used in the calculation of the quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$. There are advantages and disadvantages in using both first and second order approximations. The first-order approximations are easy to implement and quick to compute; the running time is $O(N^2)$ in the first instance (operations required to multiply the $1 \times N$ vector $\boldsymbol{y}^{(0)\prime}$ by the $N \times N$ matrix $Q^{(0)}$) and running time increases only linearly with the number of missing years involved. By contrast, the second-order approximations have complexity $O(N^3)$ since they demand the multiplication of $N \times N$ matrices, and running time increases quadratically with the number of missing years because $(T-2)(T-3)/2$ matrix multiplications are needed in total for a fit covering the period from time 0 to time T.

Nonetheless, the second-order approximations have advantages over the first-order in improved accuracy and tighter error bounds (section 4.2). They are useful for providing a check that the results from the first-order fitting procedure are satisfactory. Moreover, they are more satisfying theoretically than the first-order approximations, for the following reason. Recall from page 109 that the quantity $s_i^{(t)}(q_{ih}^{(t)})$ is equal to the probability that there is no colonization path $k_0 \to^{(0)} k_1 \to^{(1)} \ldots \to^{(t-1)} i \to^{(t)} h$ for any sites $k_0, \ldots, k_{t-1}$. This quantity is given by (11) as:

$$
\begin{aligned}
s_i^{(t)}(q_{ih}^{(t)}) \;=\; \prod_{k_0 : N_{k_0}^{(0)}=1} \prod_{k_1=1}^{N} \Big\{ 1 - q_{k_0 k_1}^{(0)} + q_{k_0 k_1}^{(0)} \prod_{k_2=1}^{N} \Big\{ 1 - q_{k_1 k_2}^{(1)} + q_{k_1 k_2}^{(1)} \prod_{k_3=1}^{N} \Big\{ 1 - q_{k_2 k_3}^{(2)} + \ldots \\
\ldots + q_{k_{t-3} k_{t-2}}^{(t-3)} \prod_{k_{t-1}=1}^{N} \Big\{ 1 - q_{k_{t-2} k_{t-1}}^{(t-2)} \, q_{k_{t-1} i}^{(t-1)} \, q_{ih}^{(t)} \Big\} \cdots \Big\} \Big\} \Big\} \, .
\end{aligned}
$$

Now given any sites k_0 at time 0 and h at time $t+1$, the possible colonization paths from k_0 to h are not independent of each other: for example if $k_0 = 1$, $h = 6$ and $t = 2$ then the following possible paths $1 \to^{(0)} 1 \to^{(1)} 1 \to^{(2)} 6$ and $1 \to^{(0)} 1 \to^{(1)} 2 \to^{(2)} 6$ do not occur independently of each other because they have common ancestry in the colonization from site 1 at time 0 to site 1 at time 1.



However, it is readily shown that the first-order approximation

$$s_i^{(t)}(q_{ih}^{(t)}) \simeq \exp\left(-\sum_{k_0:N_{k_0}^{(0)}=1} \sum_{k_1=1}^{N} \cdots \sum_{k_{t-1}=1}^{N} q_{k_0k_1}^{(0)} q_{k_1k_2}^{(1)} \cdots q_{k_{t-1}i}^{(t-1)} q_{ih}^{(t)}\right)$$

is identical to the first-order approximation when the colonization paths are assumed independent. That is, the first-order approximation to the quantity

$$\prod_{k_0:N_{k_0}^{(0)}=1} \prod_{k_1=1}^{N} \cdots \prod_{k_{t-1}=1}^{N} \left(1 - q_{k_0k_1}^{(0)} q_{k_1k_2}^{(1)} \cdots q_{k_{t-1}i}^{(t-1)} q_{ih}^{(t)}\right)$$

obtained when all colonization paths are assumed independent is the same as the first-order approximation to $s_i^{(t)}(q_{ih}^{(t)})$. The difference becomes manifest only when using second-order and higher approximations. It is more satisfying to employ approximations that distinguish between the true case and the independent case than to employ those that do not.

Taking all of these considerations into account, it is recommended that the second-order approximations should be used as far as is practicable, and certainly for times $t = 1$ and $t = 2$ when they do not require any greater computational effort than the first-order approximations. When applying the methods, the length of time required for the calculation of the quantities $\left\{q_{ih}^{(t)}\right\}$ should be monitored for each t. In addition, there might be occasions on which the instability of the second-order approximation demands that first-order approximations are used instead, although this situation has not been observed. There will generally be a clear point t beyond which it is not worthwhile to continue with the second-order approximations, either because the time taken is too great, the results differ negligibly from those using the first-order approximations, or perhaps because the regularity of unstable results is too high. This point will vary according to the number of survey sites N, the sparsity of the observed distributions and the computer processor. The first-order approximations should be used to calculate the quantities $\left\{q_{ih}^{(u)}\right\}$ for values of u greater than t.

Chapter 6

Simulation-based approaches to parameter estimation in the colonization model

Following from the analytic likelihood estimates of Chapter 5, this chapter examines two alternative approaches to parameter estimation in the colonization model. Both approaches have their basis in simulation. It is invariably easy to simulate from a model — even a very complicated one — simply by substituting values for the unknown parameters and generating random numbers to mimic the stochasticity in the system. Parameter estimates obtained by analytic means, on the other hand, often require involved calculations and might also be specific to a single problem. This is certainly true of the estimates obtained in Chapter 5. Simulation-based approaches therefore have the potential to be useful where analytic calculations would be either intractable or too time-consuming to perform; they also help to reduce the danger of working with models that are simplified to an unrealistic extent in order to accommodate the restrictions of an analytic treatment. There are of course certain disadvantages to simulation-based approaches, some of which will become clear in the course of this chapter.

In the first method presented here, the likelihood function is dispensed with altogether for parameter estimation. The second method uses simulations to facilitate the calculation of the likelihood function. Both methods will be applied to the woodlark data, introduced in Chapter 3 and analysed in Chapter 5. Although both of the approaches can be applied to models other than the colonization model, the first method is perhaps more general.

1 Full simulation technique

1.1 Methodology

The approach presented in this section relies entirely on simulation, and removes the need for likelihood calculations. This is useful for problems in which the likelihood function is very difficult or impossible to compute. In the case of the colonization model, for example, even the calculation of the one-step likelihood relies on the assumption that colonizations of and from sites are independent over a single time period. Without this assumption the calculation of the likelihood would become much more difficult and perhaps impossible, depending on the specified correlation structure of the colonization mechanism. No matter how hard it is to compute the likelihood, however, it is generally straightforward to perform simulations from any model using assumed parameter values. This fact forms the basis of the simulation method.

An informal motivation for the approach is provided here. In much of statistical modelling, parameter estimation proceeds by determination of a function of the parameters that describes the plausibility of each parameter vector in the light of the data observed. In frequentist analyses this function is the likelihood $L(\boldsymbol{x} \mid \boldsymbol{\theta})$, while in the Bayesian case it is the posterior function $f(\boldsymbol{\theta} \mid \boldsymbol{x})$: here, $\boldsymbol{x}$ is the vector of observations and $\boldsymbol{\theta}$ is the vector of parameters. In both cases the fitting of the model entails the collection of observations $\boldsymbol{x}$, the evaluation of a function of the parameters according to some rule, and the final manipulation of this function to yield parameter estimates. The last stage is typically a maximization in the case of the likelihood function, and a calculation of the posterior mean or median in the Bayesian case.

The function that is intended to represent degrees of plausibility across the parameter space is chosen to a greater or lesser extent subjectively in both frequentist and Bayesian statistics. In the Bayesian framework, subjectivity enters explicitly through the choice of the prior function $f(\boldsymbol{\theta})$: the posterior is calculated via the expression $f(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto L(\boldsymbol{x} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$. The likelihood function in frequentist statistics is also a subjective choice in some senses, although this is easily forgotten because it is so widely used. Of course choices such as these are popular because they are effective, but other possibilities need not be excluded.

The traditional concept of observations or data is also somewhat narrow. Observations of the statistical process are usually restricted to those for which the parameter values are unknown. For example, in the colonization model the observations of the colonization

process are the spatial distributions of species' presence and absence at the beginning and end of a time period. These are precisely the observations to be modelled: estimates are required for the unknown parameter values associated with them. The terminology *natural observations* will henceforth be used to denote this type of data.

In reality, there is much more information available about the process than that contained in the natural observations. Typical outputs from the model using any set of known parameters are readily obtained, simply by simulation. Comparisons of the model output under known parameters with the natural observations of the unknown parameters can be made to suggest the level of compatibility of the known parameters with the natural observations. For instance, suppose that a single natural observation of 3.0 is obtained from a normal distribution with unit variance and unknown mean; and suppose that ten thousand observations were simulated from a normal distribution with unit variance and mean 7. If all of these ten thousand observations are found to be strictly greater than 3.0, there is strong evidence that the mean behind the natural observation of 3.0 is not 7. This is the same probabilistic thinking as that behind the likelihood function: if the probability of the data to be modelled is very low under a particular set of parameter values, these do not provide good parameter estimates.

For more complicated modelling situations, simulations will be unable to provide such an unequivocal impression of probability as in the normal distribution example. Nonetheless, it is clear that a function of the parameters to measure compatibility with the natural observations can still be derived using the extra observations collected from model simulation trials. Such observations will be referred to as *simulated observations.* In essence, a *compatibility function* $\mathcal{C}(\boldsymbol{\theta} \mid \boldsymbol{y_0}, \boldsymbol{x})$ is defined to measure compatibility between the parameters $\boldsymbol{\theta}$ and the natural observations $\boldsymbol{y_0}$, on the basis of simulated observations $\boldsymbol{x}$ and their relation to $\boldsymbol{y_0}$. The compatibility function takes the place of the likelihood function, and is maximized with respect to $\boldsymbol{\theta}$ to provide the final parameter estimates.

To formalize the above, let Θ be the parameter space, and let $\boldsymbol{y}(\boldsymbol{\theta})$ be a single simulation from the model using parameters $\boldsymbol{\theta} \in \Theta$. It is assumed that the model is correct: that is, the natural observations $\boldsymbol{y_0}$ are given by $\boldsymbol{y_0} = \boldsymbol{y}(\boldsymbol{\theta_0})$ for some unknown $\boldsymbol{\theta_0} \in \Theta$. In the colonization model, for example, $\boldsymbol{y_0}$ would be the observed spatial distribution at the end of the time period, and $\boldsymbol{y}(\boldsymbol{\theta})$ would be a simulated spatial distribution at the end of the time period, obtained using the known parameters $\boldsymbol{\theta} \in \Theta$.

The simulated observation $\boldsymbol{y}(\boldsymbol{\theta})$ is compared against the natural observation $\boldsymbol{y_0}$, and the

compatibility function evaluated at the point $\boldsymbol{\theta}$ is calculated on the basis of this comparison. If $\boldsymbol{y}(\boldsymbol{\theta})$ is close to $\boldsymbol{y_0}$ in some pre-ordained respect, the parameter vector $\boldsymbol{\theta}$ is given a high compatibility score; otherwise the compatibility score is low. The compatibility of $\boldsymbol{\theta}$ with $\boldsymbol{y_0}$ is written as $\mathcal{C}\left(\boldsymbol{\theta} \mid \boldsymbol{y_0}, \boldsymbol{y}(\boldsymbol{\theta})\right)$.

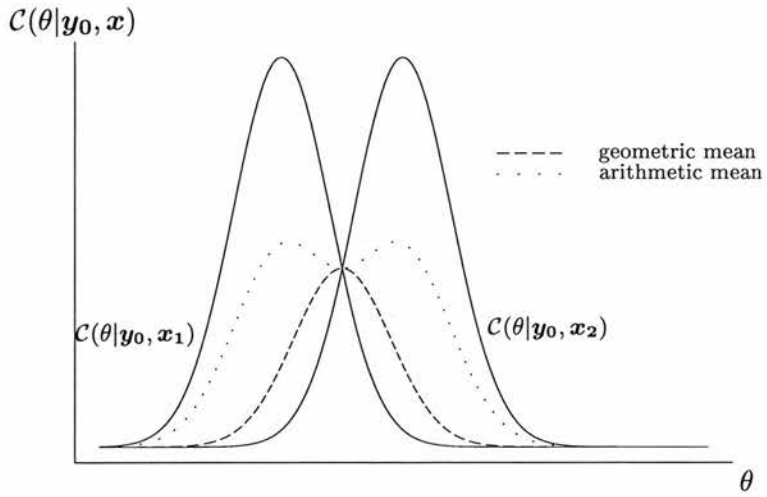
In practice, it is desirable to include more than one simulated observation for any $\boldsymbol{\theta}$ at which $\mathcal{C}$ is to be evaluated. One simulation alone might be unusual and distort the results. Instead, some way of combining the information from a number of simulated observations is required. It is recommended that this is achieved by separate evaluation of the compatibility function $\mathcal{C}$ at each simulated observation, with the final value taken to be the geometric mean of the individual results. Thus if k simulated observations $\boldsymbol{y_1}(\boldsymbol{\theta}), \ldots, \boldsymbol{y_k}(\boldsymbol{\theta})$ are obtained for the parameter vector $\boldsymbol{\theta}$, the compatibility function is given by

$$\mathcal{C}\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{y_0}, \boldsymbol{y_1}(\boldsymbol{\theta}), \ldots, \boldsymbol{y_k}(\boldsymbol{\theta})\right) = \left\{ \prod_{i=1}^{k} \mathcal{C}\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{y_0}, \boldsymbol{y_i}(\boldsymbol{\theta})\right) \right\}^{\frac{1}{k}} .$$

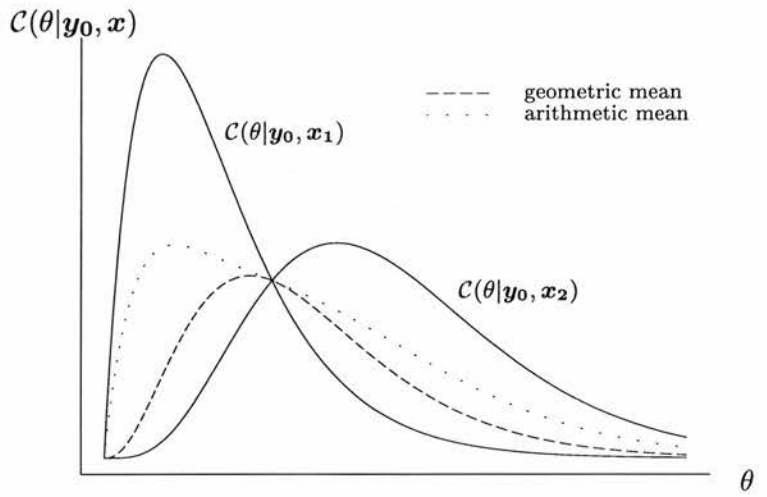
The geometric mean is suggested rather than the arithmetic mean because it tends to be more effective in combining information from different simulations. Suppose for example that one set of simulated observations $\boldsymbol{x_1} = \{\boldsymbol{y_1}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ gives maximum compatibility at the parameter vector $\boldsymbol{\theta_1}$, while a second set $\boldsymbol{x_2} = \{\boldsymbol{y_2}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ gives maximum compatibility at a different vector $\boldsymbol{\theta_2}$, and relatively low compatibility at $\boldsymbol{\theta_1}$. The best estimate of the true parameter vector on the basis of these two simulations would lie between $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$; however, the arithmetic mean tends to favour the points $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ individually, while the geometric mean takes better account of a low compatibility score from the other simulation and reaches a peak between the two points (Figure 1).

For the same reason, however, care must be taken to ensure that the compatibility function does not penalize parameter vectors too heavily for a poor performance. If the compatibility of a parameter vector $\boldsymbol{\theta}$ is zero for any one simulation, use of the geometric mean will cause the combined compatibility from several simulations to be zero, even if the performance of $\boldsymbol{\theta}$ is good in the other simulations. Although this means that only those parameters which are consistently good will be selected, it should be borne in mind that the natural observation $\boldsymbol{y_0}$ is not necessarily a representative observation from the true parameter vector $\boldsymbol{\theta_0}$.

There are a number of issues that must be resolved in any implementation of the simulation method for parameter estimation. First and most important is the choice of compatibility



(a)



(b)

Figure 1: Comparisons for scalar θ between the arithmetic and geometric means of pairs of functions, $\mathcal{C}(\theta|\boldsymbol{y_0}, \boldsymbol{x_1})$ and $\mathcal{C}(\theta|\boldsymbol{y_0}, \boldsymbol{x_2})$ (solid lines). The dashed lines show the geometric mean, and the dotted lines show the arithmetic mean. In each figure the geometric mean reaches a peak between the two peaks of the original functions, while the arithmetic mean tends to favour one or both of the original peaks individually.

function $\mathcal{C}(\boldsymbol{\theta} \mid \boldsymbol{y_0}, \boldsymbol{y(\theta)})$. The choice will always be subjective, but should be made so as to reward parameters that correctly predict the features of greatest interest in the natural observations. In the example of the colonization model, it is the distribution and range of the species at the end of the time period that are of greatest interest, and the compatibility function might be based on the best attainable match (BAM) scores derived in Chapter 4. In other examples, there might be quite different features of the natural observations that are priorities for parameter selection, such as the mean or variance of a vector of

identically distributed observations, or the extreme points of a set of data for risk control, and so on.

A second issue is the number of simulated observations to take at each $\boldsymbol{\theta}$ for which the compatibility function is evaluated. In general, a large number of simulations will give more representative results; however, there are several considerations on which the decision should be based. These include the variance of the simulated observations, the choice of compatibility function, the dimension of the parameter space and the computational expense of each simulation. If the simulated observations vary little from one simulation to the next, or if the compatibility function is relatively insensitive to the variation that occurs, only a few simulations need be taken at each point. Again, if there are a large number of parameters or if every simulation is expensive in computer time, the number of simulations should be kept small so as to ensure the maximization procedure is not too lengthy. This is especially important if the variance of the parameter estimates is to be assessed by means of the bootstrap, as is recommended.

Thirdly, a procedure must be determined for the maximization of the compatibility function $\mathcal{C}$ with respect to $\boldsymbol{\theta}$. Strictly speaking, $\mathcal{C}\left(\boldsymbol{\theta} \mid \boldsymbol{y_0}, \boldsymbol{y}(\boldsymbol{\theta})\right)$ is not a function but a random variable, since it depends on the random simulation $\boldsymbol{y}(\boldsymbol{\theta})$. As such, it would be expected to take a different value each time it was evaluated, and maximization procedures would be unlikely to converge due to the random variation throughout the parameter space. This problem may be overcome by fixing the random seed for the simulations. Simulations from any stochastic model involve the generation of random numbers to emulate the stochasticity in the system being modelled. In practice, random numbers are invariably generated according to a specified scheme: they are not truly random and are referred to as *pseudo-random* numbers (Morgan 1984). Given a starting point for the pseudo-random scheme, known as the *random seed*, the sequence of pseudo-random numbers is completely determined. By fixing the random seed at a constant value before calculating the simulated data $\boldsymbol{y}(\boldsymbol{\theta})$, the same result for $\mathcal{C}\left(\boldsymbol{\theta} \mid \boldsymbol{y_0}, \boldsymbol{y}(\boldsymbol{\theta})\right)$ is guaranteed at every evaluation.

When several simulated observations $\boldsymbol{y_1}(\boldsymbol{\theta}), \ldots, \boldsymbol{y_k}(\boldsymbol{\theta})$ are collected for each parameter vector $\boldsymbol{\theta}$, a different random seed must be used for each of the k simulations — otherwise the same results would be repeated k times. The fixing of the random seed applies when the compatibility function is evaluated at a different value of $\boldsymbol{\theta}$. Thus if the k values $s_1, \ldots, s_k$ were used as random seeds for vector $\boldsymbol{\theta_1} \in \Theta$, the seeds used for any other vector $\boldsymbol{\theta_2} \in \Theta$ should also be $s_1, \ldots, s_k$.

The method of fixing the random seed not only ensures that the compatibility function $\mathcal{C}$ is a true function, but also helps to smooth $\mathcal{C}$ to some extent. Two parameter vectors $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ that are close to each other in the parameter space Θ will generate similar simulated observations $\boldsymbol{y}(\boldsymbol{\theta_1})$ and $\boldsymbol{y}(\boldsymbol{\theta_2})$ if the same random numbers are used in the two simulations, and consequently the respective compatibilities $\mathcal{C}(\boldsymbol{\theta_1} \mid \boldsymbol{y_0}, \boldsymbol{y}(\boldsymbol{\theta_1}))$ and $\mathcal{C}(\boldsymbol{\theta_2} \mid \boldsymbol{y_0}, \boldsymbol{y}(\boldsymbol{\theta_2}))$ will be close. Although the compatibility function is never likely to be continuous at all points, the discontinuities will be tempered considerably when the random seed is fixed. This will also facilitate the maximization process.

A final issue is the choice of initial vector $\boldsymbol{\theta}$ from which to begin the maximization procedure. Given the simulated nature of the compatibility surface, in all but the most unusual circumstances it is likely to have numerous crevices, discontinuities and non-global maxima. The numerical maximization of a function usually proceeds either by finding points at which the derivative of the function is zero, or by repeated evaluation of the function until the highest point is thought to have been attained. The first of these possibilities is out of the question in the present context, since the function to be maximized is not differentiable. In the second method, an initial vector $\boldsymbol{\theta}$ is selected, the shape of the function in the vicinity of $\boldsymbol{\theta}$ is explored, and new evaluations are performed as $\boldsymbol{\theta}$ moves in directions along which the function increases. An initial choice $\boldsymbol{\theta}$ close to a non-global maximum, therefore, might cause the routine to converge wrongly by overlooking the true maximum.

The only precaution that can be taken against convergence to a non-global maximum is to start the maximization routine from many different points in the parameter space and adopt the best result. Even this does not guarantee that the correct maximum will be found, if none of the selected starting points are sufficiently well-placed (Brooks & Morgan 1994). Some care must therefore be taken in this initial selection procedure. Brooks and Morgan (1994) present an algorithm based on simulated annealing (e.g. Ripley 1988) for automatic selection of the number and positions of starting points.

The simulation method is implemented in the following sections using the colonization model and the woodlark data described in Chapter 3. In section 1.2 the basic colonization model is used, and the results obtained from the simulation method are compared against those of the maximum likelihood methods in Chapter 5. The basic colonization model is extended in section 1.3 to remove the assumption of independence of colonizations of and from sites over a single time-step. It is not possible to use the likelihood methods for this case, but the simulation method is easily employed.

The ideas outlined in this section are undoubtedly rooted more in expediency than in theoretical elegance. However, the examples in the following sections will demonstrate that their careful application can bring encouraging results.

1.2 Application to the woodlark data

The simulation method is applied here to the woodlark data using the colonization model described in Chapter 3. The data from the first and final years 1986 and 1990 will be used, so that there are three years with missing surveys. The compatibility function evaluated at any parameter vector $\boldsymbol{\theta}$ is therefore based on a comparison of the observed spatial distribution of woodlarks in 1990 with a simulated distribution for 1990, where the simulated distribution is obtained by conditioning on the observed distribution for 1986 and simulating four years ahead using the parameters $\boldsymbol{\theta}$.

Although not illustrated here, the method can be employed using any combination of years between 1986 and 1990. If more than two time-points are included in the analysis, the compatibility function spanning all of the included time-steps may be taken as the product of the compatibilities from each consecutive pair of time-points.

All of the examples in this section have a fixed number of ten simulated observations taken at each parameter vector $\boldsymbol{\theta}$. This is a suitable compromise between a quick maximization and a large sample. With ten simulated observations at each parameter vector, a typical maximization is completed in 3–10 minutes on a Sun Ultra 10 server with 192MB RAM. This allows a bootstrap procedure with 100 replicates to be completed within a reasonable time span of a few hours.

