Bayesian Models for Pest Detection

Chris Malone

2022-04-26

Contents

Chapter 1

Introduction

1.1 CHECKLIST

- Description of the problem
- Review what is known
- Identify gaps
- What I plan to do to address gaps
- How I plan to carry out investigation
- Explanation of contribution
- Outline of content

1.2 Introducing the research

1.2.1 Establish background area (biosecurity, monitoring)

1.2.2 Establish specific problem (identify gap explicitly)

1.2.3 State the research question

1.3 Outline of the thesis

Chapter 2

Literature review

2.1 Introduction

2.1.1 Description of what this chapter is about

2.1.2 Outline of the chapter

2.2 Non-literature background

2.2.1 Eradicating tropical fruit flies

Costs associated with invasive and pest species can be varied and significant. Invasive species pose economic, social and environmental costs. Fruit flies are an example of an economically costly pest. Fruit flies lay eggs in ripening fruits and vegetables. Produce that has been infested with fruit fly larvae is

not fit for sale, domestically or abroad. Further, even if produce has not been infested, local suppliers can charge a premium for supplying pest-free produce to international markets.

Governments have a responsibility monitor and eradicate potential biological invasions. In particular, the Australian government is bound by the Biosecurity act () to ensure a very low level of risk for animal and plant invasions. Effectively and efficiently controlling biological invasions requires a principled and rigorous approach to monitoring. However, monitoring biological species is difficult and costly. In particular, the cost of increasing monitoring intensity must be traded off against the cost of failing to detect an incipient invasion.

The manager's role is made more difficult when a pest population is *subdetectable*. Populations may be subdetectable because they can persist at low rates; because they are *taxonomically cryptic* (i.e. hard to distinguish from well known native species); because they are *behaviourally cryptic* (i.e. behave in ways that help them avoid detection by predators; see Kery [2002]); or because the population is suppressed by efforts to eradicate them. Matters are made worse when tests to detect the presence of a population have extremely low sensitivity for other reasons. For example, monitoring traps for fruit flies have very low attractiveness. One study of Mediterannean fruit fly found that only 0.02% of 38.8 million flies were recaptured after release, in a standard surveillance setup in Adelaide (**?**).

Suppose an outbreak of an invasive pest has occurred. Monitoring for the pest

must then be intensified and localised to the area of suspicion. If the pest has since been eradicated, governments are incentivised to declare eradication quickly. This is because

On the other hand, if eradication has *not* occured,

A necessary prerequisite for minimising the probability of failing to detect an existing invasion is to understand what can be inferred about the state of the population from what has been observed in surveillance. In particular, it is desirable to understand the likelihood that extinction has occurred, given that the species has not been detected over a certain period of time.

In Australia, Pest Free Area status is awarded to a number of . However, outbreaks of these pests are periodic. It is typical for PFA status to be suspended following an outbreak.

PFA status is economically valuable to producers in the relevant region. Firstly, there may be offshore markets which require that produce is supplied from a PFA for a given pest. Secondly, in some markets, PFA status may allow the supplier to receive a price premium for goods sold. In other words, PFA status increases the value of local produce on international markets.

- Current codes of practice suggest a lower catch rate required for an outbreak to be declared when supplementary traps are set, but probabilities have not been estimated rigorously.

2.2.2 The problem of concern in this work

As mentioned above, the manager seeks to minimise resources spent on surveillance, while minimising the chance that an eradication is declared when none has occurred. The general problem this thesis is concerned with is to infer whether or not an invasive species has been extinguished based on survey records.

2.3 Literature background

In this subsection, I will describe a number of models that have been applied to the problem of inferring species absence. Existing methods can be categorised as frequentist and Bayesian. The core difference between frequentist and Bayesian methods is in their treatment of the variables that describe the presence or absence of the species. Frequentist methods assume that the species is either absent or present with probability one. In this framework, surveillance activities are considered to be random experiments with fixed parameters. On the other hand, Bayesian methods assume that species absence is uncertain (i.e. random).

