Bayesian inference of eradication of incipient Tephritid fruit fly populations, with application to Mediterranean fruit fly

Chris Malone

2022-05-08

Contents

Chapter 1

Introduction

1.1 Control and monitoring of Tephritid fruit flies

TODO: Restructure the following paragraphs; focus more on zero-catch criteria

Biological invasions are incursions of incipient populations of plants or animals that are non-native to a given region. In the midst of the ongoing globalisation of trade, the salience of biological invasions is increasing around the world. Biological invasions are salient because they pose severe economic and environmental risks. On the economic side, they threaten "wholesale loss of agricultural, forestry, and fishery resources" (**?**). On the environmental side, they threaten to disrupt native ecological processes – outcompeting potentially endangered

natives, and creating globally homogenous, "cosmopolitan" ecological systems.

Tephritid fruit flies (*Diptera Tephritidae*) are of particularly high economic significance. Species in this genus have the potential to destroy a wide variety of horticultural produce, in extremely large amounts. Females lay eggs in plant tissue (often fruits). Eggs develop into larvae, which then feed on the plant tissue. Potential costs of Tephritid fruit fly incursions include losses to significant crop losses; materials and labour allocated to their eradication or suppression; knock-on effects to agricultural produce as a consequence of pesticide usage (Suckling et al. [2016]). Economic costs of successful fruit fly invasions in Australia have been estimated to be as high as hundreds of millions of dollars (Suckling et al. [2016], Hancock et al. [2000]).

Governments have a responsibility monitor and eradicate potential biological invasions. In particular, the Australian government is legally bound by the Biosecurity act 2015 to ensure a very low level of risk for animal and plant invasions. A core element required for controlling this risk is the development of capacities for monitoring and controlling pest outbreaks. Monitoring biological species is difficult and costly. As such,in the public interest, 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 (Meats and Smallridge [2007]).

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 suspension of pest free area status is economically costly to local producers. On the other hand, if eradication has *not* occured, declaring eradication prematurely could (a) lower the manager's credibility, and therefore value of local produce, and (b) cause larger costs longer term due to local establishment of the pest.

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.

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.

TODO: Discuss ZTC.

1.1.1 Problems this work seeks to address

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.

The present work attempts to address two gaps related to Tephritid fruit flies. First of all, I attempt to develop a model for understanding the evidential properties of survey records in arbitrary trapping grids. Second of all, I attempt to address the issue of evaluating existing PFA reinstatement dates. These are the date- and location-dependent lengths of time for which a pest must not be observed before eradication can be declared.

I seek to develop a method that can (a) evaluate proposed zero-catch criteria from a statistical point of view; (b) help to formulate new zero-catch criteria based on thresholds for the posterior probability of eradication; and (c) flexibly inform decision making, so that decision makers can weigh the cost of eradication probabilistically against the cost of further eradication attempts.

TODO: Write a paragraph about how the primary goal of the model is to infer the eradication of **spot infestations** – i.e. small, incipient infestations.

1.1.2 Outline of this work

TODO: Write outline of the whole thesis

Chapter 2

Literature review

TODO: Write introduction to this chapter

2.1 Zero-catch criteria

TODO: Explain the role of zero-catch criteria in fruit fly trapping. Cite Meats and Clift as evidence of this being the case.

"eradication of spot pest infestations, such as Medfly, is common to regions normally claiming area freedom (fruit fly exclusion zones)" (Meats 2005, p. 1335).

- Zero-catch scenarios.
- Rules of thumb.

2.2 Elaborate models

2.2.1 Elaborate models for inferring pest absence

The problem of inferring the presence or absence of a difficult-to-detect species has a literature of its own.

TODO: Discuss differences between simple and complex models.

- Problems with simple models
- Benefits of elaborate models

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.

There is also a class of simple models that could be useful here. (See Boakes et al. [2015] for a review of simple models.) Unfortunately, these models make strong assumptions about priors which are not defensible in general. For example, most assume a fixed, and either constant or declining population size (Caley et al. [2015], p. 2). This is not reasonable for the study of invasive species, whose

numbers are uncertain, and may be increasing quickly. Many models 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, many models (see, e.g., **?**, **?**, **?**) assume that we know the probability of detecting a specimen drawn randomly from the population. However, in the case of fruit flies, this quantity is difficult to study empirically. This is because the capture probability is highly dependent on the spatial layout and types of monitoring traps. In turn, the layout and type of trap are highly dependent on time and place in which the monitoring is taking place. For example, countries (and states within countries) differ in the types of traps used and the spatial density of traps. As such, while we may be able to learn about the probability of capture **per trap**, it is not possible to learn about the probability that a fly is captured in general.

