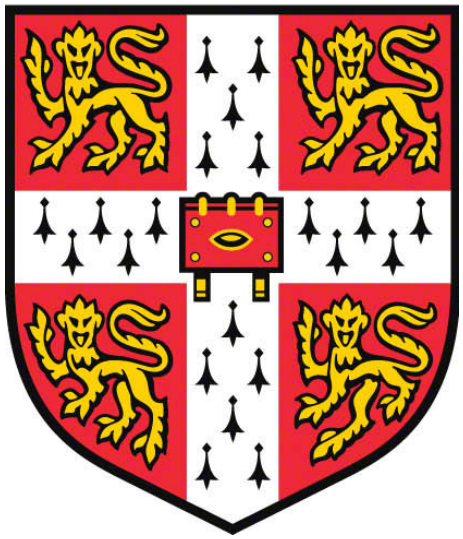
Predicting the Spread and Management of the Cassava Brown Streak Disease Epidemic



David Godding
Jesus College, University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

January 2019

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed 60,000 word limit for the Biology Degree Committee.

David Godding

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Chris Gilligan, for giving me the opportunity to work on such an important project that I could be genuinely passionate about, and for giving me the right balance of guidance and freedom throughout. I look forward to many more years working together.

Endless thanks to Dr Rich Stutt, without whom this thesis would not exist. He has been an inexhaustible source of information, clarity, sanity, and patiently repeated explanations.

Thanks to my second supervisor, Dr Nik Cunniffe, Dr Anna Szyniszewska, Dr Andrew Craig, Dr James Elderfield, and other current and former members of the Epidemiology and Modelling Group and the Theoretical and Computational Epidemiology Group. I would specifically like to thank Dr Matt Castle, who happened to be teaching my statistics course and, without whom, I would never have known about the group or taken the risk of dropping out of the lab I had previously joined.

I must also thank our many collaborators at Rothamsted, NRI, the Gates Foundation and the WAVE and CDP project members, who have been a pleasure and inspiration to work with.

A special thanks to Dr William Summers for his wisdom and guidance over the years. He has been an endless source of, occasionally intentional, amusement. Lastly, I would like to thank my family and countless friends (you know who you are).

Abstract

Cassava Brown Streak disease (CBSD) is a viral disease of cassava that causes necrosis of the edible root tissue, which reduces both consumable and marketable yield. In 2004, CBSD emerged in Uganda and has since been spreading rapidly through previously unaffected regions of East Africa and into Central Africa. Preventing spread to West Africa is a major food security and development priority, along with mitigating the impact of CBSD in endemic regions. This thesis focuses on the development of a landscape-scale spatial model of the CBSD epidemic to inform management.

Currently, there is disparate information on the epidemiology of CBSD and significant associated uncertainty. We begin with a review of CBSD from an epidemiological perspective. The review focuses on: mechanisms and rates of pathogen dispersal, surveillance, disease impact and management efficacy to inform the structure of the CBSD model.

Prior to model development, it was necessary to aggregate all available data on the historic spread of the epidemic. Minimal surveillance data were available in the literature. Therefore, it was necessary to work extensively with East African collaborators to acquire and digitise over 10 years of previously unavailable surveillance records from Uganda and surrounding countries. Extensive post-processing was performed to minimise errors in the data. In parallel with digitisation of the surveillance data, we describe work to enable digital data collection via the creation of a cassava disease surveillance app, along with extensive training. The goal was to minimise errors in data collection and reduce the time lag between disease surveillance and reporting in surveillance programmes.

The second section of the thesis describes the development, parameterisation, and validation of a stochastic, spatio-temporal epidemic model for CBSD. Using digitised Ugandan surveillance data from 2005-2010, and estimates of cassava density

throughout Uganda and immediately surrounding regions, we apply Approximate Bayesian Computation (ABC) to estimate dispersal parameters, providing methodological details on the development and validation of summary statistics. The model fitting also takes account of empirical data for vector density across Uganda and surrounding regions. The model fits the data well for the training set for 2005-2010. Survey data from Uganda and the surrounding region from 2011-2017 are then used as a rigorous independent test to validate model predictions.

The third section of the thesis describes the application of the model to address questions concerning historic, current and future epidemic spread. We use the model to identify reasons why, although there were historically high levels of CBSD infection in Malawi, negligible epidemic spread occurred into Zambia from Malawi showing that low density of cassava cultivation in south east Zambia could account for the inhibition of spread. The model does successfully predict the incursion of the epidemic into north east Zambia from the Democratic Republic of Congo (DRC). We run cross-continent simulations to predict the spatiotemporal spread of the epidemic through central Africa, including DRC and the Central African Republic, where there is very little disease surveillance and reporting for CBSD. The simulations allow us to compute the likely distributions of arrival times of the epidemic in West African countries. We also simulate rates of spread of the disease in West African countries following direct introduction for example by importation and by natural spread from adjoint countries. Finally, we simulate management interventions in Nigeria, to identify the scale and speed at which management programmes would need to be deployed to contain the epidemic.

The thesis concludes with a review of the principal results and critical assumptions underlying the results. Some proposals are presented for future work in epidemiological modelling to address practical problems of the management of CBSD.

Contents

Chapter 1

Introduction

In this thesis, we describe the development, parameterisation, and application of a landscape-scale epidemiological model of cassava brown streak disease (CBSD). Cassava is grown throughout sub-Saharan Africa and is a key food security crop. For this reason, the study of the CBSD epidemic touches on many research areas. In this chapter, we begin with a brief overview of the context of this work, in relation to food security, disease constraints, principle cassava diseases, and modelling approaches. Additional details are given throughout the thesis in the relevant chapters.

1.1 Plant disease epidemics

1.1.1 Significance of epidemics

Agricultural plant disease epidemics vary enormously in their impact. Epidemics in developed countries have an economic effect on producers and consumers. In contrast, financial constraints limit the capacity for developing countries to import food to counterbalance losses (Fry, 2012). Therefore, in addition to an economic impact, epidemics can directly reduce the availability of food in developing countries. This contributes to malnutrition, political instability, and in extreme cases, famine (Timmer, 2000).

In addition, the continuing trend towards monocultures in commercial agriculture is creating an increasingly susceptible environment for large scale pathogen spread. There are two reasons for this change in susceptibility. Firstly, the large number of

plants in a contiguous spatial location makes spread from infectious to susceptible plants easier (Fry, 2012). Secondly, monocultures artificially lack genetic diversity. This means a pathogen that infects a single host is far more likely to be able to infect others, compared with a wild context (Fry, 2012).

Improvements in crop breeding and molecular biology have potential to counteract some of these trends. However, given evolution's remarkable capacity for innovation, it would be a high risk strategy to ignore the fundamental principle of integrated pest management (IPM): that relying on a single management technology is likely to drive the evolution of resistance (Altieri and Rosset, 1999).

It is overly simplistic to assume that the only cause of food insecurity is a lack of food production for the global population (Altieri and Rosset, 1999). The world currently produces enough food per person to theoretically avoid hunger and malnutrition. Clearly, inequality plays a major role (Altieri and Rosset, 1999). However, it is also overly simplistic to assume this can be solved by the planned redistribution of resources. This would be to ignore the current political and economic realities governing national and international trade (Godfray et al., 2010). Therefore, multiple factors should simultaneously be addressed to improve food security.

It is safe to say, in the absence of radical technological or economic changes to food production, continued intensification will be a necessary component of feeding a growing population and addressing rising per capita demands (Godfray et al., 2010; Fry, 2012). This is especially true in the context of emerging pressures on the food system from climate change (Bajzelj et al., 2014). This means an increasing proportion of the population will remain reliant on intensive agriculture, enabling epidemics to have an even greater capacity to cause instability.

It is worth highlighting two historic cases in which plant diseases are known to have played a catastrophic role. Firstly, the Great Famine in Ireland from 1845-1849 caused by the oomycete, *Phytophthora infestans*, is estimated to have resulted in the death of approximately 1 million people (Carefoot and Sprott, 1967). Secondly, low rice harvests in 1942 caused by the fungus, *Cochliobolus miyabeanus*, contributed to the Bengal famine of 1943 in which 2-3 million people died (Padmanabhan, 1973).

Loss of production due to disease is currently estimated to be at least 10% (Strange and Scott, 2005). However, this average does not convey the catastrophic effect of instability in the food system. Often, unpredictable shocks can be more destructive than predictable yield losses (Savary et al., 2012). In addition, the poorest pop-

ulations are disproportionately dependent on agriculture, making them especially vulnerable to instability. However, this also presents an opportunity. Improvements to agriculture and increasing demand will reward producers and is, therefore, likely to be a key route to poverty alleviation and economic development (Collier and Dercon, 2014). Together, these factors highlight the need to promote growth and stability in the food system.

1.1.2 What can be done?

Disease management is an essential component in the broader context of maximising agricultural productivity (Madden, 2006). For plant disease epidemics, control methods are categorised based on their mechanisms of action (Table 1.1. The development of such control methods enables the real-world reduction of pathogen prevalence.

Control method category	Examples
Biological	Parasitoids
Cultural	Crop rotation; roguing
Chemical	Pesticides; fungicides
Resistance	Selective breeding; genetic engineering (e.g. RNAi)
Induced resistance	Cross protection

Table 1.1: Categorised plant disease control methods with examples (Jones, 2004; Fry, 2012; Zhou and Zhou, 2012).

Complementary to the development of control methods is the question of how best to deploy these tools. The epidemiology of infectious plant diseases addresses this concern by improving our understanding of why pathogens spread and how control methods can be optimised to control an epidemic (Madden et al., 2007; Gilligan and van den Bosch, 2008).

Designing real-world disease management programmes is a complex biological, socio-economic and political problem which commonly requires collaboration between researchers, private organisations and policy makers, potentially involving multiple countries (Savary et al., 2012). Moreover, the decision of how to respond to a specific disease is invariably made in the context of a range of competing considerations.

Mathematical models provide a formal framework to structure our understanding of a disease's epidemiology and to interrogate different scenarios regarding the spread and management of pathogens (Gilligan and van den Bosch, 2008). In the next

section, we outline the development of epidemic models from the simplest through to more complex implementations.

1.2 Epidemiological modelling

1.2.1 Compartmental models in epidemiology

In epidemiology, compartmental models are an established and flexible approach to representing infectious disease dynamics in humans, animals and plants. The original framework was pioneered by Kermack and McKendrick (1927), building on earlier developments in the application of mathematics to epidemics.

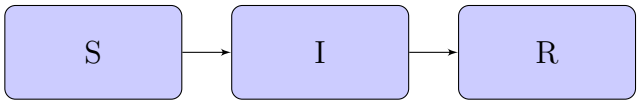


Figure 1.1: SIR model structure representing susceptible, infectious, and removed classes respectively.

The fundamental assumption of compartment models is the categorisation of host units into one of a finite number of mutually exclusive epidemiological states (Keeling and Rohani, 2011). For example, a host may be susceptible or infected; but not both, with the simulation of a compartment model tracking the movement of hosts through these states. The amount of host in each state can be treated as a proportion of the total population in a continuous model, or discrete units for individual-based models. The exact structure of a compartmental model can be informed by the epidemiology of the disease at the spatiotemporal resolution of interest (Keeling and Rohani, 2011).

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI, \\ \frac{dI}{dt} &= \beta SI - \gamma I, \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{1.1}$$

Deterministic compartmental models are commonly represented by coupled differential equations or difference equations. A simple form of compartment model is given by equations 1.1, which describe the basic SIR model (Figure 1.1), where S, I

and R represent the proportion of the population in the susceptible, infectious and recovered compartments. The parameter, γ, is the per capita rate of removal from the infectious class, with $1/\gamma$ being the average infectious period. The parameter, β is the rate of infection between pairs of susceptible and infectious individuals. The transmission term, βSI, assumes there is homogeneous mixing in the population. Despite the simplicity of the SIR model it cannot be solved analytically and, therefore, must be solved numerically. The SIR model can be readily extented to capture more complex epidemic behaviour. For example, the SEIR model introduces an exposed class, which represents an incubation period, during which an individual is infected but not yet infectious.

1.2.2 Spatially-explicit models

In many situations, the assumption of homogeneous mixing is not appropriate. Various extensions to the basic model structure can be made to incorporate heterogeneities in the host or pathogen structure (Keeling and Rohani, 2011).

Metapopulation models are one of the simplest spatial implementations. This involves splitting the population into two or more spatially separate sub-populations and defining how the sub-populations interact. Metapopulation models were originally proposed by Richard Levins around 1970 (Hanski and Gilpin, 1991). Levins-Type metapopulation models do not include dynamics within a sub-population, in contrast to later extensions of the paradigm (Keeling and Rohani, 2011).

A dispersal kernel is a common method for defining the interaction between sub-populations, being the probability of dispersal as a function of distance. Common dispersal kernels are derived from the power law and exponential family of distributions, where $P(d)$ is the probability of dispersing a distance, d. Equations 1.2 and 1.3 give examples for a possible power law and exponential kernel respectively.

$$P(d) = Ad^{-\alpha} \tag{1.2}$$

$$P(d) = Ae^{-\alpha d} \tag{1.3}$$

These kernel structures assume radial symmetry in the dispersal process. Depending on the mechanism of dispersal and environment, real-world dispersal, such as wind-borne processes, may be highly anisotropic. The potential limitations of this

assumption are discussed in detail in Chapter 6.

1.2.3 Modelling plant disease epidemics

The application of mathematical concepts to plant disease systems was pioneered in van der Plank (1963), which applied different growth equations to the development of epidemics. This led to the development of mechanistic models, often described by coupled differential equations (Madden, 2006), such as compartmental models (Section 1.2.1). The compartmental model structure was then extended to incorporate the important effect of host spatial structure. Gilligan (2002) is an early example of a spatial model applied to the sugar beet disease, rhizomania. More recent developments have resulted in the proposal of a unifying theoretical framework to analyse the invasion, persistence and management of pathogens (Gilligan and van den Bosch, 2008).

1.2.4 Stochastic models

Real-world epidemics are not deterministic or continuous processes. They are composed of many discrete epidemiological events that each have certain probabilities of occurrence.

In a stochastic framework, the ordinary differential equation (ODE) system is interpreted in terms of a set of discrete events that change the state of the system in a defined way (Keeling and Rohani, 2011). Table 1.2 illustrates the state changes associated with each of the two possible events in the SIR model. The rates are converted to probabilities, from which a weighted random sample is performed to select which event should occur next, resulting in the simulation of a continuous time discrete-state Markov model (Maroufy et al., 2012).

Event	Rate	State change
Infection	βSI	$S = S - 1, I = I + 1$
Recovery	γI	$I = I - 1, R = R + 1$

Table 1.2: SIR model events and their resultant state changes

1.2.4.1 Simulation methods

For non-trivial models, closed-form analytical solutions are often impossible to derive. In these cases, it is necessary to use simulation methods. Numerical simulation of an ODE model is deterministic and assumes the system is continuous.

The Gillespie algorithm is a continuous time, discrete-state, stochastic simulation framework (Gillespie, 1977). An adapted version of the Gillespie algorithm is used in this thesis. The simulation follows the steps outlined in Algorithm 1, where there are a set of N possible events, indexed by variable, $i, 1 \rightarrow N$, and each event occurs at rate, r_i. The simulation runs from time, $t = 0$ to $t = T$.

The algorithm depends upon two random numbers to govern the random selection of the next event and the size of the time step. The time step is a random sample from an exponential distribution. The random selection of each discrete event is weighted by the probability (rate) at which it occurs.

Algorithm 1 Gillespie's Direct Method (DM) algorithm

1: Initialize values for each compartment: S, I, R, $t = 0$
2: **while** $t < T$ **do**
3: Calculate the rates, r_i, and rate at which any event occurs: $r_{total} = \sum_{i=1}^{N} r_i$
4: Generate two uniformally distributed random numbers: $0 < U_1, U_2 < 1$
5: $p =$ next event selected by weighted random sample with U_1
6: Calculate the time step: $\delta t = \frac{-1}{r_{total}} \log(U_2)$
7: Change system state according to event p, $t = t + \delta t$

1.2.5 Parameter estimation

In cases where the objective of a model is to provide predictions about the real world, it is important that the model structure and parameters realistically describe the system (Downey, 2017). For a given model structure, the process of identifying the most appropriate parameters given the data is known as parameter estimation or model fitting. Bayesian inference provides a framework in which real-world data and prior knowledge are used to allocate probabilities to different model parameter values.

1.2.5.1 Bayes' theorem

We wish to know the probability that a given set of model parameters are true given the observed data, $P(\theta|D)$. However we can only directly calculate the probability of observing a particular set of data given a choice of model parameters, $P(D|\theta)$. Given a set of model parameters, θ and observed data, D, Bayes' theorem allows you to move from a prior belief, $P(\theta)$, about the probability of the model parameters to a posterior belief based on the prior belief and the observed data, $P(\theta|D)$ (Equation 1.5) (Kruschke, 2014).

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{1.4}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} \tag{1.5}$$

In this context, the model provides the likelihood function which takes parameters from the prior as input and returns a probability of seeing the data, given those parameters. If the prior is derived from a specific subset of probability distributions (e.g. Gaussian), the denominator integral can be solved analytically. These are known as conjugate priors. However, for many continuous prior distributions, it is impossible to solve this integral (Kruschke, 2014).

1.2.5.2 Likelihood-free methods

For complex models, the exact likelihood function may be intractable in a practical or actual sense. Additionally, the evaluation of the exact likelihood for multiple parameter sets may be computationally prohibitive. This is the often the case for large epidemiological models (Csilléry et al., 2010). For these reasons, likelihood-free methods have been developed.

Approximate Bayesian computation (ABC) is a highly parallelizable likelihood-free computational method (Beaumont, 2010). In exact likelihood methods, the likelihood function returns the exact probability density value for the given parameters. In ABC, the model is used to generate output data for the given parameters, which are then either accepted or rejected depending on how close they are to the real-world data (Toni et al., 2009).

The ABC rejection algorithm accepts or rejects parameters sampled from the prior according to a threshold distance between data simulated by the model and real-world data (Algorithm 2). One or more summary statistics are commonly used for complex data, and are calculated for the simulation and real-world data (Prangle, 2015). However, the use of summary statistics and a tolerance introduces error into the estimation of the posterior.

Algorithm 2 ABC rejection algorithm (Pritchard et al., 1999)

1: Sample θ^* from $\pi(\theta)$ ▷ Sample parameter from the prior distribution, $\pi(\theta)$
2: Simulate dataset x_{sim} from $f(x|\theta^*)$ ▷ Simulate model with sampled parameter
3: If $d(x_{real}, x_{sim}) < \epsilon$, accept θ^*, else reject
4: Return to 1.

If a summary statistic, S, is used, the distance measure in line 3 in Algorithm 2 becomes:

$$d(S(x_{real}), S(x_{sim})) < \epsilon \tag{1.6}$$

We apply ABC to the estimation of parameters for a CBSD spatial epidemic model in Chapter 4. In this chapter, we provide a detailed account of the specific ABC methodology we apply, with an emphasis on mitigating the potential sources of error.

1.3 Cassava: Importance and threats

1.3.1 Importance of cassava

Cassava is a staple crop for approximately 500 million people in sub-Saharan Africa (SSA). It is the continent's second most important crop in terms of per capita calorie intake (FAO and IFAD, 2005). With the exception of cassava, all major crops' net yields are predicted to be negatively affected by climate change. In contrast, cassava yields are largely expected to remain stable or even benefit from the climatic variations (Jarvis et al., 2012). Moreover, cassava has high levels of drought tolerance, reasonable yields in poor soils and, unlike most other crops, can be left in the ground for up to two years. This long harvest window acts as insurance against

food insecurity in the event of the loss of other crops. For these reasons, cassava is disproportionately grown and eaten by those in poorer subsistence communities.

In addition to cassava's role as a food security crop (Jarvis et al., 2012; Legg et al., 2006), increases in demand from food processing industries can be leveraged to drive economic development for smallholder communities. Therefore, it is essential to protect this long term nutritional and economic potential by identifying current and emerging threats to reliable production.

1.3.2 Threats and constraints to cassava production

For high yielding cassava cultivars, the theoretical yield under optimal conditions is approximately 40t/ha (Ntawuruhunga et al., 2006; Fermont et al., 2007). Changes in agronomic practices, availability of inputs, and choice of cultivar will all play a significant role in increasing yield. However, a number of biotic factors also have a major impact on production where these factors currently exist. Cassava mosaic disease (CMD) and cassava brown streak disease (CBSD) are believed to be two of the most significant threats to production (McCallum et al., 2017).

1.3.3 Cassava mosaic disease

Cassava mosaic disease is endemic to varying extents across all cassava growing regions in Africa (Thresh and Cooter, 2005). The disease is caused by a family of single stranded DNA cassava mosaic geminiviruses (CMGs) (McCallum et al., 2017). Infected plants have reduced tuber volume and distinct leaf patterns and deformation. Quantifying exact yield loss is complex due to the number of contributing factors. Nonetheless, Legg et al. (2006) estimated average yield loss of approximately 47% in affected regions. Cassava mosaic disease came to prominence due to an epidemic involving an especially severe recombinant strain (East African Cassava Mosaic Virus Uganda - EACMV-UG) of the virus in Uganda in the late 1980s (Colvin et al., 2006). Recent studies have revealed the vast genetic diversity of this virus. A total of 11 CMGs have been identified (Legg et al., 2015; McCallum et al., 2017), with each of the eight African strains occupying distinct geographic distributions (Bull et al., 2006). The African cassava whitefly, *Bemisia tabaci* putative species, called Sub-Saharan Africa 1, 2 and 3 (SSA1, SSA2 and SSA3), are the vectors of African CMGs (Tajebe et al., 2014).

1.3.4 Cassava brown streak disease

Currently, two distinct viral species are known to cause CBSD: cassava brown streak virus (CBSV) and Ugandan cassava brown streak virus (UCBSV) (Monger et al., 2001; Winter et al., 2010). These are collectively referred to as CBSVs. There is some evidence for slight differences in the aetiology and diversity of these two viral species, such as symptom severity (Mohammed et al., 2012). However, CBSVs do not appear to have distinct geographic distributions (Ndunguru et al., 2015). A detailed account if the epidemiology of CBSD is given in Chapter 2.

Cassava brown streak disease is believed to have been endemic in coastal East Africa and along the shore of Lake Malawi since the 1930s (Story, 1936). A single outbreak of CBSD was reported in the 1950s in Uganda, but the disease was believed to have been eliminated (Nichols, 1950). CBSD was only subsequently reported once in the following 50 years, at a single site in southern Uganda in 1994 (Thresh et al., 1994). In 2004, CBSD was observed near Kampala, Uganda (Alicai et al., 2007). Since 2004, it has spread widely throughout previously uninfected regions: all regions of Uganda, the Lake Zone of Tanzania, Rwanda, Burundi (Bigirimana et al., 2011), eastern DRC (Mulimbi et al., 2012), and was confirmed in Zambia in 2017 (Mulenga et al., 2018). No single CBSV species seems to be exclusively associated with the post-2004 epidemic (Mbewe et al., 2017). The exact cause of the 2004 Ugandan occurrence and subsequent spread of CBSD beyond the endemic region is unknown. A range of hypotheses have been proposed, such as changes to the vector population, the evolution of a new viral strain, or simply the introduction of diseased planting material from the endemic region (Legg et al., 2011).

The population of Uganda has doubled since 1995 (World Bank, 2017). This means significantly more cassava is being grown, and the frequency of cutting trade is also lik ely to be significantly higher as a result. From an epidemiological perspective, this is significant. The higher the density of cassava and rate of trade, the more susceptible an agroecology is to disease establishment and spread and more host enables proportionally greater vector populations. Moreover, more host theoretically presents an increased fitness incentive to pests, such as *B. tabaci* to specialise to feeding on cassava.

In the absence of any unique factors being confirmed as the cause of the post-2004 spread, our working hypothesis is that CBSD spread is possible in any region where cassava is present. No cultivar has been demonstrated as truly resistant (Tomlinson

et al., 2017). However, there is variability in the severity of symptoms. Therefore, we propose the factors that modulate the rates of spread in a region are: susceptibility of cultivars, vector fecundity and abundance, and frequency and spatial structure of trade.

1.3.5 Mechanisms of CBSD spread

1.3.5.1 Cutting movement

Cassava is a vegetatively propagated crop i.e. planted from stem cuttings. The majority of farmers retain a subset of cuttings after harvest and plant them in the subsequent season. This means that viruses propagate between growing seasons. Less frequently, farmers exchange cuttings locally, or purchase improved varieties from local markets. Therefore, the movement of cuttings introduces the pathogen to new regions. However, assuming farmers do not actively select for diseased material, this does not increase the amount of infected cuttings, it simply distributes it.

1.3.5.2 Transmission by *Bemisia tabaci*

Bemisia tabaci is a complex of multiple morphologically identical whitefly species (Malka et al., 2018). Together, these species are a globally significant crop pest and the vector of hundreds of different plant viruses (Gilbertson et al., 2015). Members of the *B. tabaci* species complex have been shown to vector CBSV (Maruthi et al., 2005, 2017), with the current hypothesis being that *B. tabaci* is the only significant vector.

The abundance of different members of the species complex is believed to vary across sub-Saharan Africa. However, the exact distribution and abundance of different species is not well understood (Macfadyen et al., 2017). Moreover, transmission experiments have not attempted to quantify possible differences in CBSV transmission efficiency across different *B. tabaci* species. It is, therefore, especially important to better understand regional differences in abundance and transmission efficiencies, as if significant differences exist, they would likely have a major impact on the rate at which the CBSD epidemic can spread to new regions.

Unlike cutting movement, vector transmission increases the number of infected

plants. Viruliferous *B. tabaci* spread CBSVs within a field and introduce the virus to other fields locally. Therefore, the rate of viral bulk up in a local region, and local spread is likely to be highly dependant on the the population density of *B. tabaci*. More specifically, the exact rates of transmission will depend on species abundance and their respective transmission efficiencies.

Chapter 2

An Epidemiological Audit of the Cassava Brown Streak Disease Epidemic

2.1 Introduction

Cassava is Africa's second most important staple crop after maize in terms of per capita calorie intake (FAO and IFAD, 2005). Moreover, cassava is disproportionately grown by subsistence, resource-limited farmers. This is due to its relative tolerance to pests and drought, and its capacity to grow well in poor soils without the need for expensive inputs (El-Sharkawy, 2004). Moreover, unlike most crops, cassava has a long harvestable period, meaning it can be left in the ground until required. This enables periods of scarcity to be bridged when other crops have failed or been consumed.

Cassava brown streak disease (CBSD) is a viral disease of cassava. Along with a number of benign symptoms, it causes necrosis of the edible root tissue, which can have a devastating impact on both consumable and marketable yield (Nichols, 1950). Preparing necrotic roots for consumption also demands significantly more labour (Gondwe et al., 2003). The combination of these factors means that CBSD threatens the food security of subsistence farmers, and the income of those who sell cassava or grow it commercially.

Cassava is a vegetatively propagated crop that is planted from stem cuttings. There

are two mechanisms of spread: movement of infected cuttings and vector spread via the whitefly, *Bemisia tabaci* (Legg et al., 2011). The majority of farmers retain a subset of cuttings after harvest to plant in the subsequent season, inadvertently allowing CBSD to propagate from the cuttings that contain cassava brown streak virus (CBSVs). Less frequently, farmers exchange cuttings locally, or purchase improved varieties from local markets, with the risk of introducing the virus, hence the disease, to new regions (Kansiime, 2014; Teeken et al., 2018).

Figure 2.1 provides a compartment model overview of the CBSV pathosystem. Newly infected plants slowly bulk up viral titres over a period of approximately 30 days, and are cryptically infected for approximately 60 days when foliar symptoms tend to appear (Mware et al., 2009b). Vector-borne spread is driven by whitefly, *B tabaci*, which transiently acquire the virus from cryptically or visually infected cassava plants and can disperse CBSVs to uninfected plants within a field or to neighbouring fields (Byrne, 1999; Maruthi et al., 2005, 2017). Mechanisms of reducing the spread of CBSVs include removing cryptically and visually infected cassava plants or lowering the levels of *B. tabaci* (Legg et al., 2014). Cryptic hosts are defined as by being infectious, but not showing symptoms.

Cassava brown streak disease was first identified near coastal Tanzania in the 1930s and was later reported to be endemic throughout coastal East Africa and along the shores of Lake Malawi (Story, 1936; Nichols, 1950). However, in 2004, CBSD emerged in Uganda and has since been spreading rapidly through previously unaffected regions of East Africa and into Central Africa (Alicai et al., 2007; Tomlinson et al., 2017).

In order to control the CBSD epidemic, it is necessary to understand the mechanisms of spread at different scales including vector dynamics and trade contributions. Our understanding of the epidemic depends on surveillance at the disease front, in endemic regions, and measurements of in-field dynamics.

By combining an understanding of the monitoring tools with observations of the rate and extent of spread, we can better understand the dynamics of the epidemic. This understanding will allow us to optimize management techniques for the scale and impact of the epidemic in different regions. Specifically, we address the following key epidemiological questions:

- How fast and how far is CBSD spreading?
- How to quantify CBSD incidence and spread?

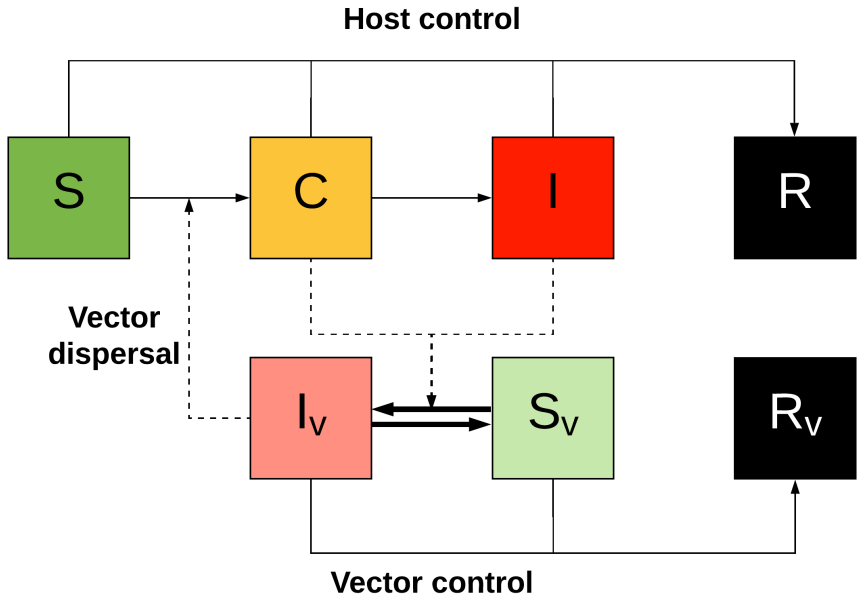


Figure 2.1: A compartment model overview of CBSV infection. Arrows indicate transitions between different states that comprise susceptible plant (S), cryptic plant (C), infected plant (I), removed plant (R), viruliferous vector (I_V), non-viruliferous vector (S_V), removed vector (R_V). Arrows indicate potential transitions between states. Arrows with dashed lines indicate that the source compartments affect the target rates.

- What effect does CBSD have on yield?
- How do we control CBSD?

For each of these epidemiological questions, we currently have imperfect information. The focus of this review is to provide an inventory of what we know from previous studies and highlight areas most in need of further research.

2.2 How fast and how far is CBSD spreading?

2.2.1 Overview of spread

Cassava brown streak disease was historically confined to the lowlands of coastal East Africa and the shores of Lake Malawi (Nichols, 1950; Legg et al., 2011). There were occurrences of the disease beyond this region in Uganda in 1945 (Nichols, 1950). However, a rigorous roguing eradication campaign was launched at the time. This was assumed to be successful, with only one subsequent visual report of CBSD outside the endemic region, at a single site in southern Uganda in 1994 (Thresh et al., 1994). In 2004 infected plants were observed near Kampala, Uganda (Alicai et al., 2007). Since 2004, the epidemic has spread from Uganda through the Lake Zone of Tanzania, western Kenya, Rwanda, Burundi and the Democratic Republic of the Congo (DRC) (Table 2.1). CBSD root symptoms were observed in eastern DRC by 2009 in the absence of foliar symptoms (Legg and Bouwmeester, 2010). Diagnostic confirmation of CBSD in eastern DRC was confirmed in 2011 (Mulimbi et al., 2012). Spread continues towards southern Africa and West Africa, with CBSD being confirmed in northern Zambia in 2017 (Mulenga et al., 2018).

It is currently unclear what initially caused the post-2004 spread of CBSD although it seems likely that higher whitefly numbers in Uganda enabled the rapid initial spread and establishment of the virus (Legg, 1999; Legg et al., 2011). In addition, McQuaid et al. (2017) have shown how high levels of vector-borne transmission can greatly increase the rate of epidemic spread, even with comparatively low levels of trade movement. Moreover, the Ugandan population has quadrupled since around 1960 (World Bank, 2018), which has likely resulted in far higher levels of net cassava production and field density. With more abundant host, the pathosystem would likely support higher vector populations. In addition, higher levels of production and the wider availability of motorised transportation would increase the frequency

Visual confirmation	Country	Molecular confirmation	Citation
2004	Uganda	2004	Alicai et al. (2007)
2006	Kenya	2006	Mware et al. (2009a)
2007	Tanzania	2008	Ntawuruhunga and Legg (2007); Mbanzibwa et al. (2009)
2009	Burundi	2011	Legg and Bouwmeester (2010); Bigirimana et al. (2011)
2009	DRC	2011	Legg and Bouwmeester (2010); Mulimbi et al. (2012)
2009	Rwanda	2014	Munganyinka et al. (2018)
2013	South Sudan	Not confirmed	Phillip Abidrabo, personal communication
2017	Zambia	2017	Mulenga et al. (2018)

Table 2.1: First year in which CBSD was visually observed (foliar or root symptoms) and the year in which presence was confirmed by molecular diagnostics. Data refer to regions beyond the endemic coastal zone and shores of Lake Malawi.

and distance of cutting movement.

There is a limited amount of survey data documenting the spread of CBSD in the DRC, making it difficult to predict when CBSD will reach West Africa by spreading through Central Africa although direct introduction by long range movement of infected cuttings is also possible. Detecting CBSD in the DRC is constrained by a lack of resources, instability and the sheer size of the country. The lack of cheap, practical molecular diagnostics makes categorical confirmation extremely difficult to validate the spread of CBSD to new regions including the occurrence of root necrosis symptoms similar to CBSD reported in Bas Congo in 2003 (Mahungu et al., 2003).

More survey data from Central Africa in general and specifically in central and western DRC, Central African Republic, Republic of Congo, Chad, and Cameroon are crucial in accurately tracking epidemic progress (Table 2.2). For a detailed overview of the historic survey data, refer to Chapter 3. In addition, multiple years of survey data additionally allow for estimation of the rate of spread of the epidemic fronts in different countries (see Chapter 4).

2.2.2 Vector-driven dispersal

Maruthi et al. (2005) experimentally tested multiple candidate insect vectors, but only achieved transmission with *B. tabaci*. In contrast, Mware et al. (2009b), achieved CBSV transmission with two species: *B. tabaci* and *Aleurodicus disperses* (Mware et al., 2009b), observing a maximum transmission rate of 40.7% with *B. tabaci* and a lower rate of 25.9% with *A. disperses*. Given that the rate of transmission with *A. disperses* is not insignificant, this study is worthy of replication. However, *A. disperses* are believed to occur in low abundance on cassava in comparison to *B. tabaci*. Given the absence of additional studies in the literature on spread by *A. disperses*, we now focus exclusively on studies concerning *B. tabaci*.

Cassava brown streak viruses are semi-persistently transmitted by *B. tabaci*. The vector can rapidly acquire the virus in 5-10 minutes of feeding on cassava and can immediately transfer the virus to other plants, which also requires only 5-10 minutes of feeding to be infected (Maruthi et al., 2017). In terms of the duration of viral retention, previous experiments have transfer of vectors from a donor, to a recipient plant, investigating whether the vector remains viruliferous over different intermediate feeding periods (Jeremiah, 2014; Maruthi et al., 2017). Jeremiah (2014) demonstrated that *B. tabaci* can retain the virus for up to 24 hours. Maruthi et al. (2017) does not clearly state a maximum retention time. A retention time of 48 hours is stated in the abstract, but this is contradicted in the results section, where it is stated that none of the experiments achieved successful transmission after a 48 hour intermediate period. Notably, to the authors knowledge, there have been no experiments to characterise viral retention in the absence of intermediate feeding. Given that CBSVs are not persistently transmitted, this experimental setup would be more representative of retention in the case of longer range wind-borne dispersal between fields.

Estimates of transmission efficiencies by *B. tabaci* differ according to the experimental system. Transmission of CBSV with a single *B. tabaci* insect was demonstrated with a success rate of 12.5% (2/16 plants) and with multiple whitefly in clip cages the transmission efficiency was 40.7% (Mware et al., 2009b). In contrast, free cage transmission experiments by Maruthi et al. (2005) did not achieve transmission with 1000 *B. tabaci* in the UK, but did succeed with the same methodology in Tanzania with only 100 *B. tabaci*, successfully infecting 22% of plants (2/9) (Free cage allows autonomous movement - size approx. 45x45x45cm). Maruthi et al. (2017) more recently achieved transmission rates of 60% with 20 whitefly. The high degree vari-

ability amongst transmission experiments is likely the result of practical challenges and categorical differences in the methodology, such as the specific viral strain and cassava cultivar. As these laboratory experiments differ significantly from field conditions, it is hard to confidently draw conclusions about the exact efficiency of *B. tabaci* as a vector in the field.

Bemisia tabaci are capable of spreading CBSVs at different scales: in-field spread observed in field trials and the harder to observe long-range dispersal between fields. Experiments on in-field spread driven by *B. tabaci* suggest that CBSD can spread at least 20m within a season from a point source (Jeremiah, 2014). Field trials in Uganda in 2009 found that *B. tabaci* spread CBSD throughout a 10m square field within a single season, replicated in two locations for multiple cultivars (Katono et al., 2015). However, the spread distance was constrained by the limited field size (Katono et al., 2015). Similar results of whole field within-season infection from clean planting material were reported by Jeremiah (2014) starting from a spreader row and spreading approximately 23m in a season. Lastly, a field experiment observed 17m spread in a season in coastal Tanzania from a spreader plot (Maruthi et al., 2016).

Inevitably, the limitations of a given experimental design lead to the introduction of error and bias. Specifically, in terms of overestimation of spread, all three experiments on in-field spread by Katono et al. (2015); Jeremiah (2014); Maruthi et al. (2016) were performed with high whitefly levels. Therefore, data on the spread rate at a broader range of whitefly abundances, and across a wider range of cultivars, is required to disentangle how whitefly abundance and cultivar influence in-field spread rate (Jeremiah, 2014; Katono et al., 2015; Maruthi et al., 2016). These experiments would then enable improved region-specific predictions of bulk up rates. In contrast, the sampling strategy can lead to underestimation of spread. For example, both Jeremiah (2014) and Maruthi et al. (2016) measured spread in a single direction, so depending on wind direction the spread values could be an underestimate.

In addition to in-field spread, whitefly are also capable of moving between different fields. The vast majority of flights are localized within the canopy of the cassava plants, with occasional exploratory flights beyond. However, once outside the shelter of the canopy, flight is largely wind-governed (Blackmer and Byrne, 1993; Byrne, 1999; Isaacs and Byrne, 1998; Colvin et al., 1998). Either through single long range flights, or multiple shorter flights, dispersal of the vector, *B. tabaci*, has been experimentally demonstrated at 7km in 12 hours (Byrne, 1999). However, Byrne (1999)

conducted experiments with the *B. tabaci* species MEAM1 in North America. Therefore, experiments should be repeated with African cassava whitefly to ensure these dispersal distances are representative of the African context.

Between-field CBSV spread has not been experimentally characterized, but given the dispersal results from Byrne (1999) and retention time of at least 24 hours found by Jeremiah (2014) it seems likely. Between field spread would also explain how the CBSD epidemic has spread approximately 100km/year in the Great Lakes region when in-field spread is limited to approximately 20m/year (Legg et al., 2011; Tomlinson et al., 2017).

2.2.3 Cutting-driven dispersal

There are two main drivers of cutting movement: first, farmers may obtain material from neighbours or local markets and second, organizations including governments, NGOs or large commercial growers may move large amounts of planting material. Due to the informal nature of smallholder agriculture, very little is quantitatively understood about how the frequency and distance of cuttings trade varies throughout the cassava growing regions of sub-Saharan Africa. Therefore, expert knowledge is the dominant source of information on trading behaviour.

The most common source of cuttings is farmer-saved seeds; however, sourcing planting material from neighbours or local markets is also common (Kansiime, 2014; Teeken et al., 2018). The relative importance of different cutting sources varies by region, and it is likely that higher levels of cutting trade occur in regions with better transport infrastructure and greater commercial demand for cassava. The movement of cuttings by larger non-governmental organizations and governments is less common but can cause large changes in the varieties grown and is rarely documented (Kansiime, 2014). As a result, the contribution of large scale cutting movements is currently entirely absent from any quantitative epidemiological analysis of the observed spread of CBSVs.

Future research should focus on characterising the scales for long-range cutting and vector-driven dispersal. For example, gaining a large-scale understanding of the network of cutting movements. As a first approximation, a network could be compiled from records of the local markets and the connections between them in terms of the direction and relative amounts of trade. For large-scale movements of cut-

tings, we recommend that a transparent database be established to house records of cutting movement along with details of the phytosanitation protocols that were used to monitor and alleviate the risk of introducing infection. In addition to a visible record of material movements, it is important that protocols and regulations be standardized and enforced to avoid the unintentional introduction of infected cuttings to previously uninfected regions.

In summary, trade should be viewed as integral to our understanding of CBSD spread and management. Disease surveillance programmes should include the quantification of cutting-mediated disease spread in addition to vector abundance data. This may take the form of a farmer questionnaire component of the survey or digital surveys, such as via SMS.

2.3 How to quantify CBSD incidence and spread?

Surveillance is the process of collecting data about disease occurrence and spread. The objectives for surveillance differ depending upon whether the pathogen is endemic or invading (Gilligan and van den Bosch, 2008; Parnell et al., 2017). In regions where CBSD is endemic the role of surveillance should be to understand the impact at a country level as well as quantifying the within-field prevalence in response to different management practices. In regions that are at risk of introduction of CBSD, optimized surveillance can minimize the lag between pathogen introduction and disease detection and reporting. Minimizing the delay in observation enables management to start sooner, which makes control of the disease much more likely.

2.3.1 Overview of surveillance

Currently, CBSD surveys target fields 3-6 months after planting and are based on visual observations of foliar symptoms and whitefly counts on roughly 30 plants per field (Sseruwagi et al., 2004). Until 2015, all surveyed plants were from the dominant cultivar grown in the field; however, more recently surveys have included an unbiased selection of plants, irrespective of cultivar. A number of countries across sub-Saharan Africa undertake routine surveillance as part of a number of projects (Table 2.2). Due to resource constraints, these programmes can only survey a small number of fields in the country relative to the amount of cassava production.

Survey Project	Country	Start Year	End Year	Survey Frequency
National	Uganda	2004	-	Annually
GLCI	Uganda	2009	2011	Annually
GLCI	Tanzania	2009	2011	Annually
GLCI	Kenya	2009	2011	Annually
GLCI	Burundi	2009	2011	Annually
GLCI	DRC	2009	2011	Annually
GLCI	Rwanda	2010	2011	Annually
CDP	Uganda	2009	2017	Every 2-3 years
CDP	Tanzania	2009	2017	Every 2-3 years
CDP	Kenya	2009	2017	Every 2-3 years
CDP	Rwanda	2009	2017	Every 2-3 years
CDP	Mozambique	2009	2017	Every 2-3 years
CDP	Malawi	2009	2017	Every 2-3 years
CDP	Zambia	2009	2017	Every 2-3 years
WAVE	DRC	2011	-	Every 2-3 years
WAVE	Nigeria	2011	-	Every 2-3 years
WAVE	Benin	2011	-	Every 2-3 years
WAVE	Togo	2011	-	Every 2-3 years
WAVE	Ghana	2011	-	Every 2-3 years
WAVE	Burkina Faso	2011	-	Every 2-3 years
WAVE	Côte d'Ivoire	2011	-	Every 2-3 years

Table 2.2: Overview of CBSD surveillance programmes. The majority of surveillance has been based exclusively on foliar symptoms. The major coordinating projects are the Great Lakes Cassava Initiative (GLCI) and Ugandan national surveys, West African Viral Epidemiology (WAVE), and Cassava Diagnostics Project (CDP).

Given the large number of countries involved in CBSD surveillance, it is important to ensure the surveillance protocols are harmonized both in terms of data collection and surveyor training. Recently, in collaboration with the University of Cambridge and Rothamsted Research, the majority of surveillance teams have started using a digital data recording platform instead of paper survey forms, significantly reducing user error by eliminating the data transcription step, automating input of GPS coordinates, error checking of input data, and standardizing the data collection form. Importantly, electronic data collection simplifies data analysis across multiple countries and time points and significantly reduces the time between data collection and dissemination.

2.3.2 Practical constraints and recommendations

There are trade-offs for all surveillance methods in terms of time, cost, and accuracy. Currently, visual surveillance is by far the most common strategy. However, beyond six months, it is reported that foliar CBSD symptoms are harder to diagnose as the mature lower leaves, which show symptoms most clearly, drop off (Alicai et al., 2007). This six month limitation on visual foliar surveillance causes a systematic under-reporting of CBSD prevalence in areas where whitefly mediated spread dominates, as the infection continues to spread and symptoms continue to develop after the survey (Katono et al., 2015). Additionally, root symptoms do not develop until later in the growing season and correlate poorly with foliar symptoms, limiting how much information current surveys can contribute to predicting yield loss and within-field incidence (Rwegasira and Rey, 2012a; Hillocks et al., 2015). Conducting surveys using molecular diagnostics removes the timing constraints and reduces the number of false negatives from cryptic infection.

Surveying foliar symptoms is the fastest and easiest option; however, the resulting information has lower accuracy and is only moderately useful for predicting yield losses. Root symptoms are the ideal way to quantify yield loss, but the surveys take much longer, limiting the number of fields that can be surveyed, and can miss infected plants or misdiagnose root rot from other causes. With the correct resources and training, molecular diagnostics have the lowest per plant error rate as they can detect cryptic infection and can be carried out on plants of any age. However, the cost for molecular diagnostics is much higher and additional error can be introduced during sample processing steps. Using in-field as opposed to lab based molecular diagnostics can reduce sample processing errors, but increases the false positive rate

due to the high sensitivity required for detecting the relatively low CBSV levels and the risk of cross contamination between samples.

The optimal strategy and combination of surveillance approaches depends on the research question and available resources. If the priority is to detect whether CBSD is present in a region, a high throughput but low accuracy surveillance technique is ideal because CBSD is likely to only be in a small number of fields, so it is important to survey as many fields at the expense of high accuracy in-field prevalence data. Thus measuring foliar symptoms is ideal because of the low overhead, low cost, and high speed. Additionally, to predict the rate of CBSD spread from a region, it is essential also to measure the relative level of whitefly in the region. Currently whitefly population abundance is addressed by counting the total number of *B. tabaci* on the top 5 leaves of all plants surveyed (Sseruwagi et al., 2004). Counting the number of nymphs has previously been used as a less time consuming and less noisy alternative to counting easily disturbed adults (Abisgold and Fishpool, 1990). Another potential extension is to measure the effect of CBSD on yield during the surveys by collecting data on root symptoms. Finally, monitoring of trade and other large movements of planting material by governments and NGOs can also be useful for predicting future spread of CBSD particularly in areas with low whitefly abundance. In regions where CBSD is endemic, more intensely studied sentinel sites are more useful than wide scale surveillance, and will be discussed in more detail in the following section.

2.4 What effect does CBSD have on yield?

CBSD can directly decrease yield by causing root necrosis, as well as through behavioural changes of farmers such as earlier harvesting (Hillocks et al., 2015). There are secondary costs in the form of increased labour to process the roots and decreased nutritional content (Ephraim et al., 2015). Cassava cultivars have varying degrees of susceptibility to CBSD as well as different responses to infection in terms of root damage (Hillocks et al., 2015). Most countries use national incidence surveys to estimate the impact; however, without improved information on cultivar distribution and root symptom susceptibility, the accuracy of estimates for the impact of CBSD on yield are severely limited.

2.4.1 Yield loss in previous studies

It is important to distinguish between studies quantifying losses at the plant level and those assessing impact on a household or region. The majority of published studies have taken the form of larger scale yield loss assessments than plant-level controlled field trials. It is important to note that these studies inevitably encounter many more uncontrolled variables than in field trials. For example, studies that characterise CBSD yield loss via uncontrolled regional assessments are complicated by time of infection, disease incidence, climate, agronomic practices, cultivar, and the burden of other of pests and diseases, such as CMD.

Gondwe et al. (2003), Hillocks et al. (2015) and Ndyetabula et al. (2016) are key examples of studies addressing yield loss at a regional scale. Across the studies there is high variability in the estimated losses due to CBSD. Hillocks et al. (2015) in Uganda, Tanzania, Malawi, and Kenya identified plants with CBSD foliar symptoms and also returned at the end of the season to measure root symptom severity. The lowest impact was a 1.9% reduction in household yield loss in Nandere, Uganda. In contrast, a 13.9% household loss in yield was recorded Natoto, Uganda. Across varieties, yield losses varied from 0 to 79% (Hillocks et al., 2015). In Malawi, Hillocks et al. (2015) observed 3.0% and 11.8% loss in the two villages they surveyed whereas Gondwe et al. (2003) observed significantly higher root losses ranging from 13-26% across villages.