2.3.1 Simple models

*** Note: Should I delete this entire subsection? I wrote it before I knew about the papers I discuss in the "Elaborate models" section.

2.3.1.1 McArdle 1990

I start by discussing the most basic approach to program design. This is the frequentist approach described by McArdle [1990]. I emphasise this method because (a) it is conceptually simple and therefore easy to describe, and (b) it is emblematic of other methods in the literature, which are similarly general and simple. First, let the *rarity* of the species $p \in [0, 1]$ be the probability that a species is detected in any given sampling unit. (Sampling units can be arranged spatially or temporally; e.g. a survey that involves checking w weeks at k locations would have wk sampling units.) Then, the number of surveys in which the species is detected is given by $X \sim \text{Binomial}(T, p)$. Accordingly, the probability of *not* detecting the species in T surveys is given by

$$\alpha = \Pr(X = 0) = 1 - (1 - p)^T.$$

The last formula allows us to compute any of the 3 quantities α, p, and t, assuming the other two are given.

Given this framework, the problem of program design becomes the following. We decide *a priori* what the smallest "rarity" p worth detecting is. McArdle supposes that if a species if sufficiently difficult to detect (while, nonetheless, $p > 0$) then it cannot be considered a member of an ecological community, and therefore not worthy of being deemed "present". Write the smallest rarity worth detecting as p_0. Then, we choose a minimum detection probability α that we are willing to accept. For example, we might wish to have chance $\alpha > 0.95$ of

detecting a species, given that it is present. Then, with the above formula, we can rearrange to get the smallest number of survey units T such that detection probability α is achieved.

Statisticians will recognise that the above is essentially power analysis for data modelled as identically and independently distributed Bernoulli trials. This analogy can be made more concrete. For any fixed rarity p, we can derive the probability of observing n or more negative surveys. This is the p-value. We can then reject the hypothesis that there is the rarity is greater than p_0, the rarity worth detecting.

2.3.1.2 Limitations of McArdle's method

Applying the above model to the problem of program design is not straightforward. Firstly, the model is fairly restrictive. It assumes that the probability of detecting the pest population is constant over time. Secondly, in the case of pest populations, it may be difficult to determine the smallest value of p that is "worth" detecting. When a pest species is cryptic, the detectability (p in this model) can be extremely small even when the population is relatively large. Further, invasive potential of the pest may be large, so that even small populations bear a large cost to the decision maker.

2.3.1.3 Similar models

There are several other models in the literature, mirroring McArdle's method in their conceptual simplicity and generality. (See Boakes et al. [2015] for a

relatively complete review.) Bayesian models are given by … and Barnes et al. These methods have the advantage that the posterior distribution is derived analytically. This means that posterior probabilities can be computed extremely efficiently. However, each model assumes a constant growth rate. However, in real contexts, growth rates are highly variable and uncertain (**?**).

2.3.2 Elaborate models

2.3.2.1 Justification for elaborate models

Recently, more elaborate models have appeared in the literature for specific eradication inference problems. Here, I discuss the motivation and justification for these models.

2.3.2.1.1 Motivation As discussed above, simple models, such as McArdle's method, are attractive for their simplicity, generality, and efficiency of computation. However, these virtues may come at the cost of biological plausibility. This is particularly the case when we have a significant degree of prior information about the species and region in question. For high stakes problems, we would like to be able to leverage existing domain knowledge about the location and species. This might mean using process models that are fine grained, scientifically plausible, and based on multiple sources of scientific knowledge and evidence. We would also like to be able to incorporate our uncertainty about the processes in question.

List of problems to discuss…

- They make strong assumptions about priors which are not defensible in general.
  - Assumption of a fixed, and either constant or declining population size (Caley et al. [2015], p. 2).
  - For example, McArdle's model assumes that there is a value of p small enough to be not worth detecting.
- They assume that we can estimate a sighting rate, which is constant as a function of the locations of the individuals.