2.3 Previous studies

The literature on elaborate computational models is fairly narrow. I know of only two attempts to biologically realistic models to infer probability of eradication from the survey record. These papers provide inspiration for the model I develop in the following chapter – although there are also some marked differences, as the reader will see. Here, I will discuss the contributions of each of these papers.

2.3.1 Red fire ants

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).

The model explicitly models the location of each agent (ant nest). Typically, when a Bayesian model is agent based, this means that there is an unknown number of parameters in the model. 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 the dimension of the parameter vector to vary between draws. To get around this problem, a generalised Gibbs sampling algorithm is used. This gets around the problem by adding a step to the Gibbs sampler in which we move between coordinate spaces.

2.3.2 Red fox

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, given a record of fox carcass sightings. They obtained data on fox carcass sightings from two sources, namely hunter kills and

road kills. These separate "observation processes" were modelled separately, so that posterior detection rates were allowed to differ 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 $[0, 1]$). This was because the fox sighting mechanism has not been studied empirically (indeed, it is not clear how it could be studied at all). 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 form of approximate Bayesian computation (ABC) 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.

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. Leveraging such information would represent the first attempt to use prior literature to set informative priors.

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.

No model has been developed for the case where prior information exists on the probability of detecting a specimen as a function of distance between the detection site and the specimen.

2.4.1 Fruit fly literature gaps

Complex models have not been developed to study the problem of inferring fruit fly eradication from. Such models, if developed, could perform a range of useful functions. For example, when used with zero-catch data, such models could be used (a) to evaluate rules of thumb for PFA reinstatement; (b) to rigorously devise new recommendations for PFA reinstatement, that are sensitive to specific features of the outbreak; and (c) to provide the decision maker with precise probabilistic estimates of the probability of eradication given any particular length of time for which flies were not detected. Further, such models, when used with real surveillance datasets, could also be used to learn about crucial properties of fruit fly species, such as their field population growth rates, and trap capture rates.

TODO: Note that this work builds on the work of **?**. The authors examine the likelihood of zero detections for various population sizes. This work expands that work by exploring the use of prior distributions on the location and population size. This allows for the use of prior information, and modelling of intensified surveillance (e.g. generalising to scenarios with variable spatial densities of traps).

2.5 Conclusion

TODO: Write conclusion of this chapter

Chapter 3

Model development

TODO: Write introduction to this chapter TODO: Write outline of this chapter

3.0.1 The scenario

TODO: Write a general description of the zero catch scenario

In the following, I describe a general probabilistic model of an ecological system. The size, location, and number of captures, for a biological population, are explicitly modelled. It is assumed that the data are a number of captures (or, alternatively, sightings or detections), and that we wish to infer posterior distributions for the parameters governing size and/or locations.

3.1 Structure of the model

At the most basic level, I propose to define a joint distribution over (a) the number of flies in a population, (b) the location of each fly, and (c) the number of flies that are detected (caught in traps). Defining notation, at each time point t, we must define a joint distribution over the number of flies N_t, the $N_t \times 2$ matrix of fly locations $\mathbf{L}_t$, and the number of flies detected y_t. (It is assumed that the frequency of time points t is determined by the frequency at which surveillance traps are checked.) I.e., the model is a joint distribution $p(\{N_t, \mathbf{L}_t, y_t\}_{t=1}^T)$. Our goal is, primarily, to infer the posterior distribution of the population size at the last time point N_T, conditional on $\mathbf{y} = \{y_t\}_{t=1}^T = \mathbf{0}_T$. To make this possible, we must specify a joint **prior** distribution over $\{N_t, \mathbf{L}_t\}$, and a likelihood, i.e. a distribution for $\mathbf{y} \mid \{N_t, \mathbf{L}_t\}_{t=1}^T$. Instead of defining a joint prior directly on $\mathbf{N} = \{N_t\}_{t=1}^T$, I define a prior distribution over the parameters that govern the process by which the population grows or decays over time. As such, an assumed stochastic growth model structures the prior over $\mathbf{N}$.

It is assumed that traps have fixed, known locations. The probability of capture for each fly, conditional on their location, is then computed as a function of the distance between that fly and each trap.

3.2 Why make the model spatially explicit?

The value of including a spatial component in the model may be questioned. This is relatively unusual in standard approaches (see, e.g., McArdle [1990]).

