Estimation of the Seed Dispersal Distribution with Genotypic Data

by

Yimei Cai

(Under the direction of Jaxk Reeves)

## Abstract

The goal of this dissertation is to formulate statistically appropriate estimators for the seed dispersal distribution function when data are collected in traps from multiple sources, both with and without genotyping, after accounting for other relevant factors, such as tree fecundity. Work along these lines has been attempted previously, under idealistic assumptions, for non-genotyped data, but this dissertation will give more general and practical results for this case. There has been almost no statistical work done on the genotyped case, so such results will be new and useful. Finally, and most important for ecologists who will in coming years have much data of these types, we propose to find a statistically appropriate estimator for the seed dispersal distribution function when the data consist of a combination of non-genotyped and genotyped seeds.

Index words: Seed dispersal, Genotyping errors, Inverse modeling

# Estimation of the Seed Dispersal Distribution with Genotypic Data

by

Yimei Cai

B.S., Peking University, China, 2000

M.S., The University Of Georgia, 2002

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2007

# Estimation of the Seed Dispersal Distribution with Genotypic Data

by

Yimei Cai

Approved:

Major Professor: Jaxk Reeves

Committee: Steve Hubbell
Nicole Lazar
William McCormick
Paul Schliekelman

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2007

# Acknowledgments

This dissertation could not have been written without the support and friendship found at University of Georgia and elsewhere. The love of family and friends provided my inspiration and was my driving force. It has been a long journey and completing this work is definitely a high point in my academic career. I could not have come this far without the assistance of many individuals and I want to express my deepest appreciation to them.

My first, and most earnest, acknowledgment must go to my major professor Jaxk Reeves. I am grateful to Dr. Reeves not only for his technical advice, but also for the encouragement and support he gave me when they were most needed, not to mention his patience with a student who took too long to finish her research. I would like to thank Professor Steve Hubbell, Professor Nicole Lazar, Professor William McCormick and Professor Paul Schliekelman for serving on my committee and for their helpful comments and suggestions.

Finally, I wish to thank my family and friends. The people I have met while in graduate school have become my closest and dearest friends and counselors, and to all of you I give my love and thanks. My parents and older brother have always believed in me and helped me reach my goals. Their support forged my desire to achieve all that I could in life. I owe them everything and wish I could show them just how much I love and appreciate them. Finally, my husband, Wei Zhao, whose love and encouragement allowed me to finish this journey, already has my heart so I will just give him a heartfelt 'thanks'.

# Table of Contents

## List of Figures

## List of Tables

# Chapter 1

## Introduction

Scientists have been studying the seed dispersal distribution for certain trees and plants for almost 100 years. Clearly, the distribution of such seeds is an important determinant of the rate of spread of a species in an area, and this information would be of considerable use to botanists, agronomists, ecologists, and others. Other things being equal, plants which can disperse their seeds over a wider range of area would have a greater chance of passing on their genetic characteristics than would plants with a smaller average dispersal area. Of course, 'other things' are rarely equal; other important characteristics of species spread include the total number of seeds produced by a plant, the typical viability of the dispersed seeds, the ability of the dispersed seeds to settle in places favorable for growth, competition from other plants, and the effects of further dispersal by animals, birds, or other means. Indeed, the 'natural' seed dispersal which we discuss here corresponds only to the small loop at the bottom center of the complete seed dispersal cycle, as described by Wang and Smith [22] and displayed in Figure 1.1.

Nonetheless, understanding the 'natural' seed dispersion characteristics of a plant species is certainly a fundamental step in understanding that species' spread history. Unfortunately, estimating seed dispersal distributions has proven to be much more difficult than one might have imagined. In the four sub-sections below, we describe four common situations in which dispersal distributions might be estimated. Except for the first case, however, the estimation is not simple, and there is a lack of consensus among experts concerning how to analyze results. The four cases examined are: (1) Single source with complete collection, (2) Single source with sampling, (3) Multiple sources with seed traps, and (4) Multiple sources with

