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

Chris Malone

2022-05-15

Contents

Chapter 1

Introduction

1.1 Control and monitoring of Tephritid fruit flies

Biological (pest) invasions are incursions of incipient populations of plants or animals that are non-native to a given region.[1] Such invasions can have high economic significance when the invading species is considered a *pest* (i.e., harmful to human concerns). In the midst of the ongoing globalisation of trade, the salience of biological invasions is increasing around the world. Invasions pose severe economic and environmental risks. On the economic side, they may threaten "wholesale loss of agricultural, forestry, and fishery resources" (Mack et al. [2000]). On the environmental side, they may threaten to disrupt native ecological processes – e.g. by threatening the extinction of endangered natives, and creating globally homogenous, "cosmopolitan" ecological systems (Ibid.).

Tephritid fruit flies (*Diptera Tephritidae*) are a genus of insects of particularly high economic significance. Species in this genus have the potential to destroy a wide variety

[1] Note that, for the purpose of this work, the term *population* will be used to mean a group of animals living together in a group. This is to be distinguished from the statistical use of the term.

of horticultural produce in large amounts. Females lay eggs in plant tissue (often fruits) which is then destroyed in the process of their development. 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]). Potential costs of Tephritid fruit fly incursions include crop losses; labour and materials associated with eradication and suppression; and knock-on effects to agricultural produce as a consequence of pesticide usage (Suckling et al. [2016]).

Some locations in Australia and the USA are internationally recognised exclusion zones for some species of fruit fly. In some of these places *spot infestations* of the relevant fly species can be common (Meats and Clift [2005]). Spot infestations are small invasions, usually met with efforts to eradicate. Eradication methods include spraying of pesticide and release of sterile insects that disrupt the flies' breeding cycle. Currently, it is widespread practice that inference of absence for such invasions is informal. In particular, pest absence is declared according to rules of thumb (Ibid.). These rules are in the form of **zero-catch criteria**. The idea is that, if zero flies are detected for some pre-specified length of time (the zero-catch criterion) we may infer with confidence, and therefore declare, that they have been eradicated from the region (see chapter 2).

There are issues with the zero-catch criteria currently used. Firstly, they are only minimally sensitive to variability between situations. For example, outbreaks two locations with the same temperatures will have the same zero-catch criteria, even though they may differ in other significant ways. Secondly, they do not explicitly or transparently take into account the low sensitivity of fruit fly traps, which has been the subject of some study (see chapter 2). Thirdly, they are based on a simplified model of fruit fly biology which does not account for variation in development times (both within and between fly

populations).

The primary goal of the present work is to develop a rigorous statistical framework for inferring pest absence after a suspected outbreak. The hope is that such a framework can be useful (a) to evaluate existing zero-catch criteria, and (b) to develop improved criteria. A secondary goal of the framework is to provide situation-specific estimates of the probability of eradication. These may be useful to the regulator, who wishes to weigh the cost of eradication probabilistically against the cost of a successful biological invasion.

The method I propose involves spatially explicit, agent based modelling of the relevant fruit fly population. This proposal is inspired by similar recent works using BAyesian models to infer probability of pest absence from real data (see chapter 2).

1.2 Outline of this work

I begin this work in chapter 2 by reviewing relevant literature. In particular, I discuss two bodies of literature. The first is on the existing literature on quantitative analysis of zero-catch criteria for tephritid fruit flies. I explain zero-catch criteria in greater detail, then discuss some ways in which they have been analysed and evaluated in academic literature. The second is on existing Bayesian models for inferring of pest absence. Two example cases from the literature are discussed as inspiration for the model developed in the present work. As far as I know, these are the only such models for this problem (inferring extirpation of a pest population) presented to date. In chapter 3, I develop, and justify, a framework for modelling tephritid fruit fly invasions. Model details are discussed at an abstract level. The focus in this chapter is to present a skeleton of the method. I.e., the basic structure of the model, and core quantitities of interest, are discussed. But the

model is not fully specifieed or fit to data. Finally, in chapter 4, I apply the methodology to the specific case of Mediterranean fruit fly, a tephritid species of particularly high economic significance in the USA and Australia. A prior distribution is defined, and inference is performed to evaluate and derive zero-catch criteria for an hypothetical case.