However, incorporating the spatial component allows us to leverage a useful source of prior information about trap efficacy. There exists a moderately large literature of **release-recapture** studies for various species of fruit fly (**?**). Release recapture studies give us two kinds of data – one one hand, we get the total proportion of flies recaptured; on the other hand, we also get the mean proportion of flies captured **per trap**, given distance between that trap and the release point. The first kind of data can be useful when the experimental setting is similar to the real-world setting for which we want to perform inference. This will be approximately the case when the types of traps, their number, and the spacing between them, are the same. However, this will often not be the case. For example, studies vary significantly in the types and number of traps used. Further, we may wish to infer eradication of pest populations in trapping systems that are genuinely novel or untested. 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 (sup [2011]).

3.2.1 Population size

As mentioned above, a prior distribution must be set for the population size N_t at each time point $t \in \{1, \dots, T\}$. A natural way to do this is to define a model of the population's change. Explicitly, I recommend to define a prior over N_1 directly. Then, define a prior on N_t the previous values of N, as well as parameters governing the growth. A simple example is the Poisson branching

process with exponential change. This is a model of the form

$$N_t \mid \{N_{t-1}, R_t\} \sim \text{Poisson}(N_{t-1} \exp\{R_t\}),$$

where a continuous prior distribution is defined over the R_t terms, which may have support over the whole real line.[1] This prior is attractive because the exponential growth model is highly popular in biology. Therefore, it should be relatively simple to set informative prior hyperparameters for the latent variable R_t. This might be through expert elicitation, or review of the relevant literature, in which population change may already be expressed in terms of growth rates.[2]

It may be worth noting that we could choose to factor in covariate information for the growth process. For example, for a given pest species, we might understand the growth of populations over time as a function of weather or rainfall. I do not discuss this possibility further in this thesis. It may also be worth noting that there are no a priori assumptions on the population dynamics for the growth model. Thanks to the generality of the sampling algorithm, it is simple to "plug in" elaborate growth models. For example, the models of Lux [2018] or Manoukis and Hoffman [2014] could be used to generate independent random draws from the population of Medfly. In this way, they can structure our prior over the population size. In this case, inference would proceed simple as in the case I outline here. This highlights key benefits of the modelling approach taken here, namely its flexibility and modularity.

[1] Alternatively, the logistic growth model, with an additional "carrying capacity" parameter, could be used.

[2] For example, see Papadopoulos et al. [2002].

3.2.2 Locations of flies

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 disc with (e.g.) 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. Firstly, when an outbreak is suspected, it is typical that information about location is available. On one hand, 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. On the other hand, if an outbreak is known or suspected, then flies must have been detected somewhere. Most likely, the locations of these detections will be known to the analyst. When the fly species has low dispersal distances (as e.g. medfly does) these detection locations are highly informative. Therefore, an informative prior, utilising this information, formally or informally, is recommended.

3.2.3 Fly locations

It is assumed the flies are typically clustered in space. This may be justifiable in practice. For a small, seed population, a population with low density will die

out due to the allee effect. However, when this assumption is not realistic, an alternative prior on locations should be considered.

Let L_c be a bivariate random variable describing the centre of the population. Let $L_{i,t}$ be a bivariate random variable describing the location of fly i at time t. It seems natural to assume that $\mathrm{E}(L_{i,t} \mid L_c) = L_c$, for any (i,t). This model assumes that the centre of the population does not move over time. Further, we can specify that $(L_{i,t} \perp\!\!\!\perp L_{i',t'}) \mid L_c$, for $(i,t) \neq (i',t')$. I.e., conditional on the centre of the population, the fly locations provide no information about each other. This gives the computational advantage that we do not need to track flies locations across time. At each time period, they scatter independently.

The benefit of these assumptions is priors can be set on the parameters governing L_c and $L_{i,t} \mid L_c$ separately. The prior on $L_{i,t}$ describes the distribution of fly dispersals. Information on this quantity, for a given species, will often be available in scientific literature.

3.2.4 Probability of capture (likelihood)

Recall from above that the number of captures (and therefore the data vector) at time point t is written as y_t, for $t \in \{1, ..., T\}$. Recall that it is assumed that the trap locations are each fixed and known with certainty. Then, it is assumed that the probability that fly i is caught in trap k at time t is given by $p_{i,k,t} = p(d_{i,k,t})$ where d is the distance between fly i and trap k at time t, and $p(\cdot)$ gives us the probability of capture as a function of distance. Then, the

probability that a fly is caught at any trap is simply

$$p_{i,t} := 1 - \prod_{k=1}^{K} (1 - p(d_{i,k,t})).$$

The probability of capture function $p(\cdot)$ is based on prior analysis of release-recapture data, already discussed, and may be deterministic or random. For example, we might regress captures on distance, from release-recapture data, and allow coefficients to vary randomly. Then, the posteriors would form the priors for the present model.