Based on 2009 survey results in the Coastal and Lake regions of Tanzania, incidence of root symptoms ranged from 0-34% in different districts with 0-29% of the roots having symptom severity score greater than three (Ndyetabula et al., 2016). For individual varieties, the incidence of root symptoms ranged from 0-75% and 0-73% of cultivars having too much necrosis to be sold; however, the household level yield loss was not quantified (Ndyetabula et al., 2016).

An important limitation of the above studies is the use of visual symptoms as a proxy for infection status, potentially excluding cryptic plants and tolerant cultivars from the infection estimate. Additionally, the studies were snapshots of the epidemic irrespective of time since disease arrival and whitefly abundance, making it difficult to infer the eventual yield loss when the region becomes endemic and to compare the snapshots of non-endemic regions.

Based on the above regional studies, plant level yield loss ranges from no loss to 75%

loss dependant on cultivar. In contrast, the regional losses due to CBSD are reported to be lower, between 1.9% to 26%, which can be largely attributed to the mixture of cultivars and the disease pressure. Due to this high variability in estimated yield loss and the methods for measuring yield loss, it is hard to draw definitive conclusions about the response of different cultivars and the broader impact of CBSD.

Ephraim et al. (2015) is a notable example of a controlled field trial contrasting the impact of CBSD infection between infected and uninfected plants across important nutritional traits and across four cultivars. The average dry matter content was reported to be 25% higher in diseased plants, being attributed to the accumulation of necrosis. Starch yield loss averaged 40% across the four cultivars. Cultivars were selected across a spectrum from those believed to be highly tolerant through to highly susceptible. However, results showed minor differences across cultivars.

2.4.2 A standardized symptom and impact framework

We propose there is a need to develop an improved experimental structure to study yield loss in order to predict current baseline yield loss and potential yield gains, in conjunction with symptom assessment. Cultivar specific yield loss should be quantified independently of the confounding factors encountered in regional studies. Separately, regional studies could collect data on the diversity and distribution of different cultivars. Results from field trials could then be used to extrapolate to the regional scale. Moreover, standardising the cultivar characterisation protocol would allow easier comparison across studies. The primary inputs of this framework are infection status, infection timing, agroecological condition, and cultivar type. The main dynamic outputs would be viral titre, foliar symptom severity over time. At harvest time, the outputs would be edible yield, marketable yield, and root quality.

Focusing exclusively on the assessment of yield, the protocol should assess healthy roots in comparison to those of infected plants and quantify the following aspects of yield: edible yield, marketable yield, and edible root quality (i.e. carbohydrate / vitamin profile). Standards for what should be counted as marketable yield and edible yield must be established. The yield assessments should then be reported by mass, and summarized as per plant averages relative to the healthy control group.

In order to extrapolate to a country-wide yield loss, there needs to be a quantitative understanding of the abundance, locations, and response of the different cultivars

grown in the country. Existing survey data have relied on expert knowledge to identify different cultivars in a field; however, these data are unlikely to be representative of the true cultivar diversity (Rabbi et al., 2015; Walker et al., 2014; Kosmowski et al., 2016). Collecting new cultivar survey data from throughout a country using microsatellite data to identify cultivars provides more accurate cultivar prevalence data to use for modelling. Connecting this to yield experiments would allow more reliable estimates of loss at a landscape scale, and predict both past and future loss if CBSD spreads. This would allow countries to choose the intensity of CBSD management programmes relative to other food security and economic threats.

2.5 How do we control CBSD?

The optimal management strategy for controlling CBSD depends on the scale of yield loss predicted in a region and the existing infrastructure. For individual farmers, control needs to be affordable, effective and reliable, as low incomes and minimal savings mean farmers are especially risk averse. For regional and country-level policy makers, management strategies need to be optimized based on their resources and the degree of epidemic establishment in the region.

2.5.1 Tools to reduce infection

Fundamentally, CBSD can be controlled by removing infected plants, such as via clean seed programmes, growing resistant varieties, and reducing the number of whitefly (Figure 2.1). At the individual farmer level, better management techniques such as preferential selection and roguing are ways to reduce the number of infected plants, growing improved varieties can potentially reduce the number of whitefly or susceptibility to infection, and early harvesting can minimize yield loss without changing the infection level. Although all of these approaches are potentially effective, the degree of farmer education needed to implement these approaches at scale may be infeasible (Legg et al., 2017; Kumakech, 2013; Chipeta et al., 2016).

At a country level, developing improved cultivars and clean seed systems can make improved varieties available to more farmers, although economic limitations can still prevent uptake by small-holder farmers (Legg et al., 2017; Kawuki et al., 2016). Modelling suggests that clean seed programs are viable when disease pressure and

vector abundance are low, and farmers carry out regular roguing, although predictions indicate that this is not sufficient to eliminate CBSD from a region after it has already established (McQuaid et al., 2016). Breeding or identifying existing cultivars with disease resistance is a strategy that worked well in the past for cassava mosaic disease (Legg et al., 2006); however, to date there are no commercially available CBSD resistant cultivars (Patil et al., 2014; Kawuki et al., 2016). CBSD varieties that are tolerant to the disease and have minimal to no root symptoms can be beneficial for improving yield in infected regions, but do not reduce the spread of disease (Hillocks et al., 2001).

An often overlooked problem in management is the large variation between foliar and root symptoms in different cultivars, which means that control strategies such as roguing and preferential selection can inadvertently select for cultivars with lower foliar symptoms without lowering the infection level (Okul Valentor et al., 2018). A separate strategy is to select cultivars that support lower numbers of whitefly to slow the spread of disease (see fig 2.1) (Rwegasira and Rey, 2012b).

In areas where spread is dominated by cultivar movement (i.e. low whitefly), trade based restrictions and education are likely to be effective (McQuaid et al., 2017). In contrast, in regions with high rates of whitefly-mediated spread, improved varieties, especially those bred for lower whitefly numbers, and tolerant varieties increase in relative importance as potential means of control. Moreover, it is essential to identify and improve incentives at the farmer level to maximize adoption rates by piloting interventions.

2.5.2 Coordinating management

Because there are so many factors that contribute to an individual management strategy being successful, optimizing when and in what combination to use the management techniques is infeasible to do at scale experimentally and should take advantage of modelling (Gilligan and van den Bosch, 2008). Although many interventions have been studied individually, there is a lack of comparative data on the relative and combined impact of different interventions in the same context. To improve the real world accuracy of predictive models, there needs to be data from representative sites in different agroecological settings with locally preferred cultivars.

The scale of management needs to be matched to the scale of the epidemic (Gilligan, 2002). In addition, imperfect adoption should be taken into account to ensure management strategies are robust to likely human behaviour. With access to quantitative data on management strategies, modelling can help address how thoroughly a given management technique needs to be implemented in a region to be effective.

When selecting an optimal strategy it is important to take into account both the extreme constraints that smallholders face and the limited government resources in addition to the yield improvement.

Chapter 3

Improving surveillance in a developing-country context

3.1 Introduction

3.1.1 Why is surveillance important?

In an ongoing epidemic, the number of infected individuals changes over time. Perfect knowledge of an epidemic is impossible as it requires constant monitoring of all individuals at all times. Surveillance programmes attempt to capture the broad trends of an epidemic, such as the geographic distribution and spatially varying intensity (Madden et al., 2007).

Surveillance is a fundamental tool for policy formulation and management. From a policy perspective, funding allocation should be weighted according to a proportional understanding of the present and future risk of a threat in relation to other threats. Without an impression of the current and predicted impact of an epidemic, there is an increased risk of resource misallocation. From a management perspective, many interventions rely on assumptions about which regions are affected or most at risk. The quality and quantity of surveillance data largely dictate our current and future predictions of prevalence and impact (Parnell et al., 2017). Moreover, the analysis of surveillance data from well structured monitoring programmes forms a feedback loop within a management programme, enabling effectiveness of different management interventions to be assessed, and strategy to be updated accordingly. As a

result, surveillance is a key component of integrated pest and disease management programmes (Fry, 2012).

As surveillance cannot give a perfect impression of an emerging epidemic, spatial statistics and epidemiological models are used to both interpolate and extrapolate beyond the available data to infer and predict the present and future epidemic distributions. Extrapolations from data have associated assumptions and methodological uncertainties, which can be tested and reduced as more data are made available.

3.1.2 Why is surveillance challenging?

Management programmes require significant financial investment. The deployment of an intervention, such as the provision of clean seed material, has a direct impact on the recipients and is publicly visible. In contrast, the value and impact of surveillance is only visible to the decision makers. Therefore, there is a political challenge when allocating funding to surveillance in place of direct intervention. This political challenge is reinforced by the multi-year nature of the epidemic. Gaining multiple snapshots of an epidemic frequently requires a long term policy commitment by a government or stakeholder.

Moreover, designing and deploying an epidemiologically-informed surveillance strategy requires qualified staff and significant coordination amongst stakeholders. In the absence of qualified personnel, questions of when and where surveillance should occur are likely to be motivated by resource and practicality constraints, leading to ad hoc decisions that are not necessarily informed by an epidemiological understanding that could improve the strategic deployment of surveillance programmes. For example, by focussing on regions of greatest risk of being infected or regions where the impact of an epidemic may be greatest (Parnell et al., 2017). A poorly designed surveillance protocol can greatly reduce utility of the output data, which then acts as evidence against the value of continued funding for surveillance.

Associated with the issues around implementation are the complex issues of data transparency, timely sharing, and ownership. If surveillance data are intended to influence the structure of management programmes, it is vital that they are made available for analysis in as close to real-time as possible.

However, the incentives of various stakeholders are often misaligned with this goal. At the governmental level, fears around disease spread can harm trade, which can

lead to governments to concealing or delaying the release of information. Similarly, businesses are concerned about public perception. Moreover, surveyors and researchers may be reluctant to share information with the wider community in a timely fashion due to fears of others analysing the data and the organisations not receiving recognition.

3.1.3 Surveillance in developing countries

A lack of financial and human resources, and poor infrastructure means that the majority of the problems previously described are exacerbated in developing countries. As a consequence, there is often minimal information, likely of poor quality, from which to monitor and appraise the risk of an emerging epidemic. In this regard, there are many similarities between diseases affecting agricultural systems in the developing world and neglected tropical diseases in humans (Baker et al., 2010).

When resources are limiting, it is increasingly important to ensure the surveillance system is designed to be as efficient as possible. For a number of reasons, this has not been the case in the context of the Cassava Brown Streak disease epidemic. This chapter addresses three fundamental points:

- What are the causes of the problems during certain previous CBSD surveillance efforts.
- How can reliable information be recovered from the historic surveillance data.
- How should future surveillance programmes be reformed to maximise the volume and utility of information being collected relative to expenditure.

In this chapter, we highlight the extensive work that was required to recover large amounts of historic surveillance data which directly enabled the parametrisation of a landscape scale epidemiological model of CBSD to be developed in Chapter 4. We believe it is important to record the significant additional time and financial investment that must be expended to make research projects such as this successful, so that similar projects can be structured accordingly.

The vast majority of efforts documented in this chapter were highly collaborative, involving the respective in-country surveillance teams and project managers, along with the support of the Epidemiology and Modelling Groups at the University of Cambridge and Rothamsted Research.

3.2 Data challenges in the context of the CBSD epidemic

The resources historically and currently allocated to understanding and managing the CBSD epidemic have been severely limited. This is largely a result of the fact that all the countries that are currently affected or predicted to be affected by the CBSD are classified as low-income economies. Moreover, despite cassava being one of the most widely grown and important food security crops throughout sub-Saharan Africa, it is of comparatively low economic value.

The vast majority of our work to digitise historic data involved the Cassava Diagnostics Project (CDP), led by Dr Joseph Ndunguru and the Ugandan National Survey data, led by Dr Titus Alicai at the National Crops Resources Research Institute, Kampala, Uganda. The CDP project lasted from 2009 to 2015 and cassava disease surveillance in multiple years in this period. The project was based in Dar es Salaam, Tanzania and worked with research teams in Tanzania, Kenya, Mozambique, Rwanda, Uganda, Malawi and Zambia. The Ugandan government, along with a number of NGOs and charities, has been funding yearly cassava disease surveillance in Uganda since the 1990s.

3.2.1 Absence of historic or current surveillance record

At the time of starting this project, there was no centralised record of when or where surveillance had taken place. None of the raw historic surveillance data were readily accessible to the wider public or for academic uses. Therefore, the initial task was to identify what surveillance had taken place and which data were still in existence. The planning and implementation of a digitisation effort was made difficult by not knowing what data existed, who had the authority to share the data, or when results based on the data could be put into the public domain.

It is certain that our digitisation efforts did not capture all surveillance that has historically occurred. During the course of the project, we became aware of a number of further datasets that are referenced in the literature or in project reports. However, beyond passing references in the literature, it is not obvious how these datasets, most of which are believed to be undigitised on hard paper copies, would be accessed or if they still exist. Some indeed are known to have been destroyed.

3.2.2 Problems with the surveillance protocol

When carrying out the surveys and digitising the resultant data, the main source of preventable errors stemmed from the design and implementation of the project. The CDP project management and senior leadership are all highly trained molecular biologists, who specialise in CBSV virology and the biology of the vector, *B. tabaci*. However, the CDP project scope put significant emphasis on disease surveillance. The reason for the emphasis on extensive regional surveillance is not clear, given the background of the project members. It seems likely that this focus was driven by the funders having identified a distinct absence of reliable, up to date, disease surveillance data. However, as none of the candidate in country experts had training in epidemiology, there was an incongruity between the project objectives and the project members specialisations. This greatly increased the probability of systematic errors occurring within the surveillance component of the project.

A lack of epidemiologically trained project members led to little attention being paid to the details of the surveillance protocol that should be used. The surveys appear to have been viewed primarily as a mechanism to collect samples for molecular analysis, rather than to generate a robust regional dataset to understand the regional disease distribution and impact. Within the field, the only guidance was a review article, Sseruwagi et al. (2004), written to address the quantification of CMD and vector counts, which does not give sufficiently precise information to act as a standard operating procedure. Moreover, the protocols in the review did not consider CBSD, which was informally added into the protocol at a later date.

Multiple cassava cultivars are often grown in a single field. Sseruwagi et al. (2004) promotes two possible in-field sampling strategies in terms of cultivar: sample exclusively from the dominant cultivar or sample plants irrespective of cultivar. Based on this review article, the CDP surveillance protocol elected to only sample from the dominant cultivar. This strategy is incorrect with regard to the stated goals of assessing either in-field disease incidence or cultivar specific disease responses. To assess differential cultivar responses, it would be necessary to collect data on the differential responses of cultivars under the same conditions, hence within the same field. In the absence of additional information on disease pressure, it would be challenging or impossible to meaningfully compare cultivar disease responses by sampling a single cultivar per field. The reason for this is that, unlike in an experimental setting, the local disease pressure is an uncontrolled covariate.

3.2.3 Protocol harmonisation and training

As transpired during the training we implemented, there was significant divergence in what was understood to be the protocol between country teams. There appears to have been minimal harmonisation of the protocol amongst country teams prior to surveys commencing. Moreover, any meetings or training that may have taken place likely only involved the country team leaders, whereas the surveys themselves were often undertaken by more junior staff. Therefore, the people implementing the survey protocol were not directly trained, and relied on second hand information from their team leader.

It is important to note that when surveyors do not understand the rationalisation of the protocol design, they are far less likely to be able to minimise errors during data collection. We found that surveyors were largely unaware of the degree to which data, such as spatial coordinates for sampled fields, were as essential as data on disease and vector levels in assessing the regional extent of the epidemic.

3.2.4 Errors due to paper form data collection

The protocol involved the collection of field data on paper forms and the subsequent digitisation of the paper records. In terms of the in-field form used to record data, no single template was used and each country team leader or a technician was responsible for creating their own data entry structure. As a consequence, the paper forms varied significantly amongst countries and across survey years within the same country. Many aspects of the forms varied, such as the layout, the exact data collected, and the units used.

A number of issues occurred at the paper record digitisation stage. In contrast to digital data entry, an additional transcription step significantly increases the potential and opportunities for human error. Additional sources of error are that there is no guarantee that the same person who carried out the survey would digitise the record. Moreover, it was clear that weeks to years often passed between collection and digitisation, or the records were never digitised. In the vast majority of cases, one or more of the following problems occurred:

- The data had not been digitised
- Original paper forms had been lost preventing cross-checking

Error	Example	Mitigation stage
Random notation error	Incorrect year	Digitisation training and post-processing
Systematic notation error	Incorrect and ambiguous coordinate notation	Digitisation training and post-processing
Random digitisation error	Typing error	Post-processing
Systematic digitisation error	Misinterpretation of coordinate system	Post-processing
Type I or type II error due to visual surveillance	False positive CBSD reporting	Analysis

Table 3.1: Errors encountered at different stages of survey data collection and digitisation, along with the stage at which retrospective mitigation could be enacted. Mitigation stages are described in detail in this section and Section 3.4

- The digitisation format was highly variable leading to errors when calculating means
- No connection was retained between paper forms and digital entry
- Data loss due to being summarised prior to digitisation to field-level summaries, rather than retaining plant-level data.
- The digitised data were unpublished, lost or otherwise unavailable

Common practice was to summarise the field-level statistics manually on the paper form prior to data entry, which led to significant amounts of information being lost and errors in calculation. Most importantly, a simple one to one correspondence between a paper form and a digital record was not conserved, which made subsequent verification difficult. Table 3.1 summarises the different stages at which errors were introduced and the stages at which they can be retrospectively mitigated.

3.3 Finding and digitising the historic data

Given the sparsity of historic surveillance in relation to the scale of the epidemic, it was essential to attempt to recover as much information as possible and minimise the prevalence of errors (Table 3.1). As previously stated, it was difficult to identify when and where cassava disease surveillance had taken place and which physical or digital records still existed. Therefore, our approach to digitising the historic data was iterative and responsive as our understanding developed through close interactions and visits to in-country teams.

Cassava brown streak disease has been present in Coastal East Africa and Malawi since at least the 1930s. In theory, a distinction should be made between surveillance that is attempting to assess the current disease distribution and impact in endemic and non-endemic regions. Notably, both Kenya and Tanzania contain endemic and post-2004 epidemic regions. However, this distinction was not reflected in the structure of surveillance programmes in East Africa. Despite our primary interest being data concerning the post-2004 epidemic region, working with the CDP the endemic data is of significant value in terms of providing a baseline for impact estimation and monitoring of management interventions.

It was decided that workshops should be organised with surveillance teams from all CDP member countries. The objective of the workshops was to return to the paper forms whenever possible, regardless of previous efforts to digitise, and provide sufficient training and supervision to minimise the introduction of digitisation errors. Hence, the digitisation process would retain traceability back to the original record for transparency and post-processing. Importantly, a single data digitisation template would be used to unify all records into a single, analysable dataset. The following overarching workflow was set for the workshops:

- For each country team, identify when surveillance had taken place.
- For each survey, identify which physical records were available for digitisation.
- Prior to digitisation, provide training to all teams to minimise errors previously encountered.
- Assign each paper form a unique ID, scan the original record, and carefully supervise the corresponding digitisation by each country team.
- Perform real-time and subsequent error checking and post-processing

Establishing and maintaining trust was of central importance throughout the workshops, and in subsequent surveillance related collaborations. The primary concerns related to worries about data ownership and the risk of data being published without acknowledgement. Where necessary, data sharing agreements were put in place. However, in-person relationship building was the main reason the workshops succeeded.

Figure 3.1: Participants at the data digitisation workshop in Tanzania, 2016.

3.3.1 Workshop 1: Dar es Salaam, Tanzania - Nov 2016

In November 2016, a data digitisation workshop was organised in conjunction with the CDP in Dar es Salaam, Tanzania. The workshop was attended by member countries of the CDP with the exception of Rwanda, who were not able to acquire visas. The only requirement prior to arrival was that all country teams brought all extant paper survey forms to the workshop.

Given often difficult conditions, such as lack of internet and equipment as well as intermittent electricity, we brought all necessary equipment to allow the digitisation to happen offline. Moreover, ensuring everyone used laptops allowed us to work through power outages. We transported a range of equipment from the UK, including unique pre-printed ID labels, a portable scanner, multiple laptops, workshop instruction manuals, and memory sticks for all participants.

We began by organising records according to their given survey year and location. Each team was tasked with uniquely labelling every form which was then scanned and a unique digital record saved. An unanticipated complication was that a single field record was often split across multiple forms. For example, the three main

collected data types, CMD, CBSD and whitefly counts, had often been recorded on separate forms. Therefore, there was not necessarily a one to one correspondence between a single field survey and a single paper form, which had to be reflected in the data input form. Importantly, the existence of multiple forms led to our realisation that the majority of survey teams had not surveyed the same 30 plants for each of the key pieces of data. Therefore, the data could not be readily analysed to identify plant level correlations.

We provided all attendees with a workshop user manual, which detailed exactly how to approach data entry into the standardised template form. We than ran detailed training sessions based upon the manual, explaining the data entry protocol. As previously stated, team members and often team leaders did not have a clear understanding of the reasons for collecting the different data. Therefore, we began by explaining which pieces of data were most important to keep free of errors from an epidemiological perspective, such as the survey year, coordinates, and the significance of the binary distinction between presence and absence of disease, in contrast to gradations of severity. From prior experience, the coordinates were especially prone to errors as three different coordinate systems were commonly used (degrees/minutes/seconds, degrees/decimal minutes, decimal degrees), but the notation for each system was often used interchangeably. Therefore, we provided extensive training on the different coordinate systems, with guidance on how to infer which system was used for a given record. Moreover, the data entry template contained extensive automated checking to minimise the probability of errors.

Throughout the workshop, we became aware of a reluctance amongst the participants to ask if they had a problem or did not understand a specific task. A related issue we encountered was a desire to infer the contents of a missing pieces of data, as opposed to recording as missing. For these reasons, we ensured that at least one of the facilitators was available at all times throughout the workshop to resolve any complications. Table 3.2 outlines the structured approach we implemented for workshop tasks and responsibilities to minimise the introduction of errors.

The workshop lasted four days. At the end of each day, a backup of the current progress was compiled across all teams and uploaded to a shared cloud storage account. Each country's data were analysed by the local and remote facilitators for systematic errors. If any were identified, corrective training was provided the next morning. The remote facilitator allowed additional real-time checking and analysis to take place, including the generation of maps, to perform a preliminary assessment

Data stage	Responsibility
Organising paper forms	Participant
Paper form labelling	Participant and facilitator
Scan form	Facilitator
Add scanned form to database	Facilitator
Data entry	Participant
Transferring digital form	Participant
Verifying digital form	Remote facilitator
Add digital form to database	Facilitator

Table 3.2: Structure of data workshop tasks with associated responsibilities, designed to minimise the introduction of errors.

of survey route.

Given the absence of a sufficiently specific in-field surveillance protocol, we designed a series of questions to identify the exact methodology that was used for each survey team in a given year. Table 3.3 summarises the responses we gathered. A key caveat is that the workshop attendees had not necessarily partaken in all the surveys in question. Therefore, the header of the table highlights where responses are limited to specific years or regions of a country.

A number of key findings emerged from the questionnaire. The answers confirmed the fact that country teams had only been surveying the dominant cultivar within each field and that the vast majority of surveys had not sampled the same plants for all three key pieces of data, CMD, CBSD and whitefly abundance. In general, the answers in Table 3.3 give an overview of the variability at all levels of the protocol, which is useful when designing harmonisation training. Moreover, this information ensured that later analysis would not make false assumptions about what could and could not be inferred from the dataset.

	Kenya (2009, 13, 15)	Malawi (2009, 12, 13)
Actively picked CMD diseased plants	No	No
Actively picked CBSD diseased plants	No	No
In field survey pattern	X	X in small fields, diagonal in large
Only surveyed dominant cultivar	Varied	No
How often different cultivars in field	Frequent	Varies by region
Reported CBSD presence on non-sampled plants	Only collect samples	Only collect samples
How were whiteflies counted	Sum on five leaves	Sum on five leaves
Which leaves selected for whitefly counts	Spiral out from inner	Fully expanded
Surveyed same plants for Whitefly and CMD	Yes	No
Surveyed same plants for CMD and CBSD	No	No
Surveyed same plants for CBSD and Whitefly	No	No
How did they choose the fields to survey	10km interval	10km interval
What time of year for surveys	Not recorded	June-August
How did they estimate field size	Not recorded	Not recorded
How did they identify cultivar name	Not recorded	Ask farmer
When is first planting season	North and West in April-May	Dec-Jan
When is second planting season	East in October	Not recorded

Table 3.3: Summary of questionnaire responses from country teams. The questionnaire designed to identify the exact methodology that was used for each survey team in a given year.

	Mozambique (2013, 15)	Tanzania (2015 – Central)
Actively picked CMD diseased plants	No	No
Actively picked CBSD diseased plants	No	No
In field survey pattern	X or Z	X
Only surveyed dominant cultivar	Yes	Yes
How often different cultivars in field	Frequent	Rarely
Reported CBSD presence on non-sampled plants	No	No
How were whiteflies counted	Sum on five leaves	Sum on five leaves
Which leaves selected for whitefly counts	Fully expanded	Fully expanded
Surveyed same plants for Whitefly and CMD	No	No
Surveyed same plants for CMD and CBSD	Yes	Yes
Surveyed same plants for CBSD and Whitefly	No	No
How did they choose the fields to survey	5-10km interval	15km interval
What time of year for surveys	April-May	June
How did they estimate field size	Estimate	Ask farmer or estimate
How did they identify cultivar name	Ask farmer	Ask farmer or own knowledge
When is first planting season	Nov	Not recorded
When is second planting season	Not recorded	Not recorded

	Tanzania (2013 – Southern)	Tanzania (2015 – Southern)
Actively picked CMD diseased plants	No	No
Actively picked CBSD diseased plants	No	No
In field survey pattern	Diagonal	M / Z / X
Only surveyed dominant cultivar	No	Mainly dominant
How often different cultivars in field	Not recorded	Sometimes
Reported CBSD presence on non-sampled plants	No	No
How were whiteflies counted	Sum on five leaves	Sum on five leaves
Which leaves selected for whitefly counts	Upper	Fully expanded
Surveyed same plants for Whitefly and CMD	Yes	No
Surveyed same plants for CMD and CBSD	No	No
Surveyed same plants for CBSD and Whitefly	No	No
How did they choose the fields to survey	10km interval	20km interval
What time of year for surveys	June	June
How did they estimate field size	Ask farmer or estimate	Estimate
How did they identify cultivar name	Ask farmer	Ask farmer
When is first planting season	Not recorded	Not recorded
When is second planting season	Not recorded	Not recorded

	Uganda (2009-15)	Zambia (2009-15)
Actively picked CMD diseased plants	No	No
Actively picked CBSD diseased plants	No	Not recorded
In field survey pattern	X	X
Only surveyed dominant cultivar	Yes	Yes
How often different cultivars in field	Frequent	Rarely
Reported CBSD presence on non-sampled plants	Yes, post-2013	Not recorded
How were whiteflies counted	Sum on five leaves	Sum on five leaves
Which leaves selected for whitefly counts	Fully expanded	Fully expanded
Surveyed same plants for Whitefly and CMD	Yes	Yes
Surveyed same plants for CMD and CBSD	Yes	Not recorded
Surveyed same plants for CBSD and Whitefly	Yes	Not recorded
How did they choose the fields to surveyed	7-10km interval	10km interval
What time of year for surveys	July, post-2013	March-April
How did they estimate field size	Estimate	Pace out
How did they identify cultivar name	Ask farmer or own knowledge	Ask farmer or own knowledge
When is first planting season	March-May	Dec
When is second planting season	Sep-Nov	Not recorded

(a) (b)

Figure 3.2: Examples of the paper records of the historic Ugandan surveillance data.

3.3.2 Workshop 2: Kigali, Rwanda - Feb 2017

As the Rwandan team could not attend the workshop in Tanzania, a second workshop was organised in Kigali in February, 2017. The same workflow was followed and we encountered similar problems and implemented the same mitigations as at the workshop in Tanzania.

3.3.3 Workshop 3: Kampala, Uganda - Apr 2017

The Ugandan team had only brought data from 2015 to the workshop in Tanzania. In addition, it transpired through discussions that a significant number of historic surveillance paper forms had been found, dating back to CMD surveillance in the 1990s. An additional data sharing agreement was signed as these data were generated from a number of projects prior to the CDP.

In total, 4223 Ugandan paper forms spanning the period of 2004-2014 were scanned, the majority in the Ugandan workshop. Given this large volume of data, the task of manually typing the remaining Ugandan forms into the standardised template was subcontracted to a company that specialise in data digitisation.

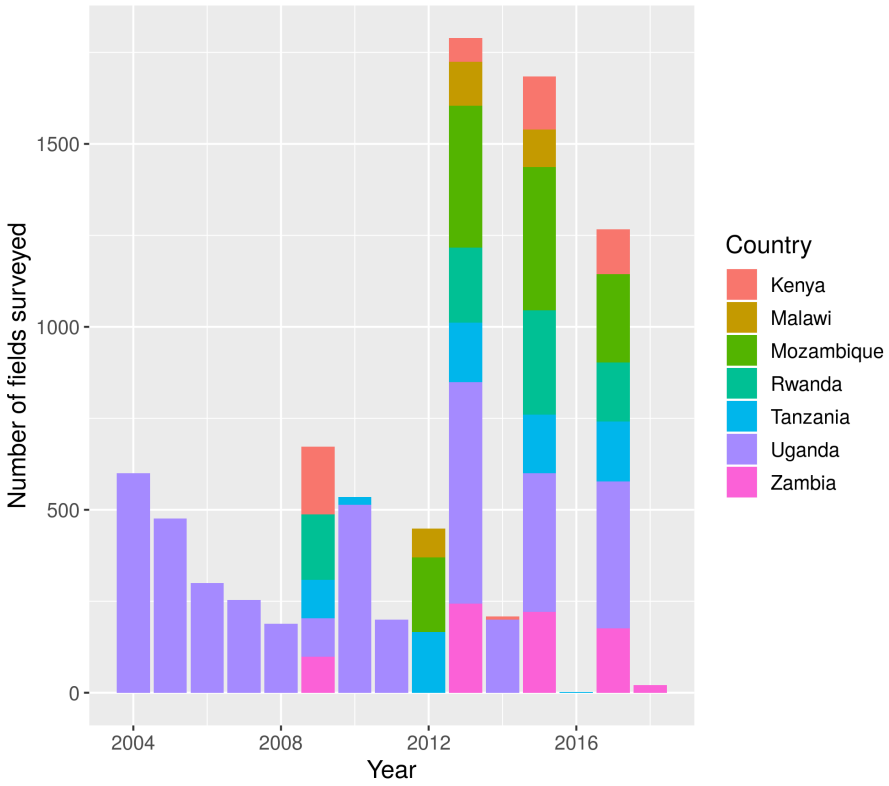


Figure 3.3: Summary of the number of field survey records from CDP member countries that were digitised, by country team each year. Data from 2017 collected digitally using the survey app.

3.3.4 Outputs of workshops and training

A total of 8645 paper field survey forms were digitised across all the workshops, which amounts to approximately 260,000 individual plant survey records. Approximately half of these records are from Uganda. Figure 3.3 summarises the number of fields surveyed by each country team each year.

Having returned to the paper forms and provided training for a standardised digitisation protocol, the resultant dataset then underwent post-processing to identify and, where possible, fix systematic and random errors. The following section gives an account of the post-processing process.

3.4 Processing raw surveillance data

In this section, we focus on the process of cleaning the Ugandan surveillance data. However, many aspects of the process were also applied to the CDP data. Given the large amount of data, we prioritised the cleaning and verification of the most vital pieces of data: coordinates, survey date, CBSD field level presence or absence, and whitefly counts.

For the raw historic surveillance data, we digitised a total of 3639 survey forms. As these forms were digitised externally, the digitisation protocol specified that all data elements were literally transcribed from the paper form contents into a template, without any manual alterations. This minimised the probability of errors if transcribers had attempted to convert the various coordinate systems used manually into a standardised decimal degree format. Consequently, the vast majority of records were not usable without additional processing. In total 273 survey records had coordinates recorded using only numeric characters. Of these 273, only 203 fell within Uganda, meaning the others were likely originally in a different coordinate system, but were incorrectly recorded.

During surveys, surveyors recorded the coordinates from a handheld GPS unit. These units are commonly capable of reporting the coordinates in at least three coordinate systems: degrees/minutes/seconds (DMS), degrees/decimal minutes (DDM), and decimal degrees (DD). The notation for DMS and DDM is comparably complicated, involving multiple non-alphanumeric characters as well as requiring the specification of the north/south and east/west state. The comparative complexity of these coordinate systems greatly increased the diversity of formats that coordinates were recorded in, as well as the probability of errors being introduced during the survey and subsequent transcription stages.

The following overarching post-processing stages were followed to clean as much data as possible:

1. Automated parsing of the raw coordinates
2. Manual interpretation of those that do not meet automated tests
3. Semi-automated testing of dates and coordinates based on the distance travelled per day, distance between points, and route along roads

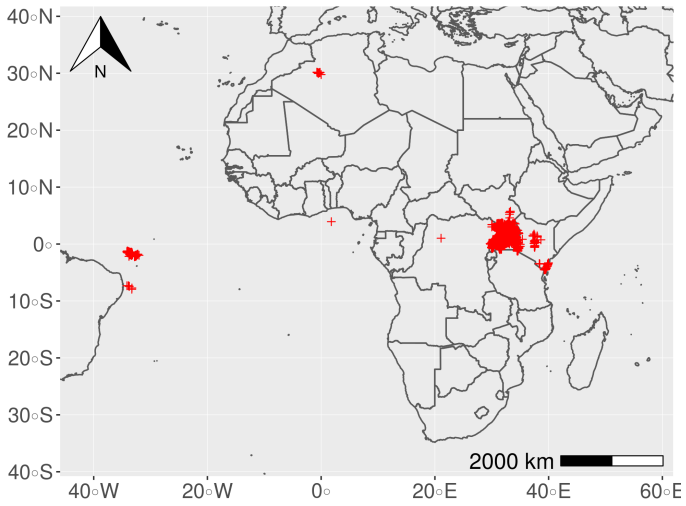


Figure 3.4: Distribution of Ugandan survey points after undergoing automated processing to extract coordinates from the raw data entered during digitisation, highlights the need for additional post-processing. 2581 of 2924 coordinates automatically converted lay within the border of Uganda.

4. Manual checking of all key aspects of field-level CBSD presence/absence, whitefly numbers and dates

5. Discard remaining ambiguous records

Given the scale of the dataset, we automated as much of the coordinate cleaning process as possible. A Python programme was written with approximately 30 different functions designed to parse 42 unique latitude notation patterns and 36 longitude patterns in the raw coordinate notation to decimal degrees. This programme automated the conversion of 2924 coordinate records, 2581 of which were located within Uganda. The 350 parsed coordinates outside Uganda were commonly caused by transcription errors. Figure 3.4 summarises the distribution of records after automated parsing.

For approximately 1000 remaining records, it was necessary to interpret the correct notation manually. Additional levels of automation were developed to streamline this process. Figure 3.5 contrasts the output for two survey years of the automated testing pre and post manual corrections.

The next stage of the data cleaning process was a manual comparison of the data against the scanned survey forms. The purpose of this check was to ensure that

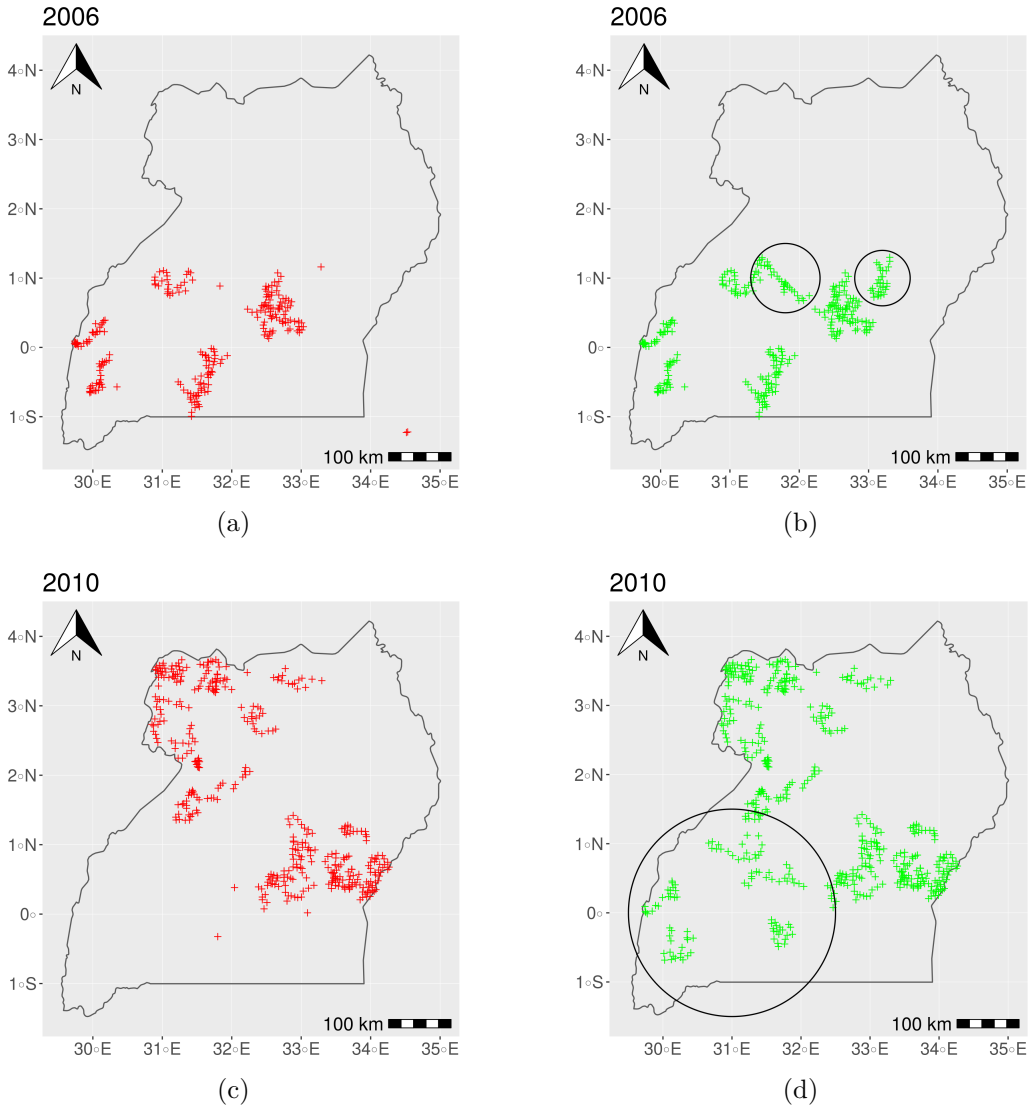


Figure 3.5: Representative examples of coordinates for which it was necessary to correct manually by returning to paper forms or inferring trends from sequential survey records. Circles highlight regions where significant corrections that were made during post-processing. a) 2006 data before and b) after corrections. c) 2010 data before and d) after corrections.

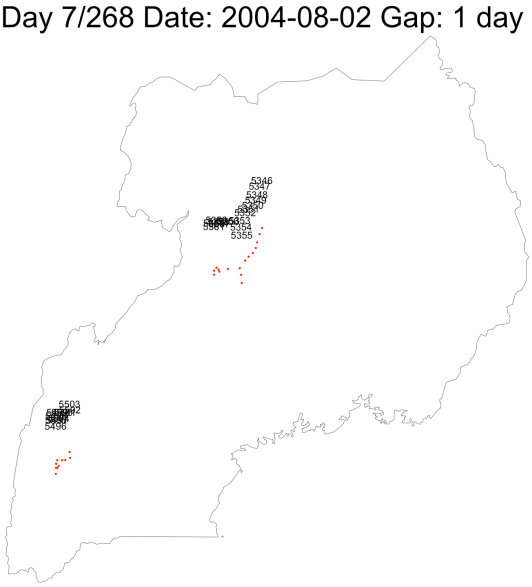
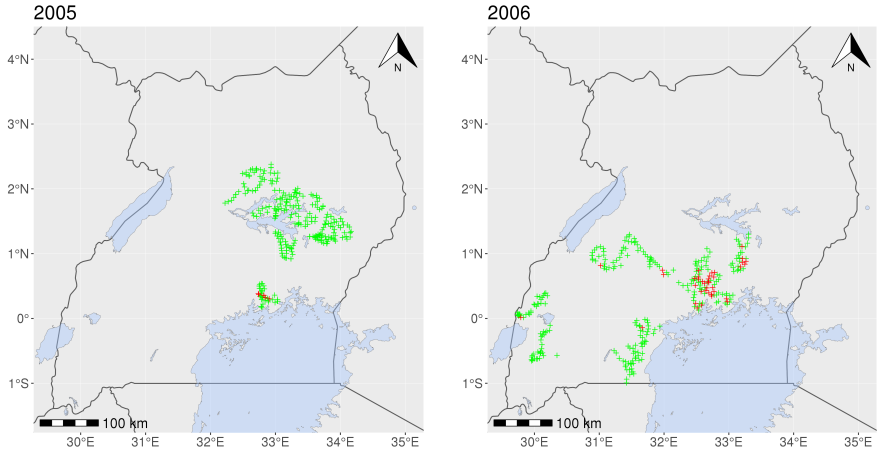


Figure 3.6: An example output plot from the R programme developed to map daily survey routes. Points indicate the location of a survey and label numbers refer to the unique form IDs assigned during scanning of the raw data records.

fields had not been mis-transcribed in terms of CBSD presence and absence, survey date and approximate per plant whitefly counts.

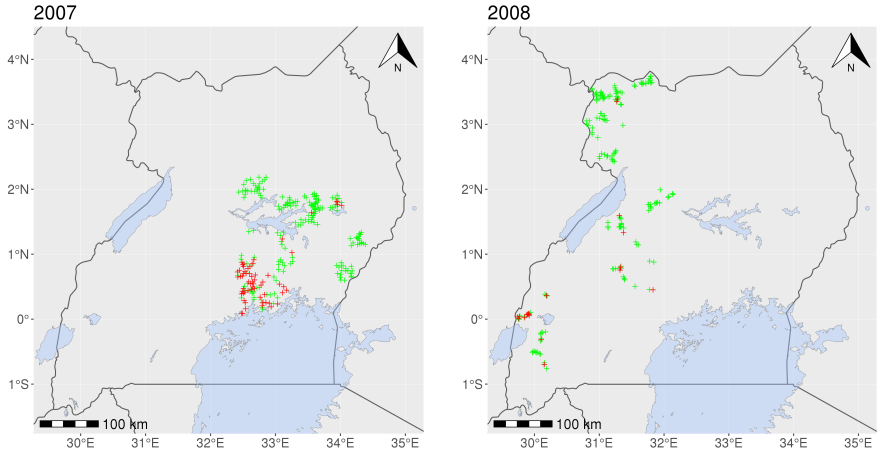
At this stage, wherever possible coordinates had been parsed and were now located within the bounds of Uganda. However, many coordinate errors were not sufficiently dramatic as to move the coordinate outside of Uganda. Therefore, an R programme was written to automatically plot the daily survey sites to ensure that all records correspond to a route that would be realistic in a single day. This led to an iterative process in which we returned to the scan of the original survey form to identify transcription errors and could infer and correct errors based on the single and multi-day spatial sequence of survey sites. Figure 3.6 gives an example of a single daily survey pattern.

Figure 3.7 presents a field level summary of CBSD symptom presence and absence in Uganda from the start of the epidemic in 2005 through to the most recent survey in 2017. The survey in 2017 was carried out using a Android survey app, the development of which is described in the subsequent section. The resultant dataset contained a digital record of the surveillance data, traceable by a unique form ID back to a scan of the original paper form.



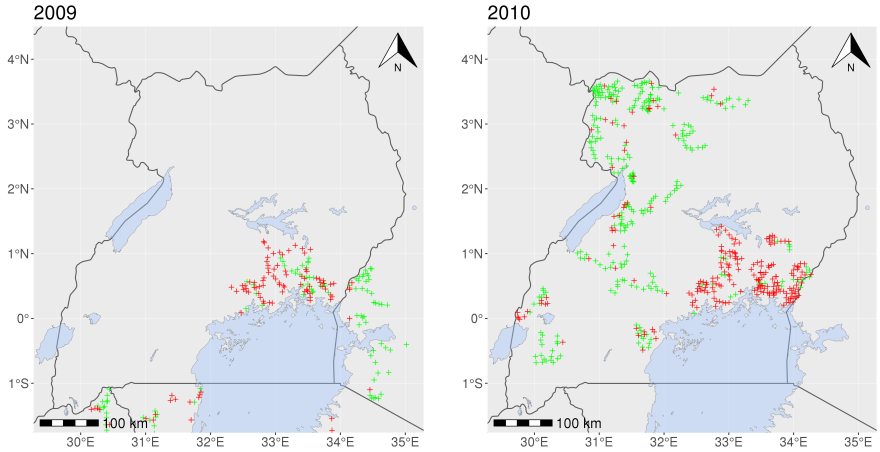
(a) 2005

(b) 2006



(c) 2007

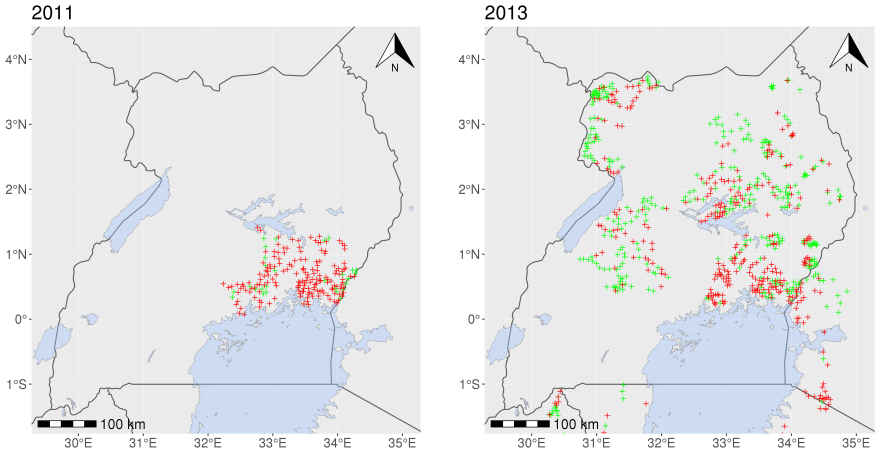
(d) 2008



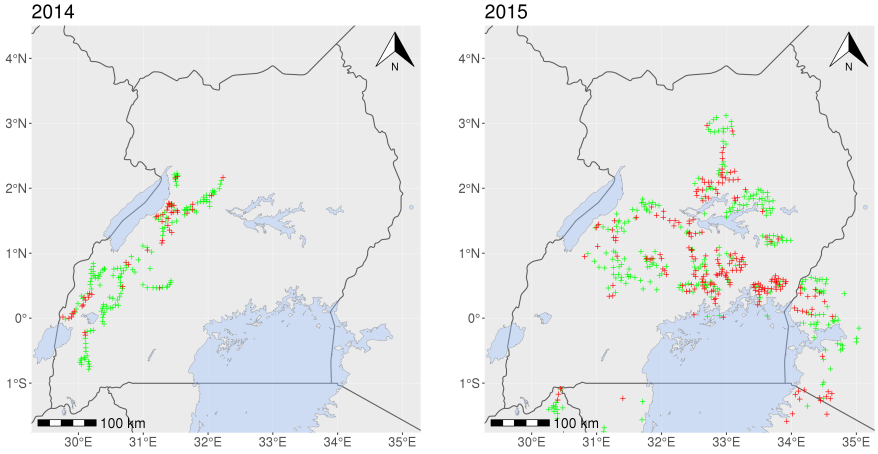
(e) 2009

(f) 2010

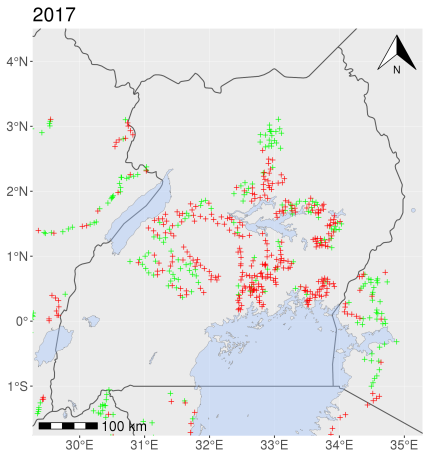
Figure 3.7



(g) 2011 (h) 2013



(i) 2014 (j) 2015



(k) 2017

Figure 3.7: Yearly maps of the Ugandan CBSD surveillance data. Red crosses indicate CBSD symptoms were identified on the 30 plants surveyed in a field and green crosses represent the absence of CBSD symptoms. Data between 2005-2015 were collected on paper forms and underwent digitisation and post-processing. Data from 2017 was collected using the Survey App, was uploaded in near real-time and required no post-processing.

3.5 Improving surveillance

Prior to the digitisation workshops, it was clear that various aspects of the surveillance protocol should be reformed to prevent the problems described in Section 3.2 from continuing. Moreover, as of 2015, a West African and DRC researcher consortium, the West African Virus Epidemiology project (WAVE), began conducting cassava virus and whitefly surveys with the ultimate goal of preventing CBSD spread to West Africa. Given the importance of data from this region, especially the DRC that likely contains the epidemic front, it was additionally important to avoid the problems that had previously occurred during East African surveillance. Above data quality, the most urgent aspect to reform was the common delay of many years between surveillance occurring and the scientific and policy community having access to these data to inform our understanding of the changing CBSD distribution.

Depending on the epidemiological goal of the survey, there are many ways in which the surveillance protocol could be optimised. For example, if the priority is finding the epidemic front, then the careful sampling of 30 plants per field may be a time consuming and ineffective way of finding low levels of infection. However, the number of country teams who were only familiar with the existing protocol meant it would be extremely challenging, time consuming and expensive to re-train surveyors to implement a significantly different protocol. We therefore took a two stage approach to reforming cassava surveillance.