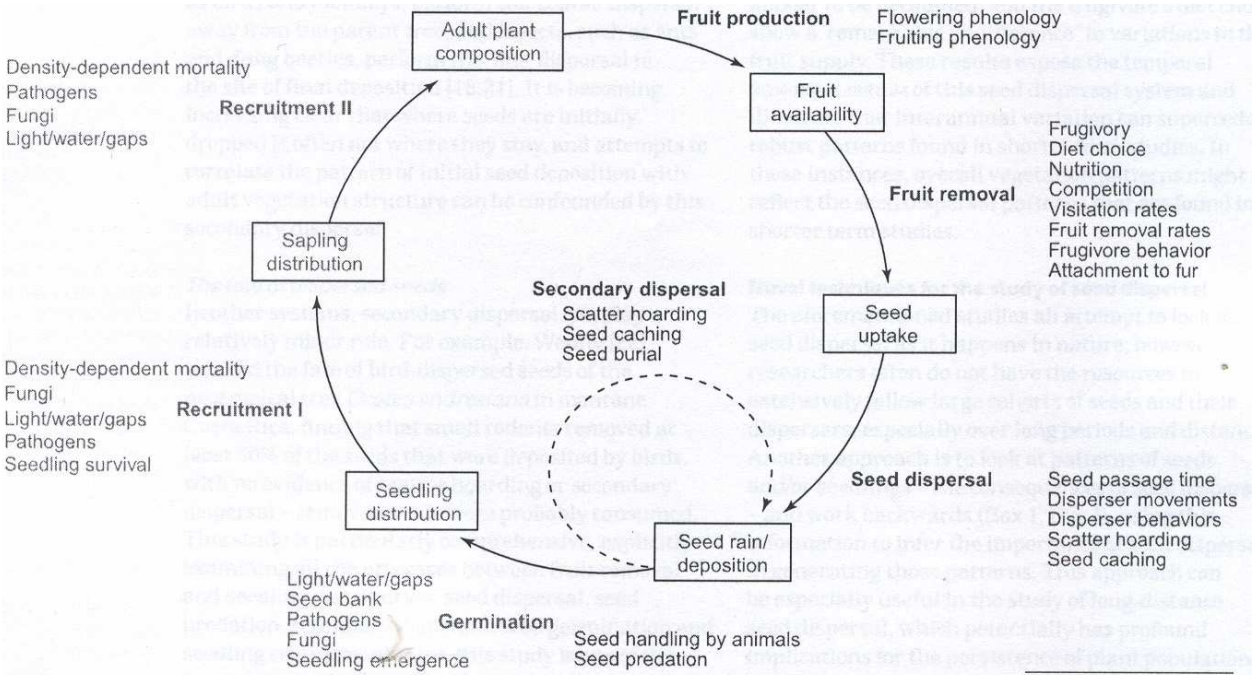


Figure 1.1: The Seed Dispersal Cycle

seed traps and genotyping. In all four cases, results could be obtained (and estimates made) from one season's worth of data, but more reliable estimates would arise from combining the results of several seasons of data, assuming that one believed that the general form of the dispersion distribution for a particular species was the same from year to year. In fact, as we shall see, this assumption is not always tenable for many real-life data sets.

## 1.1 Single Source with Complete Collection

This is the simplest, although somewhat unrealistic, method from which to estimate the natural seed dispersal distribution. A seed-producing plant or tree of the species of interest is planted (or observed by chance) in an area where no other plants or trees of this type are located. Every seed dispersed by this tree or plant during the seed dispersal season (most trees and plants disperse only during certain months of the year) is followed and its distance

from the source is recorded. For this method to be feasible, seeds must be large enough to be located by the naked eye, the terrain must be such that all seeds can be located easily, and one must be sure that the search area is large enough to contain all possible dispersed seeds from the tree or plant in question, but none from any neighboring plants or trees of this species. As one can easily imagine, the conditions necessary for such estimation are not at all likely to be met in nature, except for a few species which disperse almost all seeds (initially) at the base of the tree or plant in question. Even for such species, although the 'natural' dispersion function may be easy to estimate, the eventual dispersion (due to animals, birds, rain runoff, etc.), which is of much more practical interest, is incalculable. For most species, even obtaining the natural dispersion by this procedure is not practical, unless one were growing the plant in a greenhouse or other artificial climate, in which case the natural effects of wind would not likely be reproduced. If one were able to collect data of this sort, however, estimating the dispersal distribution would be relatively trivial. One could bin the data in some convenient units measuring distance from the base of each plant or tree and attempt to fit various two-dimensional parametric functions to the data. Unless one had strong reason to believe that there were a directional component to the data (for example, if there were a strong NW wind prevailing during the dispersal season), one would typically fit isotropic models; that is models which were a function of the distance, $r$, of a seed from its source, but not of direction. Clearly, more observations from different plants/trees of the same species within the same season or of the same plants/trees over different seasons would allow one to obtain tighter confidence intervals on the dispersal functions so estimated. If a parametric form were used, one could easily estimate both the cumulative distribution function for any distance $r$ from the base, $F(r)$, as well as the dispersal density, $f(r)$, at any point. Of course, one need not assume any parametric form at all, and one could still estimate $F(r)$ using the empirical cumulative distribution function. Examples of all of these methods are given in Section 3.3.