An intuitive distribution for y_t is the Poisson-binomial distribution. This is the distribution of successes in independent Bernoulli trials with unequal means.

Note that this model of captures takes for granted the common assumption that there is no interference between traps (see **?**). This assumption is essentially that the probability that a fly is captured at a given trap is not affected by the presence of other traps. The justification is that $p(\cdot)$ is typically estimated to decrease quickly as a function of distance. This is because traps are generally ineffective as attractants. Therefore, it is typically the case that $p(d_{i,k,t})$ is very small for all but at most one trap. Therefore, the effect of discounting the possibility of being caught there is negligible.

NOTE: Delete the next paragraph?

If the researcher has cause to believe that interference may be non-negligible, a simple correction can be applied. Without loss of generality, suppose we are

interested in a single fly in a single trapping period. Let q_k be the probability that that fly is captured at trap k, calculated using the distance function above. Let q_0 be the probability that the fly is not captured at all. The set $\{q_k\}_{k=0}^{K}$ are probabilities of exhaustive and mutually exclusive events. Therefore, we can redefine the probabilities of trap-specific capture (or no capture) as $q'_k = q_k / \sum_{i=0}^{K} q_i$.

TODO: Write conclusion of this chapter

Chapter 4

Case study: Mediterranean fruit fly

TODO: Probabilistic graphical model of the entire model (with a box to demarcate the "internal" components from the components on which a prior distribution is set).

4.1 Introduction

TODO: Rewrite this intro in light of the model proposal chapter

In the previous chapter, I proposed to address a gap in the literature by providing an elaborate model for inferring the eradication of an incipient invasive population of Tephritid fruit flies. 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.[1]

TODO:

- Note that this analysis is primarily illustrative. Performing a more specific analysis requires access to confidential data.
- 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.

TODO: Write one paragraph outlining this chapter

4.2 Medfly (Ceratitis capitata)

4.2.1 Medfly are economically important

Mediterranean fruit fly (*Ceratitis Capitata*) or *medfly* are a particularly salient species of tephritid fruit fly. Medflies have 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.).

[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.

Recently, an incursion of medfly in Adelaide, South Australia, prompted a large scale eradication effort. This comprised in part of hiring 350 special-purpose staff that set over 13,000 additional traps, and collected over 350 tonnes of fruit. The scale of the response to this outbreak indicates the perceived economic significance of this fruit fly species.

4.2.2 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

4.3 Data

In a Bayesian model, various sources of data can be used either to perform inference on parameters (i.e. infer marginal posterior distributions) or to inform prior distributions. For the present model, I propose to to perform inference on simulated data from a hypothetical scenario. However, real data is used to inform the structure of the prior distribution and the likelihood.

4.3.1 Zero-sighting surveillance data

As mentioned above, I do not use real data to estimate parameters. Instead, I model a hypothetical situation. 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. The goal of the analysis is to infer the probability of eradication for the incipient population, given that no flies detected at any point in this period. Therefore, our "data" is a vector of zeroes, with one for each vector.

Thus, the data we wish to learn from is hypothetical, or simulated. The idea is to simulate a relatively 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.

4.3.2 Prior data on capture probability: release recapture studies

In the previous chapter, I briefly discussed release-recapture experiments. These are experiments involving the release of large numbers of flies into networks of standard traps. These studies give us a useful source of information about the

probability of capturing a fly given distance between a fly and a trap.

In cases where Bayesian models have been used, data has not been available on detection rates. For example, Caley and Barry [2014] and Keith and Spring [2013] set uninformative priors on the detection rates, and attempt to learn the detection rates for data. However, because of their global economic importance, tephritid fruit flies are relatively well studied. In particular, a number of fruit fly species have been studied with release recapture experiments. Release recapture experiments involve the release of a large number of (often sterilised) fruit flies. These studies help us to learn both the dispersal patterns and tendencies of various species of fly, but also the effectiveness of various trap types and layouts.

4.4 Model

Now that I have discussed background and the data available, I turn to a detailed discussion of the model for this case. This discussion builds on the previous chapter, where the general, basic model was outlined. Here, I focus on the specific prior distributions and likelihood that are used.