The first stage involved the design, training and deployment of a survey app. Many of the issues previously encountered could be reformed by moving to digital data collection. The second stage of surveillance reform is to optimise the survey protocol for a given regions research goals and ensure surveyors are trained to maximise the efficiency and accuracy of the surveys. This is an ongoing process, and is discussed in Section 3.5.2.

3.5.1 Survey app development and training

A cross-platform, highly customisable data entry platform, iFormBuilder, was chosen for app development. This framework increased the speed at which an app could be developed and was considered to be more reliable, compatible and secure than de novo in-house development, especially given the professionally maintained database

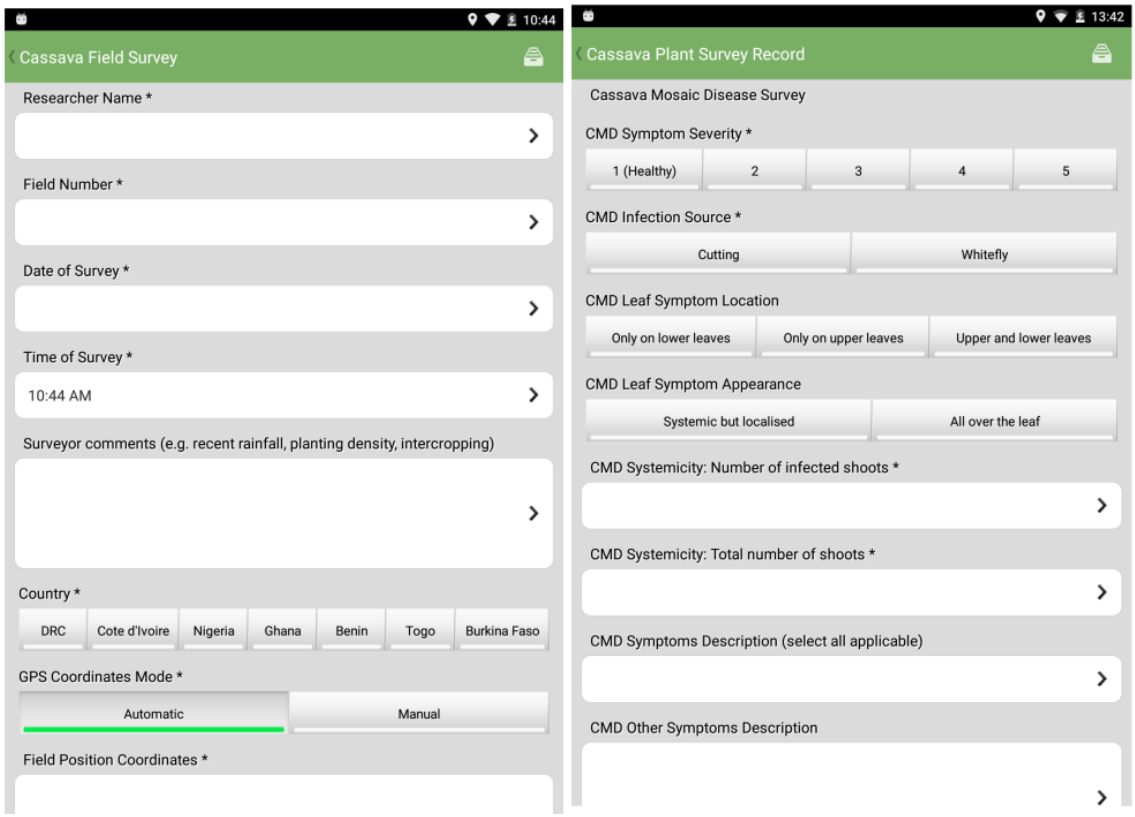
Feature	Benefit
Single template for all surveyors	Harmonisation across teams
Real-time error checking	Minimise errors Reduce time per survey
Automatic coordinates and datetime	Minimise errors Reduce time per survey
Pre-defined single button categorical answers	Reduce time per survey Minimise errors
Real or near-real time uploading	Minimise reporting delay Eliminate post-survey data entry

Table 3.4: Key benefits of digital data collection using the surveillance app in contrast to paper forms.

back end. To minimise barriers to adoption, we translated the existing protocol as closely as possible to the digital framework. Within the app, a range of beneficial features could be implemented to improve the speed and accuracy of the survey, as well as to minimise the lag between data collection and data sharing (Table 3.4). For example, key pieces of information, such as coordinates, date and time, could be recorded automatically. Beyond the benefits experienced in the field, data collection using the app eliminated the need for post-survey digitisation. Moreover, after training, the in-field collection rapidly became faster than paper form collection. Figure 3.8 gives an overview of the surveillance app interface.

The process of supporting and training surveyors to adopt digital data collection was far more involved than app development. Training was especially challenging for a number of reasons. Firstly, there are 14 countries spanning the full breadth of the continent undertaking surveillance as part of the CDP and WAVE projects. Moreover, multiple countries had more than one, largely independent, survey teams. Given the logistical challenge of travelling to all country teams, training sessions were appended to other project-related meetings whenever possible, when many project members were in a single location.

All surveillance teams were provided with at least one Android tablet. Wherever possible, these tablets were the same model to minimise platform specific variations in app behaviour and simplify the training process. When multiple country teams were attending a meeting in a single location, dedicated training sessions were organised with each team. During these sessions, attendees were guided through the process of first time app set up, followed by multiple rounds of practice data entry. This was most productive when it was possible to run through a full 30 plant sur-



(a) Field level information (b) Plant level information

Figure 3.8: Screenshots of the surveillance app interface on an Android tablet.

Figure 3.9: In field survey app training near Kampala, Uganda in 2017.

vey in a real cassava field. Importantly, it was necessary for attendees to practice multiple entries so that initial concerns and unfamiliarity were overcome and they could directly experience the benefits. This greatly increased the probability of a team adopting the app in subsequent surveys. Table 3.5 summarises the events at which app training and follow-up training was provided. In-field training occurred at all Africa based events with the exception the WAVE meeting in Abuja, Nigeria.

Notably, when in-field training was carried out with multiple survey teams in the same location, the main issues did not concern the app, but were focussed on disagreements amongst surveyors from different survey teams about how the details of the protocol should be implemented, for example, precisely how to count whitefly abundance. Therefore, despite the goal of the session being to familiarise surveyors with the app, significant progress was also made towards harmonising the protocol amongst teams. This highlighted the need for extensive harmonisation training in the event of future changes to the protocol. Moreover, training sessions allowed for the implementation of a correction to the protocol. For reasons described in the previous section, surveys had previously only sampled from the dominant cultivar in the field. To prevent the bias introduced by this method, the protocol was adapted to perform an unbiased random sample along a transect, and an detail SOP was provided, describing all aspects of the survey protocol, including this change. The app was adapted to make recording of the plant level cultivar as fast as possible.

During the training sessions, it was vital to build trust amongst facilitators and

Date	Location	Event	Attendees
Jan 16	San Diego, USA	Gates / PAG conference	CDP & WAVE
Jul 16	Abidjan, Ivory Coast	WAVE meeting	WAVE
Oct 16	Lusaka, Zambia	Whitefly meeting	CDP & WAVE representatives
Nov 16	Dar es Salaam, Tanzania	Data entry workshop	CDP
Feb 17	Kigali, Rwanda	Data entry workshop	Rwandan team
Apr 17	Kampala, Uganda	Data entry workshop	Ugandan team
Oct 17	Abuja, Nigeria	WAVE meeting	WAVE

Table 3.5: Project events where surveillance app training was provided.

attendees. Without this, surveyors were far less likely to be open about problems they were having both during the sessions and when it came to carrying out the surveys. Every effort was made to minimise the barriers for survey teams to get in touch facilitators after the workshop to discuss and receive advice on any problems they were experiencing.

Prior to the dedicated in-person training sessions, surveillance teams had access to tablets, the survey app and a detailed manual on how to set up the app and enter and upload data. However, this was insufficient to persuade any teams from WAVE of CDP to switch to digital data collection without dedicated training sessions. After the training sessions in 2016 and 2017, 16 teams across all 13 countries in East, Central and West Africa who carried out surveys in 2017 adopted the app. Figure 3.3 and 3.10 summarise the number of fields surveyed using the app by country in 2017 in East and West Africa respectively. Importantly, almost all data were uploaded within days of carrying out the surveys, in contrast to the multi-year or entire absence of wider reporting experienced previously. Moreover, the data required no further post-processing as the majority of sources of error, such as manual coordinate entry, had been replaced by automated processes. Figure 3.11 gives an overview of data uploaded through the app in 2017. The next round of surveillance in WAVE countries and a subset of East African countries is anticipated to commence in 2019.

3.5.2 Future steps: Optimising the protocol

Migrating to digital data collection and modifying the protocol to fix incorrect methods has greatly reduced the occurrence and impact of many of the issues outlined in Section 3.2 and reduced the reporting delay, for all countries, from multiple years

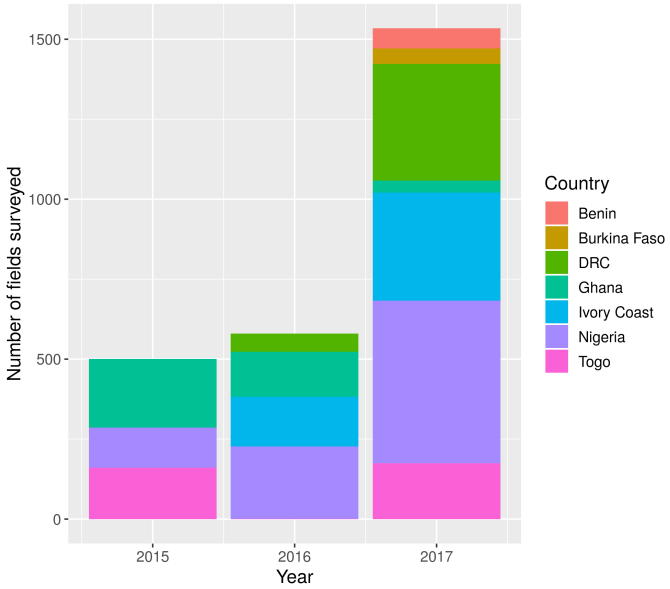


Figure 3.10: Summary of the number of field survey records from WAVE member countries that were digitised, by country team each year. Data from 2017 were collected digitally using the survey app.

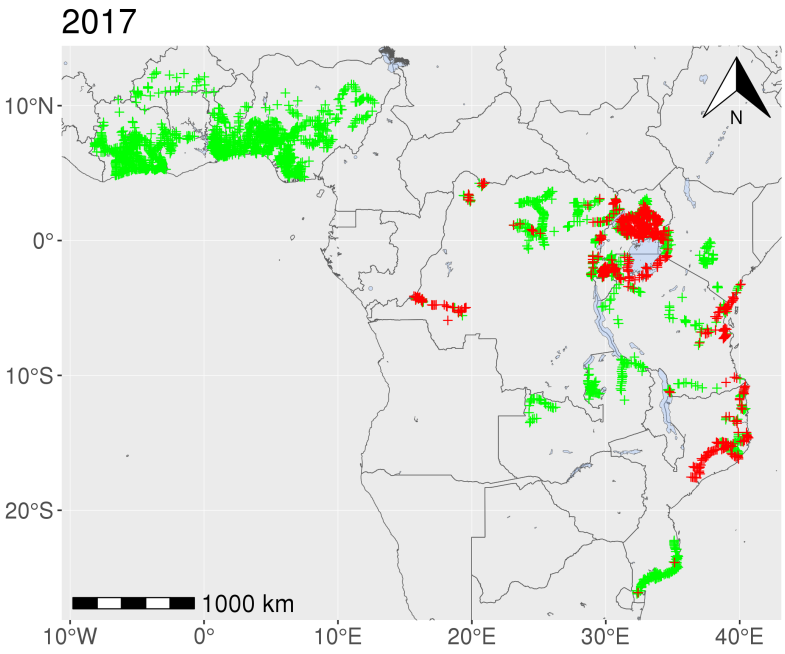


Figure 3.11: Map of the CBSD surveillance data uploaded using the survey app in 2017 from 16 teams in 13 partner countries. Unlike paper form collection, data required no additional post-processing. Red crosses indicate a CBSD symptoms were identified on the 30 plants surveyed in a field and green crosses represent the absence of CBSD symptoms.

to a few days. Importantly, within the app, it is easy to deploy modified data collection forms to all teams, allowing country specific data to be collected. Now that teams are familiar with the interface, it would be significantly easier to make changes to the protocol and deploy new data collection templates. This would greatly reduce, but not eliminate, the amount of in person training necessary, as much of the information could be clearly communicated within the app.

As highlighted in Figure 3.11, surveillance is extremely sparse or non-existent in Central Africa, specifically the DRC, Angola, Republic of Congo, CAR, Gabon and Cameroon. In coming years, the WAVE consortium is aiming to expand to include additional countries in this region in addition to the DRC. In the short-term, the entire absence of surveillance in certain countries will likely only be solved by significant additional investment by the government of a given country or an aid agency. However, two changes can be made to optimise the use exiting resources in countries currently undertaking surveillance.

Firstly, countries that are undertaking surveillance are generally designing surveillance routes to maximise the area covered. Therefore, if the goal is to find a CBSD incursion as early as possible, the location of surveyed fields within the country can be concentrated on high risk regions. The structure of the survey should therefore depend on the research question. For example, surveillance capacity in Nigeria may have a higher probability of early CBSD detection if concentrated on the eastern border with Cameroon and around major ports, such as Lagos. Nigerian surveillance strategies are explored in more detail in Chapter 5.

The second approach to optimising the use of existing resources involves changing the protocol itself. For example, it would be possible to increase the number of fields surveyed by decreasing the level of detail in a each field. The current protocol requires the careful sampling of 30 plants in the field and the collection of many potentially extraneous variables, such as the number major shoots on the plant. Moreover, given our poor understanding of true cultivar diversity or the ability of farmers or surveyors to identify cultivars correctly, it is unclear that cultivar records are meaningful. In regions that need to prioritise monitoring for CBSD presence/absence, streamlining these aspects of the protocol would allow either greater numbers of fields to be visited or more plants to be surveyed within the field. The exact trade-off between detection probability, number of plants surveyed, and number of fields surveyed depends on the incidence of CBSD in the region. We are actively working to develop in-field CBSD spread models to address this question.

In addition, counting *B. tabaci* abundance is often time consuming, especially in regions with higher populations. Currently, adult *B. tabaci* are counted from the top five leaves, which is a highly variable measure as adult *B. tabaci* fly away when disturbed. Both these issues may be improved by deploying the method described in Abisgold and Fishpool (1990), which instead involves counting *B. tabaci* nymphs to achieve a statistically representative measure of population size. Nymphs are affixed to the leaf and are therefore easier to count and a more stable measure, meaning fewer plants would need to be assessed. The method involves only counting from a subset of leaf sectors, which reduces the area to be counted. Moreover, as nymphs are affixed to the leaf, samples could be removed and stored during the survey, and counted either at a later time, either manually or automated using machine learning methods.

Nonetheless, there are limitations to the gains that can be made by streamlining the in-field survey. A large amount of a surveyors time is spend travelling between fields, especially in extremely challenging conditions, such as the DRC. Therefore, reducing the time spent in a single field may allow for a higher density of sampling along a single route, but not necessarily enable surveyors to reach sites that are inaccessible due to physical or resource limitations.

In the medium to long term, two strategies may improve our ability to generate surveillance data at the landscape scale. The first strategy involves leveraging and improving the capacity of national extension services to collect and centralise symptomatic plants observed in farmers throughout the country. In the longer term, technological solutions may deliver a more scalable solution to data collection. For example, SMS and smartphone based diagnosis and reporting systems have significant potential. Currently smallholders are unlikely to have access to smartphones but many have access to 'brick' phones with 44% average penetration across sub-Saharan Africa in 2017 (GSMA, 2017). These data sources are likely to be far noisier than active surveillance but have the potential to reach scales that would otherwise be infeasible, and their value likely depend on the system having the correct incentives to promote accurate reporting. Notably, there are existing examples of promising scalable surveillance systems being deployed for livestock disease surveillance (Hutchison et al., 2018).

Chapter 4

Development of a CBSD Spatial Epidemic Model

4.1 Introduction

4.1.1 Importance of cassava

Cassava is the second largest source of calories in Africa (FAO and IFAD, 2005). It is especially tolerant of drought and grows well in poor soils. Moreover, the tuberous root can be left in the ground for 2-3 years, meaning it is good insurance against the failure of other crops. For these reasons, it is disproportionately grown by poorer subsistence communities (FAO, 2010).

The DRC has the highest global level of per capita cassava consumption, at more than 1000 calories per person per day. Nigeria is the world's largest net producer, at over 34 million tonnes annually (FAO and IFAD, 2005). Despite the high net production of cassava in sub-Saharan Africa (SSA), the average fresh root yield in 2016 was 9.24 t/ha (Figure 4.1), significantly lower than the theoretically achievable yields of 15-40 t/ha achieved in Kenya and Uganda during farm breeding trials (Ntawuruhunga et al., 2006; Fermont et al., 2007). Many factors contribute to the sub-optimal yield, including low-yielding cultivars, lack of inputs, on-farm practices, environmental conditions, and the impact of pests and diseases (Fermont et al., 2009; Campo et al., 2011).

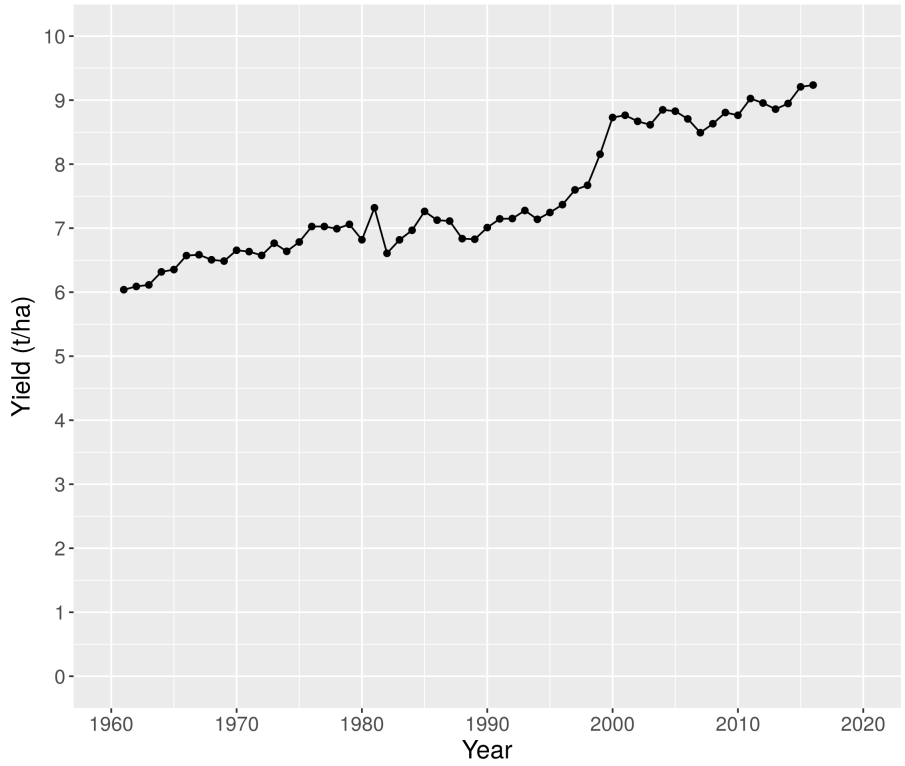


Figure 4.1: Average African yield of cassava between 1961-2016 (FAO, 2016)

4.1.2 Cassava brown streak disease

There are two primary disease threats to cassava: cassava mosaic disease (CMD) and cassava brown streak disease (CBSD). However, these two diseases are not equivalently distributed (Legg et al., 2011). Cassava mosaic disease is present throughout the continent, whereas CBSD is currently confirmed to be present in East Africa and the DRC and absent from West Africa (Legg et al., 2006; Mulimbi et al., 2012). Due to the endemic nature of CMD, programmes that attempt to increase cassava yield should contextualise the impact of CMD with other factors that reduce yield. In contrast, it makes sense to prioritise efforts to prevent the establishment of CBSD in West Africa as it is far easier to control an epidemic before it is widespread. The primary symptom of CBSD is a necrosis of the starchy root tissue (Story, 1936). The necrotic regions are inedible and it requires significant manual labour to extract what remains that is edible. Moreover, necrotic roots are unlikely to be accepted at markets (Hillocks et al., 2015).

Cassava brown streak disease was historically confined to the lowlands of coastal East Africa (Nichols, 1950; Legg et al., 2011). No significant spread was observed beyond the endemic region until 2004, when infected plants were observed near Kampala,

Uganda (Alicai et al., 2007). Since 2004, the epidemic appears to have spread from Uganda to the lake zone of Tanzania, western Kenya, Rwanda, Burundi and the DRC (Bigirimana et al., 2011; Legg et al., 2011; Mulimbi et al., 2012). Spread continues towards Southern Africa and West Africa. The disease was confirmed in northern Zambia in 2017 (Mulenga et al., 2018). Cassava brown streak disease root symptoms were observed in eastern DRC in 2009, notably in the absence of foliar symptoms (Legg and Bouwmeester, 2010). Diagnostic confirmation of CBSD in eastern DRC was confirmed in 2011 (Mulimbi et al., 2012).

The post-2009 epidemic spread of CBSD in the DRC is not well documented. This is especially problematic when addressing questions of spread to West Africa as the DRC is by far the most likely conduit. The detection of CBSD in DRC is complicated by two factors. Firstly, root necrosis that resembles CBSD has been present and reportedly spreading from south-western DRC since 2003 (Mahungu et al., 2003). In contrast to what is stated in Mahungu et al. (2003), the root necrosis is not coincident with CBSD foliar symptoms (Personal communication, T. Bakelana, 2017) and the causal agent is not currently known. Secondly, the lack of cheap, practical diagnostics makes categorical confirmation extremely difficult, without the problematic movement of material outside the country.

It is important to investigate the potential for well targeted management programmes to both minimise the spread and impact of CBSD within the DRC and reduce the rate of westward spread. This also requires a better understanding of the distribution and impact of CBSD within the DRC.

4.1.3 What can be done?

In poorer communities, there are generally fewer management options available for agricultural pests and diseases, especially for largely non-commercial crops. Therefore, it is especially important that currently disease-free regions prevent disease establishment, as eliminating an endemic disease would be costly and logistically difficult. Conversely, endemic regions should take a holistic approach, factoring in the relative importance of other factors on yield, such as CMD or improved access to farm inputs. This requires an improved understanding of the precise nature and extent of yield loss associated with CBSD, as the limited number of historic studies present a highly variable impression (Bock, 1994; Hillocks et al., 2001; Gondwe et al., 2003; Hillocks et al., 2015; Ephraim et al., 2015; Ndyetabula et al., 2016).

Any approach to disease prevention and management requires an understanding of the current disease distribution, availability and efficacy of management practices, and estimates of future spread. These may take the form of expert opinion, or be more formally approached through experimentation, surveillance and modelling.

4.1.4 Role of epidemiological modelling

Epidemiological modelling addresses the need to integrate disparate and uncertain information into a quantitatively formalised description of a pathosystem. The model can then be used to extrapolate from the given assumptions to predict the future behaviour of the system. Moreover, the model can be designed to factor in stochasticity, to estimate the range of possible future events and their relative probabilities.

Previous examples of epidemiological modelling and parameter estimation for emerging plant pathogens include an SEIR model of Bahia Bark Scaling on citrus groves in Brazil (Cunniffe et al., 2014) and an SEIDR model of Huanglongbing on citrus in Florida, USA (Parry et al., 2014), where the 'E' class represents exposure i.e. an incubation period, and the 'D' class represents detection. These are individual-based models and parameters are estimated using high resolution datasets on relatively small spatial scale. Meentemeyer et al. (2011) developed and parameterised an SI model for sudden oak death in California and Cunniffe et al. (2016) applied the model to questions of future spread. Often, due to the logistics of surveillance and data sharing, modelling of plant disease epidemics occurs retrospectively. However, due to the continental scale of the CBSD epidemic, we are in a position to apply these techniques to an ongoing epidemic.

4.1.5 Model development

Cassava is a vegetatively propagated crop. When harvesting the roots, stems are cut into multiple pieces, known as cuttings. A subset of these cuttings is planted in the next season. As multiple cuttings are obtained from a single plant, the excess cuttings can be sold as seed material or used as firewood. If infected stems are selected for planting, in the absence of intervention, CBSD will persist across cropping seasons. For this reason, an SI model appropriately describes the pathosystem in the absence of management.

There are two mechanisms of CBSV dispersal. The first mechanism is the movement of infected cuttings that introduces inoculum to a new field, but does not statistically multiply the total amount of infection in the system, if cuttings are selected randomly. The majority of cutting movements are likely to be local exchanges with neighbouring farms. However, longer range movements also occur between markets or by the movement of large amounts of planting material by governments or NGOs (Kansiime, 2014; Teeken et al., 2018). The second mechanism of dispersal is the whitefly vector, *Bemisia tabaci*, which increases the number of infected plants by transmitting viral particles between plants when feeding. Vector dispersal is also likely to be largely local, within and between nearby cassava fields. However, *B. tabaci* is readily wind dispersed and known to travel many kilometres in a matter of hours (Byrne, 1999). Within epidemiological models of agricultural systems, spatial spread is commonly described using a dispersal kernel, being the probability distribution describing the relative probability of dispersal occurring at different distances from a source. As CBSD dispersal events are primarily local, but occasionally long range, we propose a power law dispersal kernel to be the most appropriate functional form. This decision was reinforced during preliminary simulations with an exponential kernel (not shown), which proved incapable of fitting the data.

A further common practice in plant disease epidemiolgy is to define the host landscape using a discretised regular grid, known as a raster. In the case of an SI model for CBSD, each cell of the raster covers a defined spatial domain. Each cell contains a number of individual fields. Given the continental scale of the epidemic, it would be computationally infeasible and, most probably, uninformative to simulate at the plant level. Moreover, the sampling frequency and resolution of disease surveillance data does not lend itself to parameter estimation at the plant level. Accordingly, our model tracks the changes in the numbers of susceptible and infected fields over time.

The key questions to be addressed by CBSD modelling are:

- What is the current extent and density of CBSD infected fields?
- Where is CBSD likely to spread in the future?
- At what rate is CBSD likely to spread?
- Where and when should surveillance be undertaken to detect CBSD in high risk regions?

- Which strategies are likely to be most effective for managing disease-free and endemic countries, respectively?

For the reasons presented in this section, we propose a spatially explicit, stochastic susceptible-infectious (SI) metapopulation model of the CBSD epidemic. This chapter describes the development, parameterisation, and validation of the model, and finishes with a discussion of the strengths and limitations of the model.

4.2 Methods

4.2.1 Data

4.2.1.1 Surveillance data

Uganda has been carrying out national cassava disease surveys since the 1990s. The digitisation and validation of the post-2004 data is detailed in Chapter 3. As illustrated by Figure 3.7, the exact number and locations of fields surveyed varied considerably each year. The in-field surveillance methodology was based upon Sseruwagi et al. (2004), which involved the recording of CBSD symptom severity on 30 plants in a given field. As the base unit of the model is the field, for the purposes of parameter estimation, the plant-level data were processed to field-level CBSD presence/absence and mean per plant *B. tabaci* counts.

4.2.1.2 Cassava host production landscape

The host landscape model was developed in collaboration with Dr Anna Szyniszewska, Rothamsted Research. The model takes the form of a spatial grid with cells of approximately 1km^2, covering 29 major cassava producing countries in East, Central and West Africa. The model takes two forms of input data: human population data and regional cassava production statistics. Previous studies have shown that the spatial distribution of the non-urban human population is the best predictor of cassava density in sub-Saharan Africa (Carter and Jones, 1993; Ugwu and Nweke, 1996). The LandScan 2014 High Resolution Global Population Data Set is the source of the rasterised population density layer, at a resolution of 1km^2. Cassava production

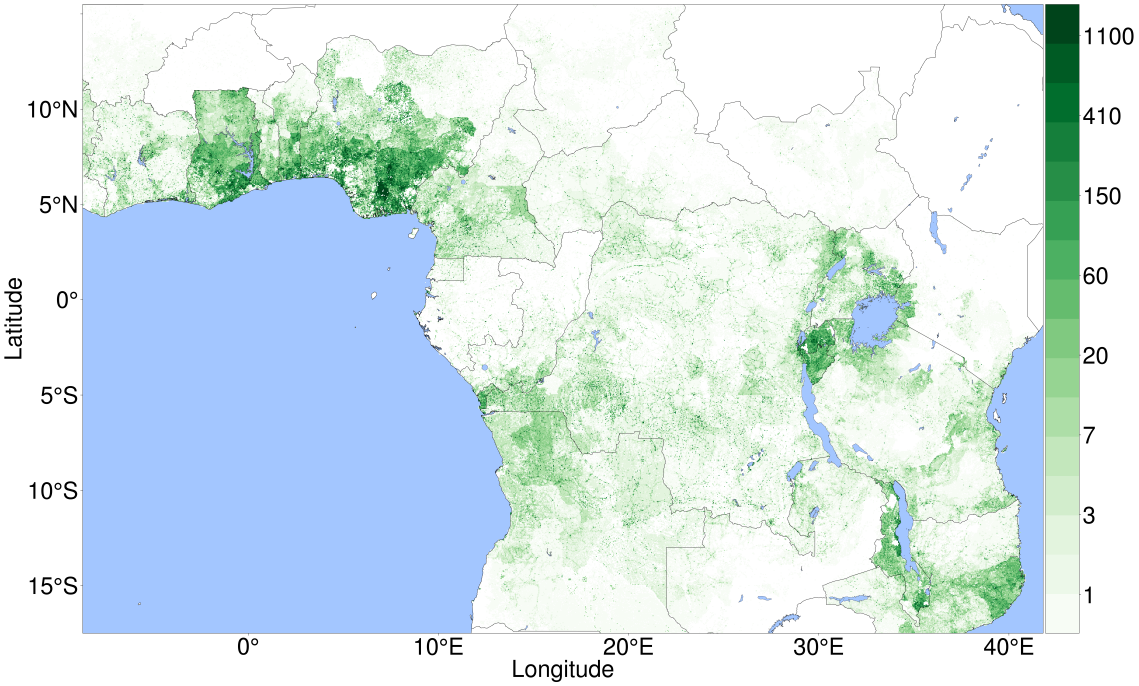


Figure 4.2: Map of the cassava host landscape, by number of fields per km^2 on a log scale.

statistics were derived from, FAO AgroMaps 2014 (kids.fao.org/agromaps) in tonnes per hectare. For countries unavailable on AgroMaps, data were obtained from national statistics produced by the Ministry of Agriculture for the relevant country. The resolution of the production data varied by region, hence, we selected the finest resolution available for each region with countries.

A linear relationship was assumed between human population density and cassava production. Therefore, for each region, the total production volume was allocated proportional to the number of inhabitants per km^2. However, spatial locations with populations greater than 5000 inhabitants per km^2 were excluded to avoid the allocation of production to urban areas. The model caps production at 1000 tonnes per km^2.

The base unit for the SI model is a field. Hence, the host production landscape model must be converted to number of fields. We assume an average per field cassava yield of 10 tonnes per hectare (FAO, 2016) and an average field size of 0.1 ha (Lose et al., 2003; Night et al., 2011; Owusu and Owusu-Sekyere, 2014). Figure 4.2 represents the number of cassava fields per km^2 across the continent on a log scale.

4.2.2 Model structure

The spread of CBSD is modelled as a spatially explicit, stochastic SI epidemic. The model is implemented using the methodology described in detail in Stutt (2015). The rasterised host landscape governs the initial distribution of susceptible fields (Figure 4.2). Spatial coupling is governed by an isotropic discrete dispersal kernel. During exploratory stages of model development, two candidate kernel functions were considered: exponential (Equation 4.1) and power law (Equation 4.2). However, the exponential kernel was rejected for the following reasons. Firstly, from an epidemiological perspective, the predominantly short range and lower probability long range dispersal described in Section 4.1.5 is better described by the functional form of a power law kernel. This perspective was reinforced during exploratory model development in which model predictions using the exponential kernel showed poor correspondence with the data.

$$K(d) = Ae^{-\alpha d} \tag{4.1}$$

$$K(d) = Ad^{-\alpha} \tag{4.2}$$

The distance between two raster cell centroids is d. The parameter, p, defines the kernel value at $d = 0$. In practical terms, p is the proportion of dispersal that remains within the source cell. A kernel cut-off distance, $D_{max} = 500$km, sets the maximum distance from the source cell that the kernel covers. For the finite set of cell centroids in the range $0 \leq d \leq D_{max}$, values are calculated based on the kernel function. A normalisation factor, A, is applied such that the sum of kernel values for $d > 0$ is equal to the value of $1 - p$. Therefore, the centroid cell kernel value is p, and the sum of the kernel is 1. The rate of dispersal from a given raster cell is βI, which is dispersed in accordance with the kernel, where β is the transmission rate. The instantaneous rate of infection at location i is governed by the equation in Table 4.1.

The model is simulated as a discrete event, continuous time stochastic process using an optimised Gillespie algorithm (Gillespie, 1977; Stutt, 2015).

Event	Rate	Effect
Infection at location i	$\left[\sum_j \beta I_j K(d_{ij})\right] S_i$	$S \to S-1$, $I \to I+1$

Table 4.1: The instantaneous rate of infection at location i, where j is indexes the force of infection on location i from all locations within the rasterised landscape, including i.

4.2.3 Initial conditions

The simulation start time, $t = 0$, corresponds with 1st January 2004. The model is seeded with initial infection equivalent to the number of infected fields observed during the 2005 survey (Figure 3.7), being the first year in which national surveys record CBSD symptomatic fields. The model time unit is years.

4.2.4 Simulated surveillance

The instantaneous state of the model is defined by the number of susceptible and infectious fields in each raster cell. The model implements a surveillance scheme that replicates the real-world surveillance structure. For all years that surveillance was carried out in Uganda, we perform one instantaneous survey at the end of the simulation year. For example, we assume the 2005 surveillance data are representative of the observed state at the end of 2005, $t = 2$. For each raster cell in the model landscape, we sum the number of fields that were surveyed in the Ugandan national survey within bounds of the 1 km^2 cell for a given survey year. Based on this real-world per cell sampling intensity, the equivalent number of fields in each model cell is randomly sampled and the numbers of sampled fields that are in each system state of susceptible and infectious are recorded.

For the majority of the Ugandan national surveys, only the dominant cultivar was surveyed in a given field. However, it is common for multiple cultivars to be grown in a single field. Cassava cultivars are known to vary in their susceptibility to CBSD and symptomaticity post-infection. Surveys only sample a small number of plants in the field. Therefore, if different cultivars in a field have different probabilities of showing symptoms and the dominant cultivar is not selected by the farmer based on CBSD response, by sampling only a single cultivar, the probability of a survey being summarised to field-level CBSD absence is higher. In addition, surveys are based

Figure 4.3: Ugandan surveillance data counts highlighting the number of fields surveyed that are reported as CBSD infected based on any cultivar in the field compared with sampling only the dominant cultivar. These data were collected during surveys between 2009 and 2014.

on above ground visual symptoms. Depending on the time of infection relative to survey time and the severity of cultivar response, there is a greater probability of not observing symptoms on a given plant. Moreover, given the subtlety of CBSD symptoms, false negatives may also occur due to human error.

Using the Ugandan surveillance data, we can estimate the probability that sampling only the dominant cultivar results in a false negative. Between 2009 and 2014 the data were collected on CBSD presence on both the dominant cultivar in the field and also on any other cultivar. The mean difference between the number of positives in the dominant cultivar compared with any cultivar is 0.167 (Figure 4.3).

Based on this calculation of false negative probability, we adapt the simulated surveillance scheme to incur false negatives. This is implemented via a binomial trial with a probability of reporting a randomly sampled infected field as uninfected of 0.15 (rounded from the calculated value of 0.167).

Event	Rate	Effect
Infection at location i	$\left[\sum_j \beta w_j I_j K(d_{ij})\right] w_i S_i$	$S \to S-1$, $I \to I+1$

Table 4.2: Modified instantaneous rate of infection equation, incorporating the vector abundance parameter, w. The location in question is indexed by i and j is indexing all locations within the rasterised landscape, including i.

4.2.5 Vector population density

In the final iteration of model development, the field-level mean *B. tabaci* abundance scores, derived by averaging for a given field the per plant *B. tabaci* counts on the top five leaves, were aggregated to modulate the rates for infection to and transmission from a given raster cell site. Specifically, the field-level mean vector abundance data from all cleaned survey data across all partner countries and years (Chapter 3) were collapsed across years into a single atemporal spatial point dataset. This point dataset then underwent inverse distance weighted (IDW) interpolation with a power value of 1.0, generating a rasterised vector abundance layer across the host landscape with a cell size of 100km^2. Field-level vector abundance mean values were capped to a value of 20 and the layer was then normalised.

The equation governing the rate of infection at a given cell site now incorporates vector abundance parameter, w, and is described in Table 4.2. We assume a linear relationship between vector abundance and its effect on infection and susceptibility. The mean vector abundance cap value of 20 and the linear relationship are informed by currently unpublished in-field CBSD modelling results developed in collaboration with other members of the Epidemiology and Modelling Group.

4.2.6 Parameter inference

The parameters to be estimated are the proportion of dispersal that remains within the source cell, p, the kernel scale parameter, α, and the transmission rate, β. We adopt a likelihood-free Bayesian approach, Approximate Bayesian Computation (ABC), which requires a distance measure between the real and simulated data, ρ, along with a tolerance, ϵ (Tavare et al., 1997; McKinley et al., 2009; Toni et al., 2009). Summary statistics, S, of the simulated and real data are used to reduce the dimensionality of complex data. A conceptual overview of ABC and the selection

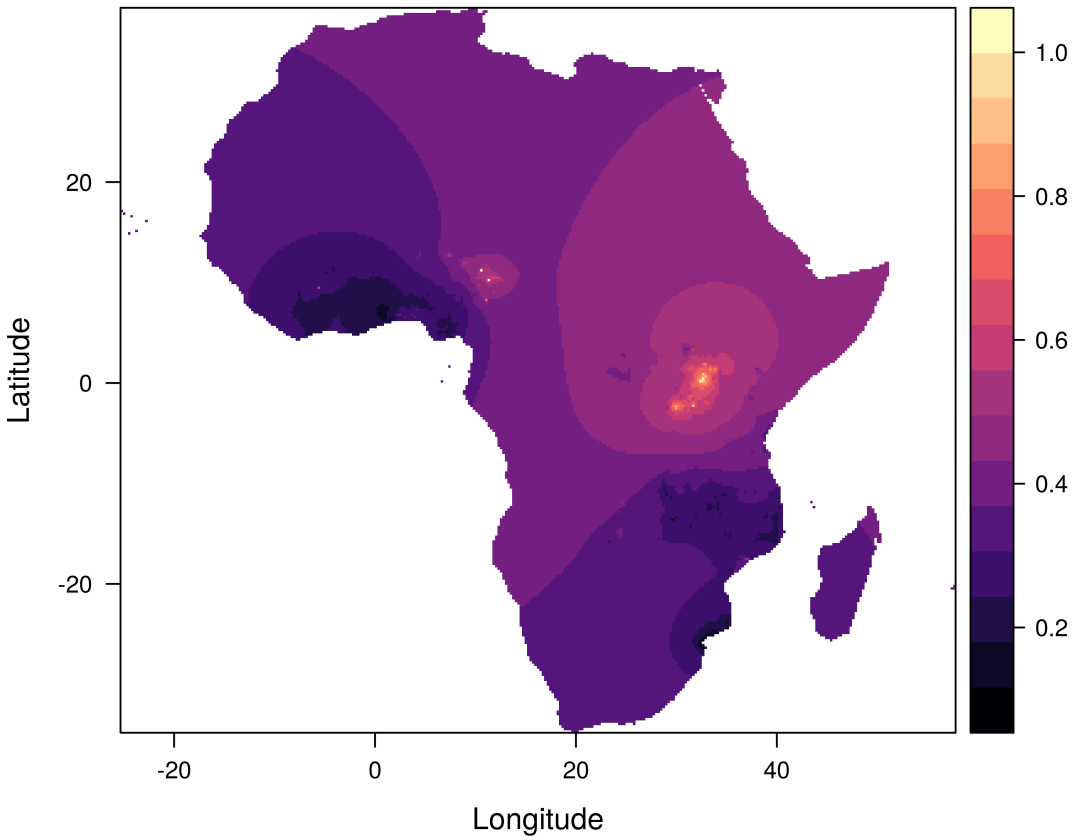


Figure 4.4: Discretised representation of the vector abundance layer, generated via the IDW interpolation of mean field *B. tabaci* count data across all surveys.

algorithm is provided in Section 1.2.5

The Ugandan surveillance data are divided into two distinct datasets. The training dataset, d_{fit}, consisting of data from 2005 to 2010, inclusive. The validation dataset, d_{val}, consists of the remaining data from 2011 to 2017. Let the model be M, the set of input parameters, θ, the prior, $p(\theta)$, simulated data, d_{sim}, summary statistic, S, and posterior $p(\theta|d_{fit})$.

4.2.6.1 Summary statistic selection overview

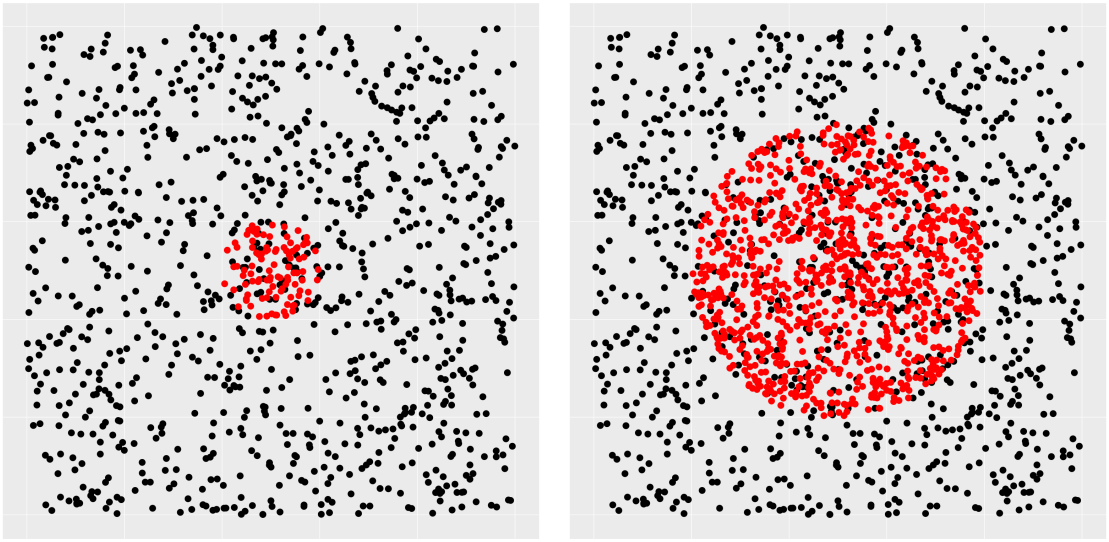
Successful fitting of models to epidemic data requires the selection of summary statistics that characterise the spatial and temporal patterns of epidemics. The summary statistics are then use to compare model simulations with empirical data.

Plant disease epidemics result from the iterative process of an infectious disease spreading from an infected plant to a susceptible one. When viewed at the landscape-scale, different epidemics have different spread characteristics. For the purposes of simplicity, we consider an SI epidemic in the absence of management. Focussing purely on spatial distribution, the epidemic may take the form of uniform wave, infecting most or all individuals as it spreads. In this case the dispersal mechanism(s) are highly localised. An example of this 'class' of epidemic is illustrated in Figure 4.5. In contrast, Figure 4.6 illustrates an epidemic with a longer average dispersal distance, which results in a patchier but more widely distributed landscape of infectious individuals.

In addition to spatial patterning, we can also differentiate epidemics by the rate at which their spatial patterns change. For example, taking Figure 4.5, the transition between the two distributions of infected individuals could occur over the course of a single day or a year. Table 4.3 summarises a set of epidemic traits that can be approximated from multiple years of surveillance data. Whilst Figures 4.5 and 4.6 illustrate epidemic spread with clearly distinct landscape-scale behaviour, it is worth emphasising that differences in epidemic dynamics exist on a continuum.

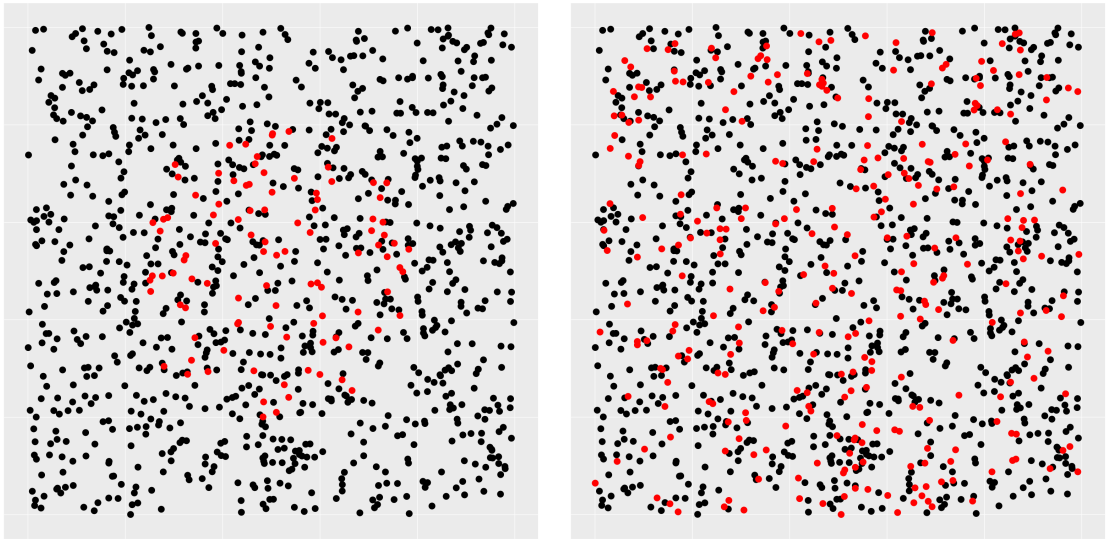
Characteristic	Description
Spatial extent	Where is the disease present and absent in the landscape
Regional incidence	What is the average disease incidence at a regional scale
Regional bulk up rate	How quickly does an introduction multiply at a regional scale
Local bulk up rate	How quickly does an introduction multiply at a local scale

Table 4.3: High level characteristics of an epidemic



(a) $t = 0$ (b) $t = 1$

Figure 4.5: Cartoon illustration of a uniform wave epidemic, characterised by an exponential dispersal kernel, as the distribution of infected individuals changes over time.



(a) $t = 0$ (b) $t = 1$

Figure 4.6: Cartoon illustration of a long range dispersal epidemic, characterised by a power law dispersal kernel, as the distribution of infected individuals changes over time.

In simple terms, the goal of estimating parameters using the Ugandan CBSD surveillance data is to calibrate the model to the dynamics of the spatiotemporal spread observed in the data (Burr and Skurikhin, 2013). Therefore, the process of summary statistic selection is to identify statistics with the ability to capture the landscape-scale spatiotemporal spread behaviour of the CBSD epidemic. However, the optimal method for selecting summary statistics and tolerances remains an open question in the likelihood-free inference literature (Beaumont, 2010; Prangle et al., 2013; Braunack-Mayer, 2013). For complex data, it may not be computationally feasible to use summary statistics that meet the statistical definition of sufficiency (Csilléry et al., 2010). Therefore, it is common practice to select summary statistics based on their ability to capture an important property of the system as informed by expert opinion (Beaumont, 2010; Csilléry et al., 2010; Burr and Skurikhin, 2013). Statistics derived in this way are described as informative summary statistics.

Given the high dimensionality of the Ugandan surveillance data, we have proposed a set of informative summary statistics, designed to discriminate between epidemics with different landscape-scale dynamic behaviour. In addition, we propose a methodology to assess the suitability of a given statistic. The methodology detailed in the following section is related to the selection procedure described in Burr and Skurikhin (2013), but adapted to the constraints of more complex stochastic models.

4.2.6.2 Assessing summary statistics using artificial data to recover known parameter values

We first identify robust summary statistics using artificial data for which we know the true value of the parameters. The process of summary statistic assessment begins by selecting model parameters from the prior distribution and running the model using these known parameters to generate artificial epidemic surveillance data, $d_{artificial}$. The location and timing of simulated surveillance for the artificial data is retained from the real-world historic Ugandan surveys.

We then perform the ABC parameter estimation procedure (Algorithm 2) to derive a posterior using a given statistic, S. As the tolerance is reduced, $\epsilon \to 0$, we assess the behaviour of the statistic to discern between different regions of parameter space and converge towards the known input parameters. To test whether a statistic is robust to the stochasticity of the model, the statistic evaluation procedure is

Parameter set	Alpha	Beta	P	# replicates	# fitting simulations
0	3.5	403.43	0.5	10	20,000
1	3	121.51	0.5	10	100,000
2	2.2	121.51	0.5	10	100,000

Table 4.4: Input parameter sets for the generation of artificial simulation data to assess summary statistic performance.

repeated to generate 10 distinct sets of artificial epidemic data for a given set of input parameters. Extending the procedure beyond 10 replicates per parameter set did not change the behaviour of each statistic. In fitting to artificial data, the selection procedure emulates the process of estimating parameters on the real-world data, but when the true parameters are known.

