Title
Statistical Methods for the Detection and Space-Time Monitoring of DNA Markers in the Pollen Cloud

Author
Marchand, Philippe

Publication Date
2013

Peer reviewed|Thesis/dissertation

Statistical Methods for the Detection and Space-Time Monitoring of DNA Markers in the Pollen Cloud

By

Philippe Marchand

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Science, Policy and Management

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ignacio H. Chapela, Chair
Professor Andrew P. Gutierrez
Professor Anthony R. Byrne

Fall 2013

Abstract

Statistical Methods for the Detection and Space-Time Monitoring
of DNA Markers in the Pollen Cloud

by

Philippe Marchand

Doctor of Philosophy in Environmental Science, Policy and Management

University of California, Berkeley

Professor Ignacio H. Chapela, Chair

The analysis of pollen grains finds applications in fields as diverse as allergology, paleoecology, apiculture and forensics. In contrast with morphological identification methods that require the visual inspection of individual pollen grains, recently-developed genetic approaches have the potential to increase both the scale and resolution of pollen analyses. In the first part of this dissertation, I describe efficient experimental designs to determine the prevalence of a genetic marker in an aggregate pollen sample from the results of DNA amplification by polymerase chain reaction (PCR). The method is based on the theory of limited dilution assays and takes into account potential sources of assay failure such as DNA degradation and PCR inhibition. In the following parts, I show how the genetic composition of air-sampled and bee-sampled pollen can be used to infer spatial characteristics of the floral landscape. Through individual-based simulations of the foraging behavior of honey bees, I obtain theoretical relationships between the genetic differentiation of pollen loads collected at a beehive and the spatial genetic structure of the plant populations visited by foragers. At a larger scale, I present a hierarchical Bayesian model that describes the distribution and spread of common ragweed in France by integrating annual pollen counts from aerobiological stations and presence data from field observations. As the capacity for pollen sampling and analysis increases, these models could be expanded to describe in more detail the biological and physical processes affecting pollen production and transport, and thus provide better predictions for ecological applications such as the control of invasive species.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to first thank my advisor, Prof. Ignacio Chapela, for his outstanding mentorship, his encouragement and guidance at key points throughout this doctoral process. I would also like to thank my dissertation committee members, Profs. Roger Byrne and Andrew Gutierrez, for their careful reading and helpful feedback on my work, as well as Ali Tonak for sharing his expertise on the genetic analysis of pollen.

I acknowledge the financial support I received for this research in the form of graduate scholarships from the Natural Science and Engineering Research Council of Canada and the Fonds de recherche du Québec – Nature et technologies. The modelling effort presented in Chapter 4 was made possible by the public availability of airborne pollen data from the Réseau national de surveillance aérobiologique (RNSA) in France and ragweed presence maps from the Communication and Information Resource Center (CIRCABC) of the European Commission.

I could not have completed this doctoral program without the continuing support of family and friends. I am especially grateful to my parents, Roger and Sylvie, who not only took a sincere interest in my work, but also helped me accomplish fieldwork in Québec during earlier stages of my research. I am thankful to my partner, Jennifer, for her contagious optimism and her understanding through the successes and challenges of the last few years.

If I had the opportunity, like so many other students, to pursue graduate studies, it was due in no small part to the existence of a public higher education system where independent researchers can thrive. I am grateful to the people who built these institutions and the students and faculty who continue to defend their existence and public character to this day.

Chapter 1

Introduction: Pollen mapping in four dimensions

In this dissertation, I introduce and develop statistical methods to determine the genetic composition of pollen samples and relate this composition to spatial and space-time characteristics of plant populations. To motivate these methods and put them into the proper context, I first review the importance of pollen analysis in the environmental sciences and discuss the potential for genetic methods to increase the capacity of pollen analyses.

1.1 The study of pollen and its applications

Pollen grains are the microgametophytes or male gamete producers of flowering plants. In the anthers, a diploid mother cell first undergoes meiosis to produce a tetrad of haploid microscopores. Each microspore undergoes mitosis to produce the two cells forming the pollen grain: a generative cell and a vegetative cell. Upon reaching a receptive stigma, the pollen grain is hydrated and germinates: the generative cell divides in two sperm cells, while the vegetative cell will form the pollen tube (McCormick 1993). In some species, the generative cell divides prior to germination to form a tricellular pollen grain: this class includes agriculturally important genera such as *Zea* (maize), *Triticum* (wheat) and *Brassica* (canola, cabbage, turnip) (Matthys-Rochon et al. 1987).

Multiple factors contribute to make pollen analysis a powerful tool for studying the spatio-temporal dynamics (distribution over time) of plant species of interest. The pollen grain's outer wall or *exine* is characterized by its high level of morphological differentiation as well as its exceptional chemical stability (Scott 1995). These properties make pollen a key source of evidence for paleoecologists and paleoclimatologists; ancient pollen preserved in lake sediments, for example, provides a historical record of the vegetation in the surrounding area. Reference texts in palynology (e.g. Erdtman 1986) describe the specific morphological

features that can be used for taxonomic identification up to the family, sometimes even genus, level.

For studies of contemporary plant dispersal and distributions, pollen sampling offers a powerful way to collect aggregate information from a landscape. Wind-pollinated species produce very large quantities of pollen – over a million grains per plant per day for maize at its peak (Jarosz et al. 2003) – and disperse it over typical distances ranging from hundreds of meters to kilometers, although there is evidence for rare long-distance transport at a continental scale (Rousseau et al. 2008). The "pollen cloud" present at a given location can be sampled either through a passive (e.g. slide deposition) or active device (e.g. sunction trap, rotating rod impaction). Colonies of social bees also accumulate pollen from their foraging area, and the study of pollen found in honey (melissopalynology) is a well-established subdiscipline of palynology (Von der Ohe et al. 2004).

Despite the challenges involved in extracting information on pollen sources from aggregate samples (such as those collected by air samplers and bees), pollen analysis offers the prospect of learning about plant sources that cannot be sampled directly by the researcher. As I will discuss in the next chapter, it is also possible to perform genetic analysis of pollen samples faster and at a cheaper cost than for bulk plant samples.

The abundance and composition of pollen in an ecosystem impacts more than just plant dispersal and reproduction. Pollen is the primary source of protein for a wide range of herbivorous insects (Romeis et al. 2005). Since the pollen of some transgenic crops, such as the Bt varieties, expresses insecticidal proteins, determining the pollen exposure from these crops is necessary to evaluate their impact on non-target insect populations (Rose et al. 2007; Stanley-Horn et al. 2001). From a human health perspective, the need to monitor and forecast the concentrations of allergenic pollen may constitute the most important and most publicly known purpose of pollen sampling and mapping.

Pollen identification has even found applications in forensics, as the detection of rare pollen types on clothing and shoes can be used to corroborate or contradict assertions on the whereabouts of individuals (Horrocks and Walsh 1998).

1.2 Genetic methods for pollen analysis

As mentioned in the previous section, visual inspection of the exine (pollen outer wall) has been the primary method for taxonomic classification of pollen. Pollen grains can be isolated from environmental samples through some version of the acetolysis procedure: samples are treated with strong acids to remove most organic contaminants, while the exine's exceptional chemical stability allows its essential morphological characteristics to be preserved. Pollen is then stained and fixed to microscope slides for identification (Agashe and Caulton 2009).

Morphological identification and counting of pollen under a microscope is a labor-intensive

process, requiring specialized training to recognize the distinctive features of various pollen types. Even in the best case, identification can only be carried to the genus level. Recently, methods based on pollen autofluoresence have been proposed to largely automate the pollen identification process and increase its taxonomic resolution, although these techniques are still at the prototype stage (Mitsumoto et al. 2009; Pan et al. 2011).

In this dissertation, I focus on the possibilities offered by the genetic analysis of pollen samples. Genetic approaches can be used to discriminate between pollen types without relying on characteristic features of the exine morphology, autofluorescence spectrum or some other visible phenotype. These methods are also versatile: the same techniques used to monitor pollen could potentially be adapted to other questions of biological interest, such as the dispersal of fungal spores or the spread of airborne diseases.

The last few years have seen an increase in the use of single-pollen DNA amplification techniques to identify the genotype of individual pollen grains (Isagi and Suyama 2011). Due to the low DNA copy number, single-pollen polymerase chain reaction (PCR) has a relatively high failure rate. As a result, the time and cost involved in applying this technique to large ecological samples can be prohibitive. However, Bektas and Chapela (2013) have shown that the use of loop-mediated isothermal amplification (LAMP) instead of traditional PCR can simplify the amplification process as well as improve the reliability of single-pollen genetic identification. These recent developments raise the prospect of a high-throughput identification method for pollen that could be used for ecological applications.

1.3 Pollen pools and flows in ecosystems

The anthers of flowering plants (and the corresponding microsporangia in gymnosperms) are the only source of pollen in the environment. Pollen is released from the anther either by the wind or animal pollinators, which constitute its two main dispersal vectors. Pollen can terminate its course on a receptive plant stigma, get degraded on the ground or some other surface, be buried and preserved in sediments, or get consumed by living organisms.

It can be useful then to think of pollen as being partitioned in different pools: pollen on anthers, in the air, on pollinators, etc. Figure 1.1 presents a graphical summary of those pools, along with the main biological and physical processes controlling the pollen distribution within and between pools. This representation is not meant to be exhaustive[1], but it provides a conceptual framework linking together different areas of pollen research.

[1]For example, it doesn't cover water transport of pollen, or "buzz pollination", where bee movement releases pollen in the air.

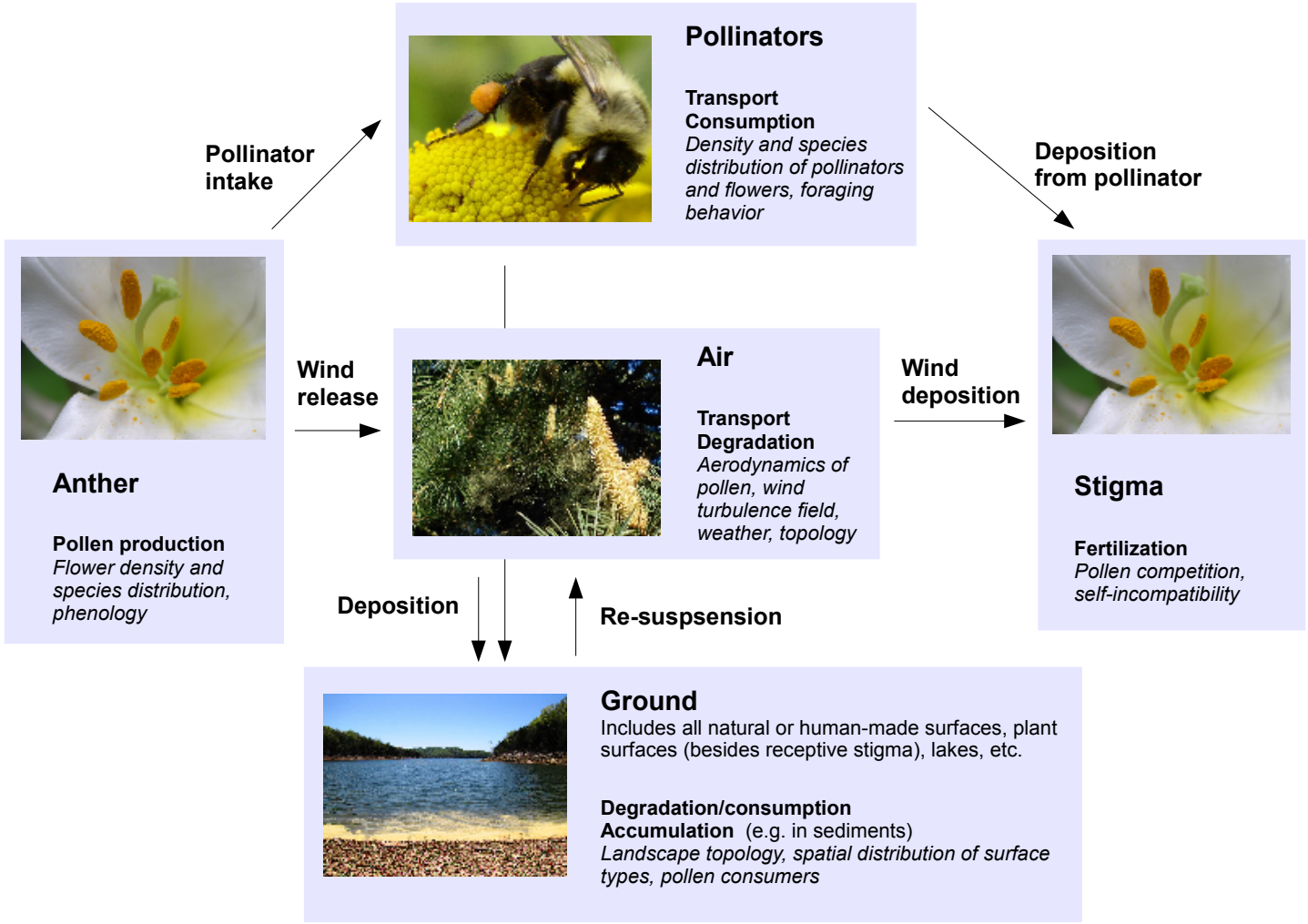

Figure 1.1: Pollen pools and flows in ecosystems. Each pollen pool (box) has distinct processes governing the production, destruction and/or transport of pollen (in bold), and these processes depend on biological and physical properties (in italics). Arrows represent movement of pollen between the main pools. Photos by Lukas Riebling (left/right), Polinizador (top), Chiswick Chap (middle) and Brian Stansberry (bottom) from Wikimedia Commons used under CC BY 3.0.

In this section, I review the sampling and analysis methods used to study three major processes: the aerial transport of pollen, its transport and consumption by animal pollinators, and its accumulation in sediments.

1.3.1 Wind dispersal

Estimating the density and composition of airborne pollen is a prerequisite for both the development of theoretical models of wind pollination and the production of pollen maps, such as those used in allergy forecasts. The Rotorod and the Hirst trap are two samplers commonly used for this purpose: the former captures airborne particles on two rotating arms covered with silicon grease, while the latter uses a vacuum pump for air intake and collects particles on an rotating drum coated with adhesive tape. In principle, pollen counts obtained from these volumetric samplers are directly proportional to the concentration in the air; in practice, their sampling efficiency can vary based on meteorological conditions or particle size (Frenz 1999, 2000). When absolute pollen concentrations are not required (e.g. one is only interested in the relative amount of pollen deposition at different sites), passive samplers of the Durham type can be a cost-effective option: a microscope slide coated with petroleum jelly is set between two parallel plates, which protect the slide from rain and ensure a more constant air flow (Durham 1944).

A comprehensive review of wind pollination models is outside the scope of this introduction; instead, I focus on presenting examples of the different modelling approaches. The majority of those studies pertain to maize: this is due in part to labelling laws for transgenic food products in specific markets (such as the European Union), which created an economic incentive for research aimed at predicting and minimizing cross-pollination between transgenic and non-transgenic fields.

Pollen dispersal models can be broadly divided in two classes: empirical or mechanistic. Empirical models select among simple parametric curves to fit dispersal data. For example, Gustafson et al. (2006) have used data from different field studies to estimate the effect of isolation distance and border rows on pollen-mediated gene flow between transgenic and non-transgenic maize fields. While these empirical models have been more successful than mechanistic ones at reproducing observed data, their functional forms and parameters values are not based on specific biological or physical processes. As a result, their conclusions cannot be readily extended to experimental conditions outside of those used to derive the model (Beckie and Hall 2008).

Mechanistic dispersal models can be classified as Eulerian or Lagrangian, based on whether they focus on the evolution of a pollen density field or the stochastic movement of individual grains, respectively.

Eulerian models are generally described as a system of partial differential equations that approximate atmospheric fluid dynamics. A basic example is given by Gaussian plume mod-

els, which reduce turbulence to a mean transport (advection) term and a diffusive term. In reality, since gusts of wind can transport pollen over exceptionally long distances, the observed dispersal functions are strongly leptokurtic, i.e. with greater probability of extreme values. Loos et al. (2003) account for this by summing two Gaussian plumes (one for short-range, one for long-range dispersal). More complex models account for some of the correlations inherent in the turbulence field, as well as the effect of heterogeneous plant canopies on air flow (Dupont et al. 2006).

Lagrangian models use stochastic differential equations (including deterministic and random terms) to describe the motion of pollen grains in three dimensions. The dispersal kernel, or fraction of pollen deposited at a given distance from its source, is obtained by simulating a large number of individual pollen trajectories. In a random walk or Brownian motion model, the change in each coordinate at each time step is taken from a normal distribution with a given mean and variance (Klein et al. 2003; Tufto et al. 1997). Random walk models (much like Gaussian plumes) reduce turbulence to a diffusive process and ignore its coherent structure. More realistic models apply a random variation to the velocity vector, rather than the position (Aylor and Flesch 2001; Aylor et al. 2003; Jarosz et al. 2004).

Both first-order (random walk) and second-order Lagrangian models require as an input some statistical knowledge of the turbulence field. These parameters can be found by fitting the model to dispersal data (Klein et al. 2003), by adding an Euleriance turbulence model (Nathan et al. 2002; Arritt et al. 2007) or by using a measured time series of the wind velocity (Kuparinen et al. 2007).

A typical field experiment used to either fit or validate dispersal models consists in a small ($<$ 100x100m) source patch located in the middle of a larger receptor field (Arritt et al. 2007; Dupont et al. 2006; Klein et al. 2003). An important question is whether the resulting dispersal curves apply for larger scales than those where they were established. One application of these dispersal curves is to predict cross-pollination and gene flow at the landscape scale, by summing the individual contributions from all pollen sources (Angevin et al. 2008). Validating these models at the landscape scale would require sampling very large quantities of pollen; this is an area where high-throughput methods for pollen analysis and classification could become useful.

1.3.2 Transport and consumption by animal pollinators

If the complexity of atmospheric turbulence limits the measurement and prediction of wind-mediated pollen transport, understanding animal-mediated transport may be an even greater challenge, based on both the diversity of pollinator species and their highly situational behavior. In this section, I review models and methods used in the study of pollinator foraging behavior and the associated pollen flow. While most of the research cited here focuses on *Hymenoptera* and particularly social bees, the range of animal pollinators extends beyond insects to vertebrates such as hummingbirds and bats (Castellanos et al. 2003; Knudsen and

Tollsten 1995).

Pollinator visits can be recorded either by observation in a fixed quadrat or by walking transects through the field, in order to estimate the visit rate per flower (Dafni 1992). Direct observation provides information on pollinator behavior at a small scale, such as the distance and direction change between flowers visited (Goulson 2000; Cresswell and Osborne 2004). Large scale parameters, such as the foraging range, must be determined by indirect means. Individuals can be marked and recaptured (Walther-Hellwig and Frankl 2000), and some larger pollinators (e.g. bumblebees) can be tracked using radar transponders (Osborne et al. 1999). In the special case of honey bees, researchers can infer the distance and direction of foraging sites by decoding the waggle dance of foragers returning to the hive (Visscher and Seeley 1982; Steffan-Dewenter and Kuhn 2003).

Some attempts have been made to find general principles underlying foraging behavior, under the topic of optimal foraging theory (Pyke 1978). This approach is based on the idea that foraging animals, including pollinators, would have evolved to search and collect resources in the most energy-efficient way. Optimal foraging theory provides an adaptative rationale for some observed patterns of behavior, but to my knowledge has not been used to make predictions about foraging range or other parameters.

Insect pollinators can be captured in the field using a net or a number of available traps. Pollen is typically collected from the insect body using cellophane tape, washing with ethanol, or ultrasound treatment. Corbicular pollen pellets, located on the hind legs of honey bees and bumblebees, are readily separated from the insect, although it should be kept in mind that this pollen is not available for pollination of further plants on the bee's path (Dafni 1992).

Neither the movement patterns of pollinators nor the analysis of their pollen loads provides us with direct information on typical pollination distances. In their model of cross-pollination in canola, Cresswell et al. (2002) use the result of an experiment where a single bumblebee was left to forage in a small patch of herbicide-resistant canola plants, then moved to a patch of non-resistant plants. By recording the order of non-transgenic plants visited and the proportion of resistant offspring, they determined the proportion of pollination to due to each previous plant in the visit sequence, which they call the "paternity shadow". This result can be combined with knowledge of foraging behavior to produce a model for gene flow for that particular combination of pollinator and pollinated species.

In addition to the social organization of their foraging behavior, another important characteristic of honey bees is the accumulation of collected pollen in the hive. On a given foraging bout, worker bees are specialized as either nectar or pollen foragers: although the former do carry some pollen and contribute to cross-pollination, the latter will visit plants specifically for their pollen (including nectarless, wind-pollinated species), collecting about 15mg of pollen in two corbicular pellets (Pouvreau 2004).

Commercial honey bee hives can be outfitted with traps to collect corbicular pollen

loads from returning foragers. In some circumstances, analyzing the composition of these pollen loads provides information on the foraging range, when one taxon has a very limited distribution in the surrounding region (Beil et al. 2008), or when researchers are able to detect genetic markers with a known source (e.g. transgenes, as in Ramsay et al. 2003).

It is estimated that a hive of 50,000 bees will consume about 40kg of pollen in a year (Pouvreau 2004). Within the hive, pollen is stored in separate combs than nectar/honey and tends to be closer to the brood area. Compared to honey, pollen tends to be collected in bursts as surrounding plants go through their pollination peak. This creates temporary accumulations of pollen in the hive; however, virtually all pollen is consumed by the end of the season, while about 40% of the honey is saved over winter (Camazine 1991). This means that pollen sampled from the hive combs at any one time may not be representative of the whole season to date. As for the trace amounts of pollen that are found in honey combs, they may have been deposited at various times by the movement of pollen-covered bees, or even from gusts of wind passing through the hive.

1.3.3 Ground deposition and accumulation

The vast majority of wind-dispersed pollen does not reach a female inflorescence, instead being deposited on some other part of the canopy, on the ground or in water. Although there is good evidence that gusts of wind can resuspend some of this deposited pollen (Tauber 1967; Aylor et al. 2003), rarely is this phenomenon explicitly considered in dispersal models (Jarosz et al. 2004).

The accumulation and conservation of pollen grains at the bottom of water basins (such as lakes and bogs) has provided an important tool for paleoecologists who investigate the history of vegetation cover at a specific site.

Pollen can be extracted from various types of sediments or even sedimentary rocks such as lignite or coal. Processing techniques vary with sample type. Sedimentary rocks are generally crushed and demineralized by soaking in strong acids, while the organic matter portion is dissolved in a heated alkali solution. As an alternative or in conjunction with chemical treatments, gravity separation techniques may be used, as pollen is typically lighter than the mineral matrix, but heavier than the organic matter fraction (Agashe and Caulton 2009).

Most efforts to relate the representation of taxa in pollen samples to the composition of vegetation surrounding the sampling sites start with the model of Prentice and Sugita (Prentice 1985; Sugita 1993, 1994). If pollen is sampled at different sites within a landscape, the quantity of pollen from species i deposited at site k (y_{ik}, the pollen load) can be described by a function of the form:

$$y_{ik} = \alpha_i \int_{z=R}^{Z_{max}} x_{ik}(z) g_i(z) \mathrm{d}z + \omega_{ik}, \tag{1.1}$$

where α_i represents the pollen productivity of species i, $x_{ik}(z)$ is the average density of

species i at a radial distance z from sampling site k, and $g_i(z)$ is the dispersal kernel (as defined in the "wind dispersal" section above) for species i. The integral is taken from the edge of the basin (assumed to be circular, with radius R) to an arbitrary limit Z_{max}. Any pollen contributed by sources outside that limit radius is included in the background pollen term, ω_{ik}.

Assuming that this background pollen contribution is constant from site to site, then the pollen load of species i is a linear function of the distance-weighted plant abundance (DWPA), the integral term in eq. (1.1). The size of the source area represented in a pollen sample increases with the size of the basin. This idea is the basis of a two-stage algorithm developed by Sugita (2007a,b) to estimate differences in species composition between sites located in the same region: pollen from large lakes is used to determine the pollen background ω_i and this background level becomes an input when determining the species composition around smaller basins, using eq. (1.1) .

With knowledge of the relative pollen production rates of each species (the α_i), the Prentice-Sugita model provides a simple link between collected pollen samples and the DWPA. However, the DWPA itself is a convolution of species distribution and dispersal dynamics, hence it is difficult to extract the species distribution itself without making simplifying assumptions (e.g. isotropic dispersal, spatial homogeneity).

Instead of reconstructing past vegetation from pollen data, an alternative approach is to simulate the pollen load resulting from a given hypothetical landscape, then compare the output with observed pollen counts. This method can incorporate more complex landscape features, including topography, heterogeneous (patchy) vegetation and non-isotropic dispersal (Bunting and Middleton 2005; Fyfe 2006). Such simulations allow researchers to determine if a species distribution is consistent with the pollen record, although it should be kept in mind that many distributions could produce very similar pollen loads.

1.4 Objectives of this dissertation

In the previous section, I have reviewed the main physical and biological processes that determine the composition of environmental pollen samples. The essential steps required to extract information on these processes from pollen analysis are shown in Fig. 1.2. This is a general workflow that could apply to other ecological questions. The underlying spatio-temporal process could be the wind-mediated dispersal function for a given species, the spatial genetic structure of a plant population, or the movement of bees during foraging, to give a few examples.

The two sets of arrows in the diagram represent the complementary uses of statistical models. At the design stage, a tentative model for the process of interest is used to choose an appropriate sampling scheme in the field, and the nature of the collected samples is used

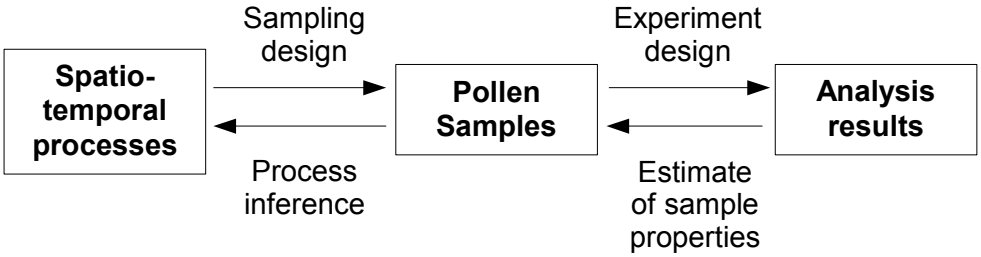

Figure 1.2: A workflow model for pollen analysis: questions about spatio-temporal processes inform the sampling design, and the nature of samples collected informs analytical methods in the laboratory. The experimental results from these methods are used to estimate the composition of samples, which is then used to make inferences about the processes of interest.

to design an experimental assay in the laboratory. At the inference stage, the same models are used to estimate properties of the sample from the assay results, and then make inference about parameters or other characteristics of the underlying spatio-temporal process.

The statistical models in this dissertation address both the problem of sample analysis and that of process inference. These methods only require the capacity to determine the presence or absence of a specific genetic marker from a pollen sample. Their aim is to answer autecological questions pertaining to the abundance, distribution and dispersal of individual species or genotypes.

In Chapter 2, I propose a method based on limited dilution assays to estimate the frequency of a specific genetic marker in a pollen sample from the binary outcome of endpoint PCR reactions. I calculate the uncertainty of this estimate under different models of the PCR error rate, and use the results to define a strategy for the design of efficient assays, i.e. those providing the most precise estimate for a given experimental effort.

The following two chapters present applications of pollen analysis to infer spatial characteristics of plant populations.

The composition of honey bees' corbicular pollen loads contains information on both the bees' foraging behavior and the floral resources surrounding the hive, but so far there has been little emphasis on describing these relationships through quantitative models. In Chapter 3, I develop an individual-based model to relate the distribution of a genetic marker in bee pollen to parameters describing foraging behavior and the spatial distribution of plants bearing that marker.

Just as fossil pollen analysis has become an essential tool for reconstructing past vegetation, contemporary pollen records could help us better understand the dynamics of extant plant populations. In Chapter 4, I discuss the use of airborne pollen records for monitoring the spread of invasive species, focusing on the case of common ragweed (*Ambrosia artemisiifolia*) in France. I use a hierarchical modelling approach to estimate the spatial and temporal variation of the plant's density from the combination of two datasets: pollen concentrations measured at a sparse set of stations over the last decade, and the presence data from a one-time national survey.

Chapter 2

Efficient designs for PCR analysis of pollen samples

In the previous chapter, I discussed some advantages of genetic methods for determining the composition of pollen samples. Only recently (around 15 years ago) have researchers developed the capacity to amplify DNA from single pollen grains, using the polymerase chain reaction (PCR). While the small number of DNA copies (2 or 3) in each pollen grain represents the main challenge of single-pollen PCR (abbreviated from hereon as spPCR), it requires a much simpler DNA extraction step, compared with methods using bulk plant tissue (Petersen et al. 1996).

A recent compilation of studies on single-pollen genotyping by Isagi and Suyama (2011) shows how spPCR can be used to analyze pollen sampled in the air, from the bodies of insects and birds, from the styles of flowers, and even ancient pollen preserved in lake sediments.

In the study of plant dispersal and pollination, a key advantage of spPCR is its phylogenetic resolution. For example, researchers succeeded in amplifying multiple microsatellite markers from individual pollen grains (Matsuki et al. 2011; Kondo et al. 2011). However, these studies also illustrate one of the drawbacks of this technique: only a very small fraction of the effective "pollen rain" can be sampled.

Consider the problem of monitoring the dispersal of a specific genetic trait or allele in an insect-pollinated species. One possibility is to collect insect pollen loads and analyze them as bulk samples through standard DNA extraction methods (see for example Ramsay et al. 2003, a study of gene flow in oilseed rape). This technique could be too coarse at the individual level, since each insect provides a single binary (presence/absence) data point. At the other extreme, one could pick individual pollen grains from each insect an perform spPCR. For 30 insects and an average of 50 grains/insect, a sample size typical of the studies cited above, this amounts to 1,500 PCR reactions, a considerable experimental effort. While this approach provides an estimate of the genetic marker frequency in each pollen load, it

could fail to detect rare alleles (e.g. present at a frequency below 1%) and could be difficult to interpret statistically. (What does it mean to "randomly" pick pollen grains from an insect?)

When researchers are interested in determining the frequency of a genetic marker in some pollen sample, an alternative approach is to perform a series of PCR reactions, while varying the number of pollen grains in each reaction. If the original sample is well-mixed, the frequency of the marker in that sample may be inferred from a statistical model. This method is analogous to the limiting dilution assays (LDA) commonly used in microbiology and immunology (Taswell 1981).

In this chapter, I develop a LDA-like model specific to pollen PCR analysis. Model parameters are determined by PCR experiments performed on samples of known composition. By calculating the model's performance (bias, variance) at estimating the unknown marker frequency as a function of different experiment designs, I provide specific criteria for efficient design of pollen PCR assays.

2.1 Statistical theory of dilution assays

Limiting dilution assays (LDA) (or serial dilution assays, as both terms are used almost interchangeably in the literature) are a type of experiment design commonly used to estimate the number or concentration of a specific type of cells in a sample, when the underlying test only produces a binary (presence/absence) response (e.g. growth medium to determine presence of culturable cells, or antibody test to determine presence/absence of a given infection). The LDA design consists in making successive dilutions of the original sample and replicating the assay multiple times at each dilution level. As theory predicts mostly positive results at low dilutions and negative results at sufficiently high dilutions, the statistical behavior between these two extreme cases can be used to estimate the initial concentration of target cells.

Statistical methods to analyze LDA results were developed in the early 1980s by Taswell (1981) and Fazekas de St. Groth (1982), although a brief discussion of the problem is included in Fisher's original work on maximum likelihood estimation (Fisher 1922). For the question of pollen identification by PCR, the parameter of interest (ϕ) is the fraction of pollen grains in the sample bearing a specific DNA marker. The equations presented here are based on Myers et al. (1994), who also consider the LDA parameter as a proportion.

The main assumption behind the classical analysis of LDA results is that the number of target cells per replicate follows a Poisson distribution, and that a positive result occurs for all replicates containing at least one target cell (the so-called "Poisson single-hit model"). The Poisson distribution is generally fit for representing rare, independent events. In the LDA context, the rarity condition means that the probabilty of any specific cell being represented in each replicate is low, or said otherwise, that each replicate only represents a small volume

of the whole sample. The independence condition requires that target cells be distributed randomly within the sample; any clustering (cells sticking closely together) or partitioning (cells "repelling" each other) effect must be negligible. The single-hit criterion excludes the possibility of either false positive (positive signal when no target cell is present) or false negative (negative signal when one or more target cells are present) errors.

2.1.1 Maximum-likelihood analysis of dilution assay data

Suppose a LDA is conducted with D dilution levels and R_d replicates at the d^{th} dilution level. Under the Poisson single-hit model, if λ_d is the average number of cells per replicate at the d^{th} dilution level and ϕ is the fraction of target cells among all cells in the sample, then the probability of getting no target cells in a replicate is $e^{-\phi\lambda_d}$. Therefore, the probability p_d of getting a positive result for any given replicate at the d^{th} dilution level is:

$$p_d = 1 - e^{-\phi\lambda_d}. \tag{2.1}$$

Assuming all replicates are independent, the probability of getting k_d positive results at the d^{th} dilution level is given by the binomial distribution with R_d tries and probability of success p_d. The joint probability of the set $\{k_d\}$ (results for all D dilution levels) is obtained by multiplying the probabilities at each level. Therefore, given the results of one LDA experiment, the likelihood ($\mathcal{L}$) of the (unknown) parameter ϕ is calculated as:

$$\mathcal{L}\big(\phi; \{k_d\}\big) = \prod_{d=1}^{D} \binom{R_d}{k_d} p_d(\phi)^{k_d} \big(1 - p_d(\phi)\big)^{R_d - k_d}, \tag{2.2}$$

where it is emphasized that p_d depends on ϕ through eq. (2.1).

The same likelihood equation (2.2) also applies for experiments where the number of tested units (e.g., cells) at each "dilution" level is known, rather than being the outcome of a Poisson process.[1] One example is the testing of seed lots for infections, when the number of seeds in each replicate can be exactly counted (Ridout 1995). If N_d is the number of tested units at the d^{th} level, then the probability of not having a single unit responding to the test is $(1-\phi)^{N_d}$, and eq. (2.1) for p_d is replaced by:

$$p_d = 1 - (1 - \phi)^{N_d}. \tag{2.3}$$

The estimate of ϕ ($\hat{\phi}$) that maximizes the likelihood of the observed data corresponds to a root (zero) of the score function U, which is the derivative of the logarithm of $\mathcal{L}$:

$$U(\phi) = \frac{\partial \ln \mathcal{L}}{\partial \phi} = \sum_{d=1}^{D} \left[\frac{k_d}{p_d(\phi)} - \frac{R_d - k_d}{1 - p_d(\phi)} \right] \frac{\mathrm{d}p_d}{\mathrm{d}\phi}. \tag{2.4}$$

[1] For simplicity, I will refer to "dilution level" and "limited dilution assay" for those experiments as well, even though the sets of replicates are not produced by successive dilutions.

In general, finding the root of U requires the use of a numerical algorithm (e.g. Newton-Raphson).

2.1.2 Bias correction for the maximum likelihood estimator

Given one set of outcomes $\{k_d\}$, applying the method above will yield a single estimate $\hat{\phi}$. While this single estimate will likely deviate from the true value of ϕ, the estimator is said to be unbiased if its expected value, taken over all possible experimental outcomes, equals the true value of the parameter: $\mathbf{E}[\hat{\phi}] = \phi$.

Unfortunately, the maximum likelihood estimator (MLE) of $\hat{\phi}$ obtained from eq. (2.4) is generally biased. This is easily illustrated in the case where $d = 1$, where the probability of getting k positives of of R replicates is a simple binomial distribution, and the MLE of p:

$$\hat{p} = \frac{k}{R}$$

is unbiased. The MLE of ϕ is found by inverting either eq. (2.1) or (2.3), depending on the model used:

$$\hat{\phi}(\hat{p}) = -\frac{\ln(1-\hat{p})}{\lambda}; \quad \text{or}$$
$$\hat{\phi}(\hat{p}) = 1 - \sqrt[N]{1-\hat{p}}.$$

In either case, ϕ is a convex function of p, so that by Jensen's inequality:

$$\mathbf{E}[\hat{\phi}(\hat{p})] \geq \hat{\phi}\big(\mathbf{E}[\hat{p}]\big) = \phi.$$

Therefore, the curvature of $p(\phi)$ combined with the unbiasedness of $\hat{p}$ results in $\hat{\phi}$ overestimating the true value of ϕ. This is shown graphically in Fig. 2.1.

A useful property of MLE is that they are asymptotically unbiased, that is, their expected value approaches the true value as the sample size n (here, n is the total number of replicates used at all levels) approaches infinity. In particular, the bias can be represented as a power series over $(1/n)$:

$$\mathbf{E}[\hat{\phi} - \phi] = \frac{b_1}{n} + \frac{b_2}{n^2} + \frac{b_3}{n^3} + \ldots, \tag{2.5}$$

where the coefficients b_n depend on the specific problem.

Different strategies have been used to correct the MLE bias at the first-order (b_1/n term in eq. (2.5)). The jackknife is a general bias-correction method that has been used in the context of LDA analysis (e.g. Does et al. 1988): it requires the computation of $\hat{\phi}$ for n subsamples of size $n-1$, each of them created by omitting a different replicate from the original sample.

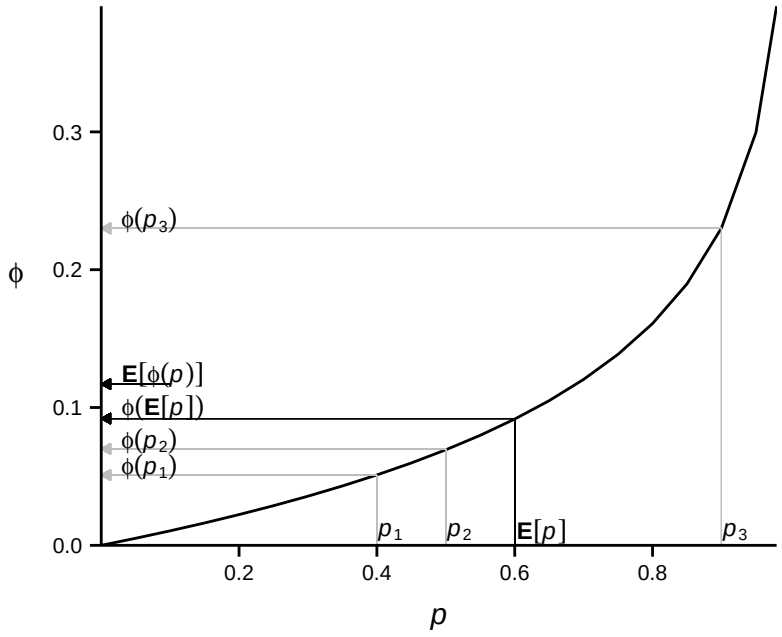

Figure 2.1: Illustration of Jensen's inequality in estimating ϕ from the observed proportion of successes, p, under a Poisson single-hit model ($\lambda = 10$). Due to the convex curvature of the function, the average estimate of ϕ, $\mathbf{E}[\phi(p)]$, is always greater than the true value of ϕ, corresponding to $\phi(\mathbf{E}[p])$.

The jackknife method, although robust, is computationally burdensome (due to the n repeats of the estimation procedure). In the context of dilution assays, Mehrabi and Matthews (1995) also found that the jackknife cannot reliably quantify the estimator's variance. Instead, they recommend using Firth's bias-corrected score function, U^* (Firth 1993):

$$U^*(\phi) = U(\phi) - \frac{1}{2\mathbf{E}\left[\frac{\partial U}{\partial \phi}\right]} \left\{ \mathbf{E}\left[\frac{\partial^2 U}{\partial \phi^2}\right] - 2\frac{\partial}{\partial \phi}\mathbf{E}\left[\frac{\partial U}{\partial \phi}\right] \right\}, \tag{2.6}$$

where U is the uncorrected score function (2.4) and all expected values are calculated over the set of experimental outcomes. The bias-corrected MLE of ϕ corresponds to the root of U^*.

2.1.3 Variance, confidence intervals and mean square error

In addition to any systematic bias, it is important to quantify the random variation of any estimator over the set of possible experimental outcomes. A general estimator for the variance of a MLE is the inverse of the Fisher information, I:

$$\hat{\sigma^2}_{\hat{\phi}} = \frac{1}{I(\phi)} = \frac{1}{-\mathbf{E}\left[\frac{\partial U}{\partial \phi}\right]}. \tag{2.7}$$

For the LDA score function (2.4), I is calculated as:

$$I(\phi) = \sum_{d=1}^{D} R_d \left[\frac{1}{p_d(\phi)\big(1 - p_d(\phi)\big)} \left(\frac{\mathrm{d}p_d}{\mathrm{d}\phi} \right)^2 \right]. \tag{2.8}$$

As defined in eq. (2.7), I corresponds to the "steepness" of the log-likelihood curve around its maximum at $\hat{\phi}$. It is sensible to expect that the MLE would be more precise if the likelihood function is very sensitive to small changes in ϕ. We also see that the Fisher information for a dilution assay (2.8) is additive: the term in square brackets can be interpreted as the amount of information provided by each of the R_d replicates at the d^{th} level.

When $\hat{\phi}$ approximately follows a normal distribution, the 100(1-α)% confidence interval can be estimated as $\hat{\phi} \pm z_{\alpha/2}\sqrt{\hat{\sigma^2}_{\hat{\phi}}}$, where $z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal distribution.

In her study on the design of LDA experiments with small sample sizes, Macken (1999) underlines two problems with these conventional estimates of the variance and confidence interval: (1) $\hat{\phi}$ is not distributed normally, and (2) the Fisher information, while providing a lower bound on the variance of an unbiased estimator, does not always indicate for which

design parameters the actual variance is minimized. Instead, she calculates the exact distribution of the MLE, $\hat{\phi}$, by determining the probability of each experimental outcome and the value of $\hat{\phi}$ that would be inferred from that outcome. Since at each level, k_d can take values from 0 to R_d, the total number of outcomes is:

$$N_{out} = \prod_{d=1}^{D}(R_d + 1). \tag{2.9}$$

If N_{out} is too large for the exact calculation to be feasible, Monte-Carlo methods can be used to generate a representative set of outcomes. Either the exact or approximate distribution of $\hat{\phi}$ can serve as a basis for calculating the bias, variance and confidence intervals for the MLE, without making any assumptions beyond the validity of the specified likelihood equation.