The method of fixing the random seed is used throughout this section. A pseudo-random number generator was used, following the algorithm

$$x_n = 27.182813\, x_{n-1} + 31.415917 \pmod 1$$

suggested in Bishop (1989), where x_n is the value of the nth random number. The compatibility function was maximized using the routine `amoeba` of Press *et al.* (1988).

Of primary interest is the choice of compatibility function $\mathcal{C}\left(\boldsymbol{\theta} \mid \boldsymbol{y}_0, \boldsymbol{y}(\boldsymbol{\theta})\right)$. As mentioned above, the best attainable match (BAM) scores of Chapter 4 provide a suitable basis for $\mathcal{C}$, since it is the spatial distribution of the data that is of greatest interest. However,

preliminary investigations quickly reveal that the BAM score by itself does not convey a sufficiently comprehensive impression of similarity. A simulated distribution of absence everywhere, for example, yields a relatively high BAM score of 242 matches out of 316 sites in 1990, while clearly failing to capture the true pattern of distribution. Such anomalies will invariably occur when the composition of the observed distribution is largely either presence or absence.

A full assessment of similarity must instead take into account the regional structure of the woodlark distribution, and the relative numbers of sites with zero and one status that are left unmatched after swaps have been performed. To facilitate this, the survey area may be divided into a number of regions tracing features of the observed distribution. A simulated distribution that is truly similar to the observed distribution will have small and roughly equal numbers of mismatched 0s and 1s in each region. (A mismatched 0 is defined as a site with zero status in the simulated distribution that is left unmatched after swaps have been performed, and a mismatched 1 is a site with status 1 in the simulated distribution that is left unmatched after swaps.)

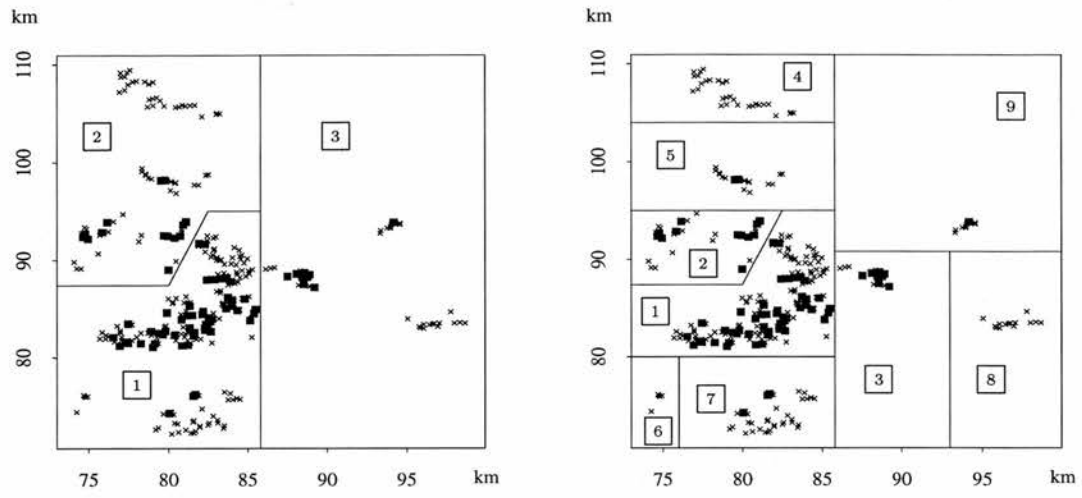
Regions of the woodlark survey area were chosen arbitrarily, with the sole condition that no clique should span more than one region. Recall that the clique of any site is the set of sites with which swaps are permitted: in the case of the woodlark data, the set of sites within 0.5 km of that site. The condition simply ensures that the number of mismatched 0s and 1s left in the region after swaps have been performed is well-defined.

Two regional divisions of the survey area were investigated, the first involving three regions and the second involving nine. These are shown in Figure 2.

Once regions have been selected, a compatibility score is required that rewards simulated distributions with low numbers of mismatched sites in each region and approximately balanced numbers of mismatched 0s and 1s. Let $n_0(r)$ be the number of mismatched 0s in region r of the simulated distribution after swaps, and let $n_1(r)$ be the number of mismatched 1s. The quantity $\left| \left(n_0(r) + n_1(r)\right) \left(n_0(r) - n_1(r)\right) \right|$ may be used as the basis for such a compatibility score, and one possible transformation is given by

$$C\left(\boldsymbol{\theta} \mid \boldsymbol{y_0}, \boldsymbol{y}(\boldsymbol{\theta})\right) = \prod_{r=1}^{R} \exp\left\{ -\sqrt{\left| \Big(n_0(r) + n_1(r)\Big) \Big(n_0(r) - n_1(r)\Big) \right|} \right\}, \qquad (1)$$

where the number of regions R is variously 3 or 9. Note that the quantities $n_0(r)$ and $n_1(r)$ $(r = 1, \ldots, R)$ stem from the comparison between the natural observations $\boldsymbol{y_0}$ and



(a) 3 regions. (b) 9 regions.

Figure 2: Regional divisions of the survey area in Thetford Forest. The distribution of woodlark presence (black squares) and absence (crosses) for the 316 suitable sites in 1990 is also shown.

the simulated observations $\boldsymbol{y}(\boldsymbol{\theta})$, and the dependence on the parameters $\boldsymbol{\theta}$ enters through the simulated data $\boldsymbol{y}(\boldsymbol{\theta})$. The square root is used in (1) to moderate the decrease in compatibility as the number of mismatched sites increases.

If required, the contributions from different regions could be weighted according to the number of sites in each. In the present study it was decided that the contribution from each region should be equal, because the criterion that the parameters should correctly predict whether or not the species had spread into a remote region, even one with only a few sites, was regarded with equal importance as the criterion that the parameters should correctly predict presence in the main concentration of sites at the centre of the survey region. Indeed, the first of these criteria was generally the more telling as to the worth of a given parameter vector.

A second possible form for the compatibility function is given by

$$\mathcal{C}\left(\boldsymbol{\theta} \mid \boldsymbol{y}_0, \boldsymbol{y}(\boldsymbol{\theta})\right) = \prod_{r=1}^{R} \exp\left\{-\sqrt{\left|\left(n_0(r) - n_1(r)\right)\right|}\right\}, \tag{2}$$

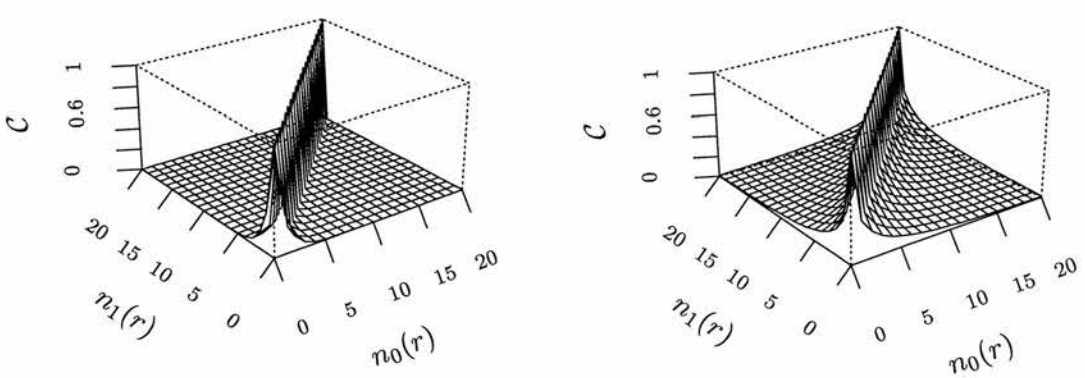
which rewards only the balance in the number of mismatched 0s and 1s, regardless of the total number of mismatches. There is a tendency for distributions with well-balanced numbers of mismatches to attain a good match overall, so the omission of an explicit term in $(n_0(r) + n_1(r))$ is not as rash as it might seem. However, the form (2) is studied

primarily for comparison with (1), since it provides more moderate fall-off as the number of mismatches increases. Predicted values of $n_0(r) = 18,\ n_1(r) = 13$, for example, incur a higher penalty under (1) than predictions of $n_0(r) = 12,\ n_1(r) = 0$, although the first of these probably represent the better prediction. This ranking is reversed under form (2).

Both of the compatibility functions (1) and (2) fail to distinguish between the case $n_0(r) = n_1(r) = 0$ and the case $n_0(r) = n_1(r) \gg 0$. In the present application this was not a concern, since it was rare for $n_0(r)$ and $n_1(r)$ to be large concurrently: usually presence was either over-predicted or under-predicted, but not both. In circumstances where the situation $n_0(r) = n_1(r) \gg 0$ might arise, one possible remedy is the addition of a constant k to the term $(n_0(r) - n_1(r))$ in (1), to give a compatibility function of the form

$$\mathcal{C}\left(\boldsymbol{\theta} \mid \boldsymbol{y_0},\, \boldsymbol{y(\theta)}\right) = \prod_{r=1}^{R} \exp\left\{-\sqrt{\left|\left(n_0(r) + n_1(r)\right)\left(n_0(r) - n_1(r) + k\right)\right|}\right\}.$$

The shapes of the compatibility functions in both of forms (1) and (2) are shown in Figure 3. The figures show the fall-off of $\mathcal{C}$ with $n_0(r)$ and $n_1(r)$ for a single region r.



(a) Form (1). (b) Form (2).

Figure 3: Shapes of two forms for the compatibility function in a single region r as $n_0(r)$ and $n_1(r)$ vary: (a) $\mathcal{C} = \exp\left\{-\sqrt{|\left(n_0(r) + n_1(r)\right)\left(n_0(r) - n_1(r)\right)|}\right\}$ as in equation (1), and (b) $\mathcal{C} = \exp\left\{-\sqrt{|\left(n_0(r) - n_1(r)\right)|}\right\}$ as in equation (2). The form in (b) decreases less sharply than that in (a).

Four fits will be provided in this section, described as follows. Each fit includes only the observed woodlark distributions from years 1986 and 1990.

Fit 1: 3 regions. Regional division as shown in Figure 2 (a). Compatibility function calculated according to equation (1), with $R = 3$.

Fit 2: 3 regions (balance). Regional division as shown in Figure 2 (a). Compatibility function calculated according to equation (2), rewarding only the balance in mismatched 0s and 1s, with $R = 3$.

Fit 3: 9 regions. Regional division as shown in Figure 2 (b). Compatibility function calculated according to equation (1), with $R = 9$.

Fit 4: 9 regions (balance). Regional division as shown in Figure 2 (b). Compatibility function calculated according to equation (2), rewarding only the balance in mismatched 0s and 1s, with $R = 9$.

The results from these four analyses are presented in Table 1. In addition, a comparison is provided with the maximum likelihood estimates obtained in Chapter 5 using the same data. The maximum likelihood estimates were calculated using Fit 5 of Chapter 5: that is, with second-order approximations for all quantities $q_{ih}^{(t)}$ and $p_i^{(t)}$ when $t < 4$, but with first-order approximations when $t = 4$ and for the likelihood. This method of likelihood calculation was also used to compute the value of the log-likelihood at the estimated parameters of each of Fits 1 to 4.

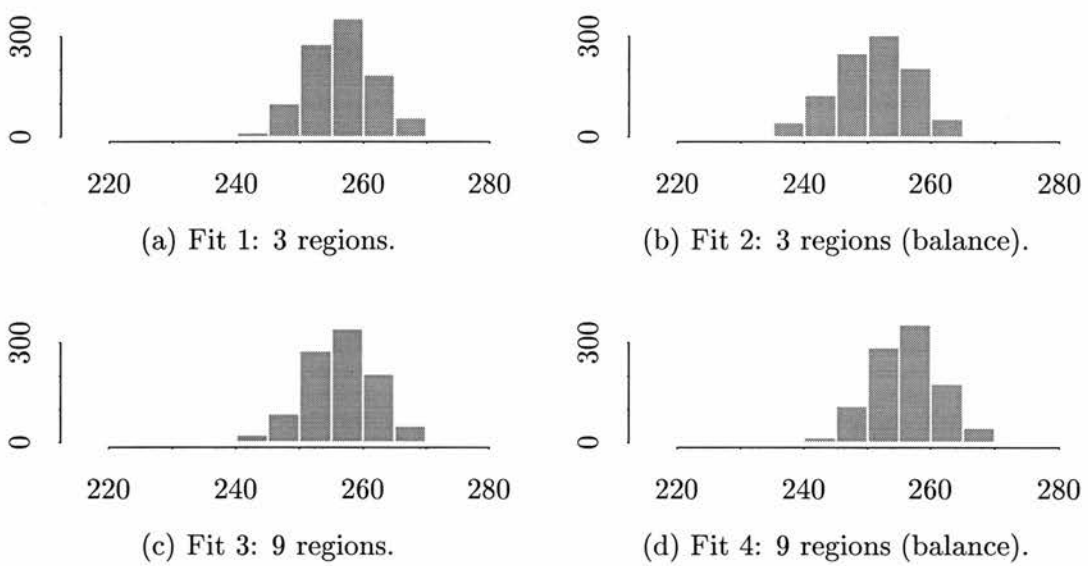
Standard error estimates for the parameters were obtained using the parametric bootstrap. For each fit, 100 distributions for 1990 were simulated from the fitted model; the compatibility function was maximized and the parameter estimates recorded. The standard error of any parameter was estimated by the sample standard error of the 100 replicates.

Best attainable match results were obtained for each of the fitted models by simulating 1000 distributions for 1990 from each one. The simulated distribution was compared against the observed woodlark distribution for 1990, and the mean best attainable match (BAM) from the 1000 trials was calculated. The spread of BAM results for each fit is shown in Figure 4, and the mean BAM is recorded in Table 1. For ease of comparison, the histograms in Figure 4 are shown on the same scale as those from the maximum likelihood estimates in Chapter 5, p. 162.

The results in Table 1 indicate that the simulation method is capable of producing estimates close to the maximum likelihood estimates, and the likelihood values associated with the final estimates are also very close to the maximum. This is particularly true of the results for Fit 3, which yield the best likelihood and BAM results. Fits 1, 3 and 4 all produce a mean BAM score greater than that of the maximum likelihood estimate, which is not surprising since the parameter estimates of these fits were selected specifically on

Method of fit	Parameter estimates						$\log L$	BAM
	a	s.e.(a)	b	s.e.(b)	p_0	s.e.(p_0)		
1: 3 regions	0.0942	0.0465	3.32	0.492	0.0434	0.0153	-133.77	257.0
2: 3 regions (balance)	0.0393	0.0499	2.79	0.755	0.0247	0.0173	-134.52	251.4
3: 9 regions	0.115	0.0391	2.97	0.620	0.0396	0.0178	-133.44	257.0
4: 9 regions (balance)	0.143	0.0786	2.59	0.813	0.0388	0.0263	-133.44	256.5
5: MLE	0.117	0.0491	2.57	0.392	0.0368	0.00880	-131.66	253.9

Table 1: Parameter estimates, estimated standard errors, log-likelihood scores and mean BAM results from the four different methods of fit. A comparison with the second-order maximum likelihood estimates corresponding to Fit 5 of Chapter 5 is provided in the final row. Standard errors (s.e.) were obtained using 100 replicates of the parametric bootstrap. Log-likelihoods were calculated using the second-order approximate method of Fit 5, Chapter 5. The final column gives the mean best attainable match from 1000 predictions using the parameters shown.



(a) Fit 1: 3 regions.

(b) Fit 2: 3 regions (balance).

(c) Fit 3: 9 regions.

(d) Fit 4: 9 regions (balance).

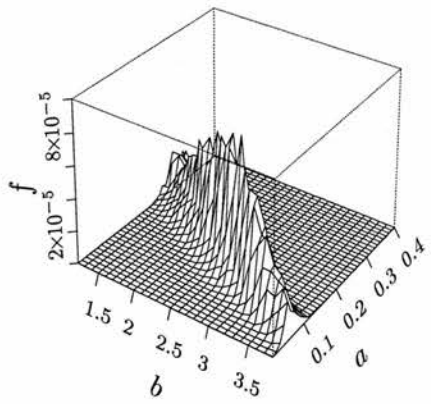
Figure 4: Best attainable match results from predicted distributions for 1990, using the parameters from four different fitting methods. Each histogram shows BAM results from 1000 simulated distributions for 1990, generated from the parameters indicated. The number of sites being compared is $N = 316$.

the basis of producing a high best attainable match. The standard error estimates from each fit are comparable with those from the likelihood method, although there is a slight tendency for variance from the simulated method to be higher. Again, this is to be ex-

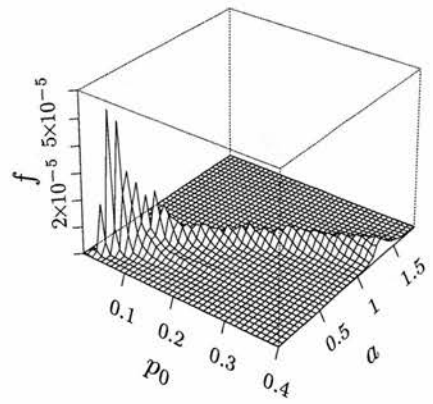
pected because of the extra variability introduced by the simulations — a phenomenon that could be investigated by increasing the number of simulated observations taken at each parameter vector $\boldsymbol{\theta}$, and observing the associated changes in variance.

In some ways it is remarkable that the *ad-hoc* method presented here can achieve results so close to those of the maximum likelihood method, given the subjectivity of the former and the extra input of theoretical development required for the latter. However, the two methods are in fact very similar in concept. Both rely on the maximization of a function of the parameters in the light of the observed data. In the likelihood case, this function is based on the probability of observed presence or absence at every point. In the simulation method the function is based on the number of matched sites over several simulations — but this is itself dependent on the probability of presence or absence at each point. The division of the survey area into smaller regions in the simulation method is an attempt to emulate the likelihood function more closely: the likelihood accounts for each site individually, whereas the simulation method relies on a summary statistic over a region. Accordingly, the results of the simulation method when nine regions were used were better on the whole than those obtained from only three regions. Changing the compatibility function once the number of regions is fixed seems to have had less bearing on the outcome of the analysis, although this could be due to the fact that the compatibility functions tried were rather similar.

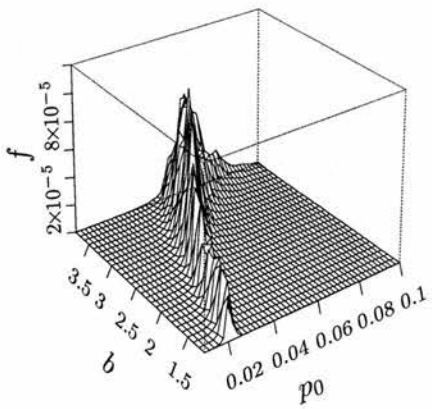
One practical problem with the application of the simulation method in this example was the difficulty in finding the true maxima of the compatibility functions. Due to the very peaked nature of the compatibility functions (1) and (2), the maximization routine would often converge to the wrong point if started at some distance from the true maximum. Figure 5 shows the shape of the compatibility surface for Fit 4 (9 regions: balance) at the maximum; the surfaces for Fits 1 and 3 are even more steep-sided. In the illustrative example given here, the problem was dealt with by conducting a number of restarts for every fit in an *ad-hoc* fashion. In a serious application of the method, a more comprehensive scheme for starting-point selection might be considered (Brooks & Morgan 1994).



(a) a, b with p_0 constant.



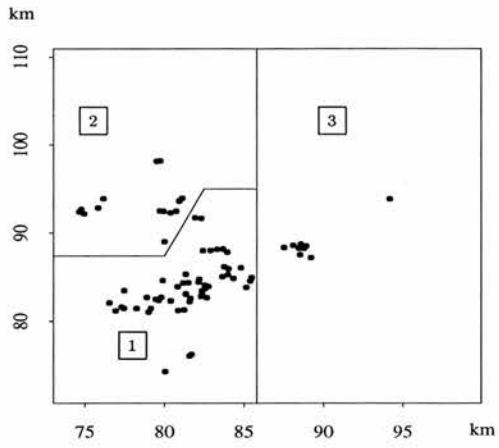
(b) a, p_0 with b constant.



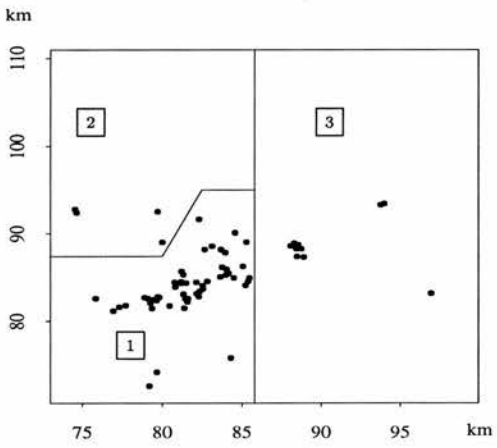
(c) b, p_0 with a constant.

Figure 5: Plots of the compatibility function (2) at the maximum under Fit 4 (9 regions: balance). For each plot, one parameter is held constant at its estimate under Fit 4, and the compatibility function is plotted as the remaining two parameters are varied.

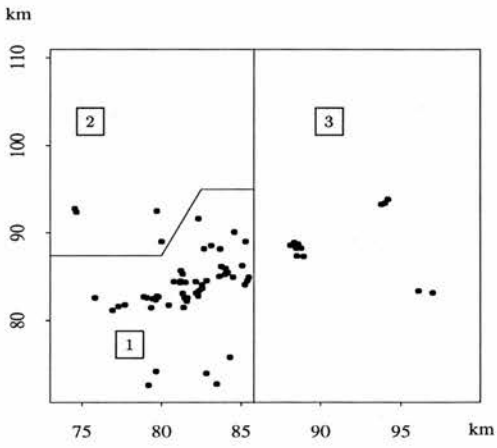
Finally, Figures 6 and 7 show typical predictions for 1990 from each of Fits 1, 2, 3 and 4, compared against the true distribution for 1990. The corresponding regional mismatch scores, $n_0(r)$ and $n_1(r)$, are given in Tables 2 and 3. The regional mismatch scores are the values on which the compatibility functions are based, and should ideally be both small and almost equal for any region. All of the predictions in Figures 6 and 7 give a reasonably good representation of the true distribution, although in all four cases the parameters were unable to reproduce the true level of presence in regions 2 and 5 — underpredicting by a total of 10 sites each time. The level of occupation in these regions is probably anomalous for the colonization model, and this is borne out by inspection of the occupation probabilities from the maximum likelihood fits of Chapter 5. Most occupied sites in regions 2 and 5 were found to have an occupation probability of 0.15 or less, which is rather low in comparison with occupied sites in other regions.



(a) True distribution.



(b) Fit 1: 3 regions.

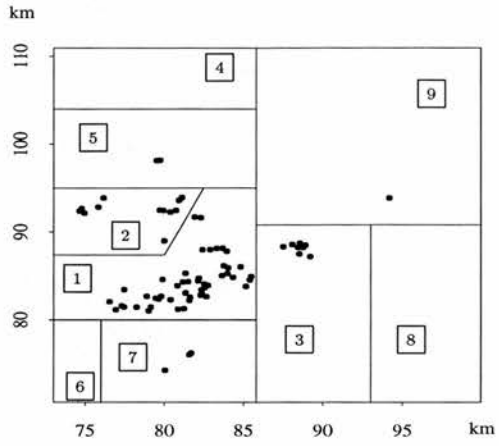


(c) Fit 2: 3 regions (balance).