Chapter 2

Literature review

2.1 Introduction

The present work responds to two literatures. Firstly, it responds to the literature on zero-catch criteria for tephritid fruit flies. In particular, I propose a methodology for rigorous, model-based evaluation and revision of zero-catch criteria. Secondly, this work responds to the literature on Bayesian inference of eradication of pest populations in general. Thus far, Bayesian models of pest extirpation have been developed for at least two scenarios. However, tephritid fruit flies raise some unique challenges, which the present work discusses. Further, the question of zero-catch criteria is not broached. In this chapter, I discuss each of these literatures, and how they inform the present work, and the problem this work seeks to address. I begin by discussing fruit fly zero-catch criteria, and how they have been analysed in the literature. I then move on to discuss the Bayesian models that inspire the approach I take in chapter 3 and 4.

2.2 Zero-catch criteria

As mentioned in chapter 1, inference of pest absence is typically performed based on rules of thumb, referred to here as *zero-catch criteria* (following Meats and Clift [2005]). In some fruit fly exclusion zones that are at high risk of invasion, a network of monitoring traps are laid out year round. These traps are checked periodically (typically every 7-14 days). These are minimum periods of time over which a fly is not detected. If we do not detect a fly for the specified (temperature-dependent) number of weeks, then we may infer and declare that the incipient population has been eradicated.

The reader may wonder why declaring eradication is considered important. The primary reason is that pest free area (PFA) status is economically valuable to producers in the relevant region. This is because there may be offshore markets which require that produce is supplied from a PFA for a given pest. Even in lieue of such a requirement, PFA status may allow the supplier to receive a price premium for goods sold. As such, PFA status increases the value of goods sold.

Zero-catch criteria vary according to the species and the international market for the relevant produce (Meats and Clift [2005]). However, it is typical that they are based on the assumed length of a generation. For example, for Mediterranean fruit fly the criterion is 12 weeks, or 4 weeks and 1 generation, whichever is longer. The length of a generation is typically a constant based on the number of "degree days" that pass. Briefly, the degree days associated with a sequence of days of length n is the sum of average temperatures for each day, minus the development threshold for the relevant species. Here, the development threshold is the lowest temperature at which we assume the species is capable of developing.[1] Degree day methods are simple but ignore important aspects of fly biology

[1] It may help to consider an example. Suppose over 3 days we observe minimum temperatures $\min_1 =$

such as humidity, host availability, and differential effects of temperature on different lifestages of fly – e.g. larvae, adult, etc. (Collier and Manoukis [2017]). Importantly, they are considered to be relatively arbitrary (see Meats and Clift [2005] and Collier and Manoukis [2017]).

As is hopefully clear from the discussion so far, zero-catch criteria are relatively informal. They assume, implicitly, that, if we have not detected a population in a certain fixed number of generations, then they must not be present at a sustainable population density. Meats and Clift [2005] provide point estimates of the probability that the zero-catch criteria get it wrong by declaring an outbreak when flies are still present at a sustainable level. However, crucial to the authors' analysis is the assumption that the zero-catch criteria are generally correct. For example, they assume that the implied acceptable rate of fly detection corresponds to the lowest viable population density for the fly species in question. The justification given is that the criteria are formulated on the basis of a significant degree of practical experience. However, it is debatable whether it is possible to learn optimal zero-catch criteria by experience. In particular, in the case of Mediterranean fruit fly, there is a long-standing debate about whether spot-infestations in California are due to repeated eradications, or instead due to repeated flare-ups of one continuous low-lying population (Carey et al. [2017]). As such, it is unclear whether existing zero-catch criteria should be taken at face value, as these authors do. Further, it is not possible to evaluate criteria using the authors method as it is based on the assumption that the zero-catch criteria are correct.