The overall performance of a biased estimator can also be characterized by its mean square error (MSE), which equals the sum of the variance and squared bias:

$$\mathbf{E}\left[\left(\hat{\phi} - \phi\right)^2\right] = \mathbf{E}\left[\left(\hat{\phi} - \mathbf{E}[\hat{\phi}]\right)^2\right] + \mathbf{E}[\hat{\phi} - \phi]^2.$$

2.1.4 Non-informative outcomes

When estimating the statistical properties of $\hat{\phi}$, outcomes with all negatives or positives (for all replicates at all levels) need to be considered separately. These outcomes will correspond to a $\hat{\phi}$ of 0 or 1 (respectively), regardless of the particular experimental design, which means they provide little information on the actual value of the parameter. Under the Poisson model, which does not explicitly treat ϕ as a proportion, all-positive outcomes are even more problematic, corresponding to $\hat{\phi} = \infty$! In practice, these experimental outcomes would not be used to infer ϕ; rather, they would lead to a repetition of the experiment at a more suitable range of dilution levels.

Two approaches have been used to avoid non-informative outcomes. With the primary concern of avoiding infinite estimates, Does et al. (1988) suggest arbitrarily replacing one of the positive tests with a negative. Macken's approach is to define a probability of non-informative assay (PNI), representing the frequency of all-negative or all-positive outcomes, then ignore these non-informative outcomes when calculating statistics from the $\hat{\phi}$ distribution. In this second approach, the PNI becomes a separate criterion that, just like the bias and variance, should be minimized in a good experimental design (Macken 1999).

2.2 Optimal design of dilution assays

While the results of a dilution assay can be analyzed in a straightforward way through maximum likelihood methods, finding optimal assay designs has proved to be a difficult problem, whether it is defined in terms of quality maximization (i.e. most precise estimate for a given experimental effort) or cost minimization (i.e. achieving a given precision level with the least effort). Optimum design theory has been concerned with defining suitable measures of efficiency - these are often functions of the Fisher information, or information matrix in the case of multiple parameters - and finding equivalences between them (Atkinson and Donev 1992).

Models where the observed response depends nonlinearly on the parameters pose a particular problem, as the optimum design depends on the (a priori unknown) values of those parameters. Dilution assays clearly fit in this category: for a given value of ϕ, there is a single dilution level - corresponding to roughly 1.6 responding cells per reaction - which maximizes the Fisher information. However, this single-level assay will become non-informative for much smaller or larger values of ϕ (Fazekas de St. Groth 1982).

As an empirical solution to this problem, Strijbosch et al. (1987) propose the use of a geometric progression of dilution levels: the concentration of cells at level d, is given by $c_0 a^{-d}$, where c_0 is the initial concentration and a is a dilution factor chosen based on the expected range of ϕ. The main advantage of this method is that its efficiency is stable over the whole range of ϕ considered, instead of peaking at a single value of the parameter. As an alternative, Abdelbasit and Plackett (1983) describe a sequential (multi-step) design: the dilution level at step 1 is chosen to optimize the information based on an initial guess of ϕ, then the estimate from step 1 becomes the initial guess for step 2, and so on. As the authors note, the efficiency of this method strongly depends on how close the initial guess is to the true value.

Bayesian approaches (Ridout 1995; Mehrabi and Matthews 1998) provide a general framework to incorporate prior information or guesses about the parameter values. This knowledge is expressed in the prior distribution of ϕ, so that the performance of a specific assay design is defined by the expected value of the efficiency criterion over that prior distribution. Maximizing the logarithm of the Fisher information or minimizing its inverse (which, as seen above, estimates the variance) are both commonly used criteria. Maximizing the Fisher information itself is not desirable, as it always produces assays with single dilution levels.[2] As an alternative to average performance measures, minimax (best performance in the worst case) or centile optimization criteria (best performance in X% of cases) can also be used within a Bayesian framework (Matthews 1998).

[2]For a given prior distribution of ϕ, there will be a dilution level which maximizes the expected information from a single reaction. Since Fisher information is additive, the information from the whole assay will be maximized when all reactions are replicates of that same dilution level.

2.3 Previous work on PCR-based dilution assays

As the polymerase chain reaction (PCR) technique of DNA amplification was popularized in the 1990s, a few authors discussed the potential of dilution assays as a way to quantify DNA copy numbers. Compared with other quantification techniques that rely on assumptions about amplification efficiencies of different sequences, methods based on LDA only require the binary (amplification or no amplification) endpoint result (Sykes et al. 1992). PCR-based assays, however, do not generally satisfy the single-hit criterion; that is, the possibility of false negatives or false positives cannot be ignored. Different approaches have been used to address this challenge.

In developing a PCR-based LDA to quantify the number of leukemic DNA copies from blood cell samples, Sykes et al. (1992) looked at two sources of amplification failure at low target copy number. To account for the possible degradation of DNA samples, they added a second set of primers for an endogenous gene (which should be present in all cells). They also observed that detection of their leukemic marker was significantly reduced when large amounts of non-leukemic DNA was added to the mix, a result they attribute to competition from homologous templates for primer binding. By performing two rounds of PCR (with nested primers), they estimate that they could detect as few as two leukemic templates in a background of over 10^5 non-leukemic genomes. Even with the added round of PCR, their data shows that adding non-specific DNA reduces the number of positive results.

The PCR-based LDA model of Rodrigo et al. (1997) incorporates false negatives and false positive rates, but treats them as constants. A more realistic model was developed by Hughes and Totten (2003), who use the results of a LDA performed at known DNA concentrations to estimate the sensitivity of the PCR reaction. Their method only assume that rate of success increases monotonically with the number of DNA templates. They do not consider the possibility of PCR inhibition at high template concentration, or the effect of non-target DNA molecules.

2.4 Limited dilution assay model for pollen PCR

In this section, I will apply the basic LDA principles to the genetic analysis of environmental pollen samples. The quantity of interest is the proportion of pollen grains containing a given DNA sequence, and the experimental test is a PCR amplification performed on varying numbers of pollen grains, using primers specific to this sequence.

The model presented here is based on the following assumptions:

1. At dilution level d, replicates are sampled in one of two ways:
    - exactly N_d pollen grains are picked randomly from the original sample ("exact N

sampling"); or

- a constant volume is taken from a well-mixed pollen suspension, obtained by dilution of the original sample; N_d is assumed to follow a Poisson distribution with an average of λ_d ("Poisson sampling").

2. The probability of amplification in a given replicate depends on the number of pollen grains present (N_d) as well as the number of grains with the marker of interest (m); this probability is written as $\Pr\left(+ \mid m, N_d\right)$.

Note that $\Pr\left(+ \mid m, N_d\right)$ represents the probability of amplification *conditional* on N_d grains being present, m of them having the target marker. It encompasses both the possibility of false positives (if $\Pr(+) > 0$ when $m = 0$) as well as false negatives (if $\Pr(+) < 1$ when $m \geq 1$).

The probability of a positive amplification at the d^{th} level, p_d, is equal to the average (expected value) of $\Pr\left(+ \mid m, N_d\right)$ over all possible values of m and N_d.

In the Poisson sampling case, N_d follows a Poisson distribution, and m follows a binomial distribution with N_d attempts and probability of success ϕ. Therefore, p_d is calculated as:

$$p_d(\phi) = \sum_{N_d=0}^{N_{max}} \frac{{\lambda_d}^{N_d} e^{-\lambda_d}}{N_d!} \sum_{m=0}^{N_d} \binom{N_d}{m} \phi^m \left(1-\phi\right)^{N_d-m} \Pr\left(+ \mid m, N_d\right). \quad \text{[Poisson]} \tag{2.10}$$

In theory, the Poisson distribution extends to arbitrarily large values of N_d (with tiny probabilities); it is truncated to some value N_{max} for computational purposes.

For the exact N sampling case, N_d is fixed, while m still follows a binomial distribution:

$$p_d(\phi) = \sum_{m=0}^{N_d} \binom{N_d}{m} \phi^m \left(1-\phi\right)^{N_d-m} \Pr\left(+ \mid m, N_d\right). \quad \text{[exact } N\text{]} \tag{2.11}$$

2.4.1 Parametric model for failed assays

In the model above, the possibility of failed assays (i.e. false positives or false negatives) is accounted for by the function $\Pr\left(+ \mid m, N_d\right)$, corresponding to the probability of amplification in a reaction where m out of N_d pollen grains have the marker of interest. Since this probability depends on both the experimental protocol used and the sample being tested (species, age, etc.), it should be determined experimentally by performing the assay on representative samples of known pollen composition. However, for the purpose of testing the model's sensitivity to failed assays, it would be useful to represent $\Pr\left(+ \mid m, N_d\right)$ in a simple parametric form. Here, I propose a parametric failed assay model that, despite its simplicity, is broad enough to encompass many of the suspected sources of false positives or negatives.

I assume the false negative rate is due to three independent components:

- DNA from each of the m pollen grains with the marker could fail to amplify (due to degradation, for example), with a probability f_{deg};
- some component of the pollen could inhibit PCR, causing the reaction to fail with probability $f_{inh}(N_d)$;
- the reaction could fail for reasons having nothing to do with the number of pollen grains, with probability f_{reac}.

Common sources of false positives in PCR include contamination (by amplicons from previous assays) and non-specific amplification (due to primer design or experimental conditions). Therefore, I will assume that the probability of false positives, f_{pos}, is independent of both m and N_d.

These combined assumptions lead to the following equation for $\Pr\big(+ \mid m, N_d\big)$:

$$\Pr\big(+ \mid m, N_d\big) = \begin{cases} f_{pos} & \text{if } m = 0, \\ \big(1 - f_{reac}\big)\big(1 - (f_{deg})^m\big)\big(1 - f_{inh}(N_d)\big) & \text{if } m \geq 1. \end{cases} \tag{2.12}$$

I still have to specify the inhibition function $f_{inh}(N_d)$. This function should be strictly increasing, approaching 0 at low pollen numbers and 1 for high pollen numbers. A logistic equation of the form:

$$f_{inh}(N_d) = \frac{1}{1 + \exp\big(-r_{inh}(N_d - N_{inh50})\big)} \tag{2.13}$$

meets these criteria and constitutes a common choice to model dose-response relationships (e.g. Haanstra et al. 1985). The parameter N_{inh50} corresponds to the value of N_d for which 50% of reactions would be inhibited, while r_{inh} determines the steepness of the curve around that point.

Figure 2.2 illustrates how the probability of success of a single reaction is affected by each of the false negative parameters, as well as their combined effect.

2.4.2 Numerical implementation

I have written a Fortran 90 program to calculate the distribution of $\hat{\phi}$ for a specific experimental design under this model. It takes as input a value of ϕ (representing the true concentration in the hypothetical sample to be tested) as well as the following design parameters:

- whether the Poisson or exact N sampling method is used;

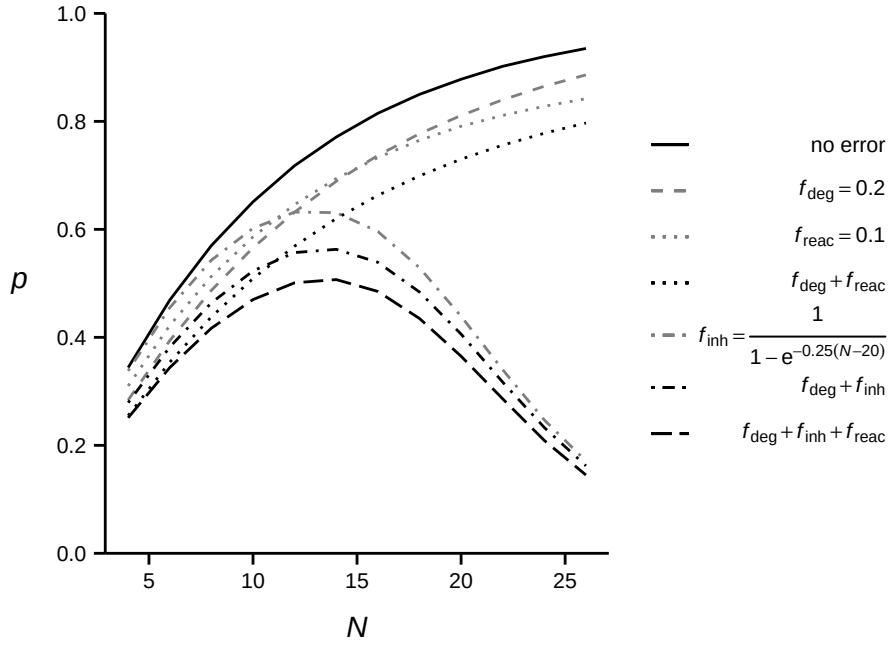

Figure 2.2: Effect of the three false negative sources (f_{deg}, f_{inh} and f_{reac}) on the probability p of a successful reaction, as a function of the number of pollen grains N (exact N sampling with $\phi = 0.1$).

- the number of dilution levels D;
- the number of replicates R_d for each level; and
- for each level, either the number of pollen grains tested by replicate N_d (exact N sampling) or the average number of grains λ_d (Poisson sampling).

The function $\Pr\big(+ \mid m, N_d\big)$ also needs to be specified in the program. To calculate p_d in the Poisson sampling case (eq. 2.10), I use $N_{max} = 2\lambda_d + 3$, which includes >99.9% of possible outcomes.

The program first calculates the probability of each outcome $\{k_d\}$:

$$\Pr\big(\{k_d\}\big) = \prod_{d=1}^{D} \binom{R_d}{k_d} p_d(\phi)^{k_d} \big(1 - p_d(\phi)\big)^{R_d - k_d}.$$

Then, for any outcome exceeding a minimum probability, the maximum likelihood estimate $\hat{\phi}$ is calculated. The minimum probability is used to avoid wasting computation time on very improbable outcomes which do not significantly affect the statistics of $\hat{\phi}$. The threshold is set at α/N_{out}, with N_{out} being the total number of possible outcomes (eq. 2.9). This ensures that the outcomes retained for the calculation of $\hat{\phi}$ represent at least 100(1-α)% of the total probability (as a conservative estimate).

The MLE is determined as the root of the corrected score function (2.6), found using Brent's algorithm (Press et al. 1996) under the assumption there is a root in the interval $\phi \in (0, 1)$. In the absence of a root, the likelihood is maximized at either 0 or 1 and the outcome is classified as non-informative (see discussion of non-informative outcomes above). Note that this definition of non-informative outcome is not equivalent to an outcome with only negative or positive results. Indeed, in the presence of a (known) false positive rate, an outcome with very few positives could be non-informative; conversely, as the results below will show, the corrected score function may have a root in $(0, 1)$ even when all replicates are positive.

The program outputs the distribution of $\hat{\phi}$ as well as summary statistics including the bias, variance and probability of non-informative assay (PNI). Confidence intervals at any desired level can be determined from the distribution itself.

Derivation of the equations used in the program (such as the corrected score function) as well as the complete source code are included as Appendix A.

I first apply this model to calculate the distribution of estimates when there are no false negatives or positives, as in the Poisson single-hit model: I call this situation the 'ideal response case'. Following this treatment of the ideal case, I will consider the problem of estimating the error parameters, and evaluate how these rates of false positives or false negatives affect the accuracy of ϕ estimates.

2.5 Model calculations in the ideal response case

2.5.1 Assays with a single dilution level

For a single dilution level, there are only two scalar design parameters: the number of pollen grains per reaction (N, or its average value λ in the Poisson sampling case) and the number of replicates (R).

Figure 2.3 shows an example of the probability mass function (pmf) predicted for $\hat{\phi}$ in a single level assay, with and without the correction applied to the score function. Unless specified otherwise, for all results presented here I set the parameter $\alpha = 10^{-3}$ (meaning that the pmf is calculated for outcomes totalling at least 99.9% of the total probability). While the results shown are for Poisson sampling, the pmf retains the same shape in the exact N sampling case.

In this example, it is clear that the distribution of parameter estimates is positively skewed and not normal. Using a quantile-quantile plot (not shown), I have verified that the distributions in Fig. 2.3 are very close to lognormal, which is consistent with the results of Macken (1999). However, it should not be assumed that this lognormal shape generalizes to

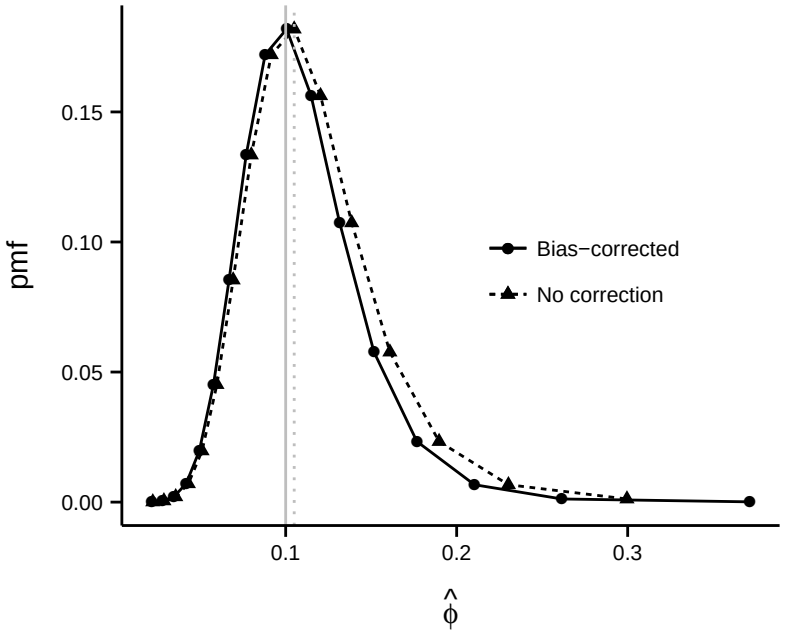

Figure 2.3: Probability mass function of the maximum likelihood estimate $\hat{\phi}$ in the ideal response case with $\phi = 0.1$, a single dilution level at $\lambda = 10$ (Poisson sampling), and $R = 20$. Vertical lines represent the expected value of $\hat{\phi}$ for each pmf.

more complex assay designs or non-ideal responses.

Another specific feature of this example is that the bias-corrected distribution includes one more outcome at the high end ($\hat{\phi} = 0.37$), which is the estimated frequency when all 20 replicates are positive; in the uncorrected case, the estimate of $\hat{\phi} = 1$ is rejected as non-informative. This frequency could be interpreted as a strict upper bound on the true value of ϕ that can be estimated from this specific design (number of replicates and number of pollen grains per replicate); that is, the assay would be useless on any sample with $\phi > 0.37$. Simulation results also show that ϕ becomes more and more underestimated at it approaches this upper bound.

Figures 2.4 and 2.5 show how, for a given frequency ϕ and number of replicates R, the mean number of pollen grains per tube $\overline{N}$ (equal to λ for Poisson sampling, and N for exact number sampling) affects the relative bias, coefficient of variation (CV) and probability of non-informative outcomes (PNI) of the estimator. The CV is the ratio of the standard deviation to the expected value:

$$CV = \frac{\sqrt{\sigma^2_{\hat{\phi}}}}{\mathbf{E}\left[\hat{\phi}\right]},$$

which is useful for comparing the relative uncertainty of estimates.

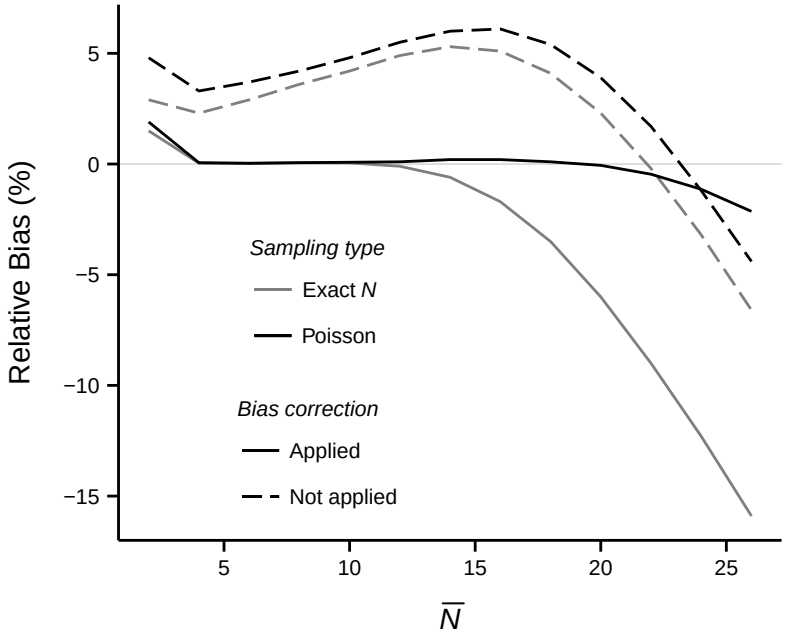

Figure 2.4: Bias of the corrected and uncorrected maximum likelihood estimator $\hat{\phi}$ for a single dilution level, as a function of the average number of pollen grains per reaction, $\overline{N}$ (for $\phi = 0.1$ and $R = 20$).

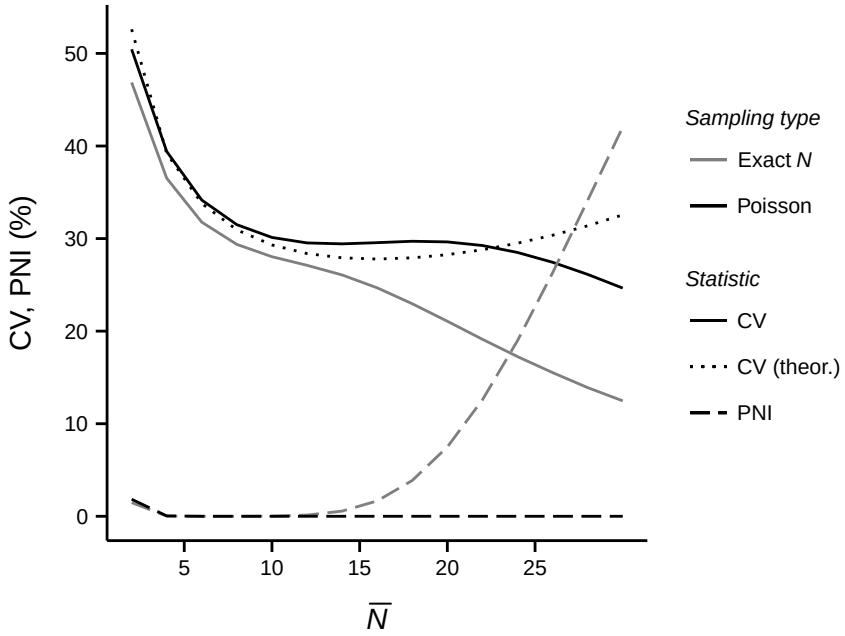

Figure 2.5: Coefficient of variation (CV) and probability of non-informative outcome (PNI) of $\hat{\phi}$ for a single dilution level, as a function of $\overline{N}$ (for $\phi = 0.1$ and $R = 20$). The dotted line shows the theoretical value (based on Fisher's I) of the CV for Poisson sampling.

In Figure 2.4, I have included the results produced with an uncorrected score function (only the corrected version will be used from this point on). For mid-values of $\overline{N}$, the uncorrected version overestimates ϕ by about 5%, while the corrected estimates fall within 0.1% of the true value. The positive bias at low values of $\overline{N}$, even in the corrected case, is due to the rejection of all negative outcomes (more common when $\overline{N}$ is low) as non-informative, rather than assigning them $\hat{\phi} = 0$ and including them in the average. Similarly, the negative bias at high $\overline{N}$ is due to all positive outcomes either being rejected, or resulting in an underestimate under the correction procedure (see above). One interesting feature of these results is that a reliable estimate can be produced at higher $\overline{N}$ using Poisson sampling; since the number of pollen grains vary from one replicate to another, there will be enough replicates with a lower N to reduce the chance of an all positive outcome.

I have displayed both the CV and PNI in the same graph (Fig. 2.5), to show the CV can be misleadingly low when many outcomes are rejected as non-informative. I have also added a theoretical curve of the CV, as calculated from Fisher's information in the Poisson sampling case (see eq. 2.7 and the accompanying discussion).

Taking into account the bias, CV and PNI, I can determine an optimal range for $\overline{N}$ for a given marker frequency (here $\phi = 0.1$). For exact N sampling, the range providing the best estimate seems to be betwen 5 and 15, corresponding to an average of 0.5 to 1.5 "positive" pollen grains per reaction. For Poisson sampling, this range extends to at least 20 (or 2 positive pollen grains per reaction). Note that the PNI at high $\overline{N}$ is low for Poisson sampling due to the correction procedure, but there is still an undesirable bias. The PNI is still high for exact N sampling.[3]

These results also show the usefulness of calculating statistics based on the complete distribution of $\hat{\phi}$ when R is not too large. As noted before, the estimation of the variance using Fisher's information leads to an optimal target of around 1.6 positive cells per replicate for the Poisson single-hit model (corresponding to $\overline{N} = 16$ in Fig. 2.5). However, for the relatively small-scale assays discussed here, it might be better to speak of an optimal range where the CV remains approximately constant. The Fisher optimum is not in the middle of that range, but rather at the high end.

Since Poisson sampling from a diluted pollen suspension adds uncertainty about the number of pollen grains per replicate, compared with picking an exact number of grains each time, it is not surprising that the coefficient of variation for Poisson sampling is higher than for exact N sampling in Fig. 2.5. This reduction of the CV by a few percentage points needs to be weighted against the additional work involved when manually picking out pollen grains from the sample, unless the experimenter has access to automated cell-sorting equipment.

As discussed before, dilution assays generate nonlinear likelihood models, so that the performance of a specific assay design depends on the value of the parameter one seeks to

[3]This is due to the difference in the shape of the function $p(\phi)$ in the Poisson and exact N cases. When the probability of success becomes large, its derivative is too close to zero in the exact N case for the correction procedure to give good results, so the outcome still has to be rejected. See Appendix A for details.

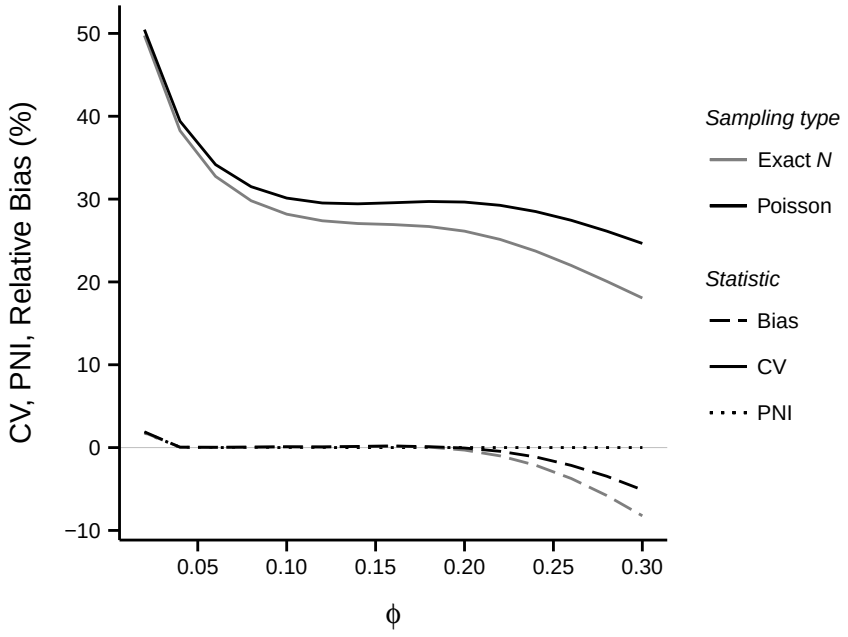

Figure 2.6: Relative bias, CV and PNI of $\hat{\phi}$ for a single dilution level, as a function of ϕ (for $\overline{N} = 10$ and $R = 20$).

estimate. Figure 2.6 shows how the relative bias, CV and PNI of the estimator are affected by the true frequency ϕ, for a given assay design ($\overline{N} = 10$ and $R = 20$). The results in the Poisson sampling case are the same as those obtained above when varying λ: assuming an ideal response (the Poisson single-hit model), the probability of success of a reaction only depends on the product $\phi\lambda$, which is also the expected number of positive cells per replicate (see eq. 2.1). However, the estimate obtained from exact N sampling "tolerates" high values of ϕ better than high values of N: above $\phi = 0.2$, the PNI is still negligible, but a negative bias appears, just as in the Poisson case.

These results show an approximately fourfold change in ϕ (from 0.05 to 0.2) where a given assay design can produce unbiased estimates, and where the variation stays within a relatively narrow range (CV between 26% and 33% for exact N sampling, or between 29% and 35% in the Poisson sampling case). This suggests a multiple stage design where ϕ is first confined to a sufficiently narrow range, followed by a single level assay with N or λ chosen by the method presented here. I will discuss this type of design further in the next section.

Based on the theoretical relationship between Fisher's information and the estimator variance (eqs. 2.7 and 2.8), I expect the variance to be proportional to $1/R$ and the standard deviation to vary as $1/\sqrt{R}$. Figure 2.7 shows how the number of replicates affects the relative bias, CV and PNI (for fixed $\phi = 0.1$ and $\overline{N} = 10$). Except for small values of R, where non-informative outcomes are common and $\hat{\phi}$ is biased (especially in the exact N sampling case), the coefficient of variation does vary as the inverse square root of R (regression not

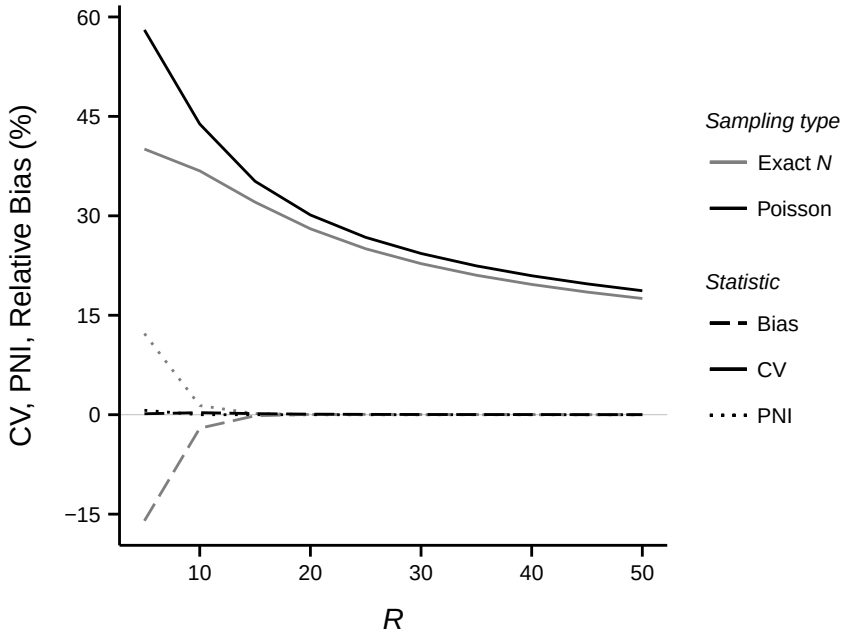

Figure 2.7: Relative bias, CV and PNI of $\hat{\phi}$ for a single dilution level, as a function of R (for $\phi = 0.1$ and $\overline{N} = 10$).

shown).

The previous results assume that the number of replicates R and the number of pollen grains per replicate $\overline{N}$ are independent design parameters. While this will often be true, there are cases when the total pollen grain number will be a limiting factor (e.g. when sampling from a very small insect pollinator). In this case, where the product $R\overline{N}$ is a design constant, is it better to keep $\overline{N}$ within its optimal range, or to decrease it in favor of more replicates? Figure 2.8 shows that under this constraint, the coefficient of variation monotonically decreases with greater R. However, the additional material and labor cost of performing more replicates may not be worth a relatively modest increase in precision.

2.5.2 Multiple level assays

Based on the discussion of optimal assay designs in section 2.2, it does not seem possible to define a single best experimental design for a multiple level dilution assay. There are multiple criteria proposed to qualify a better design. (For example, should it produce the most precise estimate on average, or in the worst case scenario?) Even if one measure of quality is ageed upon, it may not be feasible to compute it for all possible values of the design parameters.

Therefore, I will narrow my search to a particular type of two-step sequential designs,

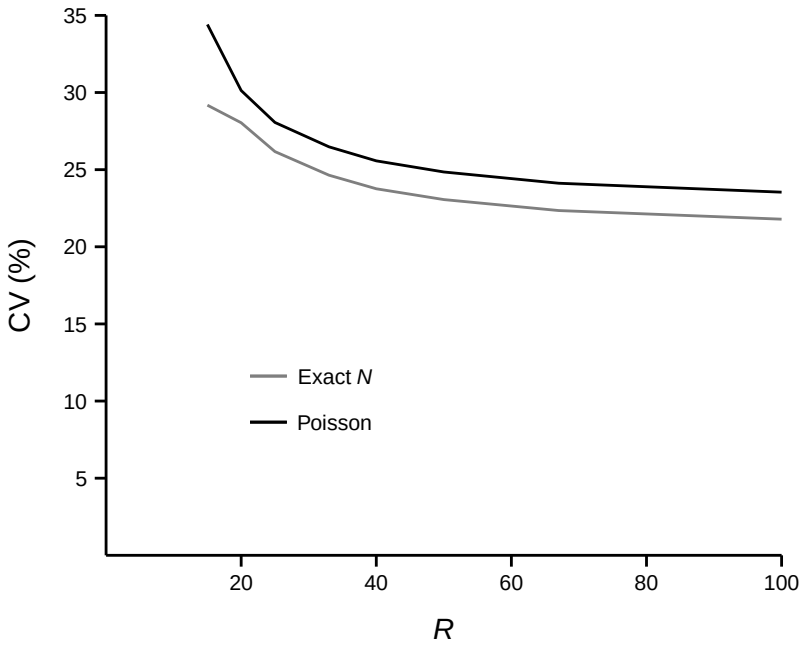

Figure 2.8: Coefficient of variation of $\hat{\phi}$ for a single dilution level as a function of R, with the product $R\overline{N}$ fixed at 200 ($\phi = 0.1$).

where step 1 is a geometric dilution series and step 2 has a single dilution level, chosen based on the estimate of ϕ obtained from step 1. This choice aims to combine the advantages of the geometric progression (which is effective for a large range of values of ϕ) and the additional precision provided by the single-dilution step. It differs from the type of multi-step designs proposed by Abdelbasit and Plackett (1983), where all steps including the first are performed at a single dilution level. Applying a two-step design in practice requires the pollen sample to be mixed and split into two parts of (nearly) identical composition, which should be feasible as long as the number of grains in the initial sample is large.

Concretely, this two-step design is specified by four parameters:

- D, the number of dilution levels for step 1;
- R_m, the number of replicates for the first, multiple-dilution step (same for all D levels);
- R_s, the number of replicates for the second, single-dilution step; and
- ϕ_{min}, the minimum value of ϕ that the assay is meant to detect (ϕ_{max} is set at 1).

For step 1, the mean number of pollen grains at level d is a^{d-1}, rounded to the nearest integer. The value of a is chosen to form a geometric progression from 1 to $1/\phi_{min}$:

$$a = \sqrt[D-1]{\frac{1}{\phi_{min}}}.$$

The mean number of pollen grains for step 2 is taken as $1/\hat{\phi}$, rounded to the nearest integer, which corresponds to one positive pollen grain per reaction. This is based on the optimal range of around 0.5 to 2 positive grains per reaction found in the previous section. Since the distribution of $\hat{\phi}$ is approximately lognormal, it makes sense to choose the midpoint of (0.5, 2) on a log scale, which is 1. I also verified that using a target above 1 positive grain per reaction (up to the Fisher information optimum of 1.6) did not improve the accuracy of the estimate.

In the following simulations, I set $\phi_{min} = 0.01$ and the total number of reactions $N_{tot} = 50$. For each set of parameters (D, R_m and R_s), I use the algorithm above to calculate the bias and variance of $\hat{\phi}$. Figure 2.9 shows the relative root mean squared error (the square root of MSE as a fraction of ϕ) for different values of D and R_m.[4] Exact N sampling was used to speed up calculations, although I have found the same pattern in a smaller-scale simulation with Poisson sampling.

To put these relative root MSE values into perspective, I note that their theoretical minimum (if all N_{tot} reactions were conducted at the optimal dilution level) is 0.14 for $\phi = 0.4$ and 0.17 for $\phi = 0.02$.

These results, which were replicated at more values of ϕ than the two shown, do not display a very sharp optimum, as many possible choices of the parameters yield comparable estimation accuracy. There is, however, a trend emerging from these results that suggests a general strategy for step 1 design: choose more dilution levels but less replicates at each level. In this particular case, six dilution levels (approx. 2.5-fold dilutions from $N = 1$ to 100) seem to suffice. This is useful knowledge, since it allows the experimenter to maximize the experimental effort at step 2, where additional reactions have the biggest impact in reducing the uncertainty about ϕ.

As an illustration of the gain in precision from the two-step assay, Fig. 2.10 shows the probability density of $\hat{\phi}$ for a true value $\phi = 0.1$, using one of the best designs found above: 6 dilution levels with 2 replicates per level at step 1, and the remaining 38 reactions at step 2. Based on a quantile-quantile plot (not shown), I found that the probability density of the estimate after step 2 closely followed a lognormal distribution, just as in the single-level case (Fig. 2.3).

None of the results from this section take into account the experimental errors associated with pollen PCR. In the following section, I will look in the impact of those errors on the design and analysis of the PCR-dilution assays.

[4]Since N_{tot} is fixed, R_s can be found from $DR_m + R_s = N_{tot}$.

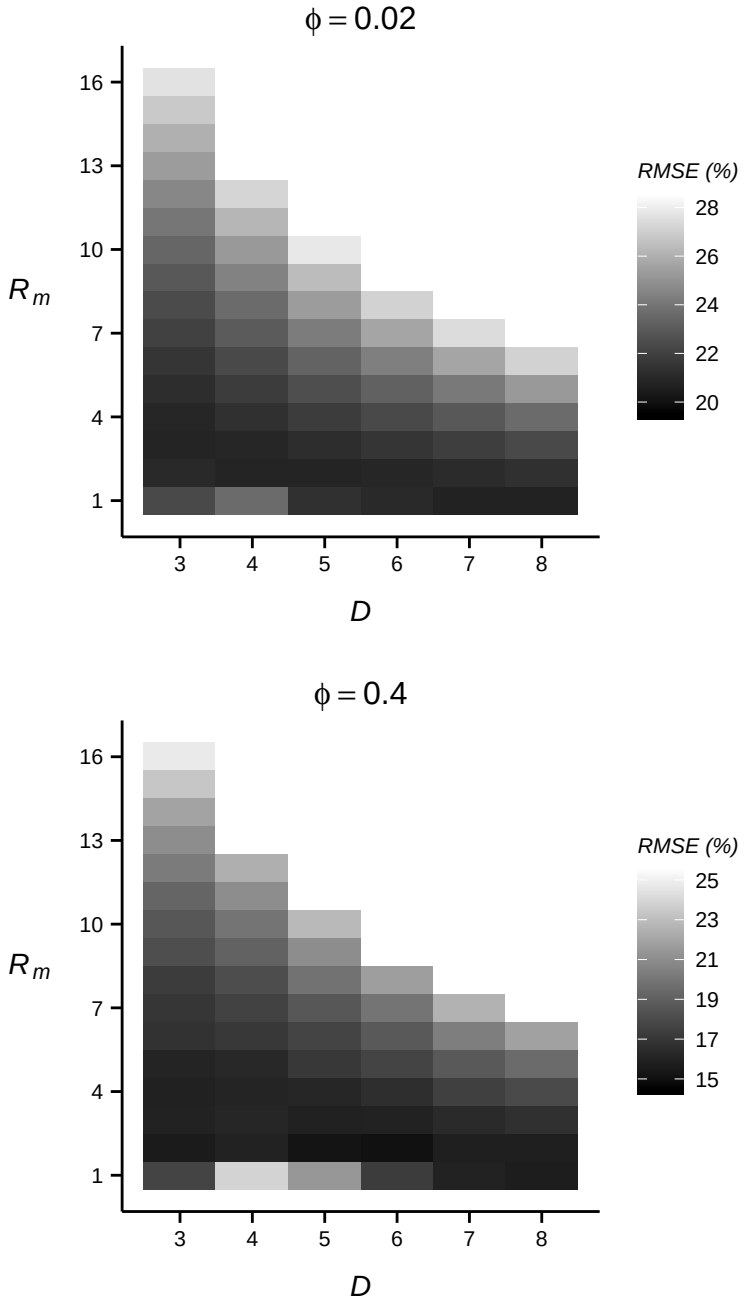

Figure 2.9: Relative root mean squared error (RMSE) of $\hat{\phi}$ as a function of the number of dilution levels (D) and the number of replicates per level (R_m) at the first step of a two-step assay, for two values of ϕ (50 total reactions, $\phi_{min} = 0.01$).

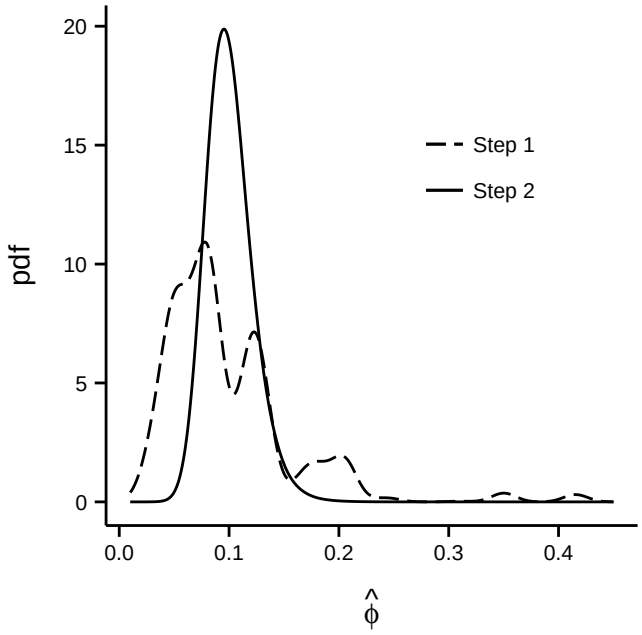

Figure 2.10: Probability density function (pdf) of $\hat{\phi}$ after step 1 and step 2 of a two-step assay, with $D = 6$, $R_m = 2$, $R_s = 38$ and $\phi_{min} = 0.01$; the true value of ϕ is 0.1.