Another issue with simpler models is their relative inflexibility with respect to the structure of the model. For example, the method of Barnes et al assumes that we know the probability of detecting a specimen drawn randomly from the population. However, in the case of fruit flies, this number is difficult to estimate, because the capture probability is highly dependent on the trapping layout, which in turn is highly dependent on location and time. For example, countries (and states within countries) differ in the types of traps used and the spatial density of traps.

2.3.2.1.2 Plausibility As (**?**) points out, most models in the literature on the program design problem are simple and general. Their simplicity means that they are parsimonious, analytically tractable, and easy to interpret. However, these properties are bought at the cost of plausibility of the model as a description of real ecological processes, which are complex and uncertain. In general, the plausibility of a model will depend on the system that is under in-

vestigation. For a given model, the analyst may have information that will allow them to specify relatively more plausible and realistic models of the processes that govern the system.

2.3.2.2 Paper 1: Keith et al

The first instance of an elaborate Bayesian model for inferring pest eradication is given by Keith and Spring [2013]. The authors use an agent-based Bayesian model to infer the distribution of fire ant nests in Brisbane. They obtained data on the locations and month of discovery for $n = 7{,}068$ nests. They also observed whether data were passively or actively discovered (e.g. by members of the public or through a targeted search).

Generalised Gibbs sampling is used to sample…

2.3.2.3 Sampling

It should be noted that both of the models above are agent based. The model of Keith et al explicitly models the location of each agent (ant nest). In each case, this leads to a difficulty in sampling algorithms. Typically, when a Bayesian model is agent based, this means that there is an unknown number of parameters in the model (and unbounded, as the size of the population typically does not have a finite bound, e.g. when it is Poisson distributed). The upshot of this is that typical Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis Hastings sampler, fail. This is because they do not allow us the dimension of the parameter vector to vary between draws.

The papers discussed above deal with this in two different ways. Keith et al use a generalisation of the Gibbs sampler, inspired by reversible jump MCMC. This gets around the problem by adding a step to the Gibbs sampler in which we move between coordinate spaces . On the other hand, Caley et al use approximate Bayesian computation (ABC).

2.3.2.4 Paper 2: Caley et al

The second instance of an elaborate Bayesian model for inferring eradication is given by Caley et al. [2015]. Caley and co-authors develop a Bayesian model of fox sightings in Tasmania. Their goal is to infer the posterior probability that foxes had been eradicated in Tasmania, given a record of fox carcass sightings. They obtained data on fox carcass sightings from hunter kills and road kills. These observation processes were modelled separately, so that posterior detection rates differed between the sighting types. Detection rates were assumed to be constant across time and location for each type of sighting. Notably, uninformative priors were set for detection rates (i.e. the probability of detecting a fox was set to be uniform on $[0, 1]$). Data consisted of a single sighting count for each location (with Tasmania divided geographically into grid cells) and each year between 2001 and 2013. Data were all zeroes with the exception of exactly four unit observations (sightings of exactly one fox).

The authors use a simple rejection algorithm to sample from the posterior. I will refer to this algorithm as exact Bayesian computation (EBC), to be distinguished from approximate Bayesian computation (ABC). ABC works by first

drawing samples of the parameter vector θ from the prior distribution $\pi(\theta)$, then second, drawing simulated data y_{sim} from the likelihood, then keeping the proposed θ if and only if y_{sim} is an approximate match with the observed data. EBC is the same, except that samples are only kept if y_{sim} is an *exact* match with the observed data. Notably, EBC is only practical in cases where the data consists entirely of discrete variables (i.e. variables with countable or finite ranges of possible values).

The authors point to the complexity of the likelihood to justify the use of ABC over more standard Monte Carlo methods such as Metropolis Hastings or Gibbs sampling. Based on their description of the model, however, it is not clear that Metropolis Hastings would not suffice.