For exposition, I break the model into the following three components: (1) The size of the population (number of individuals); (2) the locations of individuals and traps; and (3) number of individuals caught in traps, conditional on (1) and (2).

The likelihood of a given number of captures is a function of latent variables, namely the number and locations of flies. Under the prior distribution, it is

assumed that data at any given timestep are generated as follows. Firstly, nature draws a number of flies (i.e. a population size) which may or may not be based on the number of flies at the previous time step. Next, nature draws a location for each of the flies. Finally, nature draws a number of detections.

4.4.1 Population size

In this case, N_1 is the first week after the most recent fly detection. I have chosen to give N_1 the prior distribution $N_1 \mid \lambda \sim \text{Poisson}(\lambda)$, where $\lambda \sim \text{Exponential}(0.05)$. 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 (see **?**)

```
g = 1/20
init_pmf = function(n, g) g / (1+g)^(n+1)


x = 0:100


df = data.frame(x = x, y = init_pmf(x, g))


ggplot(df) +
  geom_bar(stat = 'identity', aes(x, y)) +
  xlab(bquote(italic(n))) +
  ylab(bquote('Pr('*italic(N)[1] == italic(n)*')'))
```

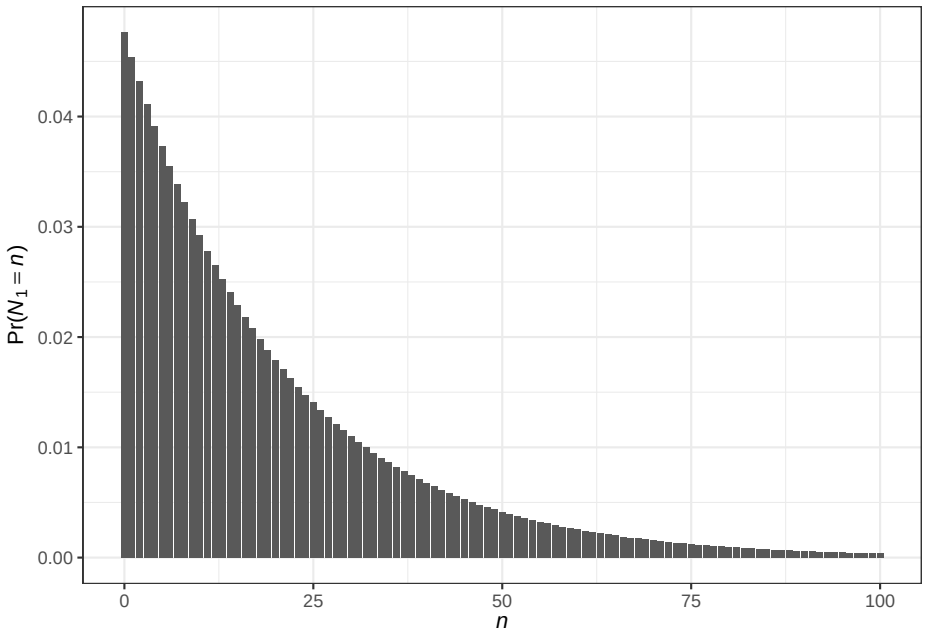

Figure 4.1: (#fig:prior_init_size)Prior distribution of initial population size

```
# Prior probability that they're already eradicated
```

As for the population sizes at later time points, I assume that growth is exponential at an uncertain rate. In particular, it is assumed that, for $t \in \{2, \ldots, T\}$, $N_t \mid N_{t-1}, R_t \sim \text{Poisson}(N_{t-1}e^{R_t})$. Here, we can give any continuous distribution for R_t. I have chosen the vaguely informative prior $R_t \overset{\text{iid}}{\sim} \text{N}(0, 0.2)$. The symmetry of this prior means that we are indifferent about whether the population is growing or declining.

The growth rate prior has been chosen to place the vast majority of density below their estimated growth rate under optimal conditions. Medfly have been estimated to grow at 8% per day, in optimal lab conditions (Papadopoulos et al.

[2002]). Under an exponential growth model this is 56% per week. This can be taken as an extremely *pessimistic* upper bound on the growth rate. In real cases, flies may fail to establish due to food scarcity, predation, and/or unsustainably low population density.

In real cases, it will be desirable to attempt to estimate R_t from data. In particular, it is known that fruit fly population growth rates are highly dependent on temperature. Therefore, it would be wise to estimate and draw values of R_t conditional on weather measurements. Choosing an empirically realistic distribution for R_t is likely to improve the efficiency of inference from the survey record.