2.6 Model calculations in the non-ideal case

For the purpose of illustrating how failed assays can be accounted for in my model, as well as how they impact its capability to estimate ϕ, I have introduced a parametric expression for the probability of success of a PCR reaction, given the number of pollen grains present (N_d) and the number of grains bearing the DNA marker of interest (m). That model assumes a constant rate of false positives (f_{pos}) and three independent sources of false negatives, corresponding to the degradation of the DNA template (f_{deg}), the inhibition of PCR by non-target DNA or other components of the pollen (f_{inh}), and a reaction failure rate independent of pollen grain number (f_{reac}):

$$\Pr\big(+ \mid m, N_d\big) = \begin{cases} f_{pos} & \text{if } m = 0, \\ \big(1 - f_{reac}\big)\big(1 - (f_{deg})^m\big)\big(1 - f_{inh}(N_d)\big) & \text{if } m \geq 1, \end{cases} \qquad \text{(2.12 revisited)}$$

with

$$f_{inh}(N_d) = \frac{1}{1 + \exp\big(-r_{inh}(N_d - N_{inh50})\big)}. \qquad \text{(2.13 revisited)}$$

Since the probability of success will depend on sample quality as well as the specific experimental protocol used, the error model needs to be calibrated by performing PCR reactions on samples of known composition (m and N_d), using pollen that is as similar as possible to the samples of interest. As an example, for a study measuring the proportion of

maize pollen in aerial samples taken around a flowering field, a calibration sample of 100% maize pollen could be collected from stamens in that field on the same day and stored in the same condition as the aerial samples.

Even if their rate of occurrence were exactly known, false positives and false negatives could reduce the precision and accuracy of the estimator $\hat{\phi}$, compared with the ideal case previously described. I will look at this effect first, before considering the additional uncertainty involved in estimating the error parameters themselves.

2.6.1 Effect of known error parameters

The inference power of dilution assays is based on a known relationship (eqs. 2.1, 2.3) between the probability of observing the response (here, PCR amplification) and the concentration of responsive cells in the sample. Different types of PCR errors will affect the precision of ϕ estimates by weakening this concentration-response link.

When a constant rate of false positives (f_{pos}) or false negatives (f_{reac}) occurs independently of the number of pollen grains present, the probability of amplification is confined to the range (f_{pos}, 1 - f_{reac}) instead of (0,1). A smaller difference in outcome between low and high pollen numbers means a less sensitive assay, therefore I expect a decrease in the precision of $\hat{\phi}$.

When each grain bearing the marker has the same probability (f_{deg}) of not having its DNA amplified, this reduces the effective marker frequency by a factor of (1 - f_{deg}). All other sources of error being equal, a sample composed of 10% marked pollen that amplifies 80% of the time ($f_{deg} = 0.2$) would produce the same distribution of outcomes as a sample with 8% marked pollen where every grain's DNA can be amplified. If the value of f_{deg} were exactly known, it should have little effect on the accuracy of $\hat{\phi}$; however, ignoring or misevaluating this source of error could be a source of systematic bias for inferring ϕ.

I calculated the impact of these PCR errors on the accuracy (measured as the relative root MSE) of a two-step dilution assay of the type presented in section 2.5.2 above, using reasonably high values of the error parameters: 5% of false positives, a constant PCR failure rate of up to 20%, as well as a per grain failure rate up to 50%. Different error scenarios are compared to the ideal case in Table 2.1, for a particular set of assay design parameters ($D = 6$, $R_s = 2$ and $R_m = 38$) that ranked consistently among the best designs with 50 total reactions. As expected, the estimator's accuracy is more sensitive to fixed errors (f_{pos} and f_{reac}). The effect of f_{deg} only becomes apparent for its largest value (0.5) and for the smallest marker frequency considered (0.02).[5] I also note that the exact N and Poisson sampling schemes achieve similar accuracies, except when the marker frequency is high ($\phi = 0.4$) so that the optimal number of pollen grains per reaction approaches 1.

[5] In that case, only 1 out of every 100 pollen grains would test positive, which corresponds to the smallest frequency (ϕ_{min}) for which the assay was designed.

	Relative root mean squared error of $\hat{\phi}$ (%)					
	$\phi = 0.02$		$\phi = 0.1$		$\phi = 0.4$	
	Exact N	Poisson	Exact N	Poisson	Exact N	Poisson
No error	21	21	20	22	15	22
$f_{deg} = 0.2$	22	22	21	23	17	21
$f_{reac} = 0.1$	24	25	26	27	21	26
$f_{reac} = 0.1$, $f_{deg} = 0.2$	25	25	26	27	22	25
$f_{reac} = 0.2$	28	28	31	32	27	32
$f_{reac} = 0.2$, $f_{deg} = 0.5$	33	33	32	33	27	29
$f_{pos} = 0.05$	22	23	21	24	16	23

Table 2.1: Effect of f_{deg}, f_{reac} and f_{pos} on the relative root mean squared error of $\hat{\phi}$, calculated for a two-step assay with $D = 6$, $R_m = 2$, $R_s = 38$ and $\phi_{min} = 0.01$.

	Relative root mean squared error of $\hat{\phi}$ (%)			
	$N_{inh50} = 20$		$N_{inh50} = 40$	
	Exact N	Poisson	Exact N	Poisson
$r_{inh} = 0.1$	42	43	29	30
$r_{inh} = 0.25$	36	37	26	26
$r_{inh} = 1$	33	34	25	26
$f_{reac} = 0.1, r_{inh} = 0.25$	38	40	28	29
$f_{reac} = 0.2, r_{inh} = 0.25$	41	44	31	32

Table 2.2: Effect of PCR inhibition on the relative root mean squared error of $\hat{\phi}$, calculated for a two-step assay with $\phi = 0.02$. Assay parameters are set as follows: $D = 6$, $R_m = 2$, $R_s = 38$, and $\phi_{min} = \frac{1}{0.6 N_{inh50}}$.

If PCR amplificiation is inhibited at high pollen numbers, as modelled by eq. (2.13), the probability of success of a reaction decreases past some critical value of $\overline{N}$ (see Fig. 2.2). This turning point depends most strongly on N_{inh50}, but also on r_{inh} and ϕ. Calculations of single-level assay outcomes with PCR inhibition produce a sharp drop in the precision and accuracy of $\hat{\phi}$ if $\overline{N}$ exceeds the same threshold (results not shown here).

In the design of a two-step dilution assay, this turning points puts an upper limit on N_D, the number of pollen grains to use at the highest concentration level. As long as the marker frequency is greater than $\phi_{min} = 1/N_D$, it is still possible to set $\overline{N}$ at step 2 to $1/\phi$ (rounded) and estimate ϕ with no loss of accuracy.

However, this accuracy decreases the further ϕ falls below ϕ_{min}, as shown in Table 2.2. In these calculations, I conservatively set the limit to 60% of N_{inh50}, so that it lies under the turning point for a broad range of values of r_{inh} and ϕ. With a N_{inh} of 20 and r_{inh} between 0.1 and 1, the root MSE for a sample with $\phi = 0.02$ (approximately one fourth of ϕ_{min}) can be 1.5 to 2 times as large as in the no error case. While a RMSE equal to 42% of the parameter value seems rather high, I note that 275 single-pollen reactions would be required

to achieve a comparable precision to the 50 reactions performed in the dilution assay.[6] Even when the number of grains per reaction is limited by PCR inhibition effects, the dilution assay remains an efficient strategy to estimate ϕ.

In the last two cases considered in Table 2.2, a fixed PCR failure rate f_{reac} is added to the model. Interestingly, the increase of RMSE due to f_{reac} is comparable to that predicted when it was the only source of error (in Table 2.1). This suggests an additive effect of these two independent types of reaction failures.

2.6.2 Estimating assay failure parameters

The discussion below focuses on the estimation of the false negative parameters, starting with f_{reac} and f_{deg} – which determine the probability of success at low pollen numbers – then adding the inhibition function f_{inh}. There are a few reasons why I ignore the false positive rate f_{pos} at this point. Since it is assumed to be constant, its estimation would only entail the replication of PCR on samples with non-target pollen. Furthermore, taking precautions against contamination and choosing specific enough PCR primers should make false positives much less frequent than false negatives.

According to my error model, the probability of amplification for a reaction with N pollen grains, all known to possess the marker of interest (i.e. $\phi = 1$), is equal to $\big(1 - f_{reac}\big)\big(1 - (f_{deg})^N\big)$. The values of N used here will be small enough (less than 10) that it seems reasonable to neglect inhibition effects.

I used the following method to simulate the estimation of these parameters and the effect of these estimates on the accuracy of the PCR dilution assay:

1. Simulate 100 outcomes of a calibration assay performed on known samples with $\phi = 1$, for some values of f_{reac} and f_{deg}. For the results presented here, the assay consists of 30, 10, 6 and 4 replicates for $N = 1$, 2, 4 and 8, respectively (50 reactions total).
2. For each of the 100 outcomes, obtain joint maximum likelihood estimates of f_{reac} and f_{deg}.
3. For each pair of (f_{reac}, f_{deg}), calculate the distribution of outcomes of a two-step dilution assay and the estimate $\hat{\phi}$ corresponding to each outcome.

I wrote a R script (see Appendix A) to perform the first two steps and generate 100 joint estimates of the error parameters, given the true values of f_{reac} and f_{deg}. At the third step, I use the same algorithm as in previous sections, except that $\hat{\phi}$ for each outcome is

[6] For $N = 275$ binomial trials with $p = 0.02$, the standard deviation of $\hat{p}$ is given by $\sqrt{\frac{p(1-p)}{N}} = 0.0084$, which is 42% of p.

(f_{reac}, f_{deg})		**Relative Bias** (mean (s.d.) in %) $\phi = 0.02$	$\phi = 0.1$	$\phi = 0.4$	**CV** (mean (s.d.) in %) $\phi = 0.02$	$\phi = 0.1$	$\phi = 0.4$
(0, 0.2)	known	2	1	0	22	21	17
	estimated	3 (9)	3 (9)	1 (9)	22 (0.4)	21 (0.3)	17 (0.1)
(0.1, 0)	known	0	-3	-5	24	26	22
	estimated	0 (10)	-4 (11)	-7 (10)	24 (0.4)	27 (2)	24 (4)
(0.1, 0.2)	known	2	-1	-3	25	26	22
	estimated	3 (12)	-2 (13)	-5 (13)	25 (1)	27 (1)	25 (5)
(0.2, 0)	known	-1	-6	-8	28	32	28
	estimated	0 (13)	-6 (15)	-10 (14)	28 (1)	34 (3)	31 (7)
(0.2, 0.5)	known	5	2	-2	31	31	27
	estimated	7 (23)	0 (21)	-7 (20)	31 (2)	31 (1)	30 (5)

Table 2.3: Relative bias and coefficient of variation (CV) of $\hat{\phi}$ predicted for a two-step assay (exact N sampling with $D = 6$, $R_m = 2$, $R_s = 38$ and $\phi_{min} = 0.01$), for known and estimated values of the error parameters f_{reac} and f_{deg}. Error parameter estimates are based on 100 simulations of the following assay: 30, 10, 6 and 4 replicates using 1, 2, 4 and 8 marked pollen grains per reaction, respectively.

determined using the estimated error parameters (whereas the distribution of outcomes is always calculated from the true values of the parameters).

The calibration assay proposed here places a majority of the experimental effort at the single grain level ($N = 1$), for two main reasons. On the one hand, the variance of binomial outcomes will usually be greater at smaller N, which motivates the use of more replicates at that level.[7] On the other hand, since the two-step dilution assay is designed to achieve an average concentration of 1 marked grain per reaction, knowing the probability of amplification in this case is particularly important.

I repeated this process for different combinations of ϕ, f_{deg} and f_{reac}, using the same two-step dilution assay design in all cases, to produce the results of Table 2.3. Since each of the 100 pairs of (f_{deg}, f_{reac}) estimates resulted in a different bias and CV of $\hat{\phi}$, I report the mean and standard deviation of those statistics in the table, and present their distribution as histograms in Fig. 2.11 (for the case where $f_{deg} = 0.2$ and $f_{reac} = 0.1$). For comparison purposes, I also include in Table 2.3 the calculated values of the bias and CV under the assumption that the error parameters are exactly known.

The frequent occurence of a large negative or positive bias in these simulation results could pose an obstacle to the use of this assay in practice. Its cause is not the biasedness of the calibration assay itself; on average, it produces accurate estimates of the error parameters. Rather, the statistical variability of the calibration assay translates into a systematic

[7]The binomial variance is proportional to $p(1-p)$, which achieves a maximum when the probability of success $p = 0.5$ and decreases as p approaches 1.

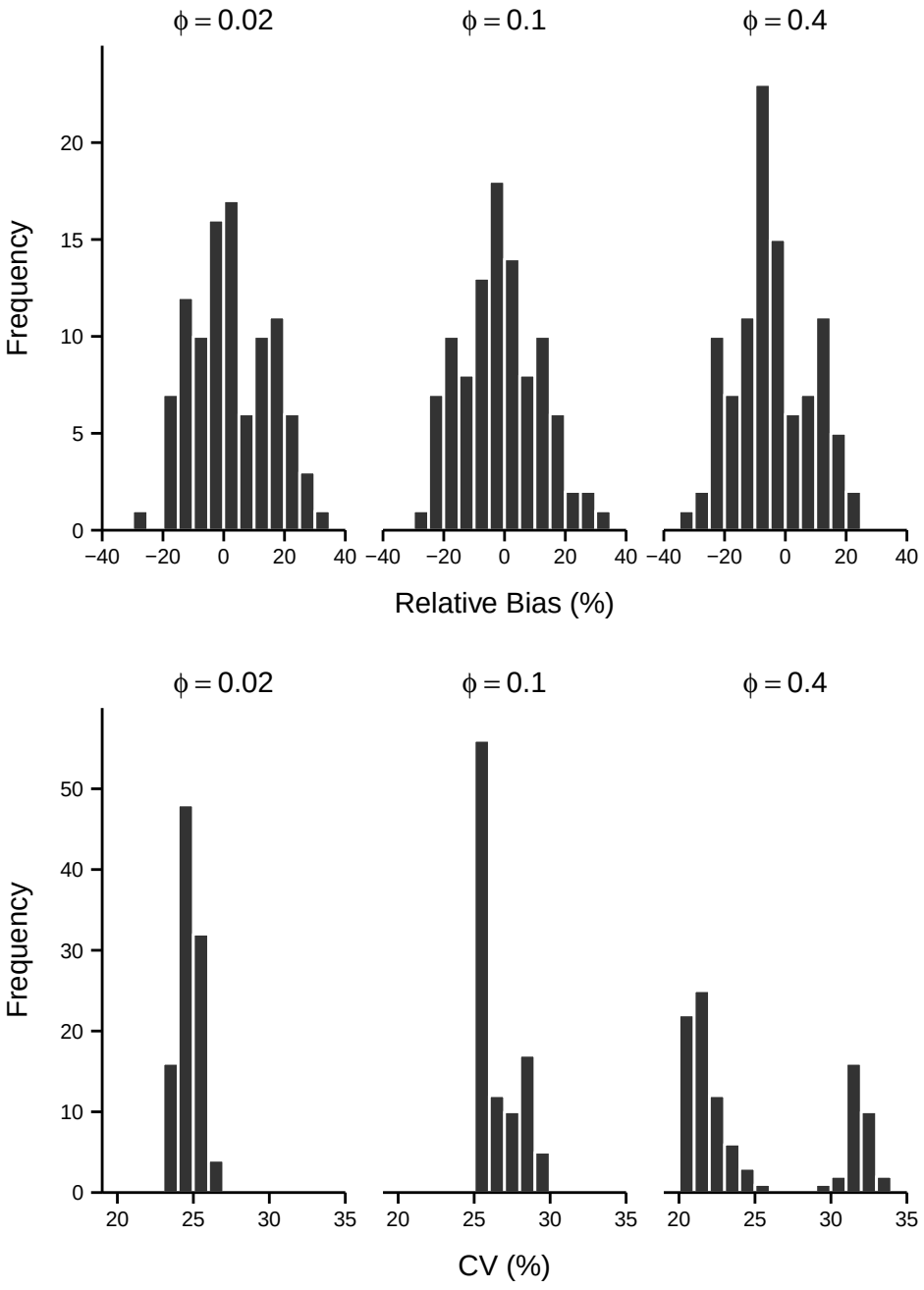

Figure 2.11: Relative bias and coefficient of variation (CV) of $\hat{\phi}$ predicted for a two-step assay with error parameters $f_{reac} = 0.1$ and $f_{deg} = 0.2$. Histograms show the distribution of each statistic for 100 simulated estimations of (f_{reac}, f_{deg}). All assay parameters are the same as in Table 2.3.

(N_{inh50}, r_{inh})		**Relative Bias** (mean (s.d.) in %)	**CV** (mean (s.d.) in %)
(20, 0.1)	known	10	37
	estimated	2 (18)	37 (1)
(20, 0.25)	known	7	33
	estimated	3 (9)	33 (1)
(20, 1)	known	6	31
	estimated	5 (0.3)	31 (0.2)
(40, 0.1)	known	3	28
	estimated	-4 (13)	28 (1)
(40, 0.25)	known	2	25
	estimated	2 (3)	25 (0.2)
(40, 1)	known	2	25
	estimated	2 (0.1)	25 (0.1)

Table 2.4: Relative bias and coefficient of variation (CV) of $\hat{\phi}$ predicted for a two-step assay (exact N sampling with $D = 6$, $R_m = 2$, $R_s = 38$ and $\phi_{min} = 0.01$), for known and estimated values of the inhibition parameters, with $\phi = 0.02$ in all cases. Inhibition parameter estimates are based on 100 simulations of a two-step calibration assay: 3 replicates each for [5, 15, ..., 95] marked pollen grains, followed by 5 replicates each for [n-4, n-2, n-1, n, n+1, n+2, n+4] grains, where n is the estimate of N_{inh50} after the first step.

error, when the estimated error parameters are used to interpret the outcome of a series of experiments performed on samples of unknown composition. These results suggest that to minimize the risk for bias, the calibration assay should include many more reactions than the subsequent dilution assays performed on individual samples (whereas I had set the total number of reaction to 50 in both cases).

I used a similar method to simulate the estimation of inhibition parameters N_{inh50} (the number of pollen grains causing a 50% failure rate) and r_{inh} (determining the steepness of the inhibition curve near N_{inh50}). I chose a two-step calibration assay, where the first step covers a large range in N to approximately locate N_{inh50}, where the second step focuses on pollen numbers near that coarse estimate of N_{inh50}. The logistic dose-response curve was fitted using the *drc* package (Ritz and Streibig 2005) in R (script included in Appendix A).

Based on the results for f_{reac} and f_{deg}, I could expect that the variability of estimates of r_{inh} and N_{inh50} would create potential for a large bias in $\hat{\phi}$. However, as shown in Table 2.4, the impact of estimation errors decreases as r_{inh} increases. The larger the value of r_{inh}, the closer N must be to N_{inh50} before inhibitory effects become significant. As discussed previously, I adapted my PCR dilution assay in the presence of inhibition, to limit the number of pollen grains by reaction to 60% of N_{inh50}. When r_{inh} is large enough, this limit is far enough from the center of the inhibition curve that a precise estimate of the steepness parameter r_{inh} becomes less critical.

Since the primary purpose of this section was to illustrate the effects of different types of PCR error, I have not addressed issues related to selecting and validating an error model. This is an area where more experimental work is needed: in my review of the literature, I could not find any systematic study of errors in pollen PCR, despite the growing number of studies using the technique.

For certain studies, it may also be inappropriate to assume that all samples would follow the same model of PCR assay failure. If any environmental covariate is suspected to significantly affect amplification success (for example, when the age of pollen varies among samples), calibration assays should take its effect into account.

2.7 Implications for the design and analysis of PCR dilution assays on pollen

In many applications of pollen analysis, each unit of sampling (e.g. pollen collected by an aerial sampler in a day, or by a bee in a single foraging trip) may consist of thousands of individual grains. When the cost of individual PCR reactions is the main limiting factor, dilution assays provide a efficient method to estimate the level of presence of a identifiable genetic type, requiring less experimental effort than single pollen methods. In this section, I synthesize the results of this chapter and outline how the PCR dilution assay method can be applied to environmental samples.

Due to the low DNA copy numbers involved in pollen PCR, the possibility of assay failure cannot be neglected. Careful estimation of the function $Pr\big(+ \mid m, N_d\big)$ – the probability of success of a PCR reaction with N_d pollen grains, m of which possess the target marker – is thus required to minimize bias in interpreting assay results. This may be done by performing calibration assays on representative samples of known composition. Results of the previous section provide guidelines in the design of these calibration assays: in particular, failure occurrences at low pollen numbers (due for example to DNA degradation) and at high pollen numbers (i.e. inhibition or competition effects) should be studied separately.

Once the error parameters have been established, the next step is to design a PCR dilution assay that could provide reliable estimates of the target marker frequency (ϕ) over a wide range of that parameter. To achieve the required balance of efficiency and robustness, I proposed a two-step assay design:

- At the first step, R_m replicates are performed at each of D dilution levels, which form a geometric progression from 1 pollen grain per tube to some maximum N_D. That maximum is determined either by the desired detection limit of the assay (i.e. ϕ_{min}) or by the necessity to avoid PCR inhibition effects. More dilution levels may be necessary to cover a larger range; my results show that $D = 6$ is sufficient for a 100-fold range

($\phi_{min} = 0.01$). The number of replicates R_m per level can be kept small (1 or 2) with no loss in efficiency.

- An estimate $\hat{\phi}$ is inferred from step 1 results, and for the second step, R_s replicates are performed at the same dilution level, chosen to obtain an average of $N = 1/\hat{\phi}$ pollen grains per reaction. The majority of the experimental effort should occur at step 2. (For example, the best designs with 50 total reactions put at least 3/4 of the effort at step 2; designs with more total reactions can afford to put an even greater proportion at step 2.)

When developing the calculations of this chapter, I assumed the replicate pollen templates at each dilution level would be obtained by one of two sampling methods. What I termed exact N sampling is not a dilution method per se, since it relies on manually picking a constant number of pollen grains for each replicate. I referred to the more traditional dilution method as Poisson sampling: the whole sample is suspended in water and the initial pollen concentration is determined (for example, using a hemocytometer), then dilutions are made from this original sample to achieve the required average number of pollen grains per replicate. When using this method, the experimenter should test the protocol to ensure the resulting number of pollen grains per replicate follows the Poisson distribution.

While exact N sampling increases the precision of the assay, this effect is more pronounced at low pollen numbers. Therefore, one option would be to use Poisson sampling only for dilution levels where N is moderately large, as exact N sampling becomes impractical. Alternatively, the experimenter may use the dilution method to produce all replicates, but examine the low N replicates under a microscope prior to PCR, to determine the exact number of pollen grains in each. Such hybrid sampling schemes, where N is exactly known for some replicates and follows a Poisson distribution for the others, can be accomodated in my model by a simple modification of eq. (2.10) or (2.11).

The estimate $\hat{\phi}$ can be inferred after each step of the PCR dilution assay using the bias-corrected maximum likelihood method. The detailed equations and source code for these calculations are in Appendix A. Since that program calculates the theoretical distribution of assay outcomes for a given assay design and true value of ϕ, it can be used to estimate dispersal measures (e.g. root mean squared error, confidence intervals) for $\hat{\phi}$.[8]

Finally, I want to emphasize that the precision and accuracy measures calculated in this chapter are only valid within the assumptions of the stated model. In particular, the use of binomial and Poisson probabilities relies on the assumption that each replicate is an independent subsample from a perfectly mixed initial sample, and the calibration procedure assumes that an experimenter can obtain some control sample of known genetic composition that will share the same error model as the samples of unwknown composition.

[8]While the true ϕ is unknown, I can assume it to be equal to the experimentally found $\hat{\phi}$ for the purpose of estimating dispersal measures. In particular, Table 2.1 above shows that the relative root MSE of the estimate remains constant over a large range of ϕ.

Chapter 3

Floristic mapping through bee pollen

This chapter focuses on the use of bee pollen sampling to obtain information about the distribution of floral resources. More specifically, I will develop an individual-based model of bee foraging behavior, with the goal of relating the spatial distribution of a genetic marker – which could identify a plant taxon or variety – to its level of presence/absence in corbicular pollen loads.

As mentioned in the introduction, both the honey bee (*Apis mellifera*) and bumblebees (*Bombus spp.*) collect pollen in the corbicula (or pollen baskets) located on their hind legs. Most published analyses of bee pollen loads are based on morphological pollen identification methods; the results have been used to estimate foraging distances (Beil et al. 2008) or to compare vegetation profiles between widely separated locations (Diaz-Losada et al. 1998).

I already discussed some of the advantages of genetic approaches to determine the composition of a pollen sample: greater phylogenetic resolution (to the species, variety or ecotype level), less reliance on subjective visual identification, and potential to benefit from advances in high-throughput single-cell PCR methods. Yet, the use of genetic methods for bee pollen analysis remains very rare. The only example I could find comes from a large-scale study of gene flow between oilseed rape (*Brassica napus*) varieties in the UK. As part of that study, Ramsay et al. (2003) performed DNA fingerprinting on whole pollen loads collected at a single hive, to show that different transgenic and non-transgenic oilseed rape varieties were represented.

Whether they use morphological or genetic identification methods, the studies I consulted report aggregate data and put little emphasis on the variation between individual pollen loads. To my knowledge, the work I report here is the first attempt to model the statistics of bee pollen composition, based on parameters of a floral resouces distribution and a bee foraging model.

In defining the models below, I refer in particular to honey bees, since they offer the most practical opportunities for pollen sampling: pollen loads can easily be collected in a

commercial bee hive, by making foragers pass through a mesh at the hive entrance. There is also a large existing literature on the foraging behavior of honey bee colonies. However, I suspect that the model could be extended to other species (such as bumblebees), or adapted to the analysis of pollen loads from bees caught in the field.

3.1 Response variables

The first step in designing the model is to choose response (output) variables that correspond to observable quantities, or statistics derived from those basic observed quantities.

After a pollen load is dissolved and mixed, the proportion of pollen grains bearing a given genetic marker can be determined through either single-pollen methods or a dilution assay as described in Chapter 2. Therefore, I will use this proportion (p, also termed frequency) of the genetic marker of interest within a single pollen load as the unit output for my model.

Under this premise, the genetic analysis of N_l pollen loads randomly picked from a single source – for example, from all the pollen loads collected by a single hive trap in an hour – would result in a data set $\{p_i | i = 1, \ldots, N_l\}$.

In addition to the full distribution of the p_i (respresented as a histogram), I will focus on two summary statistics of this distribution: the mean $\overline{p}$ and the measure of genetic differentiation F_{ST}. The latter statistic comes from population genetics, where it is known as the fixation index and represents the portion of the total genetic variance that is due to variation between, as opposed to within, populations of a species. Taking each pollen load as a "population", F_{ST} is calculated as:

$$F_{ST} = \frac{{\sigma_p}^2}{\overline{p}(1-\overline{p})},$$

where ${\sigma_p}^2$ is the variance of the p_i.

The interpretation of $\overline{p}$ must depend on the specific form of the model, but it can be generally seen as a weighted average of the marker frequency in the surrounding floral landscape, with a weight depending on factors such as distance from the hive, bee foraging preferences, etc.

The choice of F_{ST} as a response variable relates to a key hypothesis which motivated this work: namely, that the genetic differentiation between pollen loads is a function of the scale of spatial genetic structure, relative to the foraging area for a single bee. If plants bearing the marker of interest are present in homogeneous patches that are large compared with the area covered in a single foraging bout, I expect highly differentiated pollen loads ($p = 0$ or 1) and a F_{ST} close to 1. Conversely, if genetically homogeneous patches are small compared with the typical bee's foraging area, bees will collect mixed pollen loads with similar values of p, and F_{ST} should approach 0.

3.2 Modelling approach and key parameters

Both the plant distribution over a landscape as well as the foraging behavior of honey bees can be described in considerable detail. However, rather than including all factors that could affect the genetic composition of pollen loads from the outset, I adopt an incremental approach that starts with simple, perhaps unrealistic, models where both the landscape and the foraging behavior are described by a few parameters. Starting with high complexity models would make it difficult to explore the effect of individual parameters. The incremental approach may also identify features of the outcome, such as summary statistics, that are not very sensititive to the detailed form of the model used.

All models described below share the following features:

- The composition of each pollen load is determined by simulating a bee foraging bout. This individual-based modelling approach seems like a natural choice for this problem, since I am primarily interested in variation in composition between pollen loads.[1] The individual-based model also readily predicts the variability associated with a given sampling effort (number of loads analyzed).
- Each simulated bee visits a sequence of locations in (continuous) 2D-space and collects pollen from a single plant at each location.
- The probability of presence/absence of the genetic marker in each plant visited is based on its location in the field.
- The resulting marker frequency in the pollen load (p) is calculated by adding the contribution of each plant; in particular, if all plants contribute the same amount of pollen, p is the proportion of plants visited that bear the marker.

Initially, I will assume bees forage in a continuous field of plants and their paths can be described by either uncorrelated or correlated (directional) random walks. After deriving an analytical expression for F_{ST} under the simplest foraging path – an uncorrelated random walk with no preferred starting point – I will present results from two different simulation models. In the first model, the distribution of marked and unmarked plants in the field is defined by its spatial mean and correlation function, and the probability that each plant along the bee's path has the marker is calculated through a method derived from kriging (geostatistical interpolation). In the second model, plants bearing the marker are confined to geometric patches of prescribed size and shape, randomly placed within the field of non-marked plants.

At the next step, I will use more realistic descriptions of the landscape, where the floral sources tapped by the bees are organized in fields of various extent, located at different

[1] It could be objected that each bee typically holds two pollen baskets, so it is technically incorrect to treat all pollen loads collected at the hive as the product of independent bouts. However, it remains a reasonable assumption if the sample of pollen loads analyzed is a small fraction of those collected by the trap.

Bee foraging behavior	
n_s	Number of steps (plants visited) in a foraging bout
l_s	Length (or RMS length) for a step
f_j	Frequency of jumps (occasional large steps)
l_j	Length (or RMS length) for a jump
ρ	Parameter of the wrapped Cauchy distribution; represents the degree of correlation between successive steps
p_{sw}	Maximum probability of switching to a different field (fragmented landscape model)
s_i	Score of field i (fragmented landscape model)
$T_{i,j}$	Transition probability from field i to field j (fragmented landscape model)
Spatial genetic structure	
ϕ	Average marker frequency in a field
a	Characteristic length of spatial correlations (kriging-based model)
p_s	Size of genetic patches (random patch model)

Table 3.1: Main parameters describing the bee foraging model.

positions around the hive. Each field can be a uniform mix of marked and unmarked plants, or have its own spatial genetic structure. Bees start foraging at one of the fields and may move to a different field in the middle of a foraging bout. Within fields, their movement is represented by (correlated) random walks, as in the previous case.

In this chapter, the word *field* does not refer necessarily to an agricultural field, but to any discrete floral "island" in the landscape where bees may forage for pollen. In other studies, this concept might be referred to as a *patch*. However, here I will only use the word patch to describe the spatial genetic structure within a field.

In Table 3.1, I introduce some key parameters that will appear in model descriptions. Their meaning will become clear as specific models are introduced in the following sections. A RMS (root mean square) length is the square root of the mean squared value of the length.

3.3 Continuous field models

3.3.1 Analytical calculation of F_{ST} for a simple random walk

For a continuous field, the spatial genetic structure of the marker of interest can be characterized by its average frequency ϕ and a correlation function $C(s)$ defined as:

$$C(s) = \frac{\mathbf{E}\left[I(x)I(x+s)\right] - \phi^2}{\phi(1-\phi)}. \tag{3.1}$$

Here, $I(x)$ is an indicator variable equal to 1 if the marker is present in the plant at x (x here represents a position in 2D rather than just one of its coordinates), or 0 if it is absent.[2] The expected value in the numerator is an average over all pairs of points separated by a distance of s, irrespective of direction.

I assume each bee starts foraging from a random point in the field and follows a simple random walk with n_s steps, a RMS step length equal to l_s, and no preferred direction of movement. I further assume that the proportion of marked pollen in the bee's pollen load is equal to the proportion of marked plants visited. In Appendix B, I show how F_{ST} can be calculated from $C(s)$ in this particular case, with the following result:

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2} \sum_{k=1}^{n_s-1} (n_s - k) \int_0^\infty C(s) \frac{2s}{k{l_s}^2} \exp\left(-\frac{s^2}{k{l_s}^2}\right) \mathrm{d}s. \tag{3.2}$$

Given a specific form of $C(s)$, eq. (3.2) can be integrated numerically to provide a prediction of the genetic differentiation between bee pollen loads. While this expression is only strictly valid under the stated assumptions, it provides a theoretical benchmark for the simulation results presented below. Furthermore, I will show that different bee movement models produce results that, once properly re-scaled, bear strong similarity to this theoretical prediction.

3.3.2 Kriging-based model

In the previous section, I describe the spatial distribution of the marker of interest by its average frequency ϕ and its correlation function $C(s)$. A spatial distribution that is completely defined by these two characteristics is termed Gaussian stationary.[3] Gaussian stationary random fields can be simulated by a number of methods based on kriging; here I will use sequential indicator simulation (Goovaerts 1997).

First, I describe how to choose the list of points on the bee's foraging path. Starting from the origin (0,0)[4], the bee's path is simulated as a correlated random walk (CRW), which has been suggested by previous studies of animal foraging behavior (Kareiva and Shigesada 1983). At each point of the path, two random numbers (u and v) are generated from a uniform distribution over (0,1); the length Δl and turning angle $\Delta\theta$ of the next step are calculated as

$$\Delta l = l_s \sqrt{-\ln u}$$

and

$$\Delta\theta = 2 \arctan\left[\frac{1-\rho}{1+\rho} \tan\left(\pi(v - 0.5)\right)\right].$$

[2] $I(x)$ could also take a value of 0.5 for heterozygotes, although I will ignore this case for simplicity.

[3] The traditional definition of Gaussian stationarity uses the covariance function, which is just the numerator of eq. (3.1).

[4] The starting point is arbitrary in this particular model, since the kriging-based simulation only requires the distances between points.

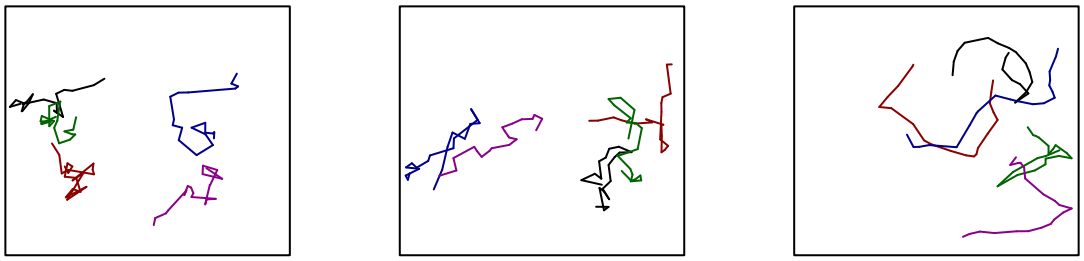

Figure 3.1: Correlated random walks simulated using different values of ρ. The number of steps ($n_s = 20$) and step length l_s are the same for all walks.

The process is repeated until n_s - 1 steps are taken (the initial point counts as step 1). Using this algorithm, the squared step length follows a χ^2 distribution with two degrees of freedom, with a scaling factor to make its RMS value equal to l_s. The turning angle follows a wrapped Cauchy distribution, with the parameter ρ indicating the tendency of the bee to maintain its direction (see Fig. 3.1 for an illustration of its effect). When $\rho = 0$, the turning angle is uniformly distributed over (-π, π) and the path is a true (uncorrelated) random walk; when $\rho = 1$, the path is a straight line (Bartumeus et al. 2005).

In testing the model, I will also try slight variations of this base algorithm, such as:

- introducing some variation to n_s from one bee to the other;
- having all steps be the same length l_s; and
- adding larger steps, or jumps, of RMS length l_j and at a frequency f_j (i.e. $f_j = 0.1$ is a jump every 10 steps).

From the matrix of distances $[d_{i,j}]$ between each pair of points in the bee's path, it is possible to compute a n_s by n_s correlation matrix $[C_{i,j}]$ from the known function $C(s)$.

The program then determines the genotype I of each plant in the path. The first plant is picked as marked ($I = 1$) with probability ϕ, or unmarked ($I = 0$) with probability 1 - ϕ. For each successive plant in the path, the probability of having the marker is a function of the previously simulated values:

$$\Pr\left[I(x_k) = 1\right] = \phi + \sum_{i=1}^{k-1} w_i\big(I(x_i) - \phi\big).$$

The purpose of kriging is to select the weights w_i so as to get the best linear unbiased predictor of $I(x)$. When the global average (here ϕ) is known, these weights are the solution

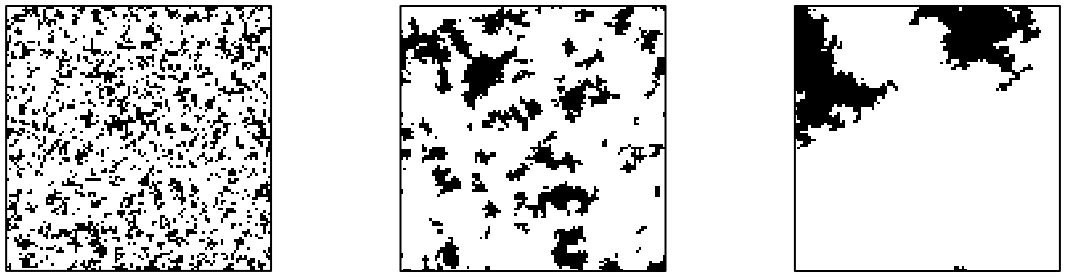

Figure 3.2: Random fields produced by sequential indicator simulation in gstat, using an exponential correlation function $C(s) = \exp(-s/a)$. Each simulation is a 100x100 field with average marker frequency (black points) $\phi = 0.25$.

of the simple kriging equations:

$$\begin{bmatrix} C_{1,1} & \cdots & C_{1,k-1} \\ \vdots & \ddots & \vdots \\ C_{k-1,1} & \cdots & C_{k-1,k-1} \end{bmatrix} \times \begin{bmatrix} w_1 \\ \vdots \\ w_{k-1} \end{bmatrix} = \begin{bmatrix} C_{1,k} \\ \vdots \\ C_{k-1,k} \end{bmatrix}.$$

Finally, the marker frequency in the pollen load, p, is taken as equal to the proportion of marked plants over all plants visited.

I implemented the algorithm described above in a Fortran 90 program, which is included in Appendix B.

To test this simulation algorithm, I will assume an exponential correlation function: $C(s) = \exp(-s/a)$, where a is a scaling parameter with units of distance. Figure 3.2 presents examples of random fields generated by that correlation function, using the R package gstat (Pebesma 2004). I note however that my algorithm does not simulate the whole field, only its values at the points visited by the bee.

I first consider the effect of a on the distribution of the marker frequency in individual pollen loads, setting the average marker frequency $\phi = 0.25$ and modelling bee paths as uncorrelated ($\rho = 0$) random walk with $n_s = 40$ steps. I set $l_s = 1$ for all simulations in this section, with no loss of generality as it amounts to expressing all other distances as multiples of the RMS bee step length.

As expected, the genetic differentiation between pollen loads increases with the scale of spatial genetic correlation. The increase in F_{ST} also corresponds to an increased occurence of genetically uniform ($p = 0$ or 1) pollen loads (Fig. 3.3). For values of a from 0.25 to 20, the simulation results agreed with the analytical result derived in Appendix B, with an average deviation of 1%.

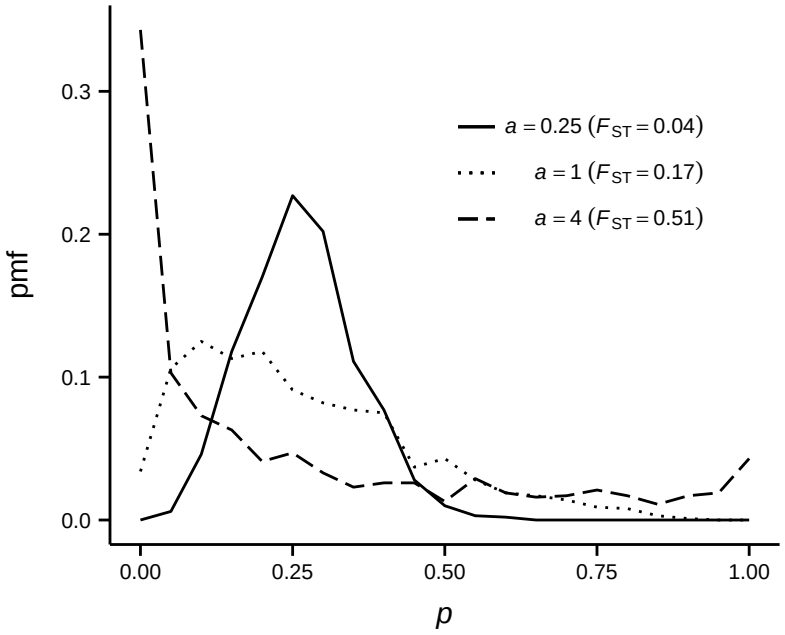

Figure 3.3: Probability mass function (pmf) of the marker frequency in individual pollen loads, from 1000 simulations using a kriging-based algorithm with exponential correlation $C(s) = \exp(-s/a)$. ($\phi = 0.25$, $n_s = 40$, $l_s = 1$ and $\rho = 0$)

I evaluated the variability of F_{ST} between simulation runs using this "baseline" set of parameters ($n_s = 40$, $l_s = 1$ and $\rho = 0$). The results (not shown here) suggest that:

- F_{ST} follows an approximately normal distribution;
- 10,000 simulated loads are sufficient to achieve a coefficient of variation of at most 1.5% of the value of F_{ST};
- neither the value of F_{ST} nor its variance are significantly impacted by either having a fixed step length l_s, or a variable number of steps (normally distributed around n_s); and
- changing the average frequency of the marker (ϕ) has no impact on F_{ST}, but does affect its variance between simulation runs.