2.4 Gaps

Above, I have discussed the two elaborate computational models in the literature for inferring eradication of an incipient biological invasion. The current work seeks to address two gaps in the literature. Firstly, elaborate Bayesian models have not been explored for inferring eradication of Tephritid fruit flies. Fruit flies pose an interesting case study, because the regulator has fine grained information about the detection system. In particular, the regulator knows the locations of each of the traps. Further, prior research investigating the efficacy of these traps exists. If used carefully, this information can be leveraged through the model's priors to learn from the zero-sighting record efficiently.

Fruit flies are interesting for a few reasons:

- A body of research exists to understand trap efficacy.
- The surveillance program is predictably structured
- The intensity of the surveillance program changes in a predictable way.

Second, no such model has addressed the question of PFA status reinstatement.

Instead, they discuss inferring eradication from an actually observed record.

Chapter 3

Case study: Ceratatis capitata

3.1 Introduction

In the previous section, a general model for inferring absence/presence of a cryptic, incipient pest species population was introduced. In this chapter, I apply the model to a specific case study. The case study is for the outbreak of Mediterranean fruit fly (Ceratitis Capitata). Prior distributions are set and justified, and posterior inference is performed using approximate Bayesian computation.

- Note that this is a mock analysis intended to improve understanding of the surveillance system.

- Note that the method can be used for a real scenario - we only need to change the priors, data and locations of the traps.
- Note that I use a simplified model of Medfly dynamics for illustrative purposes - but the method can easily incorporate more complex ABS models, and cite those models.
- Recap previous chapters
  - Explain how this chapter ties in
- Explain what this chapter is about
- OUtline this chapter

In this chapter, I present an illustrative model of medfly surveillance after an hypothetical invasion. I use a simplified model of Medfly population dynamics. However, for various species of Tephritid fruit fly (medfly included) detailed models exist. A benefit of the proposed method is that it can easily incorporate almost any model of medfly dynamics, so long as the population density.[1]

3.2 Background

3.2.1 Medfly are economically important

Mediterranean fruit fly (*Ceratitis Capitata*) or *medfly* is a fly species native to sub-Saharan Africa. It is considered to be of high economic importance, due to

[1] Note, though that the sampling method I use may not be appropriate in all cases. When the model predicts that the population size "explodes", then the rejection rate for the sampling algorithm may become very high, causing the algorithm to be highly inefficient.

its potential for destruction of fruit production (Sciarette et al. 2018). Medfly has high invasive potential, as it can adapt to a relatively large range of climates and environments, and is known to have the capability to infest the fruits of over 300 species of plants (Ibid.).

3.2.2 Medfly are cryptic

Medfly are very hard to detect at low levels. Monitoring for medfly is typically performed with the aid of lured traps (namely so-called Lynfield or Jackson traps). These traps are relatively ineffective for detecting medfly. For example, one study from the Adelaide metro area trapping grid found that only 0.02% of flies were recaptured from a release of 38.8 million flies. Further, medfly are known to have low dispersals across space. This means that low-lying populations of flies may go undetected across generations. https://onlinelibrary-wiley-com.virtual.anu.edu.au/doi/pdfdirect/10.1111/j.1570-7458.2006.00415.x

3.3 Data

3.3.1 Zero sighting surveillance data

In this section, I do not use real data to estimate parameters. Instead, I model a hypothetical situation in which we observe $\mathbf{y} = \mathbf{0}_T$ (see above). The situation is as follows: We assume that at least one fly has been detected; eradication measures have since begun and then ceased; and we now proceed with intensified monitoring, while whatever population that may exist is free to grow relatively

unhindered. By intensified monitoring, I mean that **supplementary** monitoring traps have been placed alongside the previously existing grid of **general** monitoring traps. More precisely, it is assumed that **general** exist year round in a 400×400 metre grid (DPIPWE, 2011, p. 50). The **supplementary** surveillance system consists of a set of 16 traps in a circular area, centred at the site of the first fly detection.[2] The goal of the analysis is to infer the probability of eradication for the incipient population, given no flies detected at any point in this period.