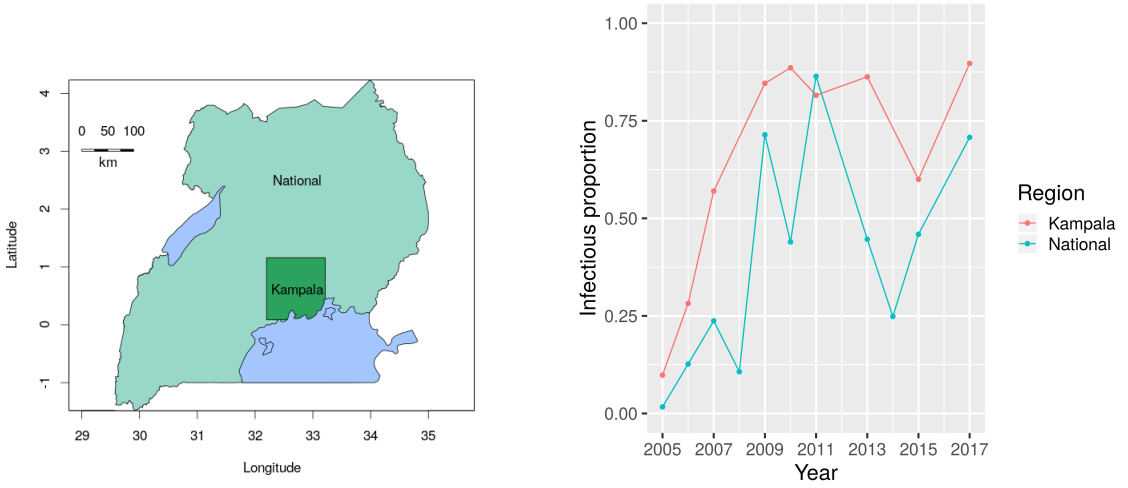
The entire process is repeated for different parameter sets to generate artificial data with different spread behaviour. The input parameter value sets are given in Table 4.4. The best performing summary statistics are then selected to estimate parameters using the real survey data.

4.2.6.3 Candidate summary statistics

A candidate set of summary statistics is proposed for assessment. Here, we define infectious proportion as the proportion of infected fields in a defined region out of the total number of sites surveyed in the region. The infectious proportion is calculated for each survey year independently. The infectious proportion tolerance is defined as a symmetric positive or negative deviation from the target value. For example, a target infectious proportion of 0.5 and a tolerance of 0.25 gives 0.5 ± 0.25. In the following three subsections, we outline the summary statistics selected to undergo assessment.

4.2.6.3.1 National and Kampala infectious proportion (National-based)

We divide Uganda into two, non-overlapping, regions. Figure 4.7 illustrates the extent of the two subregions and their corresponding infectious proportion values calculated from the Ugandan national surveillance data. The first region covers a small, densely sampled, area around Kampala. The second is what remains of Uganda, excluding the Kampala region. The infectious proportion is calculated for



(a) (b)

Figure 4.7: National-based summary statistic highlighting the extent of each sub-region and the corresponding infectious proportion calculated from the Ugandan national surveillance data.

each survey year. These statistics capture local and national bulk up dynamics.

4.2.6.3.2 Regional infectious proportion (Region-based)

Uganda is divided into four administrative regions: Central, Western, Eastern and Northern (Figure 4.8). We then calculate the infectious proportion within each region (Figure 4.9). The administrative regions coarsely correspond to the agro-ecological zones in Uganda (Okonya et al., 2014).

4.2.6.3.3 Regular grid disease presence (Grid-based)

The maximum latitude longitude extent of Uganda is divided into a regular grid, resulting in 25 quadrats covering the entire land area. For each of these sections, the survey data are processed to classify yearly presence or absence (Figure 4.11a). We then derive the year of first CBSD detection for each quadrat. This results in a timeline for each quadrat that defines the years of negative observation, up to and including the first year of CBSD observation (Figure 4.11b). Up to and including 2010, 19 quadrats contain surveillance data, and the additional four have been surveyed by 2017. Survey points from neighbouring countries that fall within the grid are included for the year 2009. The statistic score is defined as the proportion of quadrat timelines that are perfectly matched. Therefore, a perfect match is 1.0.

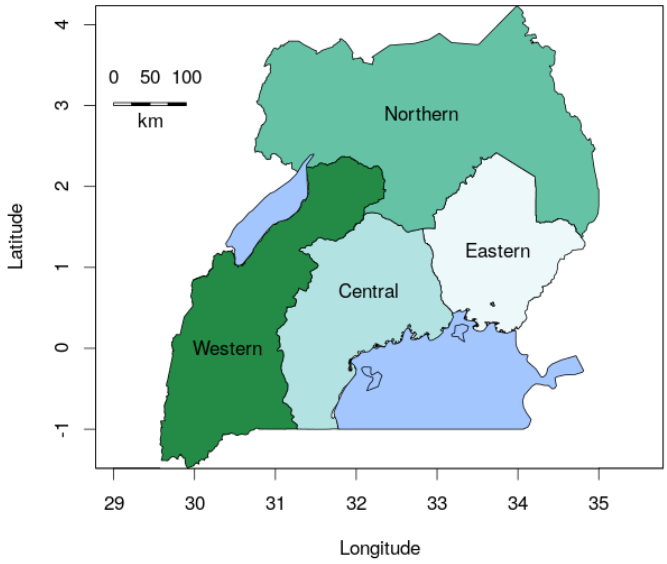
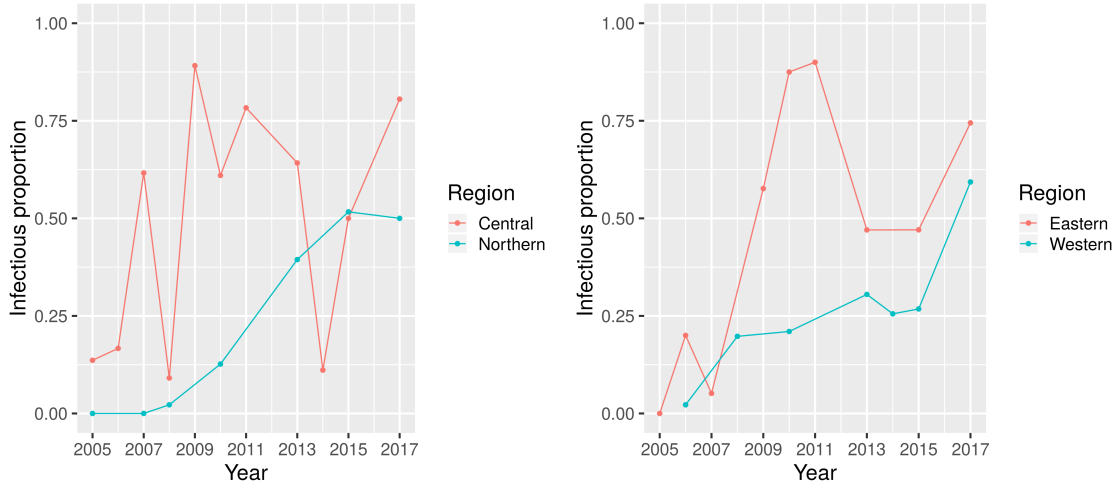


Figure 4.8: The four non-overlapping administrative regions of Uganda, from which yearly infectious proportions are calculated for the Region-based summary statistic.



(a) (b)

Figure 4.9: Region-based summary statistic infectious proportion values from sub-regions identified in Figure 4.8 calculated from the Ugandan national surveillance data.

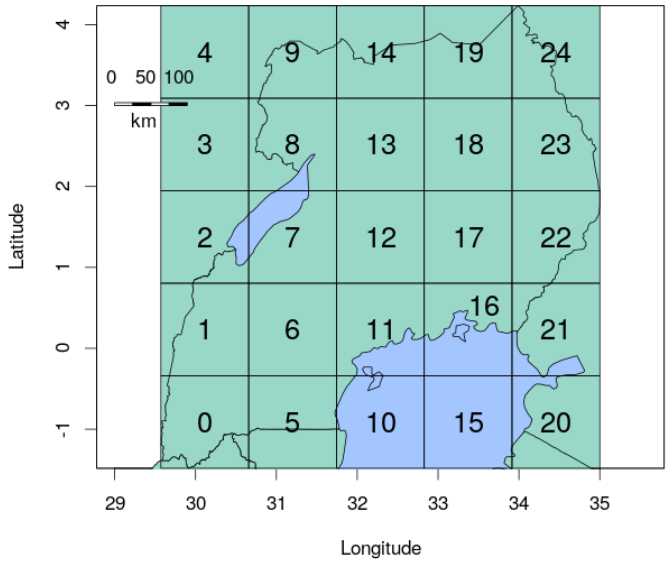


Figure 4.10: Regular quadrat grid covering the extent of Uganda. Quadrats 3, 4, and 15 are excluded as they do not contain any survey points within the fitting period. Survey points from neighbouring countries that fall within the grid are included for the year 2009.

The tolerance is applied as a deviation from 1.0.

4.2.7 Model validation using empirical data

Ensuring the selection of informative summary statistics is an important step in ABC parameter estimation. However, the true test of model performance is assessed during validation. In order to validate the model, we perform 10,000 simulations from 2004 to 2017 using parameters randomly sampled from the posterior. For simulations that pass the fitting criteria used for parameter estimation from 2005 to 2010, we assess the behaviour of the model relative to the validation dataset, d_{val}, for 2011 to 2017. We report the proportion of simulations that pass within a specified tolerance using the infectious proportion summary statistics (and undertake visual comparisons of the simulated and real-world surveys).

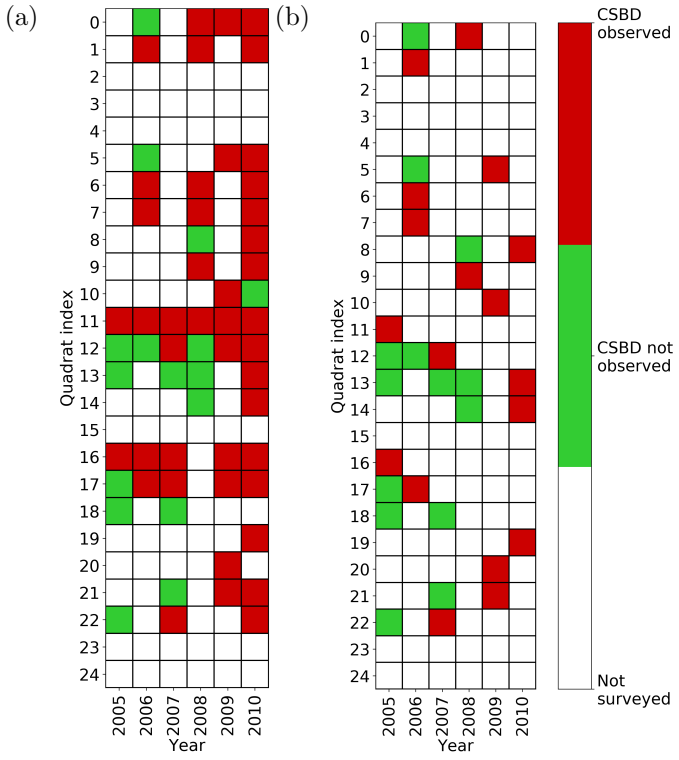


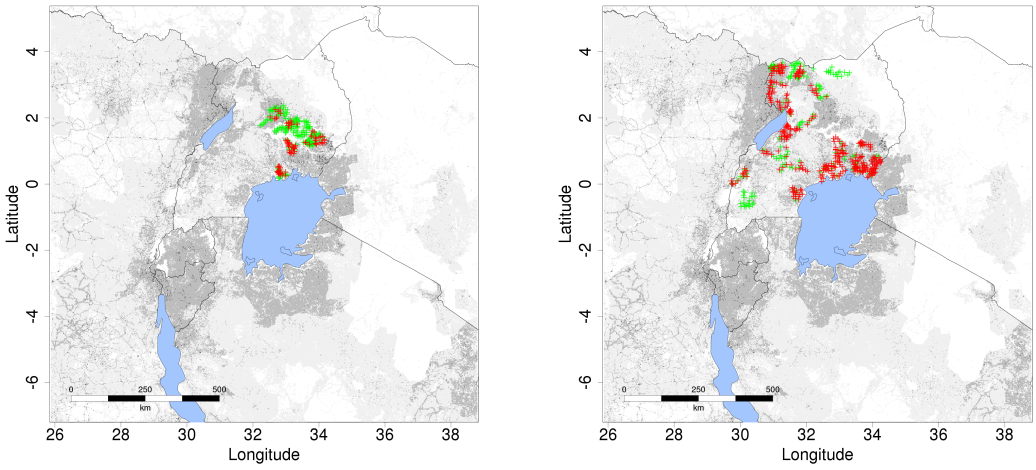
Figure 4.11: Summary of the surveillance data into 25 quadrats. (a) illustrates the full dataset, (b) shows the data for each quadrat up to the year of first detection

4.3 Results

4.3.1 Assessing summary statistics using artificial data

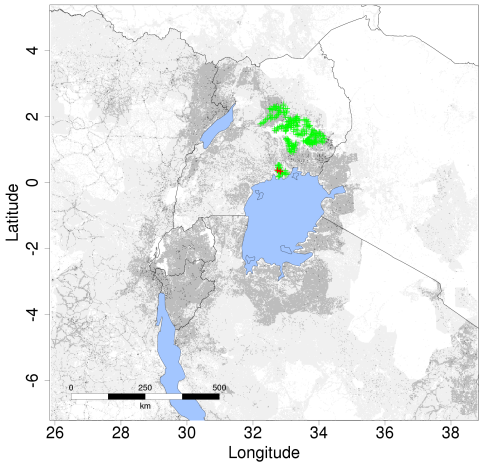
Three different parameter sets were selected from distinct regions of parameter space to test summary statistic robustness to a diversity of epidemics. The exact input parameters and the number of simulations for each round of ABC are detailed in Table 4.4. Ten seven year epidemic simulations were performed from each parameter set. Figure 4.12 shows the outputs of the simulated surveillance scheme a representative artificial simulation for each input parameter set. Importantly, each parameter set results in epidemics with visually distinct spread patterns.

For all replicate artificial epidemic datasets for each parameter set, ABC parameter estimation was implemented via the ABC rejection algorithm (Algorithm 2) and each summary statistic independently and for summary statistics in combination across a range of tolerances. As the tolerance was reduced, the posterior distribution converges towards the simulation parameters (Figures 4.13 to 4.15). Each

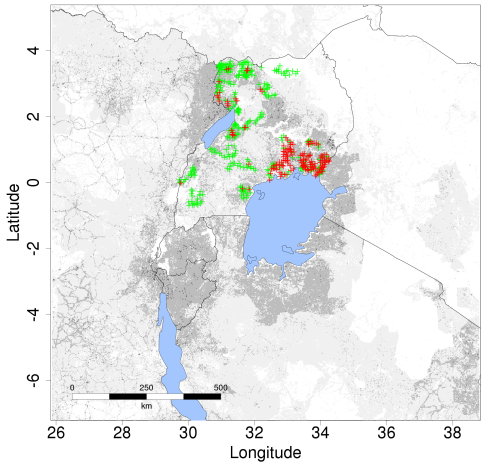


(a) Parameter set 0: 2005

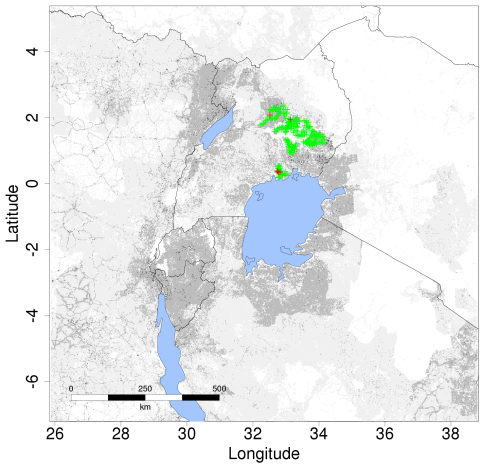
(b) Parameter set 0: 2010



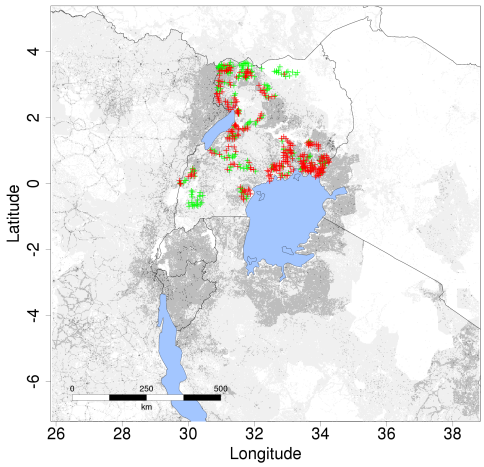
(c) Parameter set 1: 2005



(d) Parameter set 1: 2010



(e) Parameter set 2: 2005



(f) Parameter set 2: 2010

Figure 4.12: Simulated surveillance data from a single realisation of each parameter set from Table 4.4, highlighting the fundamental differences in the spatial structure of the different epidemics as the input parameters are varied. The intensity of the grey shading represents the density of cassava cultivation. Red crosses indicated CBSD detection and green crosses for a CBSD negative in the simulated surveillance.

Model	Compartments	Kernel	Host landscape	Survey error	Vector layer
model_1	SI	Power law	Yes	No	No
model_2	SI	Power law	Yes	Yes	No
model_3	SI	Power law	Yes	Yes	Yes

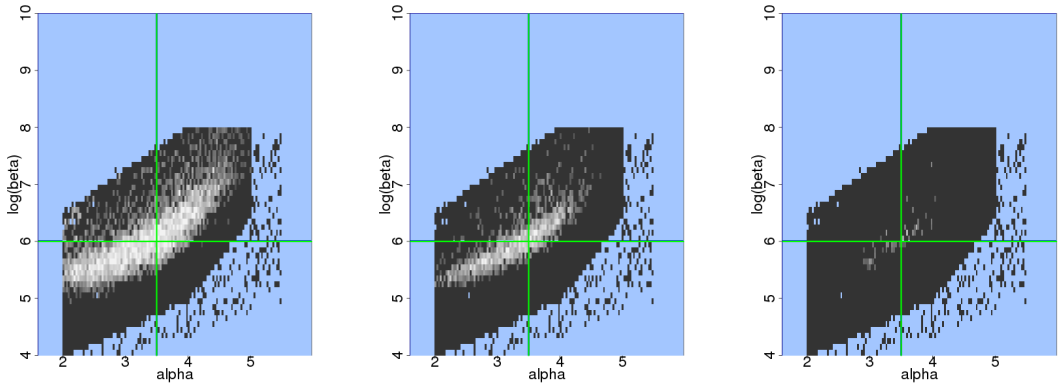
Table 4.5: Summary of model features explored during development.

summary statistic reliably recovered the known simulation parameters. In addition, the ABC procedure with a combination of the national-based and grid-based statistic performed especially well (Figure 4.16). We conclude that the national-based and region-based summary statistics capture broadly similar epidemic characteristics. The good performance of the national-based and grid-based statistics when used in combination (Figure 4.16) motivated the selection of this pair of statistics for ABC parameter estimation with the Ugandan surveillance data.

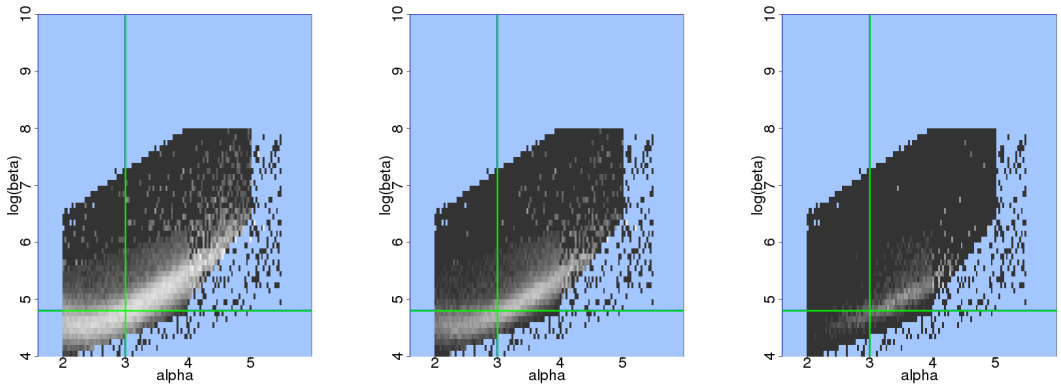
4.3.2 Estimating parameters on empirical Ugandan surveillance data from 2005-2010

Three different model structures were tested during model development, comprised of different combinations of the survey error probability and vector abundance layer components. Table 4.5 summarises the features of each of the three tested models. Model_1 is the base model without survey error or vector abundance features. Model_2 adds in a survey error probability of 0.15 (see Section 4.2.4), and Model_3 includes both the survey error probability and vector abundance layer. During early development, models were tested with exponential kernels, but proved incapable of reproducing the observed CBSD dynamics.

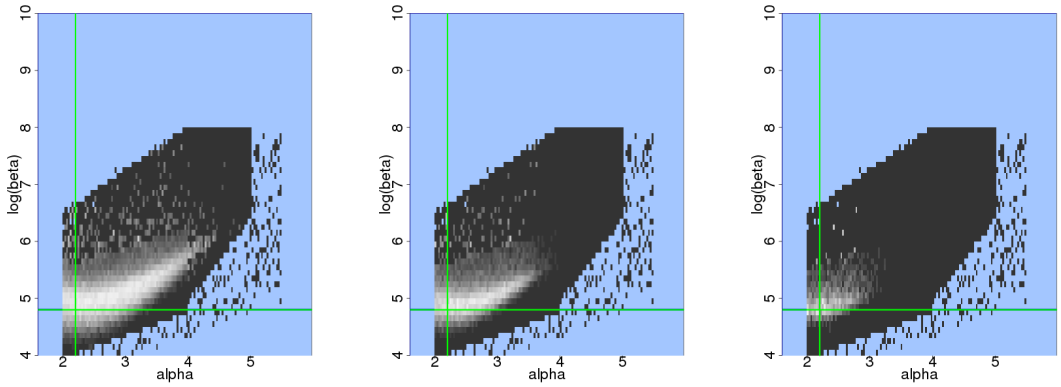
For 'model_1' and 'model_2', national and Kampala-region summary statistics were selected with a tolerance of ± 0.25. For 'model_3', the national-based and grid-based summary statistics were used in combination, to estimated model parameters for the real-world surveillance data. The tolerance applied for the national-based metric was ± 0.21 and for the grid-based metric, 0.5. Figure 4.17 shows the resultant posterior distribution for 'model_3' from approximately 200,000 simulations from the prior.



(a) ±0.3 tolerance (b) ±0.2 tolerance (c) ±0.1 tolerance

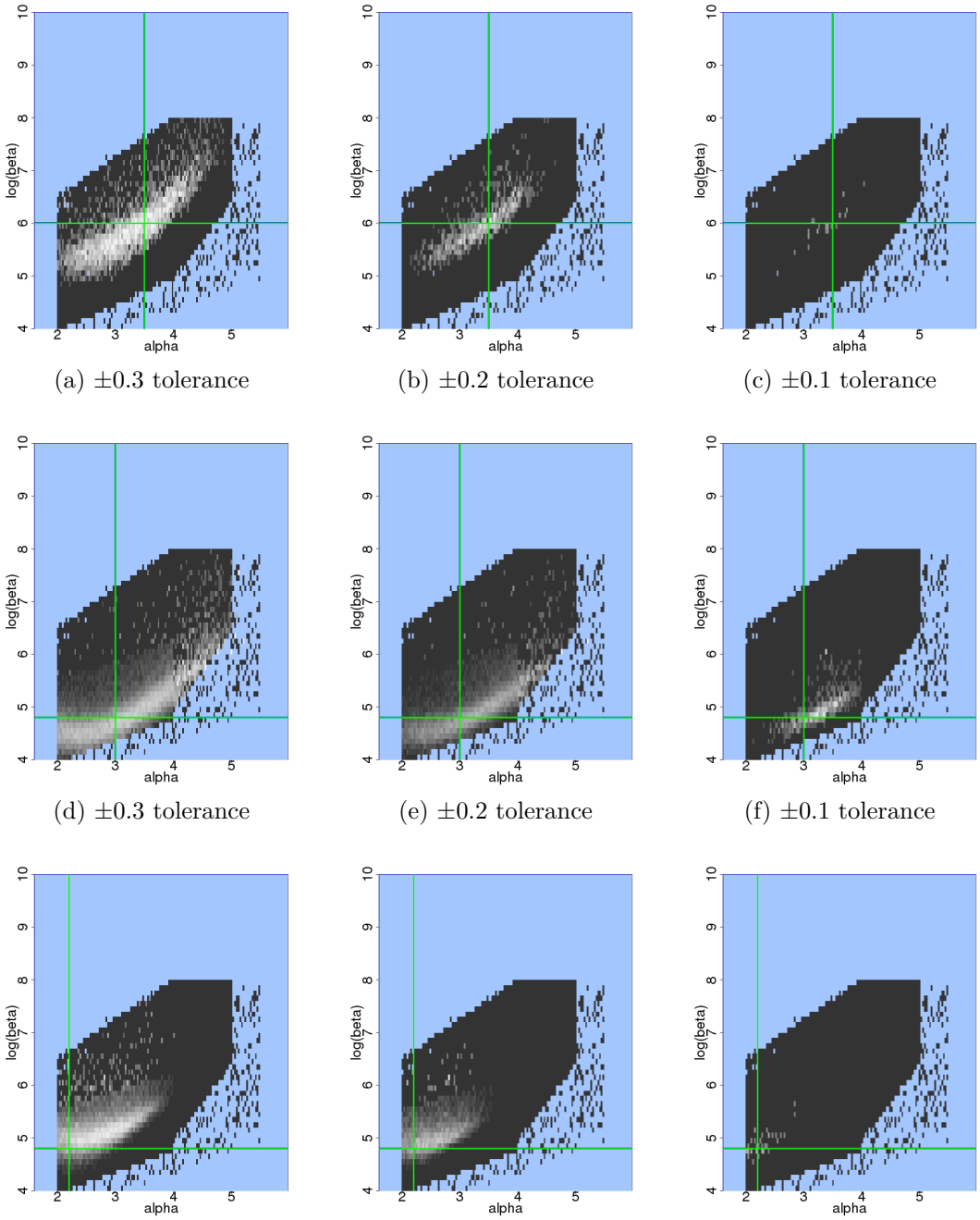


(d) ±0.3 tolerance (e) ±0.2 tolerance (f) ±0.1 tolerance



(g) ±0.3 tolerance (h) ±0.2 tolerance (i) ±0.1 tolerance

Figure 4.13: Recovering simulation parameters for three different input parameter sets using the national-based summary statistic. (a-c) is parameter set 0, (d-f) is parameter set 1, (g-i) is parameter set 2.

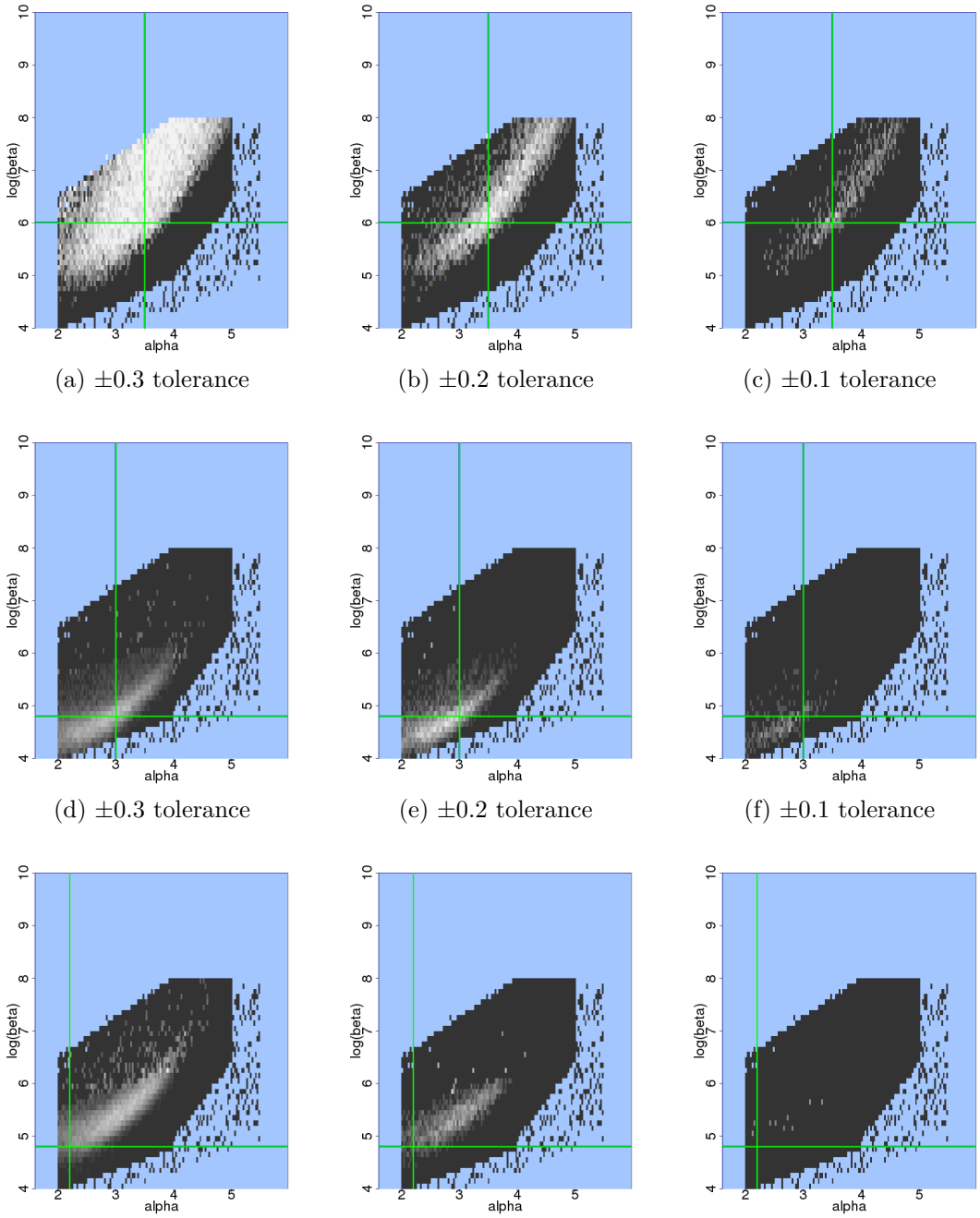


(a) ±0.3 tolerance (b) ±0.2 tolerance (c) ±0.1 tolerance

(d) ±0.3 tolerance (e) ±0.2 tolerance (f) ±0.1 tolerance

(g) ±0.3 tolerance (h) ±0.2 tolerance (i) ±0.1 tolerance

Figure 4.14: Recovering simulation parameters for three different input parameter sets using the region-based summary statistic. (a-c) is parameter set 0, (d-f) is parameter set 1, (g-i) is parameter set 2.

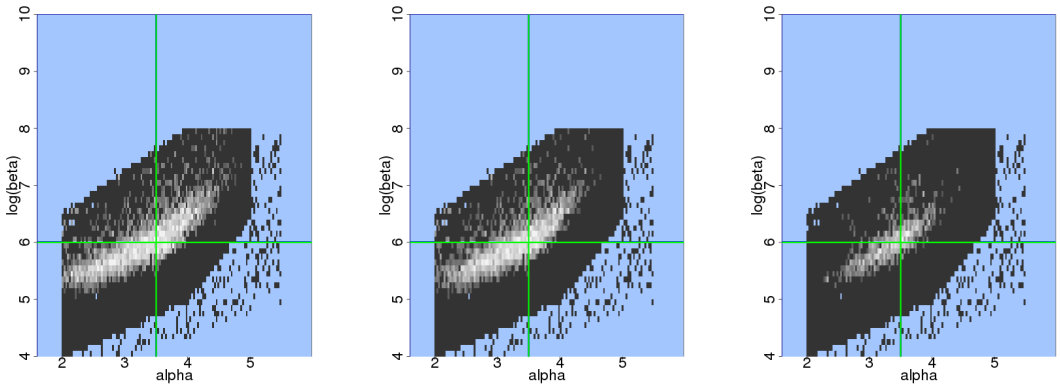


(a) ±0.3 tolerance (b) ±0.2 tolerance (c) ±0.1 tolerance

(d) ±0.3 tolerance (e) ±0.2 tolerance (f) ±0.1 tolerance

(g) ±0.3 tolerance (h) ±0.2 tolerance (i) ±0.1 tolerance

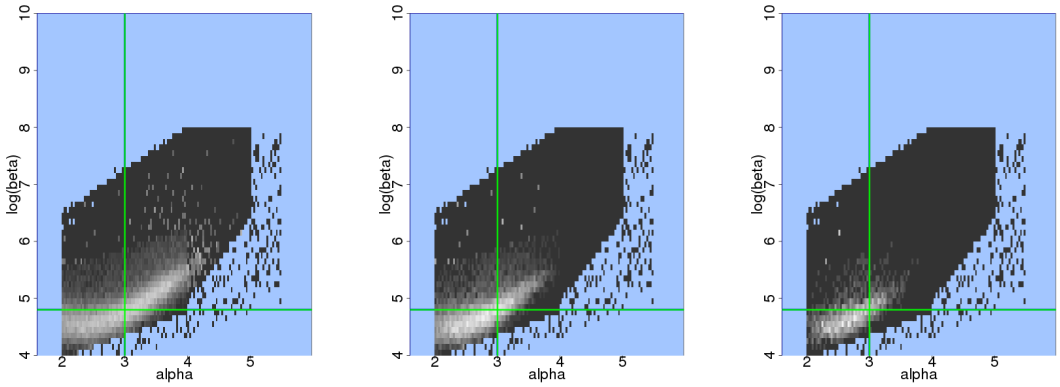
Figure 4.15: Recovering simulation parameters for three different input parameter sets using the grid-based summary statistic. (a-c) is parameter set 0, (d-f) is parameter set 1, (g-i) is parameter set 2.



(a) National-based tol ±0.25, grid-based tol 0.5.

(b) National-based tol ±0.25, grid-based tol 0.3.

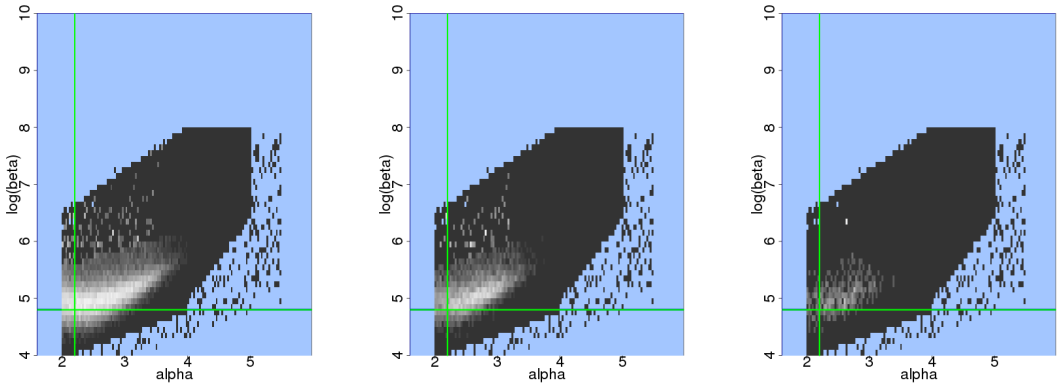
(c) National-based tol ±0.2, grid-based tol 0.2.



(d) National-based tol ±0.25, grid-based tol 0.5.

(e) National-based tol ±0.25, grid-based tol 0.3.

(f) National-based tol ±0.2, grid-based tol 0.2.



(g) National-based tol ±0.25, grid-based tol 0.5.

(h) National-based tol ±0.25, grid-based tol 0.3.

(i) National-based tol ±0.2, grid-based tol 0.2.

Figure 4.16: Recovering simulation parameters for three different input parameter sets using a combination of the national-based and grid-based summary statistics. (a-c) is parameter set 0, (d-f) is parameter set 1, (g-i) is parameter set 2.

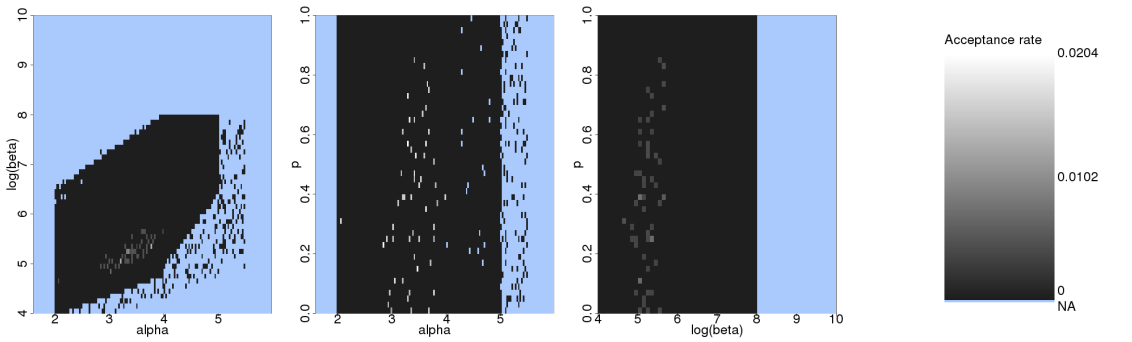


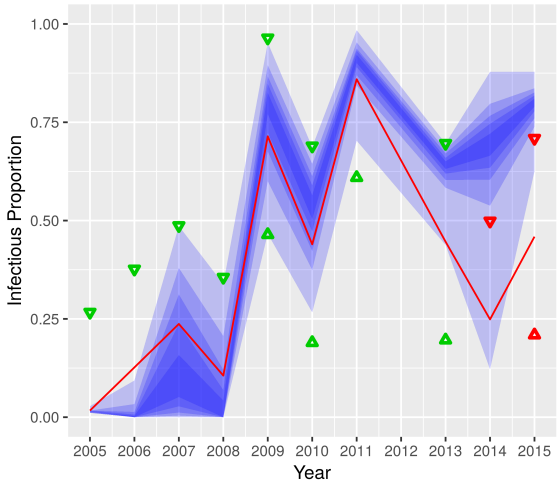
Figure 4.17: Posterior parameter estimates for 'model_3' with real-world survey data using the national-based and grid-based summary statistics at tolerances of ±0.25 and 0.5 respectively.

4.3.3 Validation simulations

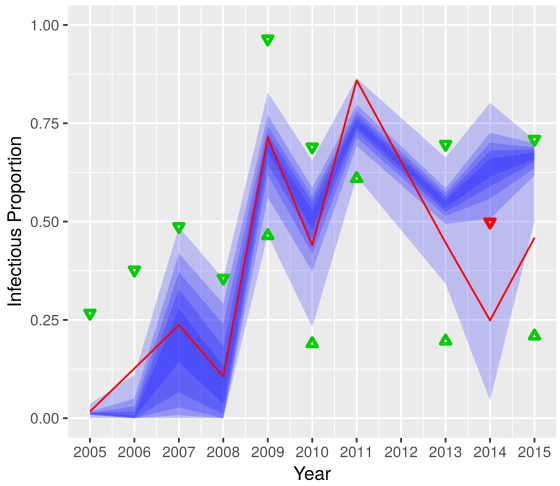
For each of the three models, 10,000 validation simulations from 2004 to 2015 were performed to assess model performance beyond the fitting period. No simulations from 'model_1' and 'model_2' passed the full validation period for the national infectious proportion statistic with a tolerance of ±0.25. Hence, constraints were removed on years 2014 and 2015 for 'model_1' and 2014 for model 'model_2' to observe the behaviour of the model in these years. For 'model_3', 67 of 10,000 validation simulations pass within a tolerance of ±0.25 during the 2005-2010 fitting period. Of the 67 candidate simulations that pass within this tolerance during the fitting period, 44 (65.7%) passed the validation period from 2011-2017 with a tolerance of ±0.25 for the national-based statistic. Figure 4.18 contrasts the performance of each model during validation.

Focussing on 'model_3', 4.19 highlights the behaviour of the simulations according to the both regional components of the national-based summary statistic in comparison with the real-world infectious proportion over the full period of 2005-2017.

Figure 4.20 illustrates the strong spatial correspondence between a single model realisation from the posterior of 'model_3' in comparison with the real-world surveillance data. The illustrations are representative of the behaviour of all simulations passing the fitting period. Moreover, surveillance data from Western Kenya, Tanzania and Rwanda is included where available (Chapter 3), which highlights the performance of the model in regions beyond Uganda.



(a) model_1



(b) model_2

(c) model_3

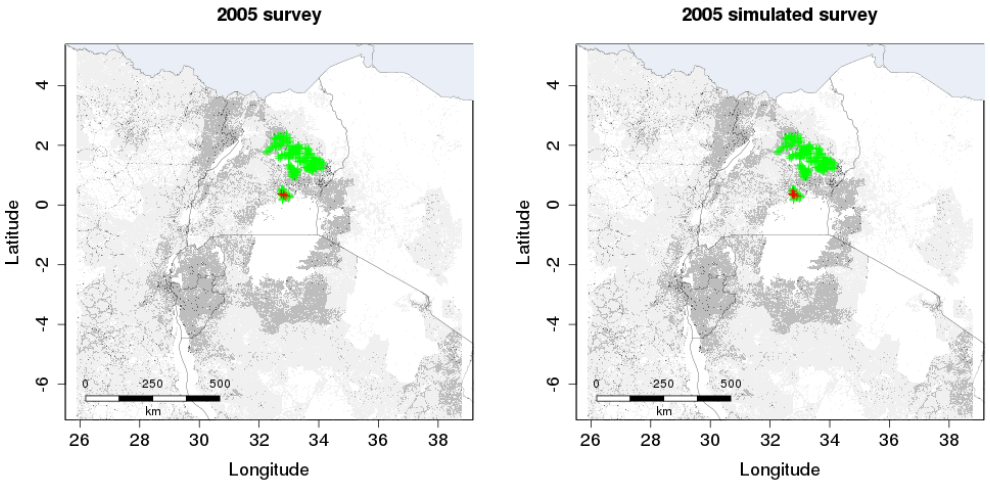
Figure 4.18: Comparison of the performance of all three tested model structures during validation. Plots summarise the predicted national infectious proportion in comparison with the real-world Ugandan infectious proportion (red line). Constraints in red for model_1 and model_2 are not imposed to prevent the elimination of all simulations.

(a) Kampala

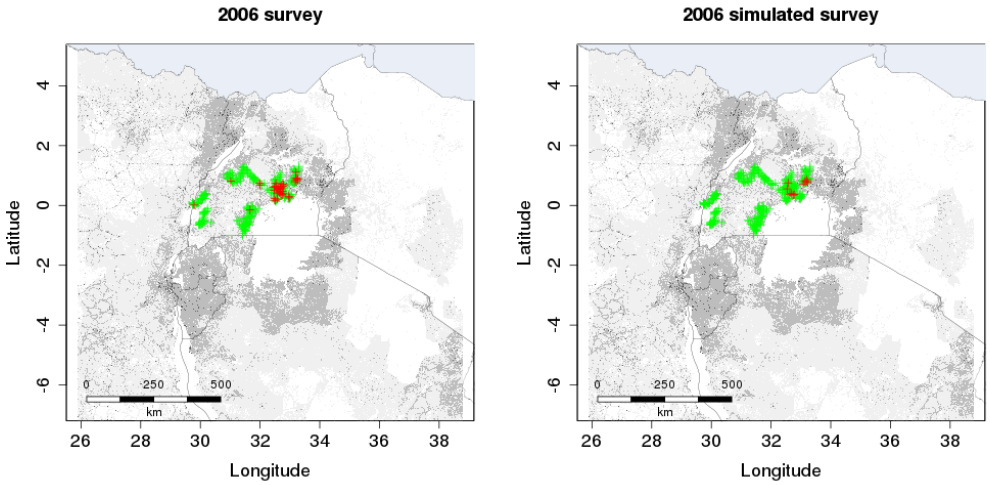
(b) National

Figure 4.19: Infectious proportion in the Kampala region and National region components of the National-based summary statistic for 'model_3' simulations that pass fitting and validation period at a tolerance of ± 0.25 for all years. The 90% percentile is indicated by the outer bounds of the blue band and increments in steps of 10%. Tolerances are indicated by green arrows. Red line indicates infectious proportion derived from the Ugandan surveillance data.