In Figure 3.4, F_{ST} is shown as a function of a for my baseline parameters, as well as variations obtained by changing n_s, adding larger steps (RMS length l_j) at a frequency f_j, or increasing ρ to create correlated random walks. The similarity between the different curves is made apparent by plotting F_{ST} against the ratio of a to the RMS distance between the first and last plants in the path, $\overline{D}$, the latter determined through one of the following equations (see Kareiva and Shigesada 1983 and Wu et al. 2000 for the correlated random

walk case):

$$\overline{D} = \begin{cases} l_s\sqrt{n_s - 1} & \text{for an uncorrelated RW,} \\ \sqrt{(n_s - 1)\left[(1 - f_j){l_s}^2 + f_j{l_j}^2\right]} & \text{for an uncorrelated RW with jumps,} \\ l_s\sqrt{(n_s - 1) + \frac{\pi}{2}\frac{\rho}{1-\rho}\left((n_s - 1) - \frac{1-\rho^{n_s-1}}{1-\rho}\right)} & \text{for a correlated RW with } \rho < 1. \end{cases}$$

To verify whether these results are limited to the exponential correlation case, I repeated the same set of simulations using a different correlation function, one which decreases linearly from 1 to 0 over a finite distance $2a$:

$$C(s) = \begin{cases} 1 - \frac{s}{2a} & \text{if } s \leq 2a, \\ 0 & \text{if } s > 2a. \end{cases}$$

Although this new correlation function produces a different F_{ST} vs. $a/\overline{D}$ curve, the shape of this curve is similarly conserved under changes of the bee foraging parameters. In the case of a simple random walk, the value of F_{ST} calculated analytically with this new $C(s)$ also agrees with the simulation output.

These simulations using the kriging-based model show a relationship between the composition of bee pollen loads (F_{ST}) and the spatial genetic structure of the field ($C(s)$) that can be predicted theoretically. This relationship is consistent for different types of random walks (one or two step sizes, correlated or uncorrelated), as long as the results are normalized by the average (specifically, RMS) distance between the first and last plants visited.

The model, however, relies on fairly stringent assumptions regarding the spatial genetic structure (Gaussian stationarity), which would not apply in most realistic cases. In the following sections, I will lift these assumptions and simulate bee paths in arbitrary field configurations.

3.3.3 Random patch model

This model, as the previous one, is based on a random arrangement of marked and unmarked plants within a single field. In this case, however, a specific 2D field configuration is created by placing geometrical patches of marked plants within a background of unmarked plants.

If the plant density over the field is assumed to be uniform, then the average frequency ϕ is equal to the proportion of the field area covered by marked patches. For example, within a 100x100 field, a frequency $\phi = 0.25$ can be attained by placing a hundred 5x5 patches, or twenty-five 10x10 patches. Since larger patches create longer-range genetic correlations, the patch size in this model plays a similar role as the parameter a used in the previous section.

For a given field configuration, an empirical correlation function can be defined using eq. (3.1), by taking the average over a large number of randomly-selected pairs of points for

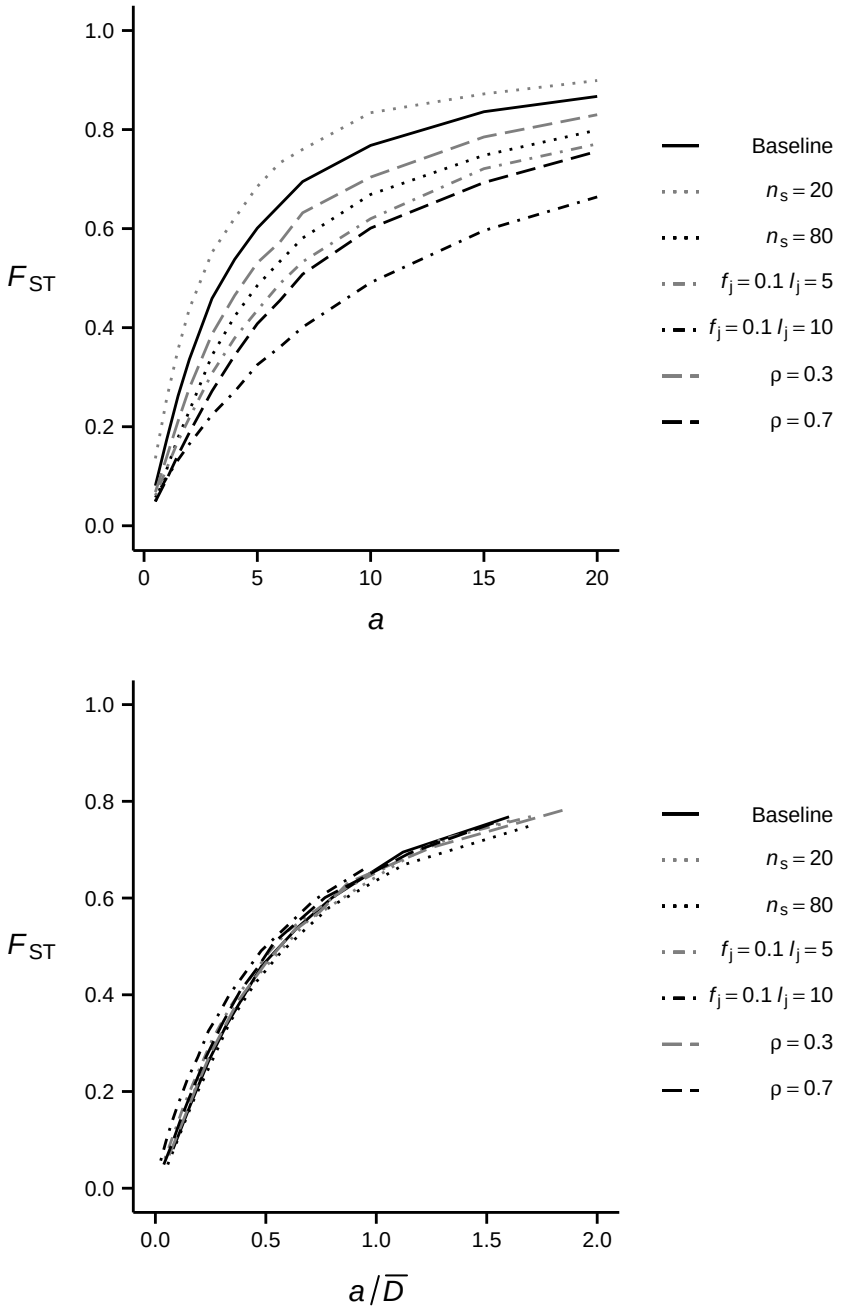

Figure 3.4: Genetic differentiation between pollen loads simulated by the kriging-based model, under an exponential correlation function and different parameters of the foraging model. Genetic differentation (measured as F_{ST}) is plotted against the correlation length a (top) and the ratio $a/\overline{D}$ (bottom), where $\overline{D}$ is the root mean square distance between the first and last plants visited. For the baseline case, $n_s = 40$ and $\rho = 0$ ($\phi = 0.25$, $l_s = 1$, and 10,000 pollen loads were simulated in all cases).

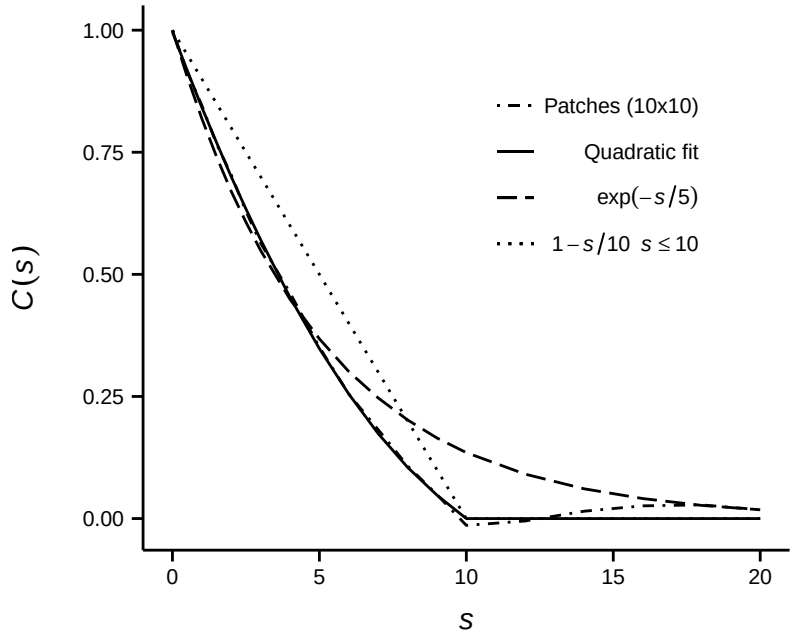

Figure 3.5: Spatial genetic correlation for a 100x100 field with twenty-five 10x10 patches of marked plants ($\phi = 0.25$). The correlation function was calculated for 18 values of s, using 100,000 random pairs of points for each s. The empirical curve is partially fit by a quadratic function (for $s \leq 10$) and compared with the exponential and linear correlation functions (with $a = 5$).

different values of s. Figure 3.5 shows the correlation function calculated from a 100x100 field with randomly placed 10x10 patches, comparing it to the exponential and linear forms of $C(s)$ used in the previous section. The empirical curve can be well approximated by a piecewise quadratic function:

$$C(s) = \begin{cases} as^2 + bs + c & \text{if } s \leq s_0, \\ 0 & \text{if } s > s_0. \end{cases}$$

I approximate $C(s)$ to be identically zero beyond a critical distance s_0 (here $s_0 = 10$, corresponding to the patch size), even though the empirical correlation becomes slightly negative, then slightly positive for larger distances.

From this approximate functional form of $C(s)$, I can predict the value of F_{ST} produced by a random walk foraging pattern by integrating eq. (3.2) (see Appendix B). Using the same baseline parameters as in the previous model ($\phi = 0.25$, $n_s = 40$, $l_s = 1$ and $\rho = 0$), I found that the predicted and simulated F_{ST} closely agree for various square and rectangular patch formats (Table 3.2). The discrepancy was greater for rectangular patches that were placed in the same orientation (all horizontal or vertical), compared to fields where both orientations were used in alternance. Since the $C(s)$ function is an average over all directions, it may not adequately describe a field with strong anisotropy (i.e. having a preferred orientation in space); however, the observed difference could also be due to a poorer fit of the quadratic

Patch size	Same orientation		Alternate orientation	
	predicted	simulation	predicted	simulation
10x10	0.601	0.607	0.597	0.582
12.5x8	0.579	0.567	0.571	0.572
16.6x6	0.552	0.536	0.516	0.515
20x5	0.492	0.455	0.489	0.490
33.3x3	0.463	0.441	0.388	0.374

Table 3.2: Comparison of F_{ST} values predicted from $C(s)$ and simulation results (10,000 pollen loads), for a 100x100 field with random rectangular patches ($\phi = 0.25$). Patches can all be in the same orientation or alternate between horizontal and vertical. Bees follow an uncorrelated random walk with $n_s = 40$ and $l_s = 1$.

function to $C(s)$.

In Figure 3.6, I compare the effect of patch size on F_{ST} for different parametrizations of the random walk. When each curve is scaled by the RMS distance from start to finish ($\overline{D}$), they are less similar than those obtained in the kriging-based model (Fig. 3.4). Furthermore, these differences cannot be explained solely by the variability of F_{ST} between simulation runs. Based on results from replicate simulations, the relative standard deviation of F_{ST} is comparable under the two models, i.e. around 1% for 10,000 simulated pollen loads.

According to these results, the relationship between F_{ST} and $C(s)$ observed in the kriging-based model, where a Gaussian random field was generated from the $C(s)$, is still valid under the random patch model, where an empirical $C(s)$ arises from the random placement of genetic patches of fixed geometry. However, the latter model is more sensitive to the exact parametrization of the foraging path.

The major assumption underlying eq. (3.2) is that every part of the field is equally likely to be visited *a priori*. This was the case in the simulations presented so far, since the starting position and direction for each bee was chosen at random. The effects of a preferred starting position will be investigated in the next type of models, which incorporate hive-to-field and field-to-field movements of the bees.

3.4 Fragmented landscape models

The previous sections focused on modelling the statistics of pollen loads gathered by bees foraging in a single field of uniform density and varying genetic structure. I will now consider the more realistic case of a hive allocating its foragers between several disjoints fields placed in a landscape.

We know from observational studies of honey bees that foragers will scout the landscape, return to the hive and recruit other foragers to the most suitable floral resources, using the

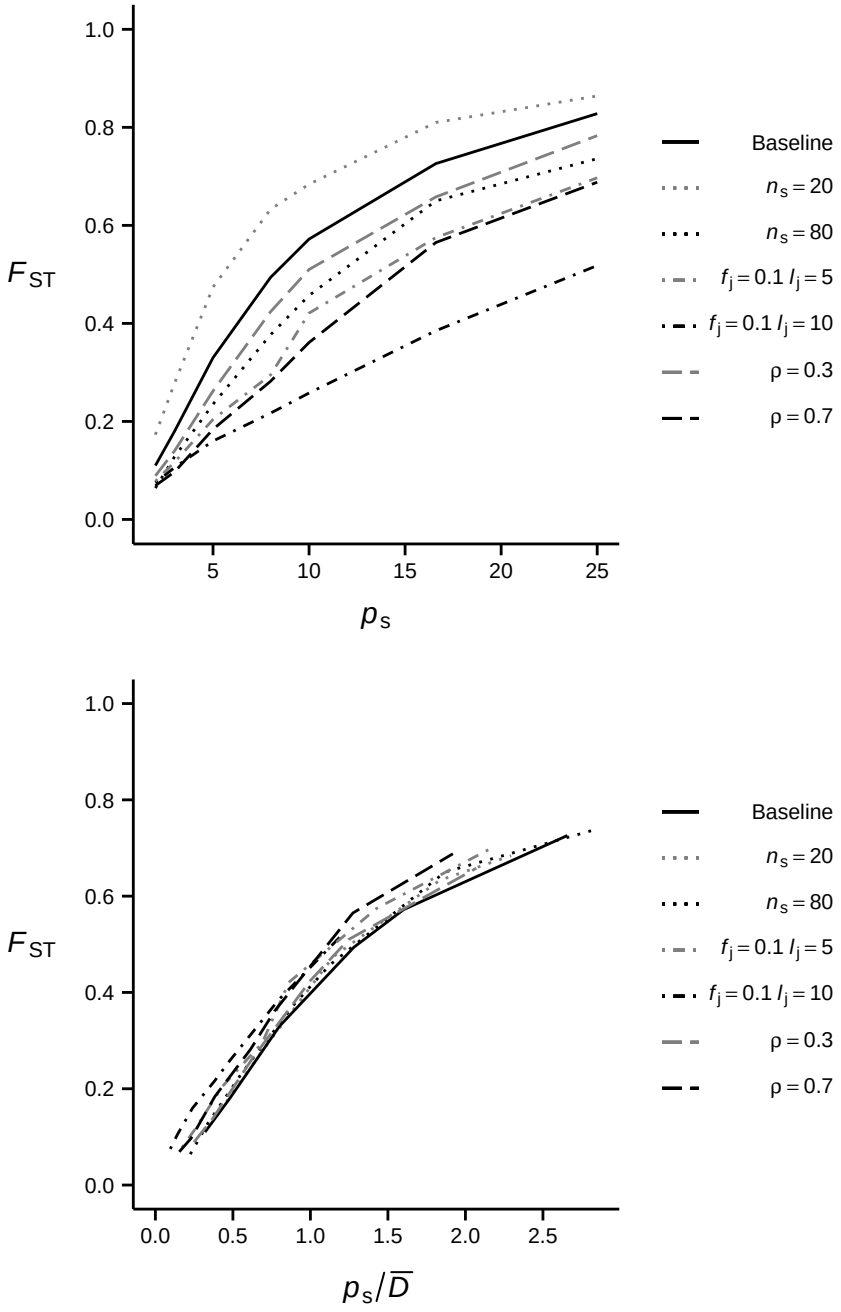

Figure 3.6: Genetic differentiation between pollen loads simulated by the random patch model (100x100 field with square patches). F_{ST} is plotted against the patch side length p_s (top) and the ratio $p_s/\overline{D}$ (bottom) for different parametrizations of the random walk. The baseline case corresponds to $n_s = 40$ and $\rho = 0$ ($\phi = 0.25$, $l_s = 1$, and 10,000 pollen loads were simulated in all cases).

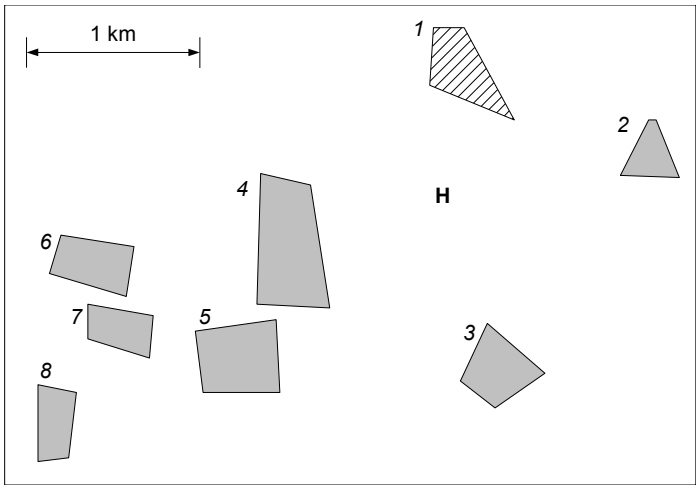

Figure 3.7: Example field layout used to illustrate the bee foraging models in a fragmented landscape, adapted with permission from Fig. 2 of Ramsay et al. (2003). Each *Brassica napus* field is numbered for reference in the text, and H denotes the bee hive position. Field 1 (hatched in this figure) was planted with a transgenic variety in the original study.

waggle dance (Visscher and Seeley 1982). The allocation of foragers between different fields, on a given day, is the outcome of this dynamic process. I will model this outcome as a "score" assigned to each field by the colony, which may depend on the floral species present, its distance from the hive, the time of year, weather, etc. My algorithm for simulating bee flights in a fragmented landscape will follow this general outline:

1. Input the landscape description (location, extent and genetic composition of each field; location of the hive) in the program.
2. Calculate a score for each field.
3. For each simulated bee, pick one of the fields to begin foraging, with probability proportional to the score. The foraging path within a field is modelled as in the previous sections.
4. (optional) With some probability, a bee may move to a different field while in the middle of a foraging bout.

Compared with the number of studies pertaining to the allocation of foragers, there is little information available in the literature to explain why a bee would split its foraging effort over disjoint fields in a single foraging trip, although it is reasonable to assume that the distance between fields is one such factor.

To illustrate model outcomes in this section, I will use a landscape layout adapted from the Ramsay et al. (2003) study previously mentioned. For that field study, the bee hive was

placed within a few kilometers of seven conventional *Brassica napus* fields and one transgenic field. To simplify calculations, I represent each field as a quadrilateral, as shown in Fig. 3.7.

3.4.1 Determination of $\overline{p}$ and F_{ST} from single-field statistics

In the case where bees remain in the same field for the duration of a foraging bout, the distribution of p in pollen loads sampled at the hive is a combination of the different fields' contributions, weighted by their scores. If, for each of the n_f fields, both the expected statistics $(\overline{p})_i$, $(F_{ST})_i$ – estimated by one of the single-field models described above – as well as the score s_i are known, the overall mean $\overline{p}$ is predicted from the law of total expectation:

$$\overline{p} = \mathbf{E}[\mathbf{E}(p|i)] = \mathbf{E}[(\overline{p})_i] = \sum_{i=1}^{n_f} s_i (\overline{p})_i, \tag{3.3}$$

and the expected F_{ST} using the law of total variance:

$$\begin{aligned} F_{ST} &= \frac{\mathbf{E}\left[{\sigma_{p|i}}^2\right] + {\sigma_{\mathbf{E}(p|i)}}^2}{\overline{p}(1-\overline{p})} \\ &= \frac{\sum_{i=1}^{n_f} s_i (F_{ST})_i (\overline{p})_i \big(1 - (\overline{p})_i\big)}{\overline{p}(1-\overline{p})} + \frac{{\sigma_{(\overline{p})_i}}^2}{\overline{p}(1-\overline{p})}. \end{aligned} \tag{3.4}$$

Note that I use $(\dots|i)$ as a shorthand notation for probabilities conditional on the bee foraging in field i. Within a given field, if all plants provide on average the same amount of pollen and have the same *a priori* probability of being visited[5], $(\overline{p})_i$ corresponds to the field-averaged marker frequency ϕ_i.

The first term in eq. (3.4) depends on genetic differentiation between pollen loads from the same field, while the second term reflects the genetic differentiation between fields.[6] Without prior information on the distribution of the genetic marker across the landscape, it may not be possible to determine the relative contribution of these two factors to the observed F_{ST} in sampled pollen loads.

To further complicate matters, it is possible that a bee forages in more than one field. I introduce field-to-field movement in the model by the way of a transition matrix T, where each element $T_{i,j}$ is the probability that a bee foraging on field i moves to field j at the next step ($T_{i,i}$ is the probability of remaining in field i). As a first approximation, I assume that these probabilities are independent of either the position of the bee in the field, or the number of flowers previously visited. The simulated foraging path thus constitutes a Markov chain with n_f states (each of the fields) and discrete time steps.[7]

[5]The same assumptions appear in the single-field models above.

[6]To calculate F_{ST} between fields in the population genetics sense, the contribution of each field should be weighted by its population, instead of the score s_i.

[7]A Markov chain is a random process where the probability of being in some state at time t only depends on the state of the process at time $t-1$. See Levin et al. (2009) for a recent presentation of the theory of Markov processes.

Assuming that field-to-field movement can be approximated as a Markov process, the probability of foraging in field i after k steps can be calculated from the score vector s (the probabilities of starting in each field) and the k^{th} power of the transition matrix:

$$\Pr\left(i \text{ at step } k\right) = \sum_{j=1}^{n_f} s_j (T^k)_{j,i}. \tag{3.5}$$

For a bee visiting a total of n_s plants in a foraging bout, the expected number of plants visited in field i (n_i) is just the sum of the probability (3.5) from step 0 (start of foraging) to n_s-1:

$$\mathbf{E}[n_i] = \sum_{k=0}^{n_s-1} \sum_{j=1}^{n_f} s_j (T^k)_{j,i}. \tag{3.6}$$

I illustrate this result for the landscape presented in Fig. 3.7, using inverse distance functions for both the scores and transition probabilities:

$$s_i = \tfrac{C_s}{d_{i,H}}\,; \qquad T_{i,j} = \begin{cases} \frac{C_t}{d_{i,j}} & \text{if } i \neq j \\ 1 - \sum_{\substack{k=1 \\ k\neq i}}^{n_f} T_{i,k} & \text{if } i = j \end{cases}. \tag{3.7}$$

In the equations above, $d_{i,H}$ is the shortest distance from the hive to field i; $d_{i,j}$ is the distance from field i to j, measured between their centroids; C_s is a normalization constant to make the s_i sum to 1; and C_t is set so that the probability of a bee switching fields is at most p_{sw} (i.e. $\min\{T_{i,i}\} = 1 - p_{sw}$). As p_{sw} increases, there are less disparities between the contributions of each field to the total pollen collected (Fig. 3.8). This can be explained by the fact that the inverse distance weights produce a symmetric transition matrix, i.e. $T_{i,j} = T_{j,i}$, and that the limiting behavior of such a Markov chain (when the number of steps becomes large enough) is to visit all states with the same probability (Levin et al. 2009).

The expected number of visits in each field can also serve to calculate the expected overall mean $\overline{p}$ from the single-field means ($\overline{p}_i$):

$$\overline{p} = \mathbf{E}[\mathbf{E}(p|\{n_i\})] = \frac{1}{n_s} \sum_{i=1}^{n_f} (\overline{p})_i \mathbf{E}[n_i]. \tag{3.8}$$

Determining the exact distribution of p in bee pollen loads within this model poses a greater challenge, as it would require the enumeration of all possible sequences of n_s steps taken in n_f fields. In Appendix B, I derive the expected value of F_{ST} in the specific case where all plants in one of the fields ($i*$), and only those plants, have the marker:

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2 \overline{p}(1-\overline{p})} \sum_{k=0}^{n_s-2} \sum_{l=k+1}^{n_s-1} \left[\sum_{j=1}^{n_f} s_j (T^k)_{j,i*} (T^{l-k})_{i*,i*} - \overline{p}^2 \right]. \tag{3.9}$$

This corresponds to the scenario depicted in Fig. 3.7, where only field 1 is planted with a transgenic variety.

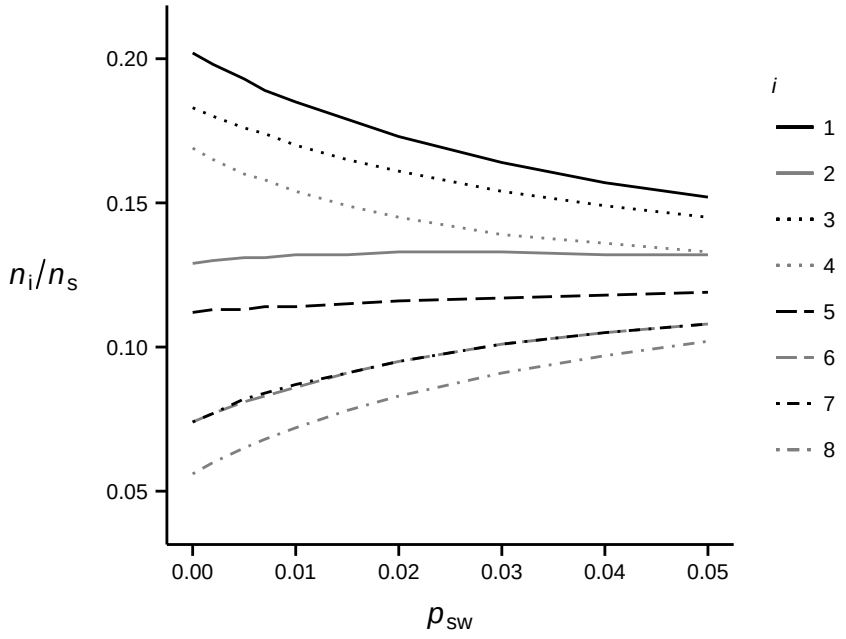

Figure 3.8: Average fraction of plant visits in each field (n_i/n_s) as a function of the maximum probability of switching fields (p_{sw}), for the landscape configuration shown in Fig. 3.7. Field scores and transition probabilities are based on inverse distance weighing (eq. 3.7) and $n_s = 100$.

3.4.2 Simulation results under specific landscape scenarios

The central problem of this chapter is to relate the composition of pollen loads collected at a bee hive to properties of the distribution of plants surrounding the hive. In a fragmented landscape model, pollen loads from different fields are aggregated and the overall statistics ($\overline{p}$, F_{ST}) do not have a simple, unique link to some property of the spatial genetic structure. However, if I can formulate specific hypotheses about this spatial genetic structure or foraging behavior, I can simulate the distribution of p in bee pollen loads predicted by these competing hypotheses and compare them with the observed bee pollen composition.

Using the landscape depicted in Fig. 3.7 from the Ramsay *et al.* study, I will compare the distribution of p in different scenarios that could result in a mixture of transgenic and conventional *Brassica napus* pollen in bee pollen loads. Before making these comparisons, I must specify the some parameters of the simulation method.

I model the movement of bees within a field as a correlated random walk, as in the continuous field case. Typical stand densities for commercial canola fields range from 30 to 140 plants/m^2 (May et al. 1994). These densities correspond to an average plant spacing of 0.08 to 0.18m, which I will consider as the lower bound for the average step size l_s. I could not find a good estimate for the number of plants visited (n_s) by honey bees in a canola field; however, Cresswell et al. (2002) estimated that bumblebees foraging on canola visit 100 to

500 inflorescences per bout. For my simulations, I chose a value at the low end of that range ($n_s = 100$) along with a relatively large step size ($l_s = 1$m, corresponding to 5-10 times the typical plant spacing). Due to the lack of data on which to base ρ, the degree of correlation of the random walk, I will use both $\rho = 0$ (uncorrelated walk) and $\rho = 0.7$ (highly-correlated walk).

I consider field 1 to be completely transgenic ($\phi = 1$) and fields 3 to 8 to be completely conventional ($\phi = 0$) in all simluations. Transgenic plants may either be absent from field 2, in which case the only source of mixed pollen loads will be movement of bees between field 1 and the other fields, or it may be present at a low frequency, which I set at $\phi = 0.1$. In the latter case, transgenic plants are either spread uniformly across the field or in randomly-placed square patches of side length p_s. The main reason I do not consider a lower value of ϕ is to get a sufficient amount of marked pollen in the results. Even the smallest field in the landscape (field 2) contains a few million plants based on its surface area ($\approx$ 55,000 m^2), so individual bees sample less than 0.01% of the field on each trip.

After the bee selects a field where it begins foraging (with probability based on the field score), the starting point within the field is either randomly selected in that field (random start method), or set to the point closest to the hive (closest point method). As in the previous section, I use inverse-distance weights for field scores and transition probabilities (eq. 3.7). However, I measure field-to-field distances $d_{i,j}$ from the current position of the bee in field i to the closest point in field j; the normalization constant C_t is still computed only once, based on the distances measured between field centroids. Transitions probabilities thus change based on the position of the bee in the field, a departure from the Markov chain assumption.

The main effect of the closest point method is to confine most bee paths to one sector of each field; as such, it would affect the composition of bee pollen loads if one or more fields are not genetically uniform. In theory, it could also affect field-to-field movement, since the transition probabilities depend on the bee's specific position in the field: however, my simulation results for uniform fields (one transgenic and seven conventional ones) show that the choosing one method or the other has very little effect on F_{ST} in that case (Fig. 3.9). In fact, even the F_{ST} values predicted by the Markov chain model (eq. 3.9), using transition probabilities based on the distance between field centroids, remain close to the simulation results that are based on the specific bee position. This is especially true if field-to-field movements are rare events (i.e. less than 1 in 100 foraging steps). Although I expected the choice of method to have a greater effect when the transgenic field has closer neighbors (e.g. field 7 compared to field 1) and thus more frequent transitions, Fig. 3.9 shows the opposite pattern.

I can now turn back to the original problem of whether or not the distribution of p in mixed pollen loads is indicative of the cause of this mixture. More specifically, I will consider the two following scenarios: (1) bees do not switch fields while foraging, but transgenic plants are present in field 2 (with $\phi = 0.1$); or (2) transgenic plants are confined to field 1, but bees

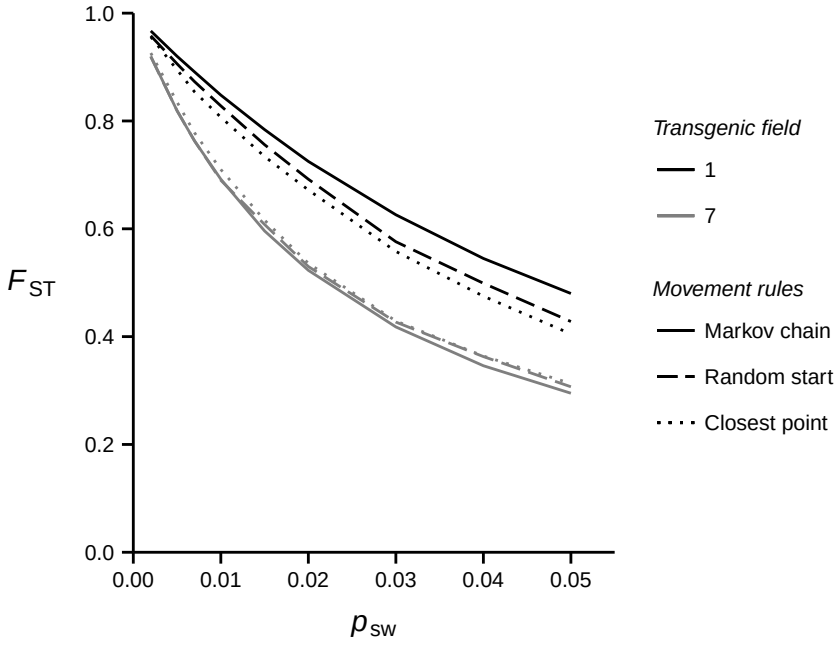

Figure 3.9: Genetic differentiation (F_{ST}) of bee pollen loads simulated for the landscape of Fig. 3.7, where only one of the fields is planted with a transgenic variety and bees can switch fields with a maximum probability of p_{sw} ($n_s = 100$, $l_s = 1$m, and 10,000 pollen loads are simulated in all cases). Bees either start foraging at a random point in the first field, or at the point closest to the hive. F_{ST} curves calculated from a Markov chain model (eq. 3.9) are included for comparison.

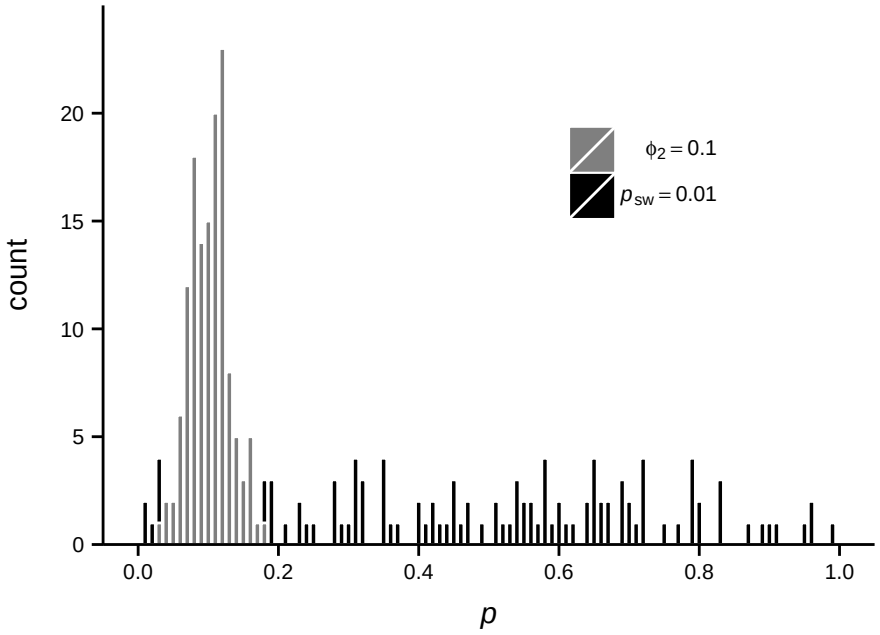

Figure 3.10: Distribution of the marker frequency p in 1,000 bee pollen loads simulated for the landscape shown in Fig. 3.7, under two different scenarios: (1) field 2 is a uniform mixture of plants with and without the marker ($\phi_2 = 0.1$); or (2) bees have some probability of switching fields after each step ($p_{sw} = 0.01$). In both scenarios, marked plants form the totality of field 1 ($\phi_i = 1$) and are absent from fields 3 to 8. Pollen loads with a p of 0 or 1 (approx. 870 in each scenario) are not shown in this figure.

can switch fields where foraging (with $p_{sw} = 0.01$).

If field 2 is a homogeneous mixture of marked and unmarked plants (i.e. no genetic patches) under the first scenario, then the resulting distribution of p is visibly different from that obtained under the second scenario (Fig. 3.10). That difference is also apparent when $\overline{p}$ are F_{ST} are calculated from the mixed pollen loads in each distribution. A uniform mixture in field 2 results in values of p clustered around ϕ ($\overline{p}_{(mixed)} = 0.10$, $F_{ST(mixed)} = 0.01$); whereas for rare occurences of field-to-field movement, there is a broad, nearly uniform distribution centered near $p = 0.5$ ($\overline{p}_{(mixed)} = 0.45$, $F_{ST(mixed)} = 0.33$).

The difference becomes less clear when field 2 has a coarse-scale spatial genetic structure, simulated here as randomly-placed square patches. Figure 3.11 shows how four statistics of p (the overall $\overline{p}$ and F_{ST}, as well as their values among mixed pollen loads only) vary with the patch size p_s, for both uncorrelated and correlated ($\rho = 0.7$) bee foraging paths. Since only about 13% of bees forage in field 2, changing the scale of genetic patches in that field has very little impact on the overall F_{ST}, which remains close to 1 as all the other fields have a ϕ of 0 or 1. Larger patches result in less mixed pollen loads, but also in higher values of $\overline{p}$ and F_{ST} for these mixed loads, closely matching those produced by field-to-field movement (e.g. for 30x30m patches, $\overline{p}_{(mixed)} = 0.44$ and $F_{ST(mixed)} = 0.33$).

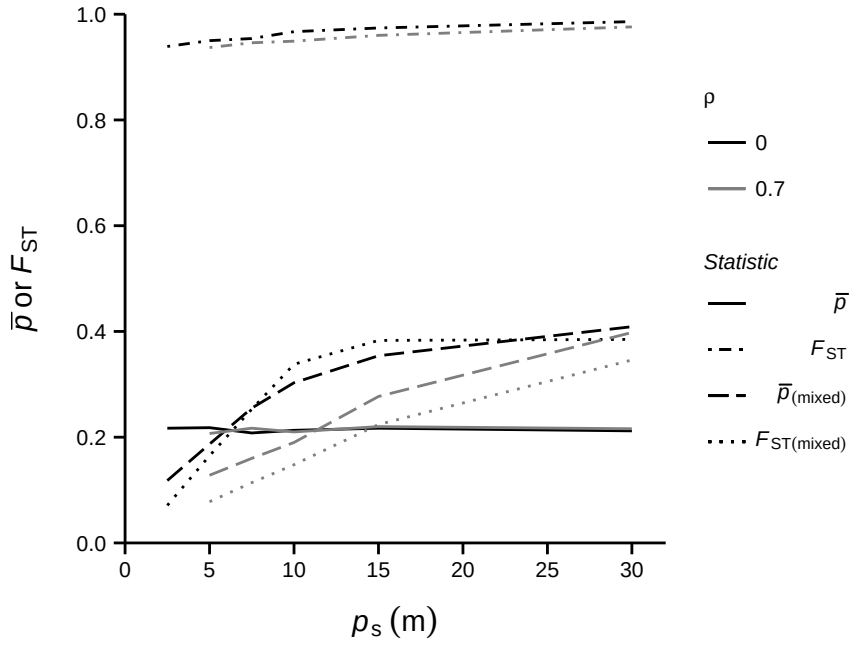

Figure 3.11: Statistics of the distribution of p in bee pollen loads simulated for the landscape of Fig. 3.7, with transgenic plants forming 10% of field 2 and arranged in randomly-placed square patches of side length p_s. Bees start foraging at a random point in the field and do not switch fields during a bout ($n_s = 100$, $l_s = 1$m, and 10,000 pollen loads are simulated in all cases).

3.5 Applications and limitations of this model

The primary objective of this chapter was to introduce a theoretical framework for the genetic analysis of bee pollen loads, one that could explain the composition of these pollen loads in terms of simple models of bee foraging and plant spatial genetic structure. In the most basic cases – when bees forage in a single, continous field, showing no preference for certain field regions or genetic types – I found that the genetic differentiation of pollen loads (measured by F_{ST}) is a direct indicator of the typical length of genetic correlations, relative to the typical distance covered by a foraging bee. For more realistic field conditions, F_{ST} results from the combined effect of within-field genetic structure, between-field differentiation and between-field movement.

At the present, there is a lack of experimental data that could be used to test this modelling approach. Although there exist multiple studies of bee pollen composition in the literature, I could not find any that report results for individual pollen loads, rather than the aggregate composition of the whole sample. The recent development of single-pollen DNA amplification methods may facilitate more detailed analyses of bee pollen. Based on my simulation results, there is much information to be gained by looking not just at the average genetic composition of pollen loads, but also at the variation between them.

In this last section, I revisit some of the key assumptions underlying my bee foraging model, and discuss how these assumptions impact the applicability of the model to field situations. I also bring forward some ideas to broaden the scope of the model, particularly along the dimension of time.

3.5.1 Intra-species or inter-species mapping

In principle, genetic markers can be used to distinguish pollen at any taxonomic level. In practice, I believe the method discussed here would be most useful for characterizing the abundance and spatial distribution of closely related species, or ecotypes/varieties within a single species. Species that can be easily differentiated by visual observation probably can be mapped without recourse to molecular methods. Furthermore, it is unlikely that a single parametrization of the model could adequately describe bee foraging in distantly related taxa.

For simplicity, I assumed that all visited plants made the same contribution to the bee's pollen load. There is certainly some stochastic variation in the quantity of pollen collected by flower that is unaccounted for in my simulation results. However, since hundreds of flowers are typically visited in a single foraging bout (Pouvreau 2004; Lotjnant et al. 2012), these fluctuations should be dampened at the pollen load level.

A more fundamental assumption of the model is that bee foraging behavior is independent of the spatial genetic structure, i.e. the relative positions of plants that have or don't have

the marker of interest. While honey bees are considered a generalist pollinator species, studies have shown that individual foragers tend to visit a single species. This behavior, termed *flower constancy*, cannot be explained solely by differences in rewards (pollen/nectar) between species or spatial clustering, although it is strongly linked to differences in flower color or morphology (Hill et al. 1997; Chittka et al. 1999). It is possible, therefore, that bees remain indifferent to a specific genetic marker that differentiates plants within a species or closely-related species, as long as the genotype does not affect flower morphology or rewards. Even in that case, the genotype may indirectly impact foraging behavior if, for example, it affects flowering time. I elaborate on this point later, when I discuss more broadly how to add the temporal dimension to the model.