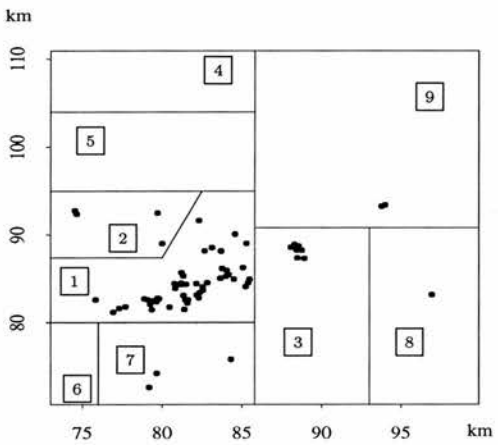
Figure 6: Observed and predicted distributions for 1990, showing the division of the survey area into three regions. Only occupied sites are shown for clarity. The observed distribution is shown in (a), and predicted distributions from the parameter estimates of Fits 1 and 2 are shown in (b) and (c) respectively.

Method of fit	Region, r					
	1		2		3	
	n_0	n_1	n_0	n_1	n_0	n_1
1: 3 regions	14	15	10	0	2	3
2: 3 regions (balance)	16	16	10	0	2	5

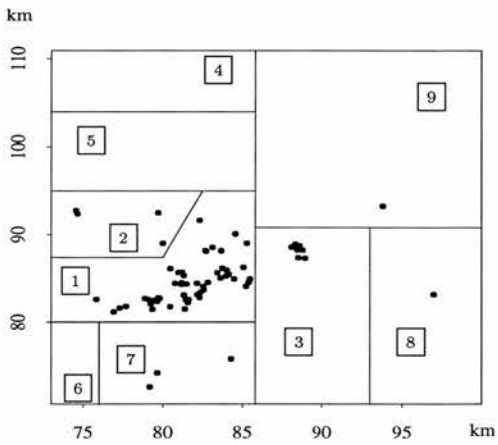
Table 2: Regional mismatch scores for the predictions from Fits 1 and 2 shown in Figure 6 (b) and (c) respectively. Within each of the three regions, the number n_0 of mismatched 0s and the number n_1 of mismatched 1s are given for the simulated distribution in 1990.



(a) True distribution.



(b) Fit 3: 9 regions.



(c) Fit 4: 9 regions (balance).

Figure 7: Observed and predicted distributions for 1990, showing the division of the survey area into nine regions. Only occupied sites are shown for clarity. The observed distribution is shown in (a), and predicted distributions from the parameter estimates of Fits 3 and 4 are shown in (b) and (c) respectively.

Method of fit	Region																	
	1		2		3		4		5		6		7		8		9	
	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1
3: 9 regions	13	13	8	0	2	1	0	0	2	0	0	0	2	2	0	1	0	1
4: 9 regions (balance)	13	16	8	0	2	1	0	0	2	0	0	0	2	2	0	1	1	1

Table 3: Regional mismatch scores for the predictions from Fits 3 and 4 shown in Figure 7 (b) and (c) respectively. Within each of the nine regions, the number n_0 of mismatched 0s and the number n_1 of mismatched 1s are given for the simulated distribution in 1990.

In conclusion, the results of this section are very encouraging; however, it is useful to have the likelihood results available as a yardstick by which to measure the performance of the method. Having verified the results against those from the likelihood method, the simulation approach may be applied with some confidence to an extension of the colonization model for which there is no tractable likelihood calculation. This will be treated in the next section. In the absence of any objective verification of the simulation results, much more care would be required in selecting the compatibility function and investigating the sensitivity of the results to different choices.

Overall, the example of this section serves to illustrate that the wealth of information that can be gathered about a stochastic process through simulation can be put to good effect in parameter estimation. A full likelihood analysis remains the preferred choice, but where this is difficult the *ad-hoc* method has the potential to be very effective.

1.3 Extension of the colonization model

The assumption in the original colonization model that colonizations of and from sites occur independently over a single time period might be considered questionable in the case of the woodlark. Woodlarks are strongly territorial birds, and it is unusual for any forestry site to be occupied by more than a single pair. Here a more realistic colonization mechanism will be suggested, although it should be stressed that the example is intended to be illustrative of the statistical methodology rather than to be a serious attempt at an ecological model. The changes render the one-step likelihood intractable, but the simulation method may readily be applied.

The new colonization mechanism is described as follows. Suppose at some time t there are K occupied sites. A random permutation ω of these sites is computed, and this is taken to be the order in which the individuals in the K sites colonize new sites at time $t+1$. Suppose the ordered sites are $i_1, \ldots, i_K$: thus $\omega(1) = i_1$ and the individuals in site i_1 at time t are the first to colonize new sites at time $t+1$, and so on. Colonizations from the first site i_1 occur exactly as in the original colonization model — site i_1 colonizes site h with probability $p_{i_1 h}$ and the event that i_1 colonizes site h_1 is independent of the event that i_1 colonizes site h_2 for any h_1, h_2. As usual, the probabilities $\{p_{ih}\}$ are parametrized as $p_{ih} = p_0 \exp(-a\,\delta_{ih} - b\,\varsigma_h)$ for i, $h = 1, \ldots, N$.

Once site i_1 has established its new colonies, the second site i_2 begins the same process. In the first instance, site i_2 also follows the rules of the original colonization model: any site

h_1 is colonized from i_2 with probability $p_{i_2 h_1}$, independently of whether or not any other site h_2 is also colonized from i_2. However, if site i_2 colonizes a site h that has already been colonized by site i_1, that colonization is temporarily displaced back to site i_2. This is intended to emulate the situation wherein a bird arrives at its preferred site, but, finding another bird already present, is forced onto a different site.

When a colonization is displaced back to site i_2, all colonization probabilities $\{p_{i_2 j}\}$ $(j = 1, \ldots, N)$ are recomputed with a new parameter c, designed to reduce the probability that a displaced bird will strike up a territory on a site that is already occupied. Let m_j be the number of times site j has already been colonized: for the time being $m_j \leq 1$ since the first site i_1 can establish at most a single colony in site j. The colonization probabilities for displaced birds are modelled as

$$p_{i_2 j} = p_0 \exp\left(-a\,\delta_{i_2 j} - b\,\varsigma_j - c\,m_j\right) ,$$

and the maximum over j of these displacement probabilities is computed. The displaced bird is assumed to move to the site j that has maximum displacement probability.

The process is repeated sequentially for all sites $i_3, \ldots, i_K$: original colonization probabilities are calculated according to the basic colonization model, and displacement probabilities with the extra parameter c are calculated whenever a colonization occurs to a site that has already been colonized by another site higher in the ordering. The aim of the analysis is the estimation of the colonization parameters a, b and p_0 and the additional displacement parameter c.

The two steps of colonization followed by displacement are needed in this model to ensure that the displacement process is visible. If the extra parameter c were placed into the colonization probabilities at the first stage, the outcome would simply be one of fewer colonizations from sites i_r low in the ordering, with no effect on the final number of occupied sites. The new model, on the other hand, allows the same number of colonizations as ever; in addition, it accounts for the process by which some of the less suitable sites are colonized once all the best sites are taken.

The implementation of the extended version of the colonization model is more time-consuming than that of the basic model, due to the sequential nature of the colonization process, and the second stage of displacement followed by re-colonization. For this reason, only five simulated observations are taken at every parameter vector $\boldsymbol{\theta}$, in contrast to the ten simulated observations taken in the previous section. Note that the parameter vector

$\boldsymbol{\theta}$ is now given by $\boldsymbol{\theta} = (a, b, p_0, c)$. The method of fixing the random seed is used here again, and the compatibility functions are those of equations (1) and (2) of section 1.2, with $R = 9$ regions. Details of the two fits are as follows.

Fit 1: 9 regions. Regional division as shown in Figure 2 (b). Compatibility function calculated according to equation (1), with $R = 9$.

Fit 2: 9 regions (balance). Regional division as shown in Figure 2 (b). Compatibility function calculated according to equation (2), rewarding only the balance in mismatched 0s and 1s, with $R = 9$.

The results are given in Table 4. Parameter estimates are provided for the four parameters a, b, p_0 and c, and standard errors are included, obtained by means of the parametric bootstrap as in the previous section. Only 50 bootstrap replicates were taken, due to the time required for each (typically 10 to 30 minutes on a Sun Ultra 10 server with 192MB RAM). No likelihood comparisons are available for the extended model, since only the basic model was amenable to likelihood calculations. Best attainable match results were obtained from 1000 simulated distributions as before; the BAM results are also shown in Figure 8.

Method of fit	Parameter estimates								BAM
	a	s.e.(a)	b	s.e.(b)	p_0	s.e.(p_0)	c	s.e.(c)	
1: 9 regions	0.0595	0.0163	6.43	0.795	0.0562	0.0256	1.71	0.500	261.0
2: 9 regions (balance)	0.115	0.0160	4.82	0.663	0.0485	0.0100	1.20	0.246	260.1

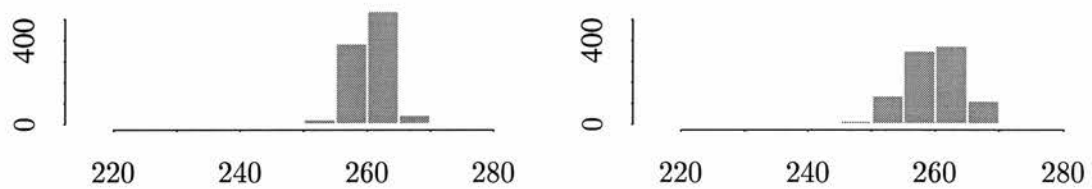
Table 4: Parameter estimates, estimated standard errors and mean BAM results from the two different methods of fit. Standard errors (s.e.) were obtained using 50 replicates of the parametric bootstrap. The final column gives the mean best attainable match from 1000 predictions using the parameters shown.

There are a number of notable features of Table 4. Firstly, the estimates of the colonization parameter b are substantially higher than those from all previous analyses, most of which have $2.5 < b < 3$. This indicates that the dependence of initial colonization on habitat suitability might be even more dramatic than suggested by the basic colonization model. The dependence enters the colonization probability p_{ih} through the component $\exp(-b\varsigma_h)$, so a high value of b will reduce the probability of colonization to all sites h that have a high value of ς_h, corresponding to poor habitat quality. If the colonization-displacement model is accurate, initial colonizations occur almost exclusively to those sites with the

best suitability scores. At $\delta_{ih} = 0$, for instance, the colonization probability p_{ih} under Fit 1 is halved as the suitability ς_h deteriorates from 0.0 to 0.11. Only 17 of the 316 available sites in 1990 attain a habitat quality score of 0.11 or below.

Those sites with poorer habitat suitability become colonized primarily through displacement. The high estimate of the displacement parameter c underlines the reluctance of woodlarks to establish territory on a site that is already occupied. Under Fit 1, for example, the maximum colonization probability of 0.056 is reduced to $0.056\exp(-1.71) = 0.010$ when a single colony is already present in the site — and the probability of a third colony being established is only $0.056\exp(-1.71 \times 2) = 0.0018$.

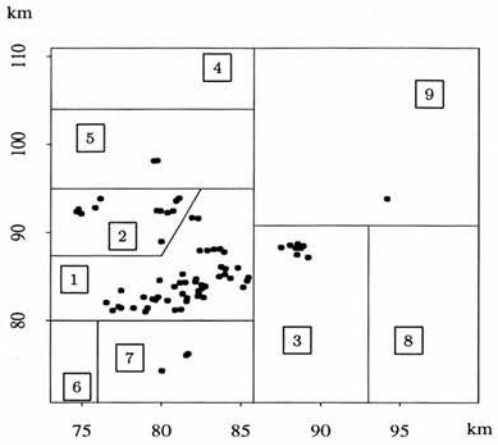
A final feature of Table 4 and Figure 8 is the increase in best attainable match scores under the extended model. The maximum likelihood estimates of Chapter 5 resulted in a typical mean BAM of around 254, while the simulation method which selects parameters on the basis of a high BAM score produced typical means of about 257 (Table 1). The extended model and extra parameter c have brought about a further improvement up to 260 or 261.



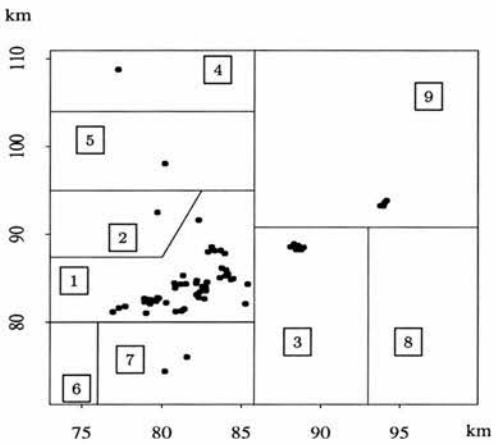
(a) Fit 1: 9 regions. (b) Fit 2: 9 regions (balance).

Figure 8: Best attainable match results from predicted distributions for 1990, using the parameters from Fits 1 and 2. Each histogram shows BAM results from 1000 simulated distributions for $N = 316$ sites in 1990, generated from the parameters indicated.

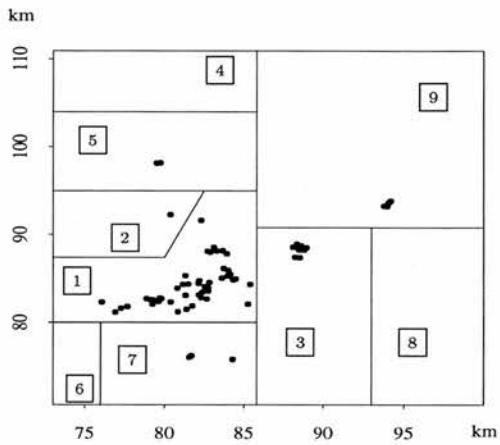
In the absence of a calculable likelihood, the usual means of assessing the significance of the new parameter c — based on Akaike's Information Criterion, for example — are not available. The improvement in BAM scores, however, can be examined. Suppose that the BAM scores under two sets of fitted parameters are normally distributed with equal but unknown variance. The equality of the two means may then be tested using the standard two-sample t-test. In the present example, Fit 2 of this section is compared against Fit 4 of section 1.2. These fits were chosen because they were obtained using the same form for the compatibility function, and the BAM scores have similar values for the sample variance and a roughly normal appearance from quantile-quantile plots. The difference between the respective means of 260.1 and 256.5 was found to be highly significant. Of course, this is due in part to the large sample size of 1000; the procedure cannot be regarded as a genuine test of the significance of the extra parameter.



(a) True distribution.



(b) Fit 1: 9 regions.



(c) Fit 2: 9 regions (balance).

Figure 9: Observed and predicted distributions for 1990, showing the division of the survey area into nine regions. Only occupied sites are shown. The observed distribution is shown in (a), and predicted distributions from the parameter estimates of Fits 1 and 2 are shown in (b) and (c) respectively.

Method of fit	Region																	
	1		2		3		4		5		6		7		8		9	
	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1	n_0	n_1
1: 9 regions	13	13	11	0	3	1	0	1	1	0	0	0	1	0	0	0	0	4
2: 9 regions (balance)	11	11	11	0	2	2	0	0	0	0	0	0	1	1	0	0	0	4

Table 5: Regional mismatch scores for the predictions from Fits 1 and 2 shown in Figure 9 (b) and (c) respectively. Within each of the nine regions, the number n_0 of mismatched 0s and the number n_1 of mismatched 1s are given for the simulated distribution in 1990.

Typical predictions from the parameter estimates of Fits 1 and 2 are shown in Figure 9, and Table 5 gives the associated breakdown of mismatched 0s and 1s over the 9 regions. The predicted distributions are a close approximation to the true distribution, although once again the occupation in region 2 has not been captured. There is also some tendency to over-predict presence in region 9.

This section has provided only a sketch of the full analysis that would be required for the extended model. A more complete treatment would demand greater attention to the outcomes using different compatibility functions, and checks on the results using a larger number of simulated observations at each parameter vector, and by changing the set values of the random seed. One possibility for an alternative compatibility function might involve keeping track of the final number of colonizations in each site, and using the best attainable match algorithm for abundance data outlined in Chapter 4. Indeed, with the dual stages of colonization and displacement, a model similar to the raw branching process described in Chapter 5 might be feasible, in which the probability that site i colonizes site h depends on the number of individuals present in site i. The displacement mechanism would ensure that the process does not run out of control, as a raw branching process might; however, the displacement stage also inhibits the calculation of the likelihood function. Such a model would therefore have to be analysed using alternatives such as the simulation method of this section.

2 Monte Carlo likelihood

A second simulation-based approach to parameter estimation is introduced in this section, in which simulated distributions are used to obtain a Monte Carlo estimate of the likelihood function for the basic colonization model. In view of the fact that the likelihood for the basic model has been approximated analytically through the work of Chapter 5, the Monte Carlo method could be regarded as obsolete for the colonization model itself; nor does it share the advantage of the full simulation approach of Section 1 in extending to the case where the one-step likelihood is incalculable. The approach is outlined briefly because the same ideas might be of use for models other than the colonization model, wherein a likelihood may be calculated over a single time-step but missing surveys cannot be accommodated.

The basis of the approach is very simple. The likelihood for the colonization model, $L\left(\boldsymbol{y}^{(T)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(0)}\right)$, is simply a discrete probability surface, and as such may be written as

the expectation of a conditional probability. For any events A and B it is generally true that $\mathbb{P}(A) = \mathbb{E}_B\{\mathbb{P}(A \mid B)\}$. In the present context, the event A may be substituted by $\boldsymbol{y^{(T)}}$, and the event B by $\boldsymbol{y^{(T-1)}}$, to give

$$L\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(0)}}\right) = \mathbb{E}\left\{L\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(T-1)}}, \boldsymbol{y^{(0)}}\right)\right\} . \tag{3}$$

Note that the expectation is taken over the conditional distribution of the random variable $\boldsymbol{y^{(T-1)}}$ given the initial distribution $\boldsymbol{y^{(0)}}$ and the parameters $\boldsymbol{\theta}$.

The conditional likelihood in (3) is precisely the one-step likelihood for the final time period, and as such is readily calculable if $\boldsymbol{y^{(T-1)}}$ is known. Furthermore, the expectation of any random variable may be approximated by a Monte Carlo estimate, which simply involves taking a number of samples from the distribution of the random variable and evaluating the sample mean. The Monte Carlo estimate of (3) is therefore given by the sample mean of $L\left(\boldsymbol{y^{(T)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(T-1)}}, \boldsymbol{y^{(0)}}\right)$ over a number of simulated values of $\boldsymbol{y^{(T-1)}}$. Simulated values of the penultimate distribution $\boldsymbol{y^{(T-1)}}$ are obtained by conditioning on the parameters $\boldsymbol{\theta}$ and the initial distribution $\boldsymbol{y^{(0)}}$.

The Monte Carlo method above is a form of stochastic integration, which is a technique that has become exceptionally popular in recent years — notably through the development of Markov chain Monte Carlo methods (McMC: Gilks *et al.* 1996). The McMC technique is an extension of the basic Monte Carlo approach to estimation of an expectation, whereby samples from the distribution of the random variable over which the expectation is to be taken are obtained by sampling from the equilibrium distribution of a carefully constructed Markov chain. Gibson (1997), and Gibson & Austin (1996), have used a variety of stochastic integration techniques — including Markov chain Monte Carlo — for the fitting of spatio-temporal models in epidemiology. The epidemiological and ecological fields are closely related, and there is much room for exchange of ideas between them. Geyer (1994) has also pioneered McMC methods for examples involving correlated and missing data.

Stochastic integration techniques are rarely straightforward to apply, since they are mostly used in cases where samples must be taken from a random variable whose distribution cannot be fully enumerated. In the case of the colonization model, however, $\boldsymbol{y^{(T-1)}}$ can be simulated directly from the model, so these difficulties are circumvented. In particular, there is no need to consider the application of a Markov chain Monte Carlo scheme in this instance.

The Monte Carlo approach to likelihood estimation was applied to the woodlark data,

and the results are given in Table 6. Only the survey data from the initial and final years 1986 and 1990 were included in the analysis, so the parameter estimates are directly comparable with those from the analytic likelihood calculations of Chapter 5. The Monte Carlo expectation was evaluated at the parameter vector $\boldsymbol{\theta}$ by taking the sample mean of ten simulated values of $\boldsymbol{y}^{(\boldsymbol{T-1})}$. The resulting approximate likelihood function was stabilized by fixing the random seed, as described in Section 1. The maximization was carried out numerically using the routine `amoeba` of Press *et al.* (1988). Variance estimation was performed using 100 replicates of the parametric bootstrap, although in the light of the observed smoothness of the estimated likelihood surface it might be feasible to use the analytic variance estimates of Chapter 5.

	Parameter estimates						
Method of fit	a	s.e.(a)	b	s.e.(b)	p_0	s.e.(p_0)	$\log L$
Monte Carlo	0.116	0.0398	2.56	0.369	0.0373	0.00994	-131.77
Analytic	0.117	0.0491	2.57	0.392	0.0368	0.00880	-131.66

Table 6: Parameter estimates, estimated standard errors and log-likelihood results from the Monte Carlo likelihood estimates of this section, and the second-order analytic maximum likelihood estimates corresponding to Fit 5 of Chapter 5. Standard errors (s.e.) were obtained using 100 replicates of the parametric bootstrap in the first fit, and using the analytic method of Chapter 5 in the second. The Monte Carlo likelihood was estimated as the sample mean of the appropriate one-step likelihood over ten simulations, while the analytic likelihood was obtained using the technique of Fit 5, Chapter 5.

The conspicuous feature of Table 6 is the similarity between the results from the Monte Carlo method of likelihood estimation and the analytic estimates of Chapter 5. Parameter estimates, standard error estimates and log-likelihood estimates are all extremely close, despite each having been derived in a different manner for the two methods. The Monte Carlo estimate of -131.77 for the log-likelihood at the final parameter values tallies closely with the analytic estimate of -131.73 at the same parameter values, calculated using the method of Fit 5 in Chapter 5. These results suggest that the Monte Carlo method has the potential to be very successful. The method is also economical in terms of both running time and programming time.

3 Concluding remarks

This chapter has presented two viable approaches to parameter estimation based on computer simulations. The full simulation technique of Section 1 has advantages in generality and flexibility of application, but disadvantages in its subjective nature and in the lack of a theoretical basis to the parameter estimates. Unlike maximum likelihood parameter estimates, there is no asymptotic theory available for the parameter estimates from the full simulation method. The variance of the estimators, and confidence regions associated with the parameters, must therefore be estimated using further simulation methods such as the bootstrap. However, the bootstrap methods themselves are restricted in their application to spatial data, due to the difficulties of resampling from the data while preserving the spatial structure. This accounts for the use of the parametric bootstrap in the examples of this chapter. The use of one simulation technique on top of another, meanwhile, can be computationally expensive.

The Monte Carlo approach to parameter estimation does not have the same flexibility as the full simulation approach, but does have the advantage of producing genuine approximations to the maximum likelihood estimates. The approach is very usable, as exemplified by the rapid convergence of the maximization process to parameter estimates very close to the maximum likelihood estimates. There is relatively little variability in the conditional likelihood $L\left(\boldsymbol{y}^{(T)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(T-1)}, \boldsymbol{y}^{(0)}\right)$ for different simulated values of $\boldsymbol{y}^{(T-1)}$, especially in the vicinity of the final parameter estimates, and this helps to smooth the estimated likelihood surface. There would probably be little difficulty in applying asymptotic techniques for variance estimation using the Monte Carlo likelihood surface, although this has not been attempted.