The data is an hypothetical survey record. We assume a fairly realistic scenario. We observe the outcomes of a surveillance process. The surveillance process is generated by weekly checks of traps that are deployed uniformly in a given area (more about the trapping arrangement below). It is assumed that no specimens are detected at any point in the survey period. In other words, the sum of all detected counts in the period is zero.

3.3.2 Prior information about capture probability

I focus on the case where the probability of capture can be elicited as a function of distance to an individual, and where a systematic surveillance system is in place (so that chance sightings are not considered).

[2] It is typical to wait until at least 2 flies have been detected near each other for an outbreak to be declared. To illustrate the method in a simplified setting, I suppose that one fly detection is sufficient.

3.4 Model

As is typical of Bayesian models, prior distributions must be specified over each of the parameters. I do not consider uninformative priors, as in practical cases, information will exist about the parameters, and should be used.

I break the model into the following three components:

1. The growth model,
2. Fly and trap locations, and
3. Detections.

3.4.1 The growth model

The third and final assumption was that the growth rate for the pest is roughly constant. This is not the case for Medfly, who reproduce in seasonal fruits. Reproduction rates are highly dependent on temperature and seasonal availability of hosts. Therefore, growth varies systematically to a large degree across the year, in ways that are somewhat well understood. Most likely, the environmental manager should use information about the weather and time of year in setting priors on the growth rate for the species.

3.4.2 Population size

3.4.3 The growth model

There are no a priori assumptions on the population dynamics for the growth model. For example, we might apply stochastic or deterministic models of

logistic, exponential, or linear growth. The only thing that matters is that we specify a joint prior distribution over the population size across time points and locations.

For the present work, I focus on the case where there exists a single incipient population of unknown size.

A natural way to set a prior on the population size at each time point t is to set a prior on the population size at the initial time point, and then assume that the population sizes at other time points are given by some (deterministic or stochastic) function of the population size at $t-1$, and the value of a covariate vector X_t, which includes variables relevant to population growth.

$$N_t = f(N_{t-1}, X_t)$$

I assume that our beliefs about the initial population size N_1 are described by a Poisson with random mean λ. I assume that λ follows an exponential distribution. This distribution for N_1 is chosen as it is a discrete distribution with right skew, and a relatively large amount of mass $f_{N_1}(x)$ at $x = 0$, corresponding to the situation where flies are already eradicated.

As for prior distributions on N_t, for $t \in \{1, ..., T\}$, a growth model is used to structure the prior. Namely, an exponential growth model is assumed, so that $N_t = \text{round}[N_{t-1} \exp\{R_t\}]$, where R_t is the growth rate at time t. The exponential growth model is chosen for its ubiquity in ecological science in general,

and in studies of fruit fly dynamics in particular. Rounding is introduced to give N_t discrete support. The growth parameter R_t is uncertain, and based on temperature.[3]

3.4.4 Locations

3.4.4.1 Fly locations

3.4.4.1.1 Central location I first discuss the option of setting a uniform prior. Setting an uninformative prior is fairly straightforward for this problem. In particular, we might assume that, beyond a certain distance from the outbreak centre (say, 1km) any existing population of Medfly is distinct from the population of interest. Therefore, we might set the prior distribution for the population location to be uniform on the surface of a disk with 1km radius around the outbreak centre.

Despite the fact that an uninformative prior is relatively straightforward to set, it is most likely not advisable in specific applications. It will typically be the case that prior information is available to the decision maker. In particular, fruit flies are heavily dependent on the availability of suitable fruit trees for survival and reproduction. Therefore, someone with local area knowledge will be able to determine the most likely locations for an existing population. Also, the supplementary zone is not chosen arbitrarily. The choice of supplementary zone will typically reflect the beliefs of the decision maker about the location of

[3]Alternatively, we could leave the rounding step out, and interpret N_t as the expected number of flies at each step. I do not consider this possibility in any further depth.

the fly population.