(a)



(b)



(c)

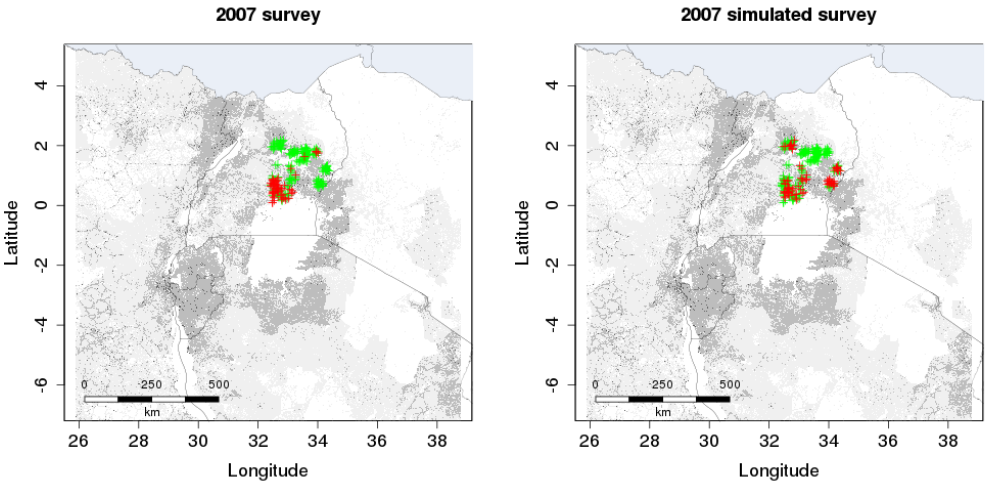
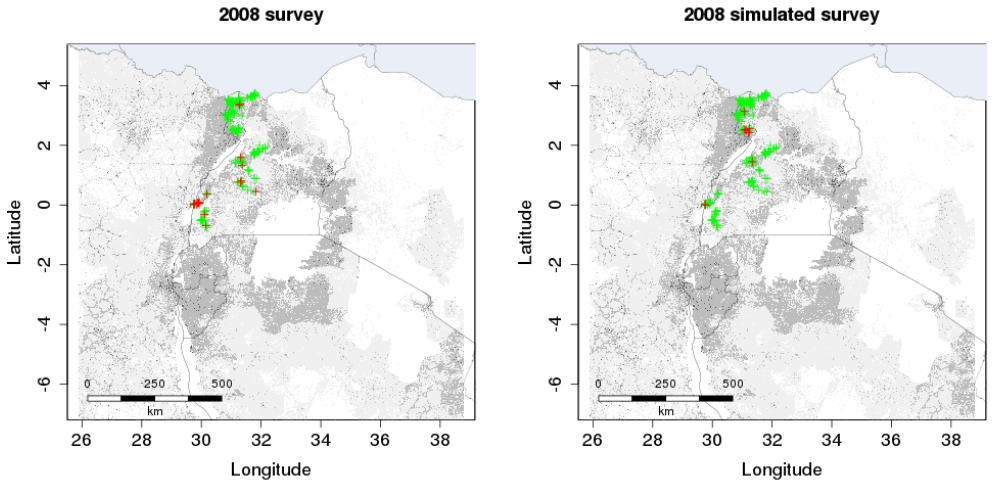
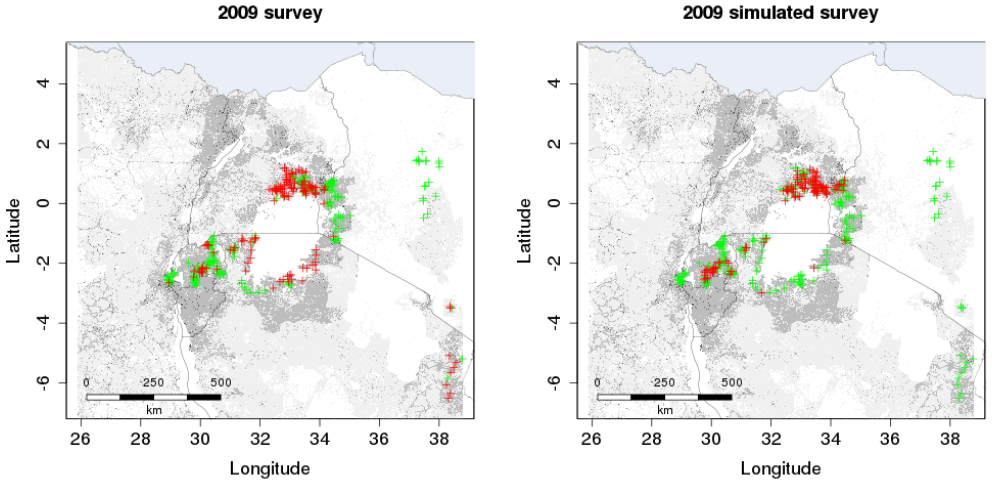


Figure 4.20

(d)



(e)



(f)

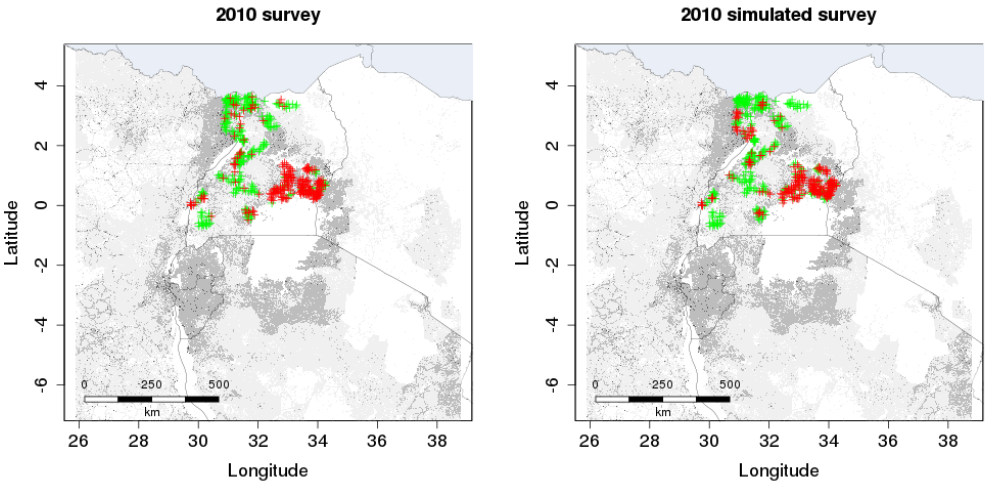
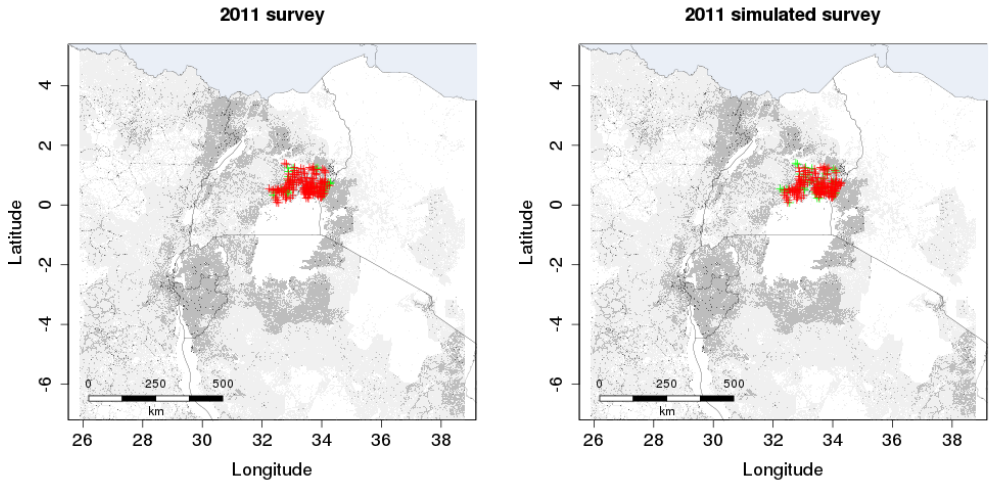
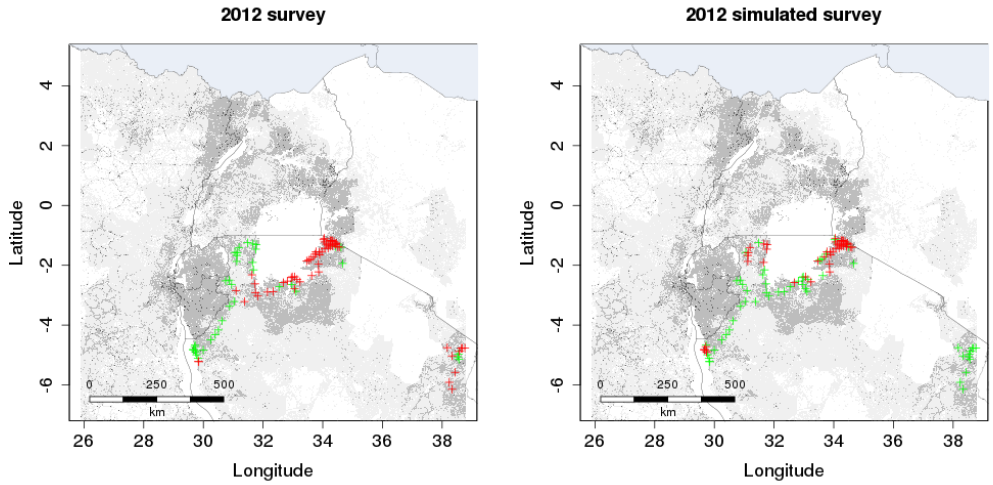


Figure 4.20

(g)



(h)



(i)

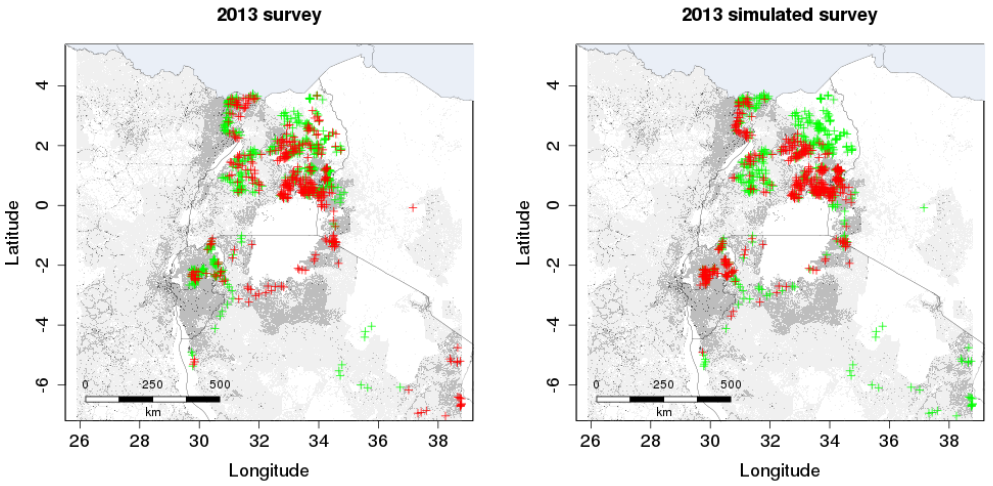


Figure 4.20

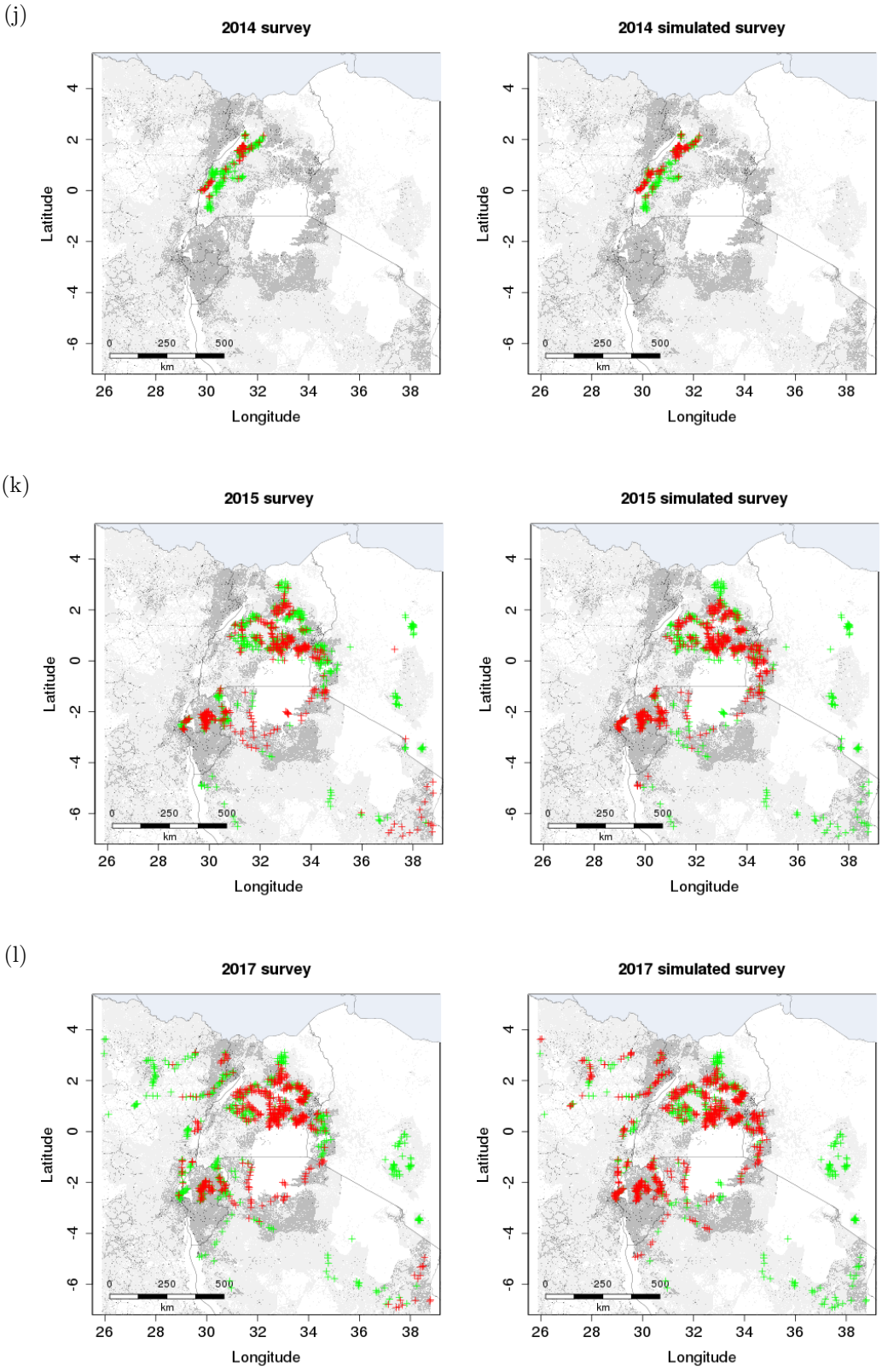


Figure 4.20: Comparison of a single validation simulation of 'model_3' that passes the fitting and validation criteria to the real-world surveillance data from 2005-2017. Red crosses indicate an observation of CBSD at the field-level in the real-world survey or simulated surveillance in the model. Green gross indicates no CBSD observed.

4.4 Discussion

We have outlined the development, parameterisation and validation of the first large-scale spatial epidemic model of cassava brown streak disease. This represents a major advancement in our ability to predict the future spread of the epidemic. During iterative model development and validation, we found significant improvements in the capacity of the model to capture the dynamics of the Ugandan epidemic when infectivity and susceptibility are modulated by whitefly abundance, which strongly indicates that vector abundance plays a role in the large-scale spread of the epidemic. This aligns with the consensus in the literature (Maruthi et al., 2005; Katono et al., 2015; McQuaid et al., 2017).

4.4.1 Real world implications

Currently, surveillance of CBSD is extremely limited relative to the scale of the epidemic and is largely absent in Central Africa, which contains the epidemic front. Moreover, there have been no attempts to spatially model or interpolate between point data from surveillance programmes to infer the wider epidemic distribution (Legg et al., 2011).

Efforts to prevent further epidemic spread in Central and West Africa will require governmental leadership. However, stakeholders and policy makers must allocate funds to address a wide range of issues. This is exacerbated in developing countries where resources are especially limited. In the absence of realistic estimates of the current epidemic distribution and anticipated rates of future spread, it is difficult to make an evidence-based case for the correct level of management action. The model presented in this chapter addresses this problem by enabling the exploration of historic, present and future epidemic distributions. Moreover, the modular components of the model, such as the vector abundance layer, can be improved and updated as more information or better methodologies are available. Therefore, the model provides a epidemiological framework to integrate, update, and test disparate sources of information about the CBSD epidemic.

Predictions of epidemic spread can significantly reduce the bounds of uncertainty for policy makers and allow them to make evidence-based assessments of the relative importance of different threats. In addition, the model allows hypothetical or exist-

ing management and surveillance strategies to be assessed (see Chapter 5), enabling limited resources to be more efficiently deployed.

4.4.2 Modelling context

Few parametrised large scale epidemiological models currently exist in the literature. Therefore, there is not yet a methodological consensus on the best parameter estimation approach. In the limited number of comparable studies, analytic likelihood MCMC has been used (Neri et al., 2014; Meentemeyer et al., 2011). However, for MCMC the likelihood becomes very costly to evaluate or may be intractable when the model is complex and the number of unobserved events in the data becomes larger. ABC overcomes this issue by approximating the likelihood function through multiple simulations of the stochastic model. To the authors knowledge, there are no large scale spatial plant epidemic models in the literature that have been parameterised when the likelihood is intractable.

The practical application of the ABC methodology has a number of sources of potential statistical error. When the tolerance is zero, ABC is functionally equivalent to analytic likelihood methods. However, for models of any complexity, it is necessary to introduce a tolerance and often summary statistics (Beaumont, 2010; Csilléry et al., 2010). The combination of informative, but not sufficient, summary statistics and a tolerance introduces a degree of error into the derived posterior. The parameter estimation methodology presented in this chapter addresses the limitations of ABC by investigating behaviour of proposed summary statistics on the recovery of known parameters using artificial data (Burr and Skurikhin, 2013). The subsequent performance of the model during validation supports the use of the selected summary statistics.

4.4.3 Future work

From a parameter estimation perspective, additional research is needed into the comparative benefits of ABC verses analytic likelihood methods. Specifically, in the context of a spatial epidemic models, a comparison of the distribution of posterior mass produced by different summary statistics in comparison with the exact likelihood posterior mass. Further research in this area has the potential to develop a canonical set of summary statistics to parametrise epidemic models a given spatial

scale.

The good correspondence between the model and Ugandan surveillance data during validation is reliant on the accuracy of the underlying host landscape and vector abundance layer. Therefore, the reliability of predictions for other regions, such as the DRC or Nigeria, depends on the quality of these data in the region in question. Future work should explore additional mechanisms of data generation, such as environmental suitability modelling. In addition, future data on the changing epidemic distribution can be incorporated into the model to update predictions. Therefore, surveillance teams and novel sources of surveillance data should be explored.

The CBSD epidemic is driven by two dispersal processes: cutting and vector movement. The model currently unifies these two processes into a single dispersal kernel, while conditioning for differences in vector density. Future attempts to model this system may parameterise distinct kernels for these two processes. However, our ability to distinguish these two dispersal processes effectively will rely on significantly more quantitative data on trade behaviour in sub-Saharan Africa. In addition, the current dispersal kernel is isotropic. A separate vector dispersal process could account for the anisotropic nature of wind-bourne dispersal (Burgin et al., 2013).

Chapter 5

Predicting the spread and management of CBSD

5.1 Introduction

Arguably, the most urgent intervention priority is to prevent the continued spread of CBSD through Central Africa into West Africa. This is because it is far more costly and logistically challenging to minimise the impact of an established epidemic than prevent establishment in the first place. Moreover, management programmes not only become more expensive, but can rapidly become economically prohibitive once an epidemic becomes too large (Cunniffe et al., 2016). In contrast, early action has a far higher probability of success and the comparative gains against future impact are at their largest.

However, there is good reason to be hesitant about investing in preventative measures, given the uncertainty around potential future threats, such as the spread of CBSD. This contrasts with current threats following invasion where there is greater certainty around their real-world impact. Allocating resources away from immediate needs to future threats is especially difficult in a highly resource limiting context, such as Central and West Africa.

Certain practices can minimise the challenges of preventative management. An important approach is to reduce the uncertainty around what is likely to happen in the future. Specifically, epidemiological modelling can be used to predict the spread, impact, and the efficacy of different interventions and surveillance programmes, tak-

ing account of uncertainty. The probability of management being implemented can be increased by identifying hybrid interventions that have present day benefits as well as fulfilling an important role in the preventative management of a potential future epidemic. Examples of these interventions are the development and improvement of infrastructure, such as seed systems and extension networks. These same infrastructural developments allow policy makers to address current problems for farmers, such as CMD, which is endemic in most cassava growing regions of West Africa, as well as improving the viability of a response to CBSD arrival.

None of this is to say that management in endemic regions is not important. However, in endemic regions, the level of funding allocated to CBSD management should be informed by its relative impact in comparison with other factors that limit production, such as managing CMD, access to markets, and access to fertilizers. Nonetheless, reducing uncertainty around which management strategies are most likely to be effective, and how best to monitor management programmes post-deployment can reduce the costs of management programmes and are readily addressed by modelling.

This chapter explores the application of the parameterised CBSD spatial model developed in Chapter 4 to a number of potential real world scenarios. These scenarios can be classified into two categories. The first is to investigate the historic dynamics of the epidemic in order to understand the behaviour of the pathosystem and to explain prior observations. The second category of scenario is to predict the future dynamics of the epidemic, with and without management. These scenarios are tailored towards addressing the urgent questions for preventative management in West Africa and Zambia, and reducing disease impact in the endemic regions.

5.2 Methodology

For each scenario, the model developed in Chapter 4 is simulated with parameters sampled from the final posterior distribution of the estimated parameters. The model simulates epidemics of CBSD on a rasterised landscape with a grid resolution of $1km^2$ with the field as the base host unit. The number of simulations per scenario is 1000, unless otherwise stated. Beyond this, the exact methodology varies depending on what needs to be investigated and details are provided in the relevant subsection. Figure 5.1 provides an overview of the region in question for each scenario presented in this chapter.

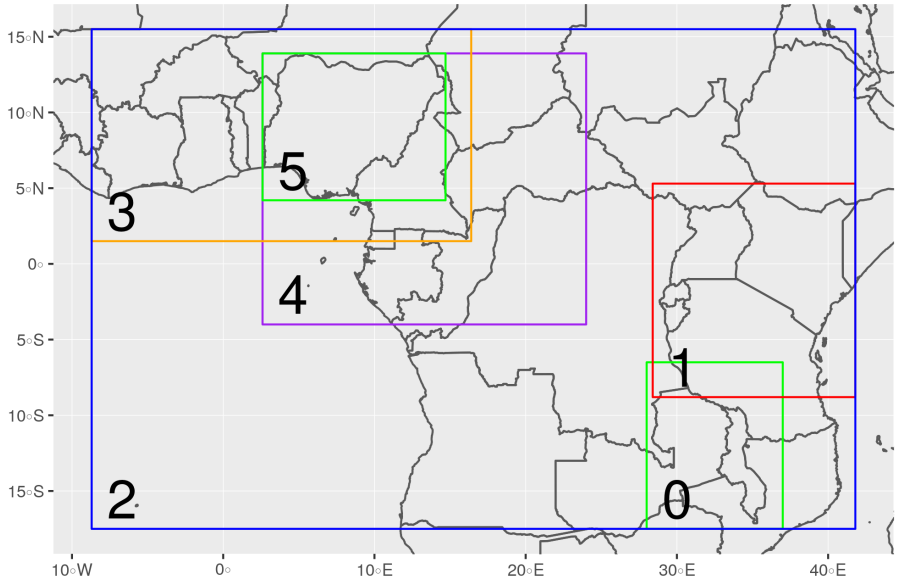


Figure 5.1: Overview of simulation extents for each scenario presented in this chapter. Extents are numbered in the bottom left and this number is referred to in the methods section of the relevant scenario.

5.3 Exploratory scenarios

5.3.1 Incursions from Malawi to Zambia

5.3.1.1 Background

CBSD was reported as widely distributed in present day Malawi since at least the 1950s (Nichols, 1950). Given that Malawi boarders Zambia, it seems surprising that CBSD has not spread and established in Zambia. We hypothesise that historically, CBSD will have spread from Malawi to Zambia. However, low vector populations, combined with low cassava density could have prevented the establishment and rapid expansion of the disease. Consequently, either by random chance or by active intervention by farmers, we would expect the disease to have been readily controlled. It is important to note that the Malawi and Zambia populations have rapidly increased since the 1960s. Malawi's population has grown from 3 million to 12.7 million between 1960 and 2009. Zambia's population increased from 3.5 to 14.4 million in the same period (World Bank, 2018). Therefore, the historic cassava density and rates of trade are likely to be far lower than in the modern day. It is unknown how significantly the abundance of *Bemisia tabaci* has varied over time in the region.

We first analyse the difference in spread rate of CBSD with present day levels cassava production in contrast to historic levels in approximately 1960. The aim of this comparison is to investigate whether changes in cassava production, driven by population growth, may have contributed to the historic absence of CBSD observations in Zambia. We do this by comparing the invasion dynamics for CBSD under the two scenarios of present-day density of cassava and reduced densities typical of the human population in the 1960s.

In addition, as of 2017, CBSD was reported to be spreading in the North of the Zambia (Mulenga et al., 2018). This region has far higher levels of cassava production than the eastern boarder with Malawi. The simulations with present day levels of production also serve to explore potential future spread in Zambia.

5.3.1.2 Methods

The model must be initialised with a distribution of infected CBSD fields. As we do not know the original sites of CBSD incursions and subsequent spread in Malawi, we have generated a credible distribution of infected fields to act as initial conditions (Figure 5.2). The extent of the simulation landscape is highlighted in extent number 0 in Figure 5.1 and Figure 5.2.

Taking this initial distribution of infected fields, two independent sets of simulations are run for 50 years with 1000 replicates. In the first set, we use the present day cassava host model as the input distribution of susceptible fields (Figure 4.2). We refer to this sub-scenario as 'present-host'. In the second set, we reduce the host density by a factor of four. This level of reduction is in line with the difference in total population for both Zambia and Malawi between the present day and 1960. We have assumed a linear relationship between human population and cassava production, hence the reduction has been made uniformly throughout the landscape. Host cells that drop below the equivalent of a single field are reduced to zero. We refer to this scenario as 'historic-host'.

As we are simulating with a hypothetical initial distribution of infected fields, we do not have a corresponding real-world year. Therefore, results for both are presented on the simulation timeline, $t = 0 \rightarrow 50$.

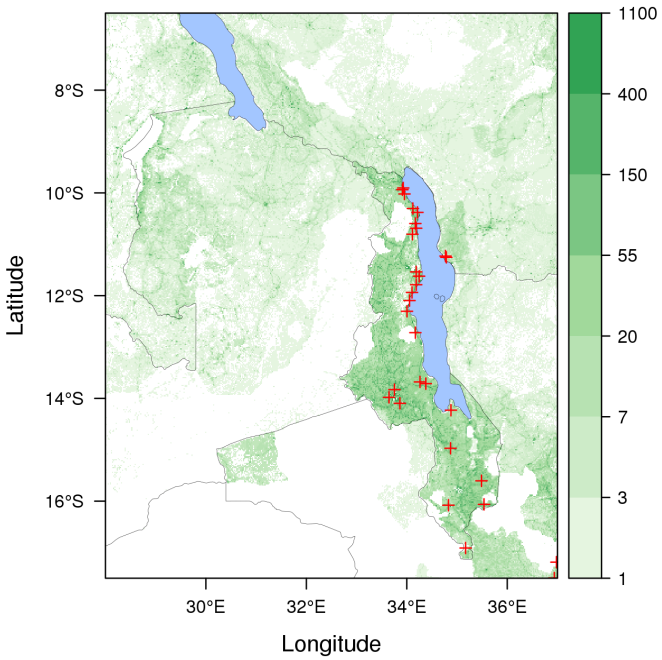


Figure 5.2: Map of initial sites of infection for 50 year simulations of CBSD spread from Malawi, indicated by red crosses. Based on 2012 surveillance data. Background represents the number of cassava fields per km^2 on a log scale, based on current estimates.

5.3.1.3 Results

Beginning with the current-host sub-scenario, Figure 5.3a shows on average, after 13 years, 25% of Malawi host cells with any host had at least one infected field. After 30 years, Malawi was reaching epidemic saturation with 75% of occupied host cells infected. Conversely, at the 50 year mark, the mean proportion of Zambia host cells infected after 50 years was 0.009%. Figure 5.3b highlights comparable trends in the proportion of fields that become infected.

Figure 5.4 illustrates a representative epidemic distribution after 50 years, while the proportion of the 1000 simulations that spread to a given quadrat after 50 years, which highlights the regions that are most frequently reached by the epidemic are summarised in Figure 5.5.

The simulations indicate that the sparse host density in eastern Zambia largely inhibits spread directly from Malawi into the bordering region of Zambia. However, the simulations do frequently predict epidemic spread along the western border of the northern region, through southern Tanzania, the DRC, and subsequently into northern Zambia. Notably, this corresponds to the region of Zambia where CBSD

has recently been observed for the first time (Mulenga et al., 2018).

Moving to the historic-host sub-scenario in which the host density has been reduced by a factor of four in line with the regional population levels in 1960. Figures 5.6 and 5.7 strongly contrasts with present day host spread rates, illustrating a much lower rate of spread within Malawi and near-zero spread into and within Zambia. Comparing figure 5.5 with 5.7 shows how the epidemic is geographically confined with a lower host density.

5.3.1.4 Conclusion

These simulation results suggest that the low historic density of cassava in the region results in significantly lower rates of spread CBSD spread within Zambia than predictions under current cassava host density. These two sub-scenarios represent lower and upper bounds for the rate of regional CBSD spread in the period between 1960 and the present day.

Our attempts to recreate historic epidemic conditions are conservative in that they do not include a number of additional factors beyond host density that would further contribute to reduced epidemic spread. For example, an additional spread constraint that is not incorporated into the model is the regional cutting trade dynamics. The structure of trade networks is generally very local, with growers predominantly re-planting from their own fields or sourcing cuttings from neighbours (Kansiime, 2014; Teeken et al., 2018). However, longer range movement of cuttings can infrequently occur, and is often associated with cultural factors, such as marriage exchanges (Delêtre et al., 2011). Given the historic near absence of mechanised transport, we can reasonably assume that the frequency of longer range cutting movement was radically lower than the present day.

In addition, data from the cross boarder trade of major food commodities suggests that trade from Malawi to Zambia is especially low (World Food Programme, 2005). Combined, these factors would further reinforce the inhibition of spread resulting purely from lower host density.

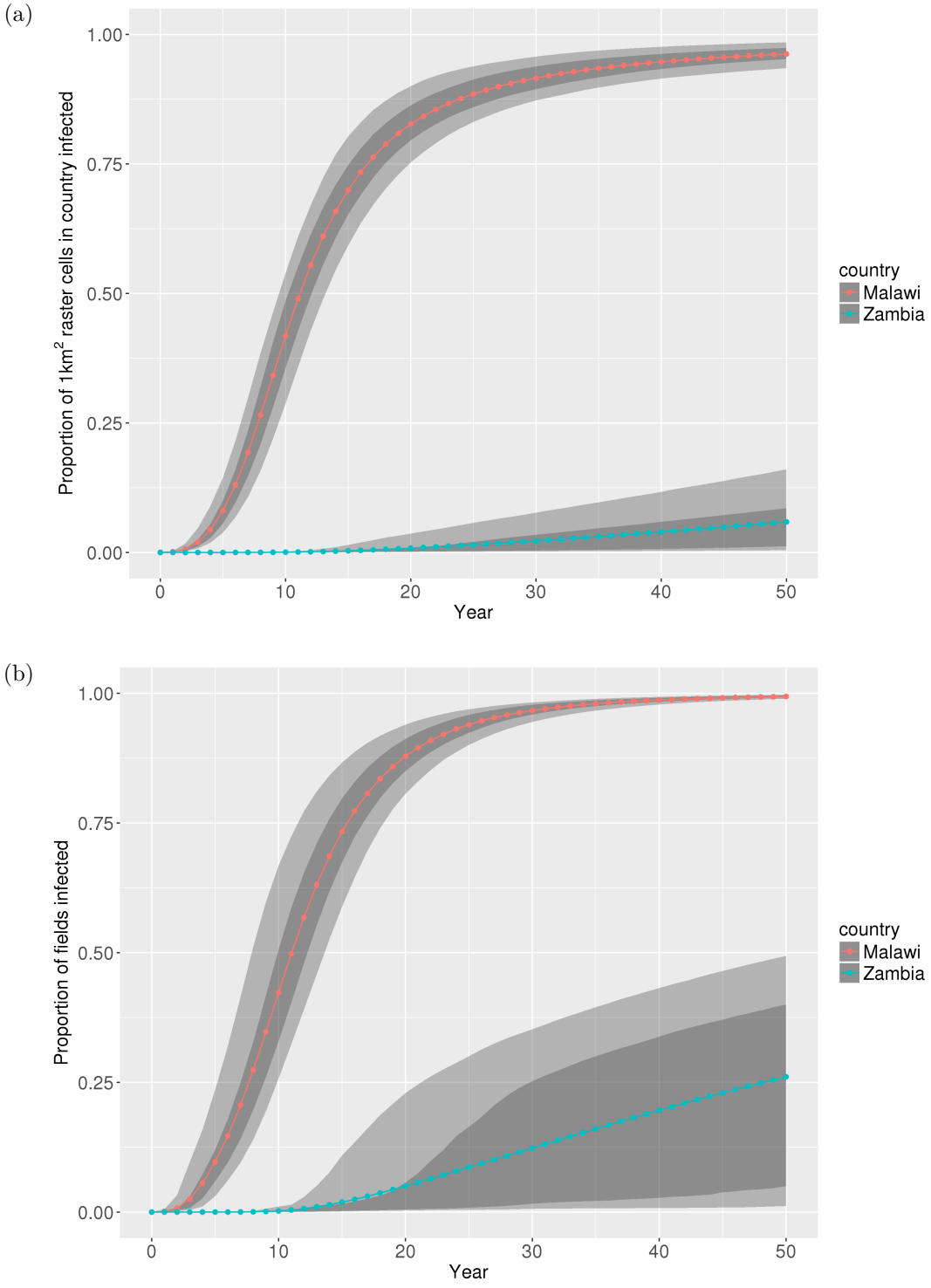


Figure 5.3: Present-host sub-scenario: Disease progress curves for Malawi and Zambia over a 50 year period (a) as a proportion of 1km^2 host raster cells containing at least one infected field and (b) as a proportion of fields in the country that are infected

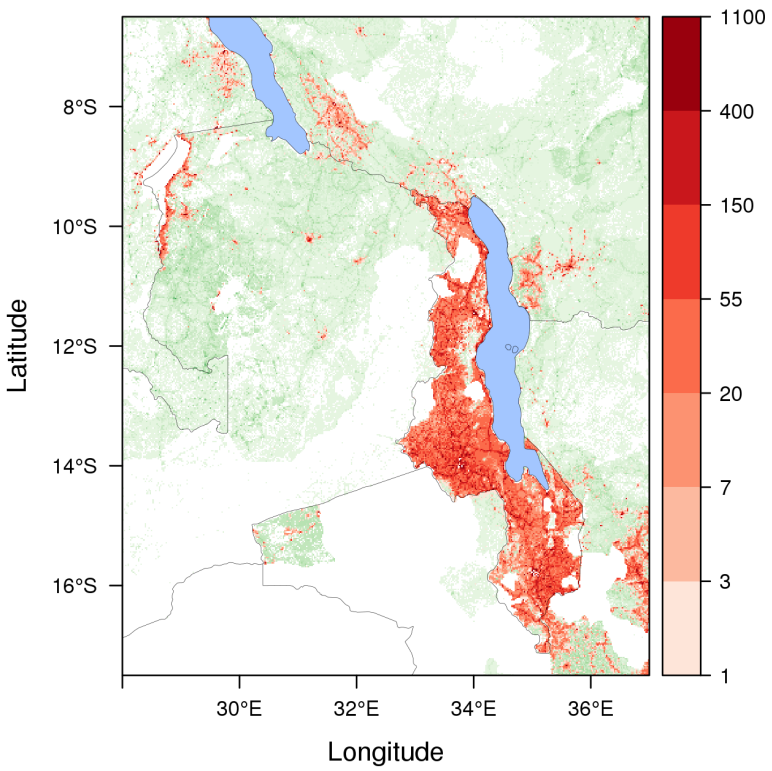


Figure 5.4: Present-host sub-scenario: Representative simulation after 50 years showing the key regions of spread. Infection is shaded red on a log scale corresponding to the absolute number of infected fields per km^2. The green background scale represents the host density as in Figure 5.2.

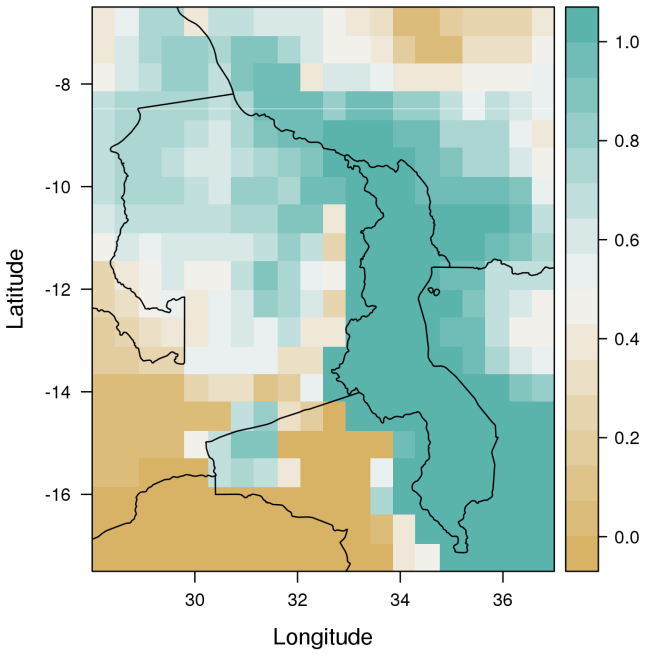


Figure 5.5: Present-host sub-scenario: Map in which each quadrat (25x25km resolution) is shaded according to the proportion of simulations that result in any infection in each quadrat after 50 years.

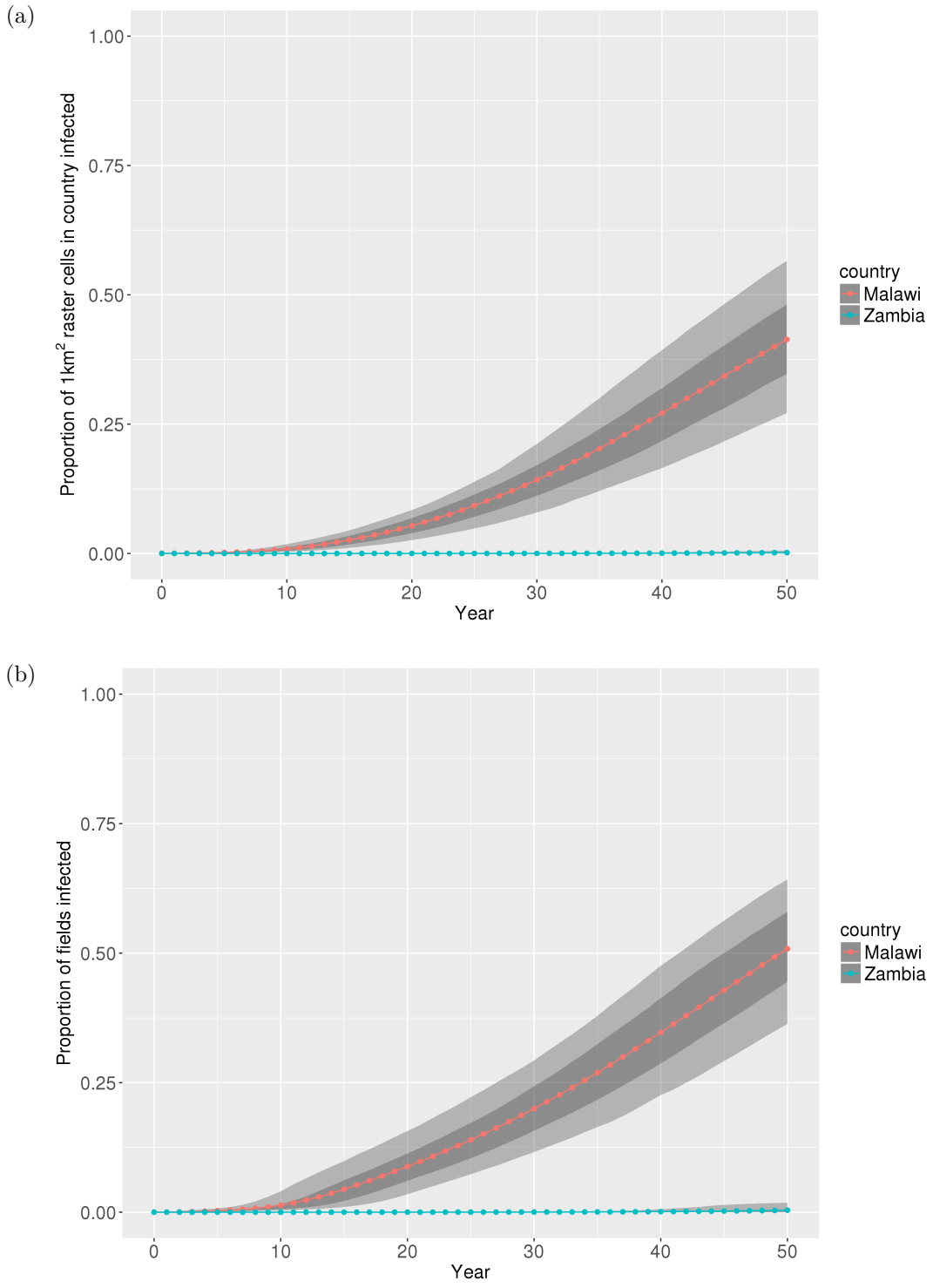


Figure 5.6: Historic-host sub-scenario: Disease progress curves for Malawi and Zambia over a 50 year period (a) as a proportion of 1km^2 host raster cells containing at least one infected field and (b) as a proportion of fields in the country that are infected

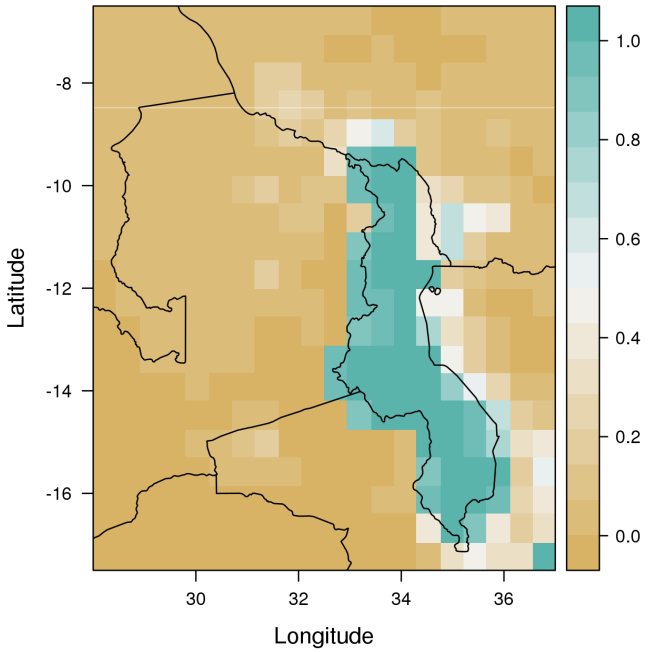


Figure 5.7: Historic-host sub-scenario: Map in which each quadrat (25x25km resolution) is shaded according to the proportion of simulations that result in any infection in each quadrat after 50 years. Cassava host density has been reduced by a factor of four in line with the approximate regional population in 1960.

5.3.2 Spread from endemic coastal region to Uganda

5.3.2.1 Background

The source of the CBSD inoculum that started the post-2004 Ugandan epidemic is not known (Alicai et al., 2007). The most probable mechanisms of introduction are human-mediated movement of infected cuttings from the endemic coastal region, or vector-mediated introduction from wild alternative hosts of CBSVs (Legg et al., 2011).

It is impossible to predict with certainty the original mechanism of introduction. However, model simulations allow us to explore the relative rates of introduction from the endemic coastal region to Uganda under different historical conditions, such as changing levels of cassava cultivation. This can provide a probabilistic insight into why the epidemic emerged in 2004.

5.3.2.2 Methods

As in the previous scenario exploring spread from Malawi to Zambia, we do not know the historic distribution of CBSD infected fields along the coastal region of Kenya and Tanzania. Therefore, we have generated a credible distribution of infected fields in this region (Figure 5.8). The extent of the simulation landscape is highlighted in extent number 1 in Figure 5.1. Simulations represent 50 years of spread from the initial sites of infection to investigate the long dynamic term behaviour of the pathosystem. As the epidemic bulks up from the initial sites of infection, the system reproduces an approximation of the endemic state of the coastal region. After the region has reached endemic levels, there is a significant disease pressure exerted on surrounding regions. This section contrasts the rate of incursions from the endemic region to Uganda under present day host densities with rate of incursion when the host density is reduced in line with the regional population in 1960.

5.3.2.3 Results

Figure 5.9 highlights the disease progress curves of 50 randomly sampled epidemics over the 50 year period. Approximately 20% of realisations under modern day host conditions do not result in any incursions into Uganda over a 50 year period (Figure

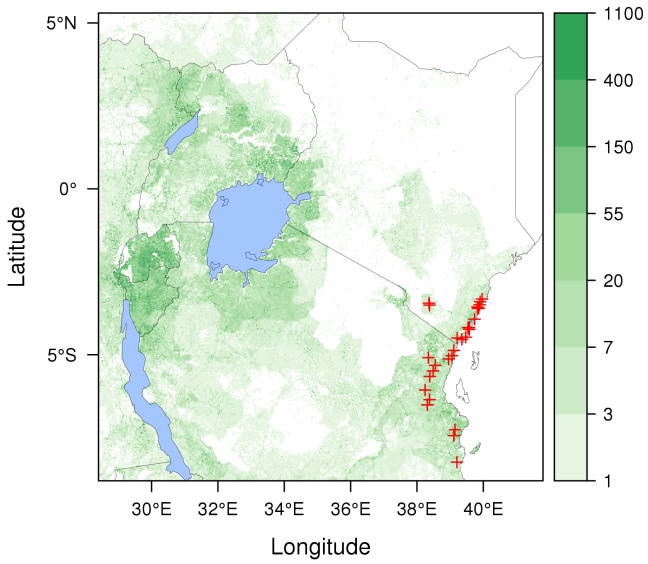


Figure 5.8: Map of initial sites of infection for 50 year simulations of CBSD spread from coastal Kenya and Tanzania, indicated by red crosses. The initial conditions, based on 2009 surveillance data, are for the endemic levels of CBSD infection. Background represents the number of cassava fields per km^2 on a log scale.

5.11a). Of the 80% of simulations that do reach Uganda, the distribution of arrival times is shown in Figure 5.10. The earliest arrival time is 7 years, with a mean of 25.6 years.

As with the previous scenario for spread from Malawi to Zambia, we then reduce the host density by a factor of four, in line with the regional population in 1960. The 50 year simulations were repeated and resulted in significantly lower spread rates. With the lower host density, 96.5% of simulations do not spread to Uganda after 50 years (Figure 5.11b).

5.3.2.4 Conclusion

Since 1960, we can hypothesise that as population density, and hence cassava cultivation, has rapidly increased, the yearly probability of incursions into Uganda has also increased significantly. This hypothesis is strongly supported by contrasting Figure 5.11a with 5.11b. The probability of an epidemic occurring under present day conditions is highly likely.

An outbreak of CBSD was reported in Uganda in the 1950s, however, control was successful (Nichols, 1950; Alicai et al., 2007). The significantly lower historic spread

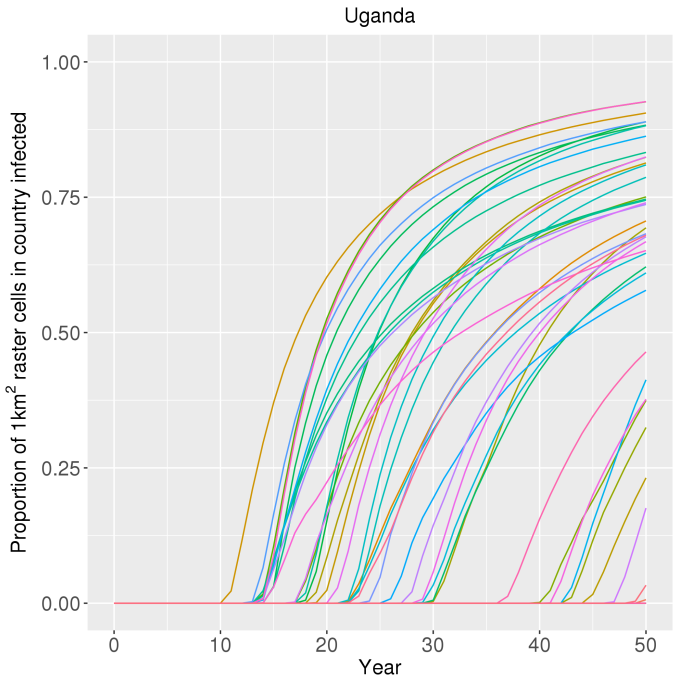


Figure 5.9: Present-host density: A random sample of 50 disease progress curves for spread from coastal East Africa to Uganda over a 50 year period in terms of proportion of 1km^2 host raster cells in Uganda that containing at least one infected field.

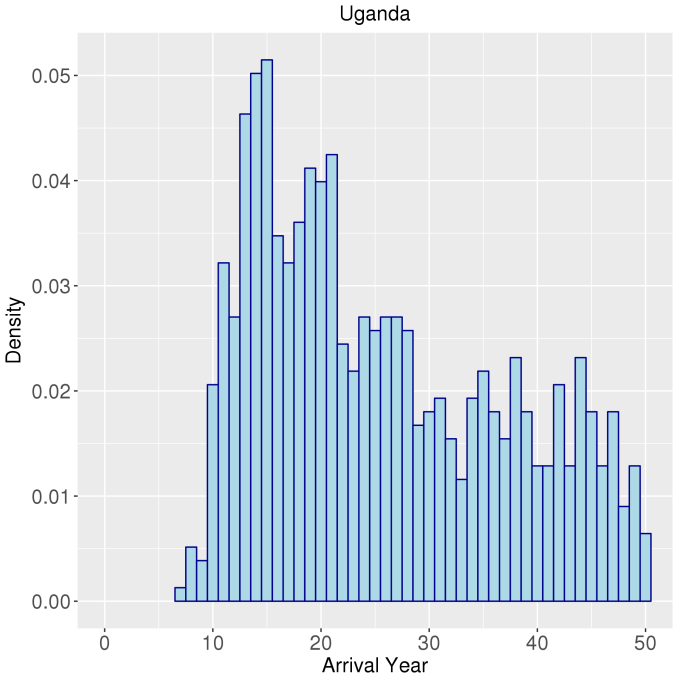


Figure 5.10: Present-host density: A histogram illustrating the distribution of first years of disease presence in Uganda, conditional on those that do spread to Uganda.

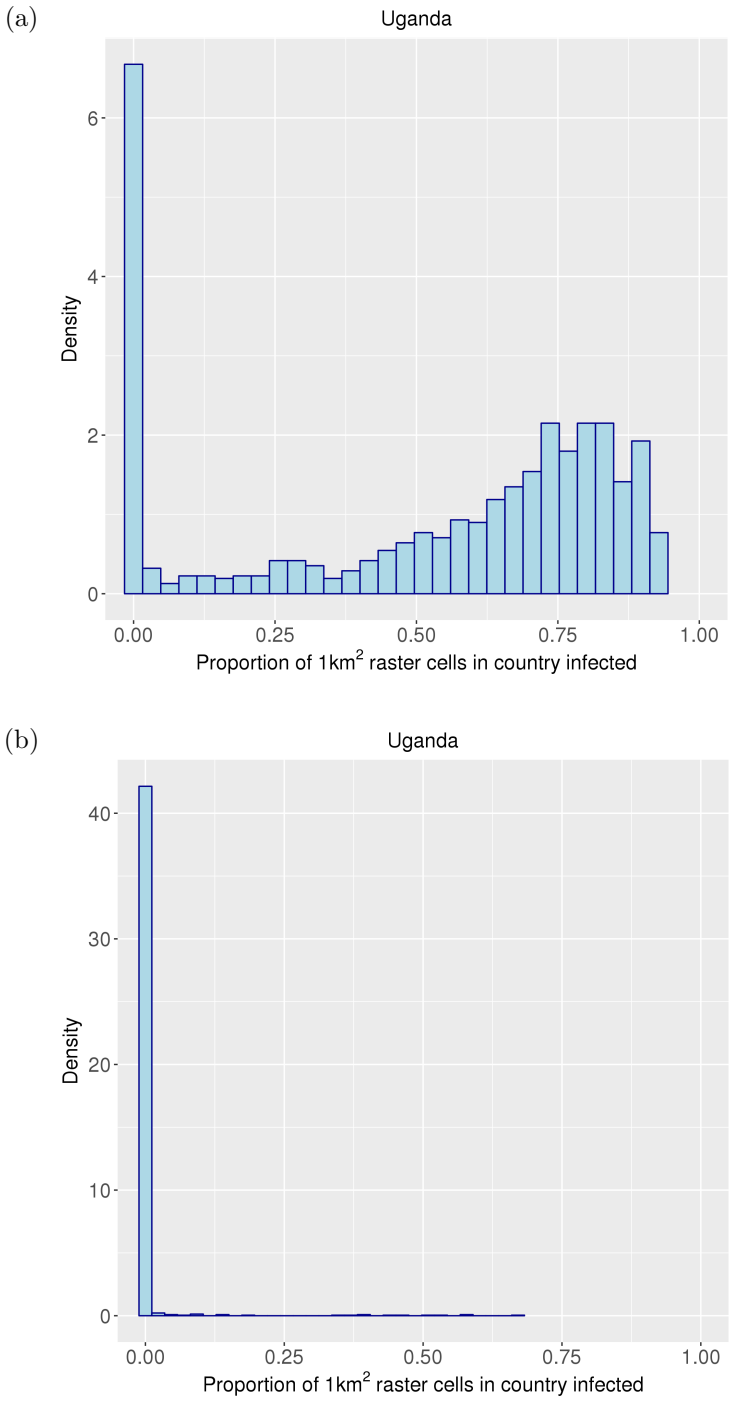


Figure 5.11: (a) Present-host density: Histogram summarising the proportion of 1km^2 host raster cells in Uganda that contain at least one infected field after 50 years of spread from the endemic coastal region across 1000 simulations. (b) Historic-host density: Equivalent histogram with cassava host density reduced by a factor of four in line with the approximate regional population in 1960.

rates, illustrated in Figure 5.11, may have made control far more viable than under present conditions. In addition, lower historic vector abundance has been reported which may have further benefited management success (Macfadyen et al., 2017). However, limited quantitative data on the extent to which vector populations have changed makes informed incorporation into simulations challenging.