In conclusion, parameter estimation via analytic calculation of the likelihood is considered the ideal approach, due to theoretical properties of the MLE such as asymptotic efficiency, consistency, normality and unbiasedness (Cox & Hinkley 1974). The Monte Carlo estimate of the likelihood is useful for occasions where a conditional likelihood is calculable and the random variable on which conditioning takes place may be simulated from the model. When there is no possibility of an analytic likelihood calculation, the full simulation method of Section 1 is likely to provide good parameter estimates, but lacks the theoretical support.

Chapter 7

Discussion and conclusions

The work of this thesis has touched on a number of points that might be of relevance to ecologists and statisticians investigating spatial and temporal patterns in wildlife distribution and abundance. This section provides a summary of the broad results and conclusions. Some of the points are specific to the material presented, while others have wider applicability in the discipline.

Analysis of population trends

Chapter 2 described an approach to the analysis of large-scale monitoring data for farmland birds, using generalized additive models. Detection of trends in bird populations is an area into which large amounts of money and effort have been directed over a period of some decades, and it is surprising that the use of a well-known statistical technique has been overlooked up to the present time. Indeed, while generalized additive models are widely used among statisticians, a literature search over the period 1995 to 1998 revealed only a handful of papers in ecological journals involving GAMs. Of these, the majority related to vegetation dynamics, and none were found that employed generalized additive models in ornithology.

Part of the reason that modern statistical methods have been slow to enter the ornithological literature is the long-term nature of the monitoring schemes from which census data arise. The British Common Birds Census was initiated over twenty years before generalized additive models first entered the statistical literature (Hastie & Tibshirani 1984; 1990); and at the time that the Mountford method was developed, an analysis of

the form presented in Chapter 2 would have been computationally infeasible (Mountford 1982). Analysts were forced to develop methods that could be handled with the tools available to them, and it can take some time for a technique that has been in place for many years to be abandoned in favour of new ideas. Nonetheless, the situation underlines a responsibility on the part of both statisticians and ecologists to ensure that the best available analysis techniques are widely known.

Similar considerations apply to survey design. The new British Breeding Bird Survey (Gregory *et al.* in press) is designed to cover a random sample of survey sites, and adopts an approach to abundance estimation based on distance sampling (Buckland *et al.* 1993). As such, it is amenable to far more rigorous analysis and interpretation than its predecessor, the CBC. Unfortunately it will be many years before sufficient data have been collected under the BBS to reveal patterns in population trend, and by that time its own design might seem out of date. Indeed, it is almost to be hoped that technology will advance sufficiently in the future to make today's survey designs and analysis methods obsolete. The overriding lesson, however, is the need for ongoing and effective communication between statisticians and ornithologists.

Parameter estimation in the colonization model

Chapter 3 introduced the colonization model, which is an example of a class of mechanistic spatio-temporal models that has received little attention in the literature to date. The primary aim of the material in Part II of the thesis was the development of statistical methodology to enable the colonization model to be fitted to data collected at irregular or infrequent points in time. This aim has been accomplished, but perhaps the most interesting outcome of the investigation lies in the comparative performances of the various methods employed.

A maximum likelihood approach to the problem was developed in Chapter 5, whereby the colonization mechanism was treated as a modified branching process, and the likelihood was calculated by means of a branching process recurrence relation. An alternative Monte Carlo likelihood approach outlined in Chapter 6 gave almost identical results. The amount of effort involved in the development of the two methods, on the other hand, bore no comparison: while some months were required for the analytic method, the Monte Carlo approach was developed, coded and applied within a single day.

In the light of this, it is difficult not to recommend that Monte Carlo methods of parameter

estimation be given the same credibility as analytic approaches. This is already becoming the case in some fields, owing to the recent surge of activity in Markov chain Monte Carlo techniques. Nonetheless, there remains a temptation to favour analytic methods — and consequently favour those models that are amenable to analytic treatment. Analytic approaches tend to be more interesting than simulation methods, and they also have a solid basis in convention — something that lends respectability to almost any practice. There is little convention surrounding simulation approaches, on the other hand, since powerful computers have only recently become widely available.

Of course, the fact that the analytic and Monte Carlo likelihood estimates coincided for the colonization model with the woodlark data does not mean that this will always be true. The Monte Carlo likelihood approach in this instance could only reasonably fail if the probabilities $L\left(\boldsymbol{y}^{(T)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y}^{(T-1)}, \boldsymbol{y}^{(0)}\right)$ were very variable for different simulated realizations of $\boldsymbol{y}^{(T-1)}$, and this was found not to be the case. For a general problem it will often happen that analytic and simulation-based approaches both involve various approximations, making it unclear which is better. For the colonization model, however, although the analytic approach of Chapter 5 was by far the more interesting exercise, the resulting methodology is less general than the Monte Carlo approach and offers no improvement in results.

Chapter 6 detailed a second simulation-based approach to parameter estimation, facilitated by the best attainable match (BAM) algorithm developed in Chapter 4. The best attainable match procedure was shown in Chapter 4 to produce a measure of the similarity of two spatial distributions that captures global likeness much more effectively than traditional measures such as the simple matching coefficient. This said, it became clear in Chapter 6 that no single summary index will ever be sufficient to convey a truly comprehensive notion of similarity in all circumstances. For a spatial distribution that is dispersed between a number of disjoint regions, such as the woodlark distribution in Thetford Forest, similarity is best ascertained by tabulating for each region the total number of mismatched sites, the number of mismatched sites with status 0, and the number of mismatched sites with status 1. Distributions are regarded as similar if the total number of mismatches is small, and the number of mismatched 0-sites is approximately equal to the number of mismatched 1-sites. The advantage of the BAM approach is still retained if the tabulation relates to those sites that remain mismatched after within-clique swaps have been performed.

The full simulation approach to parameter estimation developed in Chapter 6 used this

regional notion of similarity as the basis of a function over the parameter space, intended to measure the compatibility of each parameter vector with the observed spatial distribution. Although developed for the colonization model, the compatibility method could be used in almost any problem involving parameter estimation. The compatibility function was evaluated at the parameter vector $\boldsymbol{\theta}$ by comparing the observed distribution against a spatial distribution generated using parameters $\boldsymbol{\theta}$, and was maximized with respect to the parameters to yield parameter estimates.

The examples in Chapter 6 demonstrated that the full simulation approach is capable of producing very reasonable parameter estimates — often close to the maximum likelihood estimates. Although very encouraging, the full simulation approach is not recommended as a replacement for likelihood methods — analytic or Monte Carlo — when these are available. This is for two reasons. Firstly, there is inevitable subjectivity in the choice of compatibility function, and the final parameter estimates might be sensitive to this choice. Secondly, there is considerable theoretical advantage in the use of the likelihood function, in respect of the asymptotic properties of the maximum likelihood estimator. The full simulation approach comes into its own when dealing with a model for which a likelihood surface is not calculable, and an example of its application in this context was given in Chapter 6.

A practical problem that occurred with the full simulation method when applied to the colonization model with the woodlark data was the difficulty in finding global maxima of the compatibility function. The issue of starting point selection is paramount here, and the problem is likely to be generally applicable unless an exceptionally smooth compatibility function can be found. By contrast, there were no difficulties in finding the global maximum of the likelihood surface in the woodlark problem; the maximization routine converged to the same result from starting points all over the parameter space.

Pseudo-likelihood functions

Before closing, it is useful to include some discussion of the use of pseudo-likelihood functions for parameter estimation. The term *pseudo-likelihood* is often used to denote a function that is related to the full likelihood but differs in some respect. Examples are conditional likelihoods, marginal likelihoods and profile likelihoods (Kalbfleisch 1982; Gong & Samaniego 1981), which are designed to eliminate nuisance parameters. Besag (1975) used the term in a rather different sense, to denote a product of full conditionals over all

observations — that is, $\prod_{i=1}^{N} L\Big(y_i \Big| \{y_j : j \neq i\}\Big)$ for observations $\{y_i\}_{i=1}^{N}$. The present thesis has considered yet another pseudo-likelihood, termed the *product-likelihood*, which is given by $\prod_{i=1}^{N} L(y_i)$. Both Besag's pseudo-likelihood and the product-likelihood might be regarded as attempts to evade the enumeration of the true correlation structure of the observations $\{y_i\}$ when calculating the likelihood function.

In the case of the colonization model, it was shown in Chapter 5 that the product-likelihood is actually a first-order approximation to the full likelihood. This lent much theoretical justification to its use, and simplified the ensuing analyses considerably. Asymptotic variance estimation, for example, is easiest when the full likelihood is a product of the single-observation likelihoods. The fact that the product-likelihood is a first-order approximation to the full likelihood is also an indication that correlations between observations are fairly weak according to the colonization model. This is especially true when the model is applied to only a few time-points, as with the woodlark data. Correlations between observations arise most strongly if there is sufficient time between surveys for occupation to be firmly established in clusters of remote sites — this occupation perhaps having arisen from a single colonization to one of the sites.

The Monte Carlo likelihood approach of Chapter 6, which does incorporate the correlation structure of the observations, provided almost identical likelihood estimates to those of the analytic product-likelihood approaches. This provides further evidence that the analytic results of Chapter 5 were not adversely affected by ignoring the correlation structure of the observations, at least in the woodlark case. The use of a pseudo-likelihood function should always be regarded with a measure of caution, however, as results are not guaranteed to be as good as those found in Chapter 5. Geyer & Thompson (1992) give examples where both Besag's pseudo-likelihood and a conditional likelihood yield poor parameter estimates, and no doubt there are also instances where the same is true of the product-likelihood function.

General remarks about the colonization model

There are two final points of interest about the colonization model. The first regards its relationship to the class of stochastic processes known as contact processes (e.g. Mollison 1977; Bartholomew 1982). A contact process is one in which some phenomenon — typically a disease — is assumed to spread through a region by consequence of a series of contacts between individuals. When an individual forms a contact, the probability that its target is located at distance d from itself is given by a distribution called the *contact distribution.*

Contacts generally occur at a rate governed by some other process.

The colonization model is a form of contact process in which individuals correspond to occupied sites, and contacts to colonizations. Contacts occur at discrete intervals, and the probability that individual i makes a contact at distance d from itself, given that a contact is made, is

$$\frac{1 - \prod_{h:\,\delta_{ih}=d}(1 - p_{ih})}{1 - \prod_{h=1}^{N}(1 - p_{ih})}.$$

The contact distribution for the colonization model is therefore governed by the specified form of the probabilities $\{p_{ih}\}$.

Mollison (1977) showed that asymptotic properties of a contact process, such as the velocity of spread, differ considerably according to the form of the contact distribution. Different contact distributions also require different theoretical treatment. In the same paper it was indicated that the issues of real-time (non-asymptotic) properties of contact processes, and of parameter estimation in the processes, had both received scant attention; the same is largely true today. The present work has shown how parameter estimation may be carried out in one particular contact process, for an arbitrary contact distribution, and has provided a framework for further investigation of the short-term properties of the process.

Finally, the modified branching process used in Chapter 5 to facilitate likelihood calculations for the colonization model is in fact a form of *controlled* branching process, in which feedback from the current state is used to determine the branching mechanism. This is another area that has been identified by a number of authors as one in which research has been lacking (Mode 1971; Asmussen 1982). In the case of the colonization model, the branching probabilities were suppressed according to the number of individuals of each type that were present in the population. (Recall that the number of individuals of type i present in generation t corresponds to the number of colonies present in site i at that time.) The control was imposed for the colonization model to ensure that no matter how many individuals of type i were present in any generation, there was no greater probability of collectively producing offspring than if only one individual of type i were present. The same methods, however, could be used to control the mechanism according to other criteria.

Bibliography

Aharoni, R., Meshulam, R., & Wajnryb, B. (1995). Group-weighted matching in bipartite graphs. *Journal of Algebraic Combinatorics*, **4**, 165–171.

Aitkin, M., Anderson, D., Francis, B., & Hinde, J. (1989). *Statistical Modelling in GLIM.* Oxford University Press, Oxford.

Asmussen, S. (1982). Branching processes. In (Kotz & Johnson 1982), volume 1, pp. 316–319.

Augustin, N. H., Borchers, D. L., Clarke, E. D., Buckland, S. T., & Walsh, M. (in review). Spatio-temporal modelling for the annual egg production method of stock assessment using generalised additive models. *Canadian Journal of Fisheries and Aquatic Sciences.*

Augustin, N. H., Mugglestone, M. A., & Buckland, S. T. (1996). An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339–347.

Baillie, S. R. (1990). Integrated population monitoring of breeding birds in Britain and Ireland. *Ibis*, **132**, 151–166.

Baillie, S. R., Gregory, R. D., & Siriwardena, G. M. (1997). Farmland bird declines: patterns, processes and prospects. *BCPC Symposium Proceedings No. 69: Biodiversity and Conservation in Agriculture*, pp. 65–87.

Bak, R. P. M. & Nieuwland, G. (1995). Long-term change in coral communities along depth gradients over leeward reefs in the Netherlands-Antilles. *Bulletin of Marine Science*, **56**, 609–619.

Barnard, G. A. (1963). Discussion of Professor Bartlett's paper. *Journal of the Royal Statistical Society, Series B*, **25**, 294.

Bartholomew, D. J. (1982). *Stochastic models for social processes.* Wiley, Chichester.

Besag, J. (1977). Discussion on spatial contact models for ecological and epidemic spread (by D. Mollison). *Journal of the Royal Statistical Society, Series B*, **39**, 315–316.

Besag, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24**, 179–195.

Bishop, J. (1989). *Pascal precisely*, 2nd edition. Addison-Wesley, Wokingham, UK.

Borchers, D. L., Buckland, S. T., Priede, I. G., & Ahmadi, S. (1997). Improving the precision of the daily egg production method using generalized additive models. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 2727–2742.

Borchers, D. L., Zucchini, W., & Fewster, R. M. (in press). Mark-recapture models for line transect surveys. *Biometrics*.

Bowden, C. G. R. & Green, R. E. (1992). *The ecology and management of woodlarks on pine plantations in the Thetford and Sandlings Forests.* RSPB Research Department, The Lodge, Sandy, Bedfordshire. Research report.

Bradu, D. & Mundlak, Y. (1970). Estimation in lognormal linear models. *Journal of the American Statistical Association*, **65**, 198–211.

Bramson, M., Cox, J. T., & Durrett, R. R. (1998). A spatial model for the abundance of species. *Annals of Probability*, **26**, 658–709.

Brooks, S. P. & Morgan, B. J. T. (1994). Automatic starting point selection for function optimization. *Statistics and Computing*, **4**, 173–177.

Brualdi, R. A. & Ryser, H. J. (1991). *Combinatorial Matrix Theory.* Cambridge University Press, Cambridge.

Buckland, S. T. (1984). Monte Carlo confidence intervals. *Biometrics*, **40**, 811–817.

Buckland, S. T., Anderson, D. R., Burnham, K. P., & Laake, J. L. (1993). *Distance Sampling: Estimating Abundance of Biological Populations.* Chapman and Hall, London.

Buckland, S. T., Cattanach, K. L., & Anganuzzi, A. A. (1992). Estimating trends in abundance of dolphins associated with tuna in the eastern tropical Pacific Ocean, using sightings data collected on commercial tuna vessels. *Fishery Bulletin*, **90**, 1–12.

Buckland, S. T. & Elston, D. A. (1993). Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology*, **30**, 478–495.

Butcher, G. S. (1990). Audubon Christmas Bird Counts. In (Sauer & Droege 1990), p. 5.

Buttell, L., Cox, J. T., & Durrett, R. (1993). Estimating the critical values of stochastic growth models. *Journal of Applied Probability*, **30**, 455–461.

Cain, A. J. & Harrison, G. A. (1958). An analysis of the taxonomists' judgement of affinity. *Proceedings of the Zoological Society of London*, **131**, 85–98.

Carroll, L. (1876). *The Hunting of the Snark: an Agony in Eight Fits.* Macmillan, New York.

Chambers, J. M. & Hastie, T. J. (1993). *Statistical models in S*, chapter 7. Chapman and Hall, London.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*, chapter 9. Chapman and Hall, London.

Cox, J. T. & Durrett, R. (1988). Limit theorems for the spread of epidemics and forest fires. *Stochastic Processes and their Applications*, **30**, 171–191.

Csima, J. & Lovasz, L. (1992). A matching algorithm for regular bipartite graphs. *Discrete Applied Mathematics*, **35**, 197–203.

Daniels, H. E. (1995). A perturbation approach to non-linear deterministic epidemic waves. In D. Mollison (ed.), *Epidemic Models: their structure and relation to data.* Cambridge University Press, Cambridge.

Dekel, E. & Sahni, S. (1984). A parallel matching algorithm for convex bipartite graphs and applications to scheduling. *Journal of Parallel and Distributed Computing*, **1**, 185–205.

Digby, P. G. N. & Kempton, R. A. (1987). *Multivariate Analysis of Ecological Communities.* Chapman and Hall, London.

Droege, S. (1990). The North American Breeding Bird Survey. In (Sauer & Droege 1990), pp. 1–4.

Egerváry, E. (1955). On combinatorial properties of matrices. *Logistics Papers (Issue 11), Paper 4, George Washington University*, pp. 1–11. Translated by H. W. Kuhn.

Engel, B. & Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1–22.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, 3rd edition. John Wiley & Sons, London.

Findlay, W. & Watt, D. A. (1985). *Pascal*, 3rd edition. Pitman, London.

Ford, L. R. & Fulkerson, D. R. (1956). Maximal flow through a network. *Canadian Journal of Mathematics*, **8**, 399.

Fuller, R. J., Gregory, R. D., Gibbons, D. W., Marchant, J. H., Wilson, J. D., Baillie, S. R., & Carter, N. (1995). Population declines and range contractions among lowland farmland birds in Britain. *Conservation Biology*, **9**, 1425–1441.

Fuller, R. J., Marchant, J. H., & Morgan, R. A. (1985). How representative of agricultural practice in Britain are Common Birds Census farmland plots? *Bird Study*, **32**, 56–70.

Geissler, P. H. & Noon, B. R. (1981). Estimates of avian population trends from the North American Breeding Bird Survey. In C. J. Ralph & J. M. Scott (eds), *Estimating numbers of terrestrial birds*, pp. 42–51. Studies in Avian Biology **6**.

Geissler, P. H. & Sauer, J. R. (1990). Topics in route regression analysis. In (Sauer & Droege 1990), pp. 54–57.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 261–274.

Geyer, C. J. & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, **54**, 657–699.

Gibbons, D. W., Reid, J. B., & Chapman, R. A. (1993). *The new atlas of breeding birds in Britain and Ireland, 1988-1991.* Poyser, London.

Gibson, G. J. (1997). Markov chain Monte Carlo methods for fitting spatio-temporal stochastic models in plant epidemiology. *Applied Statistics*, **46**, 215–233.

Gibson, G. J. & Austin, E. J. (1996). Fitting and testing spatio-temporal stochastic models with applications in plant epidemiology. *Plant Pathology*, **45**, 172–184.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (eds) (1996). *Markov chain Monte Carlo in practice.* Chapman and Hall, London.

Goldfarb, D. & Hao, J. X. (1993). On the maximum capacity augmentation algorithm for

the maximum flow problem. *Discrete Applied Mathematics*, **47**, 9–16.

Gong, G. & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Annals of Statistics*, **9**, 861–869.

Gotway, C. A. & Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological and Environmental Statistics*, **2**, 157–178.

Greenwood, J. J. D. & Baillie, S. R. (1991). Effects of density dependence and weather on population changes of English passerines using a non-experimental paradigm. *Ibis*, **133**(suppl. 1), 121–133.

Gregory, R. D., Baillie, S. R., & Bashford, R. I. (in press). Monitoring breeding birds in the United Kingdom. In *Proceedings of the 13th International Conference of European Bird Census Councils*, Estonia 1995.

Hagemeijer, W. & Verstrael, T. (eds) (1994). *Bird Numbers 1992: Distribution, Monitoring and Ecological Aspects; Proceedings 12th International Conference of IBCC and EOAC.* SOVON, Beek-Ubbergen, The Netherlands.

Hammersley, J. M. & Welsh, D. J. A. (1965). First-passage percolation, subadditive processes, stochastic networks and generalized renewal theory. In J. Neyman & L. M. LeCam (eds), *Bernouilli, Bayes, Laplace Anniversary Volume*, pp. 61–110. Springer-Verlag, Berlin.

Hartsfield, N. & Ringel, G. (1990). *Pearls in graph theory: a comprehensive introduction.* Academic Press, Boston, Mass., USA.

Hastie, T. J. & Tibshirani, R. J. (1984). *Generalized additive models.* Technical Report 98, Department of Statistics, Stanford University.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hengeveld, R. (1989). *Dynamics of biological invasions.* Chapman and Hall, London.

Hiby, L. & Lovell, P. (1990). Computer aided matching of natural markings: a prototype system for grey seals. In P. S. Hammond, S. A. Mizroch, & G. P. Donovan (eds), *Individual Recognition of Cetaceans*, pp. 57–61. Report of the International Whaling Commission,

Special Issue **12**, Cambridge.

Hopcroft, J. E. & Karp, R. M. (1973). An $O(n^{2.5})$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal of Computing*, **2**, 225–231.

Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, **37**, 241–272.

James, F. C., McCulloch, C. E., & Wiedenfeld, D. A. (1996). New approaches to the analysis of population trends in land birds. *Ecology*, **77**, 13–27.

James, F. C., McCulloch, C. E., & Wolfe, L. E. (1990). Methodological issues in the estimation of trends in bird populations with an example: the Pine Warbler. In (Sauer & Droege 1990), pp. 84–97.

Jockusch, W. (1994). Perfect matchings and perfect squares. *Journal of Combinatorial Theory, Series A*, **67**, 100–115.

Kalbfleisch, J. D. (1982). Pseudo-likelihood. In (Kotz & Johnson 1982), volume 7, pp. 324–327.

Karp, R. M., Upfal, E., & Wigderson, A. (1985). Constructing a perfect matching is in random NC. In *Proceedings of the 17th ACM Symposium on Theory of Computing*, 1985.

Karzanov, A. V. (1987). Maximum matching of given weight in complete and complete bipartite graphs. *Cybernetics*, **23**, 8–13.

Kendall, D. G. (1948). A form of wave propagation associated with the equation of heat conduction. *Proceedings of the Cambridge Philosophical Society*, **44**, 591–594.

Kendall, D. G. (1965). Mathematical models of the spread of infection. In *Mathematics and Computer Science in Biology and Medicine*, pp. 213–225. M.R.C., H.M.S.O..

Kernighan, B. W. & Ritchie, D. M. (1978). *The C Programming Language.* Prentice-Hall, London.

Kim, T. & Chwa, K.-Y. (1987). An $O(n \log n \log \log n)$ parallel maximum matching algorithm for bipartite graphs. *Information Processing Letters*, **24**, 15–17.

König, D. (1936). *Theorie der endlichen und unendlichen Graphen: kombinatorische*

Topologie der Streckenkomplexe. Akademische Verlagsgesellschaft, Leipzig.

Kotz, S. & Johnson, N. L. (eds) (1982). *Encyclopedia of Statistical Sciences.* Wiley, New York; Chichester.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, **2**, 83–97.

Lenstra, J. K., Rinnooy Kan, A. H. G., & Schrijver, A. (eds) (1991). *History of mathematical programming.* Elsevier, Amsterdam.

Liang, Y. D. & Blum, N. (1995). Circular convex bipartite graphs: maximum matching and Hamiltonian circuits. *Information Processing Letters*, **56**, 215–219.

Link, W. A. & Sauer, J. R. (1997a). Estimation of population trajectories from count data. *Biometrics*, **53**, 488–497.

Link, W. A. & Sauer, J. R. (1997b). New approaches to the analysis of population trends in land birds: comment. *Ecology*, **78**, 2632–2634.

Marchant, J. H. & Gregory, R. D. (1994). Recent population changes among seed-eating passerines in the United Kingdom. In (Hagemeijer & Verstrael 1994), pp. 87–95.