Some attempts have been made to evaluate zero-catch criteria using agent based mod-

$6, \min_2 = 8$, and $\min_3 = 7$, and maximums $\max_1 = 22$, $\max_2 = 18$, $\max_3 = 19$. Assume flies don't develop when the ambient temperature is below 10°C. Then the accumulated number of degree days in this period is $\sum_{i=1}^{3} \left(\frac{\min_i + \max_i}{2} - 10 \right) = 10$.

els (see, e.g. Collier and Manoukis [2017]). These are complex computational models of fly dynamics. Such studies involve simulating the time to extirpation of outbreaks in various locations. The idea is to determine the time taken for the pest population to be eradicted under control measures. However, this method does not use information in survey records to inform this process.

2.3 Bayesian models for inferring absence of a pest population

The problem of inferring the presence or absence of a difficult-to-detect species has a literature of its own. Models date back to at least the 1990s, using relatively simple and general models to infer pest absence (see Boakes et al. [2015] and Caley and Barry [2014] for reviews). Such methods are attractive for their simplicity, generality, and efficiency of computation. 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, for which the population growth rates are uncertain, and indeed may be increasing.

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 dependent on the spatial layout and types of monitoring traps that are deployed. Spatial density and trap type vary between situations. 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.

In contrast, a class of elaborate models for inferring species extinction have recently appeared in the literature on inferring pest extirpation from survey records. In particular, two attempts have recently been made to develop 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 Metropolis-Hastings sampling, 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 the Markov chain moves between coordinate spaces.

2.3.2 Red foxes

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. 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. The authors use the sum of the observations over the observation period as the matching statistic. They kept samples only if the y simultations were an exact match with the data.

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 surveillance system is relatively simply. Traps are placed at fixed, known locations, and checked periodically, at regular time intervals. Further, prior research investigating the efficacy (i.e. sensitivity) of these traps exists (see chapter 3). If used carefully, this information can be leveraged through the model's priors to learn from the zero-sighting record efficiently. Since the authors discussed above used noninformative priors for key parameters, leveraging such information represents the first attempt to use prior literature to set informative priors for such a model.

2.4.1 Fruit fly literature gaps

Complex models have not been developed to study the problem of inferring fruit fly eradication from zero-sighting records. 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.

The present work builds on the work of Lance and Gates [1994]. In that work, the authors first estimate the probability of detection in a single trap, as a function of distance between trap and fly, based on their own experimental data. Then, they use the estimated curve to derive a point estimate of the probability of capturing one or more flies. This is done under the assumption of a uniform prior distribution for the locations of flies. Under these assumptions, the authors examine the likelihood of zero detections for various population sizes. This allows the authors to estimate how large the population needs to be before it is detected with a high degree of certainty. This work generalised that work in two ways. Firstly, a prior distribution is defined on the population size. This allows us to infer the posterior probability of eradication. Secondly, a more realistic prior is set for fly locations. In real cases, if we suspect the presence of a fly population, we will typically have some information about its whereabouts. I discuss ways for using this information to infer the probability of eradication more efficiently in chapter 3.

2.5 Conclusion

Chapter 3

General framework proposal

3.1 Introduction

3.2 The 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 given 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.

3.3 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]). It is typical, instead,

to assume that each fly has identical probability of being detected (captured). 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. In particular, studies vary significantly in the types and number of traps used. For example, Lance and Gates [1994] used Jackson traps spaced 1.6 km apart, as is standard in California. Meanwhile, Meats and Smallridge [2007] used Lynfield traps, spaced 0.4 m apart, as is standard in Adelaide. 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 (DPIPWE [2011]).

3.4 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) whether or not any flies were detected (i.e. caught in traps). Defining notation, at each time point t, we must define a

joint distribution over (a) the number of flies N_t, (b) the $N_t \times 2$ matrix of fly locations $\mathbf{L}_t$, and (c) an indicator variable y_t which is 1 if any fly was caught at time t.[1] I.e., the model is in essence a joint distribution $p(\{N_t, \mathbf{L}_t, y_t\}_{t=1}^T)$. Recall that our goal is to evaluate zero-catch criteria for declaring eradication. Therefore, to evaluate zero-catch criteria, we want to infer $\Pr(N_T = 0 \mid \{y_t\}_{t=1}^T = \mathbf{0}_T)$, where T is the established zero-catch criteria, based on degree day calculations. Call this quantity **the probability of eradication conditional on no detection**. To derive new zero-catch criteria, we want to find the smallest T such that $\Pr(N_T = 0 \mid \{y_t\}_{t=1}^T = \mathbf{0}_T) < 1 - \alpha$, for some threshold α. Here, α represents our risk tolerance. Call this quantity **the time to eradication**.