In conclusion, it seems likely that increased regional cassava density has facilitated the post-2004 Ugandan epidemic of CBSD. Subsequent within-country spread rates, and consequently the difficulty of management, may have been further exacerbated by higher vector abundance in recent years, but the extent to which the vector populations have changed is not known.

5.4 Predictive scenarios

5.4.1 Spread towards West Africa

5.4.1.1 Introduction

Cassava brown streak disease has been spreading from Uganda since 2004. The primary management priority is preventing spread to West Africa, where cassava reliance and production is especially high and CBSD is not currently present. Visual symptoms of CBSD were reported in Eastern DRC in 2009 and confirmed by molecular diagnostics in 2011 (Legg and Bouwmeester, 2010; Mulimbi et al., 2012). In 2017, visual CBSD symptoms were reported from surveillance in the North-west corner of the DRC, but to date there has been no molecular confirmation. Similarly, CBSD is reportedly present but unconfirmed in Angola as of 2018 (Personal communication, J. Pita, 2018).

Surveillance in the DRC has been extremely limited. Therefore, we do not have a good impression of the current epidemic distribution. Moreover, our confidence in the CBSD surveillance data from DRC is complicated by a number of factors, such as the absence of molecular confirmation. In this section, we first outline the factors that affect the credibility of CBSD observations in North-west DRC in 2017. Based on this assessment, we propose upper and lower bound scenarios to predict the spread rates of the CBSD epidemic towards West Africa. In the lower bound scenario, we assume the north-west DRC symptom observations are false positives.

In the upper bound scenario, we then analyse spread rates under the assumption that CBSD did reach north-west DRC by 2017.

It is important to emphasise that the following predictions are what we term the 'expanding-spread' scenario. By this we mean the simulations do not account for the direct introduction of infected material to West Africa, such as via air travel or shipping. The dispersal process in the model is comparatively local, and simulates human-mediated and vector driven local spread, such as the epidemic progress observed in Uganda. However, the landscape scale dynamics of the epidemic could be radically disrupted by the direct introduction of infected material to West Africa, such as someone flying with cassava cuttings or long distance trade. Therefore, we first explore the expanding-spread spread rate we would expect to observe in the absence of these direct events. In Section 5.4.3, we explore the consequences of a direct introduction to West Africa on spread and management.

5.4.1.2 Probability of false positives in DRC surveillance data

Unlike surveillance in Uganda, very little surveillance has taken place in the DRC. Therefore, we do not have empirical observations of the changing multi-year spatial distribution of the epidemic in the DRC. In addition, no samples from the 2017 survey in which visual symptoms were observed in the north-west have been tested using molecular diagnostics for logistical reasons. Therefore, the nearest sites in which visual symptoms have been confirmed by molecular diagnostics are located approximately 1000km away from north-west DRC, near the Eastern border of DRC (Mulimbi et al., 2012; Casinga et al., 2018). The absence of spatiotemporal spread from multi-year surveillance and any form of molecular diagnostics means we cannot be sure that the symptoms observed during surveillance are caused by CBSV infection.

5.4.1.2.1 Virus introduction from wild host

We must also consider the possibility that CBSV in this region has occurred independently of the post-2004 epidemic. For example, we know that CBSV is found and likely originated in wild host plants (Gwandu, 2014). Depending on how widely CBSVs are distributed in wild host plants, it is not inconceivable that an independent vector-mediated introduction from wild host into cassava could take place in other regions, especially in regions where vector populations have markedly increased or

adapted different feeding behaviours. Nonetheless, the historic absence of CBSVs beyond the endemic region of East Africa suggests that introductions from wild host occur with a very low probability. Moreover, molecular diagnostics from East Africa indicate that CBSD has spread at least 460 km from Uganda between 2004 and 2009. Therefore, epidemic spread of 1480km in 13 years to the north-west of DRC does not seem unrealistic if we assume spread rates in East Africa are indicative of spread rates in the DRC.

5.4.1.2.2 Long range cutting movement

An alternative possible cause of infection in north-west DRC is that it may be the result of a long-distance human mediated introduction of CBSV infected material from the endemic region. However, given that cassava is grown disproportionately by poor farmers, which is especially true in the DRC, long range movement of planting material in this region seems unlikely. Moreover, road conditions in the DRC are notoriously poor, which is also likely to reduce the probability of long range movements. Nonetheless, the impact of the direct introduction of infected material to regions of West Africa is investigated in Section 5.4.3.

5.4.1.2.3 'CBSD-like' symptoms in Bas-Congo

Since 2002, symptoms described as 'CBSD-like' have been reported in the South-west region of Bas-Congo, DRC (Mahungu et al., 2003). This is a phenomenon believed to be unique to the DRC that increases the probability of CBSD false positives from visual surveys.

However, a number of factors mitigate the probability that the CBSD symptoms in the north-west are false positives resulting from the 'CBSD-like' phenomenon. Firstly, despite a number of attempts to diagnose the phenomenon using molecular diagnostics, none have been successful. Therefore, there is a lack of evidence that CBSV, or any infectious agent is associated with the symptoms observed in Bas-Congo, hence, minimising the probability that the phenomenon is capable of spreading. Secondly, a distance of approximately 1000km separates the Bas-Congo from the observations in the north-west of the country, which further minimises the probability that CBSV symptoms observed in the north-west are in fact caused by the same unidentified process causing 'CBSD-like' in the south-west of DRC.

We conclude that there is a relatively low probability that the CBSD foliar and root

symptoms observed in the north-west of the country are associated with the 'CBSD-like' phenomena in south-west DRC. Nonetheless, regardless of the causative agent, understanding and minimising the impact of 'CBSD-like' should a be priority in its' own right.

5.4.1.2.4 DRC false positives: Conclusion

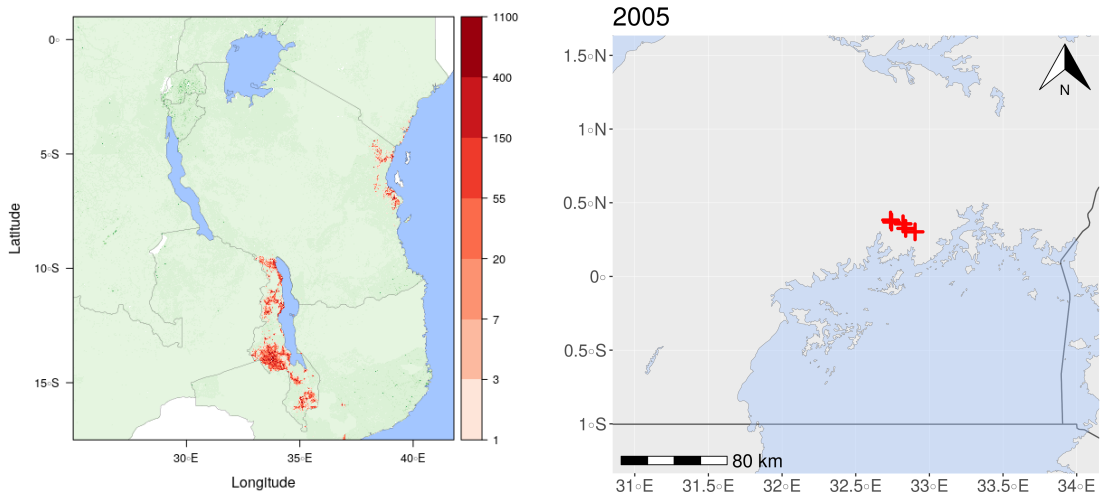
The set of complicating factors presented in the preceding section emphasises the importance of treating the DRC observations with caution. We therefore simulate cross-continental spread rates with and without assuming the CBSD symptoms observed in north-west DRC are correct. These simulations provide upper and lower bound estimates for the continued spread of the epidemic from Uganda towards West Africa.

5.4.1.3 Methods and results

Simulations are initialised with infected fields using the first recorded CBSD positive observations near to Kampala from the 2005 Ugandan surveillance data. We assume these points are representative of the infection state of Uganda in 2004. In addition to the Ugandan sites of infection, the historically endemic regions are also seeded with infection. To approximate the endemic status of these regions, a representative simulation at the 10 year mark were selected from the previous Malawi and coastal endemic scenarios. Figure 5.12 illustrates the initial distribution of infected fields, cropped to highlight the endemic regions.

The full extent of the simulation landscape is highlighted in extent number 2 in Figure 5.1. Simulation year zero corresponds to 2004. Approximately 1800 simulations were then run for 50 years, from 2004 to 2054. The first 14 years of the simulation, from 2004 to 2019, generate spatiotemporal spread that has generated the present day distribution of the epidemic.

As the model is stochastic, each of the 1000 simulations represents a different predicted spatiotemporal epidemic trajectory. We can use real-world information about the spread of CBSD post-2004 to isolate those simulations that are following a similar trajectory. In this way, we can reduce the diversity of predicted epidemic trajectories. There are two pieces of information we use to constrain the simulations:



(a) (b)

Figure 5.12: Map highlighting (a) the initial number of infected fields in the endemic region on a log scale and (b) the sites of the infected fields observed in southern Uganda in the 2005 survey. We assume the distribution of infected fields is representative of the real-world distribution in 2004.

- 2005-2017: Ugandan yearly foliar CBSD surveillance data.
- 2017: Foliar and root symptoms from surveillance in north-west DRC.

We analyse predicted spread rates towards West Africa under two different sets of assumptions. The first sub-scenario produces model predictions that correspond to the 2005-2017 Ugandan surveillance data. This represents a lower-bound estimate of 'expanding-spread' rates. For the second sub-scenario, in addition to the Ugandan data constraint, we also make the assumption that the CBSD observations in north-west DRC are accurate and result from the post-2004 spread from Uganda. This sub-scenario represents an upper bound estimate of 'expanding-spread' rates.

5.4.1.3.1 Constraining on Ugandan surveillance data

Figure 5.13 illustrates the yearly infectious proportion of fields across all available years of the Uganda surveillance data, along with the bounds of a ± 0.25 deviation from the real-world infectious proportion values. A total of 1170 of 1800 simulations fall within the ± 0.25 envelope of the yearly infectious proportion of surveyed fields in Uganda between 2005 and 2017. Figure 5.14 summarises the predicted time of epidemic arrival for a set of key countries and regions. Notably, the median predicted year of arrival in Luapula Province in northern Zambia is 2017, which corresponds

Figure 5.13: Yearly infectious proportion of fields across all available years of the Uganda surveillance data, along with the summary statistic tolerances (green ticks) of a ±0.25 deviation from the real-world infectious proportion values. Tolerances are not plotted where they exceed the 0-1 bounds of a proportion.

with the first confirmed report of CBSD symptoms in the region (Mulenga et al., 2018). Despite many outlier simulations arriving in Nigeria as early as 2024, the median arrival year is not captured in these simulations as it lies beyond 2054.

In addition to the time of arrival, the rate at which the epidemic then spreads within a region depends on the density and spatial structure of the host landscape and relative whitefly abundance. Figure 5.15 illustrates the predicted year in which a given country or region is expected to exceed 25% of fields infected with CBSD, which gives some insight into the predicted post-introduction dynamics of the epidemic.

Based on this criterion, we estimate the probabilistic present day CBSD epidemic distribution, as of 2019. Figure 5.16 is a risk map summarising the epidemic distribution across all 1170 simulations that meet the observed real-world Ugandan infectious proportion. The risk map is generated by taking the raster representing the distribution of infected fields in each simulation in the year 2019. Each raster of the predicted 2019 distribution of infected fields is converted to represent binary presence and absence of infection at each $1km^2$ cell in the landscape. The binary rasters across all simulations are then added together and the resultant raster is divided by the total number of simulations. Applying the same methodology, we generate a risk map for 2050 for the expanding-spread epidemic distribution.

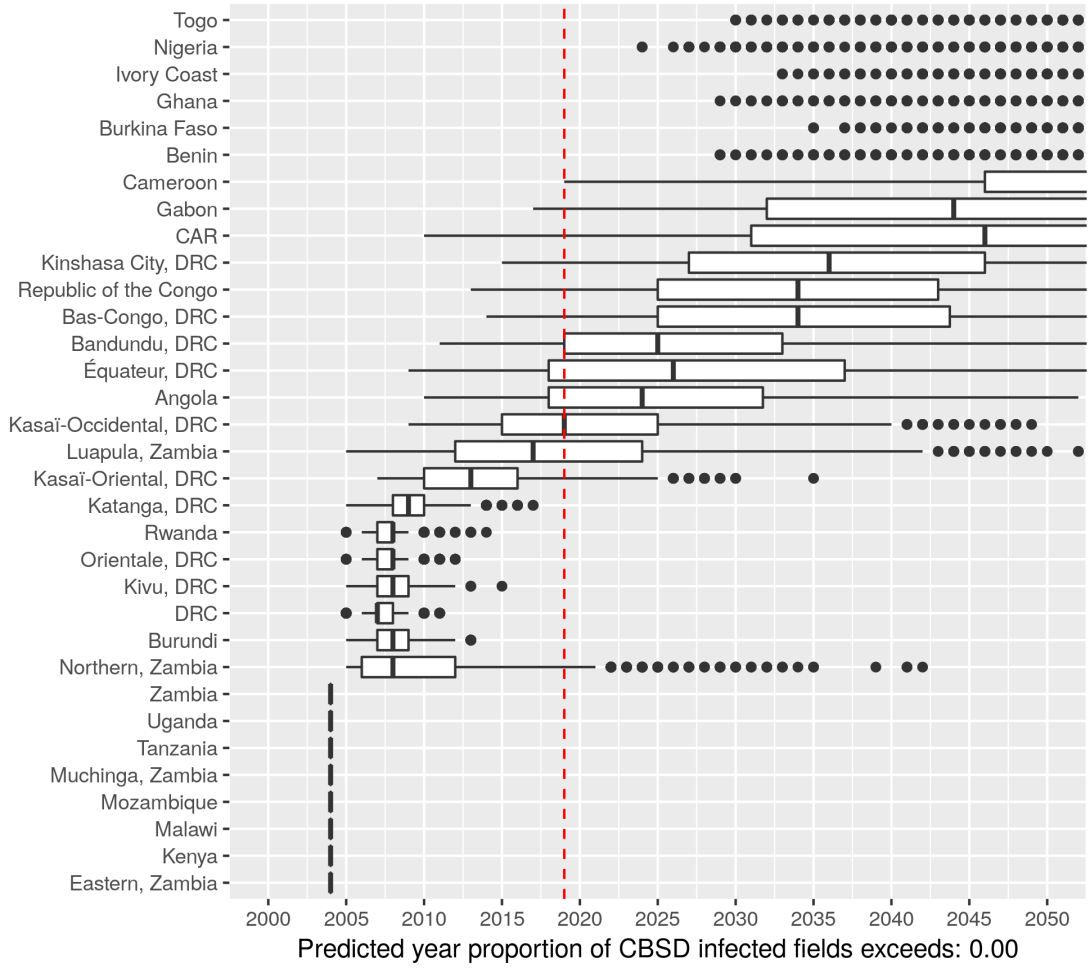


Figure 5.14: Box plot of median year of CBSD arrival in a set of key regions from 1170 simulations that satisfy the summary statistics constrained on the Ugandan surveillance data between 2005-2017. The red line denotes the year 2019. The hinges correspond to the 25th and 75th percentiles. The whiskers extend to the largest and smallest values that fall within the 1.5*IQR, where IQR is the inter-quartile range between the 25th and 75th percentiles. Points outside this range or where there are is a small number of arrivals are plotted as outliers.

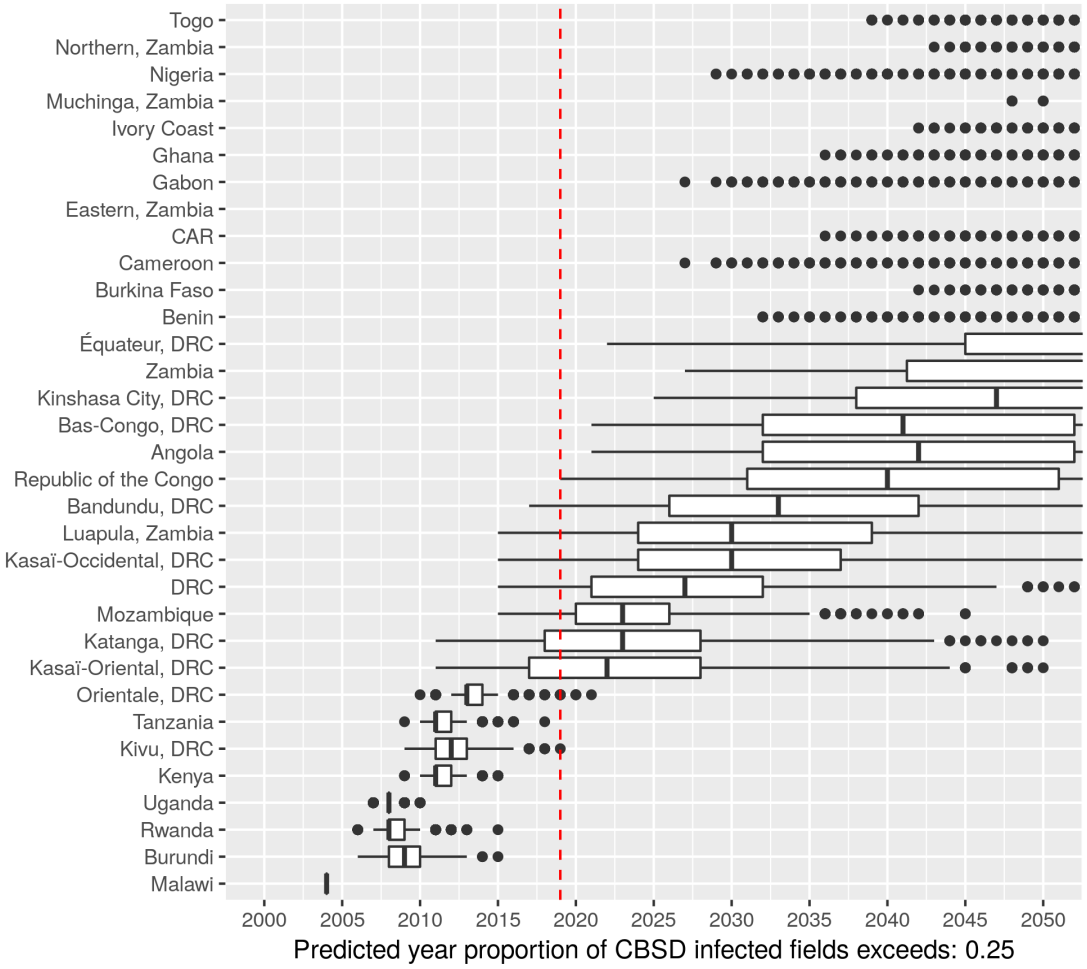
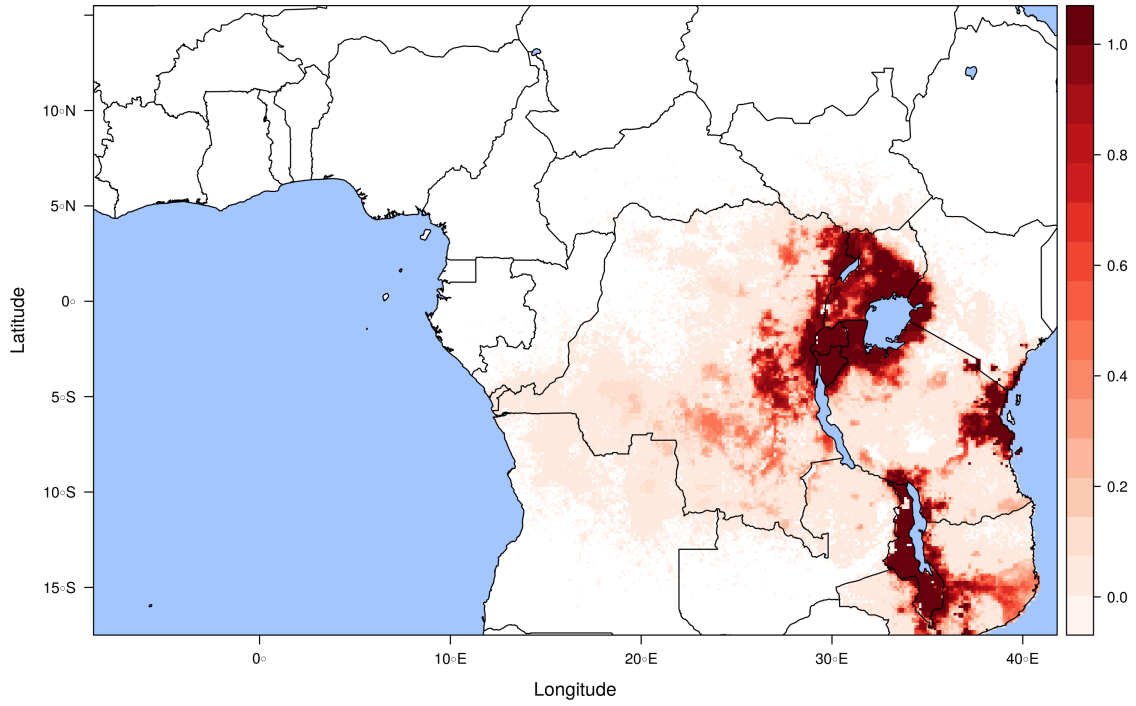
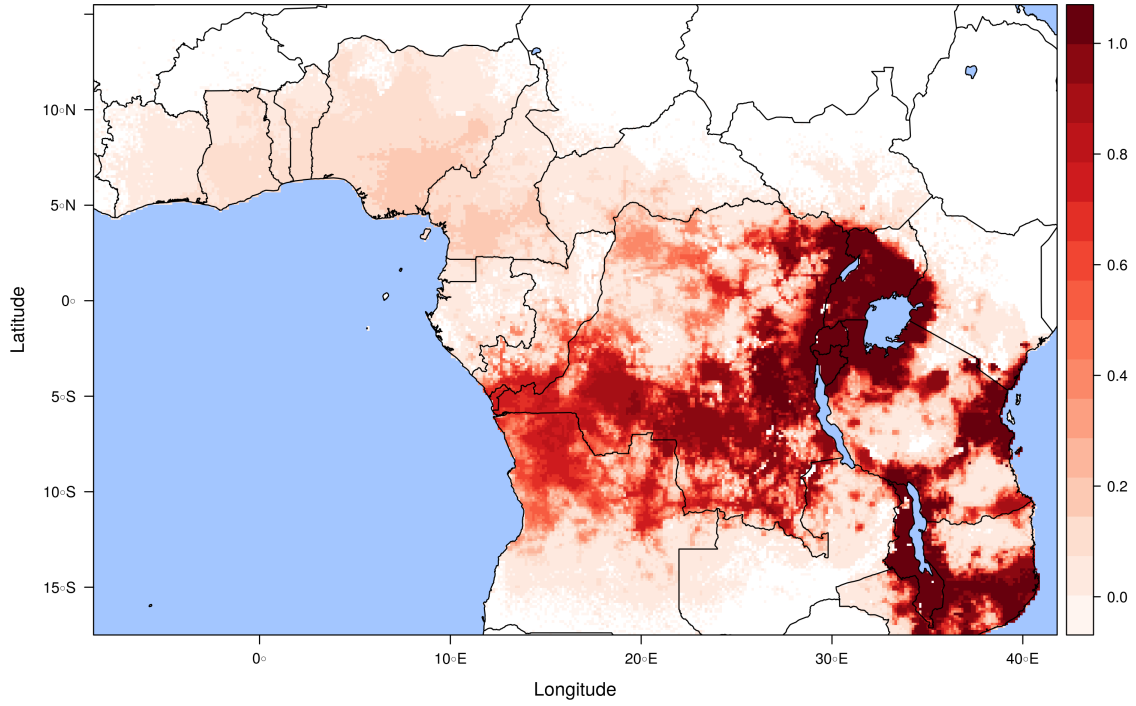


Figure 5.15: Box plot of median year in which 25% of fields in a given region or country become infected. Summarised from 1170 simulations that satisfy the summary statistics constrained on the Ugandan surveillance data between 2005-2017. The red line denotes the year 2019. The hinges correspond to the 25th and 75th percentiles. The whiskers extend to the largest and smallest values that fall within the 1.5*IQR, where IQR is the inter-quartile range between the 25th and 75th percentiles. Points outside this range or where there is a small number of arrivals are plotted as outliers.



(a) 2019 risk map



(b) 2050 risk map

Figure 5.16: Risk maps summarising the predicted probabilistic distribution of the CBSD epidemic at the start of 2019 from 1170 simulations that satisfy the summary statistics constrained on the Ugandan surveillance data between 2005-2017. Each raster cell is shaded according to the proportion of simulations that result in any infected fields in a given 100km^2 cell in the 2019 state raster.

5.4.1.3.2 Constraining on Ugandan and DRC surveillance data

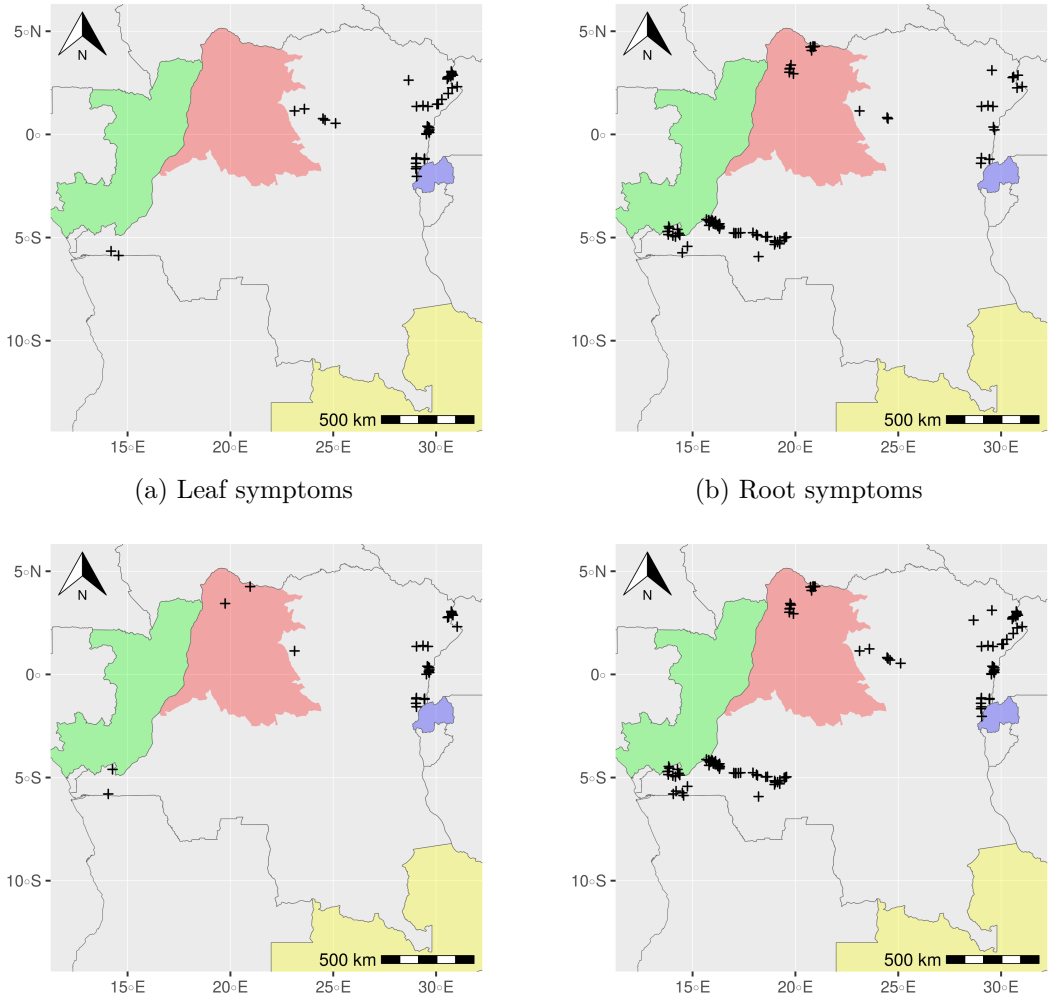
We now assume the observations in north-west DRC are correct and causally linked to the post-2004 Ugandan epidemic. Figure 5.17 is a map showing the region in which CBSD symptomatic plants were observed in 2017. We define a simulation as corresponding to the DRC observations if there are any infected fields in the north-west region of DRC by 2017. The extent of the region is highlighted in red in Figure 5.17.

Taking the subset of simulations that pass the previous criteria, we additionally constrain on the DRC criteria. A total of 269 of 1800 simulations pass both the Ugandan and DRC criteria. Figures 5.18 to 5.20 present the set of results from this subset of simulations. The earliest time of epidemic arrival in Cameroon is predicted to be 2019, with the median year being 2037 (Figure 5.18). For Nigeria, the earliest arrival is 2024, with the median year being 2046. Therefore, predictions range from earliest arrival in 5 years time to median arrival in 27 years. Notably, due to the high density of cassava production in Nigeria, the median year in which 25% of fields are infected is only 4 years later in 2050, when we allow for infection having spread to north-west DRC from Uganda by 2017.

5.4.1.4 Conclusion

The significant variability between the simulations that meet the two sets of assumptions concerning the current extent of CBSD spread in DRC highlights the urgency of validating the presence or absence of CBSD in north-west DRC. A more reliable sense of the present day distribution would narrow the variability in predictions relating to time of arrival in West Africa, hence allowing preparedness plans to be better calibrated with reality.

Nonetheless, as previously described, the scenarios in this section represent a lower-bound of predicted arrival times in West Africa. Highly unpredictable long range movements of infected planting material could directly introduce the disease. In section 5.4.3, we explore the dynamics of a direct introduction to West Africa.



(a) Leaf symptoms

(b) Root symptoms

(c) Stem symptoms

(d) Any symptoms

Figure 5.17: Surveillance data of CBSD symptoms observed in 2017 in DRC, highlighting the different symptoms observed in different locations. Sites of largely root symptom observations concentrated in the south-west, Bas-Congo, region of the country are likely associated with the unknown causal agent of 'CBSD-like' symptoms. New administrative divisions in DRC were brought in 2015. The map highlights in red the pre-2015 province of Équateur, Rwanda in blue, Zambia in yellow, and the Republic of Congo in green.

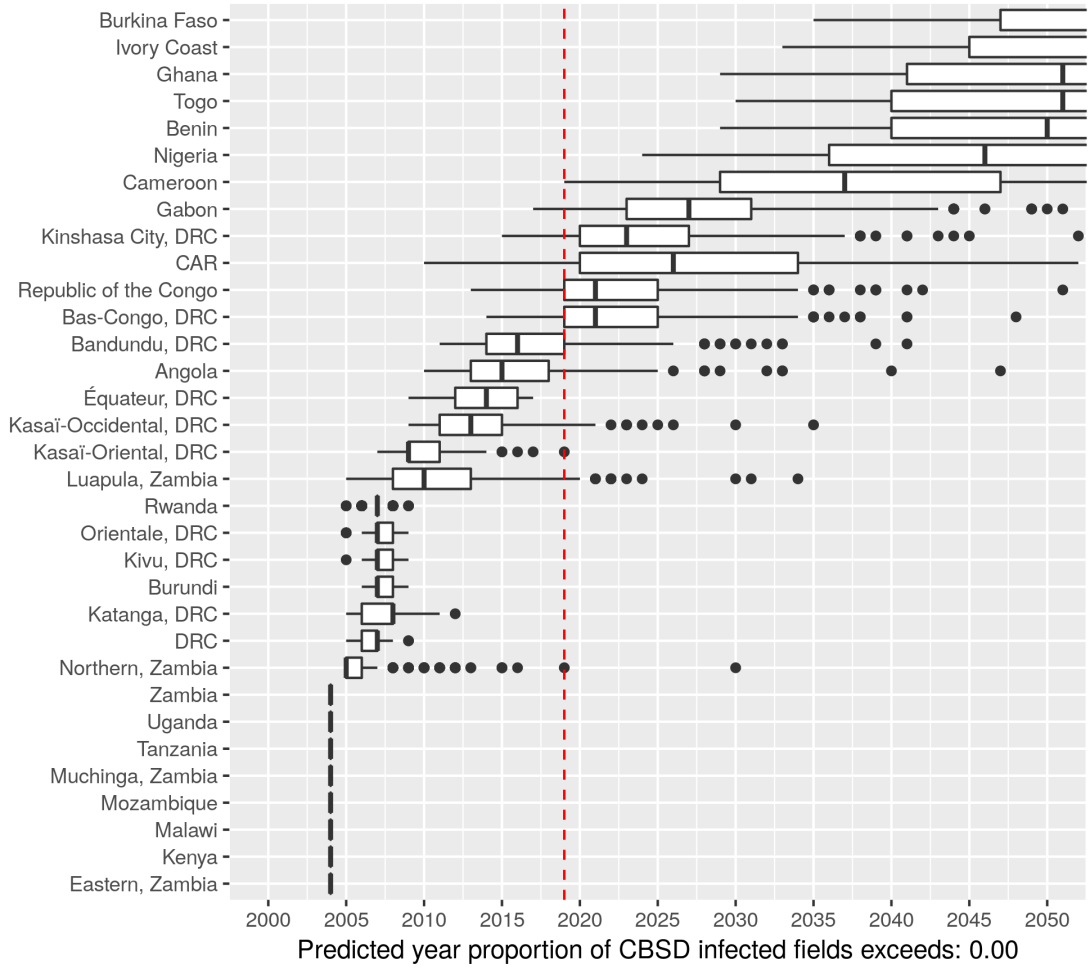


Figure 5.18: Box plot of median year of CBSD arrival in a set of key regions from 1170 simulations that satisfy the summary statistics constrained on the Ugandan surveillance data between 2005-2017 and also result in infection in the north-west region of DRC by 2017. The red line denotes the year 2019. The hinges correspond to the 25^{th} and 75^{th} percentiles. The whiskers extend to the largest and smallest values that fall within the 1.5*IQR, where IQR is the inter-quartile range between the 25^{th} and 75^{th} percentiles. Points outside this range or where there is a small number of arrivals are plotted as outliers.

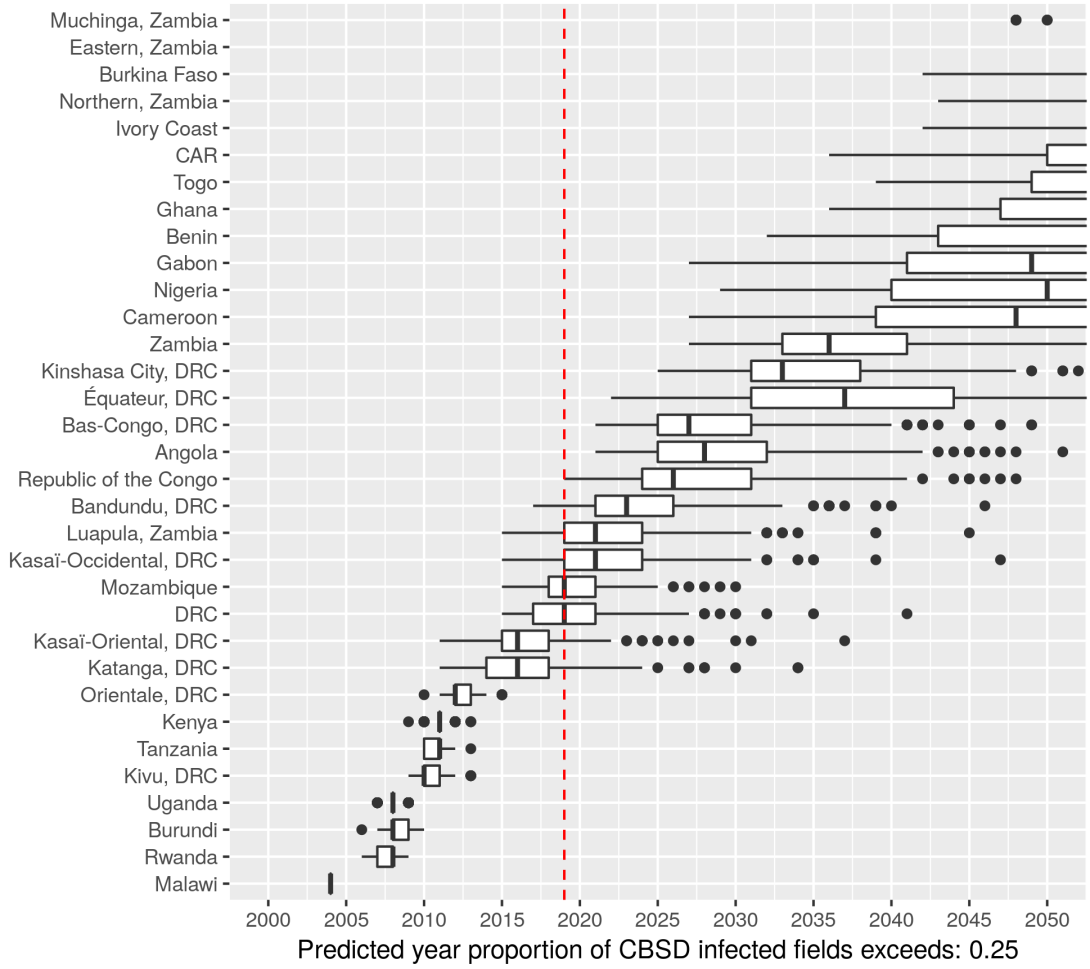
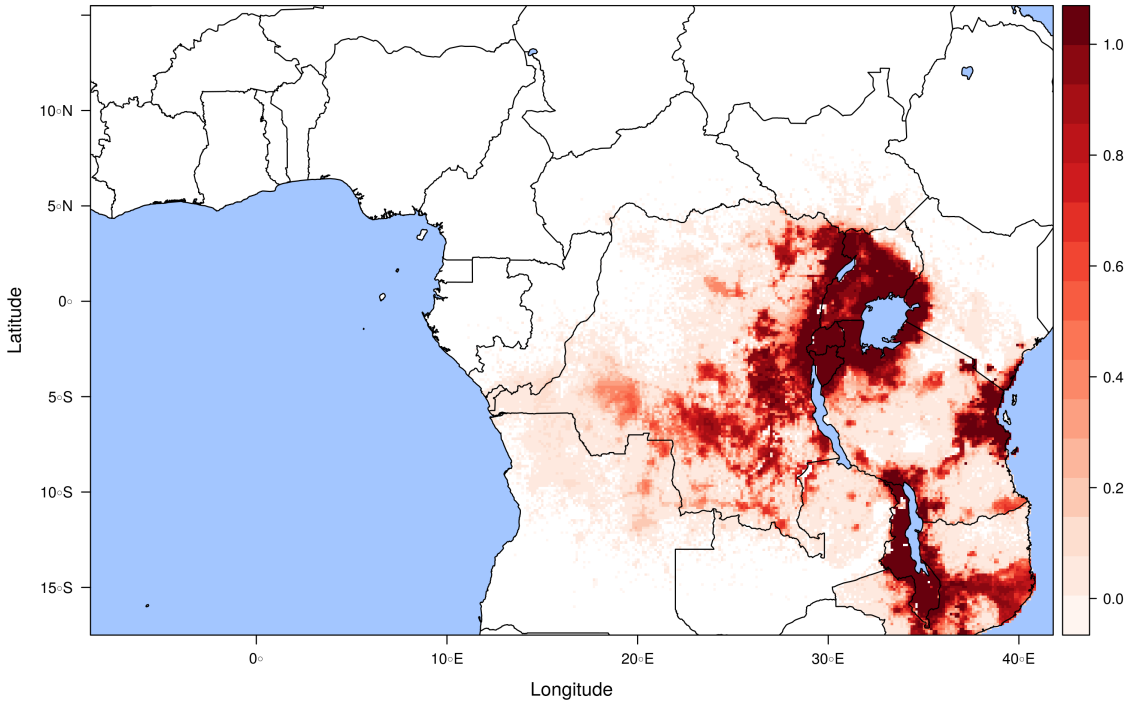
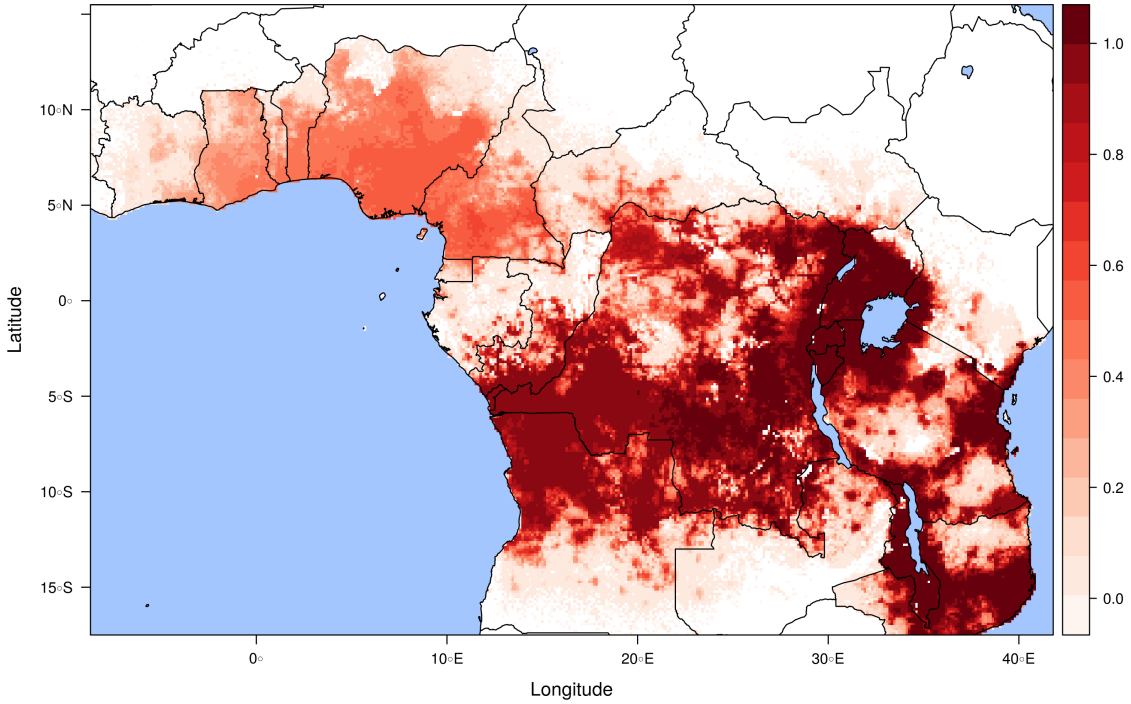


Figure 5.19: Box plot of median year in which 25% of fields in a given region or country become infected. Summarised from 1170 simulations that satisfy the summary statistics constrained on the Ugandan surveillance data between 2005-2017 and also result in infection in the north-west region of DRC by 2017. The red line denotes the year 2019. The hinges correspond to the 25th and 75th percentiles. The whiskers extend to the largest and smallest values that fall within the 1.5*IQR, where IQR is the inter-quartile range between the 25th and 75th percentiles. Points outside this range or where there is a small number of arrivals are plotted as outliers.



(a) 2019 risk map



(b) 2050 risk map

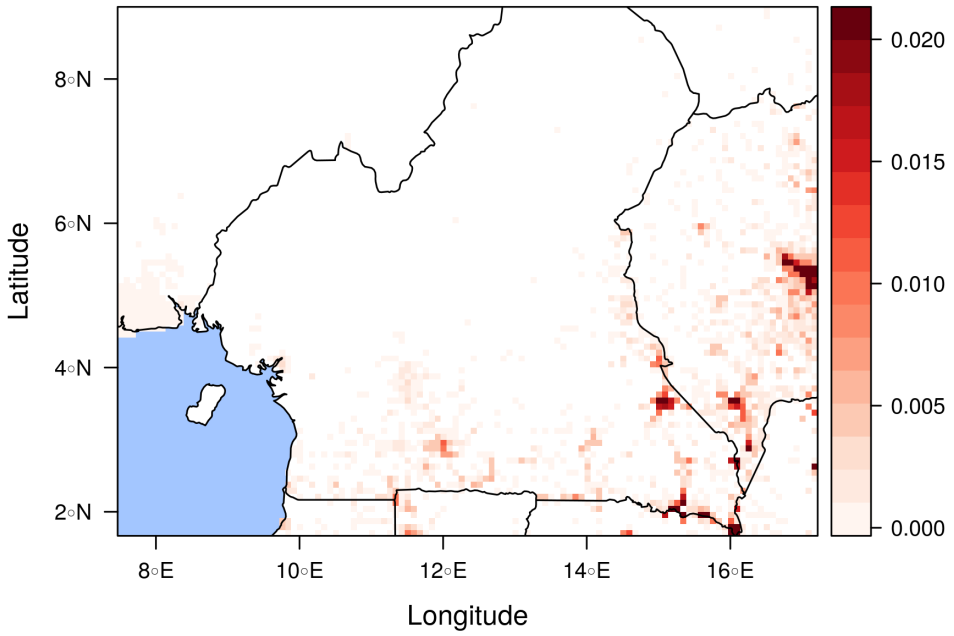
Figure 5.20: Risk map summarising the predicted probabilistic distribution of the CBSD epidemic at the start of 2019 from 1170 simulations that satisfy the summary statistics constrained on the Ugandan surveillance data between 2005-2017 and result in infection in the north-west region of DRC by 2017. Each raster cell is shaded according to the proportion of simulations that result in any infected fields in a given 100km^2 cell in the 2019 state raster.

5.4.2 Early warning surveillance sites

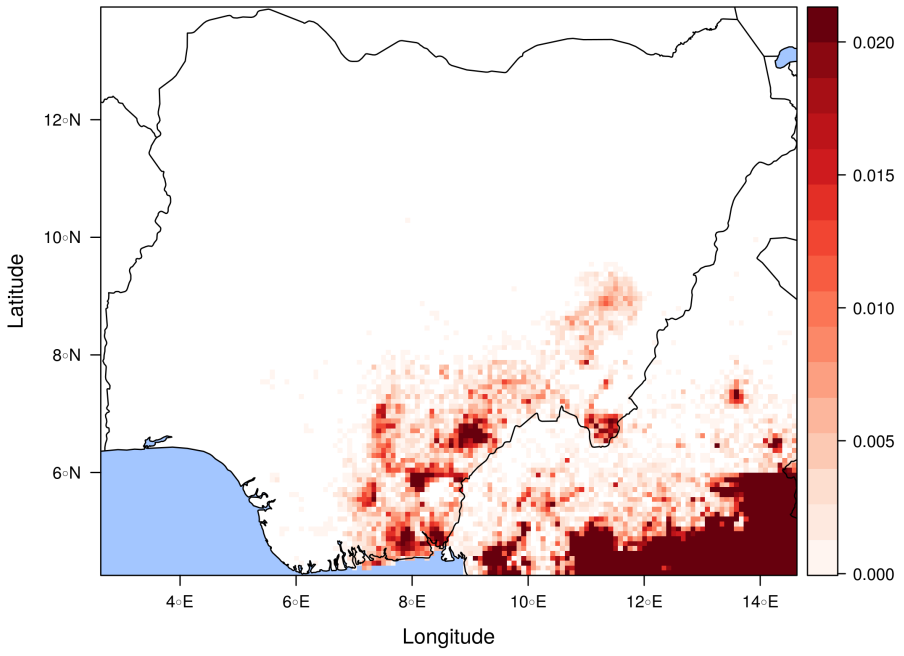
Surveillance resources in Central and West Africa should be optimised to identify the spread of CBSD into new regions as early as possible. In this section, we present risk maps that highlight the most likely sites of epidemic outbreak in Cameroon and Nigeria.

The methodology in this section is adapted from the risk map generation in the previous section. However, in this case, rather than generating the map using a fixed year, such as 2050, the infectious state raster for each simulation is the year in which the epidemic reaches a defined region. Therefore, irrespective of the predicted year of epidemic arrival, the resultant risk map highlights the most likely sites of infection in the first year.

Figure 5.21 shows the epidemic arrival risk map for Cameroon and Nigeria. In Cameroon, there are clear hotspots in the south and south-east in which surveillance could be concentrated to spot epidemic arrival. Similarly, there are key sites in south-eastern Nigeria that are most likely to first become infected under the expanding-spread spread scenario.



(a) Cameroon epidemic arrival risk map derrived from 1000 simulations that reach Cameroon within 50 years



(b) Nigeria epidemic arrival risk map derrived from 801 simulations that reach Nigeria within 50 years

Figure 5.21: Risk maps representing the distribution of infected sites from the first year the epidemic reaches the given country. Resolution is 10km. Raster values have been capped to a maximum of 0.02 to improve the visual resolution of the spatial heterogeneity amongst low valued cells.

5.4.3 Direct introduction to West Africa

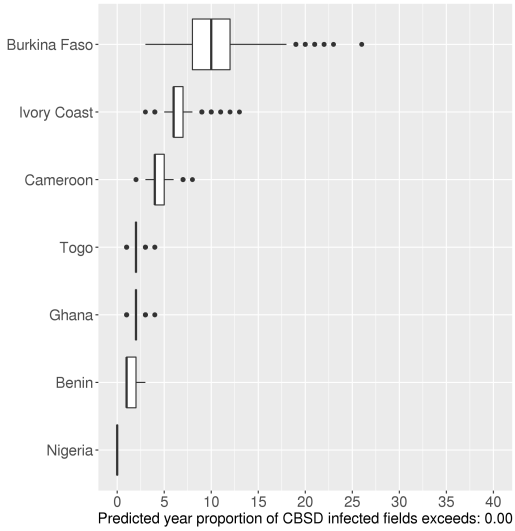
In this section, we explore the consequences of a direct human-mediated introduction of infected planting material at major ports in West Africa. To explore the dynamics of the resultant epidemic, we have selected the two most populous cities in West Africa: Lagos, Nigeria and Abidjan, Ivory Coast. These cities experience high levels of commercial flights and trade and are therefore high risk sites for human-mediated introduction.