## 1.2 Single Source with Sampling

The major drawback of the 'Single Source, Complete Collection' method is the 'complete collection' assumption. In a natural environment, it is very hard to locate all of a tree/plant's seeds. One would usually sample in some systematic way. The best theoretical procedure for wind-dispersed seeds is to randomly sample seeds as they are dispersed, but unless one has some very accurate and fine-scale camera to record randomly selected seeds' flights paths, this will not be possible. What most investigators do in such a situation is 'line transect sampling'. A small strip (several inches to one foot wide) is followed from the source tree's base to a distance far from the tree, and all seeds which land in that strip are enumerated, along with their associated distance from the source tree's base. It is typical to construct four line transects, in the North, South, East, and West directions. From data collected in this manner, it is fairly easy to construct parametric distribution and density estimates in the manner described above. The situation is slightly more complex than above in that there are typically many fewer seeds present than under complete collection, and because one must weight the sample to account for the decreased proportional radial area surveyed as the transect progresses further from the base of the tree. This weighting, combined with typically small numbers of observed seeds far from the base of the tree, can sometimes cause instability in the estimation of the dispersal function, especially if non-parametric functions are used. Parametric functions are less subject to such instability, but should be checked to see if the parametric form assumed is compatible with the data obtained. Of course, as is always the case in evaluating tail behavior of distributions, unless the sample size is huge, there will be few observations available from which to validate tail behavior.

As with the complete collection method of Section 1.1, this estimation method is relatively straight-forward, but often infeasible. For it to work, one must be able to follow the transects far enough to insure that all possible seeds could be caught. This maximum possible distance is usually unknown, and is typically under-estimated. Some instability in long-distance estimates can be removed by using wedge-shaped transects rather than line

transects (so as to keep the radial angle constant), but this procedure will become unfeasible as the distance increases. In addition, as one increases this distance, one also increases the chance that the transect includes seeds from another nearby source. In a densely populated forest or other regions, it may be impossible to avoid such a situation, leading to the estimation strategy described next.

## 1.3 Multiple Sources with Seed Traps

In many natural settings, it is almost impossible to find sources that are sufficiently separated from one another so that seeds found on a line transect can be uniquely attributed to the nearest source. In such settings, researchers have turned to a different sampling method – seed trap networks. When such a plan is used, researchers typically carefully survey a designated plot of land, recording the location and size (DBH = diameter at breast height) of all plants or trees of interest. Later, a seed trap network is set up in the vicinity of the trees. Seeds are collected over a period of time (usually the entire dispersal season) from each trap. Of course, one does not know which seeds originated from which tree, but one can make certain assumptions about the dispersal function and the number of seeds produced per tree to estimate the dispersal function. This procedure, called 'inverse modeling' by ecologists, has been an area of intense research in recent years, and is discussed in the literature review section of this dissertation.

There are many factors which enter into the construction of inverse models, with some of the most fundamental being the following:

(a) Specification of a functional form for the dispersal function

(b) Assumptions made about the number of seeds produced per source tree

(c) Sampling corrections to account for placement of traps relative to trees

(d) Corrections for seeds which might have been dispersed from off-site trees

From reviewing the literature, it appears that those conducting research in this area have spent the majority of their time discussing topic (a), although it appears that corrections due to factors (b), (c), or (d) may completely dominate many proposed functional form improvements. As is the case with any of these methods, statistical verification for proposed models can be provided by demonstrating that the same model, with appropriate random error terms, will fit the data for several different seasons of data. Unfortunately, in many cases, this does not appear to be what happens, thus demonstrating inadequacies with many of these models. If no good general model can be found, or if the random component due to individual tree effects or season effects is too large, one might argue that it does not really make much sense to estimate any 'dispersal function' for a species. Sections 4.1-4.3 illustrate some of the above topics for a data set collected over 16 years from a biological research station operated in Panama by the Smithsonian Tropical Research Institute.