A Bayesian framework is assumed for inference. As such, to infer the above quantities, 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 $\mathbf{N}$.

3.4.1 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 stochastic model of the population's change. Explicitly, I recommend to define a prior over N_1 directly. Then, for each $t \in \{1, 2, \dots, T\}$, define a prior on N_t, conditional on the previous values $\{N_i\}_{i=1}^{t-1}$, as well as parameters governing the growth. A simple example, which I employ in chapter 4, is the Poisson branching process with exponential growth or decay. This is a model of

[1] It is assumed that the frequency of time points t corresponds to the frequency at which surveillance traps are checked.

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.[2] This model has the attractive property that, when we condition on $\{R_t\}_{t=1}^T$ but marginalise out $\{N_t\}_{t=1}^{T-1}$, N_T has mean $\mathbb{E}(N_1) \exp\left\{\sum_{t=1}^T R_t\right\}$.[3] This prior is attractive because the exponential growth model is popular in biology and ecology. Given this, in many cases, it should be relatively simple to set informative prior hyperparameters for the latent variable R_t. We might do this through expert elicitation, or through review of the relevant literature, in which population change may already be expressed in terms of growth rates.[4]

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. In principle, it should be 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 is outside the scope of this thesis, and I do not explore it further. However, this highlights key benefits of the modelling

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

[3] Note, however that $N_T \mid \{R_t\}_{t=1}^T$ is not Poisson distributed, and has variance strictly larger than its mean.

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

approach taken here, namely its flexibility and modularity.

3.4.2 Fly locations

As mentioned previously, I propose to explicitly model the location of each fly. As a consequence, at each time t, a prior distribution must be set on each of the N_t flies. Accounting for prior beliefs about the location of each fly introduces substantial complexity to the model. However, this can be simplified significantly, as I hope to demonstrate here.

Note that, in typical situations, we will not be interested in inferring the posterior distributions of fly locations. As such, this is a nuisance parameter.

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.

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.

3.4.3 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 $t \in \{1, \ldots, T\}$. 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 $q'_k = q_k / \sum_{i=0}^{K} q_i$.

3.5 Conclusion

TODO: Write conclusion of this chapter

Chapter 4

Case study: Mediterranean fruit fly (*C. Capitata*)

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]

[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.

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.). 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 populatons 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 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 **?**).

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 $t \in \{2, \ldots, T\}$, $N_t \mid N_{t-1}, R_t \sim$