5.4.3.1 Methods

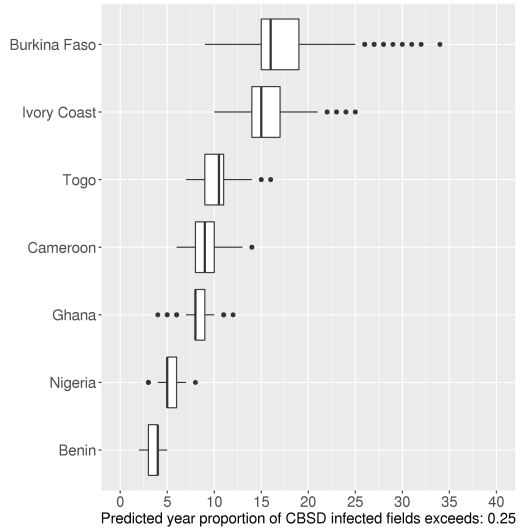
The equivalent of a single infected field is seeded at the start of the simulation in the vicinity of the city in question. The full extent of the simulation landscape is highlighted in extent number 3 in Figure 5.1. The simulation is then run for 40 years. For each scenario, 1000 realisations are simulated.

5.4.3.2 Results

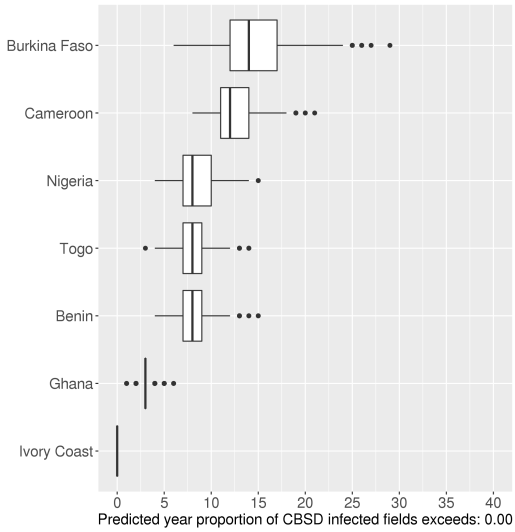
Figure 5.22 represents the rate at which the epidemic would be likely to spread after an introduction in either Nigeria or Ivory Coast. Secondary regional spread from an introduction to Nigeria is slightly faster than Ivory Coast, reflecting higher densities of cassava and greater degrees of connectivity amongst areas of cassava cultivation. However, in both scenarios, it is striking how rapidly the epidemic spreads throughout the region. In either case, the median rate of epidemic spread is predicted to reach all countries in the region in less than 15 years post-introduction. Similarly, median predictions for the year in which the proportion of infected fields for all countries in the region exceeds 0.25 is approximately less than 16 years after and introduction to Nigeria and 21 years after and introduction to Ivory Coast (Figure 5.22).



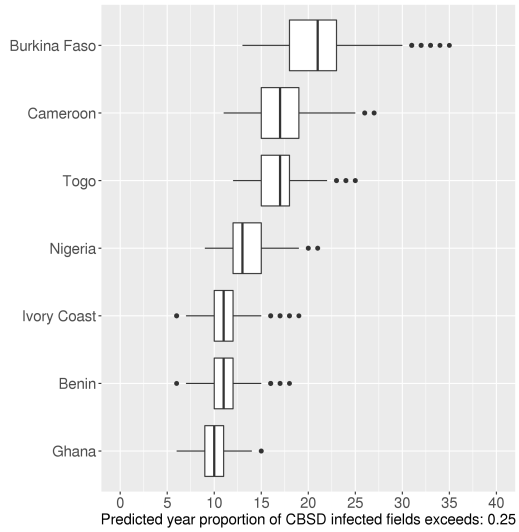
(a) Nigeria - 0.0



(b) Nigeria - 0.25



(c) Ivory Coast - 0.0



(d) Ivory Coast - 0.25

Figure 5.22: Bar plots representing the year in which the infectious proportion in a given country exceeds the specified value, of 0.0, referring to any infected fields in the country, hence time of first arrival, and 0.25, referring to 25% of fields becoming infected. Figures (a) and (b) represent simulations where infected planting material was introduced near Lagos, Nigeria. Figures (c) and (d) represent an introduction near Abidjan, Ivory Coast.

5.5 Management scenarios

5.5.1 Overview of management options

The management scenarios in this section build upon the expanding-spread and human-mediated spread scenarios in the previous section. We explore a number of scenarios that represent real-world interventions that could be implemented in West Africa, with the aim of providing a preliminary evaluation of the comparative effectiveness of different management strategies. A range of CBSD management tools and techniques already exist and many more are under development. However, given the resource constraints in all impacted countries, only a few tools and techniques are viable. We begin with a brief overview of the range of tools available to policy makers, along with a summary of their benefits and downsides, focussing upon their likely impact and how they can be incorporated into the model. We then describe a set of scenarios in which these tools are deployed to minimise epidemic spread and predict the efficacy of these interventions.

5.5.1.1 Reduce cassava production

The first method of susceptible host reduction is to replace cassava production with a different crop. This reduces the amount of host in the pathosystem. There are gradations of this strategy, in which the amount of cassava grown in a region is reduced, rather than eliminated. The upside of this strategy is either partially or fully to reduce reliance on cassava in order to reduce the risk of high yield loss. This must be weighed against the production risks of the alternative crops. The downside of this strategy is that it is difficult to persuade farmers to significantly change their growing behaviour in this way. Moreover, this would also disrupt the attempts of many countries to develop their cassava industries.

Within the model, the cassava host landscape represents the number of cassava fields at a 1km^2 resolution. The number of fields being grown can be uniformly reduced by a given factor, entirely removed, or specific regions can be targeted.

5.5.1.2 Reduce host susceptibility

The second method of control is to grow less susceptible, more resistant cultivars. There are two potential mechanisms that create the effect of resistance, and again, it is possible to consider gradations of this strategy. A cultivar may have a reduced probability of becoming infected with CBSV per viruliferous vector feeding event. Alternatively, post-infection, the cultivar may suppress viral titre. In both cases, increased resistance is associated with a reduction in the ability of the virus to spread from an infected plant. If a cultivar has a lower probability of becoming infected, fewer plants in the landscape will be infected, and therefore fewer can act as a source of inoculum. Alternatively, with a lower viral titre, there is likely to be a lower per feeding event probability of a vector acquiring the virus.

Note that the definition of resistance makes no reference to the impact of the disease on the plant in terms of its utility to humans. A cultivar that displays less severe symptoms is often described as tolerant (Fry, 2012). Therefore, despite these effects being locally beneficial, tolerance does not change the dynamics of an ongoing epidemic.

Ideally, planting material that is entirely resistant to all CBSV strains would be used. However, fully resistant cultivars identified in breeding trials are not yet widely available. Alternatively, cultivars that are partially resistant to infection and less readily pass the virus to the vector would still be an improvement. The viability of this strategy depends upon the availability and cost of improved planting material. The benefits of this strategy are that the level of required cultural change is lower than attempting to change the crop being grown. Nonetheless, cultural problems may still arise in terms of changes to other characteristics, such as taste or starch content.

Within the model, a reduction in host susceptibility could be incorporated in an analogous way to reducing the amount of host, as described in the previous section. Alternatively, partial resistance could be modelled through an influence on the transmission parameters, as with vector density.

5.5.1.3 Reduce vector abundance

As outlined in the model development chapter, the relative abundance of the vector, *B. tabaci*, has a significant causal effect on the epidemic dynamics. Reducing the

vector abundance reduces the rate at which the virus is dispersed between plants, which ultimately reduces the rate of epidemic growth. Therefore, the number of infected plants is reduced which means the epidemic has a reduced impact. Moreover, by slowing the rate of spread, other management strategies are likely to be effective. This could be achieved by breeding varieties for low vector abundance traits or via the systemic or conventional use of insecticides.

A reduction in vector abundance can be modelled by reducing the values in the vector abundance layer. This can be done in a specific region or throughout the landscape. However, within the model, trade and vector dispersal are integrated in a single dispersal kernel. Therefore, changes to the vector abundance layer will affect the combined rate of vector and trade dispersal. Nonetheless, we anticipate that unifying the two dispersal processes has a minimal effect on the validity of predictions with reduced vector abundance. The reason for this is that the impact of trade dispersal is highly dependant on vector abundance, hence the two processes are largely interconnected. Movement of infected cuttings does not, on average, increase the total number of infected plants in the landscape. It is vector activity that increases the number of infected plants. Therefore, in regions with low vector abundance, there are likely to be lower levels of infected plants, hence a lower probability of an infected cutting being exchanged.

5.5.1.4 Trade restrictions

The movement of infected planting material introduces the virus to new regions. Therefore, reducing the movement of planting material would reduce the probability of a human-mediated introduction to a new region. Our models show that, in principle, the introduction of a single infected field to Nigeria or Ivory Coast is likely to result in the rapid regional spread of the epidemic. Given the limited resources available to enforce trade restrictions within and between countries, efforts should be concentrated on the highest risk locations. For example, planting material transported by air could directly introduce the virus to West Africa. Therefore, it is important to consider the likely cost-effectiveness of introducing or improving quarantine checks at airports. Comparably localised movements of planting material occur throughout the continent and are therefore much harder to prevent, requiring significant political backing. However, cross border checks may be possible.

Charities, NGOs and governments are often responsible for the movement of large

amounts of planting material over relatively long distances (Ntawuruhunga and Legg, 2007). Therefore, it is important to ensure that these activities do not inadvertently introduce infected material to new regions.

As the model currently unifies trade and vector dispersal into a single kernel, we do not have an explicit mechanism to impose trade restrictions. However, there are possible indirect mechanisms, such as reducing the maximum dispersal distance in defined regions, such as near country borders. In doing so, we would be assuming that dispersal events above the given threshold distance are occurrences of trade movements, as opposed to long range dispersal by *B. tabaci*.

5.5.1.5 Responsive epidemic management

In newly infected or endemic regions, additional strategies can be deployed to reduce the number of infected plants in the pathosystem. These strategies often require both new infrastructural developments or investment in extension agent programmes to educate farmers (Fry, 2012).

Clean seed programmes can be set up to disseminate uninfected planting material, ideally from an improved cultivar. Governments can disseminate or farmers can buy this material between planting seasons to partially or fully replace their fields with clean material. Clean seed programmes are challenging to implement as they require the large-scale reliable production and certification of clean planting material (Legg et al., 2017). In addition, there are comparably high overheads for governments if materials are disseminated for free or for the farmer if they are sold.

An alternate strategy without the high overhead costs is preferential selection, which requires farmers to actively select cuttings from uninfected plants for the next season (Fargette, 1995). However, this strategy can only work if the in-field prevalence is low enough that there are sufficient uninfected plants at the end of the season. Moreover, preferential selection also requires that the farmer is sufficiently trained to identify CBSD symptoms.

Within a planting season, farmers can actively rogue infected plants. This relies on the farmer's ability to correctly diagnose an infected plant. Depending on the number of infected plants in the field and the vector abundance, roguing can reduce or eliminate infection (Legg et al., 2017).

5.5.1.6 Integrated disease management

In West Africa, the goal of management is to prevent the introduction and establishment of CBSD. However, preventing or eliminating all incursions may be difficult, and most likely impossible, given both the scale of the spreading epidemic and of cassava production. Therefore, management programmes should be designed to minimise the probability of an introduction and also to minimise the rate of spread in the event of an introduction. This has been achieved in previous examples of effective plant disease management programmes when a number of different control strategies that complement one another have been deployed. Fry (2012) presents examples of robust management in agricultural systems in the USA, often involving multiple pathogens, such as potato production in the north-east, and corn production in the midwest. In addition, it is vital for surveillance to be incorporated into management programmes to provide reliable information to guide decision making (Parnell et al., 2017). The model provides a framework to explore combinations of management strategies to assess which are likely to be most robust and effective in managing the CBSD epidemic.

5.5.2 Preventative management in Eastern Nigeria

5.5.2.1 Introduction

In section 5.4.1, we simulate the expanding-spread scenario in which the epidemic moves through Central Africa and enters Nigeria along the south and central regions of the eastern border (Figure 5.21). Incursions in the north are less frequent due to low cassava host density in those regions.

A buffer region along the eastern border of Nigeria has been postulated at previous working meetings of the West African consortium, WAVE, as a possible management strategy. Accordingly, in this scenario, we explore a pre-emptive control strategy by simulating a buffer region of various widths along this border. It is unlikely that a buffer could entirely prevent introductions in the long run. However, if the rate of epidemic spread could be slowed, it would buy time for other responsive interventions to be deployed.

A number of interventions could produce the effect of a buffer region along the border. From an epidemiological perspective, a full buffer is the entire removal of

susceptible host and a partial buffer would be a reduction. This could be achieved by changing the crop being grown or deploying resistant material. Entirely replacing all susceptible material would be extremely challenging and would require significant political backing, resources and public compliance. This scenario attempts to assess whether such a draconian measure would be an efficient or effective use of resources.

5.5.2.2 Methods and results

Within the model, we remove susceptible host within Nigeria along the border of Cameroon. Simulations explore the effects of a range of different sized buffer regions (Figure 5.23). For each buffer size, 1000 simulations are run for 20 years. The simulations are initialised using a representative simulation state from cross-continental spread simulation. Specifically, we select the distribution of infected fields from the year prior to epidemic arrival in Nigeria (Figure 5.24). The full extent of the simulation landscape is highlighted in extent number 4 in Figure 5.1. Within this section, we do not provide a corresponding real-world year, hence we refer to simulation years, $t = 0 \rightarrow 20$. However, $t = 1$ approximates to the year in which CBSD would spread from Cameroon to Nigeria in the absence of a buffer.

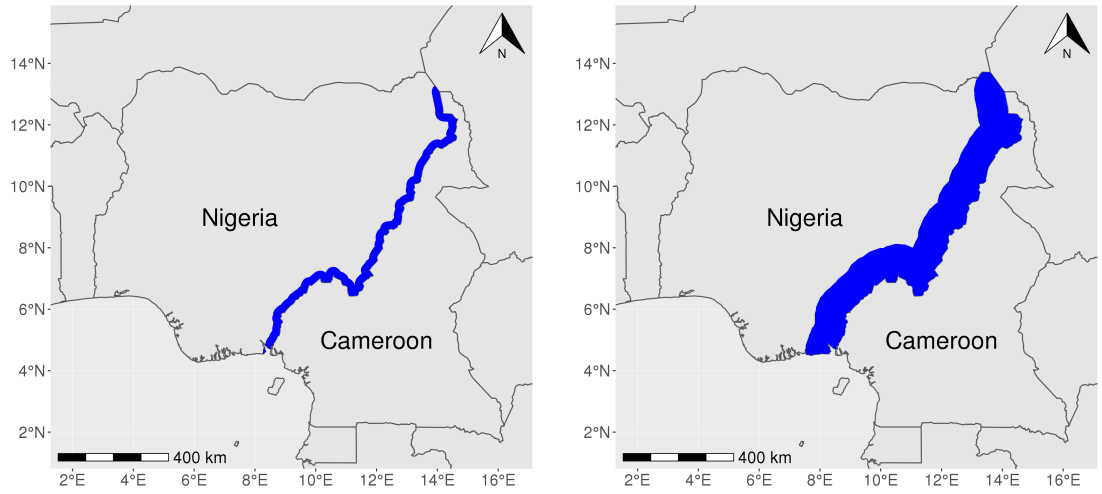
Figure 5.25 summarises the impact of different buffer region sizes on epidemic arrival and subsequent bulk up. The median arrival time for buffers of size 0.25 and 0.5 decimal degrees is year 3, which is the same for the control group with no buffer. Simulations with a buffer of 1.0 decimal degrees (>110km) have a median arrival time of year 4. The largest buffer region of 2.0 decimal degrees (>220km) still has a relatively small impact, moving the median year of arrival to year 6. Moreover, only 2 simulations out of 1000 with a buffer size of 2.0 decimal degrees did not result in epidemic arrival in Nigeria within 20 years. For all other buffer sizes, all 1000 simulations reach Nigeria in the 20 year simulation period. Table 5.1 quantifies the approximate proportion of cassava fields in Nigeria that would need to be removed or replaced to produce a buffer region of a given size.

5.5.2.3 Conclusion

A buffer region along the border would require extremely high logistical and resource investments. However, the results presented in this scenario indicate that even large buffer regions would have minimal impact of the predicted year of epidemic arrival or

Buffer region width (decimal degrees)	Buffer region width (approx. km)	Proportion of fields removed
0	0	0
0.25	27.5	0.02
0.5	55	0.06
1	110	0.19
2	220	0.51

Table 5.1: Summary of the sizes of the different simulated buffer regions. The number of fields removed in the buffer region is given as a proportion of the total number of fields in Nigeria without a buffer region.



(a) 0.25 decimal degree buffer

(b) 1.0 decimal degree buffer

Figure 5.23: Maps highlighting the extent of two buffer regions of size (a) 0.25 and (b) 1.0 decimal degrees in diameter. We explore the impact on a 20 year epidemic of removing susceptible host from the buffer region. See Figure 5.24 for the initial distribution of infected fields in the landscape

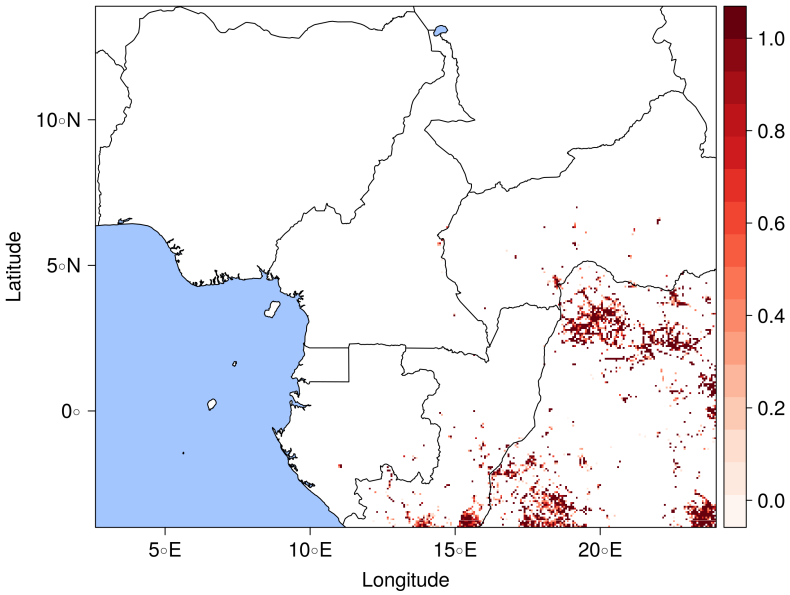
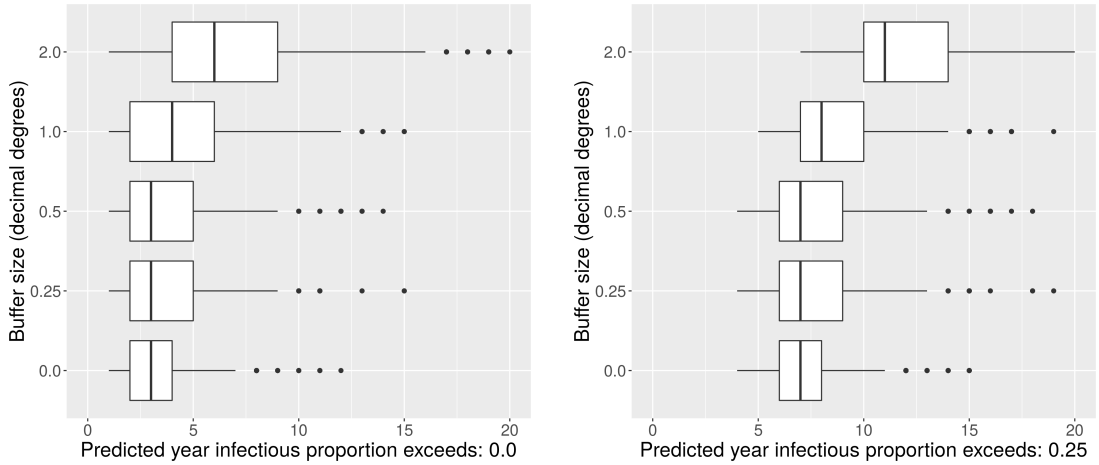


Figure 5.24: Distribution of infected fields as a proportion of the host landscape at a resolution of 1km^2 (see Figure 4.2) used as initial conditions for buffer region simulations.



(a) Year of epidemic arrival

(b) Year the proportion of infected fields exceeds 0.25

Figure 5.25: Box plots contrasting the impact of differently sized Nigerian border buffer regions on the predicted year of epidemic arrival and bulk up. (a) summarises the year in which the infectious proportion in Nigeria exceeds the specified value, of 0.0, referring to any infected fields in the country, hence time of first arrival. (b) refers to the year in which the proportion infected fields in the Nigerian host landscape exceeds 0.25.

subsequent spread. Moreover, we have assumed that cassava in this region could be entirely replaced with alternative crops or resistant cultivars, which given the scale of cassava growth in the border region and the current unavailability of resistant cultivars at scale, would likely be impossible to achieve in reality. We conclude that it would be an ineffective use of resources to implement this strategy as a preventative measure, incurring significant disruption for minimal benefit. Similar conclusions have been reached when considering control involving buffer zones in other pathosystems, such as Sudden Oak Death in California, USA (Filipe et al., 2012).

5.5.3 Surveillance strategies and responsive management

5.5.3.1 Introduction

The results from the buffer region scenario suggest that a large but static management strategy is unlikely to control the epidemic. We now explore responsive management strategies, which require information on the changing epidemic distribution to guide subsequent control measures.

The WAVE consortium is currently lobbying for the major cassava growing countries of West Africa to develop preparedness plans in anticipation of CBSD incursions. Here, we focus on the aspect of preparedness relating to controlling, and ideally eliminating, CBSD post-introduction, as opposed to measures to reduce the probability of an original incursion.

A post-introduction CBSD control strategy requires 1) a mechanism find sites of infection as early as possible and 2) the ability rapidly deploy effective management resources. Both of these components can be strengthened in advance through the tactical deployment of resources.

In this scenario, we perform a preliminary analysis towards understanding where resources should be invested to improve Nigeria's surveillance capacity in anticipation of CBSD being introduced. Specifically, we contrast two hypothetical strategies: increasing the level of surveillance via teams of trained surveyors, or developing the national capacity for farmers and extension workers to report infection.

In this first order analysis, we focus on contrasting the two surveillance strategies. In doing so, we make a number of unrealistic, simplifying, assumptions about the exact

implementation of surveillance and control. For example, we assume that available surveillance capacity is deployed uniformly throughout Nigeria, rather than being tactically concentrated in high risk regions (Thompson et al., 2016). In terms of control, we assume that the identification and reporting of an infected field leads to the elimination of CBSD from that field and any others in the same $1km^2$, after which, the fields are CBSD resistant. This is analogous to the provision and mandatory replacement of cassava with a resistant cultivar. This likely over estimates the efficacy of providing resistant clean seed material. Yet it is also unlikely that a single control strategy will be pursued, given that a combination of complementary control strategies are more likely to be effective than a single strategy in isolation (Fry, 2012).

In future analyses, it would be possible to explore more nuanced, cost-effective, and realistic optimisations to the surveillance and control strategy in Nigeria.

5.5.3.1.1 Surveillance option 1: Trained surveyors

The current strategy in practice in Nigeria is to deploy teams of surveyors to a subset of fields within a region to visually assess the infection status of each field and collect samples for molecular diagnostics. Currently, surveillance is largely uniform within the country as it is also required to gather information on the national incidence of CMD. A downside of this surveyor-driven strategy is that it can become expensive to gather large amounts of information at scale.

5.5.3.1.2 Surveillance option 2: Farmer and extension service reporting

In contrast, by leveraging the extension services, disease outbreaks can be reported at scale directly from farmers and extension workers. Whilst requiring significant initial investment to improve and expand the currently limited extension service, this strategy has the potential to reach far more fields than would be possible with active surveillance. Moreover, from a political perspective, strengthening the extension system has many benefits beyond improving the national CBSD surveillance capacity, such as improving agronomic practices leading to higher yields and more trade.

Nonetheless, it would be far more logistically challenging to improve the extension services than scale up the ongoing surveyor-driven surveillance activities. Moreover,

farmers and extension workers would not be trained to the same level as specialised surveyors. Therefore, reporting via the extension services has a higher risk of problems relating to data accuracy and low rates of participation.

5.5.3.2 Methods

We take the stochastic, spatially-explicit SI model as previously used to model CBSD spread and introduce parameters to explore surveillance and control scenarios. The full extent of the simulation landscape is highlighted in extent number 5 in Figure 5.1. Table 5.2 describes the key parameters that modulate the behaviour of surveillance and control within the model. For the purposes of interpretive clarity, we define the DetectionParameter as the product of two directly measurable parameters (Equation 5.1).

The parameter, IdentifyInfectionProb, could be measured by testing surveyors, extension workers or farmers to quantify their average diagnostic accuracy. Whilst the ReportingProb parameter would be largely predictable with surveillance teams, it could be quantified amongst extension workers and farmers via in-person or phone-based surveys.

$$\text{DetectionParameter} = \text{IdentifyInfectionProb} \times \text{ReportingProb} \tag{5.1}$$

We have constructed two contrasting scenarios to explore the relative efficacy of focussing resources on strengthening active surveillance capacity via teams of surveyors versus strengthening the capacity for farmers to report via an extension system.

Beginning with the farmer reporting scenarios, we assume resources have been invested into strengthening the national extension system to train farmers to identify CBSD symptoms and the extension workers pass infection reports from farmers upstream to a central authority, such as the regional government. By assuming a degree of farmer education on CBSD symptoms, we set IdentifyInfectionProb to 0.5. The value of this parameter is approximated in relation to the empirically derived average value of 0.85 from the Ugandan National surveillance data described in the model development chapter. The value of 0.5 may be a conservative estimate as farmers will have the benefit of observing the much more obvious root symptoms, whereas surveyors only tend to observe the more subtle foliar symptoms. However, we do not have data to quantify the effect precisely of farmer education on accurate

Parameter	Description
DetectionFrequency	Frequency of surveys (years)
DetectionFirst	Simulation year when surveillance begins
IdentifyInfectionProb	Probability that an infected field is diagnosed as infected
ReportingProb	Probability that the diagnosis of an infected field is reported
DetectionParameter	Probability that a surveyed infected field is reported as infected
DetectionSampleProportion	Proportion of fields surveyed within a given site (1 km^2)
ControlCullRadius	Number of cells beyond the control site to also control (A value of 0.0 controls only the original site)
ControlCullEffectiveness	Proportion of infected fields within site to perform control on

Table 5.2: Description of model parameters relating to surveillance and control specification.

disease reporting.

Having fixed all other parameters, we investigate the consequences of increasing the ReportingProb parameter, being the probability that a farmer-identified infected field is reported to the extension services (Table 5.3). We explore what we estimate to be a potential viable range of reporting probabilities, between 1% and 20%. Increasing this parameter is analogous to improving the reach of the extension service and the associated capacity for farmers CBSD observations to be reported centrally. For the purposes of simplicity and exploratory analysis, we assume that a positive CBSD report leads to removal of all cassava within the $1km^2$ and its replacement with resistant planting material or an alternative crop.

In contrast, within the surveyor reporting scenario, we assume that the probability of an observation being reported is 1.0 and we set the probability of identifying infection value to 0.85, as empirically quantified. We then sweep over the DetectionSampleProportion parameter, which is analogous to increasing the number of fields the national surveillance team visits in a given year (Table 5.3). In 2017, the Nigerian survey teams surveyed 860 fields out of an approximate 56 million in total. This equates to a proportion of approximately 1.5e-05. We take this as the lower bound DetectionSampleProportion parameter value and increment the parameter by an order of magnitude each step.

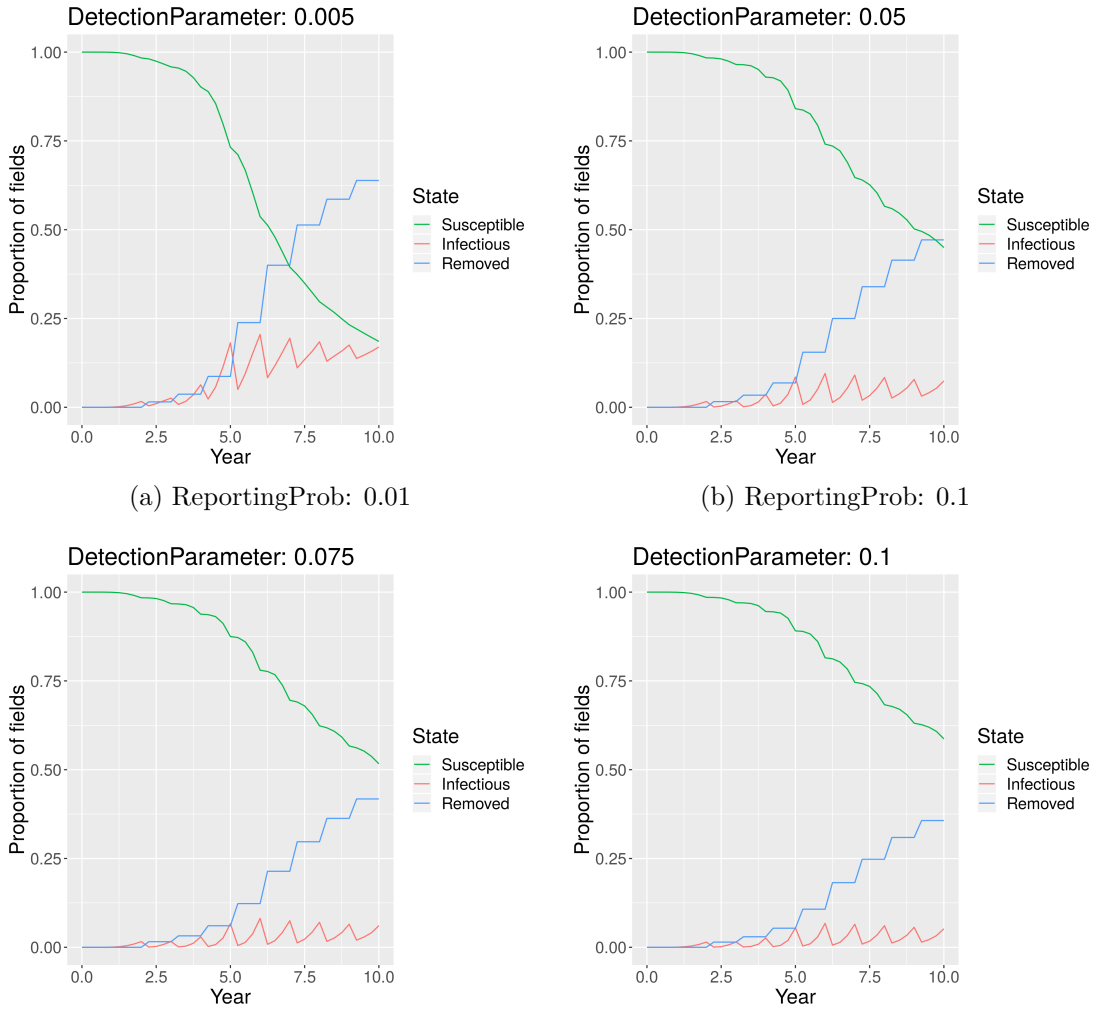
Surveillance strategy	Parameter	Value
	DetectionFrequency	1.0
	DetectionFirst	2.0
	DetectionParameter	Equation 5.1
	ControlCullRadius	0.0
	ControlCullEffectiveness	1.0
Surveyor	IdentifyInfectionProb	0.85
Surveyor	ReportingProb	1.0
Surveyor	DetectionSampleProportion	{1.5e-05, 1.5e-04, 0.0015, 0.015}
Farmer/Extension	IdentifyInfectionProb	0.5
Farmer/Extension	ReportingProb	{0.01, 0.05, 0.1, 0.15, 0.2}
Farmer/Extension	DetectionSampleProportion	1.0

Table 5.3: Surveillance and management parameters. Global parameters except where specified for the trained surveyor or farmer/extension system reporting scenario.

5.5.3.3 Results

Figure 5.26 and 5.27 track the median proportion of host in each state across 1000 simulations per parameter value. As would be expected, increasing the reach of surveillance generally reduces the infectious proportion and the total amount of host that must be removed and replaced each year. However, the impact of increasing active surveyor-based surveillance by multiple orders of magnitude appears to be relatively small in terms of the impact on the amount of host that must be removed and replaced, bringing the yearly proportion of fields that need to be replaced from 8.4% to 7.6% across a three orders of magnitude increase in active surveillance (Table 5.5). Moreover, there appears to be a minimal reduction in the yearly epidemic size as active surveillance is increased.

In contrast, increases in the reach of the extension services via the ReportingProb parameter appear to result in sizeable improvements in epidemic control and reducing the total amount of host that needs to be replaced. A 1% reporting probability (DetectionParameter: 0.005) requires 8.0% of fields to be replaced per year down to a 20% reporting probability (DetectionParameter: 0.1) which requires only 4.5% to be replaced per year. Figure 5.28 highlights this distinction in the efficacy of the two strategies by contrasting the total proportion of fields to be replaced over the ten year period as a function of parameter value for each scenario.



(a) ReportingProb: 0.01

(b) ReportingProb: 0.1

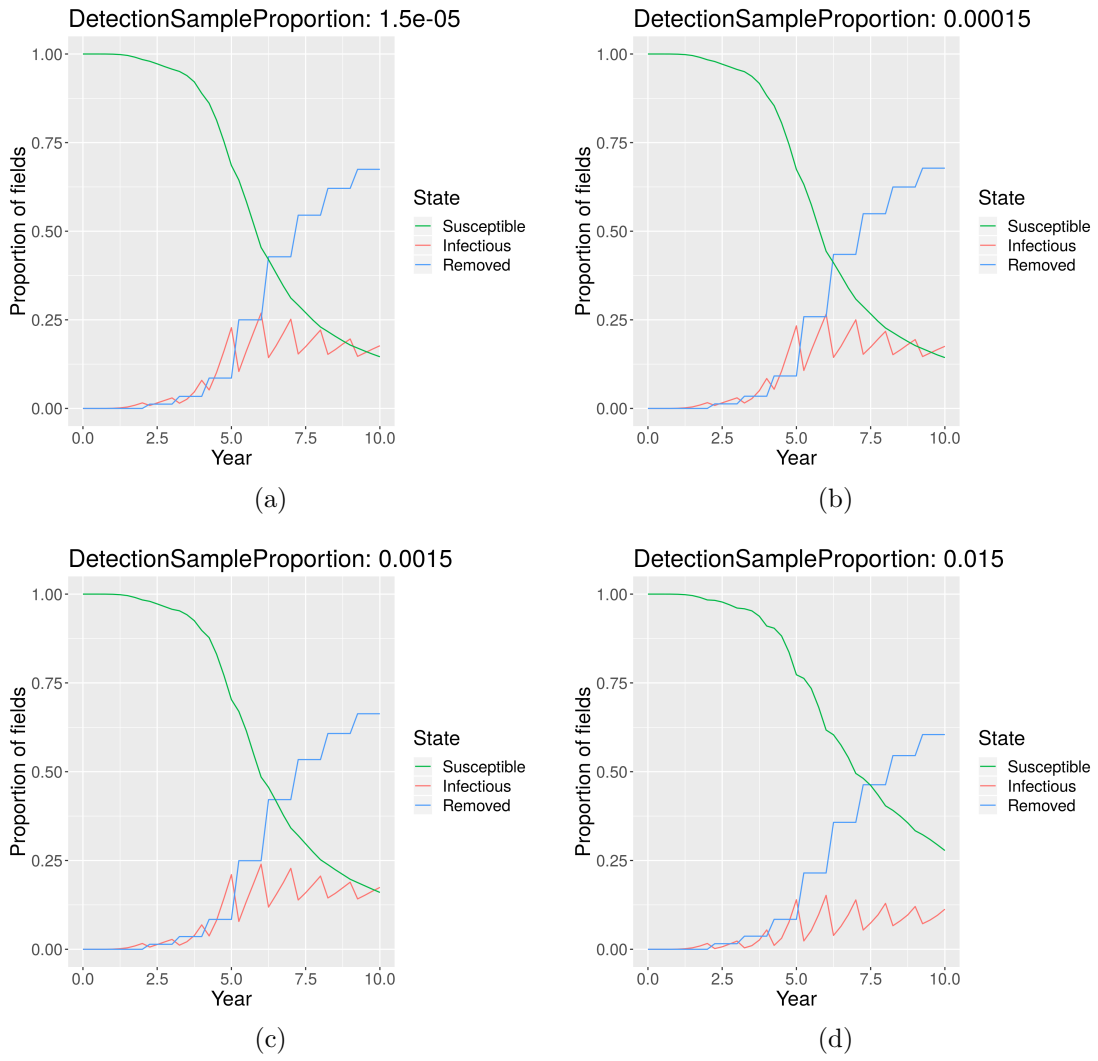
(c) ReportingProb: 0.15

(d) ReportingProb: 0.2

Figure 5.26: Median epidemic state over time for the farmer reporting scenario in Nigeria as the ReportingProb parameter, and consequently the DetectionParameter, are increased.

DetectionParameter	PropRemoved	PropPerYear	Hectares	HectaresPerYear
0.005	0.639	0.08	3,576,000	447,000
0.025	0.559	0.07	3,128,000	391,000
0.05	0.471	0.059	2,639,000	329,000
0.075	0.418	0.052	2,338,000	292,000
0.1	0.357	0.045	1,998,000	249,000

Table 5.4: Mean epidemic statistics across 1000 simulations per parameter for the farmer reporting scenario in Nigeria as the ReportingProb parameter, and consequently the DetectionParameter, are increased (Equation 5.1).

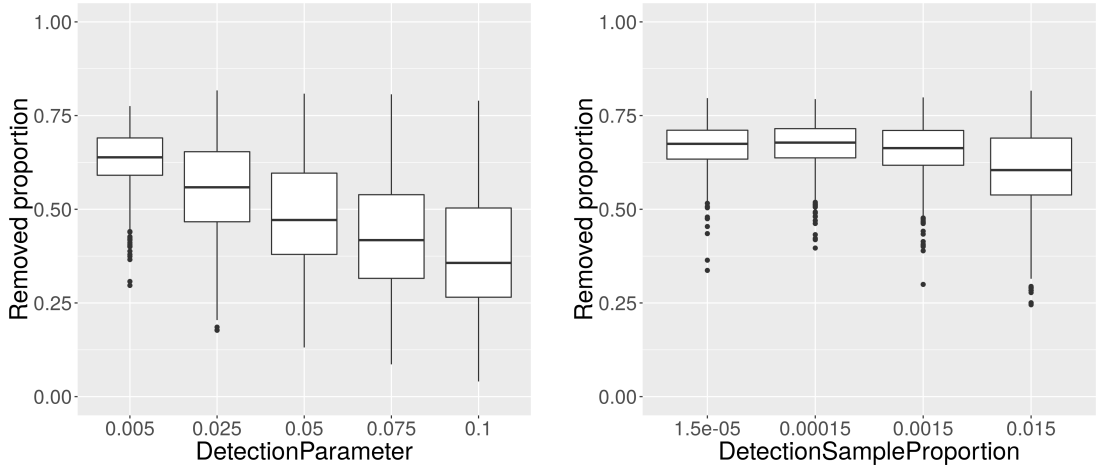


(a) (b) (c) (d)

Figure 5.27: Median epidemic state over time for the surveyor reporting scenario in Nigeria as the DetectionSampleProportion parameter is increased.

DetectionSampleProportion	PropRemoved	PropPerYear	Hectares	HectaresPerYear
1.5E-05	0.675	0.084	3,777,000	472,000
0.00015	0.678	0.085	3,796,000	474,000
0.0015	0.663	0.083	3,713,000	464,000
0.015	0.605	0.076	3,385,000	423,000

Table 5.5: Mean epidemic statistics across 1000 simulations per parameter for surveyor reporting scenario in Nigeria as the DetectionSampleProportion parameter is increased.



(a) (b)

Figure 5.28: Box plots representing the total amount of resistant material that is deployed over the 10 year epidemic and control simulations with different parameter values for (a) the ReportingProb parameter, and consequently, the DetectionParameter for the farmer/extension reporting scenario and (b) the DetectionSampleProportion for the trained surveyor-based scenario.

5.5.3.4 Conclusion

The results in this section draw attention to the contrasting efficacy of the two surveillance strategies at the national level. Not only does the extension worker/farmer reporting model appear to be far more effective, improvements to the extension service would have many other benefits beyond those relating to CBSD.

However, it is important to note the simplifying assumptions that have been made. Specifically, we have assumed that the surveillance effort is deployed uniformly across the country. Therefore, significant gains could be achieved for each strategy if surveillance capacity were more intelligently guided to key regions. Moreover, we have not explored the likely possibility of both strategies to work synergistically to guide the deployment of control more efficiently than via each strategy in isolation.

In addition, we have assumed that if control were implemented, it fully replaces all host in the 1km^2 with fully resistant material or alternate crops. Under present day conditions, this would not be possible due the lack of truly resistant material available. However, these scenarios allow us to approximate the level of resistant material production that would be necessary to meet the yearly demands.

5.5.4 Reducing vector abundance

5.5.4.1 Introduction

As demonstrated in Chapter 4, vector abundance is an integral feature of the model and significantly modulates the dynamics of the predicted epidemics. In this section, we quantify the effects of reducing vector abundance on the rate of epidemic spread within Nigeria. In the absence of a control process that actively removes infected cassava plants, the epidemic would continue to spread. However, reducing vector abundance could in principle slow the rate of epidemic spread, which reduces the severity of the shock to the food production system, meaning smallholders have more time to adapt to the changes. A reduced rate of epidemic spread also gives policy makers additional time to coordinate the deployment of management capacity.

Vector abundance has not been a focus of cassava breeding programmes. However, it would be possible to promote the adoption of pre-existing cultivars that support lower vector counts, whilst simultaneously selecting for this trait in breeding programmes along with viral resistance.

5.5.4.2 Methods

We take the same infectious distribution used to demonstrate spread from a single infected field near Lagos, Nigeria (Section 5.4.3). We then generate five modified input vector abundance layers by multiplying the original vector abundance raster by a set of reduction coefficients: 0.9 to 0.5 in steps of 0.1. Holding all other inputs constant, we run 1000 simulations for a period of 20 years per reduced vector abundance layer.

5.5.4.3 Results

Figure 5.29 represents the distribution of years that the epidemic results in 25% of Nigerian fields becoming infected with CBSD. Under present day vector abundances, we predict the median time the epidemic reaches 25% of fields to be 4.75 years, when initialised with a single infected field near Lagos. In contrast, reducing the average vector abundance by 50% significantly slows the epidemic spread, reaching 25% of fields after 18.5 years.

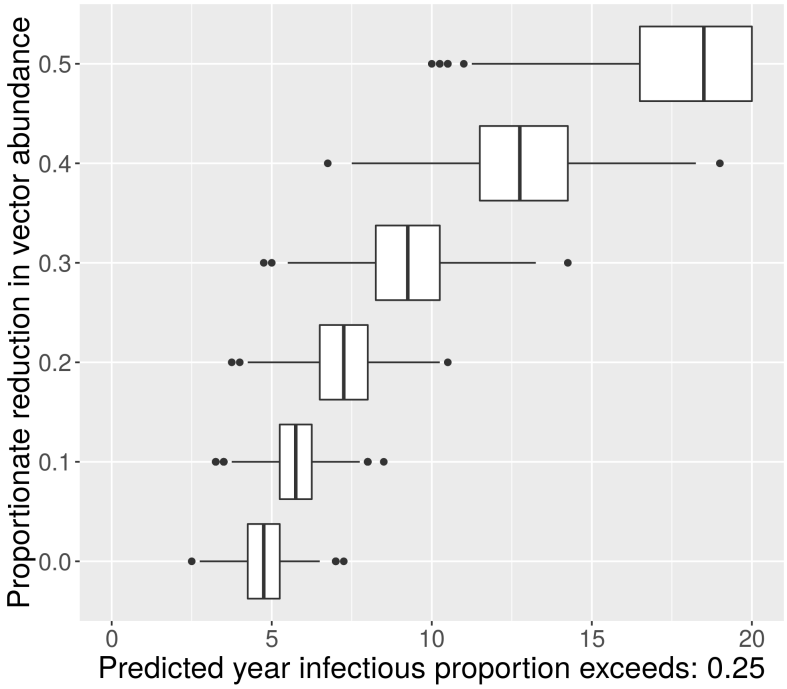


Figure 5.29: The distribution of years in which the epidemic results in 25% of Nigerian fields becoming infected with CBSD, as the vector abundance layer is proportionately reduced.

5.5.4.4 Conclusion

Whilst resistance to both CMD and CBSD are each a major focus of breeding programmes (Kawuki et al., 2016), it is also important to consider the facilitating role of the vector in these epidemics. Moreover, beyond known threats to production, new viral diseases may emerge at any time and will likely be dispersed by known cassava pests. Therefore, breeding for lower vector populations on cassava would likely be a preventative measure against the emergence and establishment of new threats.

5.6 Discussion

In this chapter, we have applied the epidemic model to a range of scenarios that address important questions relating to our understanding of the historic and future behaviour of the CBSD epidemic, with and without management for the disease and vector. We have shown that the historic confinement of the epidemic to the endemic coastal region of East Africa and Malawi can be explained in terms of changes in

cassava production alone. This is not to say that other factors, such as changes in vector abundance, did not play a role, but they are not required to explain the historic trends.

We then applied the model to predicting the spread of the epidemic from the endemic region and Uganda in 2004 through to 2050. As the model is stochastic, multiple simulations generate a distribution of epidemic trajectories. We have constrained the set of predictions according to two sets of assumptions about the real-world epidemic spread: those that conform to the Ugandan surveillance data, and those that conform to both the Ugandan data and observations in western DRC, about which there is less certainty. Respectively, these predictions can be viewed as two possible lower and upper bound estimates of the expanding spread across the continent.

Under the Uganda and DRC assumptions, it seems likely from the predicted 2019 epidemic distribution, that CBSD has reached Northern Zambia and spread through southern DRC, into the north-east corner of Angola. Under the same assumptions, we predict the median year of arrival in Nigeria to be 2046. However, the earliest predicted year of expanding-spread arrival in Nigeria is 2024, via the cross-continental route.

The expanding-spread predictions do not account for disruptive events, such as the direct movement of infected material to West Africa. We simulate the effects of introducing a single field in the vicinity of the two largest cities in West Africa. Our analysis suggests that an introduction near Lagos, Nigeria results in rapid spread throughout the region, with a median prediction of arrival in Ivory Coast of 6 years post introduction.

Subsequently, we undertook a preliminary exploration of the viability of different management and surveillance strategies to prevent spread of CBSD in West Africa. We first considered a buffer strategy, in which we hypothesise the removal or replacement of cassava fields with resistant varieties along the Nigerian border with Cameroon. The results show that a buffer has a minimal impact, only delaying the arrival and bulk-up of the epidemic by a few years at the most, even with a buffer in excess of 100km in diameter. Given the minimal impact and large investment that would be required to implement a buffer region, we conclude that it is not a viable strategy for cost-effective control.

We then moved from a fixed control strategy to dynamic surveillance coupled with control. We compared two scenarios in which surveillance information is gathered by

active surveillance, involving teams of trained surveyors, or through farmer education and reporting, via the extension services. In each of these scenarios, we explored the impact of increasing the scale of surveillance on the national incidence of the epidemic over time and the amount of resistant material that would be required to achieve a given level of control. The results indicate that increasing the level of active surveillance by multiple orders of magnitude does not result in substantial improvements to the efficacy of control. In contrast, improving the level of farmer reporting via the extension system reliably reduces the total amount of resistant clean seed material that is required to reduce the rate of spread and impact of the epidemic. Throughout this section, we focus on the management and control as a means to reduce the impact of the epidemic, as opposed to necessarily eradicating an introduction, as eradication would likely demand unrealistically high levels of intervention.

We finish by quantifying the impact of reductions to vector abundance on epidemic spread. Taking as initial conditions a single infected field outside of Lagos, Nigeria, we ran simulations in which the vector abundance was reduced in multiple increments. Our results indicate that reducing vector abundance by 50% delays the rate of bulk up to 25% of fields by 13.75 years from 4.75 years post-introduction to 18.5 years.

5.6.1 Model limitations

Certain aspects of the model structure constrain the way in which different scenarios can be explored. Firstly, the base unit of the model is the field. In the absence of management, each field is in a binary state of susceptible or infected. Fundamentally, management interventions in the model either replace fields in a given area with susceptible host, or reduce vector abundance. At the landscape scale, the model captures the dynamics of the epidemic well.

However, at the field level, there are behaviours that are not directly captured by the model. For example, if a low number of plants are infected in a susceptible field, the infection may not persist into the second season, assuming the rate of within season bulk-up by vector activity is not high. This is because only a subset of cuttings are selected to replant in the next season. Therefore, the probability of replanting infected cuttings scales with the number of infected plants at harvest time. This type of reversal is likely to occur in regions with low infection pressure and low vector

density. However, the model does modulate field susceptibility and infectiousness by whitefly abundance. Therefore, in regions with low whitefly, susceptible fields are less likely to become infected per challenge, and rates of dispersal from the field are lower.