When prior information exists, setting the prior distribution to be uninformative may cause us to underestimate the likelihood of observing captures in the supplementary surveillance zone. The overall effect will be to inflate $\Pr(N_T = 0 \mid \mathbf{y} = \mathbf{0}_T)$.

To update on detection location when the first fly is detected at a trap (say trap k) we can use a trick. The trick is to model the probability of the first detection being at trap k as the probability that a fly is detected at k in one period conditional on exactly one fly total being detected in that period. The benefit of this model is that it does not depend on how many weeks it took to get the first detection (which would require information about how long flies have been around before the first detection). See appendix for more details.

A mathematical trick can be used to derive a prior in some cases. Suppose we have K traps indexed by $k \in \{1, \ldots, L\}$. Suppose also that we have a prior distribution over the population size N, given by $N \sim \text{Poisson}(\lambda)$, with $\lambda \sim \text{Exponential}(1/20)$. Here we assume no change in population size over time. Now, we suppose that each trap k is "competing" to catch the first trap each week. We suppose that the trap at the centre of the grid was the first to catch a fly, and we want to use this information. Define the random variable

$$C_k = \begin{cases} 1 & \text{a fly is caught in trap } k \text{ before any other trap} \\ 0 & \text{otherwise.} \end{cases}$$

Under these assumptions, $L \mid C_k = 1$ is the distribution of L, given that a fly was caught in trap k before any other trap.

Whether or not we can analytically derive the posterior density depends on the probability of capture function $p(x)$. In the case we consider here, the function cannot be integrated, and so I resort to sampling. Under the above assumptions, the posterior resembles the convolution of a normal and a uniform distribution (see figure). See appendix for more details.

3.4.4.1.2 Dispersals I assume that flies in the population are dispersed in space around the central location L. Let $D_{i,t}$ denote the location of fly i at time t, relative to the population centroid L. It is assumed that the population centroid does not change over time (i.e. L is independent of t, and everything else in the model). However, $D_{i,t}$ is independent of $D_{i',t'}$, for any $(i',t') \neq (i,t)$. Thus, our belief is essentially that flies are shuffled around at each time point, so that a fly's location at $t-1$ tells us nothing about its location at t, except through the information both reveal about L. This assumption justifies not tracking individual flies across time – whether a fly lives across time periods, or instead dies and is replaced, are equivalent scenarios under this model.

3.4.4.2 Trap locations

3.4.5 The detection model

The second key assumption is that we can estimate the probability of capturing a randomly selected fly from data. This is difficult in the case of Medfly. For fruit

flies, capture probability is typically estimated from data taken from release-recapture studies. In these studies, the researcher obtains a large collection of sterilised specimens, and releases them at a single point in space. Then, the

These experimental data can be useful when the trapping setup is similar to the setup we want to draw inference about. However, this will often not be the case. For example, studies vary in the number and type of traps used (SEE NOTE). Further, we may wish to infer eradication of pest populations in trapping systems that are highly unlike those in studies. For example, after an outbreak has occurred, and eradication measures have been stopped, it is common to set up supplementary trapping units to intensify monitoring and increase the likelihood of detecting flies, conditional on their presence in the area. (CITATION).

- Source for supplementary trapping: <https://nre.tas.gov.au/Documents/Review_of_IR_fo

3.4.5.1 The surveillance (detection) model

It is assumed that surveillance events occur at regular time intervals $t \in \{1, \cdots, T\}$.

3.4.5.2 The trapping arrangement

- Show plot of the trapping grid.
- Note that the number of traps is assumed fixed as it would be in a real scenario.
- Note that the trap locations would be used in a real analysis.

3.4.5.3 Maths

Finally, I discuss the model for detecting individuals. Conditional on N_t, L, and $D_{i,t}$.