4.4.2 Spatial location

As mentioned in the previous chapter, the model is spatial, insofar as likelihood of detection depends on distances between traps and individual flies.

TODO: Explain that the centre of the grid is the site of the last detection.

4.4.2.1 Population location

TODO: Restructure this section so that it is primarily about fly locations.

The centre of of the population is assumed to be located at the two dimensional vector L. I have set the prior on L to be

$$L \sim \text{Normal}_2(\mathbf{0}_2, 160^2 I_2),$$

where Normal_2 is the bivariate normal distribution, $\mathbb{0}_2$ is the two-dimensional zero vector and I_2 is the 2×2 identity matrix. This prior reflects a prior belief that the centre of the population is highly likely to be found somewhere on a disc with. E.g., we believe that there is a 95% chance that the centre of the population is within $1.96 \cdot 160 \approx 320$ metres of the centre of the grid.[2] Recall that this makes sense because we believe that the flies are within the distribution.

4.4.2.2 Individual fly dispersals

I assume that, for any given fly, their prior distance from the central location is described by

$$D_{i,t} \sim \text{Normal}_2(\mathbf{0}_2, 12.5^2 I_2)$$

where the notation is defined as in the previous section above. The variance of this distribution

The normal distribution is chosen for a few reasons. Firstly, it is conceptually simple and intuitive to parameterise. Secondly, the location of fly i at time t, namely $L + D_{i,t}$, has a simple marginal distribution, thanks to the fact that the sum of normal random variables is itself normal. Thirdly, for mean zero normal random variables, distances to the origin have a known distribution. For example, let

$$\mathbf{X} = (X_1, \dots, X_p)^\top \sim \mathbf{N}_p(0, \sigma^2 I),$$

where I is the $p \times p$ identity matrix. Then the length of $\mathbf{X}$, $\|\mathbf{X}\|^{1/2} \sim$

[2] For details about how this prior was arrived at, see appendix.

Gamma(??????) **?** (see appendix). This allows us to compare and calibrate the distribution against experimental results (see appendix). This, in turn, makes elicitation of priors simpler and more intuitive.

It may be worth noting that, in real cases, the assumed prior distribution on dispersals may not be reasonable. For example, dispersals may have non-zero mean (due to strength and direction) or non-spherical covariance matrix (i.e. non-equal variances and/or non-zero covariances).[3] In applied cases, it may be possible to use information about wind directions to set more informative priors. This is beyond the scope of this work.

4.4.3 Trap locations

```
grid_size = 4
general_grid_1d = seq(-grid_size / 2, grid_size / 2) * 400
general_grid = expand.grid(general_grid_1d, general_grid_1d)
general_traps = as.matrix(general_grid)


d = 200 / sqrt(2)
supp_grid_1d = seq(-d, d, length.out = 4)
supp_traps = expand.grid(supp_grid_1d, supp_grid_1d)
all_traps = rbind(general_traps, supp_traps)


supp_indicator = rep(c('General', 'Supplementary'), c(nrow(general_traps), nrow(supp_t
```

[3] This is argued by Baker et al. [1986].

```
traps_df = data.frame(all_traps, factor(supp_indicator, c('Yes', 'No')))


# Plot of traps
ggplot(traps_df) +
  geom_circle(aes(x0=0, y0=0, r=200), colour = 'grey') +
  geom_point(aes(Var1, Var2, col=supp_indicator)) +
  coord_fixed() +
  xlab('X coordinate (metres)') +
  ylab('Y coordinate (metres)') +
  labs(col = 'Trap type') +
  scale_x_continuous(breaks = (-4:4)*200) +
  scale_y_continuous(breaks = (-4:4)*200)
```

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** surveillance traps are placed year round in a 400 $\times$ 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.[4]

[4]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.