3.5.2 Incorporating variation in plant density

Although my model treats foraging as a series of discrete steps representing a bee visiting individual plants, each field in the landscape is modelled as a uniform plant cover. The continuum approximation seems appropriate, as it would be unncessarily tedious to simulate spatial coordinates for individual plants; however, even within a continuum representation, it would be possible to incorporate some variation of plant density within a field.

In the single field simulations above, I set the parameter l_s (representing the average distance between successive plants on the foraging path) to unity, which means that all distances are represented in "bee step" units. Therefore, the results of these simulations do not apply only in the case of fields with uniform plant density, but more generally, to any case where the foraging step length scales with the plant spacing in the field. This latter assumption is supported by empirical observations: for example, Levin and Kerster (1969) found that for multiple plant species, the average bee step between successive plants is a linear relationship of the average plant spacing. Using plant spacing as the natural unit of length means that the spatial coordinate system is "stretched" in low density areas and "compressed" in high density areas. This is not a problem as long as all other distances in the model (the distance correlation function $C(s)$, field and patch sizes, etc.) are defined consistenly with that choice of natural units.

3.5.3 Inferring bee foraging behavior in field trials

I developed this model with the goal of inferring an unknown spatial genetic structure using known properties of the bee foraging behavior. The same modelling approach can be used to infer bee foraging behavior from pollen loads collected in a field with known spatial genetic structure. For example, to reproduce one of the simulated designs used here, square patches of some cultivar A could be planted in a larger field of a genetically distinct cultivar B. This type of field design is already being used to study the scale of gene flow in agricultural crops. Given the uncertainty about many aspects of the foraging behavior model, testing and

parametrizing the model in controlled field trials would be an important first step, before it can be applied to more natural systems.

3.5.4 Accounting for time

Throughout this chapter, I have implicitly assumed that parameters of bee foraging behavior and the spatial distribution of plants are fixed properties of the system, ignoring both their seasonal and year-to-year variation.

In describing the effect of foraging behavior and spatial genetic structure on the composition of bee pollen loads, I also did not specify over which time period the pollen load sample should be collected. Is there any sampling period over which the static assumptions of this model would hold? Choosing a short sampling period (e.g. one day) ensures that all pollen loads were collected in the same floral landscape, but may result in a poor spatial coverage, since the honey bee colony only exploits a small portion of the surrounding resources on any given day (Visscher and Seeley 1982). Pollen loads sampled over the whole season or year would be more representative of the floral landscape, but their analysis requires a model that explicitly accounts for phenological variation: the probability that different plants are visited in the same bout would not only depend on their spatial separation, but also on a possible temporal separation between their flowering times.

Studying the level of daily and seasonal fluctuations in bee pollen loads is not only useful to determine the adequate sampling period for a "snapshot" of floral resources, it could also serve to detect trends in pollen data over multiple years by filtering out these fluctuations. Such long-term changes in floral resources may be due for example to population expansion or decline, land-use changes in managed landscapes, or phenological changes driven by climate.

Chapter 4

Invasive species mapping from aerial pollen counts and field observation data: the case of common ragweed in France

As I discussed in Chapter 1, fossil pollen analysis is an essential tool for inferring long-term vegetation changes at the local, regional or continental scales. Yet there are few examples of the use of contemporary pollen records, specifically those from aerial samplers, to better understand the dynamics of extant plant populations.

Allergy risk assessment remains the main application of airborne pollen and spore monitoring. As a result, the list of taxa monitored by aerobiological stations reflect the major allergenic pollen types[1], while modelling efforts focus on the spatial interpolation of pollen concentrations between stations (e.g. Alba et al. 2006; DellaValle et al. 2012) and the short-term forecasting of these concentrations (Aznarte et al. 2007; Efstathiou et al. 2011).

The limited time extent and poor spatial coverage of airborne pollen records, combined with their high level of stochastic noise, contribute to the difficulty of detecting demographic patterns. Even if volumetric pollen samplers were invented in the 1950s (Hirst 1952), the creation and expansion of pollen monitoring networks, such as the European Aeroallergen Network (Nilsson 1988), mostly occured in the last 30 years. Pollen stations are typically located in major cities and separated by tens to hundreds of kilometers.

Despite these limitations, aerobiological samples may be useful in monitoring invasive plants, fungal pathogens and other fast-changing biological populations with airborne propagules. A fixed air sampler effectively pools information about propagule sources in its vicinity.

[1]Typically, pollen concentrations are reported for many tree genera but comparatively fewer herbaceous species. For example, most grasses are lumped into a generic *Poaceae* category.

With the development of high-throughput, automated methods of genetic identification, the analysis of air samples could become a faster, cheaper alternative to field surveys.

In this chapter, I use a hierarchical Bayesian model to describe the density distribution of common ragweed (*Ambrosia artemisiifolia*) in France during the last decade, based on measured airborne pollen concentrations for that period as well as the estimation of the plant's current range from a 2011 survey. This specific case was chosen due to ragweed's status as a major invasive weed in Europe, its known expansion in south-eastern France over the last 30 years, and the public availability of pollen data from 61 stations in metropolitan France. A hierarchical modelling framework is particularly well-suited to the task of relating an underlying space-time process (the dispersal and growth of the ragweed population) to the two different sets of measurements.

After a brief review of the chosen modelling approach and the current state of knowledge on the distribution and spread of *A. artemisiifolia* in Europe, I describe the specific datasets used and perform a preliminary analysis of the space-time variation in the pollen data. Based on the results of this analysis, I formulate a hierarchical model and estimate its parameters in the JAGS software; distribution maps of *A. artemisiifolia* over time are produced from the fitted model. Finally, I use a cross-validation technique to evaluate the model's performance in predicting pollen counts. The results not only demonstrate the potential and limits of existing pollen records for monitoring an invasive plant, but also provide information on how much additional sampling would be required for better predictions.

4.1 Hierarchical Bayesian models and their ecological applications

The hierarchical Bayesian modelling approach in ecology stems from two key advances:

- the use of Bayesian networks to express complex systems in terms of simple conditional probability relationships; and
- the development of Markov chain Monte-Carlo (MCMC) methods, specifically Gibbs sampling, to compute an approximate marginal probability density for any parameter in a Bayesian network.

Bayesian networks are graphs where nodes represent observed or unobserved variables and links represent probabilistic relationships between these variables. They were originally introduced by Judea Pearl (1985) as a tool to represent related hypotheses and study how new information on a specific variable affects the probability (or degree of belief, in a Bayesian interpretation) of other hypotheses in the network. In many applications, these networks are referred to as hierarchical Bayesian models due to their multilevel structure: the parameters

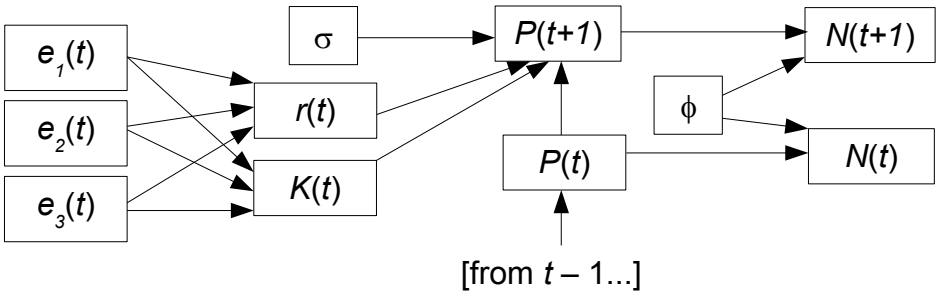

Figure 4.1: Example of a Bayesian network in the context of population dynamics. The observed count N is Poisson distributed with respect to the true population P and a detection probability ϕ. The population follows a logistic growth equation with Gaussian noise of magnitude σ. The logistic parameters r and K depend on time-varying environmental variables e_1 to e_3. Only part of the network is shown, relating the populations at time t and $t+1$.

used to explain data at one level of the model are themselves modelled at the level below as a function of other parameters or measured quantities.[2]

For example, consider the modelling of a population time series by a logistic equation with growth rate r and carrying capacity K: $\Delta P/\Delta t = rP(1-P/K)$. A linear regression of $(\Delta P/\Delta t)/P$ against P could be used to estimate r and K, assuming that the noise term (representing stochastic fluctuations or measurement errors) is normally distributed and independent of P. However, if the number of individuals observed each year is only a small fraction of the true population, and if the parameters r and K depend on a set of measured environmental variables, the single-level regression can be replaced by the following hierarchical model:

- The observed number of individuals at time t, $N(t)$ follows a Poisson distribution, with a mean depending on both the population $P(t)$ and a probability of detection ϕ.
- The population $P(t)$ depends on its previous year's value $P(t-1)$, the growth rate $r(t)$ and the carrying capacity $K(t)$ through the logistic growth equation, with the addition of a random Gaussian fluctuation (mean 0, standard deviation σ).
- The parameters $r(t)$ and $K(t)$ are related to environmental variables (say $e_1(t)$, $e_2(t)$ and $e_3(t)$) through a linear regression.

I illustrate part of the corresponding Bayesian network in Fig. 4.1.

[2]Hierarchical models are not limited to Bayesian networks and also include, for example, multilevel linear regression and generalized linear models (see Gelman and Hill 2007 for a comprehensive overview).

A fundamental property of Bayesian network is that the joint probability distribution of all nodes is a simple product of conditional probabilities (Pearl 1985):

$$\Pr\big(X_1, X_2, \ldots, X_n\big) = \prod_{i=1}^{N} \Pr\big(X_i \mid par(X_i)\big). \tag{4.1}$$

In the preceding equation, $par(X)$ is the set of X's parents. For nodes with no parents (e.g. σ in 4.1), the conditional distribution is replaced with a prior distribution.

Let D denote the set of observed ("data") nodes and H denote the remaining ("hidden") set of nodes; the objective is to estimate H from the data. Using Bayes' theorem, the posterior distribution of H (given the data) is proportional to the product of the likelihood of the data (given H) and the prior distribution of H: $\Pr(H \mid D) \propto \Pr(D \mid H) \Pr(H)$.[3] Although both terms on the right side can be determined from (4.1), finding the proportionality factor involves an integral over all possible joint values of the nodes in H, which cannot be directly computed in most cases. The purpose of Markov chain Monte Carlo (MCMC) methods is to draw a representative sample from the posterior distribution without determining the exact form of that distribution.

The common principle of all MCMC techniques is to generate a random sequence of values for a random variable, where the choice of each value in the sequence depends on the previous one (hence the Markov chain) and on partial knowledge of the underlying probability distribution. The Gibbs sampling algorithm (Geman and Geman 1984; Gelfand and Smith 1990) is particularly useful for Bayesian networks as it relies on the knowledge of conditional probabilities. Let $X_i^{(j)}$ denote the value of node i after the j^{th} iteration of the algorithm. Starting from a set of initial values $\{X_1^{(0)}, X_2^{(0)}, \ldots, X_n^{(0)}\}$, the Gibbs sampler goes through each unobserved node and picks a new value from its full conditional distribution (i.e. given the current values of all other nodes, including the observed ones): $\Pr(X_1^{(1)} \mid \{X_2^{(0)}, \ldots, X_n^{(0)}\})$, $\Pr(X_2^{(1)} \mid \{X_1^{(1)}, X_3^{(0)}, \ldots, X_n^{(0)}\})$ and so forth. With enough iterations of the algorithm, the sequence of values generated for a node will converge to its true posterior distribution.

In a Bayesian network, the full conditional distribution of a node X only depends on its parents $par(X)$, its children $child(X)$ and their parents (Lunn et al. 2000):

$$\Pr\big(X \mid \text{all other nodes}\big) \propto \Pr\big(X \mid par(X)\big) \prod_{Y \in child(X)} \Pr\big(Y \mid par(Y)\big).$$

The proportionality factor in this equation can be determined if the right-hand side corresponds to a known probability density (e.g. it is composed of a single term or the product of two conjugate distributions). Otherwise, it may be evaluated with other MCMC algorithms such as Metropolis-Hastings (Metropolis et al. 1953) or slice sampling (Neal 2003), which only require knowledge of the distribution up to a constant factor.

[3] In this context, the distribution of a set refers to the joint distribution of all nodes in the set.

The estimation of posterior distributions for Bayesian networks using Gibbs sampling has been automated in software packages such as WinBUGS (Lunn et al. 2000), OpenBUGS and JAGS (Plummer 2003). Given the model description and data, these programs automatically select appropriate sampling methods for each node.

In ecology, hierarchical Bayesian models have been particularly useful in the study of spatial and spatio-temporal pattern including biological invasions. Based on data from the North American Breeding Bird Survey over a 35-year period, Wikle (2003) modelled the spread of house finches in the eastern US as a diffusion process (with the diffusion coefficient varying in space) combined with exponential growth and random fluctuations in time. The survey bird counts were assumed to be Poisson distributed with a mean depending on the population value at the specific location and time. Hooten and Wikle (2005) used an extended version of this model to explain the spread of European collared-dove in the same area, by adding a carrying capacity parameter and the effect of human population as a covariate. When dealing with presence/absence information rather than counts, the (hidden) population variable can be linked to the binary observation data using a logistic function, as was done by Zheng and Aukema (2010) to model mountain pine beetle outbreaks in British Columbia.

Other applications of this approach in ecology include the inference of tree fecundity parameters from tree size data and seed samples (Clark et al. 2004); the modelling of species distributions as a function of climate and habitat parameters (Gelfand and al. 2006); and the prediction soil properties at a regional scale from small plot measurements (Kaye et al. 2008).

There are few precedents for the use of hierarchical models to relate pollen counts to vegetation at a large scale. Paciorek and McLachlan (2009) use the pollen composition of 23 sediment cores to infer the historical prevalence of 9 tree species in a 192x192km area of New England. First, they fit a model relating the proportion of each pollen type to the proportion of trees of the corresponding species, using tree census data from two different periods (contemporary and colonial). That model accounts for differences in pollen production between species and includes both short-distance (within the same 12x12km grid square) and long-distance dispersal, the latter determined by a Gaussian spread kernel. The fitted parameters of that model are then used to interpolate the prevalence of each species in the intermediary time periods for which no tree data is available. They assume spatial correlations between tree species in nearby cells, but do not model any temporal process.

The study I present in this chapter differs from that of Paciorek and McLachlan (2009) in several ways. I focus on the aerial pollen data for a single species during a short (10 year) contemporary time period. Rather than assuming a static spatial correlation model at each point in time, I use a dynamic space-time process (diffusion-growth) similar to Wikle (2003). Although I do include some knowledge of the distribution of the plant based on observations in the field, this presence data is less informative than a tree survey using systematic sampling.

4.2 Current knowledge of the distribution of common ragweed in Europe

From its native range in North America, common ragweed was introduced in Europe over a century ago. Using herbarium records, Chauvel et al. (2006) retraced its history in France, concluding that its current distribution is the result of multiple introduction events. This is consistent with the genetic analysis of Genton et al. (2005), which shows that French ragweed populations are descended from several distinct North American populations.

A thorough review and synthesis of the distribution of *Ambrosia artemisiifolia* in the European Union was performed by Bullock et al. (2012). They combined national survey data to produce maps of the known range of common ragweed, up to a 10x10km grid scale. Although their maps do not include information about the density of the plant across its range, a separate map of airborne pollen concentrations suggest higher densities in eastern Europe (Hungary, Ukraine) as well as parts of France and Italy. They also discuss the primary long-distance dispersal vectors for ragweed seeds, both natural (primarily water as the seeds are not adapted for wind transport) and human (agricultural machinery, soil transport, contamination of seed imports).

Recent studies aimed at explaining the spread of ragweed combine species distribution modelling – based on climatic and habitat suitability – and dynamic dispersal on a grid. In the model of Smolik et al. (2010), the probability of ragweed spreading to a new grid cell depends on both its suitability and the distance from the nearest infested cell. They fit the model using the observed distribution of the plant in Austria for 1990 and 2005. At the European scale, Bullock et al. (2012) develop a more complex metapopulation model that includes the probability of extinction from a cell, re-emergence from the seed bank and long-distance dispersal, the latter roughly estimated as being proportional to the volume of agricultural imports from other European countries and the US. For dispersal between neighbor cells, both studies assume that the dispersal distances follow a Gaussian kernel and estimate the standard deviation of that kernel to be around 20 km.

Based on measurements of ragweed pollen away from a small experiment field, Raynor et al. (1970) determined that the integrated concentration of airborne pollen at plant height decreased by 80% in the first 100m from the source. Even if only a small fraction of the total pollen output is transported over long distances, this phenomenon is essential to explain the presence of airborne pollen in regions where *A. artemisiifolia* is absent. Using the particle trajectory simulator HYSPLIT, Zemmer et al. (2012) estimated the geographic distribution of sources for *Ambrosia* pollen sampled in Istanbul during a month; their results suggest that a significant proportion originated in Ukraine. Similarly, Zink et al. (2012) found that up to 20% up the *Ambrosia* pollen collected during a few days in Germany came from Hungary. Building on the distribution maps produced by Bullock et al. (2012), Prank et al. (2013) simulated pollen concentrations for a whole season in Europe. Their results support the idea that long-distance dispersal may lead to rare episodes of high *Ambrosia* pollen counts in

any region, although the total counts for a season remain well correlated with local ragweed presence.

4.3 Description of data sources

4.3.1 Distribution of *Ambrosia artemisiifolia* in France

As part of their report: "Assessing and controlling the spread and the effects of common ragweed in Europe", Bullock et al. (2012) synthesized current knowledge of the distribution of *A. artemisiifolia* in the European Union with maps showing areas of known ragweed presence in a 10x10km grid. Their data for France was collected in a 2011 survey performed by regional botanical conservatories (Petermann 2011). As the report does not describe the sampling methodology and acknowledges differences in sampling efforts between and within regions, these records are a form of presence-only data; in particular, they are not indicative of the relative abundance of ragweed populations at different locations.

Besides *A. artemisiifolia*, two other *Ambrosia* species (*A. psilostachya* and *A. trifida*) have been observed in Europe, primarily in England, Belgium, Netherlands and Northern Germany (Bullock et al. 2012). Since both have been very rarely observed in France, I neglect them in the analysis below and assume that *Ambrosia* airborne pollen counts are representative of the distribution of common ragweed exclusively.

4.3.2 Airborne pollen records

The French Réseau national de surveillance aérobiologique (RNSA) collects and publishes daily pollen concentrations for 21 taxa, as measured by Hirst pollen traps. In this study, I used *Ambrosia* pollen data for a 12-year period (2000-2011) during which the RNSA network grew from 39 to 61 pollen stations.[4]

For each station-year pair, I calculated the annual pollen index (API) – the sum of daily pollen concentrations over the year – for *Ambrosia* and the total API for all pollen types. If each pollen trap was operational for the whole season, the API would be a good indicator of yearly fluctuations and trends; however, many of the time series suffer from missing data. These gaps in the daily pollen record can be a few weeks long and are due to various causes, such as broken samplers being sent for repair or the absence of the sole person charged with operating the sampler (Cassagne 2008).

[4]The RNSA database is published at http://www.pollens.fr.

4.3.3 Selection of model area

Based on both the plant survey and pollen data, I chose to limit the scope of my model to a 520x520km area from central to southeastern France, which encompasses the main concentrations of ragweed along the Rhône and Loire valleys and most pollen stations where significant quantities of *Ambrosia* pollen were sampled (Fig. 4.2). I did not use the records from three stations within that area: Périgueux and Tours, which had a week or more of missing data for every *Ambrosia* season; and Briançon, which perhaps due to its high-altitude location receives very little *Ambrosia* pollen. All of the 29 remaining stations[5] in the selected area had an average API of at least 20 grains/m^3 for *Ambrosia*, compared with only 5 of the 29 stations located outside that area.

Out of the possible 348 station-year pairs (29 stations over 12 years), 290 have daily pollen records. To avoid underestimation of the *Ambrosia* API due to missing data, I removed 33 records that had over 15 missing days in August and September – corresponding to the main ragweed pollen season – and kept a separate list of 51 records with 7 to 14 missing days during the same months to be included for now but possibly excluded later in the analysis. Here I define a missing day as one where no pollen was observed for any taxon, even if it is not strictly impossible for a working sampler to receive zero pollen grains in a day during high pollen season.

4.4 Preliminary analysis

To inform the design of a space-time hierarchical model in the next section, I separately analyze the temporal variation of the API for fixed locations and the spatial variation of the mean API between stations. All calculations in this section were performed in R (R Development Core Team 2008).

4.4.1 Annual pollen index variation by station

Since the API values are strictly positive and their overall distribution is right-skewed, I first model them as being lognormally distributed at each station:

$$\log API_{s,t} = \log \mu_s + \epsilon_{s,t}, \tag{4.2}$$

where $API_{s,t}$ is the measured API for station s on year t, μ_s is a station-specific mean and the $\epsilon_{s,t}$ are normal i.i.d. (independent and identically distributed) residuals. A plot of the fitted residuals against the standard normal quantiles (Fig. 4.3) supports the lognormal

[5] Agen, Aix-en-Provence, Angoulême, Annemasse, Annecy, Aurillac, Avignon, Besançon, Bourg-en-Bresse, Bourges, Castres, Châlon-sur-Saône, Chambéry, Clermont-Ferrand, Dijon, Gap, Grenoble, Lyon, Marseille, Montluçon, Montpellier, Nevers, Nîmes, Poitiers, Roussillon, St-Étienne, Toulon, Toulouse, Vichy.

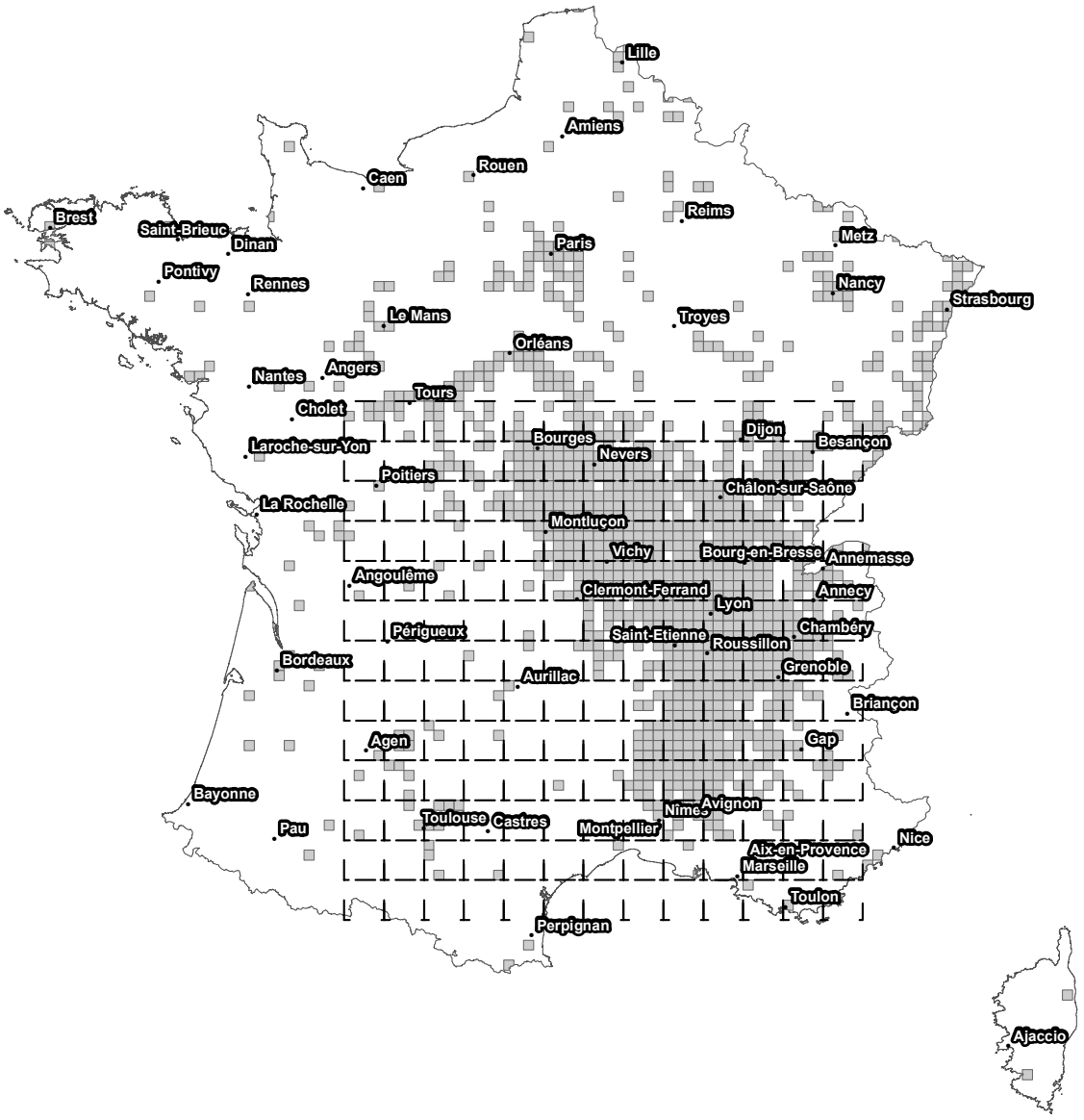

Figure 4.2: Map of France showing the *Ambrosia artemisiifolia* presence data from Bullock et al. (2012) (10x10km grey squares), the point locations of RNSA pollen stations and the 40x40km prediction grid (dashed lines) used in this model. The map was produced in ArcGIS using a Lambert azimuthal equal-area projection.

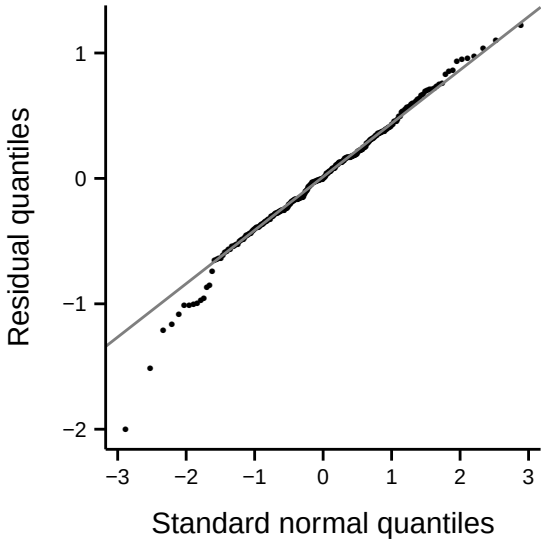

Figure 4.3: Quantile-quantile plot for the residuals of the lognormal regression model (4.2) fitted to *Ambrosia* annual pollen index (API) data.

hypothesis: of the 257 residuals, only the 14 most negative ones lie under their expected values. This discrepancy cannot be explained solely by missing data, as only 5 of the 14 outliers were part of 51 worst records included (those with 7 to 14 missing days during the ragweed pollen season).

The model above assumes that the mean pollen levels are stationary over time. Neither the addition of a global temporal trend (i.e. using the year as a linear predictor) nor the inclusion of station-year interaction terms (representing local trends) significantly improves its explanatory power, as determined by analysis of variance (ANOVA) and Akaike's information criterion (AIC). However, a model incorporating the year as a factor:

$$\log API_{s,t} = \log \mu_s + \alpha_t + \epsilon_{s,t}$$

yields a small but significant ($p = 0.0002$ in ANOVA) increase in explanatory power over (4.2), with the adjusted coefficient of determination (R^2) increasing from 0.85 to 0.87 and the AIC decreasing from 396 to 376. This suggests than in addition to stochastic fluctuations at each station, pollen concentrations might increase or decrease across the study area from year to year; at this point, it is not possible to determine whether this variation stems from changes in the ragweed population, or other factors (e.g. number of rainy days) affecting pollen production and transport.

I also looked for the existence of temporal trends in the API recorded at individual stations. Due to the small number of points in each regression, I determined the significance of the trends using a permutation test, as implemented in the lmPerm R package (Wheeler 2010). The resulting p-value represents the probability that a trend of that magnitude be obtained for a random permutation of the points in the time series. Of the 24 stations with

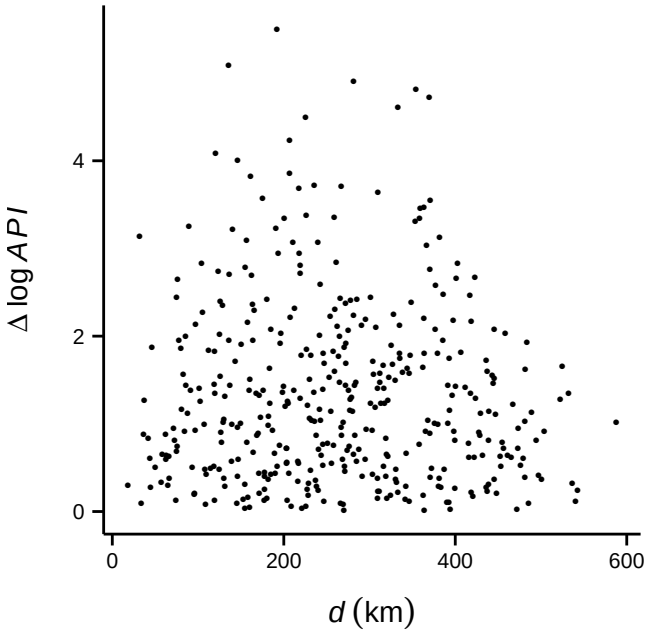

Figure 4.4: Absolute difference in mean log API between pollen stations as a function of their separation distance. Distances are based on the Lambert azimuthal equal-area projected coordinates (Fig. 4.2).

at least 6 yearly pollen records, only 2 had p-values under 0.05 ($p = 0.009$ for Annecy, $p = 0.046$ for Nevers), not a very significant result when accounting for the multiple tests.

These results are not necessarily incompatible with the hypothesis that ragweed populations are growing and spreading in France. It is possible that 12 years of data are insufficient to detect a trend above the stochastic fluctuations in the annual pollen index. Areas where the plant is spreading may also not be adequately sampled by any of the pollen stations.

4.4.2 Spatial variation in mean API

Having shown that the log-transformed API are approximately normally distributed around a station-specific mean, I now focus on the variation of the mean log API in space.

As I discussed in the introduction to this chapter, geostatistical approaches such as kriging have previously been used to interpolate the long-term pollen counts between monitoring stations. Kriging methods model the random field as a Gaussian stationary process, defined by a global mean and a spatial covariance function. In the present exemple, kriging assumptions would imply that stations closer to each other would record more similar pollen counts. However, a plot of the difference in mean log API as function of the distance between pollen stations (Fig. 4.4) does not reveal any such pattern.

If common ragweed was introduced in France through multiple events, as suggested by recent genetic evidence (Genton et al. 2005), and has spread locally except for rare occurrences of long-distance dispersal, then the absence of a large-scale covariance structure in the spatial distribution of the plant is not surprising. For this reason, in my hierarchical model for *Ambrosia* pollen counts presented in the next section, the underlying plant density does not follow a spatial covariance function. Instead, the spread of plants to nearby cells (local diffusion) creates short-scale spatial correlations over time.

4.5 Hierarchical model specification

The purpose of this modelling effort is to represent the distribution and spread of *Ambrosia artemisiifolia* in central and south-east France based on pollen counts and presence observations. To this end, I divide the 520x520km model area in a 13x13 prediction grid of 40x40km cells (Fig. 4.2). I chose this grid scale based on both the spatial resolution of the pollen dataset – with the exception of Aix-en-Provence and Marseille, all pairs of stations are separated by over 30km – and computation time limitations. To limit the effect of boundary conditions on predicted densities in the area of interest, I add a 40km (1 cell) buffer on each side to create a 15x15 grid.

In this section, I describe the four main components of my hierarchical Bayesian model:

- a dynamic space-time model to describe how the density of *A. artemisiifolia* plants in each cell evolves from year to year;
- a model to relate the *Ambrosia* API measured at each station to the density of plants in nearby cells;
- a model to relate the presence data (in 10x10km cells) from the 2011 survey to the plant density (in 40x40km cells); and
- prior distributions for the model parameters and the initial plant density values.

I use the following conventional notation to describe probability relationships: $X \mid Y \sim \mathcal{N}(\mu(Y), \sigma^2)$ signifies that conditional on the value of Y, X follows a normal distribution with a mean of μ (which depends on Y) and a variance of σ^2. A uniform distribution over the interval (a, b) is noted as $\mathcal{U}(a, b)$, whereas a binomial distribution with n trials and probability of success p is noted as $\mathcal{B}(n, p)$

4.5.1 Plant population dynamics

Let $\rho_{t,i,j}$ denote the relative density[6] of ragweed on year t for the grid cell in column i, row j ($i, j \in \{1, \ldots, 15\}$). Following Wikle (2003), I model the time evolution of ρ using a diffusion-growth process:

$$\frac{\partial \rho}{\partial t} = \delta \left(\frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2} \right) + \alpha \rho, \tag{4.3}$$

where α is the growth rate (which may be negative) and δ is the diffusion coefficient. The discretization of eq. (4.3) using a 1 year temporal step and a 1 cell spatial step results in:

$$\rho_{t,i,j} = (1 - 4\delta + \alpha)\, \rho_{t-1,i,j} + \delta \left(\rho_{t-1,i-1,j} + \rho_{t-1,i+1,j} + \rho_{t-1,i,j-1} + \rho_{t-1,i,j+1} \right). \tag{4.4}$$

For cells along the boundary, (4.4) is modified so that diffusion does not occur across grid edges. For example:

$$\begin{aligned} \rho_{t,1,1} &= (1 - 2\delta + \alpha)\, \rho_{t-1,1,1} + \delta \left(\rho_{t-1,2,1} + \rho_{t-1,1,2} \right); \quad \text{and} \\ \rho_{t,1,j} &= (1 - 3\delta + \alpha)\, \rho_{t-1,1,j} + \delta \left(\rho_{t-1,2,j} + \rho_{t-1,1,j-1} + \rho_{t-1,1,j+1} \right). \end{aligned}$$

This condition is equivalent to assuming that the ragweed density is the same on both sides of the boundary (i.e. its derivative is zero). One alternative, assuming that the density itself is zero outside the boundary, would be in contradiction with the presence data.[7]

This simple diffusion-growth model is not meant to accurately represent the mechanisms underlying the spread of *A. artemisiifolia*. For example, the actual growth and diffusion rates should vary within the study area due to differences in habitats, potential to spread via roads and rivers, etc. Although I will consider variations of the model such as a time-varying α, the relatively coarse scale of the pollen data limits the number of parameters than can be added without overfitting.

4.5.2 From plant density to pollen index

Based on the preliminary analysis results (section 4.4 above), I choose a lognormal distribution for the annual pollen index:

$$\log API_{s,t} \mid \mu_{s,t}, \sigma_p \sim \mathcal{N}(\log \mu_{s,t}, {\sigma_p}^2). \tag{4.5}$$

The variance σ_p is a global parameter in the model, while the mean $\mu_{s,t}$ depends on the relative plant density around station s on year t.

[6] Note that expressing the model in terms of absolute instead of relative densities would require the independent determination of a proportionality constant between plant density and pollen production.

[7] Although not shown on Fig. 4.2, the known range of ragweed extends across the eastern border into Switzerland (Bullock et al. 2012).

Since most of the *Ambrosia* pollen comes from local sources, I model $\mu_{s,t}$ as a weighted mean of the density for the grid cell where station s is located (i_s,j_s) and its 8 nearest neighbors:

$$\mu_{s,t} = \sum_{i=i_s-1}^{i_s+1} \sum_{j=j_s-1}^{j_s+1} w_{s,i,j}\rho_{t,i,j}. \tag{4.6}$$

The $w_{s,i,j}$ are calculated from a dispersal kernel k evaluated at the distance between the location of station s and the center of cell (i,j):

$$w_{s,i,j} = \frac{k(x_s - x_{i,j}, y_s - y_{i,j})}{\sum_{i=i_s-1}^{i_s+1} \sum_{j=j_s-1}^{j_s+1} k(x_s - x_{i,j}, y_s - y_{i,j})},$$

where the denominator ensures the sum of weights is 1. I will use two different forms for k, a Gaussian kernel:

$$k(x, y) = \exp\left(-\frac{x^2 + y^2}{2a^2}\right)$$

and an exponential one:

$$k(x, y) = \exp\left(-\frac{\sqrt{x^2 + y^2}}{a}\right).$$

Both forms are specified by a single parameter a, which represents the scale of dispersal and shares the units of x, y (in this case, kilometers). Estimating a along with the other parameters in the full hierarchical model would greatly increase the volume of calculations at each iteration of the MCMC algorithm. For this reason, I pre-compute the weights $w_{s,i,j}$ for a given a and include these weights as constants in the hierarchical model. I then estimate the parameter by performing separate model runs for different values of a and comparing their fit using the deviance information criterion (DIC).

4.5.3 Incorporating presence data

Pollen counts alone may not be sufficient to infer plant density across the model area: there are much less pollen stations (29) than cells in the prediction grid (169), and parts of the grid – such as most of the Massif Central around Aurillac – are poorly covered by the monitoring network. To use the additional information provided by the common ragweed presence data, I relate this binary dataset to the underlying plant density through a logistic regression model.

Since the presence data is defined on a 10x10km grid, each prediction grid cell contains are $n_{sub} = 16$ subcells in which ragweed was either observed or not observed. I assume that probability $\phi_{i,j}$ of detecting the plant in a subcell of cell (i, j) is a logistic function of the log plant density in that cell, for the year when the survey was realized (here $t_{obs} = 12$,

corresponding to 2011):

$$\phi_{i,j} = \frac{1}{1 + \exp\big(-b(\log \rho_{t_{obs},i,j} - c)\big)}. \tag{4.7}$$

The parameter c is the midpoint of the logistic curve (corresponding to $\phi = 0.5$) while b determines the steepness of the curve around its midpoint.

Under this assumption, the number of subcells in cell (i, j) where ragweed was observed $(n_{i,j})$ follows a binomial distribution:

$$n_{i,j} \mid \phi_{i,j} \sim \mathcal{B}(n_{sub}, \phi_{i,j}). \tag{4.8}$$

Although the logistic model does not account for the variation in plant density or sampling effort between subcells, it might still constitute the best option without any additional information on that variation. Furthermore, if the presence data is a poorer predictor of plant density than the pollen data, then the latter will automatically have more influence on posterior estimates of plant density in the hierarchical model.

4.5.4 Prior distributions

To complete the hierarchical model specification, I define uniform prior distributions for the five global parameters and for the logarithm of the plant density at $t = 1$ (the year 2000): $\sigma_p \sim \mathcal{U}(0, 2)$; $\alpha \sim \mathcal{U}(-0.2, 0.2)$; $\delta \sim \mathcal{U}(0, 0.2)$; $b \sim \mathcal{U}(0, 10)$; $c \sim \mathcal{U}(0, 10)$; and $\log \rho_{1,i,j} \sim \mathcal{U}(-5, 10)$.

I set a broad range for each prior, based on physical restrictions and the results of the preliminary analysis of the dataset. In particular, α and δ are limited to 0.2 in magnitude to ensure that the term $(1 + \alpha - 4\delta)$ in eq. (4.4) is non-negative. The priors for $\log \rho_{1,i,j}$ and c are wider than the range of observed log API values (1.5 to 9). The standard deviation of the lognormal API distribution (σ_p) should be close to 0.5, based on the quantile-quantile plot in Fig. 4.3.

These priors can be reviewed following initial model runs, for example if the posterior distribution of any parameter approaches the limits of the prior, indicating that the hypothesized range may be too narrow.

The full model is summarized by a network diagram in Fig. 4.5.

4.5.5 Model variations

To determine how different components of the hierarchical model affect both the parameter estimates and model fit, I introduce a few variants of the full model:

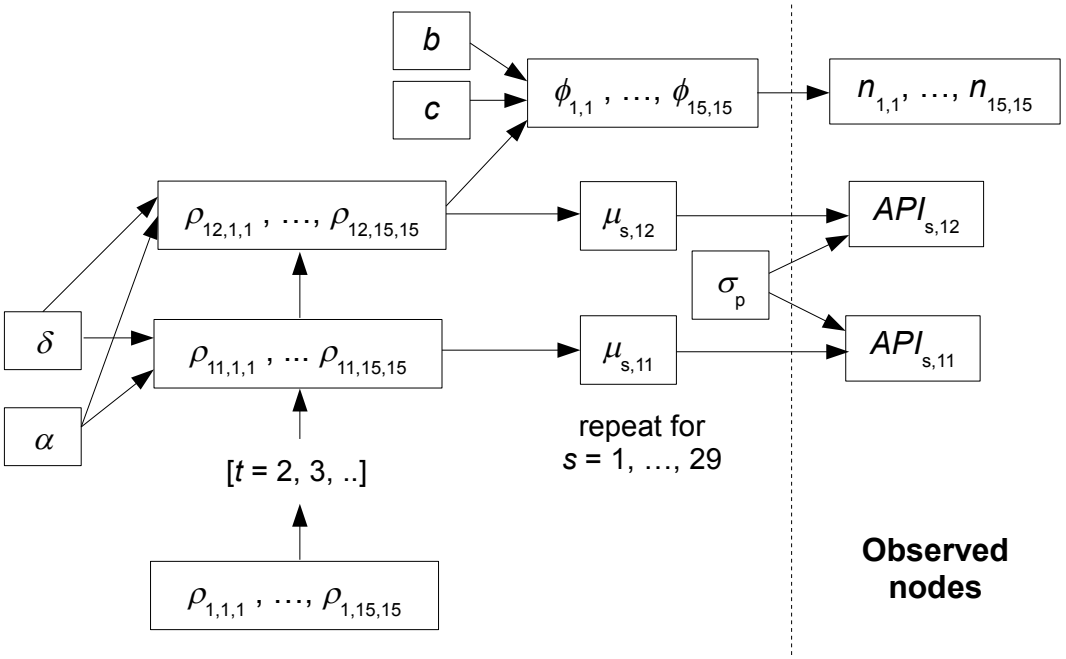

Figure 4.5: Bayesian network for the ragweed dispersal model. The density of the plant in each grid cell ($\rho_{t,i,j}$) evolves according to a diffusion-growth model with parameters α and δ. The measured annual pollen index (API) at each station is distributed lognormally, with a mean $\mu_{s,t}$ depending on the value of ρ in nearby cells. The gridded presence data ($n_{i,j}$) for the year 2011 ($t = 12$) is related to ρ through a logistic regression with parameters b and c.