## 1.4 Multiple Sources with Seed Traps and Genotyping

The major difficulty encountered in progressing from single plant sources to multiple plant sources is that one does not know which seeds in a trap are associated with which source tree. However, in recent years, this situation has changed. Using micro-satellite DNA data, one can genetically identify (usually uniquely) all source trees within a certain area. Similarly, one can genotype the maternal tissue attached to trapped seeds and match them to sources, thus determining exactly how far each trapped seed traveled, and presumably improving upon the estimation of the dispersal distribution. Of course, many of the same problems noted above are still present, but perhaps to a lesser extent. For example, one must still deal with difficulties caused by the inclusion of seeds from off-site (non-genotyped) sources, but one now has an idea of how often this occurs.

The idea of using micro-satellite DNA to match seeds to sources is relatively new and has been done only a few times (as of 2007), but there is no doubt that it will become more common, especially as the cost of genotyping decreases and more technicians become

proficient at doing it. There are certainly difficulties associated with genotyping, among which are cost (in terms of time and money) to sample sources and seeds, decisions concerning which micro-satellite loci to use in identifying unique individuals, and difficulties involved with extracting DNA from seeds and matching it to the possible parent library. Nonetheless, this is clearly the future wave of seed dispersal research, and it behooves statisticians to become involved at an early stage in formulating procedures to use such data in seed dispersal distribution estimation.

## 1.5 Goals of Dissertation

A goal of this dissertation is to formulate statistically appropriate estimators for the seed dispersal distribution function when data are collected in traps from multiple sources, both with and without genotyping, after accounting for other relevant factors, as described above. Work along these lines has been attempted previously, under idealistic assumptions, for non-genotyped data, but this dissertation will give more general and practical results for this case. There has been almost no statistical work done on the genotyped case, so such results will be new and useful. Finally, and most important for ecologists who will in coming years have much data of these types, we propose to find a statistically appropriate estimator for the seed dispersal distribution function when the data consist of a combination of non-genotyped and genotyped seeds. We will illustrate our methods with two of the first data sets of this type ever collected, as explained in further detail in Section 3.2.

# Chapter 2

## Literature Review of Seed Dispersal Modeling

Dispersal influences many key aspects of plant biology, including population dynamics, evolution of population, metapopulation dynamics, biological invasions, and the dynamics and diversity of ecological communities (Cain et al. [2]). Understanding these effects requires descriptions of dispersal at local and regional scales and statistical models that permit estimation (Clark et al. [4]). Two challenges hinder prediction of dispersal within natural communities ([4]). The first is finding models which accurately describe dispersal across a range of spatial scales. The second is the development of statistical methods for estimation and model testing. Seed dispersal studies have often used curve-fitting techniques to estimate dispersal kernels based on seed collections made at known locations in the field (Portnoy and Willson [18]; Willson [23]; Nathan and Muller-Landau [16]). There are a number of methods available for determining seed dispersal curves, but by far the most economical is the inverse modeling approach pioneered by Ribbens et al. [19]. Under this approach, maximum liklihood methods are used to estimate the terms of the dispersal function. The inverse modeling approach has now been used in a number of different studies, but with disagreement among practitioners over the most appropriate functional form of the dispersal curve. Many functional forms have been used to describe how offspring abundances vary with distance from the parent tree. A general functional form which characterizes many of these distributions was introduced by Clark et al. [5] :

$$f(r) = \frac{1}{N} \exp\left[-\left(\frac{r}{\alpha}\right)^c\right] \tag{2.1}$$

where $r$ is the distance traveled, $\alpha$ is a dispersion parameter, $c$ is a dimensionless shape parameter, and $N$ is the normalization constant obtained by integrating arc-wise and with distance:

$$\begin{aligned} N &= \int_0^\infty \oint_{2\pi} \exp\left[-\left(\frac{r}{\alpha}\right)^c\right] r dr d\theta \\ &= 2\pi \int_0^\infty \exp\left[-\left(\frac{r}{\alpha}\right)^c\right] r dr = \frac{2\pi\alpha^2\Gamma(2/c)}{c}, \end{aligned}$$

where

$$\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} e^{-z} dz$$

is the gamma function. The kernel can be concave at the source and fat tailed ($c \leq 1$) or convex at the source and platykurtic ($c > 1$). This flexible density includes as special cases almost all common simple dispersal models with modes at the origin. The (2-dimensional) exponential corresponds to $c = 1$ (Willson [23])

$$f(r) = \frac{1}{2\pi\alpha^2} \exp\left[-\frac{r}{\alpha}\right], \tag{2.2}$$

with $c = 2$ corresponding to the 2-dimensional Gaussian Distribution

$$f(r) = \frac{1}{\pi\alpha^2} \exp\left[-\left(\frac{r}{\alpha}\right)^2\right]. \tag{2.3}$$

Clark et al. [5] used the Gaussian model (c=2), with fecundities assumed proportional to basal area, to fit a number of tree species in Southern Appalachian forests.