Chapter 4

(Appendix) Full model statement

Population size

Initial no. of flies:	$N_1 \mid \lambda \sim \text{Pois}(\lambda)$, where $\lambda \sim \text{Exponential}(1/20)$	
Rate of increase:	$R_t \sim \text{Normal}(\mu_t, \sigma_t^2)$,	$t \in \{2, \dots, T\}$
No. of flies:	$N_t := \text{round}\{N_{t-1} \exp(R_t)\}$	$t \in \{2, \dots, T\}$

Fly locations

Popn. loc.:	$L^{(U)} \sim \text{Uniform}^2(200, 600)$	
	$L^{(N)} \sim \text{Normal}^2(0, \sigma)$	
	$L := L^{(U)} + L^{(N)}$	
Fly dispersals:	$D_{i,t} \sim \text{Normal}(0, 20)$	$i \in \{1, \dots, N_t\}$, $t \in \{1, \dots, T\}$
Fly locations:	$L_{i,t}^{\text{fly}} := L + D_{i,t}$	$i \in \{1, \dots, N_t\}$, $t \in \{1, \dots, T\}$

Detection model

No. traps:	$K \in \mathbb{N}_+$	
Trap locations:	L_k^{trap}	$k \in \{1, \dots, K\}$
Dist. btw. fly i and trap k at time t:	$\delta_{i,k,t} := \|L_k^{\text{trap}} - L_{i,t}^{\text{fly}}\|$	$i \in \{1, \dots, N_t\}$, $k \in \{1, \dots, K\}$, $t \in \{1, \dots, T\}$
Individ. cap. prob.:	$p_{i,t} = 1 - \prod_{k=1}^{K} (1 - p(\delta_{i,k,t}))$,	$i \in \{1, \dots, N_t\}$,

4.1 Computing the posterior distribution

In this section, I discuss the problem of computing the posterior distribution, given a survey record. Above, I stated that the model could be defined flexibly. Without restrictions on the form of the growth and detection models, the posterior may be analytically intractable. In other words, we will not be able to write out the posterior density or mass as a function of the data and prior distributions. Such situations are common in the Bayesian framework, because of the tendency for the posterior density or mass to depend, implicitly or explicitly, on analytically intractable integrals.

In section 2, the model I outlined, given by Barnes et al., had a known analytic solution. In other words, the posterior probability of eradication could be computed as a relatively simple function of the number of negative surveys recorded (i.e., the data), and prior distributions on the population size at each time point and location.

So far, we have talked about situations when sampling is required for inference. Further problems arise when the model is *agent-based*. In other words, when we include uncertainty about individual-level features in the model. In this case, the detection probability is random, even when the location and population size is known. In other words, the probability of detecting at least one individual is a function of the number of individuals, and also their individual (random) properties. This is a situation in which "the number of things you do not know is one of the things you do not know" (Richardson and Green, 1997).

4.1.1 Analogy with mixed models

4.1.2 Sampling algorithms when the number of unknowns is unknown

As mentioned above, standard MCMC algorithms for Bayesian inference will not work when the number of parameters is random. In this subsection, I discuss strategies for sampling from the posterior when this is the case. Firstly, there exist extensions to classical MCMC algorithms for the case where the number of parameters is random. Green (1995) outlines a method he calls "reversible jump MCMC". This involves adding a step to the Metropolis Hastings algorithm, where . A second approach is to use approximate Bayesian computation (ABC). In this work, I focus on this method, as it has some nice properties for the issues we are concerned with here.

4.1.3 Explanation of ABC

4.1.4 ABC models

ABC methods, as exemplified by Caley's model, are a promising approach. Here, I will give a brief description of the family of sampling algorithms known collectively as Approximate Bayesian Computation (ABC). In its simplest form, ABC is a form of rejection sampling.

- Intro to sampling algorithms
- Origin of ABC

- Motivation for ABC
- Extensions of ABC
- Benefits of ABC
  - Can incorporate ABS

4.1.5 Motivation for ABC

There are two reasons for sampling with ABC. Firstly, the likelihood may be unattractive to work with due to its complexity. Secondly, when our model incorporates uncertainty about the individual members of a population, whose size is itself uncertain, standard sampling techniques do not work.