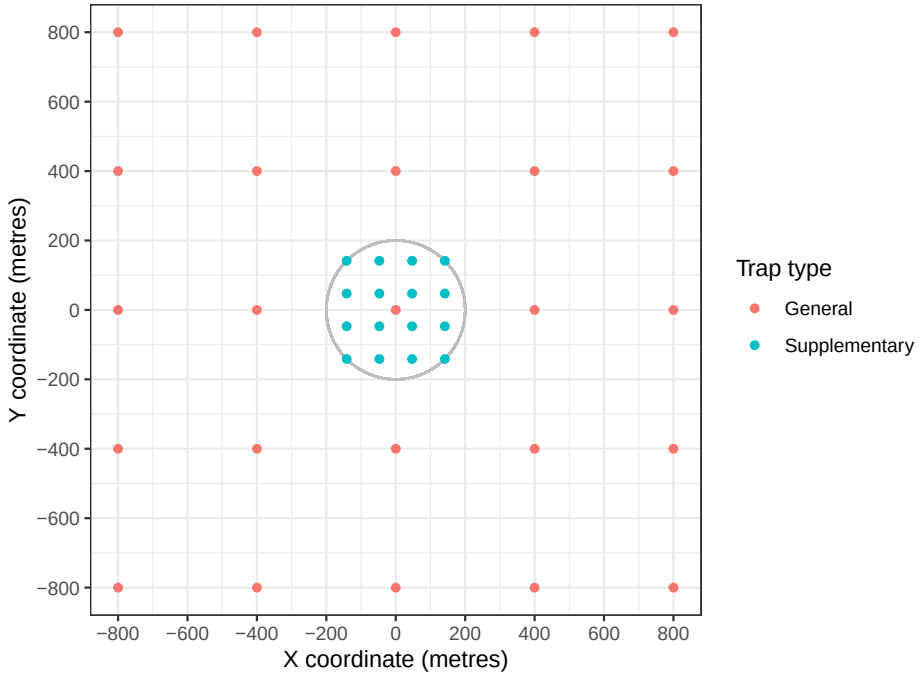

Figure 4.2: (#fig:trap_grid)Illustration of the hypothetical trapping grid. Grey circle represents 200m radius disc surrounding the site of the most most recently detected medfly specimen.

4.4.3.1 Probability of capture

I assume that the probability of capture for a given fly in a given trap is

$$p(x) = \begin{cases} ax^{-b}, & d > 1 \\ a & 0 \geq d \geq 1 \end{cases}$$

where $a = 0.4702111$, $b = 1.37$, and x is the distance between the fly and the trap at the start of the period. Thresholding is introduced because (a) the function does not yield valid probabilities for small enough x. In general, the use of a negative exponential function is not ideal here. This function is used for convenience as it is provided by the authors of the study, and data is not publicly accessible. In the context of a more detailed analysis, it may be beneficial to approximate this function with a more "well behaved" function. This is not within the scope of this thesis.

TODO: mention that the probability of capture is adjusted, because the study gives us the lifetime probability of capture.

The results of release recapture studies are highly variable, and gaining access to data may be difficult for older studies. In this case, it may be desirable to incorporate uncertainty about the probability function. This could be done, for example, by assigning prior distributions to the coefficients a and b. I do not explore this any further here, as this model exposition is intended for the purposes of illustration only.

4.5 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.

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).

4.5.1 ABC models

A simple rejection algorithm is used to draw samples from the posterior distribution.

Here, I have used EBC for intuitiveness, ease of implementation, and the fact

that it is relatively efficient for this problem. However, other methods exist that may be worth exploring, for analogous problems where the rejection rate of ABC is higher. There are at least two known methods for sampling from the posterior when the dimension of the parameter space is uncertain. These are the reversible jump MCMC sample (Green [1995]) and the generalised Gibbs sampler (**?**). These may be worth exploring for problems where the rejection rate is high for EBC.

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. However, for the current model, the likelihood is known, and relatively simple to write out.

On the other hand, 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.

10,000 samples were taken. The acceptance rate for sampling was 0.68. This high level of acceptance is due to the low likelihood of captures, across most of the high prior density region of the model space.

- How long did sampling take?

- What was the rejection rate?

4.6 Results

- The fly free period is 12 weeks or 28 days and one generation, whichever is longer (Meats and Clift [2005]). In summer, a Medfly generation takes 28-34 days (dpi). Therefore, the period I look at is over 12 weeks. However, this may be different based on the period that the manager is interested in.

The posterior probability of extinction after 12 weeks is approximately 0.684.[5]

4.7 Discussion

Here I discuss limitations and objections.

4.7.1 Drawbacks

- We do not get a posterior distribution over the probability of eradication.

TODO: Section on objections to/limitations of the model

- Objection: the model is subjective
- Objection: The model is too sensitive to priors.
  - Defence in Caley 2015

[5] Unfortunately, it is not straightforward to visualise the prior or posterior density of this quantity.

4.7.2 Future work

- Sensitivity analysis for N_1
- Sensitivity analysis for growth rates
- Model checking for the growth process
- Estimating growth process parameters from data.