Marchant, J. H., Hudson, R., Carter, S. P., & Whittington, P. (1990). *Population trends in British breeding birds.* British Trust for Ornithology, Tring, Hertfordshire.

Marriott, F. H. C. (1979). Barnard's Monte Carlo tests: how many simulations? *Applied Statistics*, **28**, 75–77.

Maynard Smith, J. (1974). *Models in Ecology.* Cambridge University Press, Cambridge.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*, 2nd edition. Chapman and Hall, London.

McKean, H. P. (1975). Application of Brownian motion to the equation of Kolmogorov-Petrovskii-Piscunov. *Communications on Pure and Applied Mathematics*, **28**, 323–331.

Metz, J. A. J. & van den Bosch, F. (1995). Velocities of epidemic spread. In D. Mollison (ed.), *Epidemic Models: their structure and relation to data.* Cambridge University Press, Cambridge.

Mode, C. J. (1971). *Multitype Branching Processes.* Elsevier, New York.

Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society, Series B*, **39**, 283–326.

Morgan, B. J. T. (1984). *Elements of Simulation.* Chapman and Hall, London.

Mountford, M. D. (1982). Estimation of population fluctuations with application to the Common Birds Census. *Applied Statistics*, **31**, 135–143.

Mountford, M. D. (1985). An index of population change with application to the Common Birds Census. In B. J. T. Morgan & P. M. North (eds), *Statistics in Ornithology*, pp. 121–132. Springer-Verlag, Berlin.

Nelder, J. A. & Mead, R. (1965). *Computer Journal*, **7**, 308–313.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, **135**, 370–384.

North, P. M. (1977). A novel clustering method for estimating numbers of bird territories. *Applied Statistics*, **26**, 149–155.

North, P. M. (1979). A novel clustering method for estimating numbers of bird territories: an addendum. *Applied Statistics*, **28**, 300–301.

O'Connor, R. J. & Shrubb, M. (1986). *Farming and Birds.* Cambridge University Press, Cambridge.

Pannekoek, J. & van Strien, A. (1996). *TRIM (TRends and Indices for Monitoring data).* Research paper 9634, Statistics Netherlands, Voorburg, The Netherlands. Freeware available from http://neon.vb.cbs.nl/sec_lmi_e/incpro/tri001p1.htm.

Peach, W. J. & Baillie, S. R. (1994). Implementation of the Mountford indexing method for the Common Birds Census. In (Hagemeijer & Verstrael 1994), pp. 653–662.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1988). *Numerical Recipes in C.* Cambridge University Press, Cambridge.

Renshaw, E. (1991). *Modelling biological populations in space and time.* Cambridge University Press, Cambridge.

Ripley, B. D. (1988). *Statistical inference for spatial processes.* Cambridge University Press, Cambridge.

SAS Institute Inc. (1996). *SAS/STAT Software: Changes and Enhancements through release 6.11.* SAS Institute Inc., Cary, NC, USA.

Sauer, J. R. & Droege, S. (eds) (1990). *Survey designs and statistical methods for the estimation of avian population trends.* US Fish and Wildlife Service, Biological Report, **90**(1).

Sauer, J. R. & Geissler, P. H. (1990). Annual indices from route regression analyses. In (Sauer & Droege 1990), pp. 58–62.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719–727.

Shiloach, Y. & Vishkin, U. (1982). An $O(n^2 \log n)$ parallel max-flow algorithm. *Journal of Algorithms*, **3**, 128–146.

Siriwardena, G. M., Baillie, S. R., Buckland, S. T., Fewster, R. M., Marchant, J. H., & Wilson, J. D. (1998). Trends in the abundance of farmland birds: a quantitative comparison of smoothed Common Birds Census indices. *Journal of Applied Ecology*, **35**, 24–43.

Statistical Sciences Inc. (1993). *S-PLUS Reference Manual.* StatSci, a division of Mathsoft, Inc., Seattle.

Taub, S. R. (1990). Smoothed scatterplot analysis of long-term Breeding Bird Census data. In (Sauer & Droege 1990), pp. 80–83.

ter Braak, C. J. F., van Strien, A. J., Meijer, R., & Verstrael, T. J. (1994). Analysis of monitoring data with many missing values: which method? In (Hagemeijer & Verstrael 1994), pp. 663–673.

Thomas, L. (1996). Monitoring long-term population change: why are there so many analysis methods? *Ecology*, **77**, 49–58.

Tucker, G. M. & Heath, M. F. (1994). *Birds in Europe: their conservation status.* BirdLife International, Cambridge. (BirdLife International Series No. 3).

Appendix A

Error in linear and quadratic exponential approximations to $s_i^{(t)}(x)$ for $t = 1$ and $t = 2$

When $t = 1$ or $t = 2$, the function $s_i^{(t)}(x)$ takes the form

$$s_i^{(t)}(x) = \prod_{k=1}^{M} (1 - \alpha_k x), \tag{1}$$

where $0 \leq \alpha_k \leq 1$ for all k, $0 \leq x \leq 1$, and M is an integer. The linear exponential approximation to $s_i^{(t)}(x)$ is given by

$$\exp\left\{-\left(\sum_{k=1}^{M} \alpha_k\right) x\right\},$$

and the quadratic exponential approximation by

$$\exp\left\{-\left(\sum_{k=1}^{M} \alpha_k\right) x - \left(\sum_{k=1}^{M} \alpha_k^2\right) \frac{x^2}{2}\right\}.$$

The function $s_i^{(t)}(x)$, and the associated exponential approximations, are symmetric in the components of $\boldsymbol{\alpha}$ in that any permutation of the components will not affect the result. Throughout much of this section, the order of components of $\boldsymbol{\alpha}$ is ignored: any permutation of $(\alpha_1, \ldots, \alpha_M)$ is considered indistinguishable from $(\alpha_1, \ldots, \alpha_M)$.

Definition: Let $S_1 = \sum_{k=1}^{M} \alpha_k$ and $S_2 = \sum_{k=1}^{M} \alpha_k^2$. The *linear error function* ϵ_1 is defined as

$$\epsilon_1(x) = \exp\{-S_1 x\} - \prod_{k=1}^{M}(1 - \alpha_k x), \tag{2}$$

and the *quadratic error function* ϵ_2 is defined as

$$\epsilon_2(x) = \exp\left\{-S_1 x - S_2\frac{x^2}{2}\right\} - \prod_{k=1}^{M}(1 - \alpha_k x). \tag{3}$$

The aim of the material in this section is to provide upper and lower bounds for $\epsilon_1(x)$ and $\epsilon_2(x)$, subject to fixed S_1 and S_2. Given S_1 and S_2, those $\alpha_1, \ldots, \alpha_M$ are found that extremize $\epsilon_1(x)$ and $\epsilon_2(x)$ and satisfy $\sum_{k=1}^{M} \alpha_k = S_1$ and $\sum_{k=1}^{M} \alpha_k^2 = S_2$.

Note that the linear and quadratic error functions are always non-negative, since

$$\begin{aligned}
\prod_{k=1}^{M}(1 - \alpha_k x) &= \exp\left\{\sum_{k=1}^{M}\log(1 - \alpha_k x)\right\} \\
&= \exp\left\{-\sum \alpha_k x - \sum \alpha_k^2\frac{x^2}{2} - \sum \alpha_k^3\frac{x^3}{3} - \ldots\right\} \\
&< \exp\left\{-\sum \alpha_k x - \sum \alpha_k^2\frac{x^2}{2}\right\} \qquad \text{(quadratic approximation)}, \\
&< \exp\left\{-\sum \alpha_k x\right\} \qquad \text{(linear approximation)}.
\end{aligned}$$

Attention is focused on absolute rather than relative errors, since this is most appropriate in the present context. For example, a probability with a true value of 10^{-80} and estimated value of 10^{-40} has a massive relative error associated with it, but for practical purposes both values are essentially zero and the absolute error is negligible.

1 Linear error function

Recall $\epsilon_1(x) = \exp(-S_1 x) - \prod_{k=1}^{M}(1 - \alpha_k x)$. It is required to find $\alpha_1, \ldots, \alpha_M$ to extremize $\epsilon_1(x)$ for fixed $x > 0$, subject to $\sum_{k=1}^{M} \alpha_k = S_1$. Since S_1 is fixed, the maximum of $\epsilon_1(x)$ occurs at the minimum of $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^{M}(1 - \alpha_k x)$ subject to $\sum_{k=1}^{M} \alpha_k = S_1$, and *vice versa.*

Definition: Let U_r be the unit hypercube in $\mathbb{R}^r$: that is,

$$U_r = \{\boldsymbol{\alpha} \in \mathbb{R}^r : 0 \leq \alpha_k \leq 1 \ \ \forall k = 1, \ldots, r\}.$$

Further let P_r be the perimeter of U_r:

$$P_r = \{\boldsymbol{\alpha} \in \mathbb{R}^r : \exists\, j \text{ such that } \alpha_k \in \{0,1\} \ \ \forall k \neq j, \text{ and } 0 \leq \alpha_j \leq 1\}\,.$$

Proposition 11 *(i) The maximum of* $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^{M}(1-\alpha_k x)$, *subject to* $\sum_{k=1}^{M}\alpha_k = S_1$, *occurs at* $\alpha_1 = \ldots = \alpha_M = S_1/M$.

(ii) The minimum of $f_M(\boldsymbol{\alpha})$, *subject to* $\sum_{k=1}^{M}\alpha_k = S_1$, *lies on the perimeter* P_M *of the M-dimensional hypercube* U_M*: that is, after permutation of components of* $\boldsymbol{\alpha}$ *if necessary,*

$$\boldsymbol{\alpha} = (\underbrace{1, \ldots, 1}_{\lfloor S_1 \rfloor}, S_1 - \lfloor S_1 \rfloor, 0, \ldots, 0)\,. \tag{4}$$

($\lfloor w \rfloor$ denotes the integer down from w: that is, the greatest integer r such that $w \geq r$.)

Proof by induction. First consider the case where $M = 2$. Let $\alpha = \alpha_1$ and $f_2(\alpha) = (1-\alpha x)(1-(S_1-\alpha)x)$. Then

$$f_2(\alpha) = (1 - S_1 x) + S_1 x^2 \alpha - x^2 \alpha^2\,.$$

(i)
$$\frac{df_2}{d\alpha} = S_1 x^2 - 2x^2\alpha \;\Rightarrow\; \alpha = \frac{S_1}{2} \text{ when } \frac{df_2}{d\alpha} = 0\,.$$

$$\frac{d^2 f_2}{d\alpha^2} = -2x^2 < 0 \;\Rightarrow\; \alpha = \frac{S_1}{2} \text{ gives a maximum of } f_2\,.$$

S_1 is the sum of two numbers in the interval $[0, 1]$, so $0 \leq S_1 \leq 2$. The solution $\alpha = \frac{S_1}{2}$ therefore lies within the allowed range for α, so (i) is true when $M = 2$.

(ii) The shape of $f_2(\alpha)$ is shown in Figure 1. From the figure it is clear that $f_2(\alpha)$ is unimodal and increases monotonically up to the mode, decreases monotonically away from the mode. The minimum of f_2 must therefore occur on the boundary of allowable α: that is, at $(\alpha_1, \alpha_2) = (S_1, 0)$ if $S_1 \leq 1$, or at $(\alpha_1, \alpha_2) = (1, S_1 - 1)$ if $S_1 > 1$ as in Figure 1. Therefore (ii) is true when $M = 2$.

Now suppose that (i) and (ii) are true for $f_r(\boldsymbol{\alpha})$ for all $r \leq M-1$. Consider $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^{M}(1-\alpha_k x)$.

(i) Suppose $(\alpha_1^*, \ldots, \alpha_M^*)$ maximizes $f_M(\boldsymbol{\alpha})$ subject to $\sum_{k=1}^{M}\alpha_k = S_1$. Then $(\alpha_1^*, \ldots, \alpha_{M-1}^*)$ must maximize $f_{M-1}(\boldsymbol{\alpha})$ subject to $\sum_{k=1}^{M-1}\alpha_k = S_1^*$, where $S_1^* = S_1 - \alpha_M^*$.

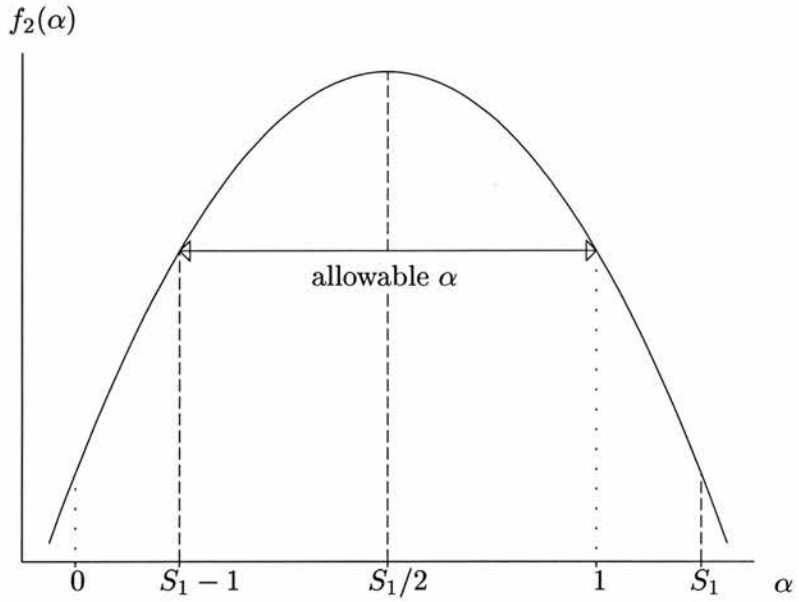


Figure 1: Shape of $f_2(\alpha) = (1 - S_1 x) + S_1 x^2 \alpha - x^2 \alpha^2$. The maximum occurs at $\alpha = S_1/2$. The allowable range for α is $[\max\{S_1 - 1, 0\}, \min\{1, S_1\}]$, because α satisfies the two expressions $0 \leq \alpha \leq 1$ and $0 \leq S_1 - \alpha \leq 1$. The minimum of f_2 occurs on the boundary of allowable α, which is either $\alpha = 1$, $\alpha_2 = S_1 - 1$ as shown here for $S_1 > 1$, or $\alpha = 0$, $\alpha_2 = S_1$ if $S_1 \leq 1$.

From the inductive hypothesis,

$$\alpha_1^* = \ldots = \alpha_{M-1}^* = \frac{S_1^*}{M-1}.$$

Similarly, $(\alpha_{M-1}^*, \alpha_M^*)$ must maximize $f_2(\boldsymbol{\alpha}) = \prod_{k=M-1}^{M}(1 - \alpha_k x)$ subject to $\alpha_{M-1} + \alpha_M = \alpha_{M-1}^* + \alpha_M^*$. By the inductive hypothesis, it follows that $\alpha_{M-1}^* = \alpha_M^*$, and therefore

$$\alpha_1^* = \ldots = \alpha_M^* = \frac{S_1}{M},$$

proving the inductive hypothesis.

(ii) Suppose $(\alpha_1^*, \ldots, \alpha_M^*)$ minimizes $f_M(\boldsymbol{\alpha})$ subject to $\sum_{k=1}^{M} \alpha_k = S_1$. Then $(\alpha_1^*, \ldots, \alpha_{M-1}^*)$ must minimize $f_{M-1}(\boldsymbol{\alpha})$ subject to $\sum_{k=1}^{M-1} \alpha_k = S_1^*$, where $S_1^* = S_1 - \alpha_M^*$.

By the inductive hypothesis, $(\alpha_1^*, \ldots, \alpha_{M-1}^*)$ is on the perimeter P_{M-1} of the $(M-1)$-dimensional hypercube U_{M-1}: so without loss of generality assume that $\alpha_1^*, \ldots, \alpha_{M-2}^* \in \{0, 1\}$ and $0 \leq \alpha_{M-1}^* \leq 1$.

Suppose α_{M-1}^* is 0 or 1: then $(\alpha_1^*, \ldots, \alpha_M^*)$ is on the perimeter P_M of the M-dimensional hypercube U_M, and the inductive hypothesis holds.

Now suppose that $0 < \alpha^*_{M-1} < 1$. In this case, $\alpha_{M-1} = \alpha^*_{M-1}$, $\alpha_M = \alpha^*_M$ must minimize $(1-\alpha_{M-1}x)(1-\alpha_M x)$ subject to $\alpha_{M-1}+\alpha_M = \alpha^*_{M-1}+\alpha^*_M$, and therefore by the inductive hypothesis lies on P_2. Since $0 < \alpha^*_{M-1} < 1$, it follows that $\alpha^*_M = 0$ or $\alpha^*_M = 1$. Thus $(\alpha^*_1, \ldots, \alpha^*_M) \in P_M$ and the inductive hypothesis holds. □

Proposition 11 provides the required results about the linear error function $\epsilon_1(x)$. The maximum linear error $\epsilon_1(x)$ for fixed x, subject to $\sum_{k=1}^M \alpha_k = S_1$, occurs at the minimum of $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^M (1-\alpha_k x)$: that is, when $\boldsymbol{\alpha} = (\underbrace{1, \ldots, 1}_{\lfloor S_1 \rfloor}, S_1 - \lfloor S_1 \rfloor, 0, \ldots, 0)$.
The maximum linear error is therefore given by

$$\max_{\boldsymbol{\alpha}} (\epsilon_1(x)) = \exp\{-S_1 x\} - \Big(1-x\Big)^{\lfloor S_1 \rfloor} \Big(1 - \big(S_1 - \lfloor S_1 \rfloor\big)x\Big) \,. \tag{5}$$

The minimum linear error $\epsilon_1(x)$ for fixed x, subject to $\sum_{k=1}^M \alpha_k = S_1$, occurs at the maximum of $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^M (1-\alpha_k x)$: that is, when $\alpha_1 = \ldots = \alpha_M = S_1/M$. The minimum linear error is therefore given by

$$\min_{\boldsymbol{\alpha}} (\epsilon_1(x)) = \exp\{-S_1 x\} - \Big(1 - \frac{S_1}{M}\, x\Big)^M \,. \tag{6}$$

2 Quadratic error function

The quadratic error function is $\epsilon_2(x) = \exp(-S_1 x - S_2\, x^2/2) - \prod_{k=1}^M (1-\alpha_k x)$. It is required to find $\alpha_1, \ldots, \alpha_M$ to extremize $\epsilon_2(x)$ for fixed $x > 0$, subject to the two conditions $\sum_{k=1}^M \alpha_k = S_1$ and $\sum_{k=1}^M \alpha_k^2 = S_2$. Since S_1 and S_2 are fixed, the maximum of $\epsilon_2(x)$ occurs at the minimum of $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^M (1-\alpha_k x)$ subject to $\sum_{k=1}^M \alpha_k = S_1$ and $\sum_{k=1}^M \alpha_k^2 = S_2$, and *vice versa.*

It is useful to visualize the range of possible values for $(\alpha_1, \ldots, \alpha_M)$ under the constraints $\sum_{k=1}^M \alpha_k = S_1$ and $\sum_{k=1}^M \alpha_k^2 = S_2$. Let the vector $\mathbf{1} \in \mathbb{R}^M$ be defined as

$$\mathbf{1} = \underbrace{(1, 1, \ldots, 1)}_{M}{}^T \,.$$

The first constraint indicates that the scalar product $\boldsymbol{\alpha} \cdot \mathbf{1}$ is constant: i.e. that all allowable vectors $\boldsymbol{\alpha}$ lie on the plane with normal $\mathbf{1}$ that lies at distance $S_1/\sqrt{M}$ from the origin. The second constraint stipulates that $\|\boldsymbol{\alpha}\|$ is constant: in fact $\|\boldsymbol{\alpha}\|^2 = S_2$, so $\boldsymbol{\alpha}$ must lie on

the M-dimensional sphere with radius $\sqrt{S_2}$. The intersection of the sphere with the plane gives a circle of allowable $\boldsymbol{\alpha}$, shown in Figure 2 for the 3-dimensional case. The cube in Figure 2 represents the further constraint that $0 \leq \alpha_k \leq 1$ for all $k = 1, \ldots, M$.

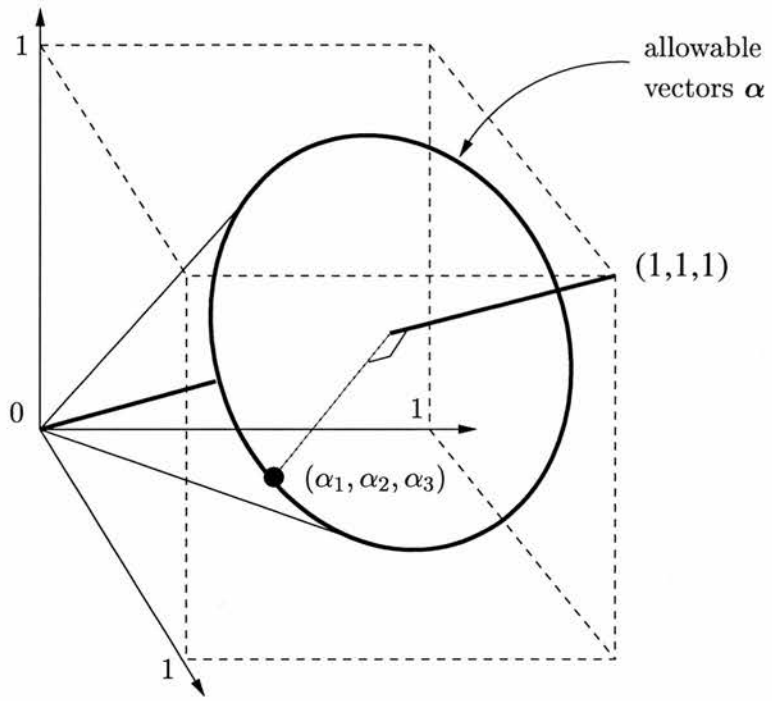


Figure 2: Allowable vectors $\boldsymbol{\alpha}$ in the 3-dimensional case subject to the constraints $\sum_{k=1}^{3} \alpha_k = S_1$ and $\sum_{k=1}^{3} \alpha_k^2 = S_2$. The allowable $\boldsymbol{\alpha}$ lie on the circular rim of the cone shown, perpendicular to the vector $(1,1,1)^T$. The further constraint that $0 \leq \alpha_k \leq 1\ \forall k$ is satisfied by those $\boldsymbol{\alpha}$ that lie within the dashed cube U_3.

The quantities S_1 and S_2 are assumed such that there exists at least one solution to the constraints $\sum_{k=1}^{M} \alpha_k = S_1$, $\sum_{k=1}^{M} \alpha_k^2 = S_2$ and $0 \leq \alpha_k \leq 1\ \forall k$. For example, the original $\boldsymbol{\alpha}$ that gives rise to S_1 and S_2 is a solution. This implies a relationship between S_1 and S_2 summarized in the following lemma.

Lemma 1 *If solutions to (a)* $\sum_{k=1}^{M} \alpha_k = S_1$, *(b)* $\sum_{k=1}^{M} \alpha_k^2 = S_2$ *and (c)* $0 \leq \alpha_k \leq 1\ \forall k$ *exist, then (i)* $S_2 \leq S_1^2$ *and (ii)* $S_1^2 \leq M S_2$.

Proof (i) Suppose $\sqrt{S_2} > S_1$. The extremal points of the plane $\sum_{k=1}^{M} \alpha_k = S_1$ in the positive quadrant of $\mathbb{R}^M$ occur at the permutations of $(S_1, 0, \ldots, 0)$, which have length $S_1 < \sqrt{S_2}$. The plane therefore lies inside the sphere $\|\boldsymbol{\alpha}\| = \sqrt{S_2}$ in the positive quadrant (Figure 3), and there are no solutions to (a), (b) and (c). Hence $S_2 \leq S_1^2$.