Two things should be noted before moving on. Firstly, the reader may note that the justification given for ABC is unusual. ABC is relatively recent technique for sampling in cases where traditional sampling techniques fail. Standard techniques for MCMC, such as Metropolis-Hastings and Gibbs sampling, assume that the likelihood function is known and can be easily computed. This may not be the case if, for example, computing the likelihood requires us to integrate out a latent variable, but the likelihood is not integrable with respect to that variable.

Secondly, the reader may note that there exist other methods for sampling from the posterior.

4.1.6 ABS models

Agent based simulation (sometimes referred to as individual based simulation) is an approach to quantitative modelling. A key motivation for using ABS is that

A sampler needs to move between points in different dimensional spaces.

Interestingly, the standard justifications for and against ABC do not apply to the case under consideration. Firstly, the standard justification for ABC is that it allows for inference when the likelihood function is "intractable" - i.e., unknown, uncomputable or otherwise difficult to work with. Standard techniques for MCMC, such as Metropolis-Hastings and Gibbs sampling, assume that the likelihood function is known and can be easily computed.

Secondly, the standard drawback for ABC is that it ensures that we can typically only draw from the posterior approximately. Under standard conditions, we must define a criteria for similarity between simulated and observed data. This is typically done by specifying a summary statistic $S(\mathbb{y})$, and a similarity measure $\rho(S(\mathbb{y}), S(\mathbb{y}'))$ defined over the space spanning our data $\mathbb{y}$. We reject a sample if we observe $\rho(\mathbb{y}_{\text{observed}}, \mathbb{y}_{\text{simulated}}) > \epsilon_0$, where ϵ_0

4.1.7 Online sampling

The method under consideration allows for **online** sampling. To update the process, we simply take our existing posterior draws, and then draw the next time step from those draws. The result is a set of posterior draws of the entire

process up to that point.

4.1.8 Extensions of ABC

4.2 Results

4.2.1 Probability of extinction after 12 weeks

4.3 Discussion

Here I discuss limitations and objections.

4.3.1 Limitations

4.3.1.1 Slow sampling

4.3.2 Objections

4.3.2.1 Bayesian models are too subjective

4.3.2.2 Bayesian models are too sensitive to priors

- Defence in Caley 2015

Chapter 5

Appendix: Proof of ABC procedure

Here, I give a proof that the simple ABC rejection algorithm yields independent draws from the posterior distribution. Recall that the algorithm works by drawing samples of θ from the prior distribution with density $\pi(\theta)$. Then, for each draw of θ, we draw a data vector y_{sim} from the likelihood $l(\theta \mid y_{\text{sim}})$. Finally, we keep the sample if we observe that $y_{\text{sim}} = y_{\text{obs}}$ (where y_{obs} is the data vector we actually observed) and reject it otherwise. Then, the draws that we keep have distribution $f_{\text{ABC}}(\theta) = \pi(\theta) \cdot l(\theta \mid y_{\text{obs}})$, since our draws from the prior and likelihood are independent.

[1]

[1] Credit is due to (this StackExchange post)[https://stats.stackexchange.com/questions/380076/proof-of-approximate-exact-bayesian-computation.].

Bibliography

Elizabeth H Boakes, Tracy M Rout, and Ben Collen. Inferring species extinction: the use of sighting records. *Methods in Ecology and Evolution*, 6(6):678–687, 2015.

Peter Caley, David SL Ramsey, and Simon C Barry. Inferring the distribution and demography of an invasive species from sighting data: the red fox incursion into tasmania. *PLoS One*, 10(1):e0116631, 2015.

Jonathan M Keith and Daniel Spring. Agent-based bayesian approach to monitoring the progress of invasive species eradication programs. *Proceedings of the National Academy of Sciences*, 110(33):13428–13433, 2013.

Marc Kery. Inferring the absence of a species: a case study of snakes. *The Journal of wildlife management*, pages 330–338, 2002.

Brian H McArdle. When are rare species not there?. *Oikos*, 57(2):276–277, 1990.