As we will see, for many populations it is clear that no simple functional form will work uniformly. The primary reason for this is that there is a certain proportion of seeds which appear to be dispersed at much greater distances than others. How to model this long distance dispersal (LDD) has become a major concern of those involved with inverse modeling. Clark [3] discussed the fact that some proportion of LDD is needed to account for the fast spread of species noted from palaeontological data. A kernel that accurately describes

dispersal at both local and long-distance scales is obtained by characterizing the seed shadow as a composite process, summarized by a continuous range of dispersal parameters $\alpha$. So a mixture of Gaussian (c=2) and 'fat tail' (c=0.5) can explain historical data, even with a very low mixing proportion. Of course, from a statistical point of view, estimating mixture distributions can become very tricky, with many combinations of functional forms and mixing functions yielding almost equivalent fits. Ribbens et al. [19] used a special case with $c = 3$ and a lower kurtosis to yield a simple functional form which might fit data observed in practice without the need to resort to mixture distributions. The mixture of Gaussian and 'fat tail' is a reasonable model for a restricted set of conditions. And it loosely fits field data for most of the tree species(Clark [3]). But the model is most sensitive to seeds dispersed over short distances, and it fails to describe sporadic seed dispersed over long distances: the tail of the kernel is essentially overlooked. So Clark et al. [4] modified the model of (2.1), so that a prior distribution for the dispersion parameter $\alpha$ is used, thus allowing more flexibility. With

$$A \equiv \frac{u}{\alpha^2}$$

where $A$ is gamma-distributed with shape parameter $p$:

$$f(A; p) = \frac{A^{p-1} e^{-A}}{\Gamma(p)},$$

the new kernel becomes

$$f(r) = \int_0^\infty f(r|A) f(A) dA = \frac{p}{\pi u \left[1 + \frac{r^2}{u}\right]^{p+1}}. \tag{2.4}$$

However, this so-called '2-dimensional-t'(2Dt) model, like all of the above, also assumes that the mode of the distribution occurs at the point of origination ($r = 0$). While this seems to be a reasonable assumption, there are many cases where it does not yield the best fit. The 2Dt is unstable at extreme values of p, and instability occurs when data are sparse (animal-dispersed types) and when data are of limited extent (Clark et al. [4]). Stoyan & Wagner [20] claimed the 2Dt was superior to the Weibull. Meanwhile, other authors have simply adopted one or another of these functions, intuiting, perhaps, that they will perform about equally

well. Nonetheless, the choice of the function is critical; as noted by Nathan & Muller-Landau [16], some functions have fat tails that are too thin to permit meta-population persistence.

A possibly serious deficiency of curve-fitting techniques, including inverse modeling methods (Ribbens et al. [19]; Clark et al. [5]; Clark et al. [4]), is that low frequency LDD events can be masked, and therefore underestimated, by the high frequency of short distance dispersal events (Turchin [21]). Failure to detect LDD events will generally cause one to underestimate the tail of the dispersal distribution. Use of molecular genetic markers is one of the new techniques that have great potential to facilitate direct measurement of actual dispersal (Wang and Smith [22]). The most beneficial aspect of using genotyped data is the fact that it will allow immediate identification of seeds which have dispersed far from their sources, greatly aiding in the estimation of the LDD component.