(ii) Suppose that $\sqrt{S_2} < S_1/\sqrt{M}$. The shortest distance between the plane $\sum_{k=1}^{M} \alpha_k = S_1$ and the origin is given by the scalar product $d = \boldsymbol{\alpha} \cdot \boldsymbol{n}$, where $\boldsymbol{n} = \frac{1}{\sqrt{M}}\mathbf{1}$ is the unit normal

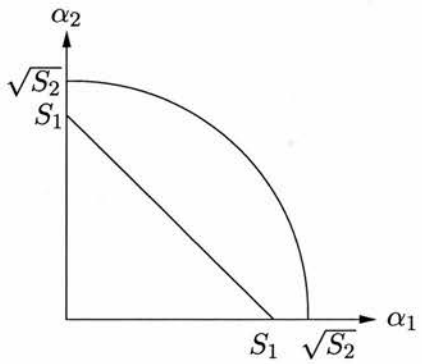


Figure 3: Cross-section of the surfaces $\sum_{k=1}^{M} \alpha_k = S_1$ and $\sum_{k=1}^{M} \alpha_k^2 = S_2$ in the (α_1, α_2)-plane when $\sqrt{S_2} > S_1$. The plane and sphere do not intersect in the positive quadrant.

vector to the plane. This distance is $d = S_1/\sqrt{M}$; so if $\sqrt{S_2} < S_1/\sqrt{M}$ then the sphere $\|\boldsymbol{\alpha}\| = \sqrt{S_2}$ lies underneath the plane in the positive quadrant (Figure 4), so there are no solutions to (a), (b) and (c). Hence $\sqrt{S_2} \geq S_1/\sqrt{M}$ and therefore $S_1^2 \leq MS_2$. □

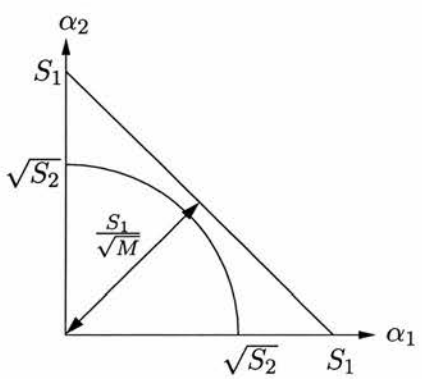


Figure 4: Cross-section of the surfaces $\sum_{k=1}^{M} \alpha_k = S_1$ and $\sum_{k=1}^{M} \alpha_k^2 = S_2$ in the (α_1, α_2)-plane when $\sqrt{S_2} < S_1/\sqrt{M}$. The plane and sphere do not intersect in the positive quadrant.

The main results of this section will be proved by induction, so it is useful at this stage to examine the behaviour of $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^{M}(1 - \alpha_k x)$ subject to $\sum_{k=1}^{M} \alpha_k = S_1$ and $\sum_{k=1}^{M} \alpha_k^2 = S_2$ for an initial case of M. With two constraints S_1 and S_2, the initial case of $M = 2$ does not provide sufficient information for the inductive proofs. Instead it is necessary to study f_M for $M = 3$.

Lemma 2 *Let $f_3(\boldsymbol{\alpha}) = (1 - \alpha_1 x)(1 - \alpha_2 x)(1 - \alpha_3 x)$, and suppose there exists at least one solution to the following constraints: (a) $\sum_{k=1}^{3} \alpha_k = S_1$, (b) $\sum_{k=1}^{3} \alpha_k^2 = S_2$ and (c) $0 \leq \alpha_k \leq 1$ $\forall k = 1, 2, 3$. Then*

(i) the global maximum of f_3 subject to (a) and (b) occurs at $\boldsymbol{\alpha}^* = (a_3, b_3, b_3)$, where

$$a_3 = \frac{S_1}{3} - \frac{2}{3}\sqrt{\frac{3S_2 - S_1^2}{2}}, \qquad b_3 = \frac{S_1}{3} + \frac{1}{3}\sqrt{\frac{3S_2 - S_1^2}{2}}.$$

The global minimum of f_3 subject to (a) and (b) occurs at $\boldsymbol{\alpha}^{**} = (c_3, c_3, d_3)$, where

$$c_3 = \frac{S_1}{3} - \frac{1}{3}\sqrt{\frac{3S_2 - S_1^2}{2}}, \qquad d_3 = \frac{S_1}{3} + \frac{2}{3}\sqrt{\frac{3S_2 - S_1^2}{2}}.$$

(ii) if $\boldsymbol{\alpha}^* \notin U_3$ then the maximum of f_3 within U_3 occurs at $(a_2, b_2, 0)$, where

$$a_2 = \frac{S_1}{2} - \frac{1}{2}\sqrt{2S_2 - S_1^2}, \qquad b_2 = \frac{S_1}{2} + \frac{1}{2}\sqrt{2S_2 - S_1^2},$$

and $(a_2, b_2, 0)$ lies in U_3.

(iii) if $\boldsymbol{\alpha}^{**} \notin U_3$ then the minimum of f_3 within U_3 occurs at $(c_2, d_2, 1)$, where

$$c_2 = \frac{S_1 - 1}{2} - \frac{1}{2}\sqrt{2(S_2 - 1) - (S_1 - 1)^2}, \qquad d_2 = \frac{S_1 - 1}{2} + \frac{1}{2}\sqrt{2(S_2 - 1) - (S_1 - 1)^2},$$

and $(c_2, d_2, 1)$ lies in U_3.

Proof Let $\alpha = \alpha_1$ and let α_2, α_3 satisfy $\alpha_2 + \alpha_3 = S_1 - \alpha$, $\alpha_2^2 + \alpha_3^2 = S_2 - \alpha^2$. Solving these expressions for α_2, α_3 yields

$$\alpha_2 = \frac{S_1 - \alpha + \sqrt{2(S_2 - \alpha^2) - (S_1 - \alpha)^2}}{2}, \qquad \alpha_3 = \frac{S_1 - \alpha - \sqrt{2(S_2 - \alpha^2) - (S_1 - \alpha)^2}}{2}. \tag{7}$$

Let $f(\alpha) = (1 - \alpha x)(1 - \alpha_2 x)(1 - \alpha_3 x)$ with α, α_2 and α_3 as above.

Routine algebra gives

$$f(\alpha) = -x^3\alpha^3 + x^3 S_1 \alpha^2 - \frac{1}{2}x^3(S_1^2 - S_2)\alpha + 1 - S_1 x + \frac{1}{2}x^2(S_1^2 - S_2). \tag{8}$$

(i)
$$\frac{df}{d\alpha} = -3x^3\alpha^2 + 2x^3 S_1 \alpha - \frac{1}{2}x^3(S_1^2 - S_2),$$

so for any fixed $x \neq 0$,

$$\frac{df}{d\alpha} = 0 \;\Rightarrow\; \alpha = \frac{S_1}{3} \pm \frac{1}{3}\sqrt{\frac{3S_2 - S_1^2}{2}}.$$

Further,

$$\frac{d^2f}{d\alpha^2} = -6x^3\alpha + 2x^3S_1\,,$$

so when $\alpha = \frac{S_1}{3} + \frac{1}{3}\sqrt{\frac{3S_2-S_1^2}{2}}$, $\frac{d^2f}{d\alpha^2} = -2x^3\sqrt{\frac{3S_2-S_1^2}{2}} < 0$, and when $\alpha = \frac{S_1}{3} - \frac{1}{3}\sqrt{\frac{3S_2-S_1^2}{2}}$, $\frac{d^2f}{d\alpha^2} = 2x^3\sqrt{\frac{3S_2-S_1^2}{2}} > 0$.

The global maximum of f therefore occurs at

$$\alpha \;=\; \alpha_1 \;=\; \frac{S_1}{3} + \frac{1}{3}\sqrt{\frac{3S_2 - S_1^2}{2}} \;=\; b_3\,,$$

and by substitution into (7),

$$\alpha_2 = \frac{S_1}{3} + \frac{1}{3}\sqrt{\frac{3S_2 - S_1^2}{2}} \;=\; b_3\,, \qquad \alpha_3 = \frac{S_1}{3} - \frac{2}{3}\sqrt{\frac{3S_2 - S_1^2}{2}} \;=\; a_3\,,$$

so that the global maximum of f_3 subject to (a) and (b) occurs at $\boldsymbol{\alpha}^* = (a_3, b_3, b_3)$ as stated in (i).

Similarly, the global minimum of f occurs at

$$\alpha \;=\; \alpha_1 \;=\; \frac{S_1}{3} - \frac{1}{3}\sqrt{\frac{3S_2 - S_1^2}{2}} \;=\; c_3\,,$$

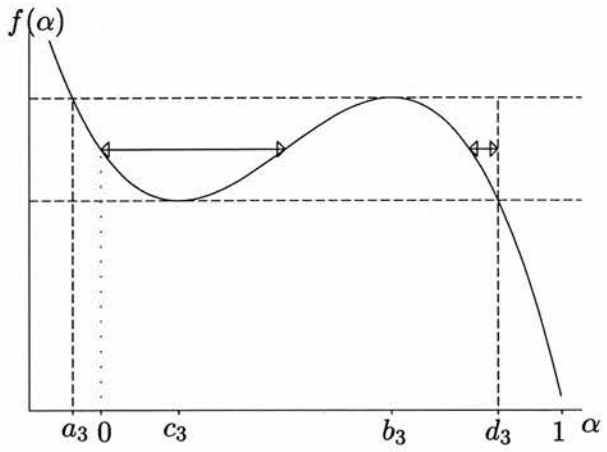
and substituting once again into (7) yields

$$\alpha_2 = \frac{S_1}{3} + \frac{2}{3}\sqrt{\frac{3S_2 - S_1^2}{2}} \;=\; d_3\,, \qquad \alpha_3 = \frac{S_1}{3} - \frac{1}{3}\sqrt{\frac{3S_2 - S_1^2}{2}} \;=\; c_3\,,$$

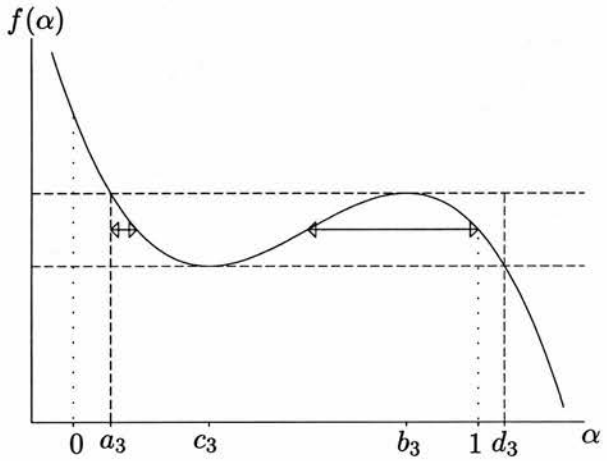
so that the global minimum of f_3 subject to (a) and (b) occurs at $\boldsymbol{\alpha}^{**} = (c_3, c_3, d_3)$ as stated in (i).

(ii) Suppose the vector $\boldsymbol{\alpha}^*$ does not lie inside the unit hypercube U_3. The shape of the curve $f(\alpha)$ is shown in Figure 5 under various scenarios. Now $f(\alpha_1) = (1-\alpha_1 x)(1-\alpha_2 x)(1-\alpha_3 x)$ where α_2 and α_3 solve $\alpha_2 + \alpha_3 = S_1 - \alpha_1$, $\alpha_2^2 + \alpha_3^2 = S_2 - \alpha_1^2$. It is clear that $f(\alpha_1) = f(\alpha_2) = f(\alpha_3)$, and thus if $(\alpha_1, \alpha_2, \alpha_3)$ lies in U_3 then a horizontal line that cuts the curve f at the point α_1 must also cut the curve f at the points α_2 and α_3. For α_1 to be an admissible solution to the constraints (a), (b) and (c), therefore, the curve f must be cut three times between 0 and 1 by a horizontal line passing through $f(\alpha_1)$.

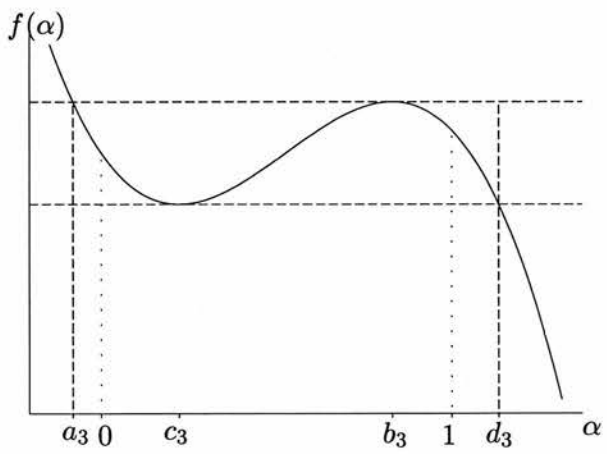
By examination of the curves in Figure 5, it is clear that a number of relations must hold



(a) $a_3 < 0$: solutions marked by arrows.



(b) $d_3 > 1$: solutions marked by arrows.



(c) $f(0) < f(1)$: no solutions.

Figure 5: Possible shapes for the curve $f(\alpha)$. Admissible values of α are those at which a horizontal line may be drawn that cuts the curve 3 times between 0 and 1. The range of admissible values is indicated by horizontal arrows on the diagrams.

if there exist any solutions to the constraints (a), (b) and (c), notably:

$$c_3 \geq 0, \qquad b_3 \leq 1, \qquad f(0) \geq f(1)\,. \tag{9}$$

It is also clear that the maximum of f must occur at $\alpha = 0$ if $a_3 < 0$; and if $a_3 < 0$ and no solutions exist inside U_3 when $\alpha = 0$ then there can be no solutions in U_3 altogether.

More formally, when $\alpha = 0$, (7) gives

$$\alpha_2 = b_2 = \frac{S_1}{2} + \frac{1}{2}\sqrt{2S_2 - S_1^2}\,, \quad \alpha_3 = a_2 = \frac{S_1}{2} - \frac{1}{2}\sqrt{2S_2 - S_1^2}\,.$$

It is evident that $b_2 \geq 0$, and it follows from relation (i) of Lemma 1 that $a_2 \geq 0$. The condition $b_2 \leq 1$ may be derived from the relation $f(0) \geq f(1)$ of (9), from which also $a_2 \leq 1$ since $a_2 \leq b_2$.

The point $(a_2, b_2, 0)$ therefore lies inside U_3, and the statement (ii) of Lemma 2 is proved.

(iii) The proof follows from the same arguments as (ii). If $d_3 > 1$ then the minimum of f within U_3 must occur at $\alpha = 1$, in which case from (7) the minimum vector is $(c_2, d_2, 1)$ as stated in (iii). Once again, this vector must lie in U_3 if $d_3 > 1$, since if there are no solutions at $\alpha = 1$ there can be no solutions at all. □

Motivated by Lemma 2, three further lemmas are proposed before the main results of the section are derived.

Lemma 3 *For any integer $M \geq 2$, let the constraints (a), (b) and (c) be defined as (a) $\sum_{k=1}^{M} \alpha_k = S_1$, (b) $\sum_{k=1}^{M} \alpha_k^2 = S_2$, (c) $0 \leq \alpha_k \leq 1 \;\; \forall\, k = 1, \ldots, M$. Suppose there exists at least one solution to (a), (b) and (c). Then*
(i) a solution to (a) and (b) of the form $\boldsymbol{\alpha} = (a, \underbrace{b, \ldots, b}_{M-1})$ with $a \leq b$ exists, and is unique up to permutation of components. The solution is given by

$$a = \frac{S_1}{M} - \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,, \qquad b = \frac{S_1}{M} + \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,;$$

(ii) a solution to the constraints (a) and (b) of the form $\boldsymbol{\alpha} = (\underbrace{c, \ldots, c}_{M-1}, d)$ with $c \leq d$ exists, and is unique up to permutation of components, given by

$$c = \frac{S_1}{M} - \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,, \qquad d = \frac{S_1}{M} + \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,.$$

Proof (i) The constraints (a) and (b) become

$$a + (M-1)b = S_1\,, \qquad a^2 + (M-1)b^2 = S_2\,,$$

whence $a = S_1 - (M-1)b$, and substituting this in the second expression gives the quadratic equation $(S_1 - (M-1)b)^2 + (M-1)b^2 = S_2$. The solution of the quadratic equation is

$$b = \frac{S_1}{M} \pm \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,.$$

Substituting $a = S_1 - (M-1)b$ yields

$$a = \frac{S_1}{M} \mp \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,.$$

There is therefore a unique solution satisfying $a \leq b$, given by

$$a = \frac{S_1}{M} - \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,, \qquad b = \frac{S_1}{M} + \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}$$

as required, since then $a \leq S_1/M \leq b$. The solutions for a and b are real, by Lemma 1.

(ii) c and d are given by the two remaining roots from part (i): that is,

$$c = \frac{S_1}{M} - \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,, \qquad d = \frac{S_1}{M} + \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,.$$

This gives $c \leq S_1/M \leq d$ as required. □

Lemma 4 *For any integer $M \geq 2$, let the constraints (a), (b) and (c) be defined as in Lemma 3. Suppose there exists at least one solution to (a), (b) and (c), and let $b(M, S_1, S_2) = \frac{S_1}{M} + \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}$ and $c(M, S_1, S_2) = \frac{S_1}{M} - \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}$. Then*

(i) $b(M, S_1, S_2) \leq 1$,
(ii) $c(M, S_1, S_2) \geq 0$.

Proof (i) Let $a(M, S_1, S_2) = \frac{S_1}{M} - \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}$. By Lemma 3, the vectors of the form $(a(M, S_1, S_2), b(M, S_1, S_2), \ldots, b(M, S_1, S_2))$ give the unique solutions to (a) and (b) of the form $(a, b, \ldots, b)$ with $a \leq b$. These solutions lie on the outside edges of the M-dimensional hypercube with side $b(M, S_1, S_2)$, illustrated in Figure 6 for the case $M = 3$.

Let $a = a(M, S_1, S_2)$ and $b = b(M, S_1, S_2)$. Suppose that there exists a solution $\boldsymbol{\alpha}$ to (a)

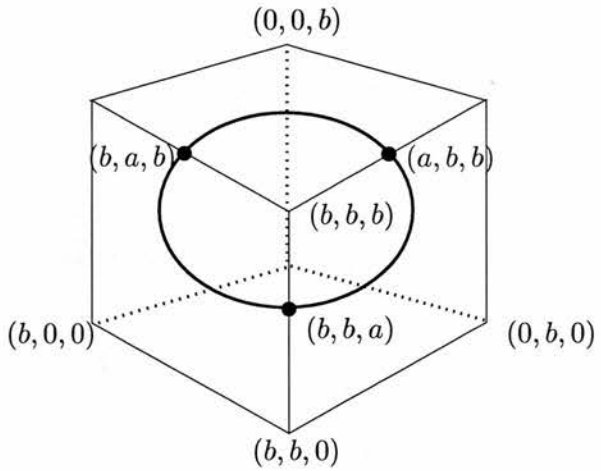


Figure 6: Illustration that no solutions lie strictly inside the hypercube of side $b(M, S_1, S_2)$ when $M = 3$. The bold circle is the surface of solutions to the constraints (a) and (b), and lies on the outside edge of the cube. The circle does not intersect the interior of the cube at any point.

and (b) strictly inside the hypercube with side b: that is, $\boldsymbol{\alpha}$ must satisfy $0 \leq \alpha_k < b \ \forall k = 1, \ldots, M$. Without loss of generality, $\alpha_1 = a + \delta$ for some $\delta > 0$, and $\alpha_2 = (b - \epsilon_1)$, ..., $\alpha_M = (b - \epsilon_{M-1})$ for some $\epsilon_1, \ldots, \epsilon_{M-1} > 0$.

By constraint (a),

$$(a + \delta) + (b - \epsilon_1) + \ldots + (b - \epsilon_{M-1}) = a + (M - 1)b \quad \Rightarrow \quad \epsilon_1 + \ldots + \epsilon_{M-1} = \delta \,,$$

and by constraint (b), substituting also the expression above,

$$\begin{aligned} (a + \delta)^2 + (b - \epsilon_1)^2 + \ldots + (b - \epsilon_{M-1})^2 &= a^2 + (M - 1)b^2 \\ \Rightarrow \quad 2a\delta + \delta^2 - 2b\delta + \epsilon_1^2 + \ldots + \epsilon_{M-1}^2 &= 0 \\ \Rightarrow \quad 2\delta^2 + 2\delta(a - b) &= 2(\epsilon_1\epsilon_2 + \epsilon_1\epsilon_3 + \ldots + \epsilon_{M-2}\epsilon_{M-1}) > 0 \\ \Rightarrow \quad \delta(\delta + a - b) &> 0 \,. \end{aligned}$$

Since $\delta > 0$, it follows that $\delta + a - b > 0$ also. But $\delta + a = \alpha_1 \leq b$ by hypothesis. Thus there are no solutions to the constraints (a) and (b) that lie strictly inside the hypercube of side b in $\mathbb{R}^M$. Hence $b(M, S_1, S_2) \leq 1$, otherwise there are no solutions within the hypercube of side 1, which is U_M.

(ii) Now consider $c(M, S_1, S_2) = \frac{S_1}{M} - \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}$. By Lemma 1, if there exists at least

one solution to the constraints (a), (b) and (c) then $S_1^2 \geq S_2$. Thus

$$MS_1^2 \geq MS_2 \quad \Rightarrow \quad (M-1)S_1^2 \geq MS_2 - S_1^2 \quad \Rightarrow \quad S_1 \geq \sqrt{\frac{MS_2 - S_1^2}{M-1}}\,,$$

whence $c(M, S_1, S_2) \geq 0$. □

Lemma 5 *Let* $a(M, S_1, S_2) = \frac{S_1}{M} - \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}$ *and* $d(M, S_1, S_2) = \frac{S_1}{M} + \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}$, *and suppose that* $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ *is a solution to the constraints (a), (b) and (c) of Lemma 3. Then each* α_k *satisfies*

$$a(M, S_1, S_2) \leq \alpha_k \leq d(M, S_1, S_2)\,.$$

Proof Suppose α_1 is considered fixed. The remaining components $\alpha_2, \ldots, \alpha_M$ satisfy $\sum_{k=2}^{M} \alpha_k = (S_1 - \alpha_1)$ and $\sum_{k=2}^{M} \alpha_k^2 = (S_2 - \alpha_1^2)$. Applying the result of Lemma 1 in $M-1$ dimensions, this gives

$$(S_1 - \alpha_1)^2 \leq (M-1)(S_2 - \alpha_1^2)\,,$$

whence, by expanding and rearranging,

$$M\alpha_1^2 - 2S_1\alpha_1 + S_1^2 - (M-1)S_2 \;\leq\; 0\,.$$

Equality holds in the above quadratic if

$$\alpha_1 = \frac{S_1}{M} \pm \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\;;$$

that is, when $\alpha_1 = a(M, S_1, S_2)$ or $d(M, S_1, S_2)$. Therefore, for the inequality to hold, $a(M, S_1, S_2) \leq \alpha_1 \leq d(M, S_1, S_2)$, and the same is true of $\alpha_2, \ldots, \alpha_M$ by symmetry. □

The main results for the quadratic exponential approximation are now presented.