- In the static model, the plant density ρ is constant in time so that eq. (4.4) is replaced with $\rho_{t,i,j} = \rho_{t-1,i,j}$. By contrast, dynamic versions of the model have a time-varying ρ.
- Instead of a constant growth rate α, I consider a version with no growth (α set at 0) and another one where α varies by year. This is motivated by the the preliminary analysis of *Ambrosia* pollen concentrations, which did not show evidence for a constant increase or decrease over time.
- The "pollen only" model ignores the presence data to infer ρ from the pollen dataset alone.

4.6 Model fitting and validation methods

I estimated posterior distributions from the model using JAGS (Just Another Gibbs Sampler) and the rjags package in R (Plummer 2003). Each model run included two MCMC chains initialized using different values sampled from the parameters' prior distributions. After a number of burn-in steps performed in adaptive mode (a function in JAGS that optimizes the sampling algorithm), I verified convergence of the two chains using the diagnostic test of Gelman and Rubin (1992). For each parameter to be estimated, the test compares its variance within each MCMC chain (W) to the variance between the chains' mean estimates (B) and outputs a potential scale reduction factor. If this factor is close to 1, then B is small relative to W, indicating that the multiple chains have converged to a common posterior distribution.

To compare the fit of different model specifications, JAGS can calculate their deviance information criterion (DIC) from the output of multiple MCMC chains. The DIC was proposed by Spiegelhalter et al. (2002) as a means to estimate the predictive power of hierarchical models. It is defined as the sum of the mean deviance ($\overline{D}$, a quantity proportional to the negative log-likelihood of the data) and the effective number of parameters (p_D). While a lower $\overline{D}$ indicates that the model maximizes the likelihood of the observed data, p_D serves to penalize more complex models that may overfit the specific data observed and thus lose some ability to predict new data.

After selecting a best model based on DIC estimates, I use a cross-validation technique to evaluate how well that model can predict pollen concentrations between monitoring stations. For each of the 29 stations, I run the model without the data from that station and compare the resulting estimate of log μ_s (the mean log API) to its estimate for the full dataset, including station s. I also compare the cross-validation estimates of my model to those obtained by ordinary kriging (spatial interpolation) of the observed log API, as performed by the Geostatistical Analyst extension in ArcGIS.

4.7 Results

Unless stated otherwise, all posterior estimates and DIC values reported here are based on 10,000 iterations of two MCMC chains. I started each model run with 10,000 burn-in adaptive steps; this was sufficient to reach a Gelman-Rubin potential scale reduction factor of under 1.01 for all parameters, indicating good convergence. The resulting posterior parameter distributions also seem close to normal, although the distribution of δ is slightly asymmetric (Fig. 4.6).

These posterior parameter distributions are not completely independent. Specifically, I found a moderate positive correlation ($R^2 \approx 0.3$) between δ, the diffusion coefficient and b, the steepness of the logistic curve describing the probability of detecting ragweed as a function of plant density. Here is one potential explanation for this correlation, as well as the observation that b estimates are lower for the static model (Table 4.1) in which δ is effectively zero. The effect of diffusion is to equalize plant densities across the grid: with a narrower range of plant densities, explaining the same range of presence/absence probabilities requires a logistic curve with a steeper slope.

As I mentioned at the model specification stage, the mean API at each station is modelled as a linear combination of the plant densities in neighboring cells (eq. 4.6), where the weights are pre-computed based on some dispersal kernel. To choose a type of kernel (Gaussian or exponential) and an appropriate value of the dispersal scale parameter a, I perform separate MCMC simulations for each model and compare their DIC. I also compare the fit of my dynamic model to a static model where the plant density is constant in time (Table 4.1). For the static models, the best pollen dispersal kernel (lowest DIC) is Gaussian with a=10km. The DIC for dynamic models depends less on the value of a, but is significantly lower than under the static case due to the difference in p_D.[8] Based on these results, I use a Gaussian kernel with a = 10km from this point on.

None of the models in Table 4.1 include a growth rate α. When this parameter is added to the dynamic model, its posterior estimate is slightly negative and includes 0 (mean -0.04, s.d. 0.02). When α is allowed to vary from year to year, none of the estimates were significantly different from zero (means from -0.12 to 0.11, s.d. from 0.07 to 0.1). Furthermore, adding α does not significantly alter the other parameter estimates or lower the DIC: therefore, I consider it to be superfluous for this particular model.

Separately, I verified that removing the 51 records with between 7 and 14 missing days in the *Ambrosia* pollen season did not improve the model fit; therefore, I included these records in all results presented here.

The units of the diffusion coefficient δ are (grid cells)^2/year. To compare its value to the spread parameters in Bullock et al. (2012) and Smolik et al. (2010), I use the well-known

[8] I believe dynamic models have less free parameters because the plant densities in different cells are linked over time through the diffusion process.

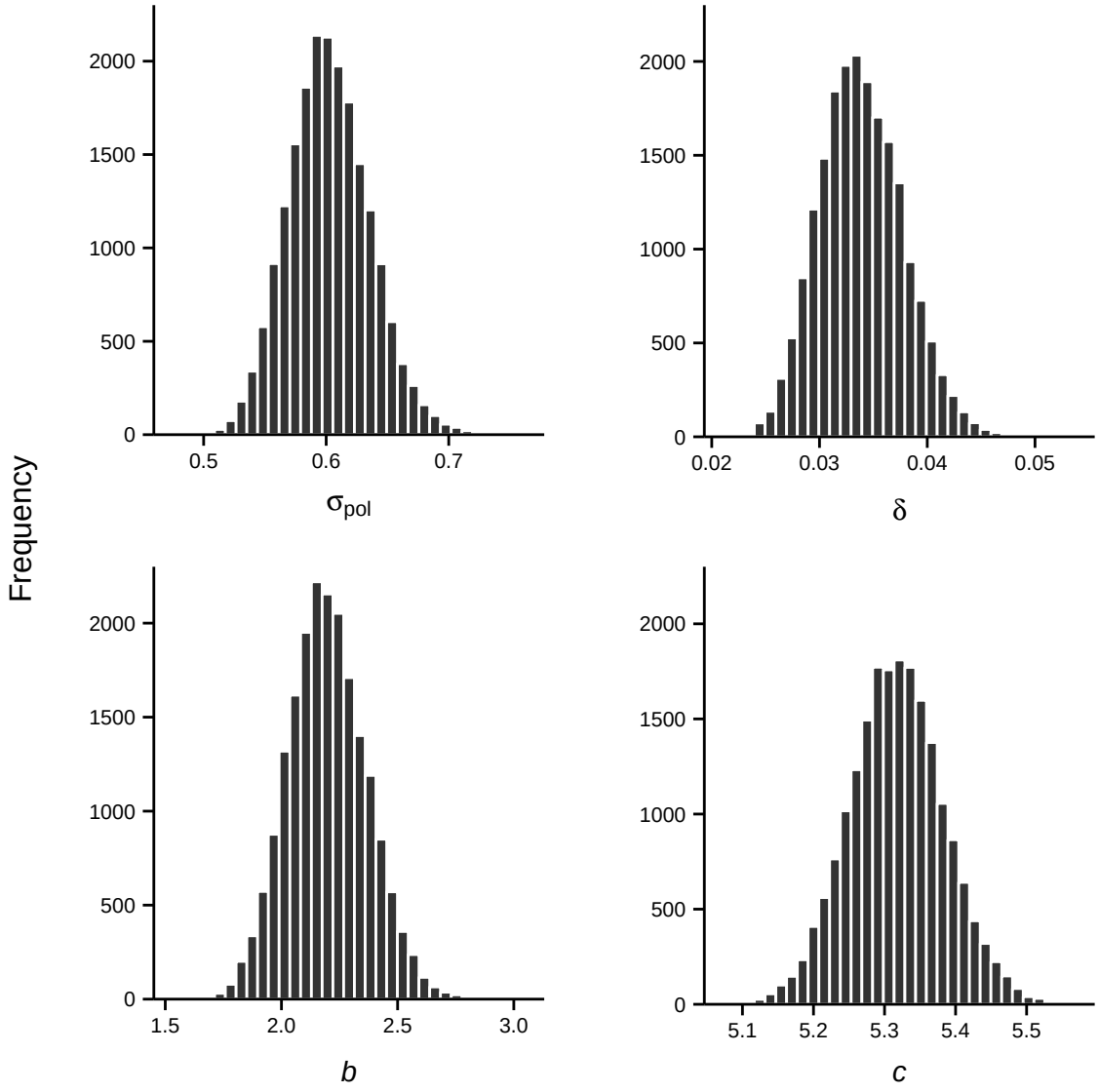

Figure 4.6: Posterior distributions for the parameters of the ragweed dispersal model, using a Gaussian kernel with $a = 10$km for pollen dispersal. Each distribution is based on 20,000 total iterations (10,000 each of two parallel MCMC chains).

Model		Posterior means (s.d.)				Relative fit		
S/D	Kernel	σ_p	δ	b	c	$\overline{D}$	p_D	DIC
S	G-5km	0.51 (0.03)		0.85 (0.07)	5.2 (0.1)	1112	180	1292
S	G-10km	0.52 (0.03)		0.85 (0.07)	5.0 (0.1)	1108	174	1282
S	G-20km	0.56 (0.03)		0.86 (0.08)	4.7 (0.1)	1195	204	1399
S	E-5km	0.53 (0.03)		0.86 (0.07)	4.9 (0.1)	1149	163	1312
S	E-10km	0.54 (0.03)		0.84 (0.07)	4.7 (0.1)	1184	176	1360
D	G-5km	0.59 (0.03)	0.033 (0.004)	2.1 (0.2)	5.34 (0.07)	1121	104	1225
D	G-10km	0.60 (0.03)	0.034 (0.004)	2.2 (0.2)	5.32 (0.07)	1125	106	1231
D	G-20km	0.65 (0.03)	0.042 (0.004)	2.2 (0.2)	5.27 (0.06)	1147	86	1233

Table 4.1: Comparison of posterior parameter estimates and relative fit – as measured by the deviance information criterion (DIC) – for static (S) and dynamic (D) models of the ragweed distribution with different pollen dispersal kernels (e.g. G-10km is a Gaussian kernel with a=10km; E denotes an exponential kernel). Results are calculated from 10,000 iterations of two parallel MCMC chains.

result that diffusion from a point source for a time t causes a Gaussian spread with a standard deviation $\sigma = \sqrt{2\delta t}$. For a t of one year, the estimated range of 0.25–0.45 for δ (Fig. 4.6) corresponds to a σ of 0.22–0.3 cells, or 9–12 km.

The next set of figures show the posterior means and standard deviations of log ρ on the 13x13 prediction grid, for both the static model (Fig. 4.7) and four years of the dynamic model (Figs. 4.8 – 4.9). While both models use the presence data to supplement the measured pollen concentrations, the static model cannot really interpolate ρ in areas with poor pollen station coverage, as indicated by the very low mean estimates of ρ and their high standard deviations. The estimates for the first year (2000) of the dynamic model show similar problems: this may be due to the lack of data at the time (only 11 out of 29 pollen stations reporting), which when combined with a broad uniform prior may lead to a very dispersed posterior distribution.

For a "pollen only" model (not shown here), the dynamic estimates of ρ in 2011 do not significantly differ from the static estimates. The posterior estimate of δ in that case also approaches zero (mean 0.003, s.d. 0.001). This suggests that the smoothing of the density field observed in Fig. 4.8 requires both a dynamic diffusion model and the incorporation of presence data.

In the cross-validation test, I removed one station at a time from the model and recorded the mean log API (log μ) predicted for 2011 at that station. Averaging over the 29 stations, I obtained a mean error (difference between the predicted log μ and the observed log API in 2011) of -0.007 and a root mean square (RMS) error of 0.86; the latter corresponds to an over- or underestimation of the observed API (not log) by a factor of 2.4, which may seem high. However, this value should be compared to the year-to-year fluctuations in pollen index that occur even if we know the station mean: the scale of this fluctuations corresponds to σ_p in my model and was estimated at around 0.6 (or a factor of 1.8). As a further comparison, I

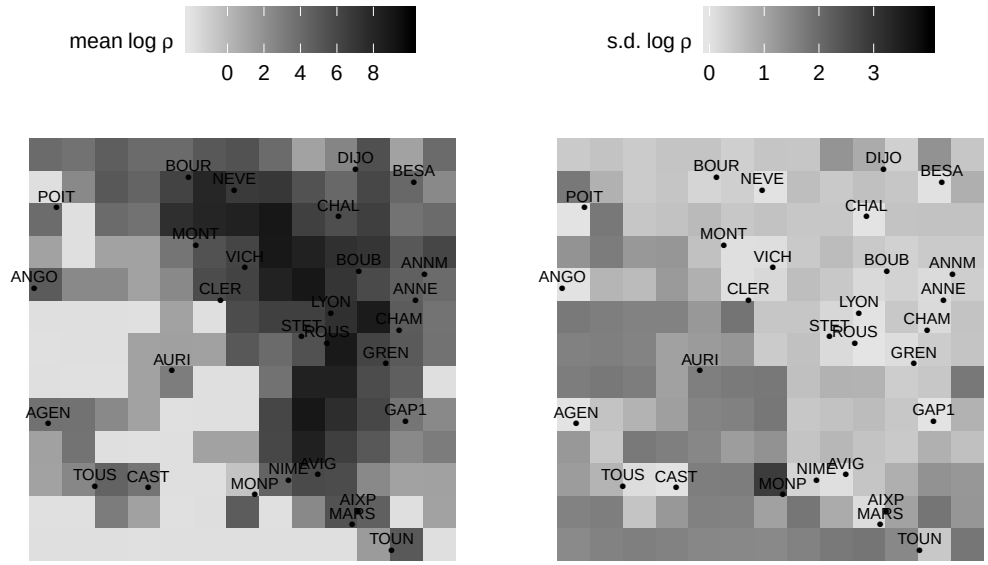

Figure 4.7: Posterior distribution (mean and standard deviation) of log ρ over the 13x13 prediction grid, under a static model with a Gaussian kernel ($a = 10$km) for pollen dispersal. The location and abbreviated name of each pollen station are indicated. Results are calculated from 10,000 iterations of two parallel MCMC chains.

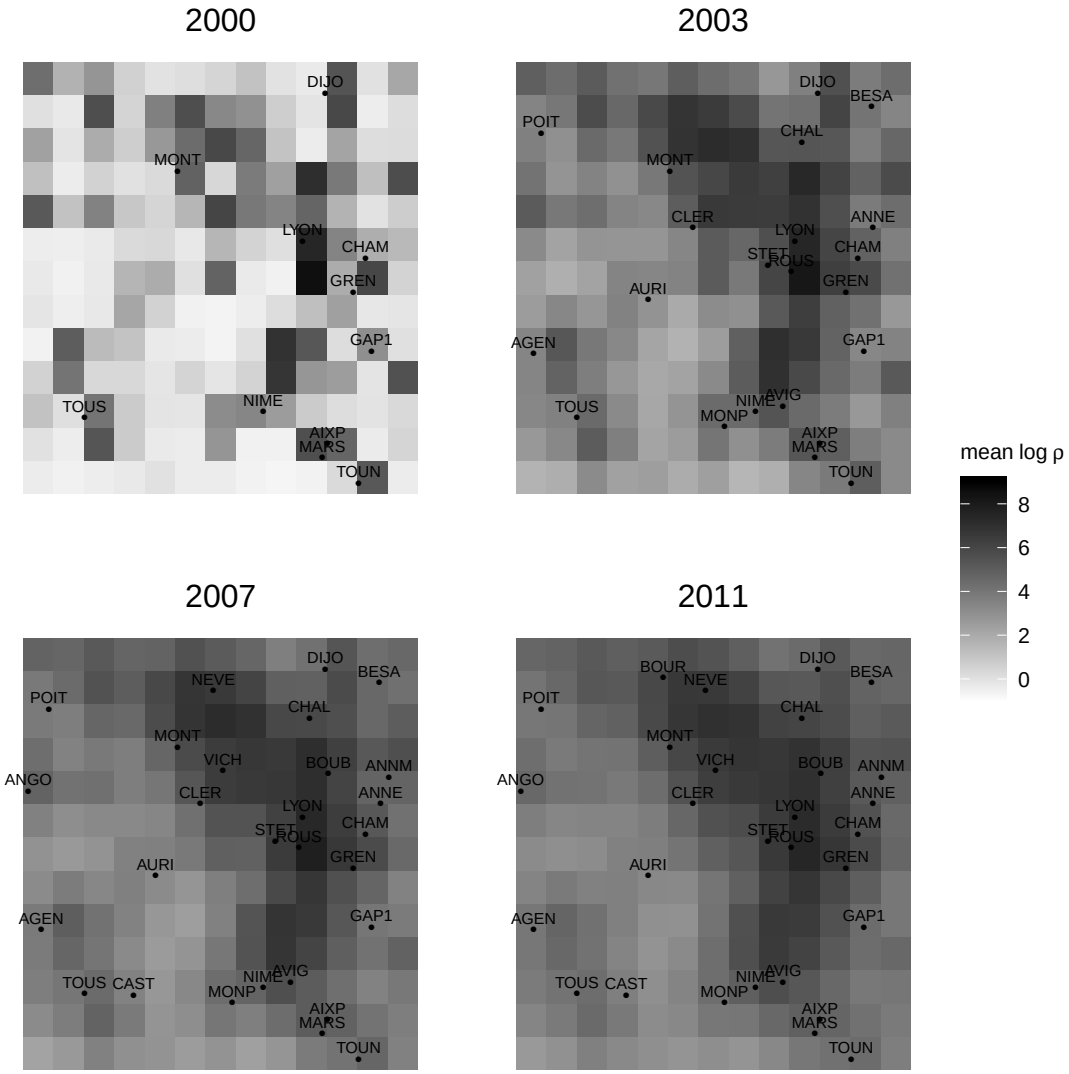

Figure 4.8: Posterior mean of log ρ over the 13x13 prediction grid at four different times, under a dynamic model with a Gaussian kernel ($a = 10$km) for pollen dispersal. The locations indicated on each map represent the pollen stations in operation during that year. Results are calculated from 10,000 iterations of two parallel MCMC chains.

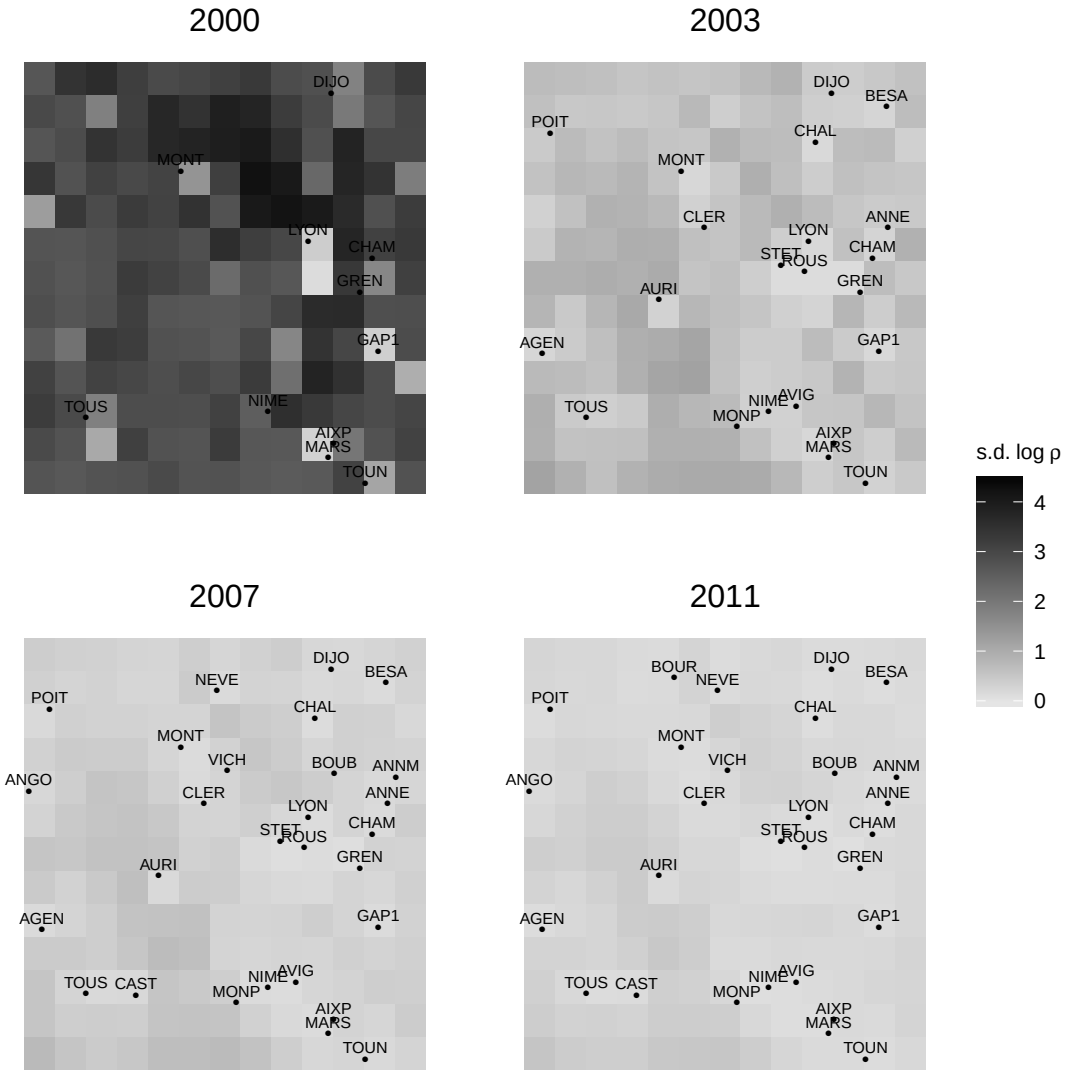

Figure 4.9: Posterior standard deviation of log ρ over the 13x13 prediction grid at four different times, under a dynamic model with a Gaussian kernel ($a = 10$km) for pollen dispersal. The locations indicated on each map represent the pollen stations in operation during that year. Results are calculated from 10,000 iterations of two parallel MCMC chains.

also fit an ordinary kriging model to the 2011 log API values in ArcGIS. Its cross-validation mean error (0.095) and RMS error (1.038) were both higher than those calculated under my model.

4.8 Discussion

In this chapter, I implemented a hierarchical modelling approach to create maps of the relative density of a known invasive weed, by combining records from sparse pollen stations and binary presence data from field surveys.

The parameter estimates for the dynamic model do not indicate significant growth or decline of the ragweed population in France over the last decade, but support some degree of diffusion – movement from the most densely populated areas to the least densely populated areas. The estimated range of δ in my model corresponds to a Gaussian dispersal kernel with a standard deviation of 9–12 km, which is similar to the 13 km estimate obtained by Smolik et al. (2010) when fitting a simple diffusion of ragweed presence in Austria.

Is it unlikely that diffusion occurs uniformly across space, so my global estimate of δ probably underestimates the real diffusion coefficient by "averaging in" regions where no diffusion occurs. When combining a diffusion process with a species distribution model describing the suitability of different habitats to ragweed, both Smolik et al. (2010) and Bullock et al. (2012) obtained higher standard deviations (20–25km) for the dispersal kernel. However, species distribution modelling also shows that the potential distribution of ragweed in Europe extends far past its current range (Bullock et al. 2012; Chapman et al. 2013); this would suggest that the spread of the plant is limited by dispersal opportunities rather than climatic constraints or lack of suitable habitats. Future modelling efforts could focus on varying the diffusion coefficient in space based on known vectors of dispersal such as rivers and major transportation networks for agricultural products.

Besides diffusion, other processes could explain the equalization of ragweed density across its range. For example, there may be more control measures taken to remove the weed in the most severely affected areas.

Part of the observed spatial smoothing of plant density over time (Fig. 4.8) may also be an artifact of the model specification. Due to the low spatial resolution of the pollen data, the presence information is necessary to produce a smooth density estimate, but it only available for 2011. Therefore, its influence will decrease the further the model goes back in time. Furthermore, the lesser amount of pollen data in 2000 and the uniform prior distribution on ρ contribute to the irregularity of the estimated density field for that year. This effect would be reduced if the plant presence data were available for a second year, preferably close to the beginning of the time series.

To integrate the binary presence data from a 2011 survey into my model, I assumed that

the probability of detection depended only on the ragweed density in each area, without accounting for the fact that not all areas were equally well surveyed. This unequal sampling effort may explain why, for example, the presence data suggests a concentration of ragweed around Paris (Fig. 4.2) even though the annual pollen index recorded there, as in the rest of northern France, was quite low. Despite these shortcomings of the ragweed presence records, the additional information they provide compensates for the large distance between pollen stations, and cross-validation results show that my model can predict 2011 pollen counts better than a spatial interpolation of the pollen data.

In theory, the hierarchical modelling approach could be used to fit the pollen monitoring data to a complex model of plant population dynamics and dispersal. In the case studied here, the spatial resolution of the data is rather coarse compared with the scale at which the relevant biological and physical processes occur. With a denser network of monitoring stations, it could be feasible to infer the distribution of an invasive plant from pollen records alone. Even in that case, micro-climatic effects related to the specific placement of pollen stations (sampler height, surrounding topography, urban/rural areas, etc.) need to be examined to avoid local biases.

Chapter 5

Conclusion

A broad focus of this dissertation is the use of pollen analysis to infer characteristics of plant populations. My research was motivated by the ongoing development of high-throughput methods for the genetic identification of pollen grains; I chose to discuss not only the possibilities offered by existing data and methods, but also those that could result from an increase in the volume and phylogenetic resolution of pollen data. As I discuss below, another objective of this work was to apply new data-model integration techniques to the study of pollen data.

In Chapter 2, I described a dilution-PCR assay to estimate the proportion of a pollen sample bearing a specific genetic marker. The assay design aims to minimize the error in the final estimate – including both sampling error and DNA amplification failure – for a given number of PCR reactions. If the cost of each PCR determination is constant, this method is more efficient that testing single-pollen grains. However, as single-pollen PCR methods evolve and both their error rate and the cost per reaction decrease, the same statistical methods I used can help identify the point at which single-pollen approaches become more efficient than dilution assays.

The bee foraging model presented in Chapter 3 constitutes an example of the additional information that could be gathered from a more fine-grained analysis of pollen samples. While most studies of bee pollen composition focus on the aggregate representation of pollen types at the hive level, I have shown that the genetic differentiation between pollen loads provides information about the scale of the spatial genetic structure of floral resources, relative to the length of bee foraging paths. To test and parametrize this model, I would need to analyze pollen loads from bees foraging in a field with known spatial genetic structure.

The capacity to perform high-throughput pollen analysis would provide even more benefits at the biogeographic scale. One of the main challenges to the inference of vegetation based on pollen data is the poor spatial resolution of existing sampling networks. The *Ambrosia* pollen data I used in Chapter 4 represents a best case scenario, since *Ambrosia* is

one of a few widely monitored allergenic pollen taxa and the 61 pollen stations in France provide a comparatively good spatial coverage.[1] Yet, these records cannot capture biological processes occurring at the scale of less than a few kilometers. High-throughput genetic methods would allow researchers to monitor the spread of a greater range of invasive plant species; the same methods applied to fungal spores could be used to monitor plant diseases.

In Chapter 1, I discussed the range of models for pollen dispersal and noted a divide between empirical approaches, based on fitting parametric functions to observed dispersal curves, and mechanistic approaches, based on modelling the postulated physical and biological processes affecting dispersal. Recent advances in ecological modelling aim to bridge this gap by relating multi-layered, process-based models to dispersal data.

Individual-based models simulate individual dispersal events from a population and calculate summary statistics over multiple simulations to compare with the observed dispersal patterns. Gilioli et al. (2013) used this approach when fitting a model of the spread of chestnut gall wasp in Europe, with the summary statistic being the area colonized by the wasp over time. The individual-based model I developed in Chapter 3 predicts the genetic differentiation between bee pollen loads from individual simulations of foraging paths.

Since the work of Wikle (2003), hierarchical Bayesian models have become increasingly popular in dispersal ecology, as they allow the direct estimation of multiple parameters in a multi-layered model combining different sources of data. In Chapter 4, I studied the application of this type of model to infer the spread of an invasive weed from pollen data. For models describing dynamic space-time processes, an alternative to the hierarchical Bayesian approach is to use data assimilation methods such as the Kalman filter (Cressie and Wikle 2002). Although commonly used in hydrology and atmospheric science, data assimilation techniques are still rarely applied to ecological problems.

[1] For example, the American National Allergy Bureau has 85 stations in the contiguous United States, an area over 14 times the size of metropolitan France.

Bibliography

Abdelbasit, K.M. and R.L. Plackett. 1983. "Experimental Design for Binary Data." *Journal of the American Statistical Association* 78: 90-98.

Agashe, S.N. and E. Caulton. 2009. *Pollen and Spores: Applications with Special Emphasis on Aerobiology and Allergy*. Enfield, NH: Science Publishers.

Alba, F., D. Nieto-Lugilde, P. Comtois, C.D. de la Guardia, C. de Linares and L. Ruiz. 2006. "Airborne-pollen map for *Olea europaea* L. in eastern Andalusia (Spain) using GIS: estimation models." *Aerobiologia* 22: 107-116.

Angevin, F., E.K. Klein, C. Choimet, A. Gauffreteau, C. Lavigne, A. Messean and J.M. Meynard. 2008. "Modelling Impacts of Cropping Systems and Climate on Maize Cross-Pollination in Agricultural Landscapes: The MAPOD Model." *European Journal of Agronomy* 28: 471-484.

Arritt, R.W., C.A. Clark, A.S. Goggi, H. Lopez Sanchez, M.E. Westgate and J.M. Riese. 2007. "Lagrangian Numerical Simulations of Canopy Air Flow Effects on Maize Pollen Dispersal." *Field Crops Research* 102: 151-162.

Atkinson, A.C. and A.N. Donev. 1992. *Optimum Experimental Designs*. Oxford: Clarendon Press.

Aylor, D.E. and T.K. Flesch. 2001. "Estimating Spore Release Rates Using a Lagrangian Stochastic Simulation Model." *Journal of Applied Meteorology* 40: 1196-1208.

Aylor, D.E., N.P. Schultes and E.J. Shields. 2003. "An Aerobiological Framework for Assessing Cross-Pollination in Maize." *Agricultural and Forest Meteorology* 119: 111-129.

Aznarte, J.L., J.M. Benítez Sánchez, D.N. Lugilde, C. de Linares Fernández, C.D. de la Guardia and F.A. Sánchez. 2007. "Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models." *Expert Systems with Applications* 32: 1218-1225.

Bartumeus, F., M.G.E. de la Luz, G.M. Viswanathan and J. Catalan. 2005. "Animal Search Strategies: A Quantitative Random-Walk Analysis." *Ecology* 86: 3078-3087.

Beckie, H.J. and L.M. Hall. 2008. "Simple to Complex: Modelling Crop Pollen-Mediated Gene Flow." *Plant Science* 175: 615-628.

Beil, M., H. Horn and A. Schwabe. 2008. "Analysis of Pollen Loads in a Wild Bee Community (Hymenoptera: Apidae) - a Method for Elucidating Habitat Use and Foraging Distances." *Apidologie* 39: 456-467.

Bektas, A. and I. Chapela. 2013. "Loop-Mediated Isothermal Amplification of Single Pollen Grains." Manuscript submitted for publication.

Bullock, J.M. et al. 2012. "Assessing and Controlling the Spread and the Effects of Common Ragweed in Europe". Final Report of Project ENV.B2/ETU/2010/0037. Wallingford, UK: National Environment Research Council.

Bunting, M.J. and D. Middleton. 2005. "Modelling Pollen Dispersal and Deposition Using HUMPOL Software, Including Simulation Windroses and Irregular Lakes." *Review of Palaeobotany and Palynology* 134: 185-196.

Camazine, S. 1991. "Self-Organizing Pattern Formation on the Combs of Honey Bee Colonies." *Behavioral Ecology and Sociobiology* 28: 61-76.

Cassagne, E. 2008. "Prévision journalière des pollens sur le territoire national français, avec un objectif d'information sanitaire des populations allergiques" [Daily Pollen Forecast on the French National Territory, with the Objective of Providing Health Information to Allergic Populations]. Ph.D. diss., Université de Bourgogne, Dijon, France.

Castellanos, M.C., P. Wilson and J.D. Thomson. 2003. "Pollen Transfer By Hummingbirds and Bumblebees, and the Divergence of Pollination Modes in *Penstemon*." *Evolution* 57: 2742-2752.

Chapman, D.S., T. Haynes, S. Beal, F. Essl and J.M. Bullock. 2013. "Phenology Predicts the Native and Invasive Range Limits of Common Ragweed." *Global Change Biology*. Advance online publication. doi: 10.1111/gcb.12380.

Chauvel, B., F. Dessaint, C. Cardinal-Legrand and F. Bretagnolle. 2006. "The Historical Spread of *Ambrosia artemisiifolia* L. in France from Herbarium Records." *Journal of Biogeography* 33: 665-673.

Chittka, L., J.D. Thomson and N.M. Waser. 1999. "Flower Constancy, Insect Psychology, and Plant Evolution." *Naturwissenschaften* 86: 361-377.

Clark, J.S., S. LaDeau and I. Ibanez. 2004. "Fecundity of Trees and the Colonization-Competition Hypothesis." *Ecological Monographs* 74: 415-442.

Cressie, N. and C.K. Wikle. 2002. "Space-Time Kalman Filter." In *Encyclopedia of Environmetics*, ed. A.H. El-Shaarawi and W.W. Piegorsch, 2045-2049. Chichester: Wiley.

Cresswell, J.E., J.L. Osborne and S.A. Bell. 2002. "A Model of Pollinator-Mediated Gene Flow between Plant Populations with Numerical Solutions for Bumblebees Pollinating Oilseed Rape." *Oikos* 98: 375-384.

Cresswell, J.E. and J.L. Osborne. 2004. "The Effect of Patch Size and Separation on Bumblebee Foraging in Oilseed Rape: Implications for Gene Flow." *Journal of Applied Ecology* 41: 539-546.

Dafni, A. 1992. *Pollination Ecology: A Practical Approach*. Oxford: Oxford University Press.

DellaValle, C.T., E.W. Triche and M.L. Bell. 2012. "Spatial and temporal modeling of daily pollen concentrations." *International journal of biometeorology* 56: 183-194.

Diaz-Losada, E., G. Ricciardelli-D'Albore and M. Pilar Saa-Otero. 1998. "The Possible Use of Honeybee Pollen Loads in Characterising Vegetation." *Grana* 37: 155-163.

Does, R.J.M.M., L.W.G. Strijbosch and W. Albers. 1988. "Using Jackknife Methods for Estimating the Parameter in Dilution Series." *Biometrics* 44: 1093-1102.

Dupont, S., Y. Brunet and N. Jarosz. 2006. "Eulerian Modelling of Pollen Dispersal over Heterogeneous Vegetation Canopies." *Agricultural and Forest Meteorology* 141: 82-104.

Durham, O.C. 1944. "The Volumetric Incidence of Atmospheric Allergens (II). Simultaneous Measurements by Volumetric and Gravity, Slide methods." *Journal of Allergy* 15: 226-234.

Efstathiou, C., S. Isukapalli, and P. Georgopoulos. 2011. "A mechanistic modeling system for estimating large-scale emissions and transport of pollen and co-allergens." *Atmospheric Environment* 45: 2260-2276.

Erdtman, G. 1986. *Pollen morphology and plant taxonomy: Angiosperms*. Leiden, Netherlands: Brill.

Fazekas de St. Groth, S. 1982. "The Evaluation of Limiting Dilution Assays." *Journal of Immunological Methods* 49: R11-R23.

Firth, D. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80: 27-38.

Fisher, R.A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society A* 222: 309-368.

Frenz, D.A. 1999. "Comparing Pollen and Spore Counts Collected with the Rotorod Sampler and Burkard Spore Trap." *Annals of Allergy, Asthma and Immunology* 83: 341-347.

Frenz, D.A. 2000. "The Effect of Windspeed on Pollen and Spore Counts Collected with the Rotorod Sampler and Burkard Spore Trap." *Annals of Allergy, Asthma and Immunology* 85: 392-394.

Fyfe, R. 2006. "GIS and the Application of a Model of Pollen Deposition and Dispersal: A New Approach to Testing Landscape Hypotheses Using the POLLANDCAL Models." *Journal of Archaeological Science* 33: 483-493.

Gelfand, A.E. and A.F.M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85: 398-409.

Gelfand, A.E., J.A. Silander Jr., S. Wu, A. Latimer, P.O. Lewis, A.G. Rebelo and M. Holder. 2006. "Explaining Species Distribution Patterns through Hierarchical Modeling." *Bayesian Analysis* 1: 41-92.

Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gelman, A. and D.B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7: 457-472.

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721-741.

Genton, B.J., J.A. Shykoff and T. Giraud. 2005. "High Genetic Diversity in French Invasive Populations of Common Ragweed, *Ambrosia artemisiifolia*, as a Result of Multiple Sources of Introduction." *Molecular Ecology* 14: 4275-4285.

Gilioli, G., S. Pasquali, S. Tramontini and F. Riolo. 2013. "Modelling Local and Long-Distance Dispersal of Invasive Chestnut Gall Wasp in Europe." *Ecological Modelling* 263: 281-290.

Goovaerts, P. 1997. *Geostatistics for Natural Resource Evaluation*. Oxford: Oxford University Press.

Goulson, D. 2000. "Why Do Pollinators Visit Proportionally Fewer Flowers in Large Patches?" *Oikos* 91: 485-492.

Greenleaf, S.S., N.M. Williams, R. Winfree and C. Kremen. 2007. "Bee Foraging Ranges and Their Relationship to Body Size." *Oecologia* 153: 589-596.

Gustafson, D.I., I.O. Brants, M.J. Horak, K.M. Remund, E.W. Rosenbaum and J.K. Soteres. 2006. "Empirical Modeling of Genetically Modified Maize Grain Production Practices to Achieve European Union Labeling Thresholds." *Crop Science* 46: 2133-2140.

Haanstra, L., P. Doelman and J.H. Oude Voshaar. 1985. "The Use of Sigmoidal Dose Response Curves in Soil Ecotoxicological Research." *Plant and Soil* 84: 293-297.

Hill, P.S.M., P.H. Wells and H. Wells. 1997. "Spontaneous Flower Constancy and Learning in Honey Bees as a Function of Colour." *Animal Behavior* 54: 615-627.

Hirst, J.M. 1952. "An Automatic Volumetric Spore Trap." *Annals of Applied Biology* 39: 257-265.

Hooten, M.B. and C.K. Wikle. 2005. "A Hierarchical Bayesian Non-Linear Spatio-Temporal Model for the Spread of Invasive Species with Application to the European Collared-Dove." *Environmental and Ecological Statistics* 15: 59-70.

Horrocks, M. and K.A.J. Walsh. 1998. "Forensic Palynology: Assessing the Value of Evidence." *Review of Palaeobotany and Palynology* 103: 69-74.

Hughes, J.P. and P. Totten. 2003. "Estimating the Accuracy of Polymerase Chain Reaction-Based Tests Using Endpoint Dilution." *Biometrics* 59: 505-511.

Isagi, Y. and Y. Suyama, eds. 2011. *Single-Pollen Genotyping*. Tokyo: Springer.

Jarosz, N., B. Loubet, B. Durand, A. McCartney, X. Foueillassar and L. Huber. 2003. "Field Measurements of Airborne Concentration and Deposition Rate of Maize Pollen." *Agricultural and Forest Meteorology* 119: 37-51.

Jarosz, N., B. Loubet and L. Huber. 2004. "Modelling Airborne Concentration and Deposition Rate of Maize Pollen." *Atmospheric Environment* 38: 5555-5566.

Kareiva, P.M. and N. Shigesada. 1983. "Analyzing Insect Movement as a Correlated Random Walk." *Oecologia* 56: 234-238.

Kaye, J.P., A. Majumdar, C. Gries, A. Buyantuyev, N.B. Grimm, D. Hope, G.D. Jenerette, W.X. Zhu and L. Baker. 2008. "Hierarchical Bayesian Scaling of Soil Properties across Urban, Agricultural, and Desert Ecosystems." *Ecological Applications* 18: 132-145.

Klein, E.K., C. Lavigne, X. Foueillassar, P.H. Gouyon and C. Laredo. 2003. "Corn Pollen Dispersal: Quasi-Mechanistic Models and Field Experiments." *Ecological Monographs* 73: 131-150.

Knudsen, J.T. and L. Tollsten. 1995. "Floral Scent in Bat-Pollinated Plants: A Case of Convergent Evolution." *Botanical Journal of the Linnean Society* 119: 45-57.

Kondo, T., S. Nishimura, Y. Naito, Y. Tsumura, T. Okuda, K. Kit, S. Ng, S.L. Lee, N. Muhammad, N. Nakagoshi and Y. Isagi. 2011. "Can Tiny Thrips Provide Sufficient Pollination Service During a General Flowering Period in Tropical Rainforest?" In *Single-Pollen Genotyping*, ed. Y. Isagi and Y. Suyama, 63-81. Tokyo: Springer.

Kuparinen, A., F. Schurr, O. Tackenberg and R.B. O'Hara. 2007. "Air-Mediated Pollen Flow from Genetically Modified Conventional Crops." *Ecological Applications* 17: 431-440.

Levin, D.A. and H.W. Kerster. 1969. "The Dependence of Bee-Mediated Pollen and Gene Dispersal upon Plant Density." *Evolution* 23: 560-571.

Levin, D.A., Y. Peres and E.L. Wilmer. 2009. *Markov Chains and Mixing Times*. Providence: American Mathematical Society.

Lojtnant, C.L., B. Boelt, S.K. Clausen, C. Damgaard, P. Kryger, A. Novy, M. Philipp, C.H. Ingvordsen and R.B. Jorgensen. 2012. "Modelling Gene Flow between Fields of White Clover with Honeybees as Pollen Vectors." *Environmental Modeling and Assessment* 17: 421-430.

Loos, C., R. Seppelt, S. Meier-Bethke, J. Schiemann and O. Richter. 2003. "Spatially Explicit Modelling of Transgenic Maize Pollen Dispersal and Cross-Pollination." *Journal of Theoretical Biology* 225: 241-255.