The model makes two simplifying assumptions about the mechanisms of dispersal. Both trade and vector dispersal are integrated in a single kernel. Existing studies from Nigeria and Uganda suggest that cutting trade is predominantly local (Kansiime, 2014; Teeken et al., 2018), therefore, the unified kernel is unlikely to be prohibitive. However, whilst we assume that trade behaviour is broadly similar amongst smallholders across sub-Saharan Africa, there is little or no quantitative data to reinforce this assumption. As agriculture becomes more commercialised, trade of cuttings over larger distances will likely become more frequent. If required, the model could be extended to include two, separately parametrised dispersal kernels or an explicit network dispersal model. Doing so would require additional data, particularly on trade behaviour, which could be collected by, for example, trade questionnaires.

As demonstrated by vector abundance simulations in Section 5.5.4, epidemic spread predictions are highly dependent on the vector abundance component of the model. The spatial vector abundance layer was generated empirically, by interpolating between average *B. tabaci* counts from surveys across the continent. Therefore, our confidence in the vector abundance layer decreases with distance from real world survey points. Moreover, methodological variations between survey teams or temporal changes have not been taken into account, as well as possible variability in the relative distribution and capacity of different members of the *B. tabaci* species complex to vector CBSVs. We argue, however, that there is advantage in using approximate data, as here, in order to begin to appraise the likely effectiveness of alternative strategies for disease management.

The vector abundance layer could be improved by combining the empirical observations with an environmental suitability model, to generate more accurate predictions of abundance in the regions between survey points. Moreover, if the members of the *B. tabaci* complex are shown to have differential CBSV transmission efficiencies, this could also be incorporated if data on the differential species distributions were available.

Moreover, there are notable climatic differences between Uganda, where the model was parametrised, and the other cassava growing countries across the continent.

However, we feel that across all model components, the differences in climate will predominantly affect vector abundance, and are therefore incorporated via the vector abundance layer. It is possible that higher levels of daily rainfall may affect the probability and distance of vector dispersal. However, there are currently not data in the literature to quantify this possible effect on dispersal.

Lastly, the model assumes cassava throughout the continent is equivalently susceptible to infection by CBSVs. Whilst this is certainly an oversimplification, there is very little available information on the relative susceptibility of cultivars to CBSV or the continental distribution of different cultivars. For example, we could hypothesise that cultivars in West Africa are far less susceptible, or even resistant, to CBSD, which would radically change the dynamics of the unfolding epidemic. If available, information of this kind could be readily integrated into the host structure of the model.

5.6.2 Concluding remarks

In conclusion, the results presented in this chapter provide estimates of the present and future distributions of the epidemic. We highlight the rapid rate at which the epidemic would be likely to spread, once introduced to West Africa and explore management scenarios to identify which approaches are likely to be most effective. In the absence of experimental data, simplifying assumptions have been made about the abundance and transmission capacity of members of the *B. tabaci* complex, as well variability in cassava cultivar susceptibility. Therefore, it is important that epidemiologically relevant datasets in these areas be generated as they may change fundamental assumptions about the pathosystem and, consequently, the interpretation of epidemic dynamics. For now, however, we believe the analyses presented in this section, subject to limiting assumptions, are an appropriate start from which to screen and develop contingency plans to prepare for the invasion of CBSD into West Africa.

Chapter 6

Discussion

6.1 The importance of the cassava brown streak disease epidemic

Cassava is the second most important crop in sub-Saharan Africa, in terms of per capita calorie intake (FAO and IFAD, 2005). Unlike maize, the top source of calories, cassava production is so high, in large part, because of its reliability under poor soil conditions and drought. Moreover, cassava roots do not need to be harvested at a specific time and can be left in the ground for multiple years (Alves, 2002). With most smallholder households not having access to affordable electricity, hence refrigeration, this trait is invaluable.

High levels of food insecurity plague many regions of sub-Saharan Africa, driven by interconnected political, economic, and agricultural factors (Devereux and Maxwell, 2001). Together, these factors drive instability and sabotage economic development. Moreover, this is likely to be the case for many years to come, with climate change predicted to further destabilise the region (Lobell et al., 2008).

The current reliance on cassava, along with its many benefits as a reliable source of nutrition, presents a strong case for production to be protected from current and emerging threats, and where possible, increased. Each country is likely to have a different set of priorities in terms of protecting and increasing cassava production. In the context of the CBSD epidemic, certain countries, such as Tanzania and Uganda, can be considered endemic. In contrast, countries such as Nigeria and those further

west are currently disease-free. However, cassava brown streak disease (CBSD) continues to spread across the continent (Tomlinson et al., 2017). In endemic countries, the focus should be to maximise edible and marketable cassava production. Therefore, the extent to which resources are targeted to CBSD management depends on an assessment of other production inhibiting factors, such as cassava mosaic disease (CMD) and low yielding cultivars. In contrast, in disease-free countries, there is an opportunity to make infrastructural and agricultural changes that could reduce the probability and impact of CSBD establishment.

However, for endemic and disease-free regions, there is a distinct absence of an epidemiologically-informed framework to aid decision making around the optimal CBSD management strategy at local, national and regional scales. The focus of this thesis has been to contribute to the development of such a framework through the synthesis and application of epidemiological information and mathematical modelling.

6.2 Summary of results

In this thesis, we have described the development, parameterisation, and application of a landscape-scale epidemiological model of cassava brown streak disease. We began in Chapter 2 with an audit of the literature on the post-2004 CBSD epidemic, focussing on key epidemiological questions relating to the extent and rate of spatial spread, the limitations of current surveillance methods on our understanding of the epidemic, the impact of CBSD on yield, and the options for management. We highlight a lack of data in the literature on the rate and distances of cutting movements and vector dispersal. Moreover, there are comparatively few studies on the distribution and diversity of the cassava cultivars being grown and their corresponding susceptibility to CBSD, from which to infer large-scale yield loss.

In order to parameterise an epidemiological model, it was essential to have data on the historic spatiotemporal spread of the CBSD epidemic. Chapter 3 describes the process of collating sufficient historic surveillance data for model parameterisation. Initially, very little CBSD surveillance data were available for analysis. The vast majority of data in the literature were summarised to regional statistics without coordinates. Moreover, the raw data that did exist were stored on paper forms in research institutes across multiple countries in East Africa. We therefore undertook an iterative process of identifying and digitising all available data, with Uganda being

by far the largest source of multiple successive yearly snapshots of the epidemic. Extensive post-processing was then performed to minimise random and systematic errors in the digitised dataset. The process of digitising historic paper forms, with the associated multi-year reporting delay, made a strong case for future surveys to adopt digital data collection. We describe the process of developing a digital survey app for CBSD, CMD and whitefly data collection and highlight how in-person training was key to teams in both East and West Africa adopting the app.

We then combined the epidemiological information synthesised in Chapter 2 with the digitised historic surveillance data in Chapter 3 to formulate and parameterise a spatially-explicit, stochastic model of CBSD, as described in Chapter 4. The scale of the pathosystem and the amount of unobserved historic spread meant that methods of parameter estimation using explict likelihoods parameter would not be viable. Therefore, we adopted the more flexible parameter estimation method of Approximate Bayesian Computation (ABC), which replaces the need to formulate the likelihood function directly with multiple simulations of the model (Csilléry et al., 2010).

Approximate Bayesian Computation requires simulation output to be assessed according to a distance measure from the real-world data, with parameters from simulations below a defined threshold passing the rejection algorithm (Algorithm 2). The distance measure commonly involves the use of summary statistics on the simulation output and real-world data. The use of summary statistics and necessity of a tolerance introduces error to the estimation of the posterior distribution. Moreover, the optimal methodology for the derivation of summary statistics remains an open question in the parameter estimation literature (Beaumont, 2010). In order to minimise the introduction of error, we devised a methodology to assess the viability of a given summary statistic. This methodology involved the generation of artificial surveillance data with known parameters. Parameter estimation with the summary statistic in question was then performed to assess the capacity of the given summary statistic to facilitate the recovery of the known parameters. This methodology was applied to evaluate and select a set of robust summary statistics for parameter estimation using the digitised Ugandan national surveillance data.

An SI compartment model structure was selected as the foundation of the CBSD model as cassava is planted from cuttings taken in the previous season, hence replanting of CBSV infected cuttings enables the propagation of the virus across planting seasons. In order to generate a spatial landscape of cassava production, sub-national

cassava production statistics were used to disaggregate levels of production across the given regions according to human population density. From this foundation, the model underwent iterative development and validation to assess the viability of extensions to the model.

The foundational model structure, comprising an SI compartmental structure, with infection rate given in Table 4.1, and the cassava host landscape, was used for parameter estimation. However, subsequent validation simulations from the derived posterior overestimated the levels of infection in the validation dataset in the years 2014 and 2015. The first model extension involved accounting for known error in the survey data. The surveillance protocol implemented for surveys in the digitised dataset specified that only the dominant cultivar in the field should be surveyed. This requirement introduced a bias into the analysis of CBSD presence/absence at the field level. In addition to specifically surveying the dominant cultivar, more recent Ugandan surveys had also noted whether CBSD was present in any other cultivar in the field. Using these data, an approximate average error of 15% was calculated to account for the bias in recording infection only on the dominant cultivar in each field. Inclusion of this error rate in the simulated surveillance scheme, via a binomial trial, into the model improved correspondence between simulated epidemic surveillance data and real-world surveillance data. However, the model still overestimated spread when compared with the 2014 surveillance data.

In the final iteration of model development, the abundance of the vector, *B. tabaci*, was incorporated into the model. Field level vector count averages from surveillance data in East, Central, and West Africa were combined into a spatial vector abundance layer using inverse distance weighted interpolation. For a given spatial location, the corresponding vector abundance layer value then modulated the probability of infection spreading from and to the fields at that location. The incorporation of the vector abundance layer significantly improved correspondence of the predicted surveillance data with real-world observations. During validation, of the simulations that corresponded to data in the fitting period, 2005-2010, 65.7% passed the validation period criteria, from 2011 to 2017. Moreover, predictions showed strong spatial correspondence with the presence/absence patterning in the real-world surveillance data. The improvements in model performance through the incorporation of vector abundance reinforces the consensus in the literature of the importance of *B. tabaci* to the epidemiology of CBSD (Katono et al., 2015; McQuaid et al., 2017).

The model, as developed and validated in Chapter 4, represents a significant advance in our ability to make predictions regarding the spread of the CBSD epidemic in different environments. We, therefore, applied the model to address a number of questions about the historic, current, and predicted spread of CBSD, as well as undertaking a preliminary exploration of management strategies. Key results include an assessment of the cross-continental spread rate under two assumptions around the validity of CBSD symptom observations in north-west DRC in 2017 that are far from other recorded sites of infection in DRC. If we assume these observations are false positives, we generate a probabilistic prediction of the current day, and future, epidemic distribution. Specifically, under this assumption, it seems likely that the epidemic has made progress through southern DRC towards northern Zambia and Angola. However, it seems unlikely to have reached Angola itself or western DRC and beyond. The earliest year of epidemic arrival in Nigeria via the cross-continental route is 2024. However, there is a relatively low probability of spread to West Africa by 2050.

In contrast, if we assume CBSD observations in north-west DRC are correct, the present day epidemic distribution appears likely to have reached the north-eastern border of Angola, made significant progress through south-western DRC and reached central DRC further north. The earliest predicted year of arrival in Nigeria remains 2024. However, the median simulation now predicts arrival in Nigeria in 2046, with a 25% percentile value of 2036.

Cross-continental spread provides a lower-bound estimate of CBSD arrival times in West Africa. It would be entirely possible for infected planting material to be directly introduced to West Africa, most likely via air or sea. We predict the rate of CBSD spread if a single field were planted with infected cassava outside Lagos, Nigeria or Abidjan, Ivory Coast. In each case, the epidemic is predicted to spread rapidly throughout the region. For example, the median year of arrival in Ivory Coast from an incursion in Nigeria is estimated to be 6 years.

Lastly, we perform a preliminary exploration of the comparative efficacy of a set of management strategies. Results indicate that strengthening the capacity for disease reporting via the extension system is likely to be more effective in terms of reducing the size of the epidemic and minimising the amount of resistant planting material that would be required to achieve this outcome. Moreover, an initial assessment of the impact on epidemic spread of a reduction in vector abundance indicates that a 50% reduction in vector abundance on cassava in Nigeria would result in a median

delay of 13.75 years to the year in which spread from a single infected field outside Lagos causes national CBSD incidence to reach 25%.

6.3 Future work

The model in its current form can be applied to address many important and practical questions regarding CBSD spread and management. However, given the scale and complexity of the CBSD pathosystem, the model developed and anlysed here is best viewed as the first step of an iterative process to formalise and centralise our understanding of the CBSD epidemic. Improvements and extensions to the model can be made if either pertinent epidemiological data are generated through experimentation and surveillance, or via improvements on the parameter estimation methodology described in this thesis. We now provide a preliminary assessment of the most beneficial advances that could be made in these two domains. In addition, we highlight infrastructural and technological advances that, if deployed, could greatly aid management efforts.

6.3.1 Experimental studies

6.3.1.1 Quantifying yield loss

Within the existing literature, there are few studies that experimentally quantify the yield loss associated with CBSD. Those that exist largely disagree on the severity, and sometimes nature, of yield loss (Chapter 2). Addressing the question of yield loss has two precursor requirements. Firstly, an understanding of the diversity and distribution of cultivars grown, informed by genetic tests as opposed to expert opinion. Secondly, a standardised methodology is required for the quantification of different types of yield loss in a given cultivar.

Given the scale of the task involved in characterising the true cultivar diversity across the continent, expert opinion would be useful in the short term to identify what are likely to be the most commonly grown cultivars in a given country. For example, selecting the top five cultivars encountered during surveillance to undergo yield loss evaluation. The WAVE consortium is currently attempting to characterise the response to CBSV infection of major cultivars from West Africa in Uganda (J.

Pita, personal communication, 2018). In the longer term, molecular methods should be adopted to increase the objectivity of assessing cultivar diversity.

An improved understanding of the impact of CBSD is not essential to managing the epidemic. However, by either overestimating or underestimating the threat, there is an increased risk of the misallocation of resources. Moreover, information on cultivar distributions and cultivar specific yield loss could be readily incorporated into the epidemiological model, to predict the large-scale yield loss associated with a simulated epidemic.

6.3.1.2 Understanding long-range dispersal

A number of existing studies quantify in-field vector-borne spread of CBSD (Katono et al., 2015; Maruthi et al., 2017). However, there have been very few, if any, studies aimed at quantifying between-field spread of CBSVs by vector dispersal, the approximate distances of between field vector migration, and very little information on the movement of cuttings.

Whilst it may be too challenging to quantify the rate at which viruliferous *B. tabaci* successfully spread infection to other fields, it may be possible to infer this rate by combining transmission experiments with dispersal studies. Firstly, it would be necessary to carry out retention-time experiments with intermediate fasting, as opposed to with intermediate feeding, as with those in the literature (Jeremiah, 2014; Maruthi et al., 2017). A fasting period would be more representative of the conditions experienced during between-field wind-borne dispersal. In previous studies, viruliferious *B. tabaci* were allowed to feed on intermediate host for the duration of the retention period prior to transfer onto cassava plants that will later be tested for infection. As CBSVs are semi-persistently transmitted, it is unsurprising that the vector rapidly loses viral particles when feeding. Secondly, the rate of successful migration from one cassava field to those in the surrounding are could be quantified in additional dispersal studies, similar to those carried out in Byrne (1999) and Colvin et al. (1998). Specifically, it would be of value to characterise the dispersal rate, direction, and distance as a function of wind direction and strength.

6.3.2 Improving management

6.3.2.1 Surveillance in Central Africa

Both the CDP and WAVE projects have generated large amounts of surveillance data across a vast region. However, there is a distinct lack of surveillance occurring throughout Central Africa (Chapter 3). The inclusion of the DRC in the WAVE consortium is certainly a positive step. Yet, given the extremely challenging conditions of the DRC, surveillance has been limited and largely concentrated on small regions. Additional data supported by molecular diagnostics from the DRC, the Republic of Congo, Gabon, Cameroon and CAR on the CBSD distribution and vector abundance would greatly improve our ability to assess the rate of CBSD spread towards West Africa.

Moreover, it is important that whilst focussing on the prevention of spread to West Africa, we do not neglect the more immediate threat of yield loss that CBSD will cause in Central Africa. The DRC has the highest per capita dependence on cassava and has some of the highest rates of poverty in the world (FAO and IFAD, 2005). Therefore, efforts should also be made to put in place realistic management strategies to minimise the burden of CBSD in these regions.

6.3.2.2 Management tools

Currently, there are no commercially available CBSD-resistant cassava cultivars (Tomlinson et al., 2017). Clearly, the development of such varieties would be invaluable to management efforts. In addition, breeding for low vector supporting cultivars would help reduce the rate of CBSD spread, as well as minimising the risk of future viral threats adapting to cassava from wild host plants.

However, the development of a cultivar does not solve the challenging issues of scaling and dissemination. Improvements to the national extension services and clean seed programmes will aid this process. Wherever possible, the market should be leveraged to incentivise the adoption of new technologies, as opposed to the long-term reliance on charity funding.

6.3.2.3 Preparedness and infrastructure

The main themes of this thesis have been around CBSD surveillance and predicting the spread and management of the CBSD epidemic, emphasising what could potentially be done to minimise the spread and impact of the disease. However, it is more important than ever to ensure that improvements in our ability to develop preparedness and management plans translate to real-world applications.

Currently, projects working on cassava improvement are highly fragmented, primarily being funded by many different NGOs, foundations, charities, and governments. There is no central coordinating body to ensure that ongoing and future projects are all pulling in the same direction. It is, therefore, vital that management efforts increasingly become informed by a common understanding of the optimal management strategy. The potential for an African Plant Health Initiative (APHI), currently under consideration by the Gates Foundation, would, if realised, be an ideal candidate for such a role.

6.3.3 Model improvements

6.3.3.1 Model extensions

As described in Chapter 4, the model currently unifies these two dispersal mechanisms of vector-borne and cutting movement, into a single dispersal kernel. It would be possible to add a second kernel in the model to allow discrimination between insect vector and trade movements. However, ensuring that the two dispersal processes are truly disentangled would require additional data on the vector dispersal characteristic and trade behaviour in sub-Saharan Africa. In addition, the current dispersal kernel is isotropic, meaning we assume a symmetric dispersal process. However, insect vector movements beyond the cassava canopy of are largely governed by wind direction. Therefore, future work could focus on the development of an anisotropic vector dispersal process that accounts for the unique characteristics of a wind-bourne dispersal process (Burgin et al., 2013).

6.3.3.2 Parameter estimation

The ABC parameter estimation methodology presented in Chapter 4 has been designed to minimise the introduction of error in the derived posterior distribution. However, there remains an element of subjectivity in the selection of summary statistics. Additional research could be undertaken across pathosystems to develop a canonical set of summary statistics to describe epidemics at the landscape-scale.

The movement from analytic to approximate likelihood methods allows a far wider variety of models to be developed and parameterised (Csilléry et al., 2010). In conjunction with advances in summary statistic selection, it would be valuable to undertake additional research to better understand the comparative performance of analytic verses approximate likelihood methods in large-scale, stochastic, spatial epidemic models (Irvine and Hollingsworth, 2018), and specifically, assess the degree to which differences in the resultant posterior distribution change model predictions.

6.4 Final remarks

Where present, cassava brown streak disease reduces food security and inhibits agricultural development. The development of management options, such as improved varieties, and the strategic deployment of available resources, as aided by epidemiological modelling, has the potential to minimise the impact of this epidemic across the continent, affecting millions of people. Therefore, all stakeholders urgently need to adopt a shared, systematic mechanism of both understanding and updating our understanding of the future impact of CBSD and optimal management strategies. Moreover, advances made in the domain of coordinated CBSD policies simultaneously contribute to the stability of the food system and the long-term prospects for agricultural development.

Bibliography

Abisgold, J. D. and Fishpool, L. D. C. (1990). A method for estimating population sizes of whitefly nymphs (*Bemisia tabaci* genn.) on cassava. *Tropical Pest Management*, 36(3):287–292.

Alicai, T., Omongo, C. A., Maruthi, M. N., Hillocks, R. J., Baguma, Y., Kawuki, R., Bua, A., Otim-Nape, G. W., and Colvin, J. (2007). Re-emergence of cassava brown streak disease in Uganda. *Plant Disease*, 91(1):24–29.

Altieri, M. A. and Rosset, P. (1999). Ten reasons why biotechnology will not ensure food security, protect the environment and reduce poverty in the developing world. *http://www.agbioforum.missouri.edu/v2n34/v2n34a03-altieri.htm*.

Alves, A. A. C. (2002). Cassava botany and physiology. In Hillocks, R. J. and Thresh, J. M., editors, *Cassava: biology, production and utilization*, pages 67–89. CABI, Wallingford.

Bajzelj, B., Richards, K., M. Allwood, J., Smith, P., S. Dennis, J., Curmi, E., and Gilligan, C. A. (2014). Importance of food-demand management for climate mitigation. *Nature Climate Change*, 4:924–929.

Baker, M. C., Mathieu, E., Fleming, F. M., Deming, M., King, J. D., Garba, A., Koroma, J. B., Bockarie, M., Kabore, A., Sankara, D. P., and Molyneux, D. H. (2010). Mapping, monitoring, and surveillance of neglected tropical diseases: towards a policy framework. *The Lancet*, 375(9710):231–238.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406.

Bigirimana, S., Barumbanze, P., Ndayihanzamaso, P., Shirima, R., and Legg, J. P. (2011). First report of cassava brown streak disease and associated Ugandan cassava brown streak virus in Burundi. *New Disease Reports*, 24(26):2044–0588.

Blackmer, J. L. and Byrne, D. N. (1993). Flight behaviour of *Bemisia tabaci* in a vertical flight chamber: effect of time of day, sex, age and host quality. *Physiological Entomology*, 18(3):223–232.

Bock, K. R. (1994). Studies on cassava brown streak virus disease in Kenya. *Tropical Science*, 34(1):134–145.

Braunack-Mayer, L. (2013). Approximate Bayesian Computation and Summary Statistic Selection in Epidemic Models. page 16.

Bull, S. E., Briddon, R. W., Sserubombwe, W. S., Ngugi, K., Markham, P. G., and Stanley, J. (2006). Genetic diversity and phylogeography of cassava mosaic viruses in Kenya. *The Journal of General Virology*, 87(Pt 10):3053–65.

Burgin, L. E., Gloster, J., Sanders, C., Mellor, P. S., Gubbins, S., and Carpenter, S. (2013). Investigating incursions of bluetongue virus using a model of long-distance *Culicoides* biting midge dispersal. *Transboundary and emerging diseases*, 60(3):263–72.

Burr, T. and Skurikhin, A. (2013). Selecting summary statistics in Approximate Bayesian Computation for calibrating stochastic models. *BioMed Research International*, 2013:1–10.

Byrne, D. (1999). Migration and dispersal by the sweet potato whitefly, *Bemisia tabaci*. *Agricultural and Forest Meteorology*, 97(July):309–316.

Campo, B. V. H., Hyman, G., and Bellotti, A. (2011). Threats to cassava production: known and potential geographic distribution of four key biotic constraints. *Food Security*, 3(3):329–345.

Carefoot, G. L. and Sprott, E. R. (1967). *Famine on the wind: man's battle against plant disease*. Rand McNally. Google-Books-ID: 11FJAAAAMAAJ.

Carter, S. E. and Jones, P. G. (1993). A model of the distribution of cassava in Africa. *Applied Geography*, 13(4):353–371.

Casinga, C. M., Monde, G., Shirima, R. R., and Legg, J. (2018). First report of mixed infection of cassava brown streak virus and Ugandan cassava brown streak virus on cassava in north-eastern Democratic Republic of Congo. *Plant Disease*.

Chipeta, M. M., Shanahan, P., Melis, R., Sibiya, J., and Benesi, I. R. M. (2016). Farmers' knowledge of cassava brown streak disease and its management in Malawi. *International Journal of Pest Management*, 0(0):1–10.

Collier, P. and Dercon, S. (2014). African Agriculture in 50 Years: Smallholders in a Rapidly Changing World? *World Development*, 63:92–101.

Colvin, J., Fishpool, L., Fargette, D., Sherigton, J., and Fauquet, C. (1998). Bemisia tabaci (Hemiptera: Aleyrodidae) trap catches in a cassava field in Côte d'Ivoire in relation to environmental factors and the distribution of African cassava mosaic disease. *Bulletin of Entomological Research*, 88:369.

Colvin, J., Omongo, C. A., Govindappa, M. R., Stevenson, P. C., Maruthi, M. N., Gibson, G., Seal, S. E., and Muniyappa, V. (2006). Host-plant viral infection effects on arthropod-vector population growth, development and behaviour: management and epidemiological implications. *Advances in Virus Research*, 67:419–452.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., and François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410–418.

Cunniffe, N. J., Cobb, R. C., Meentemeyer, R. K., Rizzo, D. M., and Gilligan, C. A. (2016). Modeling when, where, and how to manage a forest epidemic, motivated by sudden oak death in California. *Proceedings of the National Academy of Sciences*, 113(20):5640–5645.

Cunniffe, N. J., Laranjeira, F. F., Neri, F. M., DeSimone, R. E., and Gilligan, C. A. (2014). Cost-effective control of plant disease when epidemiological knowledge is incomplete: modelling bahia bark scaling of citrus. *PLoS computational biology*, 10(8):e1003753.

Delêtre, M., McKey, D. B., and Hodkinson, T. R. (2011). Marriage exchanges, seed exchanges, and the dynamics of manioc diversity. *Proceedings of the National Academy of Sciences*, 108(45):18249–18254.

Devereux, S. and Maxwell, S. (2001). *Food security in sub-Saharan Africa.* ITDG Publishing, London.

Downey, A. (2017). *Modeling and Simulation in Python.* Green Tea Press.

El-Sharkawy, M. A. (2004). Cassava biology and physiology. *Plant Molecular Biology*, 56(4):481–501.

Ephraim, N., Yona, B., Evans, A., Sharon, A., and Titus, A. (2015). Effect of cassava brown streak disease (CBSD) on cassava (*Manihot esculenta* Crantz) root

storage components , starch quantities and starch quality properties. *International Journal of Plant Physiology and Biochemistry*, 7(2):12–22.

FAO (2010). Cassava Diseases in Africa: A Major Threat to Food Security. Technical report.

FAO (2016). FAOSTAT statistics database.

FAO and IFAD (2005). A review of cassava in Africa with country case studies on Nigeria , Ghana , the United. *A review of cassava in Africa with country case studies on Nigeria , Ghana , the United Republic of Tanzania, Uganda and Benin. Proceedings of the validation forum on the global cassava development strategy Volume 2*, page 357.

Fargette, V. (1995). Simulation of the effects of host resistance, reversion, and cutting selection on incidence of African cassava mosaic virus and yield losses in cassava. *Phytopathology*.

Fermont, A., Obiero, H., Van Asten, P. J., Baguma, Y., and Okwuosa, E. (2007). Improved cassava varieties increase the risk of soil nutrient mining: an ex-ante analysis for western Kenya and Uganda. pages 511–520.

Fermont, A., van Asten, P., Tittonell, P., van Wijk, M., and Giller, K. (2009). Closing the cassava yield gap: An analysis from smallholder farms in East Africa. *Field Crops Research*, 112(1):24–36.

Filipe, J. A. N., Cobb, R. C., Meentemeyer, R. K., Lee, C. A., Valachovic, Y. S., Cook, A. R., Rizzo, D. M., and Gilligan, C. A. (2012). Landscape epidemiology and control of pathogens with cryptic and long-distance dispersal: sudden oak seath in southern Californian forests. *PLOS Computational Biology*, 8(1):e1002328.

Fry, W. E. (2012). *Principles of Plant Disease Management*. Academic Press. Google-Books-ID: n1kxOTitCAgC.

Gilbertson, R. L., Batuman, O., Webster, C. G., and Adkins, S. (2015). Role of the Insect Supervectors Bemisia tabaci and Frankliniella occidentalis in the Emergence and Global Spread of Plant Viruses. *Annual Review of Virology*, 2(1):67–93.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Gilligan, C. and van den Bosch, F. (2008). Epidemiological models for invasion and persistence of pathogens. *Annual Review of Phytopathology*, 46:385–418.

Gilligan, C. A. (2002). An epidemiological framework for disease management. In *Advances in Botanical Research*, volume 38, pages 1–64. Academic Press.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and Toulmin, C. (2010). Food Security: The challenge of feeding 9 billion people. *Science*, page 1185383.

Gondwe, F. M. T., Mahungu, N. M., Hillocks, R. J., Raya, M. D., Moyo, C. C., Soko, M. M., Chipungu, F. P., and Benesi, I. R. M. (2003). Economic losses experienced by small-scale farmers in Malawi due to cassava brown streak virus disease. pages 28–38. Natural Resources International Ltd.

Gwandu, C. B. (2014). *Epidemiological aspects of cassava brown streak disease in field grown cassava in Coastal regions of Tanzania*. PhD thesis, PhD Thesis, University of Agriculture, Morogoro, Tanzania.

Hanski, I. and Gilpin, M. (1991). Metapopulation dynamics: brief history and conceptual domain. *Biological Journal of the Linnean Society*, 42(1-2):3–16.

Hillocks, R., Maruthi, M., Kulembeka, H., Jeremiah, S., Alacho, F., Masinde, E., Ogendo, J., Arama, P., Mulwa, R., Mkamilo, G., Kimata, B., Mwakanyamale, D., Mhone, A., and Benesi, I. (2015). Disparity between Leaf and Root Symptoms and Crop Losses Associated with Cassava Brown Streak Disease in Four Countries in Eastern Africa. *Journal of Phytopathology*, 2004:86–93.

Hillocks, R. J., Raya, M. D., Mtunda, K., and Kiozia, H. (2001). Effects of brown streak virus disease on yield and quality of cassava in Tanzania. *Journal of Phytopathology*, 149:389–394.

Hutchison, J., Mackenzie, C., Madin, B., Happold, J., Leslie, E., Zalcman, E., Meyer, A., and Cameron, A. (2018). New approaches to aquatic and terrestrial animal surveillance: The potential for people and technology to transform epidemiology. *Preventive Veterinary Medicine*.

Irvine, M. A. and Hollingsworth, T. D. (2018). Kernel-density estimation and approximate Bayesian computation for flexible epidemiological model fitting in Python. *Epidemics*, 25:80–88.

Isaacs, R. and Byrne, D. (1998). Aerial distribution, flight behaviour and eggload: their inter-relationship during dispersal by the sweetpotato whitefly. *Journal of Animal Ecology*, pages 630–639.

Jarvis, A., Ramirez-Villegas, J., Campo, B. V. H., and Navarro-Racines, C. (2012). Is cassava the answer to African climate change adaptation? *Tropical Plant Biology*, 5(1):9–29.

Jeremiah, S. C. (2014). *The Role of the Whitefly (Bemisia tabaci gennadius) in the Transmission and Spread of Cassava Brown Streak and Cassava Mosaic Viruses in the Field and Semi Field Conditions*. PhD thesis, University of Dar es salaam.

Jones, R. A. (2004). Using epidemiological information to develop effective integrated virus disease management strategies. *Virus Research*, 100(1):5–30.

Kansiime, M. (2014). Baseline report on farmers' access to seed and other planting materials. Integrated seed sector development programme in Uganda. page 32.

Katono, K., Alicai, T., Baguma, Y., Edema, R., Bua, A., and Omongo, C. (2015). Influence of host plant resistance and disease pressure on spread of cassava brown streak disease in Uganda. *American Journal of Experimental Agriculture*, 7(5):284–293.

Kawuki, R. S., Kaweesi, T., Esuma, W., Pariyo, A., Kayondo, I. S., Ozimati, A., Kyaligonza, V., Abaca, A., Orone, J., Tumuhimbise, R., Nuwamanya, E., Abidrabo, P., Amuge, T., Ogwok, E., Okao, G., Wagaba, H., Adiga, G., Alicai, T., Omongo, C., Bua, A., Ferguson, M., Kanju, E., and Baguma, Y. (2016). Eleven years of breeding efforts to combat cassava brown streak disease. *Breeding Science*, 66(4):560–571.

Keeling, M. J. and Rohani, P. (2011). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press. Google-Books-ID: LxzILSuKDhUC.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society*, 115(772):700–721.

Kosmowski, F., Aragaw, A., Kilian, A., Ambel, A. A., Ilukor, J., Yigezu, B., and Stevenson, J. (2016). Varietal identification in household surveys : results from an experiment using DNA fingerprinting of sweet potato leaves in southern Ethiopia. Technical Report WPS7812, The World Bank.

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.

Kumakech, A. (2013). Knowledge on cassava disease management: The case of cassava brown streak disease awareness in Northern Uganda. *African Journal of Plant Science*, 7(12):597–601.

Legg, J. and Bouwmeester, H. (2010). Cassava Disease Surveillance Surveys 2009 Mapping Report Great Lakes Cassava Initiative Text and Mapping. Technical report, IITA.

Legg, J., Ndalahwa, M., Yabeja, J., Ndyetabula, I., Bouwmeester, H., Shirima, R., and Mtunda, K. (2017). Community phytosanitation to manage cassava brown streak disease. *Virus Research*.

Legg, J. P. (1999). Emergence, spread and strategies for controlling the pandemic of cassava mosaic virus disease in east and central Africa. *Crop Protection*, 18(10):627–637.

Legg, J. P., Jeremiah, S. C., Obiero, H. M., Maruthi, M. N., Ndyetabula, I., Okao-Okuja, G., Bouwmeester, H., Bigirimana, S., Tata-Hangy, W., Gashaka, G., Mkamilo, G., Alicai, T., and Lava Kumar, P. (2011). Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. *Virus Research*, 159(2):161–70.

Legg, J. P., Kumar, P. L., Kanju, E. E., Tennant, P., Fermin, G., and Others (2015). Cassava brown streak. *Virus Diseases of Tropical and Subtropical Crops*, page 42.

Legg, J. P., Owor, B., Sseruwagi, P., and Ndunguru, J. (2006). Cassava mosaic virus disease in East and Central Africa: Epidemiology and management of a regional pandemic. *Advances in Virus Research*, 67(06):355–418.

Legg, J. P., Shirima, R., Tajebe, L. S., Guastella, D., Boniface, S., Jeremiah, S., Nsami, E., Chikoti, P., and Rapisarda, C. (2014). Biology and management of *Bemisia* whitefly vectors of cassava virus pandemics in Africa. *Pest management science*, 70(April):1446–1453.

Lobell, D. B., Burke, M. B., Tebaldi, C., Mastrandrea, M. D., Falcon, W. P., and Naylor, R. L. (2008). Prioritizing climate change adaptation needs for food security in 2030. *Science*, 319(5863):607–610.

Lose, S. J., Hilger, T. H., Leihner, D. E., and Kroschel, J. (2003). Cassava, maize and tree root development as affected by various agroforestry and cropping systems in Bénin, West Africa. *Agriculture, Ecosystems & Environment*, 100(2):137–151.

Macfadyen, S., Paull, C., Boykin, L. M., De Barro, P., Maruthi, M. N., Ghosh, S., Otim, M., Kalyebi, A., Vassão, D. G., Sseruwagi, P., Tay, W. T., Delatte, H., Seguni, Z., Colvin, J., and Omongo, C. A. (2017). Cassava whitefly, *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae), in sub-Saharan African farming landscapes: a review of the factors determining abundance. *Bulletin of Entomological Research*.

Madden, L. V. (2006). Botanical epidemiology: some key advances and its continuing role in disease management. *European Journal of Plant Pathology*, 115(1):3–23.

Madden, L. V., Hughes, G., and Bosch, F. v. d. (2007). *The study of plant disease epidemics.* American Phytopathological Society (APS Press), St. Paul.

Mahungu, N., M. Bidiaka, Tata, H., Lukombo, S., and N'Luta, S. (2003). Cassava brown streak disease-like symptoms in Democratic Republic of Congo.

Malka, O., Santos-Garcia, D., Feldmesser, E., Sharon, E., Krause-Sakate, R., Delatte, H., van Brunschot, S., Patel, M., Visendi, P., Mugerwa, H., Seal, S., Colvin, J., and Morin, S. (2018). Species-complex diversification and host-plant associations in Bemisia tabaci: A plant-defence, detoxification perspective revealed by RNA-Seq analyses. *Molecular Ecology*, 27(21):4241–4256.

Maroufy, H. E., Omari, L., and Taib, Z. (2012). Transition Probabilities for Generalized SIR Epidemic Model. *Stochastic Models*, 28(1):15–28.

Maruthi, M. N., Hillocks, R. J., Mtunda, K., Raya, M. D., Muhanna, M., Kiozia, H., Rekha, A. R., Colvin, J., and Thresh, J. M. (2005). Transmission of Cassava brown streak virus by *Bemisia tabaci* (Gennadius). *Journal of Phytopathology*, 153(5):307–312.

Maruthi, M. N., Jeremiah, C. S., Mohammed, I. U., and Legg, J. P. (2016). Virus-vector relationships and the role of whiteflies, *Bemisia tabaci*, and farmer practices in the spread of cassava brown streak viruses. *Phytopathology*.

Maruthi, M. N., Jeremiah, S. C., Mohammed, I. U., and Legg, J. P. (2017). The role of the whitefly, *Bemisia tabaci* (Gennadius), and farmer practices in the spread of cassava brown streak ipomoviruses. *Journal of Phytopathology*, 00:1–11.

Mbanzibwa, D. R., Tian, Y. P., Tugume, A. K., Mukasa, S. B., Tairo, F., Kyamanywa, S., Kullaya, a., and Valkonen, J. P. T. (2009). Genetically distinct

strains of Cassava brown streak virus in the Lake Victoria basin and the Indian Ocean coastal area of East Africa. *Archives of Virology*, 154(2):353–359.

Mbewe, W., Tairo, F., Sseruwagi, P., Ndunguru, J., Duffy, S., Mukasa, S., Benesi, I., Sheat, S., Koerbler, M., and Winter, S. (2017). Variability in P1 gene redefines phylogenetic relationships among cassava brown streak viruses. *Virology Journal*, 14:118.

McCallum, E. J., Anjanappa, R. B., and Gruissem, W. (2017). Tackling agriculturally relevant diseases in the staple crop cassava (*Manihot esculenta*). *Current Opinion in Plant Biology*, 38:50–58.

McKinley, T., Cook, A. R., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1).

McQuaid, C. F., Sseruwagi, P., Pariyo, A., and Bosch, F. v. d. (2016). Cassava brown streak disease and the sustainability of a clean seed system. *Plant Pathology*, 65(2):299–309.

McQuaid, C. F., van den Bosch, F., Szyniszewska, A., Alicai, T., Pariyo, A., Chikoti, P. C., and Gilligan, C. A. (2017). Spatial dynamics and control of a crop pathogen with mixed-mode transmission. *PLoS computational biology*, 13(7):e1005654.

Meentemeyer, R. K., Cunniffe, N. J., Cook, A. R., Filipe, J. A. N., Hunter, R. D., Rizzo, D. M., and Gilligan, C. a. (2011). Epidemiological modeling of invasion in heterogeneous landscapes: spread of sudden oak death in California (1990–2030). *Ecosphere*, 2(2):art17.

Mohammed, I. U., Abarshi, M. M., Muli, B., Hillocks, R. J., and Maruthi, M. N. (2012). The Symptom and Genetic Diversity of Cassava Brown Streak Viruses Infecting Cassava in East Africa.

Monger, W. A., Seal, S., Isaac, A. M., and Foster, G. D. (2001). Molecular characterization of the cassava brown streak virus coat protein. *Plant Pathology*, 50(4):527–534.

Mulenga, R. M., Boykin, L. M., Chikoti, P. C., Sichilima, S., Ng'uni, D., and Alabi, O. J. (2018). Cassava brown streak disease and ugandan cassava brown streak virus reported for the first time in Zambia. *Plant Disease*, 102(7):1410–1418.

Mulimbi, W., Phemba, X., Assumani, B., Kasereka, P., Muyisa, S., Ugentho, H., Reeder, R., Legg, J. P., Laurenson, L., Weekes, R., and Thom, F. E. F. (2012).

First report of Ugandan cassava brown streak virus on cassava in Democratic Republic of Congo. *New Disease Reports*, 26:11–11.

Munganyinka, E., Ateka, E. M., Kihurani, A. W., Kanyange, M. C., Tairo, F., Sseruwagi, P., and Ndunguru, J. (2018). Cassava brown streak disease in Rwanda, the associated viruses and disease phenotypes. *Plant Pathology*, 67(2):377–387.

Mware, B., Ateka, E., and Songa, J. (2009a). Transmission and distribution of cassava brown streak virus disease in cassava growing areas of Kenya. *Journal of Applied Biosciences*, pages 864–870.

Mware, B., Narla, R., Amata, R., Olubayo, F., Songa, J., Kyamanyua, S., and Ateka, E. M. (2009b). Efficiency of cassava brown streak virus transmission by two whitefly species in coastal Kenya. *Journal of General and Molecular Virology*, 1(4):40–45.

Ndunguru, J., Sseruwagi, P., Tairo, F., Stomeo, F., Maina, S., Djinkeng, A., Kehoe, M., and Boykin, L. M. (2015). Analyses of twelve new whole genome sequences of cassava brown streak viruses and ugandan cassava brown streak viruses from east africa: diversity, supercomputing and evidence for further speciation. *PLoS One*, 10(10):e0139321.

Ndyetabula, I. L., Merumba, S. M., Jeremiah, S. C., Kasele, S., Mkamilo, G. S., Kagimbo, F. M., and Legg, J. P. (2016). Analysis of interactions between cassava brown streak disease symptom types facilitates the determination of varietal responses and yield losses. *Plant Disease*, 100(7):1388–1396.

Neri, F. M., Cook, A. R., Gibson, G. J., Gottwald, T. R., and Gilligan, C. A. (2014). Bayesian analysis for inference of an emerging epidemic: citrus canker in urban landscapes. *PLoS Computational Biology*, 10(4):e1003587.

Nichols, R. F. W. (1950). The brown streak disease of cassava. *The East African Agricultural Journal*, 15(3):154–160.

Night, G., Asiimwe, P., Gashaka, G., Nkezabahizi, D., Legg, J., Okao-Okuja, G., Obonyo, R., Nyirahorana, C., Mukakanyana, C., Mukase, F., Munyabarenzi, I., and Mutumwinka, M. (2011). Occurrence and distribution of cassava pests and diseases in Rwanda. *Agriculture, Ecosystems & Environment*, 140(3-4):492–497.

Ntawuruhunga, P. and Legg, J. (2007). New spread of cassava brown streak virus disease and its implications for the movement of cassava germplasm in the east and central african region. *USAID, Crop Crisis Control Project C3P*.

Ntawuruhunga, P., Ssemakula, G., Ojulong, H., Bua, A., Ragama, P., Kanobe, C., and Whyte, J. (2006). Evaluation of advanced cassava genotypes in uganda. *African Crop Science Journal (ISSN: 1021-9730) Vol 14 Num 1*, 14.

Okonya, J., Mwanga, R., Syndikus, K., and Kroschel, J. (2014). Insect pests of sweetpotato in Uganda: Farmers' perceptions of their importance and control practices. *SpringerPlus*, 3:303.

Okul Valentor, A., Ochwo-Ssemakula, M., Kaweesi, T., Ozimati, A., Mrema, E., Mwale, E. S., Gibson, P., Achola, E., Edema, R., Baguma, Y., and Kawuki, R. (2018). Plot based heritability estimates and categorization of cassava genotype response to cassava brown streak disease. *Crop Protection*, 108:39–46.

Owusu, V. and Owusu-Sekyere, E. (2014). Assessing the determinants of adoption of improved cassava varieties among farmers in the Ashanti region of Ghana. *Africa Development And Resources Research Institute (Adrri) Journal*, 5:92–104.

Padmanabhan, S. Y. (1973). The Great Bengal Famine. *Annual Review of Phytopathology*, 11(1):11–24.

Parnell, S., van den Bosch, F., Gottwald, T., and Gilligan, C. A. (2017). Surveillance to inform control of emerging plant diseases: an epidemiological perspective. *Annual Review of Phytopathology*, 55:591–610.

Parry, M., Gibson, G. J., Parnell, S., Gottwald, T. R., Irey, M. S., Gast, T. C., and Gilligan, C. A. (2014). Bayesian inference for an emerging arboreal epidemic in the presence of control. *Proceedings of the National Academy of Sciences of the United States of America*, 111(17):6258–62.

Patil, B. L., Delhi, N., Delhi, N., Patil, B. L., Patil, B. L., Legg, J. P., Kanju, E., and Fauquet, C. M. (2014). Cassava brown streak disease: a threat to food security in Africa. *Journal of General Virology*, 96(5):956–968.

Prangle, D. (2015). Summary Statistics in Approximate Bayesian Computation. *arXiv:1512.05633*. arXiv: 1512.05633.

Prangle, D., Blum, M. G. B., Popovic, G., and Sisson, S. A. (2013). Diagnostic tools of approximate Bayesian computation using the coverage property. *arXiv:1301.3166*. arXiv: 1301.3166.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

Rabbi, I. Y., Kulakow, P. A., Manu-Aduening, J. A., Dankyi, A. A., Asibuo, J. Y., Parkes, E. Y., Abdoulaye, T., Girma, G., Gedil, M. A., Ramu, P., Reyes, B., and Maredia, M. K. (2015). Tracking crop varieties using genotyping-by-sequencing markers: a case study using cassava (*Manihot esculenta* Crantz). *BMC Genetics*, 16:115.

Rwegasira, G. and Rey, C. M. E. (2012a). Relationship between symptoms expression and virus detection in cassava brown virus streak-infected plants. *Journal of Agricultural Science*, 4(7):e1916–9760.

Rwegasira, G. M. and Rey, M. E., C. (2012b). Response of selected cassava varieties to the incidence and severity of cassava brown streak disease in tanzania. *Journal of Agricultural Science*, 4(7):237–245.

Savary, S., Ficke, A., Aubertot, J.-N., and Hollier, C. (2012). Crop losses due to diseases and their implications for global food production losses and food security. *Food Security*, 4.

Sseruwagi, P., Sserubombwe, W. S., Legg, J. P., Ndunguru, J., and Thresh, J. M. (2004). Methods of surveying the incidence and severity of cassava mosaic disease and whitefly vector populations on cassava in Africa: a review. *Virus Research*, 100(1):129–42.

Story, H. (1936). Virus diseases of East African plants. VI-A progress report on studies of the disease of cassava. *East African Agricultural Journal*, 2:34–39.

Strange, R. N. and Scott, P. R. (2005). Plant disease: a threat to global food security. *Annual Review of Phytopathology*, 43(1):83–116.

Stutt, R. O. J. H. (2015). *Large Scale Epidemiological Modelling of Invading Plant Pathogens*. PhD thesis, University of Cambridge.

Tajebe, L. S., Boni, S. B., Guastella, D., Cavalieri, V., Lund, O. S., Rugumamu, C. P., Rapisarda, C., and Legg, J. P. (2014). Abundance, diversity and geographic distribution of cassava mosaic disease pandemic-associated Bemisia tabaci in Tanzania. *Journal of Applied Entomology*, (DECEMBER 2014):n/a–n/a.

Tavare, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

Teeken, B., Olaosebikan, O., Haleegoah, J., Oladejo, E., Madu, T., Bello, A., Parkes, E., Egesi, C., Kulakow, P., Kirscht, H., and Tufan, H. A. (2018). Cassava trait

preferences of men and women farmers in Nigeria: implications for breeding. *Economic Botany*, 72(3):263–277.

Thompson, R. N., Cobb, R. C., Gilligan, C. A., and Cunniffe, N. J. (2016). Management of invading pathogens should be informed by epidemiology rather than administrative boundaries. *Ecological Modelling*, 324:28–32.

Thresh, J. M. and Cooter, R. J. (2005). Strategies for controlling cassava mosaic virus disease in Africa. *Plant Pathology*, 54(5):587–614.

Thresh, J. M., Fargette, D., and Otim-Nape, G. W. (1994). The viruses and virus diseases of cassava in Africa. *African Crop Science Journal*.

Timmer, C. P. (2000). The macro dimensions of food security: economic growth, equitable distribution, and food price stability. *Food Policy*, 25(3):283–295.

Tomlinson, K., Bailey, A., Alicai, T., Seal, S., and Foster, G. (2017). Cassava brown streak disease: historical timeline, current knowledge and future prospects. *Molecular Plant Pathology*, 19(5):1282–1294.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian Computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Ugwu, B. O. and Nweke, F. I. (1996). Determinants of cassava distribution in Nigeria. *Agriculture, Ecosystems & Environment*, 60(2):139–156.

van der Plank, J. E. (1963). *Plant Diseases: Epidemics and Control*. Elsevier. Google-Books-ID: HqzSBAAAQBAJ.

Walker, T., Alene, A., Ndjeunga, J., Labarta, R., Yigezu, Y., Diagne, A., Andrade, R., Muthoni, R., De Groote, H., Mausch, K., and others (2014). Measuring the effectiveness of crop improvement research in sub-Saharan Africa from the perspectives of varietal output, adoption, and change: 20 crops, 30 countries, and 1150 cultivars in farmers' fields. *Rome: CGIAR Independent Science and Partnership Council Secretariat*.

Winter, S., Koerbler, M., Stein, B., Pietruszka, A., Paape, M., and Butgereitt, A. (2010). Analysis of cassava brown streak viruses reveals the presence of distinct virus species causing cassava brown streak disease in East Africa. *Journal of General Virology*, 91(5):1365–1372.

World Food Programme (2005). Informal Cross Border Food Trade in Southern Africa, January 2005. Technical report, World Food Programme.

Zhou, C. and Zhou, Y. (2012). Strategies for viral cross protection in plants. *Methods in Molecular Biology (Clifton, N.J.)*, 894:69–81.