Proposition 12 *Let* $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^{M}(1 - \alpha_k x)$ *and let the constraints (a), (b) and (c) in* $\mathbb{R}^M$ *be defined as (a)* $\sum_{k=1}^{M} \alpha_k = S_1$, *(b)* $\sum_{k=1}^{M} \alpha_k^2 = S_2$, *(c)* $0 \leq \alpha_k \leq 1 \;\; \forall\, k = 1, \ldots, M$ *(that is,* $\boldsymbol{\alpha} \in U_M$*).*

Suppose there exists at least one solution to (a), (b) and (c). Then
(i) The global maximum of f_M *for fixed* x, *subject to the constraints (a) and (b), occurs*

at $\boldsymbol{\alpha}^* = (a^*, \underbrace{b^*, \ldots, b^*}_{M-1})$ *where* $a^* \leq b^*$, *and*

$$a^* = a(M, S_1, S_2) = \frac{S_1}{M} - \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}, \quad b^* = b(M, S_1, S_2) = \frac{S_1}{M} + \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}.$$

(ii) If $(a^*, b^*, \ldots, b^*)$ *does not satisfy the constraint (c), then the maximum of* f_M *subject to (a), (b) and (c) occurs at* $\boldsymbol{\alpha}'_{\boldsymbol{n}} = (a', \underbrace{b', \ldots, b'}_{n-1}, 0, \ldots, 0)$ *for some* n, *where* $a' = a(n, S_1, S_2)$ *and* $b' = b(n, S_1, S_2)$.

(iii) There exists some $n \leq M$ *for which* $\boldsymbol{\alpha}'_{\boldsymbol{n}}$ *lies inside* U_M, *and* $n = \left\lceil \frac{S_1^2}{S_2} \right\rceil$. *This is the unique value of* n *for which* $\boldsymbol{\alpha}'_{\boldsymbol{n}} \in U_M$ *and* $a(n, S_1, S_2) > 0$.

($\lceil w \rceil$ *denotes the integer up from* w: *that is, the least integer* r *such that* $w \leq r$.*)*

Proof By induction.

(i) The result is true when $M = 3$ by Lemma 2. Suppose the result holds for f_{M-1} with $M - 1 \geq 3$. Now consider the M-dimensional case. Let $(\alpha_1^*, \ldots, \alpha_M^*)$ be the point at which the global maximum of f_M occurs, subject to the constraints (a) and (b). Further let $S_1^* = S_1 - \alpha_M^*$, $S_2^* = S_2 - (\alpha_M^*)^2$.

The vector $(\alpha_1^*, \ldots, \alpha_{M-1}^*)$ must be the point at which the global maximum of f_{M-1} occurs in $M - 1$ dimensions, subject to the revised constraints $\sum_{k=1}^{M-1} \alpha_k = S_1^*$, $\sum_{k=1}^{M-1} \alpha_k^2 = S_2^*$; since if there were a vector $\boldsymbol{\beta}^*_{\boldsymbol{M-1}}$ such that $f_{M-1}(\boldsymbol{\beta}^*_{\boldsymbol{M-1}}) > f_{M-1}((\alpha_1^*, \ldots, \alpha_{M-1}^*))$ then the vector $(\boldsymbol{\beta}^*_{\boldsymbol{M-1}}, \alpha_M^*)$ would satisfy the constraints (a) and (b) and $f_M(\boldsymbol{\beta}^*_{\boldsymbol{M-1}}, \alpha_M^*) > f_M((\alpha_1^*, \ldots, \alpha_M^*))$. By the inductive hypothesis, therefore, $(\alpha_1^*, \ldots, \alpha_{M-1}^*) = (a^{**}, b^{**}, \ldots, b^{**})$ where $a^{**} \leq b^{**}$, $a^{**} = a(M-1, S_1^*, S_2^*)$, $b^{**} = b(M-1, S_1^*, S_2^*)$.

Consider the two cases (1) $a^{**} = b^{**}$, (2) $a^{**} < b^{**}$.

(1) $(\alpha_1^*, \ldots, \alpha_M^*) = (b^{**}, \ldots, b^{**}, \alpha_M^*)$. Omit the first component $\alpha_1^* = b^{**}$. This leaves the vector $(\underbrace{b^{**}, \ldots, b^{**}}_{M-2 \geq 2}, \alpha_M^*)$, which again must be a global maximum subject to revised constraints in $M - 1$ dimensions. By the inductive hypothesis, the singleton α_M^* in this vector is the smaller element: so $\alpha_M^* \leq b^{**}$, and following permutation of components $(\alpha_1^*, \ldots, \alpha_M^*)$ is of the required form $(a, b, \ldots, b)$ with $a \leq b$. By Lemma 3, a and b have the forms given in the proposition.

(2) When $a^{**} < b^{**}$, then $(\alpha_1^*, \ldots, \alpha_M^*) = (a^{**}, b^{**}, \ldots, b^{**}, \alpha_M^*)$. Omit the component

$\alpha^*_{M\ 1} = b^{**}$. This leaves the vector $(a^{**}, \underbrace{b^{**}, \ldots, b^{**}}_{M-3\geq 1}, \alpha^*_M)$, which must be a global maximum subject to revised constraints in $M-1$ dimensions. By the inductive hypothesis, this vector is of the form $(a, \underbrace{b, \ldots, b}_{M-2})$ with $a \leq b$, and it follows that $\alpha^*_M = b^{**}$ and the original vector $(\alpha^*_1, \ldots, \alpha^*_M) = (a^{**}, b^{**}, \ldots, b^{**})$ is of the required form. By Lemma 3, a and b have the forms given in the proposition.

(ii) The result is true for $M = 3$ by Lemma 2. Suppose the result holds for f_{M-1} with $M - 1 \geq 3$ and consider the M-dimensional case. Let $(\alpha^*_1, \ldots, \alpha^*_M)$ be the maximum of f_M subject to the three constraints (a), (b) and (c).

If any $\alpha^*_k = 0$ the problem reduces immediately to $M-1$ dimensions and the result holds by the inductive hypothesis. Suppose instead that $\alpha^*_k > 0\ \forall k$, and suppose that the global maximum does not lie in U_M, otherwise the result is trivially true with $n = M$. Omit the component α^*_M: by the inductive hypothesis, and because all $\alpha^*_k > 0$, the remaining $(M-1)$-dimensional vector $(\alpha^*_1, \ldots, \alpha^*_{M-1})$ is a maximum in U_M of the form $(a^*, \underbrace{b^*, \ldots, b^*}_{M-2\geq 2})$ where $0 < a^* \leq b^*$. Now omit $\alpha^*_{M-1} = b^*$. Using the inductive hypothesis again, $(\alpha^*_1, \ldots, \alpha^*_{M-2}, \alpha^*_M) = (a^{**}, b^{**}, \ldots, b^{**})$ is a maximum in U_M where $0 < a^{**} \leq b^{**}$.

It follows that $a^{**} = a^*$, $b^{**} = b^*$, whence the vector $(\alpha^*_1, \ldots, \alpha^*_M)$ is given by $(a^*, b^*, \ldots, b^*)$ where $a^* \leq b^*$. As the unique solution of this form, $(\alpha^*_1, \ldots, \alpha^*_M)$ must be the global maximum, contradicting the proposition that the global maximum does not lie in U_M. Thus either some $\alpha^*_k = 0$, or the global maximum lies in U_M.

(iii) The result holds trivially for $M = 2$: if there exists a solution to the three constraints (a), (b) and (c) in $\mathbb{R}^2$, there must exist one of the form (a, b) with $a \leq b$. Further, from Lemma 1 it is known that $1 \leq S_1^2/S_2 \leq 2$, so that $\lceil S_1^2/S_2 \rceil = 1$ or 2. From Figure 3 it is clear that $S_1^2/S_2 = 1$ if and only if the only solutions to (a), (b) and (c) are $(S_1, 0)$ and $(0, S_1)$, in which case the result holds with $n = 1$. If $S_1^2/S_2 > 1$ the result holds with $n = 2$.

Suppose that for all integers $r \leq M-1$, if there are solutions to the constraints (a), (b) and (c) in $\mathbb{R}^r$, then there exist a and b such that $a \leq b$ and $(a, \underbrace{b, \ldots, b}_{n-1}, 0, \ldots, 0) \in U_r$, where $n = \lceil S_1^2/S_2 \rceil$. By Lemma 3, the values of a and b are given by $a = a(r, S_1, S_2)$, $b = b(r, S_1, S_2)$.

Now consider $r = M$. There are two cases:

(1) $\lceil S_1^2/S_2 \rceil \leq M-1$,

(2) $M-1 < n = \lceil S_1^2/S_2 \rceil \leq M$.

(1) Let $\alpha_M^* = 0$. This gives the $(M-1)$-dimensional case where the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{M-1})$ satisfies $\sum_{k=1}^{M-1} \alpha_k = S_1$ and $\sum_{k=1}^{M-1} \alpha_k^2 = S_2$. In order to apply the inductive hypothesis, it is also necessary to show that there exist solutions to these constraints within the $(M-1)$-dimensional hypercube U_{M-1}.

For this, note that

$$\frac{S_1^2}{S_2} \leq M-1 \quad \Rightarrow \quad MS_1^2 \leq M(M-1)S_2 \quad \Rightarrow \quad S_1^2 \leq (M-1)(MS_2 - S_1^2)\,,$$

whence $a(M, S_1, S_2) \leq 0$. If $a(M, S_1, S_2) = 0$, the vector $(0,\, b(M, S_1, S_2), \ldots, b(M, S_1, S_2))$ is itself a solution to the constraints (a) and (b) within U_{M-1}, as $b(M, S_1, S_2) \leq 1$ by Lemma 4. Suppose instead that $a(M, S_1, S_2) < 0$. The point $(a(M, S_1, S_2), b(M, S_1, S_2), \ldots, b(M, S_1, S_2))$ is a solution to the constraints (a) and (b) that lies outside of the unit hypercube U_M. However, by hypothesis there exists at least one solution to (a) and (b) inside U_M, say $\boldsymbol{\gamma}$. Furthermore, for $M > 2$ the solutions to (a) and (b) occur at the intersection of the hyperplane defined by (a) with the sphere defined by (b), and therefore lie on a continuous surface that bounds a convex region. When $M = 3$, for example, the surface of solutions is the circle shown in Figure 2, while for $M = 4$ it is a sphere, and so on. Let $\mathcal{C}$ be the surface of solutions to (a) and (b).

Since $\mathcal{C}$ contains points that lie both inside and outside the hypercube U_M, and is continuous, it must pass through a face of U_M. Suppose that $\mathcal{C}$ passed only through faces given by $\alpha_k = 1$ for some k. These will be referred to as 1-faces. Suppose further that $\mathcal{C}$ intersects the face with $\alpha_M = 1$ at $(\beta_1, \beta_2, \ldots, 1)$. If $\beta_1 = \ldots = \beta_{M-1}$ then by Lemma 3 this point is $(c, \ldots, c, d)$ where $c = c(M, S_1, S_2)$ and $d = 1$. By Lemma 5, d is the maximum value that any of the components of the points on $\mathcal{C}$ can take, so the surface $\mathcal{C}$ must touch the 1-face at $(c, \ldots, c, 1)$ and $\mathcal{C}$ contains no points beyond this face. It may therefore be assumed that there exist at least two distinct values among $\beta_1, \ldots, \beta_{M-1}$. Assume without loss of generality that $\beta_1 \neq \beta_2$.

By symmetry, $\mathcal{C}$ also intersects the 1-face at $(\beta_2, \beta_1, \ldots, 1)$, and indeed at any other permutation of the components $\beta_1, \ldots, \beta_{M-1}$. Intuitively, since the number of permutations is even, this means that every time $\mathcal{C}$ leaves the 1-face it must re-enter it again (Figure 7). Note that $\mathcal{C}$ must lie outside of U_M between the two permutations of the $\{\beta_i\}$, since $\mathcal{C}$ is the boundary of a convex region of $\mathbb{R}^M$.

Now the surface $\mathcal{C}$ is known to pass through the point $(a, b, \ldots, b)$ where $a = a(M, S_1, S_2) < 0$ and $b = b(M, S_1, S_2) \leq 1$. If $\mathcal{C}$ passes through this point without intersecting any 0-faces of U_M, then it must do so between leaving and re-entering a 1-face. This is not possible, since $\mathcal{C}$ is the boundary of a convex region of $\mathbb{R}^M$. It follows that $\mathcal{C}$ must pass through a 0-face of U_M at some point. Specifically, as any component α_k varies from $a(M, S_1, S_2) < 0$ to $\gamma_k > 0$, the surface $\mathcal{C}$ passes through the face of U_M with $\alpha_k = 0$. Solutions to the constraints (a) and (b) therefore exist within the $(M-1)$-dimensional hypercube U_{M-1}, since U_{M-1} corresponds to a 0-face of U_M.

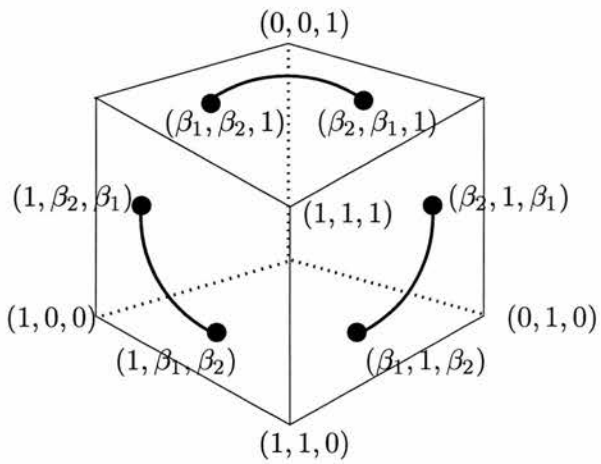


Figure 7: Illustration of the situation where the surface of solutions $\mathcal{C}$ intersects only the 1-faces of U_M, when $M = 3$. The bold circular segments denote those parts of $\mathcal{C}$ lying outside of U_3. Whenever $\mathcal{C}$ leaves U_3, it must re-enter on the same face.

The inductive hypothesis may now be applied for case (1) of page 240. Since $\lceil S_1^2/S_2 \rceil \leq M-1$, and there exist solutions to the constraints (a), (b) and (c) in $\mathbb{R}^{M-1}$, by the inductive hypothesis there exists $n = \lceil S_1^2/S_2 \rceil$ such that $(a', \underbrace{b', \ldots, b'}_{n-1}, 0, \ldots, 0) \in \mathbb{R}^{M-1}$ lies in U_{M-1}, where $a' = a(n, S_1, S_2)$ and $b' = b(n, S_1, S_2)$. Thus $(a', \underbrace{b', \ldots, b'}_{n-1}, 0, \ldots, 0, 0) \in \mathbb{R}^M$ lies in U_M.

(2) Suppose now that $M - 1 < n = \lceil S_1^2/S_2 \rceil \leq M$. Let

$$\alpha_1^* = a(M, S_1, S_2) = \frac{S_1}{M} - \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,.$$

Now

$$\frac{S_1^2}{S_2} > M - 1 \quad \Rightarrow \quad MS_1^2 > M(M-1)S_2 \quad \Rightarrow \quad S_1^2 > (M-1)(MS_2 - S_1^2)\,,$$

whence $\alpha_1^* > 0$. Further, $S_1 \leq M$ since S_1 is the sum of M numbers in $[0, 1]$, so $\alpha_1^* \leq 1$.

Thus $0 < \alpha_1^* \leq 1$.

Let

$$\alpha_2^* = \ldots = \alpha_M^* \;\; = \;\; b(M, S_1, S_2) \;\; = \frac{S_1}{M} + \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,.$$

The vector $(\alpha_1^*, \ldots, \alpha_M^*)$ satisfies the constraints (a) and (b). By Lemma 4, $b(M, S_1, S_2) \leq 1$, and clearly $b(M, S_1, S_2) \geq 0$. Therefore $(\alpha_1^*, \ldots, \alpha_M^*) \in U_M$ and the inductive hypothesis holds with $n = M$.

It remains to show that $n = \lceil S_1^2/S_2 \rceil$ is the unique value such that $\boldsymbol{\alpha}'_{\boldsymbol{n}} \in U_M$ and $a(n, S_1, S_2) > 0$. If $r < n$ then $r < S_1^2/S_2$ so $rS_2 - S_1^2 < 0$ and the quantities $a(r, S_1, S_2)$ and $b(r, S_1, S_2)$ are not real. When $r > n$,

$$\frac{S_1^2}{S_2} \leq r - 1 \quad \Rightarrow \quad rS_1^2 \leq r(r-1)S_2 \quad \Rightarrow \quad S_1^2 \leq (r-1)(rS_2 - S_1^2)\,,$$

whence $a(r, S_1, S_2) \leq 0$. □

Proposition 13 *Let $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^{M}(1 - \alpha_k x)$ and once again let the constraints (a), (b) and (c) in $\mathbb{R}^M$ be defined as (a) $\sum_{k=1}^{M} \alpha_k = S_1$, (b) $\sum_{k=1}^{M} \alpha_k^2 = S_2$, (c) $0 \leq \alpha_k \leq 1 \;\; \forall k = 1, \ldots, M$ (that is, $\boldsymbol{\alpha} \in U_M$).*

Suppose there exists at least one solution to (a), (b) and (c). Then
(i) The global minimum of f_M for fixed x, subject to the constraints (a) and (b), occurs at $\boldsymbol{\alpha}^* = (\underbrace{c^*, \ldots, c^*}_{M-1}, d^*)$ where $c^* \leq d^*$, and

$$c^*(M, S_1, S_2) = \frac{S_1}{M} - \frac{1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,, \qquad d^*(M, S_1, S_2) = \frac{S_1}{M} + \frac{M-1}{M}\sqrt{\frac{MS_2 - S_1^2}{M-1}}\,.$$

(ii) If $(c^*, \ldots, c^*, d^*)$ does not satisfy the constraint (c), then the minimum of f_M subject to (a), (b) and (c) occurs at $\boldsymbol{\alpha}'_{\boldsymbol{n}} = (\underbrace{c', \ldots, c'}_{n-1}, d', \underbrace{1, \ldots, 1}_{M-n})$ for some n, where $c' = c(n, S_1 - (M-n), S_2 - (M-n))$ and $d' = d(n, S_1 - (M-n), S_2 - (M-n))$.

(iii) There exists some $n \leq M$ for which $\boldsymbol{\alpha}'_{\boldsymbol{n}}$ lies inside U_M, and $n = \left\lceil \frac{(M-S_1)^2}{(S_2 - 2S_1 + M)} \right\rceil$. This is the unique value of n for which $\boldsymbol{\alpha}'_{\boldsymbol{n}} \in U_M$ and $d(n, S_1 - (M-n), S_2 - (M-n)) < 1$.

Proof (i) and (ii) are proved in the same way as (i) and (ii) of Proposition 12.

(iii) First note that $n = \left\lceil \frac{(M-S_1)^2}{(S_2 - 2S_1 + M)} \right\rceil$ satisfies $1 \leq n \leq M$: the first inequality is clear,

while for the second, Lemma 1 gives $S_1^2 \leq MS_2$, from which

$$\begin{aligned} M^2 - 2MS_1 + S_1^2 &\leq M^2 - 2MS_1 + MS_2 \\ \Rightarrow \quad (M - S_1)^2 &\leq M(S_2 - 2S_1 + M) \\ \Rightarrow \quad \tfrac{(M-S_1)^2}{(S_2-2S_1+M)} &\leq M\,. \end{aligned}$$

Now consider part (iii) of the proposition. The result holds for $M = 2$ as follows. If there exists a solution to the three constraints (a), (b) and (c) in $\mathbb{R}^2$, there must exist one of the form (c, d) with $c \leq d$. The quantity $n = \left\lceil \frac{(2-S_1)^2}{(S_2-2S_1+2)} \right\rceil$ is equal to either 1 or 2. Routine algebra shows that $n = 1$ if and only if $\sqrt{S_2} \geq \sqrt{(S_1-1)^2+1}$, whence by examining Figure 8 and using Pythagoras' rule, it is clear that $n = 1 \iff d \geq 1$. However, since there are at least one and at most two solutions to (a), (b) and (c) in $\mathbb{R}^2$, at (c, d) and (d, c), it follows that $d \leq 1$. Thus $n = 1 \iff d = 1$ and the result holds for $M = 2$.

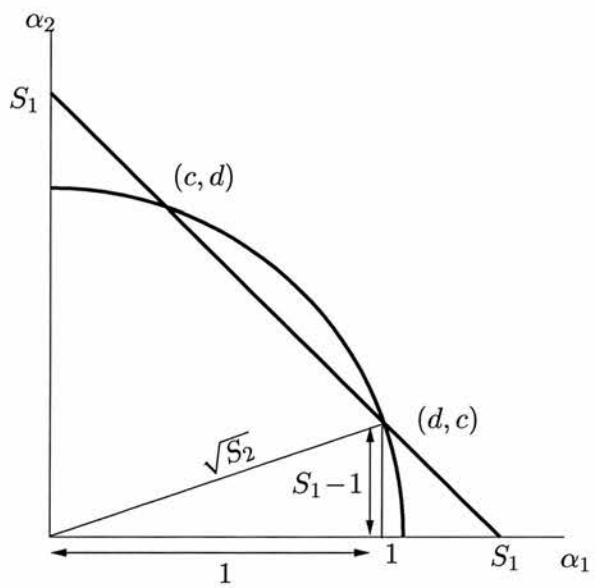


Figure 8: Intersection of the circle $\alpha_1^2 + \alpha_2^2 = S_2$ with the line $\alpha_1 + \alpha_2 = S_1$. If there are solutions to these constraints in U_2, and if $\sqrt{S_2} \geq \sqrt{(S_1-1)^2+1}$, then the point of intersection (d, c) must occur at $d = 1$ by Pythagoras' law.

The inductive hypothesis is now formulated. Suppose that for all integers $r \leq M - 1$, if there are solutions to the constraints (a), (b) and (c) in $\mathbb{R}^r$, then there exist c and d such that $c \leq d$ and $(\underbrace{c, \ldots, c}_{n-1}, d, \underbrace{1, \ldots, 1}_{r-n}) \in U_r$, where $n = \lceil \frac{(r-S_1)^2}{S_2-2S_1+r} \rceil$, and this vector is a solution to (a), (b) and (c). By Lemma 3, the values of c and d are given by $c = c(r, S_1 - r + n, S_2 - r + n)$, $d = d(r, S_1 - r + n, S_2 - r + n)$.

Now consider $r = M$. There are two cases:
(1) $n = \lceil \frac{(M-S_1)^2}{S_2-2S_1+M} \rceil \leq M-1$,
(2) $M-1 < n \leq M$.

(1) Let $\alpha_M^* = 1$. This leaves the remaining $(M-1)$ components to sum to $S_1 - 1$ and their squares to sum to $S_2 - 1$. Now

$$\left\lceil \frac{((M-1)-(S_1-1))^2}{(S_2-1)-2(S_1-1)+(M-1)} \right\rceil = \left\lceil \frac{(M-S_1)^2}{S_2-2S_1+M} \right\rceil \leq M-1\,,$$

and consequently this is the $(M-1)$-dimensional case where the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{M-1})$ satisfies $\sum_{k=1}^{M-1} \alpha_k = S_1 - 1$ and $\sum_{k=1}^{M-1} \alpha_k^2 = S_2 - 1$. Once again, it is necessary to show that there exist solutions to these constraints within the $(M-1)$-dimensional hypercube U_{M-1} in order to apply the inductive hypothesis.