TODO: Write conclusion of this chapter

Chapter 5

Appendices

5.1 Appendix 1: 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.

5.2 Appendix 2: Full model statement

Population size

Initial no. of flies:	$N_1 \mid \lambda \sim \text{Pois}(\lambda)$, where	
	$\lambda \sim \text{Exponential}(0.05)$	
Number of flies:	$N_t \mid N_{t-1} \sim \text{Poisson}\{N_{t-1}\exp(R)\}$, where	
	$R \sim \text{Normal}(0, 12.5^2)$,	$t \in \{2, \dots, T\}$

Fly locations

Population location:	$L \sim \text{Normal}_2(\mathbf{0}_2, 160^2 I_2)$	
Fly locations:	$L^{\text{fly}}_{i,t} \mid L \sim L + \text{Normal}_2(\mathbf{0}_2, 30^2 I_2)$	$i \in \{1, \dots, N_t\}$, $t \in \{1, \dots, T\}$

Detection model

Number of traps:	$K \in \mathbb{N}_+$	
Trap locations:	$L^{\text{trap}}_k \in \mathbb{R}$	$k \in \{1, \dots, K\}$
Dist. btw. fly i and trap k at time t:	$\delta_{i,k,t} := \lVert L^{\text{trap}}_k - L^{\text{fly}}_{i,t} \rVert$	$i \in \{1, \dots, N_t\}$, $k \in \{1, \dots, K\}$, $t \in \{1, \dots, T\}$
Individ. cap. prob.:	$p_{i,t} = 1 - \prod_{k=1}^{K}(1 - p(\delta_{i,k,t}))$,	$i \in \{1, \dots, N_t\}$, $t \in \{1, \dots, T\}$
	$\mathbf{p}_t := [p_{i,t}]_{i=1}^{N_t}$	$t \in \{1, \dots, T\}$
No. of captures:	$y_t \mid \theta \sim \text{Poisson-binomial}(N_t, \mathbf{p}_t)$,	$t \in \{1, \dots, T\}$
	$\mathbf{y} := [y_t]_{t=1}^{T}$	

5.3 Appendix 3: Population location prior

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 $k \in \{1, \dots, L\}$. Suppose also that we have a prior distribution over the population size N, given by $N \sim \text{Poisson}(\lambda)$, with $\lambda \sim$

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.

Bibliography

Mediterranean fruit fly life cycle and biology. URL https://agric.wa.gov.au/n/911.

Review of import requirements for fruit fly host produce from mainland australia, 2011. URL https://nre.tas.gov.au/Documents/Review_of_IR_for_FruitFly.pdf.

PS Baker, AST Chan, and MA Jimeno Zavala. Dispersal and orientation of sterile ceratitis capitata and anastrepha ludens (tephritidae) in chiapas, mexico. *Journal of applied ecology*, pages 27–38, 1986.

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 and Simon C Barry. Quantifying extinction probabilities from sighting records: inference and uncertainties. *PLoS One*, 9(4):e95857, 2014.

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.

Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

DL Hancock, R Osborne, S Broughton, and P Gleeson. Eradication of bactrocera papayae (diptera: Tephritidae) by male annihilation and protein baiting in queensland, australia. 2000.

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.

Slawomir A Lux. Individual-based modeling approach to assessment of the impacts of landscape complexity and climate on dispersion, detectability and fate of incipient medfly populations. *Frontiers in Physiology*, 8:1121, 2018.

Nicholas C Manoukis and Kevin Hoffman. An agent-based simulation of extirpation of ceratitis capitata applied to invasions in california. *Journal of pest science*, 87(1):39–51, 2014.

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

A Meats and AD Clift. Zero catch criteria for declaring eradication of tephritid fruit flies: the probabilities. *Australian Journal of Experimental Agriculture*, 45(10):1335–1340, 2005.

A Meats and CJ Smallridge. Short-and long-range dispersal of medfly, ceratitis capitata (dipt., tephritidae), and its invasive potential. *Journal of Applied Entomology*, 131(8):518–523, 2007.

NT Papadopoulos, Byron I Katsoyannos, and JR Carey. Demographic parameters of the mediterranean fruit fly (diptera: Tephritidae) reared in apples. *Annals of the Entomological Society of America*, 95(5):564–569, 2002.

David Maxwell Suckling, John M Kean, Lloyd D Stringer, Carlos Cáceres-Barrios, Jorge Hendrichs, Jesus Reyes-Flores, and Bernard C Dominiak. Eradication of tephritid fruit fly pest populations: outcomes and prospects. *Pest management science*, 72(3):456–465, 2016.