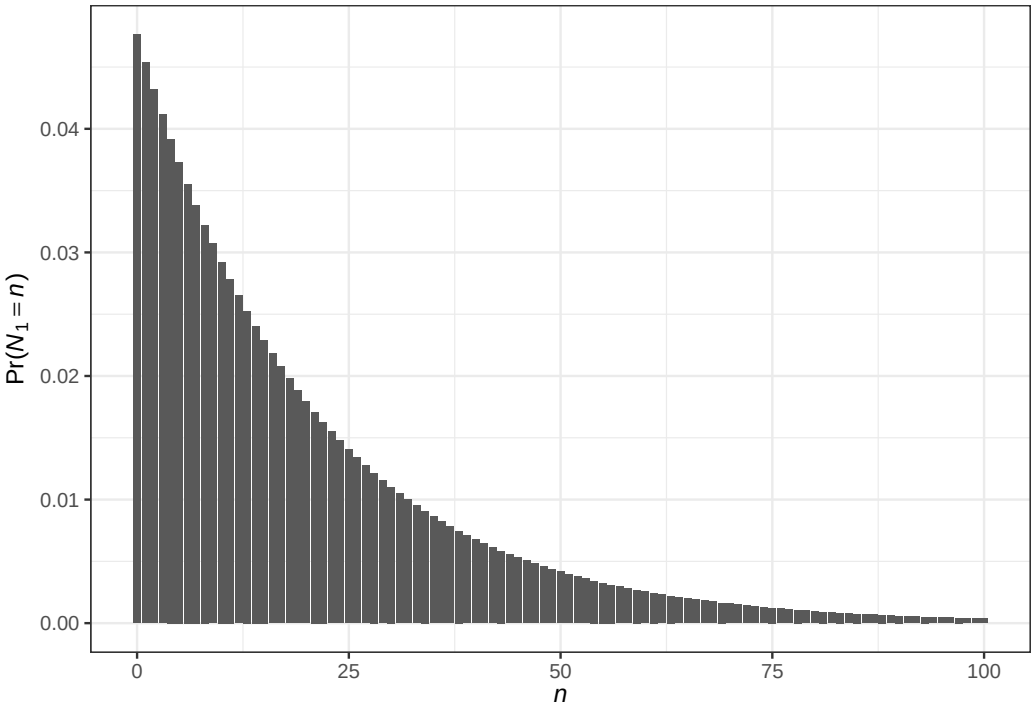

Figure 4.1: Prior distribution of initial population size

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 upper bound on the growth rate. In the wild, flies may fail to establish due to food scarcity, predation, and/or unsustainably low population density.

In practical, applied 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, if possible, fly population growth rates should be estimated conditional on weather. 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

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

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

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

Let $L_{i,t}$ denote the location of fly i at time t, as in the previous chapter. I assign this quantity the prior distribution

$$L_{i,t} \mid L_c \sim \text{Normal}_2(L_c, 12.5^2 I_2)$$

where the notation is defined as in the previous section above. The variance of this distribution is chosen on the basis of prior analysis of dispersal data (see appendix).

Note: Do I need to prove the claims in the next paragraph?

The normal distribution is chosen for a few reasons. Firstly, it is conceptually simple and intuitive to parameterise. Secondly, the location has a simple marginal distribution,

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

thanks to the fact that a normal random variable with a normal mean is itself normal. Thirdly, the distance between a normal random variable and its mean has a known distribution. In particular, for any fly i, the squared distance between $L_{i,t}$ and L_c, conditional on L_c, has the gamma distribution with parameters $\alpha = 1$, and $\beta = 12.5^2/2$. Similarly, the distance between $L_{i,t}$ and the origin is the same but with $\beta = (12.5^2 + 160^2)/2$. Knowing this allows us to easily compare and calibrate the distribution against experimental results. This, in turn, makes elicitation of priors simpler and more intuitive. (It allows us to visualise, for example, the distribution of distances to the origin and to the population centre.)

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] Further, it may not be reasonable to assume that the fly population cannot move across time. Ideally, the importance of these assumptions should be checked based on the case at hand (e.g. properties of the situation or species in question). Extending the model in various ways to meet these problems is beyond the scope of this work, but warrants further investigation.

4.4.3 Trap locations

As mentioned previously, it is assumed that trap locations have fixed, known locations. In particular, we assume that monitoring is intensified for the first 6 weeks ($t \in \{1, \ldots, 6\}$). By intensified monitoring, I mean that **supplementary** monitoring traps have been placed alongside the previously existing grid of **general** monitoring traps. More precisely,