To show this, note that

$$n = \left\lceil \frac{(M-S_1)^2}{S_2-2S_1+M} \right\rceil \leq M-1 \iff (M-S_1)^2 \leq (M-1)(S_2-2S_1+M)\,, \tag{10}$$

whereupon rearrangement yields

$$\frac{(S_1-1)^2}{S_2-1} \leq M-1\,. \tag{11}$$

But the proof of Lemma 5 demonstrates that $\alpha \in \mathbb{R}$ and $(S_1-\alpha)^2/(S_2-\alpha^2) \leq M-1$ if and only if $a(M,S_1,S_2) \leq \alpha \leq d(M,S_1,S_2)$. Putting $\alpha = 1$ and using (10) and (11) gives

$$n \leq M-1 \iff d(M,S_1,S_2) \geq 1 \geq a(M,S_1,S_2) \tag{12}$$

If $d(M,S_1,S_2) = 1$, the vector $(c(M,S_1,S_2), \ldots, c(M,S_1,S_2), 1)$ is itself a solution to the constraints (a) and (b) on a 1-face of U_M, as $c(M,S_1,S_2) \geq 0$ by Lemma 4. Suppose instead that $d(M,S_1,S_2) > 1$. The point $(c(M,S_1,S_2), \ldots, c(M,S_1,S_2), d(M,S_1,S_2))$ is a solution to the constraints (a) and (b) that lies outside of the unit hypercube U_M. By hypothesis there exists at least one solution to (a) and (b) inside U_M, say $\boldsymbol{\gamma}$. The surface $\mathcal{C}$ of solutions to the constraints (a) and (b) must therefore pass through a face of U_M. Using arguments analogous to those of Proposition 12 part (iii), the surface $\mathcal{C}$ must pass through a 1-face of U_M, and therefore there exist solutions to the revised constraints (a′) $\sum_{k=1}^{M-1} \alpha_k = S_1 - 1$ and (b′) $\sum_{k=1}^{M-1} \alpha_k^2 = S_2 - 1$ in the $(M-1)$-dimensional hypercube U_{M-1}.

The inductive hypothesis may now be applied for case (1). Since $\left\lceil \frac{((M-1)-(S_1-1))^2}{(S_2-1)-2(S_1-1)+(M-1)} \right\rceil = \left\lceil \frac{(M-S_1)^2}{(S_2-2S_1+M)} \right\rceil \leq M-1$ and there exist solutions to the constraints (a'), (b') and (c) in $\mathbb{R}^{M-1}$, by the inductive hypothesis there exists $n = \left\lceil \frac{((M-1)-(S_1-1))^2}{(S_2-1)-2(S_1-1)+(M-1)} \right\rceil = \left\lceil \frac{(M-S_1)^2}{S_2-2S_1+M} \right\rceil$ such that $(\underbrace{c,\ldots,c}_{n-1},d,\underbrace{1,\ldots,1}_{M-1-n}) \in \mathbb{R}^{M-1}$ lies in U_{M-1}, and $c \leq d$. Hence $(\underbrace{c,\ldots,c}_{n-1},d,\underbrace{1,\ldots,1,1}_{M-n}) \in \mathbb{R}^{M}$ lies in U_M, and the inductive hypothesis holds.

(2) Suppose now that $M-1 < \frac{(M-S_1)^2}{(S_2-2S_1+M)} \leq M$, so that $\left\lceil \frac{(M-S_1)^2}{S_2-2S_1+M} \right\rceil = M$. Let

$$\alpha_M^* = d(M,S_1,S_2) = \frac{S_1}{M} + \frac{M-1}{M}\sqrt{\frac{MS_2-S_1^2}{M-1}}\,.$$

Clearly $\alpha_M^* \geq 0$, and $\alpha_M^* < 1$ by (12); so $0 \leq \alpha_M^* < 1$.

Let

$$\alpha_1^* = \ldots = \alpha_{M-1}^* \;=\; c(M,S_1,S_2) \;= \frac{S_1}{M} - \frac{1}{M}\sqrt{\frac{MS_2-S_1^2}{M-1}}\,.$$

The vector $(\alpha_1^*,\ldots,\alpha_M^*)$ satisfies the constraints (a) and (b). By Lemma 4, $c(M,S_1,S_2) \geq 0$, and clearly $c(M,S_1,S_2) \leq 1$. Therefore $(\alpha_1^*,\ldots,\alpha_M^*) \in U_M$ and the inductive hypothesis holds with $n = M$.

It remains to show that $n = \left\lceil \frac{(M-S_1)^2}{(S_2-2S_1+M)} \right\rceil$ is the unique value such that $\boldsymbol{\alpha_n'} \in U_M$ and $d(n, S_1-(M-n), S_2-(M-n)) < 1$.

If $r < n$ then

$$r < \frac{(M-S_1)^2}{S_2-2S_1+M}\,.$$

Manipulation of this expression gives

$$r(S_2-M+r) - (S_1-M+r)^2 < 0\,,$$

and consequently

$$d(r, S_1-M+r, S_2-M+r) = \frac{S_1-M+r}{r} + \frac{r-1}{r}\sqrt{\frac{r(S_2-M+r)-(S_1-M+r)^2}{r-1}}$$

is not real. Similarly, $c(r, S_1-M+r, S_2-M+r)$ is not real.

If $r > n$ then

$$\frac{(M-S_1)^2}{S_2-2S_1+M} \leq r-1\,,$$

which after rearrangement gives

$$\frac{(S_1-(M-r)-1)^2}{(S_2-(M-r)-1)} \leq r-1\,. \tag{13}$$

By Lemma 5, if $\alpha \in \mathbb{R}$ and $(S_1-(M-r)-\alpha)^2/(S_2-(M-r)-\alpha^2) \leq r-1$, then $\alpha \leq d(r, S_1-(M-r), S_2-(M-r))$. Putting $\alpha = 1$ and using (13) demonstrates that $d(r, S_1-(M-r), S_2-(M-r)) \geq 1$ when $r > n$. Therefore n is the unique value for which $\boldsymbol{\alpha}'_{\boldsymbol{n}} \in U_M$ and $d(n, S_1-(M-n), S_2-(M-n)) < 1$. □

Propositions 12 and 13 provide the required results about the quadratic error function $\epsilon_2(x)$. The maximum quadratic error $\epsilon_2(x)$ for fixed x, subject to $\sum_{k=1}^{M}\alpha_k = S_1$ and $\sum_{k=1}^{M}\alpha_k^2 = S_2$, occurs at the minimum of $f_M(\boldsymbol{\alpha}) = \prod_{k=1}^{M}(1-\alpha_k x)$: that is, when

$$\boldsymbol{\alpha} = (\underbrace{c, \ldots, c}_{n-1}, d, \underbrace{1, \ldots, 1}_{M-n})\,;$$

and

$$c = c(n, S_1-M+n, S_2-M+n) = \frac{S_1-M+n}{n} - \frac{1}{n}\sqrt{\frac{n(S_2-M+n)-(S_1-M+n)^2}{n-1}}\,,$$

$$d = d(n, S_1-M+n, S_2-M+n) = \frac{S_1-M+n}{n} + \frac{n-1}{n}\sqrt{\frac{n(S_2-M+n)-(S_1-M+n)^2}{n-1}}\,,$$

and

$$n = \left\lceil \frac{(M-S_1)^2}{(S_2-2S_1+M)} \right\rceil .$$

The maximum quadratic error is therefore given by

$$\max_{\boldsymbol{\alpha}} \Big(\epsilon_2(x)\Big) = \exp\left\{-S_1 x - S_2\frac{x^2}{2}\right\} - \Big(1-cx\Big)^{n-1}\Big(1-dx\Big)\Big(1-x\Big)^{M-n}\,. \tag{14}$$

The minimum quadratic error $\epsilon_2(x)$ for fixed x, subject to $\sum_{k=1}^{M}\alpha_k = S_1$ and $\sum_{k=1}^{M}\alpha_k^2 = S_2$, occurs at the maximum of $f_M(\boldsymbol{\alpha}) = \coprod_{k=1}^{M}(1-\alpha_k x)$: that is, when

$$\boldsymbol{\alpha} = (a, \underbrace{b, \ldots, b}_{n-1}, \underbrace{0, \ldots, 0}_{M-n})\,;$$

and

$$a = a(n, S_1, S_2) = \frac{S_1}{n} - \frac{n-1}{n}\sqrt{\frac{nS_2-S_1^2}{n-1}}\,,$$

$$b = b(n, S_1, S_2) = \frac{S_1}{n} + \frac{1}{n}\sqrt{\frac{nS_2 - S_1^2}{n-1}},$$

and

$$n = \left\lceil \frac{S_1^2}{S_2} \right\rceil .$$

The minimum quadratic error is therefore given by

$$\min_{\boldsymbol{\alpha}} \Big(\epsilon_2(x)\Big) = \exp\left\{-S_1 x - S_2 \frac{x^2}{2}\right\} - \Big(1 - ax\Big)\Big(1 - bx\Big)^{n-1} . \tag{15}$$

Remark The results of this section may be used to derive general inequalities for samples of numbers between 0 and 1. Let the sample $\alpha_1, \ldots, \alpha_M$ satisfy $0 \le \alpha_k \le 1 \quad \forall k$, and let $S_1 = \sum_{k=1}^{M} \alpha_k$ and $S_2 = \sum_{k=1}^{M} \alpha_k^2$. The mean of the sample is $\hat{\mu} = S_1/M$, and the sample standard deviation is given by

$$\hat{\sigma} = \sqrt{\frac{\sum_{k=1}^{M}(\alpha_k - \bar{\alpha})^2}{M-1}} = \sqrt{\frac{MS_2 - S_1^2}{M(M-1)}} .$$

Thus

$$a(M, S_1, S_2) = \hat{\mu} - \frac{M-1}{\sqrt{M}}\hat{\sigma}, \qquad b(M, S_1, S_2) = \hat{\mu} + \frac{1}{\sqrt{M}}\hat{\sigma},$$

$$c(M, S_1, S_2) = \hat{\mu} - \frac{1}{\sqrt{M}}\hat{\sigma}, \qquad d(M, S_1, S_2) = \hat{\mu} + \frac{M-1}{\sqrt{M}}\hat{\sigma},$$

The result of Lemma 5 indicates that $a(M, S_1, S_2) \le \alpha_k \le d(M, S_1, S_2)$ for all k: that is, the maximum and minimum of a sample of M numbers between 0 and 1 can lie no further than $(M-1)\hat{\sigma}/\sqrt{M}$ from the sample mean $\hat{\mu}$, where $\hat{\sigma}$ is the sample standard deviation.

Lemma 4 gives the further results that

$$\hat{\mu} \ge \frac{1}{\sqrt{M}}\hat{\sigma}, \qquad \hat{\mu} \le 1 - \frac{1}{\sqrt{M}}\hat{\sigma}.$$

The maximum and minimum of the function $f(\boldsymbol{\alpha}) = \prod_{k=1}^{M}(1-\alpha_k)$, subject to fixed mean and variance of the sample $\{\alpha_k\}_{k=1}^{M}$, are provided by Propositions 12 and 13 respectively. If the sample $\alpha_1, \ldots, \alpha_M$ represents the success probabilities for M independent Bernoulli trials, then these bounds give the maximum and minimum probability of failure in all trials, subject to fixed mean and variance of the probabilities.

Practical uses for these results have not been found.

Appendix B

Glossary of notation and terminology for Chapter 5

The glossary is divided into miscellaneous symbols, Greek symbols, Roman symbols and words or phrases. Page numbers are given where appropriate.

Miscellaneous symbols

$\lceil\ \rceil$: integer up from, e.g. $\lceil x \rceil$ is the least integer r such that $x \leq r$.

$\lfloor\ \rfloor$: integer part of, e.g. $\lfloor x \rfloor$ is the greatest integer r such that $x \geq r$.

$\emptyset$: the empty set, $\{\ \}$.

$\rightarrow^{(t)}$: colonization notation. The event $i \rightarrow^{(t)} h$ is the event that individuals in site i at time t or their offspring have colonized site h at time $t+1$. (Page 103).

$\not\rightarrow^{(t)}$: non-colonization. The event $i \not\rightarrow^{(t)} h$ is the event that individuals in site i at time t or their offspring have not colonized site h at time $t+1$. (Page 103).

$\mathbb{E}$: symbol denoting expectation, e.g. $\mathbb{E}(X)$ is the expectation of the random variable X.

$\mathbb{E}_{\boldsymbol{\theta}}$: expectation given parameters $\boldsymbol{\theta}$, e.g. $\mathbb{E}_{\boldsymbol{\theta}}(X)$ is the expectation of the random variable X when the values of the parameters governing the distribution of X are $\boldsymbol{\theta}$.

$\mathbb{P}$: symbol denoting probability, e.g. $\mathbb{P}(A)$ is the probability of event A occurring.

$\hat{\mathbb{P}}$: estimated probability, e.g. $\hat{\mathbb{P}}(A)$ is the estimated probability that event A occurs.

$\mathbb{R}^M$: space of M-dimensional real vectors.

Greek symbols

δ_{ih} : the distance between two sites i and h in the survey region. (Page 48).

ϵ_1 : linear error function for the linear exponential approximation to the function $s_i^{(t)}$ when $t = 1$ or 2. The function is given by $\epsilon_1(x) = \exp\{-S_1 x\} - \prod_{k=1}^{M}(1-\alpha_k x)$, where $\exp\{-S_1 x\}$ is the linear exponential approximation to $s_i^{(t)}(x)$, and $\prod_{k=1}^{M}(1-\alpha_k x)$ is the shorthand expression for $s_i^{(t)}(x)$ when $t = 1$ or 2. (Page 123).

ϵ_2 : quadratic error function for the quadratic exponential approximation to the function $s_i^{(t)}$ when $t = 1$ or 2. The function is given by $\epsilon_2(x) = \exp\{-S_1 x - S_2 x^2/2\} - \prod_{k=1}^{M}(1-\alpha_k x)$, where $\exp\{-S_1 x - S_2 x^2/2\}$ is the quadratic exponential approximation to $s_i^{(t)}(x)$, and $\prod_{k=1}^{M}(1-\alpha_k x)$ is the shorthand expression for $s_i^{(t)}(x)$ when $t = 1$ or 2. (Page 123).

$\boldsymbol{\theta}$: the vector of parameters that enter the colonization model through the colonization probabilities $\{p_{ih}\}$. (Page 51).

ς_h : quantity measuring habitat quality in the site h. Used in the colonization probabilities p_{ih}. (Page 48).

$\varsigma_h^{(t)}$: habitat quality in site h at time t, when there is time-dependence.

Roman symbols

$\boldsymbol{a}$: vector of length N denoting inclusion in the set of sites A. The ith component is $a_i = 0$ if site $i \notin A$, $a_i = 1$ if site $i \in A$. (Page 146).

a : one of the parameters of the colonization probabilities used to illustrate the colonization model. The colonization probability is illustrated by the form $p_{ih} = p_0 \exp(-a\,\delta_{ih} - b\,\varsigma_h)$. (Page 48).

b : one of the parameters of the colonization probabilities used to illustrate the colonization model. The colonization probability is illustrated by the form $p_{ih} = p_0 \exp(-a\,\delta_{ih} - b\,\varsigma_h)$. (Page 48).

g : function given by $g(x) = \log s_i^{(t)}(x)$ for a site i and time t. The function g is not indexed by i and t for notational convenience. (Page 111).

g' : first derivative of the function g with respect to its single argument.

g'' : second derivative of the function g with respect to its single argument.

g_A : function given by $g_A(x) = \log s_A^{(t)}(x)$ for a set of sites A and time t. The function g_A is not indexed by A and t for notational convenience. (Page 145).

g'_A : first derivative of the function g_A with respect to its single argument.

g''_A : second derivative of the function g_A with respect to its single argument.

$i_r(\boldsymbol{\theta})$: information matrix for a single observation. (Page 176).

$i_\cdot(\boldsymbol{\theta})$: information matrix for a full sample. Given by $i_\cdot(\boldsymbol{\theta}) = \sum_{r=1}^{N} i_r(\boldsymbol{\theta})$ when the full-sample likelihood is the product of the likelihoods for single observations.

$I\{\ \}$: indicator function, e.g. $I\{A\}$ takes value 1 if event A occurs, value 0 if A does not occur.

$I\{k_u \rightarrow^{(u)} k_{u+1}\}$: indicator function taking value 1 if a single colony in the site k_u at time u colonizes site k_{u+1} at time $u+1$. (Page 105).

$L\left(\boldsymbol{y^{(t_2)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(t_1)}}\right)$: conditional likelihood of the distribution $\boldsymbol{y^{(t_2)}}$ given the observed distribution $\boldsymbol{y^{(t_1)}}$, where $t_1 < t_2$. The likelihood is a function of the unknown parameters $\boldsymbol{\theta}$.

$\hat{L}\left(\boldsymbol{y^{(t_2)}} \,\middle|\, \boldsymbol{\theta},\ \boldsymbol{y^{(t_1)}}\right)$: estimated value of $L\left(\boldsymbol{y^{(t_2)}} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{y^{(t_1)}}\right)$.

$M^{(t_1,\ldots,t_r)}$: matrix formed as the element-wise square of the matrix $Q^{(t_1)}Q^{(t_2)}\ldots Q^{(t_r)}$. The $(i,\, h)$ component of $M^{(t_1,\ldots,t_r)}$ is given by $M^{(t_1,\ldots,t_r)}[i,\, h] = m_{ih}^{(t_1,\ldots,t_r)} = \left(Q^{(t_1)}Q^{(t_2)}\ldots Q^{(t_k)}[i,\, h]\right)^2$. (Page 115).

$m_{ih}^{(t_1,\ldots,t_r)}$: $(i,\, h)$ component of the matrix $M^{(t_1,\ldots,t_r)}$. (Page 115).

N : number of sites in the survey.

$N_i^{(t)}$: random variable denoting number of colonies in site i at time t. (Page 103).

$N_A^{(t)}$: random variable denoting the total number of colonies in the set of sites A at time t. Given by $N_A^{(t)} = \sum_{i \in A} N_i^{(t)}$. (Page 144).

p_{ih} : the colonization probability p_{ih} is the probability that individuals or their offspring in site i at any time-point have colonized site h by the next time-point. (Page 48).

$p_{ih}^{(t)}$: probability of colonization from site i at time t to site h at time $t+1$, when there is time-dependence. When there is no time-dependence, $p_{ih}^{(t)}$ is simply written p_{ih}.

$p_i^{(t)}$: probability that site i is occupied at time t. (Page 103).

p_0 : one of the parameters of the colonization probabilities used to illustrate the colonization model. The colonization probability is illustrated by the form $p_{ih} = p_0 \exp(-a\,\delta_{ih} - b\,\varsigma_h)$. (Page 48).

$q_{ih}^{(t)}$: the colony-scale colonization probability that a single colony present in site i at time t will colonize site h at time $t+1$. (Page 100).

$Q^{(t)}$: matrix with $(i,\, h)$ component $Q^{(t)}[i,\, h] = q_{ih}^{(t)}$. (Page 114).

$Q^{(t)}[i,\, h]$: $(i,\, h)$ component of the matrix $Q^{(t)}$.

$Q^{(t_1)}Q^{(t_2)}\ldots Q^{(t_r)}[i,\, h]$: $(i,\, h)$ component of the matrix product $Q^{(t_1)}Q^{(t_2)}\ldots Q^{(t_r)}$.

$\boldsymbol{r_A}^{(t_1,\ldots,t_n)}$: vector with ith component $\boldsymbol{r_A}^{(t_1,\ldots,t_n)}[i] = \left(Q^{(t_1)}\ldots Q^{(t_n)}\boldsymbol{a}[i]\right)^2$, where $\boldsymbol{a}$ is the vector denoting inclusion of set A (see glossary entry for $\boldsymbol{a}$). (Page 146).

$s_i^{(t)}$: function given by $s_i^{(t)}(x) = \mathbb{E}\left\{(1-x)^{N_i^{(t)}}\right\}$ for a site i and time t. (Page 104).

$s_A^{(t)}$: function given by $s_A^{(t)}(x) = \mathbb{E}\left\{(1-x)^{N_A^{(t)}}\right\}$ for a set of sites A and time t. (Page 145).

S_1 : the sum $\sum_{k=1}^{M}\alpha_k$, where $\prod_{k=1}^{M}(1-\alpha_k x)$ is a shorthand for the function $s_i^{(t)}(x)$ when $t = 1$ or 2. (Page 122).

S_2 : the sum $\sum_{k=1}^{M}\alpha_k^2$, where $\prod_{k=1}^{M}(1-\alpha_k x)$ is a shorthand for the function $s_i^{(t)}(x)$ when $t = 1$ or 2. (Page 122).

T : the time-point of the final survey.

$\boldsymbol{u^{(t)}}$: vector of length N with transpose $\boldsymbol{u^{(t)\prime}} = \boldsymbol{y^{(0)\prime}}Q^{(0)}\ldots Q^{(t-1)}$. (Page 155).

$V_i^{(t)}(k,u)$: random variable denoting the number of colonies established in site i at time t, beginning from a single ancestor colony in site k at time u. (Page 104).

$V_A^{(t)}(k,u)$: random variable denoting the number of colonies established in the set of sites A at time t, beginning from a single ancestor colony in site k at time u. Given by $V_A^{(t)}(k,u) = \sum_{i\in A} V_i^{(t)}(k,u)$. (Page 145).

$\boldsymbol{y^{(t)}}$: vector of length N containing the observed distribution at time t. The ith component is 1 or 0 according to whether site i is occupied or unoccupied at time t.

$\boldsymbol{y^{(t)'}}$: vector transpose of the vector $\boldsymbol{y^{(t)}}$.

$y_i^{(t)}$: ith component of $\boldsymbol{y^{(t)}}$, giving the status of site i at time t. Value is 0 if site i is unoccupied at time t, 1 if site i is occupied at time t.

$\boldsymbol{z^{(t)}}$: vector of length N with transpose $\boldsymbol{z^{(t)'}} = \boldsymbol{y^{(0)'}} Q^{(0)} \dots Q^{(t-2)} M^{(t-1)}$. (Page 156).

Words and phrases

BAM : best attainable match. Developed in Chapter 4.

colonization : the process of individuals moving from one site at one time-point to another site at the next time-point. (Page 101).

colonization probability : the colonization probability p_{ih} is the probability that individuals or their offspring in site i at any time-point have colonized site h by the next time-point. (Page 48).

colony : a colony is established as a result of every colonization. Each colony has unique ancestry and acts independently of every other colony. (Page 101).

colony-scale colonization probability : the probability $q_{ih}^{(t)}$ that a single colony present in site i at time t will colonize site h at time $t+1$. (Page 100).

colony-scale polynomial : the polynomial $s_i^{(t)}(x) - 1 + p_{ih} p_i^{(t)}$. (Page 109).

extinction probability : the extinction probability at time t of a set A of sites is the probability that all sites in A are unoccupied at time t. (Page 144).

first-order approximation : see linear exponential approximation.

linear exponential approximation : approximation to the function $s_i^{(t)}(x)$ by an exponential function of the form $\exp(\beta x)$, where β is the coefficient of x in the Taylor expansion of $\log\left(s_i^{(t)}(x)\right)$ about 0. Also referred to as the first-order approximation. (Page 118).

product-likelihood : likelihood obtained by multiplication of the individual probabilities or likelihoods of correlated observations. (Page 151).

quadratic exponential approximation : approximation to the function $s_i^{(t)}(x)$ by an exponential function of the form $\exp(\beta x + \gamma x^2)$, where β is the coefficient of x in the Taylor expansion of $\log\left(s_i^{(t)}(x)\right)$ about 0, and γ is the coefficient of x^2 in the Taylor expansion. Also referred to as the second-order approximation. (Page 118).

second-order approximation : see quadratic exponential approximation.

site-scale colonization probability : see colonization probability. Refers to movement of individuals between sites, rather than between colonies within sites.

time-point : the dividing points of generations. At each time-point, a survey either took place or is considered missing. The distributions $\boldsymbol{y}^{(0)}, \ldots, \boldsymbol{y}^{(T)}$ therefore correspond to the time-points $0, \ldots, T$.

time-step : the period between time-points denotes one time-step.