Lunn, D.J., A. Thomas, N. Best and D. Spiegelhalter. 2000. "WinBUGS – a Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing* 10: 325-337.

Macken, C. 1999. "Design and Analysis of Serial Limiting Dilution Assays with Small Sample Sizes." *Journal of Immunological Methods* 222: 13-29.

Matsuki, Y., M. Tomita and Y. Isagi. 2011. "Pollination Efficiencies of Insects Visiting *Magnolia obovata*, as Determined by Single-Pollen Genotyping." In *Single-Pollen Genotyping*, ed. Y. Isagi and Y. Suyama, 17-32. Tokyo: Springer.

Matthews, J.N.S. 1998. "Alternative Criteria for Optimal Design of Limiting Dilution Assays." *Statistics in Medicine* 17: 2733-2746.

Matthys-Rochon, E., P. Vergne, S. Detchepare and C. Dumas. 1987. "Male Germ Unit Isolation from Three Tricellular Pollen Species: *Brassica oleracea*, *Zea mays*, and *Triticum aestivum*." *Plant Physiology* 83: 464-466.

May, W.E., D.J. Hume and B.A. Hale. 1994. "Effects of Agronomic Practices on Free Fatty Acid Levels in the Oil of Ontario-Grown Spring Canola." *Canadian Journal of Plant Science* 74: 267-274.

McCormick, S. 1993. "Male Gametophyte Development." *The Plant Cell* 5: 1265-1275.

Mehrabi, Y. and J.N.S. Matthews. 1995. "Likelihood-Based Methods for Bias Reduction in Limiting Dilution Assays." *Biometrics* 51: 1543-1549.

Mehrabi, Y. and J.N.S. Matthews. 1998. "Implementable Bayesian Designs for Limiting Dilution Assays." *Biometrics* 54: 1398-1406.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller. 1953. "Equations of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21: 1087-1092.

Mitsumoto, K., K. Yabusaki, K. Kobayashi and H. Aoyagi. 2009. "Development of a Novel Real-Time Pollen-Sorting Counter Using Species-Specific Pollen Autofluorescence." *Aerobiologia* 26: 99-111.

Myers, L.E., L.J. McQuay and F.B. Hollinger. 1994. "Dilution Assay Statistics." *Journal of Clinical Microbiology* 32: 732-739.

Nathan, R., G.G. Katul, H.S. Horn, S.M. Thomas, R. Oren, R. Avissar, S.W. Pacala and S.A. Levin. 2002. "Mechanisms of Long-Distance Dispersal of Seeds by Wind." *Nature* 418: 409-413.

Neal, R.M. 2003. "Slice sampling." *Annals of statistics* 705-741.

Nilsson, S. 1988. "Preliminary Inventory of Aerobiological Monitoring Stations in Europe." *Aerobiologia* 4: 4-7.

Osborne, J.L., S.J. Clark, R.J. Morris, I.H. Williams, J.R. Riley, A.D. Smith, D.R. Reynolds and A.S. Edwards. 1999. "A Landscape Scale Study of Bumble Bee Foraging Range and Constancy, Using Harmonic Radar." *Journal of Applied Ecology* 36: 519-533.

Paciorek, C.J. and J.S. McLachlan. 2009. "Mapping Ancient Forests: Bayesian Inference for Spatio-Temporal Trends in Forest Composition." *Journal of the American Statistical Association* 104: 608-622.

Pan, Y.-L., S.C. Hill, R.G. Pinnick, J.M. House, R.C. Flagan and R.K. Chang. 2011. "Dual-Excitation-Wavelength Fluorescence Spectra and Elastic Scattering for Differentiation of Single Airborne Pollen and Fungal Particles." *Atmospheric Environment* 45: 1555-1563.

Pearl, J. 1985. "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning." *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*. Irvine, CA.

Pebesma, E.J. 2004. "Multivariable Geostatistics in S: The gstat Package." *Computers and Geosciences* 30: 683-691.

Petermann, A. 2011. "Cartographie Nationale de l'Ambroisie (*Ambrosia artemisiifolia* L.)" [National Mapping of Common Ragweed]. Paris: Fédération des Conservatoires botaniques nationaux.

Petersen, G., B. Johansen and O. Seberg. 1996. "PCR and Sequencing from a Single Pollen Grain." *Plant Molecular Biology* 31: 189-191.

Plummer, M. 2003. "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria.

Pouvreau, A. 2004. *Les insectes pollinisateurs* [Insect Pollinators]. Paris: Delachaux et Niestlé.

Prank, M., D.S. Chapman, J.M. Bullock, J. Belmonte, U. Berger, A. Dahl, S. Jäger, I. Kovtunenko, D. Magyar, S. Niemelä, A. Rantio-Lehtimäki, V. Rodinkova, I. Sauliene, E. Severova, B. Sikoparija and M. Sofiev. 2013. "An Operational Model for Forecasting Ragweed Pollen Release and Dispersion in Europe." *Agricultural and Forest Meteorology* 182-183: 43-53.

Prentice, C. 1985. "Pollen Representation, Source Area, and Basin Size: Toward a Unified Theory of Pollen Analysis." *Quaternary Research* 23: 76-86.

Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery. 1996. *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing*, $2^n d$ ed. Cambridge: Cambridge University Press.

Pyke, G.H. 1978. "Optimal Foraging: Movement Patterns of Bumblebees between Inflorescences." *Theoretical Population Biology* 13: 72-98.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. URL http://www.r-project.org.

Ramsay, G., C. Thompson and G. Squire. 2003. "Quantifying Landscape-Scale Gene Flow in Oilseed Rape." Final Report of DEFRA Project RG0216. London: Department for Environment, Food and Rural Affairs.

Raynor, G.S., E.C. Ogden and J.V. Hayes. 1970. "Dispersion and Deposition of Ragweed Pollen from Experimental Sources." *Journal of Applied Meteorology* 9: 885-895.

Ridout, M.S. 1995 "Three-Stage Designs for Seed Testing Experiments." *Journal of the Royal Statistical Society C: Applied Statistics* 44: 153-162.

Ritz, C. and J.C. Streibig. 2005. "Bioassay Analysis Using R." *Journal of Statistical Software* 12(5).

Rodrigo, A.G., P.C. Goracke, K. Rowhanian and J.I. Mullins. 1997. "Quantitation of Target Molecules from Polymerase Chain Reaction-Based Limiting Dilution Assays." *AIDS Research and Human Retroviruses* 13: 737-742.

Romeis, J., E. Städler and F.L. Wäckers. 2005. "Nectar- and Pollen-Feeding by Adult Herbivorous Insects." In *Plant-Provided Food for Carnivorous Insects*, ed. F.L. Wäckers, P.C.J. van Rijn and J. Bruin, 178-219. Cambridge: Cambridge University Press.

Rose, R., G.P. Dively and J. Pettis. 2007. "Effects of Bt Corn Pollen on Honey Bees: Emphasis on Protocol Development." *Apidologie* 38: 368-377.

Rousseau, D.-D., P. Schevin, J. Ferrier, D. Jolly, T. Andreasen, S.E. Ascanius, S.-E. Hendriksen and U. Poulsen. 2008. "Long-Distance Pollen Transport from North America to Greenland in Spring." *Journal of Geophysical Research* 113: G02013.

Scott, R.J. 1995. "Pollen Exine - the Sporopollenin Enigma and the Physics of Pattern." In *Molecular and Cellular Aspects of Plant Reproduction*, ed. R.J. Scott and A.D. Stead, 49-82. Cambridge: Cambridge University Press.

Smolik, M.G., S. Dullinger, F. Essl, I. Kleinbauer, M. Leitner, J. Peterseil, L.-M. Stadler and G. Vogl. 2010. "Integrating Species Distribution Models and Interacting Particle Systems to Predict the Spread of an Invasive Alien Plant." *Journal of Biogeography* 37: 411-422.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin and A. Van der Line. 2002. "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society B* 64: 583-616.

Stanley-Horn, D.E., G.P. Dively, R.L. Hellmich, H.R. Mattila, M.K. Sears, R. Rose, L.C.H. Jesse, J.E. Losey, J.J. Obrycki and L. Lewis. 2001. "Assessing the Impact of Cry1Ab-Expressing Corn Pollen on Monarch Butterfly Larvae in Field Studies." *Proceedings of the National Academy of Sciences of the United States of America* 98: 11931-11936.

Steffan-Dewenter, I. and A. Kuhn. 2003. "Honeybee Foraging in Differentially Structured Landscapes." *Proceedings of the Royal Society B* 270: 569-575.

Strijbosch, L.W.G., W.A. Burman, R.J.M.M. Does, P.H. Zeinen and G. Groenewegen. 1987. "Limiting Dilution Assays: Experiment Design and Statistical Analysis." *Journal of Immunological Methods* 97: 133-140.

Sugita, S. 1993. "A Model of Pollen Source Area for an Entire Lake Surface." *Quaternary Research* 39: 239-244.

Sugita, S. 1994. "Pollen Representation of Vegetation in Quaternary Sediments: Theory and Method in Patchy Vegetation." *Journal of Ecology* 82: 881-897.

Sugita, S. 2007a. "Theory of Quantitative Reconstruction of Vegetation I: Pollen from Large Sites REVEALS Regional Vegetation Composition." *The Holocene* 17: 229-241.

Sugita, S. 2007b. "Theory of Quantitative Reconstruction of Vegetation II: All You Need Is LOVE." *The Holocene* 17: 243-257.

Sykes, P.J., S.H. Neoh, M.J. Brisco, E. Hughes, J. Condon and A.A. Morley. 1992. "Quantitation of Targets for PCR by Use of Limiting Dilution." In *The PCR Technique: Quantitative PCR*, ed. J.W. Larrick, 81-91. Westborough, MA: Eaton.

Taswell, C. 1981. "Limiting Dilution Assays for the Determination of Immunocompetent Cell Frequencies I. Data Analysis." *Journal of Immunology* 126: 1614-1619.

Tauber, H. 1967. "Investigations of the Mode of Pollen Transfer in Forested Areas." *Review of Palaeobotany and Palynology.* 3: 277-286.

Tufto, J., S. Engen and K. Hindar. 1997. "Stochastic Dispersal Processes in Plant Populations." *Theoretical Population Biology* 52: 16-26.

Visscher, P.K. and T.D. Seeley. 1982. "Foraging Strategy of Honeybee Colonies in a Temperate Deciduous Forest." *Ecology* 63: 1790-1801.

Von der Ohe, W., L. Persano Oddo, M.L. Piana, M. Morlot and P. Martin. 2004. "Harmonized Methods of Melissopalynology." *Apidologie* 35: S18-S25.

Walther-Hellwig, K. and R. Frankl. 2000. "Foraging Habits and Foraging Distances of Bumblebees, *Bombus* spp. (Hymenoptera, Apidae), in an Agricultural Landscape." *Journal of Applied Entomology* 124: 299-306.

Wheeler, R.E. 2010. *lmPerm*. The R Project for Statistical Computing. URL http://www.r-project.org.

Wikle, C.K. 2003. "Hierarchical Bayesian Models for Predicting the Spread of Ecological Processes." *Ecology* 84: 1382-1394.

Wilson, J.D. 2000. "Trajectory Models for Heavy Particles in Atmospheric Turbulence: Comparison with Observations." *Journal of Applied Meteorology* 39: 1894-1912.

Wu, H., B.-L. Li, T.A. Springer and W.H. Neill. 2000. "Modelling Animal Movement as a Persistent Random Walk in Two Dimensions: Expected Magnitude of Net Displacement." *Ecological Modelling* 132: 115-124.

Zelterman, D., A. Tulupyev, R. Heimer and N. Abdala. 2009. "Statistical Design for a Small Serial Dilution Series." *Statistics in Medicine* 29: 411-420.

Zemmer, F., F. Karaca and F. Ozkaragoz. 2012. "Ragweed Pollen Observed in Turkey: Detection of Sources Using Back Trajectory Models." *Science of the Total Environment* 430: 101-108.

Zheng, Y. and B.H. Aukema. 2010. "Hierarchical Dynamic Modeling of Outbreaks of Mountain Pine Beetle Using Partial Differential Equations." *Environmetrics* 21: 801-816.

Zink, K., H. Vogel, B. Vogel, D. Magyar and C. Kottmeier. 2012. "Modeling the Dispersion of *Ambrosia artemisiifolia* L. Pollen with the Model System COSMO-ART." *International Journal of Biometeorology* 56: 669-680.

Appendix A

Equations and source code implementing the limited dilution assay model

This appendix derives the equations and presents the Fortran and R code used to perform calculations under the limited dilution assay model presented in Chapter 2.

A.1 Derivation of the corrected score function

First I re-introduce the score function for a limited dilution assay (2.4):

$$U(\phi) = \sum_{d=1}^{D} \left[\frac{k_d}{p_d(\phi)} - \frac{R_d - k_d}{1 - p_d(\phi)} \right] \frac{\mathrm{d}p_d}{\mathrm{d}\phi}, \tag{A.1}$$

where d is the number of dilution levels, R_d is the number of replicates (i.e. PCR tubes) at level d, k_d is the number of positives recorded out of R_d replicates, and p_d is the probability of getting a positive result at level d. This probability of success is a function of ϕ, the (unknown) frequency of pollen bearing the marker of interest.

Since the root of this function will be a biased estimate of ϕ, I will use the bias-corrected score function, U^*:

$$U^*(\phi) = U(\phi) - \frac{1}{2\mathbf{E}\left[\frac{\partial U}{\partial \phi}\right]} \left\{ \mathbf{E}\left[\frac{\partial^2 U}{\partial \phi^2}\right] - 2\frac{\partial}{\partial \phi}\mathbf{E}\left[\frac{\partial U}{\partial \phi}\right] \right\}. \tag{A.2}$$

Evaluating U^* requires the first two derivatives of U:

$$\frac{\partial U}{\partial \phi} = \sum_{d=1}^{D} \left[\left(\frac{k_d}{p_d(\phi)} - \frac{R_d - k_d}{1 - p_d(\phi)} \right) \frac{\mathrm{d}^2 p_d}{\mathrm{d}\phi^2} - \left(\frac{k_d}{p_d(\phi)^2} + \frac{R_d - k_d}{\big(1 - p_d(\phi)\big)^2} \right) \left(\frac{\mathrm{d}p_d}{\mathrm{d}\phi} \right)^2 \right] \tag{A.3}$$

and

$$\frac{\partial^2 U}{\partial \phi^2} = \sum_{d=1}^{D} \left[\left(\frac{k_d}{p_d(\phi)} - \frac{R_d - k_d}{1 - p_d(\phi)} \right) \frac{\mathrm{d}^3 p_d}{\mathrm{d}\phi^3} - 3 \left(\frac{k_d}{p_d(\phi)^2} + \frac{R_d - k_d}{\big(1 - p_d(\phi)\big)^2} \right) \frac{\mathrm{d}p_d}{\mathrm{d}\phi} \frac{\mathrm{d}^2 p_d}{\mathrm{d}\phi^2} \right.$$
$$\left. -2 \left(\frac{R_d - k_d}{\big(1 - p_d(\phi)\big)^3} - \frac{k_d}{p_d(\phi)^3} \right) \left(\frac{\mathrm{d}p_d}{\mathrm{d}\phi} \right)^3 \right]. \tag{A.4}$$

Since both derivatives are linear functions of the observed values k_d, their expectations can be calculated simply by replacing k_d with $\mathbf{E}(k_d) = p_d(\phi) R_d$:

$$\mathbf{E}\left[\frac{\partial U}{\partial \phi} \right] = -\sum_{d=1}^{D} \frac{R_d}{p_d(\phi)\big(1 - p_d(\phi)\big)} \left(\frac{\mathrm{d}p_d}{\mathrm{d}\phi} \right)^2 \tag{A.5}$$

and

$$\mathbf{E}\left[\frac{\partial^2 U}{\partial \phi^2} \right] = -\sum_{d=1}^{D} \left[\frac{3R_d}{p_d(\phi)\big(1 - p_d(\phi)\big)} \frac{\mathrm{d}p_d}{\mathrm{d}\phi} \frac{\mathrm{d}^2 p_d}{\mathrm{d}\phi^2} \right.$$
$$\left. +2R_d \left(\frac{1}{\big(1 - p_d(\phi)\big)^2} - \frac{1}{p_d(\phi)^2} \right) \left(\frac{\mathrm{d}p_d}{\mathrm{d}\phi} \right)^3 \right]. \tag{A.6}$$

The final term needed is the derivative of (A.5):

$$\frac{\partial}{\partial \phi} \mathbf{E}\left[\frac{\partial U}{\partial \phi} \right] = -\sum_{d=1}^{D} \left[\frac{2R_d}{p_d(\phi)\big(1 - p_d(\phi)\big)} \frac{\mathrm{d}p_d}{\mathrm{d}\phi} \frac{\mathrm{d}^2 p_d}{\mathrm{d}\phi^2} \right.$$
$$\left. +R_d \left(\frac{1}{\big(1 - p_d(\phi)\big)^2} - \frac{1}{p_d(\phi)^2} \right) \left(\frac{\mathrm{d}p_d}{\mathrm{d}\phi} \right)^3 \right]. \tag{A.7}$$

Substituting (A.5) to (A.7) into (A.2), I obtain a succinct expression for the bias correction to U:

$$U^*(\phi) = U(\phi) + \frac{\sum_{d=1}^{D} \frac{R_d}{p_d(\phi)\left(1 - p_d(\phi)\right)} \frac{\mathrm{d}p_d}{\mathrm{d}\phi} \frac{\mathrm{d}^2 p_d}{\mathrm{d}\phi^2}}{2 \sum_{d=1}^{D} \frac{R_d}{p_d(\phi)\left(1 - p_d(\phi)\right)} \left(\frac{\mathrm{d}p_d}{\mathrm{d}\phi} \right)^2}. \tag{A.8}$$

When p_d is a strictly increasing, concave function of ϕ, the correction term in (A.8) is negative, which is consistent with the finding that the uncorrected score function would produce an overestimate of ϕ (see Fig. 2.1).

As discussed in Chapter 2, $p_d(\phi)$ takes the following form when the number of pollen grains per sample (N_d) follows a Poisson distribution with mean λ_d:

$$p_d(\phi) = \sum_{N_d=0}^{N_{max}} \frac{{\lambda_d}^{N_d} e^{-\lambda_d}}{N_d!} \sum_{m=0}^{N_d} \binom{N_d}{m} \phi^m \left(1-\phi\right)^{N_d-m} \Pr\left(+ \mid m, N_d\right). \tag{A.9}$$

Its first two derivatives, which enter the calculation of the corrected score function, are:

$$\frac{\mathrm{d}p_d}{\mathrm{d}\phi} = \sum_{N_d=0}^{N_{max}} \frac{{\lambda_d}^{N_d} e^{-\lambda_d}}{N_d!} \sum_{m=0}^{N_d} \binom{N_d}{m} \phi^m \left(1-\phi\right)^{N_d-m} \Pr\left(+ \mid m, N_d\right) \left[\frac{m}{\phi} - \frac{N_d-m}{1-\phi}\right] \tag{A.10}$$

and

$$\frac{\mathrm{d}^2p_d}{\mathrm{d}\phi^2} = \sum_{N_d=0}^{N_{max}} \frac{{\lambda_d}^{N_d} e^{-\lambda_d}}{N_d!} \sum_{m=0}^{N_d} \binom{N_d}{m} \phi^m \left(1-\phi\right)^{N_d-m} \Pr\left(+ \mid m, N_d\right) \left[\left(\frac{m}{\phi} - \frac{N_d-m}{1-\phi}\right)^2 - \frac{m}{\phi^2} - \frac{N_d-m}{(1-\phi)^2}\right]. \tag{A.11}$$

If the number of pollen grains in each reaction, N_d, is exactly known, then the previous three equations are simply modified by removing the first sum and the Poisson probability term.

A.2 Calculation of dilution assay properties in Fortran

I have appended below the Fortran 90 program I wrote to calculate the distribution of $\hat{\phi}$, depending on the true value of ϕ and the dilution assay design (R_d and N_d or λ_d for each of the d levels).

I decided to use Brent's algorithm from Press et al. (1996) to find the root of the corrected score function, instead of the Newton-Raphson method. While the latter converges at a faster rate in some cases, it requires evaluating the function's derivative, which in this case adds a significant computational cost: the correction term for the derivative of U^* has six sum terms compared with the two sums in (A.8). When testing both algorithms, I found Brent's method to produce faster results.

Even with a stable root finding function, a problem can occur at values of ϕ close to 1, when the derivative of p_d approaches zero. Machine roundoff errors can result in the wrong sign determination for the derivative of p_d, and thus for the score function (A.1) itself.

Since under the assumptions of my model, p_d is a strictly increasing function of ϕ, I decided to set a very small positive number as the lower bound for its derivative. Because this lower bound is arbitrary, I do not apply the bias correction in this case; the result is that some outcomes with ϕ very close to 1 will be classified as non-informative (though no more than would be if I did not include a bias correction at all). Classifying an outcome as non-informative is still preferrable to obtaining a nonsense estimate due to a roundoff sign error.

```
MODULE mod_pollen
    USE nrtype  ! Module from Numerical Recipes
    IMPLICIT NONE
    INTEGER(I4B), PARAMETER :: FILENUM = 99
    INTEGER(I4B), PARAMETER :: NMAX = 200 ! Max value for either np or n_rep in PROGRAM pollenPCR
    INTEGER(I4B), PARAMETER :: NDMAX = 20 ! Max value for nd in PROGRAM pollenPCR
    INTEGER(I4B), PARAMETER :: NSAMP = 100 ! How many samples from parameter estimate distribution
    REAL(DP), PARAMETER :: LOW = 0.001_dp, HIGH = 0.999_dp ! phi_est must be between LOW and HIGH
    REAL(DP), PARAMETER :: EPS = 1.0D-6 ! Absolute accuracy required for root finding function
    REAL(DP), PARAMETER :: ALPHA = 1.0D-3 ! Outcomes with probability < alpha / #outcomes are ignored
    LOGICAL, PARAMETER :: POISSON = .false. ! If true, np follows a Poisson distr.; if false, np is constant
    LOGICAL, PARAMETER :: TWO_STEP = .true. ! Two-step or one-step design
    REAL(DP), DIMENSION(0:NMAX, 0:NMAX) :: bin_coef ! Array to pre-compute binomial coefficients (n k)
    REAL(DP), DIMENSION(0:NMAX, 0:NMAX/2) :: pois_dist ! Array to pre-compute Poisson distribution P(k,lambda)
    REAL(DP), DIMENSION(0:NMAX, 0:NMAX) :: p_pos, p_pos_e
        ! p_pos(m,n) [P(+|m,N_d) in text] is the probability of a positive PCR
        ! when m grains out of n have the marker
        ! p_pos_e is the estimate of p_pos (for inference)
    INTEGER(I4B) :: nd ! [D in text] Number of dilution levels
    INTEGER(I4B), DIMENSION(1:NDMAX) :: n_rep, mean_np, n_pos
        ! [Respectively R_d, N_d and k_d in text] Number of replicates, number of grains per replicate
        !  and number of positive amplifications by dilution level
    LOGICAL :: is_inferring ! if is_inferring, parameter estimates are used instead of true values

    REAL(DP):: phi_min ! Min possible value for "real" phi (i.e. limit of detection)
    REAL(DP) :: fpos_t, freac_t, fdeg_t, rinh_t, ninh_t ! True parameters of error model
    REAL(DP) :: fpos_e, freac_e, fdeg_e, rinh_e, ninh_e ! Estimated parameters of error model
END MODULE mod_pollen


PROGRAM pollenPCR
    USE nrtype; USE mod_pollen
    IMPLICIT NONE
    INTERFACE
        SUBROUTINE pol2s_sim(phi,avg_phi,var_phi,p_info)
        USE nrtype; USE nr; USE mod_pollen
        IMPLICIT NONE
        REAL(DP), INTENT(IN) :: phi
        REAL(DP), INTENT(OUT) :: avg_phi, var_phi, p_info
        END SUBROUTINE pol2s_sim

        SUBROUTINE array_init
        USE nrtype; USE mod_pollen
        IMPLICIT NONE
        END SUBROUTINE array_init
    END INTERFACE

    INTEGER(I4B) :: i
    REAL(DP) :: phi, avg, var, pinf
    REAL(DP), DIMENSION(1:NSAMP) :: fp_val, fr_val, fd_val, r_val, n_val

    ! Input error parameters form file
    open(unit=FILENUM, file='err_param.txt', action='read')
    do i = 1,NSAMP
```

```
        read(FILENUM,*) fp_val(i), fr_val(i), fd_val(i), r_val(i), n_val(i)
    end do
    close(FILENUM)

    open(unit=FILENUM, file='pollenPCR.csv', action='write')

    ! Initialize phi, two-step assay parameters and true error parameters
    phi = 0.02
    nd = 6
    n_rep(1:nd) = 2
    n_rep(nd+1) = 38

    fpos_t = 0.05
    freac_t = 0.1
    fdeg_t = 0.2
    rinh_t = 1.0
    ninh_t = 40

    ! Pre-compute bin_coef, pois_dist and p_pos tables
    call array_init

    do i = 1, NSAMP
        fpos_e = fp_val(i)
        freac_e = fr_val(i)
        fdeg_e = fd_val(i)
        rinh_e = r_val(i)
        ninh_e = n_val(i)
        phi_min = 1.0_dp / (0.6_dp*real(ninh_e, DP))
        call array_init
        call pol2s_sim(phi,avg,var,pinf)
        write(FILENUM,*) real(fpos_e,SP), real(freac_e,SP), real(fdeg_e,sp),  &
            real(rinh_e,SP), real(ninh_e,SP), real(avg,SP), real(var,SP), real(pinf,SP)
    end do
    close(FILENUM)

END PROGRAM pollenPCR


SUBROUTINE pol2s_sim(phi,avg_phi,var_phi,p_info)
    USE nrtype; USE nr; USE mod_pollen  ! Module nr includes zbrent function from Numerical Recipes
    IMPLICIT NONE
    INTERFACE
        FUNCTION func(x)
        USE nrtype; USE mod_pollen
        IMPLICIT NONE
        REAL(DP), INTENT(IN) :: x
        REAL(DP) :: func
        END FUNCTION func

        SUBROUTINE calc_p(phi,lambda,p,d_p,d2_p)
        USE nrtype; USE mod_pollen
        IMPLICIT NONE
        REAL(DP), INTENT(IN) :: phi
        INTEGER(I4B), INTENT(IN) :: lambda
        REAL(DP), INTENT(OUT) :: p, d_p, d2_p
        END SUBROUTINE calc_p
    END INTERFACE

    REAL(DP), INTENT(IN) :: phi
    REAL(DP), INTENT(OUT) :: avg_phi, var_phi, p_info
    INTEGER(I4B) :: d, i, n_outcomes
    LOGICAL :: done
    REAL(DP) :: phi_est, p_outcome, p_step1, a  ! a is geometric progression factor
    REAL(DP) :: d1p, d2p, dl_left, dl_right
    REAL(DP), DIMENSION(1:NDMAX) :: p_succ
```

```
! Calculates the max. likelihood estimate of phi (the marker frequency),
! its variance and the probability of an informative assay
! for a two-step dilution PCR experiment:
! Step 1 - logarithmic progression with nd levels for mean_np(d),
!         n_rep(d) is the number of replicates at level d
! Step 2 - all replicates at same level, chosen from result of step 1
!          (only if TWO_STEP = TRUE)

! Dilution levels follow a geometric progression (mean_np goes from 1 to 1/phi_min)
a = (1.0_dp / phi_min)**(1.0_dp / (nd - 1))
do d = 1, nd
    mean_np(d) = nint(a**(d-1))
end do

n_pos(1:nd) = 0
n_outcomes = product(n_rep(1:nd) + 1) ! Number of possible Step 1 outcomes

avg_phi = 0.0_dp
var_phi = 0.0_dp
p_info = 0.0_dp

! Calculate probability of individual PCR success for given value of phi at each dilution level
! (d1p and d2p are placeholders)
is_inferring = .false.
do d = 1, nd
    call calc_p(phi, mean_np(d), p_succ(d), d1p, d2p)
end do
is_inferring = .true.

!  Loop over Step 1 outcomes
done = .false.
do while (.not. done)

    p_outcome = 1.0_dp

    do d = 1, nd
        ! Deal with p(+) = 0 and 1 separately (to avoid 0^0 errors)
        if ( p_succ(d) == 0.0_dp .and. n_pos(d) == 0 .or.  &
             p_succ(d) == 1.0_dp .and. n_pos(d) == n_rep(d) ) cycle
        if ( p_succ(d) == 0.0_dp .and. n_pos(d) > 0 .or.   &
             p_succ(d) == 1.0_dp .and. n_pos(d) < n_rep(d) ) then
            p_outcome = 0.0_dp
            exit
        end if

        p_outcome = p_outcome * bin_coef(n_rep(d), n_pos(d)) * p_succ(d)**n_pos(d) &
            * (1.0_dp - p_succ(d))**(n_rep(d) - n_pos(d))
    end do

    ! Calculate the estimate of phi if this outcome passes the min. probability threshold
    if ( p_outcome > ALPHA / n_outcomes ) then
        ! If the score function (log-likelihood derivative) is positive for at phi = LOW
        ! and negative at phi = HIGH, then there is a root between LOW and HIGH, find it
        dl_left = func(LOW)
        dl_right = func(HIGH)
        if (dl_left < 0) then
            phi_est = 0.0_dp
        else if (dl_right > 0) then
            phi_est = 1.0_dp
        else
            phi_est = zbrent(func, LOW, HIGH, EPS)

            if (TWO_STEP) then
                ! Step 2 design (add a new dilution level)
                p_step1 = p_outcome
                nd = nd + 1
```

```
                    mean_np(nd) = nint(1.0_dp / max(phi_est, phi_min))
                    is_inferring = .false.
                    call calc_p(phi, mean_np(nd), p_succ(nd), d1p, d2p)
                    is_inferring = .true.

                    ! Cycle through Step 2 outcomes, find phi_est as above
                    do i = 0, n_rep(nd)
                        n_pos(nd) = i
                        if ( p_succ(d) == 0.0_dp .and. n_pos(d) > 0 .or.  &
                            p_succ(d) == 1.0_dp .and. n_pos(d) < n_rep(d) ) cycle
                        if ( p_succ(d) > 0.0_dp .and. p_succ(d) < 1.0_dp ) then
                            p_outcome = p_step1 * bin_coef(n_rep(nd), n_pos(nd)) &
                                * p_succ(nd)**n_pos(nd) * (1.0_dp - p_succ(nd))**(n_rep(nd) - n_pos(nd))
                        end if
                        dl_left = func(LOW)
                        dl_right = func(HIGH)
                        if (dl_left < 0) then
                            phi_est = 0.0_dp
                        else if (dl_right > 0) then
                            phi_est = 1.0_dp
                        else
                            phi_est = zbrent(func, LOW, HIGH, EPS)
                            ! Update global stats
                            p_info = p_info + p_outcome
                            avg_phi = avg_phi + p_outcome * phi_est
                            var_phi = var_phi + p_outcome * phi_est**2
                        end if
                    end do
                    nd = nd - 1
                end if

            end if
        end if

        ! Cycle to next Step 1 outcome (test every combination of n_pos(d) for the nd dilution levels)
        d = 1
        do while ( d < nd .and. n_pos(d) == n_rep(d) )
            n_pos(d) = 0
            d = d + 1
        end do
        if ( n_pos(d) == n_rep(d) ) then
            done = .true.
        else
            n_pos(d) = n_pos(d) + 1
        end if

    end do

    ! Calculate average and variance
    avg_phi = avg_phi / p_info
    var_phi = var_phi / p_info - avg_phi**2

END SUBROUTINE pol2s_sim


FUNCTION func(x)
    USE nrtype; USE mod_pollen
    IMPLICIT NONE
    INTERFACE
        SUBROUTINE calc_p(phi,lambda,p,d_p,d2_p)
        USE nrtype; USE mod_pollen
        IMPLICIT NONE
        REAL(DP), INTENT(IN) :: phi
        INTEGER(I4B), INTENT(IN) :: lambda
        REAL(DP), INTENT(OUT) :: p, d_p, d2_p
        END SUBROUTINE calc_p
```

```
    END INTERFACE

    REAL(DP), INTENT(IN) :: x
    REAL(DP) :: func
    REAL(DP), PARAMETER :: TINY = 1.0D-12
    INTEGER(I4B) :: d
    REAL(DP) :: pval, pderiv, p2deriv
    REAL(DP) :: f_mult
    REAL(DP) :: sum1, sum2, cmult

    ! Given a marker frequency (here x) and the observed data ( n_pos(1,..,nd) ),
    ! this subroutine returns the bias-corrected score function (return value func)
    ! for use in the Numerical Recipes zbrent function.

    ! Uses calc_p to calculate the prob. of success of a single reaction (and its derivatives)

    func = 0.0_dp
    sum1 = 0.0_dp
    sum2 = 0.0_dp

    do d = 1, nd
        call calc_p(x, mean_np(d), pval, pderiv, p2deriv)
        ! Keep pval at least TINY away from 0 or 1 to prevent division by zero
        if (pval < TINY) pval = TINY
        if (pval > 1.0_dp - TINY) pval = 1.0_dp - TINY
        ! Keep pderiv > TINY to avoid sign-changing roundoff error
        if (pderiv < TINY) pderiv = TINY

        f_mult = real(n_pos(d),DP)/pval - real(n_rep(d) - n_pos(d), DP)/(1.0_dp - pval)
        func = func + pderiv * f_mult

        ! Quantities used to calculate bias correction
        if (pderiv > TINY) then
            cmult = real(n_rep(d),DP) / (pval * (1.0_dp - pval))
            sum1 = sum1 + cmult * pderiv * p2deriv
            sum2 = sum2 + cmult * pderiv**2
        end if
    end do

    ! Apply bias correction
    if (pderiv > TINY) then
        func = func + sum1 / (2.0_dp * sum2)
    end if

END FUNCTION func


SUBROUTINE calc_p(phi,lambda,p,d_p,d2_p)
    USE nrtype; USE mod_pollen
    IMPLICIT NONE
    REAL(DP), INTENT(IN) :: phi
    INTEGER(I4B), INTENT(IN) :: lambda
    REAL(DP), INTENT(OUT) :: p, d_p, d2_p
    INTEGER(I4B) :: k, np
    REAL(DP) :: pval, pderiv, p2deriv
    REAL(DP) :: partial_p, p_bin, dp_bin, d2p_bin, dp_mult, d2p_mult

    ! Calculate the probability of success p(phi) as well as its first two derivatives (d_p, d2_p)
    pval = 0.0_dp
    pderiv = 0.0_dp
    p2deriv = 0.0_dp

    if (POISSON) then

        do np = 0, min(2*lambda+3, NMAX)
            p_bin = 0.0_dp
```

```
            dp_bin = 0.0_dp
            d2p_bin = 0.0_dp
            do k = 0, np
                if (is_inferring) then
                    partial_p = bin_coef(np,k) * phi**k * (1.0_dp - phi)**(np - k) * p_pos_e(k,np)
                else
                    partial_p = bin_coef(np,k) * phi**k * (1.0_dp - phi)**(np - k) * p_pos(k,np)
                end if
                dp_mult = real(k,DP)/phi - real(np - k, DP)/(1.0_dp - phi)
                d2p_mult = - real(k,DP)/phi**2 - real(np - k, DP)/(1.0_dp - phi)**2
                p_bin = p_bin + partial_p
                dp_bin = dp_bin + partial_p * dp_mult
                d2p_bin = d2p_bin + partial_p * ( dp_mult**2 + d2p_mult )
            end do
            pval = pval + pois_dist(np,lambda) * p_bin
            pderiv = pderiv + pois_dist(np,lambda) * dp_bin
            p2deriv = p2deriv + pois_dist(np,lambda) * d2p_bin
        end do

    else

        np = lambda
        do k = 0, np
            if (is_inferring) then
                partial_p = bin_coef(np,k) * phi**k * (1.0_dp - phi)**(np - k) * p_pos_e(k,np)
            else
                partial_p = bin_coef(np,k) * phi**k * (1.0_dp - phi)**(np - k) * p_pos(k,np)
            end if
            dp_mult = real(k,DP)/phi - real(np - k, DP)/(1.0_dp - phi)
            d2p_mult = - real(k,DP)/phi**2 - real(np - k, DP)/(1.0_dp - phi)**2
            pval = pval + partial_p
            pderiv = pderiv + partial_p * dp_mult
            p2deriv = p2deriv + partial_p * ( dp_mult**2 + d2p_mult )
        end do

    end if

    p = pval
    d_p = pderiv
    d2_p = p2deriv

END SUBROUTINE calc_p


SUBROUTINE array_init
    USE nrtype; USE mod_pollen
    IMPLICIT NONE
    INTEGER(I4B) :: k, m, n, kmax, lambda

    ! Compute and store all binomial coefficients (n k) with n and k up to NMAX
    bin_coef(0,0) = 1
    bin_coef(1,0) = 1
    bin_coef(1,1) = 1

    do n = 2, NMAX
        if (mod(n,2) == 0) then
            kmax = n/2
        else
            kmax = (n-1)/2
        end if
        bin_coef(n,0) = 1
        bin_coef(n,n) = 1
        do k = 1, kmax
            bin_coef(n,k) = bin_coef(n-1,k-1) + bin_coef(n-1,k)
            bin_coef(n,n-k) = bin_coef(n,k)
        end do
```

```
    end do

    ! If needed, compute and store the Poisson probabilities (k;lambda) up to lambda = NMAX/2
    if (POISSON) then
        do lambda = 1, NMAX/2
            pois_dist(0,lambda) = exp(-REAL(lambda,DP))
            do k = 1, NMAX
                pois_dist(k,lambda) = pois_dist(k-1,lambda) * REAL(lambda,DP) / REAL(k,DP)
            end do
        end do
    end if

    ! Set the probability of a positive PCR result, given n total grains of which m have the marker
    ! p_pos is the true value, p_pos_e is the estimated value for inference
    do n = 0, NMAX
        do m = 0, n

            if (m == 0) then
                p_pos(m,n) = fpos_t
                p_pos_e(m,n) = fpos_e
            else
                p_pos(m,n) = (1.0_dp - freac_t) * (1.0_dp - fdeg_t**m) &
                    / (1.0_dp + exp(rinh_t*(n-ninh_t)))
                p_pos_e(m,n) = (1.0_dp - freac_e) * (1.0_dp - fdeg_t**m) &
                    / (1.0_dp + exp(rinh_e*(n-ninh_e)))
            end if

        end do
    end do

END SUBROUTINE array_init
```

A.3 Error parameter estimation in R

To determine how the uncertainty about PCR false negative parameters (f_{deg}, f_{reac}, r_{inh} and N_{inh50}) affects the inference of ϕ in Chapter 2, I assumed these error parameters were estimated by a calibration assay performed on samples with $\phi = 1$. Given true values of the error parameters, the R scripts below generate samples of parameter estimates from the calibration assay; these samples are then taken as input in the Fortran program above. I used R for this part of the program to take advantage of its built-in functions for joint parameter estimation, as well as the *drc* package (Ritz and Streibig 2005) for estimation of dose-response curves.

```
### R Script 1 ###
# Finds max. likelihood estimate of error parameters f_reac and f_deg
# using limited dilution assays with all-positive samples

# Min. anx max. parameter values (avoid errors at 0 and 1)
low <- 0.001
high <- 0.999

# Assay parameters: r(d) replicates with n(d) grains each at dilution level d
nsamp <- 100
n <- c(1,2,4,8)
r <- c(30,10,6,4)
```

```
a_t <- 0.8   #  a = 1 - f_reac
b_t <- 0.5   #  b = f_deg

sim_res <- function(r, n) {
rbinom( nsamp, r, a_t * (1 - b_t^n) )
}
n_pos <- mapply(sim_res, r, n)

get_est <- function(x) {
k <- x

fbin <- function(x) {
a <- x[1]
b <- x[2]
p <- a * (1 - b^n)
-sum(k * log(p) + (r - k) * log(1 - p))
}

res <- optim(c(0.5,0.5), fbin, method = "L-BFGS-B", lower = rep(low,2), upper = rep(high,2))
res$par
}
par_est <- apply(n_pos, 1, get_est)

# Output (f_reac,f_deg) samples

outfile <- "err_param.txt"
samples <- cbind(1 - par_est[1,],par_est[2,])
write.table(samples,row.names=F,col.names=F,file=outfile)


### R Script 2 ###

# Finds max. likelihood estimate of error parameters r_inh and N_inh50
# using limited dilution assays with all-positive samples

# Load dose-response curve package
library(drc)

# Assay parameters: r(d) replicates with n(d) grains each at dilution level d
nsamp <- 100
n1 <- c(5,15,25,35,45,55,65,75,85,95)
r1 <- rep(3,10)
r2 <- rep(5,7)
r12 <- c(r1,r2)

n_inh <- 40
r_inh <- 1.0
a <- 1.0

sim_res <- function(r, n) {
rbinom( nsamp, r, a/(1+exp(r_inh*(n-n_inh))) )
}
n_pos <- mapply(sim_res, r1, n1)
nsamp <- 1

get_est <- function(x) {
# Get step 1 estimate of n_inh
lreg <- drm(x/r1~n1, weights=r1, type="binomial", logDose=exp(1),
fct=LL2.3(fixed=c(NA,a,NA), names=c("r_e","a_e","n_e")) )
n_est <- round(coef(lreg)[2])
# Draw step 2 outcome
n2 <- c(n_est-4, n_est-2, n_est-1, n_est, n_est+1, n_est+2, n_est+4)
n_pos2 <- mapply(sim_res, r2, n2)
# Combine results of both steps for estimation
n12 <- c(n1,n2)
n_pos12 <- c(x, n_pos2)
```

```
lreg <- drm(n_pos12/r12~n12, weights=r12, type="binomial", logDose=exp(1),
 fct=LL2.3(fixed=c(NA,a,NA), names=c("r_e","a_e","n_e")) )
coef(lreg)
}
par_est <- apply(n_pos, 1, get_est)

outfile <- "err_param.txt"
inh_vals <- cbind(par_est[1,],par_est[2,])
write.table(inh_vals,row.names=F,col.names=F,file=outfile)
```

Appendix B

Equation derivations and source code for the bee foraging model

B.1 Calculation of F_{ST} from $C(s)$ under a random walk

Here I derive an analytical expression for the expected genetic differentiation (F_{ST}) between bee pollen loads collected under the most simple scenario from Chapter 3:

- each bee begins foraging at a random position within a continuous field of plants and performs a random walk with n_s steps of RMS (root mean square) length l_s;
- the marker of interest is present at an average frequency ϕ and follows a spatial correlation function $C(s)$ (see eq. 3.1); and
- the marker frequency in the pollen load is equal to the fraction of marked plants among those visited.