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

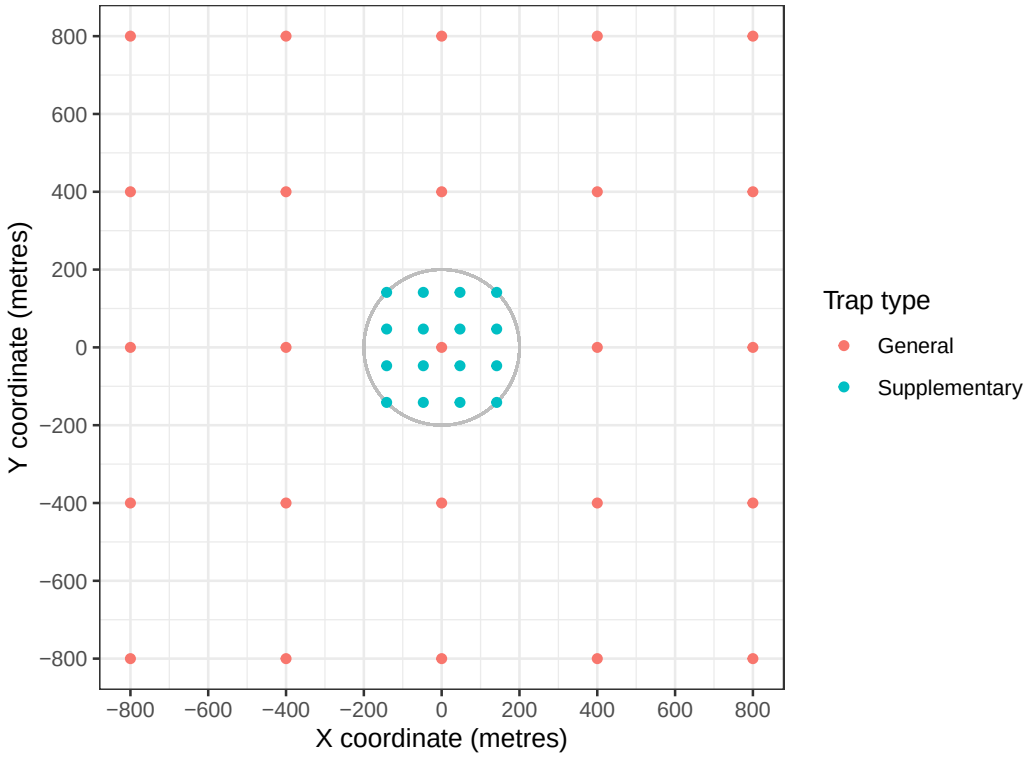

Figure 4.2: Illustration of the hypothetical trapping grid. Grey circle represents 200m radius disc surrounding the site of the most most recently detected medfly specimen.

it is assumed that **general** surveillance traps are placed 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.[4] See **?**.

4.4.3.1 Probability of capture

I assume that the probability of capture for any given fly in any 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

[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.

this model exposition is intended for the purposes of illustration only.

4.5 Sampling procedure

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 Approximate Bayesian Computation

A simple rejection algorithm is used to draw samples from the posterior distribution (see previous chapter). Instead of matching on a statistic, such as the sum of the observations, the rejection rule was set so that data were only kept in cases where the simulated data

vector was an exact match with the actual data (the hypothetical zero-catch record).

Here, I have used ABC 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. The acceptance rate can be interpreted as a numerical approximation to the prior probability of observing zero captures.

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 Limitations

- 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

4.7.2 Future work

- Sensitivity analysis for N_1
- Sensitivity analysis for growth rates
- Model checking for the growth process

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

- Estimating growth process parameters from data.

TODO: Write conclusion of this chapter

Chapter 5

Appendices

TODO: Add appendix on location prior

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

Detection model

Number of traps:	$K \in \mathbb{N}_+$	
Trap locations:	$L_k^{\text{trap}} \in \mathbb{R}$	$k \in \{1, \dots, K\}$
Dist. btw. fly i and trap k at time t:	$\delta_{i,k,t} := \lVert L_k^{\text{trap}} - L_{i,t}^{\text{fly}} \rVert$	$i \in \{1, \dots, N_t\}$, $k \in \{1, \dots, K\}$, $t \in \{1, \dots, T\}$
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 $\lambda \sim \text{Exponential}(1/20)$. 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.

Bibliography

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

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.

James R Carey, Nikolas Papadopoulos, and Richard Plant. The 30-year debate on a multi-billion-dollar threat: Tephritid fruit fly establishment in california. *American Entomologist*, 63(2):100–113, 2017.

Travis Collier and Nicholas Manoukis. Evaluation of predicted medfly (ceratitis capitata) quarantine length in the united states utilizing degree-day and agent-based models. *F1000Research*, 6, 2017.

DPIPWE. 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.

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.

DR Lance and DB Gates. Sensitivity of detection trapping systems for mediterranean fruit flies (diptera: Tephritidae) in southern california. *Journal of Economic Entomology*, 87 (6):1377–1383, 1994.

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.

Richard N Mack, Daniel Simberloff, W Mark Lonsdale, Harry Evans, Michael Clout, and Fakhri A Bazzaz. Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological applications*, 10(3):689–710, 2000.

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.