If x_i is the position of the plant visited at step i, the last assumption can be expressed as:

$$p = \frac{1}{n_s} \sum_{i=1}^{n_s} I(x_i). \tag{B.1}$$

Inserting this expression in the definition of F_{ST}, I get:

$$F_{ST} = \frac{\overline{p^2} - \overline{p}^2}{\overline{p}(1-\overline{p})} = \frac{\frac{1}{{n_s}^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} \mathbf{E}[I(x_i)I(x_j)] - \phi^2}{\phi(1-\phi)}. \tag{B.2}$$

Note that $\overline{p} = \phi$ follows from the first assumption above, which results in all plants having the same chance of being visited. The last equation is equivalent to:

$$F_{ST} = \frac{1}{{n_s}^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} \left[\frac{\mathbf{E}[I(x_i)I(x_j)] - \phi^2}{\phi(1-\phi)} \right] = \frac{1}{{n_s}^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} C_{ij}, \tag{B.3}$$

where C_{ij} is the correlation in marker presence between plants i and j on the bee's path. Noting that $C_{ii} = 1$ and $C_{ij} = C_{ji}$, I re-express (B.3) as:

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2} \sum_{i=1}^{n_s-1} \sum_{j=i+1}^{n_s} C_{ij}. \tag{B.4}$$

Since I assume knowledge of $C(s)$, the correlation in marker presence between two points separated by a distance s, I can calculate C_{ij} by taking the expected value of $C(s)$ over the distribution of S_{ij}, the distance between the i^{th} and j^{th} plants visited:

$$C_{ij} = \int_0^\infty C(s) \Pr\big(S_{ij} = s\big) \mathrm{d}s. \tag{B.5}$$

In the case of a simple random walk, displacements in x and y are normally distributed at each step: $\Delta X \sim \mathcal{N}\big(0, \frac{1}{2}{l_s}^2\big)$ and $\Delta Y \sim \mathcal{N}\big(0, \frac{1}{2}{l_s}^2\big)$. As a result, the displacement in each direction between steps i and j (with $j > i$) is distributed as $\mathcal{N}\big(0, \frac{1}{2}(j-i){l_s}^2\big)$, which means that S_{ij} follows a Rayleigh distribution:

$$\Pr\big(S_{ij} = s\big) = \frac{2s}{(j-i){l_s}^2} \exp\left(-\frac{s^2}{(j-1){l_s}^2}\right). \tag{B.6}$$

Combining eqs. (B.4) to (B.6), I obtain:

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2} \sum_{i=1}^{n_s-1} \sum_{j=i+1}^{n_s} \int_0^\infty C(s) \frac{2s}{(j-i){l_s}^2} \exp\left(-\frac{s^2}{(j-1){l_s}^2}\right) \mathrm{d}s, \tag{B.7}$$

which by setting $k = j - i$ simplifies to:

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2} \sum_{k=1}^{n_s-1} (n_s - k) \int_0^\infty C(s) \frac{2s}{k{l_s}^2} \exp\left(-\frac{s^2}{k{l_s}^2}\right) \mathrm{d}s. \tag{B.8}$$

I append below the results of integrating eq. (B.8) for several forms of $C(s)$ that will be used in Chapter 3.[1]

Exponential correlation:

$$C(s) = \exp\left(-\frac{s}{a}\right)$$

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2} \sum_{k=1}^{n_s-1} (n_s - k) \left[1 - \frac{\sqrt{\pi k} l_s}{2a} \exp\left(\frac{k{l_s}^2}{4a^2}\right) \left(1 - \operatorname{erf}\left(\frac{\sqrt{k} l_s}{2a}\right)\right)\right]$$

[1] Integrals were calculated using the Wolfram Mathematica online integrator (integrals.wolfram.com). In all equations, $\operatorname{erf}(x)$ refers to the Gaussian error function.

Linear correlation:

$$C(s) = \begin{cases} 1 - \frac{s}{2a} & \text{if } s \leq 2a, \\ 0 & \text{if } s > 2a. \end{cases}$$

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2} \sum_{k=1}^{n_s - 1} (n_s - k) \left[1 - \frac{\sqrt{\pi k} l_s}{4a} \operatorname{erf}\left(\frac{2a}{\sqrt{k} l_s} \right) \right]$$

Quadratic correlation:

$$C(s) = \begin{cases} as^2 + bs + c & \text{if } s \leq s_0, \\ 0 & \text{if } s > s_0. \end{cases}$$

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2} \sum_{k=1}^{n_s - 1} (n_s - k) \left[c + \left(1 - \exp\left(-\frac{{s_0}^2}{k {l_s}^2} \right) \right) a k {l_s}^2 + \frac{b}{2} \sqrt{\pi k} l_s \operatorname{erf}\left(\frac{s_0}{\sqrt{k} l_s} \right) \right]$$

B.2 Derivation of equation 3.9

In this section, I derive an expression for F_{ST} in a fragmented landscape model with n_f fields, where field $i*$ is filled with marked plants ($\phi = 1$) and all other fields are filled with unmarked plants ($\phi = 0$). Field-to-field movement is modelled as a Markov process: the probability of a bee starting to forage in field j is s_j, while the probability of moving to field j after visiting a plant in field i is $T_{i,j}$. As previously, I assume that the marker frequency in a bee pollen load, p, corresponds to the fraction of plants visited that bear the marker of interest.

First, I express F_{ST} in terms of correlations between successive steps on the bee's path (see eqs. B.3 and B.4 above):

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2} \sum_{k=0}^{n_s - 2} \sum_{l=k+1}^{n_s - 1} \frac{\mathbf{E}[I(x_k) I(x_l)] - \overline{p}^2}{\overline{p}(1 - \overline{p})}, \tag{B.9}$$

where $\overline{p}$ can be calculated from eq. (3.8). Note that the first plant visited is indexed by 0 (instead of 1) here, to match the notation for Markov chains.

Since $I(x)$ is an indicator function, the product $I(x_k)I(x_l)$ is equal to 1 if both the k^{th} and l^{th} plants have the marker and 0 otherwise. Therefore, its expected value corresponds to the probability that both plants are located in field $i*$:

$$\mathbf{E}[I(x_k) I(x_l)] = \Pr\big(i* \text{ at step } k \text{ AND } i* \text{ at step } l \big), \tag{B.10}$$

or using conditional probabilities:

$$\mathbf{E}[I(x_k) I(x_l)] = \Pr\big(i* \text{ at step } k \big) \Pr\big(i* \text{ at step } l \mid i* \text{ at step } k \big). \tag{B.11}$$

Both terms in eq. (B.11) can be calculated from the vector of initial probabilities s and the powers of the transition matrix T:

$$\mathbf{E}[I(x_k)I(x_l)] = \left(\sum_{j=1}^{n_f} s_j (T^k)_{j,i*}\right) (T^{l-k})_{i*,i*}. \tag{B.12}$$

Inserting this result into (B.9) completes the derivation:

$$F_{ST} = \frac{1}{n_s} + \frac{2}{{n_s}^2 \overline{p}(1-\overline{p})} \sum_{k=0}^{n_s-2} \sum_{l=k+1}^{n_s-1} \left[\sum_{j=1}^{n_f} s_j (T^k)_{j,i*} (T^{l-k})_{i*,i*} - \overline{p}^2\right]. \tag{B.13}$$

B.3 Bee foraging simulations in Fortran

In this section, I present the Fortran 90 source code used to generate the simulation results of Chapter 3. There are three different programs appended below, but they all share the following subroutines for producing the steps of correlated or uncorrelated random walks.

```
SUBROUTINE ucw_step(dx,dy)
    USE nrtype
    IMPLICIT NONE
    REAL(DP), INTENT(OUT) :: dx, dy
    REAL(DP) :: u, v, s

    ! Generate a step (dx,dy) of random direction and unit average length
    ! uses the polar version of the Box-Muller transformation
    ! (R^2 = dx^2 + dy^2 is distributed khi^2(0.5)/2)

    do
        call random_number(u)
        call random_number(v)
        u = 2.0_dp * u - 1.0_dp
        v = 2.0_dp * v - 1.0_dp
        s = u**2 + v**2
        if (s > 0.0_dp .and. s < 1.0_dp) exit
    end do
    s = sqrt(-log(s)/s)
    dx = u * s
    dy = v * s

END SUBROUTINE ucw_step


SUBROUTINE crw_step(rho,dl,dtheta)
    USE nrtype
    IMPLICIT NONE
    REAL(DP), INTENT(IN) :: rho
    REAL(DP), INTENT(OUT) :: dl, dtheta
    REAL(DP) :: u, v

    ! Generates a step (length dl, turning angle dtheta) of a
    ! correlated random walk, with dl^2 ~ khi^2(2)/2 (average of 1)
    ! and rho is the parameter of the wrapped Cauchy distribution
    ! (see Bartumeus et al. 2005: http://esapubs.org/archive/ecol/E086/168/appendix-A.htm)
```

```
    call random_number(u)
    call random_number(v)
    dl = sqrt(-log(u))
    dtheta = 2.0_dp * atan( (1.0_dp-rho)/(1.0_dp+rho) * tan(PI_D*(v - 0.5_dp)) )

END SUBROUTINE crw_step
```

The first program (below) implements the kriging-based model, where bees forage in a continuous field characterized by a spatial genetic correlation function. To solve the matrix equations and obtain the kriging weights, I used the Cholesky decomposition algorithm from Press et al. (1996).

```
PROGRAM bee_krig
    USE nrtype
    IMPLICIT NONE

    INTEGER(I4B), PARAMETER :: FILENUM = 99
    INTEGER(I4B), PARAMETER :: NSIM = 10000 ! Number of simulated flights
    INTEGER(I4B), PARAMETER :: NMAX = 100 ! Max value for ns (steps per flight)
    LOGICAL, PARAMETER :: CRW = .false. ! Is the random walk correlated?

    INTEGER(I4B) :: ns, fj ! ns: number of steps in random walk, fj: jump frequency
    INTEGER(I4B) :: i, j
    REAL(DP) :: rnd, pr, fst, mean
    REAL(DP) :: p, ls, lj ! Average marker frequency (p); step length (ls) and jump length (lj) in random walk
    REAL(DP) :: a ! Length parameter in correlation function
    INTEGER(I4B), DIMENSION(1:NMAX) :: m ! Marker presence(1) or absence(0) at each sampling point
    REAL(DP), DIMENSION(1:NMAX) :: x, y, theta ! Coords. and incoming angle of each point in random walk
    REAL(DP), DIMENSION(1:NMAX, 1:NMAX) :: cor_m ! Correlation matrix between points (i,j) (upper triangular)
    REAL(DP), DIMENSION(1:NMAX, 1:NMAX) :: weights ! Matrix of kriging weights:
            ! the column weights(:,j) is the vector of weights at step j
    REAL(DP), DIMENSION(1:NSIM) :: p_sim ! Frequency of maker in each simulated load

    ! Calculates the frequency distribution of a genetic marker in bee pollen loads
    ! taken from a spatially correlated plant population.
    ! p is the average frequency of the marker, and the spatial correlation function is
    ! included in the subroutine calc_weights.
    ! Bee flights (number of points and distances between them) are modelled in subroutine sim_flight.

    ns = 40
    fj = 10 ! Jump frequency
    ls = 1.0_dp
    lj = 1.0_dp ! Jump length
    p = 0.25_dp
    cor_m = 0.0_dp
    a = 1.0_dp

    call random_seed
    open(unit=FILENUM, file='beesim.csv', action='write')

    do i = 1, NSIM

        call sim_flight     ! Obtain correlation matrix
        call calc_weights   ! Compute kriging weight matrix
        ! First plant is (+) with probability p
        call random_number(rnd)
        if (rnd < p) then
            m(1) = 1
        else
            m(1) = 0
        end if
```

```
        do j = 1, ns-1
            ! Calculate probability of (+) by kriging
            pr = sum(weights(1:j,j)*real(m(1:j),DP)) + (1.0_dp - sum(weights(1:j,j)))*p
            call random_number(rnd)
            if (rnd < pr) then
                m(j+1) = 1
            else
                m(j+1) = 0
            end if
        end do

        p_sim(i) = sum(m(1:ns))/real(ns,DP)
        write(FILENUM,*) p_sim(i)

    end do

    mean = sum(p_sim(1:NSIM))/NSIM
    fst = (sum(p_sim(1:NSIM)**2)/NSIM - mean**2) / (mean*(1.0_dp - mean))
    write(*,*) fst
    close(FILENUM)

CONTAINS

    SUBROUTINE sim_flight
        USE nrtype
        IMPLICIT NONE
        INTERFACE
            FUNCTION corr(d)
            USE nrtype
            IMPLICIT NONE
            REAL(DP), INTENT(IN) :: d
            REAL(DP) :: corr
            END FUNCTION corr

            SUBROUTINE ucw_step(dx,dy)
            USE nrtype
            IMPLICIT NONE
            REAL(DP), INTENT(OUT) :: dx, dy
            END SUBROUTINE ucw_step

            SUBROUTINE crw_step(rho,dl,dtheta)
            USE nrtype
            IMPLICIT NONE
            REAL(DP), INTENT(IN) :: rho
            REAL(DP), INTENT(OUT) :: dl, dtheta
            END SUBROUTINE crw_step
        END INTERFACE

        INTEGER(I4B) :: i, j
        REAL(DP) :: dx, dy, dl, dtheta, rho, rnd

        x(1) = 0.0_dp
        y(1) = 0.0_dp
        theta(1) = 0.0_dp
        rho = 0.7_dp
        cor_m(1,1) = 1.0_dp

        do j = 2, ns

            if (CRW) then
                ! Generate a correlated random walk step (unit average length)
                call crw_step(rho,dl,dtheta)
                theta(j) = theta(j-1) + dtheta
                ! Bring theta back in (-pi, pi) if necessary
                if (abs(theta(j)) > PI_D) theta(j) = theta(j) - sign(TWOPI_D, theta(j))
```

```
            x(j) = x(j-1) + ls * dl * cos(theta(j))
            y(j) = y(j-1) + ls * dl * sin(theta(j))
        else
            ! Generate an uncorrelated random walk step (unit average length)
            call ucw_step(dx,dy)
            ! Every fj steps, take a jump (lj)
            call random_number(rnd)
            if ( mod(j,fj) == 0 ) then
                x(j) = x(j-1) + lj * dx
                y(j) = y(j-1) + lj * dy
            else
                x(j) = x(j-1) + ls * dx
                y(j) = y(j-1) + ls * dy
            end if
        end if

        do i = 1, j
            cor_m(i,j) = corr(sqrt((x(i)-x(j))**2 + (y(i)-y(j))**2))
        end do

    end do

END SUBROUTINE sim_flight

FUNCTION corr(d)
    USE nrtype
    IMPLICIT NONE
    REAL(DP), INTENT(IN) :: d
    REAL(DP) :: corr

    ! Exponential correlation function
    corr = exp(-d/a)

    ! Linear
    ! corr = max( 1.0_dp - d/(2.0_dp*a), 0.0_dp )

END FUNCTION corr


SUBROUTINE calc_weights
    USE nrtype; USE nr ! Module nr includes the Cholesky decomposition functions from Numerical Recipes
    IMPLICIT NONE

    INTEGER(I4B) :: j
    REAL(DP), DIMENSION(1:NMAX) :: b, diag

    ! The vector of weights after j steps, w = weights(1:j,j) is the solution of
    ! A * w = b, where A = cor_m(1:j,1:j) and b(i) = cor_m(i,j+1).
    ! Since cor_m is symmetric, we can use Cholesky decomposition.

    weights(1,1) = cor_m(1,2)
    do j = 2, ns-1
        b(1:j) = cor_m(1:j,j+1)
        call choldc(cor_m(1:j,1:j),diag(1:j))
        call cholsl(cor_m(1:j,1:j),diag(1:j),b(1:j),weights(1:j,j))
    end do

END SUBROUTINE calc_weights

END PROGRAM bee_krig
```

The next program implements the random patch model, where bees forage in continuous field with rectangular patches of marked plants.

```
PROGRAM bee_patch
    USE nrtype
    IMPLICIT NONE
    INTERFACE
        SUBROUTINE ucw_step(dx,dy)
        USE nrtype
        IMPLICIT NONE
        REAL(DP), INTENT(OUT) :: dx, dy
        END SUBROUTINE ucw_step

        SUBROUTINE crw_step(rho,dl,dtheta)
        USE nrtype
        IMPLICIT NONE
        REAL(DP), INTENT(IN) :: rho
        REAL(DP), INTENT(OUT) :: dl, dtheta
        END SUBROUTINE crw_step
    END INTERFACE
    INTEGER(I4B), PARAMETER :: NMAX = 1000, NPAIRS = 100000, NSIM = 10000, NSMAX = 100
    INTEGER(I4B), PARAMETER :: FILENUM = 99
    LOGICAL, PARAMETER :: CRW = .true.  ! True: correlated random walk, false: uncorrelated
    LOGICAL, PARAMETER :: ALTERN = .true. ! Alternates patch orientation (horizontal/vertical) if true

    INTEGER(I4B) :: i, j, k, np, ns, fj, total
        ! np: number of patches, ns: number of steps in random walk, fj: jump frequency
    REAL(DP) :: mean, var, corr, fst
    REAL(DP) :: fs, pxs, pys  ! Field side (fs x fs), patch side (pxs x pys)
    REAL(DP) :: x1, y1, x2, y2 ! Coords for random points
    REAL(DP) :: ls, lj, rho ! Parameters for bee flights
    REAL(DP) :: xb, yb, theta, xtry, ytry, theta_try, dx, dy, dl, dtheta ! Coords for bee flights
    REAL(DP), DIMENSION(1:NMAX) :: xp, yp  ! (x,y) Position for left-hand corner of each patch
    REAL(DP), DIMENSION(1:NSIM) :: p_sim
    INTEGER(I4B), DIMENSION(0:NSMAX) :: nm_hist  ! Histogram of nm (number of marked plants visited)
    REAL(DP), DIMENSION(18) :: dval = (/ 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0, &
            6.0, 7.0, 8.0, 9.0, 10.0, 12.0, 14.0, 16.0, 18.0, 20.0 /)
        ! Values of d to calculate empirical correlation

    ns = 40
    fj = 10
    ls = 1.0_dp
    lj = 1.0_dp
    fs = 100.0_dp
    pxs = 10.0_dp
    pys = 10.0_dp
    np = 25
    rho = 0.7

    do k = 0, ns
        nm_hist(k) = 0
    end do

    ! Global mean and variance for marker presence
    mean = REAL(np,DP) * pxs * pys / fs**2
    var = mean*(1.0_dp - mean)

    ! Set up random field
    call random_seed
    do i = 1, np
        if (ALTERN .and. mod(i,2) == 0) then    ! switch rectangle orientation for even patches
            pick_xy: do
                call random_number(xp(i))
                call random_number(yp(i))
                xp(i) = xp(i) * (fs - pys)
                yp(i) = yp(i) * (fs - pxs)
                do j = 1, i-1
                    if (overlap(i,j)) cycle pick_xy
                end do
```

```
            exit pick_xy
        end do pick_xy
    else
        pick_xy: do
            call random_number(xp(i))
            call random_number(yp(i))
            xp(i) = xp(i) * (fs - pxs)
            yp(i) = yp(i) * (fs - pys)
            do j = 1, i-1
                if (overlap(i,j)) cycle pick_xy
            end do
            exit pick_xy
        end do pick_xy
    end if
end do

! For each distance in dval, calculate correlation from random NPAIRS
open(unit=FILENUM, file='fieldcorr.csv', action='write')
do i = 1, size(dval)
    corr = 0.0_dp
    do j = 1, NPAIRS
        call random_number(x1)
        call random_number(y1)
        x1 = x1 * fs
        y1 = y1 * fs
        do
            call random_number(theta)
            theta = theta * 2.0_dp * PI
            x2 = x1 + dval(i) * cos(theta)
            y2 = y1 + dval(i) * sin(theta)
            if (    x2 > 0.0 .and. x2 < fs &
             .and.  y2 > 0.0 .and. y2 < fs ) exit
        end do
        corr = corr + value(x1,y1) * value(x2,y2)
    end do
    corr = (corr/real(NPAIRS,DP) - mean**2)/var
    write(FILENUM,*) dval(i), corr
end do
close(FILENUM)

! Simulate NSIM bee flights in field
do i = 1, NSIM
    call random_number(xb)
    call random_number(yb)
    call random_number(theta)
    xb = xb * fs
    yb = yb * fs
    theta = theta * TWOPI_D - PI_D
    total = value(xb,yb)    ! Total tracks the # of marked plants visited

    do j = 2, ns
        if (CRW) then
            call crw_step(rho, dl, dtheta)
            theta_try = theta + dtheta
            if (abs(theta_try) > PI_D) then ! Bring theta back in (-pi, pi) if necessary
                theta_try = theta_try - sign(TWOPI_D, theta_try)
            end if
            xtry = xb + ls * dl * cos(theta_try)
            ytry = yb + ls * dl * sin(theta_try)
            ! If CRW step goes outside field, try uncorrelated random walk
            if ( xtry < 0.0 .or. xtry > fs &
            .or. ytry < 0.0 .or. ytry > fs ) then
                do
                    call ucw_step(dx, dy)
                    xtry = xb + ls * dx
                    ytry = yb + ls * dy
```

```
                    if (    xtry > 0.0 .and. xtry < fs &
                     .and.  ytry > 0.0 .and. ytry < fs ) exit
                end do
            end if
            xb = xtry
            yb = ytry
            theta = theta_try
            total = total + value(xb,yb)
        else
            do
                call ucw_step(dx, dy)

                ! Every fj steps, take a jump (lj)
                if (mod(j,fj) == 0) then
                    xtry = xb + dx * lj
                    ytry = yb + dy * lj
                else
                    xtry = xb + dx * ls
                    ytry = yb + dy * ls
                end if
                if (    xtry > 0.0 .and. xtry < fs &
                .and.  ytry > 0.0 .and. ytry < fs ) exit
            end do
            xb = xtry
            yb = ytry
            total = total + value(xb,yb)
        end if
    end do
    nm_hist(total) = nm_hist(total) + 1
    p_sim(i) = REAL(total,DP) / ns
end do

! Output mean marker frequency and FST to screen, histogram of nm to file
mean = sum(p_sim(1:NSIM))/NSIM
fst = (sum(p_sim(1:NSIM)**2)/NSIM - mean**2) / (mean*(1.0_dp - mean))
write(*,*) mean, fst
open(unit=FILENUM, file='fieldsim.csv', action='write')
do k = 0, ns
    write(FILENUM,*) k, nm_hist(k)
end do
close(FILENUM)

CONTAINS

FUNCTION overlap(i,j)
    USE nrtype
    IMPLICIT NONE
    INTEGER(I4B), INTENT(IN) :: i, j
    LOGICAL :: overlap

    ! Check if rectangles i and j overlap
    ! If ALTERN, even-numbered rectangles have pxs and pys switched

    overlap = .TRUE.
    if (ALTERN) then
        if ( mod(i,2) == 1 .and. mod(j,2) == 1 ) then
            if (   xp(i) + pxs < xp(j) .or. xp(j) + pxs < xp(i) &
              .or. yp(i) + pys < yp(j) .or. yp(j) + pys < yp(i) ) overlap = .FALSE.
        else if ( mod(i,2) == 0 .and. mod(j,2) == 1 ) then
            if (   xp(i) + pys < xp(j) .or. xp(j) + pxs < xp(i) &
              .or. yp(i) + pxs < yp(j) .or. yp(j) + pys < yp(i) ) overlap = .FALSE.
        else if ( mod(i,2) == 1 .and. mod(j,2) == 0 ) then
            if (   xp(i) + pxs < xp(j) .or. xp(j) + pys < xp(i) &
              .or. yp(i) + pys < yp(j) .or. yp(j) + pxs < yp(i) ) overlap = .FALSE.
        else
            if (   xp(i) + pys < xp(j) .or. xp(j) + pys < xp(i) &
```

```
                .or. yp(i) + pxs < yp(j) .or. yp(j) + pxs < yp(i) ) overlap = .FALSE.
            end if
        else
            if (   xp(i) + pys < xp(j) .or. xp(j) + pys < xp(i) &
              .or. yp(i) + pxs < yp(j) .or. yp(j) + pxs < yp(i) ) overlap = .FALSE.
        end if

END FUNCTION overlap


    FUNCTION value(xc,yc)
     USE nrtype
IMPLICIT NONE
        REAL(DP), INTENT(IN) :: xc, yc
        INTEGER(I4B) :: value, i

        ! Returns the value (0 or 1) at point (xc,yc)
        ! If ALTERN, even-numbered rectangles have pxs and pys switched
        value = 0
        do i = 1, np
            if (ALTERN .and. mod(i,2) == 0) then
                if (    xc > xp(i) .and. xc < xp(i) + pys &
                  .and. yc > yp(i) .and. yc < yp(i) + pxs ) then
                    value = 1
                    exit
                end if
            else
                if (    xc > xp(i) .and. xc < xp(i) + pxs &
                  .and. yc > yp(i) .and. yc < yp(i) + pys ) then
                    value = 1
                    exit
                end if
            end if
        end do

    END FUNCTION value

END PROGRAM bee_patch
```

Finally, the program below is used to simulate bee foraging in the fragmented landscape of section 3.4.2.

```
PROGRAM bee_landscape
    USE nrtype
    IMPLICIT NONE
    INTEGER(I4B), PARAMETER :: NFMAX = 10, NPMAX = 200, NSMAX = 500, NSIM = 10000
    INTEGER(I4B), PARAMETER :: FILENUM = 99
    LOGICAL, PARAMETER :: RAND_START = .true.   ! Whether or not bees start at random point in field
                                                ! if .false., they start at the point closest to hive

    INTEGER(I4B) :: nf, np, ns  ! nf: # of fields, np: # of patches (field 2), ns: # of steps in walk
    REAL(DP) :: xl, yl ! dimensions of landscape
    REAL(DP) :: xh, yh ! (x,y) coordinates of the hive
    REAL(DP) :: pxs, pys ! patch size
    REAL(DP) :: p_sw ! probability to switch field within a foraging bout
    REAL(DP) :: pmult ! Multiplier used to normalize probabilities of field switching
    INTEGER(I4B), DIMENSION(1:NSIM) :: nm  ! number of marked plants visited by each simulated bee
    INTEGER(I4B), DIMENSION(0:NSMAX) :: freq_nm ! Frequency of paths with nm marked plants
    REAL(DP), DIMENSION(1:NFMAX) :: phi_f ! marker frequency for each field
    REAL(DP), DIMENSION(1:NFMAX) :: score, cumul_sc ! score of each field and cumulative score
    REAL(DP), DIMENSION(1:NFMAX,1:5) :: xf, yf ! (x,y) coordinates of the 4 points in each field
                                               ! point 5 is a repeat of point 1 for geometric functions
```

```
    REAL(DP), DIMENSION(1:NPMAX) :: xp, yp ! (x,y) position for upper-left corner of each patch

    INTEGER(I4B) :: i
    REAL(DP) :: mean, var, fst, mean_m, var_m, fst_m
        ! '_m' statistics are calculated over mixed pollen loads only (i.e. excluding p=0 and p=1)

    open(unit=FILENUM, file='fieldsim_struct.csv', action='write')

    ns = 100
    p_sw = 0.0_dp
    call init_field
    call sim_bees

    ! Output histogram of nm and statistics
    mean = 0.0_dp
    var = 0.0_dp
    do i = 1, ns-1
        freq_nm(i) = count(nm(:) == i)
        mean = mean + freq_nm(i) * i
        var = var + freq_nm(i) * i**2
    end do
    freq_nm(0) = count(nm(:) == 0)
    freq_nm(ns) = count(nm(:) == ns)

    mean_m = mean / (sum(freq_nm(1:ns-1)) * ns)
    var_m = var / (sum(freq_nm(1:ns-1)) * ns**2) - mean_m**2
    fst_m = var_m / (mean_m * (1.0_dp - mean_m))

    mean = mean + freq_nm(ns) * ns
    var = var + freq_nm(ns) * ns**2
    mean = mean / (NSIM * ns)
    var = var / (NSIM * ns**2) - mean**2
    fst = var / (mean * (1.0_dp - mean))

    write(*,*) mean, fst
    write(*,*) mean_m, fst_m
    do i = 0, ns
        write(FILENUM,*) i, freq_nm(i)
    end do
    close(FILENUM)

CONTAINS

FUNCTION dist(x1,y1,x2,y2)
        USE nrtype
        IMPLICIT NONE
        REAL(DP), INTENT(IN) :: x1, y1, x2, y2
        REAL(DP) :: dist

        ! Distance between (x1,y1) and (x2,y2)
        dist = sqrt((x2-x1)**2 + (y2-y1)**2)

    END FUNCTION dist


    FUNCTION in_field(x,y,i)
        USE nrtype
        IMPLICIT NONE
        REAL(DP), INTENT(IN) :: x, y
        INTEGER(I4B), INTENT(IN) :: i
        LOGICAL :: in_field
        INTEGER(I4B) :: j

        ! Checks if point (x,y) is in field i
        in_field = .false.
        do j = 1, 4
```

```
        if ( (     ((yf(i,j)   <= y) .and. (y < yf(i,j+1)))  &
              .or. ((yf(i,j+1) <= y) .and. (y < yf(i,j))) )  &
             .and. ( x < xf(i,j) + (xf(i,j+1)-xf(i,j))*(y - yf(i,j))/(yf(i,j+1)-yf(i,j)) ) ) then
            in_field = .not. in_field
        end if
    end do

END FUNCTION in_field


FUNCTION dist_to_field(x,y,i)
    USE nrtype
    IMPLICIT NONE
    REAL(DP), INTENT(IN) :: x, y
    INTEGER(I4B), INTENT(IN) :: i
    REAL(DP) :: dist_to_field
    INTEGER(I4B) :: j
    REAL(DP) :: t, d, xproj, yproj

    ! Calculates the shortest distance between (x,y) and field i
    ! (assuming the point is outside the field)

    dist_to_field = xl*yl ! Initialize to some large value
    ! Calculate distance from point to each segment, keep smallest value
    do j = 1, 4
        t = (x - xf(i,j))*(xf(i,j+1) - xf(i,j)) + (y - yf(i,j))*(yf(i,j+1) - yf(i,j))  &
            / ( (xf(i,j+1) - xf(i,j))**2 + (yf(i,j+1) - yf(i,j))**2 )
        if (t <= 0) then
            d = dist(x,y,xf(i,j),yf(i,j))
        else if (t >= 1) then
            d = dist(x,y,xf(i,j+1),yf(i,j+1))
        else
            xproj = xf(i,j) + t*(xf(i,j+1) - xf(i,j))
            yproj = yf(i,j) + t*(yf(i,j+1) - yf(i,j))
            d = dist(x,y,xproj,yproj)
        end if
        dist_to_field = min(dist_to_field, d)
    end do

END FUNCTION dist_to_field


SUBROUTINE closest_point(x,y,i,xc,yc,theta)
    USE nrtype
    IMPLICIT NONE
    REAL(DP), INTENT(IN) :: x, y
    INTEGER(I4B), INTENT(IN) :: i
    REAL(DP), INTENT(OUT) :: xc, yc, theta
    INTEGER(I4B) :: j
    REAL(DP) :: t, d, dmin, xproj, yproj

    ! Finds the point (xc,yc) in field i that is closest to (x,y)
    ! assuming (x,y) is outside the field
    ! theta is the angle between (x,y) and (xc,yc)

    dmin = xl*yl ! Initialize to some large value
    ! Finds the segment closest to (x,y), pick projection on that segment or one of the two ends
    do j = 1, 4
        t = (x - xf(i,j))*(xf(i,j+1) - xf(i,j)) + (y - yf(i,j))*(yf(i,j+1) - yf(i,j))  &
            / ( (xf(i,j+1) - xf(i,j))**2 + (yf(i,j+1) - yf(i,j))**2 )
        if (t <= 0) then
            xproj = xf(i,j)
            yproj = yf(i,j)
        else if (t >= 1) then
            xproj = xf(i,j+1)
            yproj = yf(i,j+1)
```

```
        else
            xproj = xf(i,j) + t*(xf(i,j+1) - xf(i,j))
            yproj = yf(i,j) + t*(yf(i,j+1) - yf(i,j))
        end if

        d = dist(x,y,xproj,yproj)
        if (d < dmin) then
            dmin = d
            xc = xproj
            yc = yproj
        end if
    end do

    theta = atan2(yc - y, xc - x)

END SUBROUTINE closest_point


FUNCTION overlap(i,j)
    USE nrtype
    IMPLICIT NONE
    INTEGER(I4B), INTENT(IN) :: i, j
    LOGICAL :: overlap

    ! Check if rectangles i and j overlap
    if (   xp(i) + pxs < xp(j) .or. xp(j) + pxs < xp(i) &
      .or. yp(i) + pys < yp(j) .or. yp(j) + pys < yp(i) ) then
        overlap = .FALSE.
    else
        overlap = .TRUE.
    end if

END FUNCTION overlap


FUNCTION value(xc,yc)
    USE nrtype
    IMPLICIT NONE
    REAL(DP), INTENT(IN) :: xc, yc
    INTEGER(I4B) :: value, i

    ! Returns the value (0 or 1) at point (xc,yc), for field 2 only
    value = 0
    do i = 1, np
        if (    xc > xp(i) .and. xc < xp(i) + pxs &
          .and. yc > yp(i) .and. yc < yp(i) + pys ) then
            value = 1
            exit
        end if
    end do

END FUNCTION value


SUBROUTINE init_field
    USE nrtype
    IMPLICIT NONE
    INTEGER(I4B) :: i, j
    REAL(DP), DIMENSION(1:NFMAX) :: xc, yc ! centroid coords. for each field
    REAL(DP), DIMENSION(1:NFMAX,1:NFMAX) :: trans, ctrans ! transition probability of moving
                            ! from field i to field j, and cumulative probability

    ! Ramsay (2003) field setup
    nf = 8
    xl = 240 ; yl = 180
    xh = 108 ; yh = 108
```

```
    xf(1,:) = (/ 103, 104, 112, 125, 103 /) ; yf(1,:) = (/  78,  63,  63,  87,  78 /)
    xf(2,:) = (/ 160, 162, 168, 155, 160 /) ; yf(2,:) = (/  87,  87, 102, 102,  87 /)
    xf(3,:) = (/ 118, 133, 120, 111, 118 /) ; yf(3,:) = (/ 140, 153, 162, 155, 140 /)
    xf(4,:) = (/  59,  72,  77,  58,  59 /) ; yf(4,:) = (/ 101, 104, 136, 135, 101 /)
    xf(5,:) = (/  42,  63,  64,  44,  42 /) ; yf(5,:) = (/ 142, 139, 158, 158, 142 /)
    xf(6,:) = (/   7,   4,  24,  26,   7 /) ; yf(6,:) = (/ 117, 127, 133, 120, 117 /)
    xf(7,:) = (/  14,  31,  30,  14,  14 /) ; yf(7,:) = (/ 135, 138, 149, 144, 135 /)
    xf(8,:) = (/   1,  11,   9,   1,   1 /) ; yf(8,:) = (/ 156, 158, 175, 176, 156 /)
    phi_f(1:8) = (/ 1.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0 /)

    ! Place patches in field 2
    np = 180
    pxs = 0.25_dp
    pys = 0.25_dp
    call random_seed
    do i = 1, np
        pick_xy: do
            call random_number(xp(i))
            call random_number(yp(i))
            xp(i) = xp(i) * (13.0_dp - pxs) + 155.0_dp
            yp(i) = yp(i) * (15.0_dp - pys) + 87.0_dp
            if (.not. in_field(xp(i),yp(i),2)) cycle pick_xy
            do j = 1, i-1
                if (overlap(i,j)) cycle pick_xy
            end do
            exit pick_xy
        end do pick_xy
    end do

    ! Score proportional to inverse distance from hive
    score(1) = 1.0_dp / dist_to_field(xh,yh,1)
    cumul_sc(1) = score(1)
    do i = 2,nf
        score(i) = 1.0_dp / dist_to_field(xh,yh,i)
        cumul_sc(i) = cumul_sc(i-1) + score(i)
    end do
    score(1:nf) = score(1:nf) / cumul_sc(nf)
    cumul_sc(1:nf) = cumul_sc(1:nf) / cumul_sc(nf)

    ! Transition probability between fields prop. to inverse dist. between centroids
    ! However, will be recomputed during simulations based on exact position of bee and closest point
    do i = 1, nf
        trans(i,i) = 0.0_dp
        xc(i) = sum(xf(i,1:4))/4
        yc(i) = sum(yf(i,1:4))/4
        do j = 1, i-1
            trans(i,j) = 1.0_dp / dist(xc(i),yc(i),xc(j),yc(j))
            trans(j,i) = trans(i,j)
        end do
    end do

    pmult = 1.0_dp / maxval( sum(trans(1:nf,1:nf), dim=2) )
    trans(1:nf,1:nf) = trans(1:nf,1:nf) * pmult
    do i = 1, nf
        ctrans(i,1) = trans(i,1)
        do j = 2, nf
         ctrans(i,j) = ctrans(i,j-1) + trans(i,j)
        end do
    end do

END SUBROUTINE init_field


SUBROUTINE sim_bees
    USE nrtype
    IMPLICIT NONE
```

```
INTERFACE
    SUBROUTINE ucw_step(dx,dy)
    USE nrtype
    IMPLICIT NONE
    REAL(DP), INTENT(OUT) :: dx, dy
    END SUBROUTINE ucw_step

    SUBROUTINE crw_step(rho,dl,dtheta)
    USE nrtype
    IMPLICIT NONE
    REAL(DP), INTENT(IN) :: rho
    REAL(DP), INTENT(OUT) :: dl, dtheta
    END SUBROUTINE crw_step
END INTERFACE
INTEGER(I4B) :: i, j, k, cur_f     ! cur_f is current field
REAL(DP) :: rnd, rnd2
REAL(DP) :: rx1, rx2, ry1, ry2 ! (x,y) coordinates of a rectangle surrounding field
REAL(DP) :: xb, yb, theta ! current coords. and orientation of bee
REAL(DP) :: ls ! step length
REAL(DP) :: rho, dl, dtheta, dx, dy ! CRW and UCW variables
REAL(DP) :: x_try, y_try, theta_try ! Attempted next (x,y,theta) step
REAL(DP), DIMENSION(1:NFMAX) :: trans, ctrans ! transition prob. of moving to field i, and cumulative

call random_seed

do i = 1, NSIM

    ! Pick first field
    call random_number(rnd)
    do k = 1, nf
        if (rnd <= cumul_sc(k)) then
            cur_f = k
            exit
        end if
    end do

    if (RAND_START) then

        ! Pick bee starting point within field
        rx1 = minval(xf(cur_f,1:4))
        rx2 = maxval(xf(cur_f,1:4))
        ry1 = minval(yf(cur_f,1:4))
        ry2 = maxval(yf(cur_f,1:4))
        do
            call random_number(xb)
            call random_number(yb)
            xb = xb * (rx2 - rx1) + rx1
            yb = yb * (ry2 - ry1) + ry1
            if ( in_field(xb, yb, cur_f) ) exit
        end do
        ! Pick random theta in (-PI, PI)
        call random_number(theta)
        theta = theta * TWOPI_D - PI_D

    else  ! Start at closest point

        call closest_point(xh, yh, cur_f, xb, yb, theta)

    end if

    ! Forage ns steps
    rho = 0.7_dp
    ls = 0.05_dp
    nm(i) = 0
    do j = 1, ns
```

```
                ! Take step
                call crw_step(rho, dl, dtheta)
                theta_try = theta + dtheta
                if (abs(theta_try) > PI_D) then ! Bring theta back in (-pi, pi) if necessary
                    theta_try = theta_try - sign(TWOPI_D, theta_try)
                end if
                x_try = xb + ls * dl * cos(theta_try)
                y_try = yb + ls * dl * sin(theta_try)
                ! If CRW step goes outside field, try uncorrelated random walk
                if ( .not. in_field(x_try, y_try, cur_f) ) then
                    do
                        call ucw_step(dx, dy)
                        x_try = xb + ls * dx
                        y_try = yb + ls * dy
                        if ( in_field(x_try, y_try, cur_f) ) exit
                    end do
                end if
                xb = x_try
                yb = y_try
                theta = theta_try

                ! Find plant genotype
                if (cur_f == 2) then
                    nm(i) = nm(i) + value(xb,yb)
                else
                    call random_number(rnd)
                    if (rnd <= phi_f(cur_f)) nm(i) = nm(i) + 1
                end if

                ! Field change
                call random_number(rnd)
                if (rnd <= p_sw) then
                    ! Recalculate trans based on current position of bee
                    do k = 1, nf
                        if (k .ne. cur_f) then
                            call closest_point(xb, yb, k, x_try, y_try, theta_try)
                            trans(k) = pmult / dist(xb, yb, x_try, y_try)
                        end if
                    end do
                    trans(cur_f) = 0.0_dp
                    ctrans(1) = trans(1)
                    do k = 2, nf
                        ctrans(k) = ctrans(k-1) + trans(k)
                    end do

                    ! Pick new field
                    call random_number(rnd2)
                    do k = 1, nf
                        if (rnd2 <= ctrans(k)) then
                            cur_f = k
                            call closest_point(xb, yb, k, x_try, y_try, theta_try)
                            xb = x_try
                            yb = y_try
                            theta = theta_try
                            exit
                        end if
                    end do
                end if

            end do

        end do

    END SUBROUTINE sim_bees

END PROGRAM bee_landscape
```