The development of molecular markers has provided the study of dispersal with new, potentially powerful tools (Ouborg [17]). There are several molecular markers, with differing degrees of variability, which, when analyzed, can yield different levels of resolution. The application of highly variable molecular markers, such as microsatellites, has facilitated the development of so-called 'direct' genetic methods. If all the potential parents in a plant population can be sampled, parentage analyses can be performed whereby the parents of individual seeds or seedlings can be determined. This technique, although potentially time and energy intensive, is extremely powerful – it provides a direct method of measuring individual dispersal events. Godoy and Jordano [10] performed this analysis for *Prunus mahaleb* trees and seeds. By sequencing the woody endocarp of the dispersed seeds and comparing those sequences to the genotypes of all the potential parent trees in the population, they were able to find unambiguous matches for 78 of the 95 seeds (82%) analyzed. Furthermore, their results indicate that strong distance limitation of seed delivery combined with infrequent long-distance dispersal events can cause extreme heterogeneity in the landscape pattern of genetic makeup, and a marked mosaic of multiple parentage for the seeds delivered to a particular patch. Even when it is not feasible to sample all potential parents, direct methods can

still be applied via assignment tests that assess the likelihood that an individual originated from each of the sampled source populations (Wang and Smith [22]).

Genotyping errors occur when the genotype determined after molecular analysis does not correspond to the real genotype of the individual under consideration. In practice, genotyping errors are defined as the differences observed between two or more molecular genotypes obtained independently from the same sample (Bonin et al. [1]). Virtually every genetic data set includes some erroneous genotypes. They can be generated at every step of the genotyping process and by a variety of factors. For some microsatellites, the main source of errors is allelic dropout (Constable et al. [7]; Jeffery et al. [12]; Creel et al. [8]), but human factors are non-negligible error generators. Therefore, tracking genotyping errors and identifying their causes is necessary to clean up data sets and validate the final results according to the precision required.

# Chapter 3

# Problem Description

## 3.1 Notation

As noted in Section 1.5, our focus in this dissertation will be on the general situation where there are multiple sources, with multiple traps collected over multiple time periods. This applies to the situation described in Sections 1.3 and 1.4 of this document, with the fundamental difference being that in Section 1.4, some proportion of the seeds collected in each trap is subjected to genetic analysis, so that matches to parents can be made. Let us use the following notation:

Let the index $i$, $i = 1, 2, ...I$ refer to the known sources in the area of collection.

Let the index $j$, $j = 1, 2, ...J$ refer to the traps in the area of collection.

Let the index $t$, $t = 1, 2, ...T$ refer to the seasons (usually years) over which data are collected.

Let the index $k$, $k = 1, 2, ...K(t)$ refer to the $k$th seed examined in season $t$. This index is relevant only when the seeds are genotyped.

The $(x, y)$ coordinates of each known source and trap are assumed to be known, so let $d_{ij}$ represent the Euclidean distance from source $i$ to trap $j$.

Let $S(j, t)$ represent the number of seeds caught in trap $j$ during season $t$.

If one is dealing with completely unambiguously genotyped seed data, then one can express

$$S(j,t) = \sum_{i=0}^{I} S(i,j,t) \tag{3.1}$$

where $S(i, j, t)$ represent the number of seeds caught in trap $j$ during season $t$, which have been genotyped to match source $i$. The $'i = 0'$ term reflects the seeds whose genotypes do

not match any of the known sources and are thus believed to have originated from unknown sources outside the study area.

## 3.2 Motivating Example

To illustrate the above in an actual example, consider the case of FDP (Forest Dynamics Plot), a biological research station on Barro Colorado Island (BCI), Panama operated by the Smithsonian Tropical Research Institute (STRI). Since 1985, a $50ha$ ($1000m \times 500m$) rectangular plot has been surveyed periodically to record all shrubs or trees whose stems are greater than 1 $cm$ diameter at breast height ('DBH') in this preserve. Each shrub or tree is identified by species, and various characteristics such as $(x, y)$ location coordinates, DBH, canopy cover, etc. are recorded. A complete census of the $50ha$ region was performed in 1982 and has been re-performed every 5 years from 1985 to 2005. Since 1986, a 200-trap trap-network has been established on the FDP site. Each trap is an $0.5m^2$ nylon net situated about $1.5m$ above the ground to collect wind-blown, bird-dispersed and some animal-dispersed seeds. The distribution of the traps throughout the $50ha$ area is shown in the plot in Figure 3.1. Note that the traps are not evenly spread throughout the region; they tend to be near trails for ease of collection. The traps are inspected weekly by employees of the FDP, who then sort the seeds by species. In the main example which is illustrated in this document, we consider seeds collected from *Jacaranda copaia* (Bignoniaceae) at the FDP site. The traps denoted by 'x' are 98 'gap-traps' that are not part of the offical FDP network. These were set up in 2000; their use will be explained later.

*Jacaranda copaia* is a large canopy tree (up to $45m$ tall) and is a characteristic species of Neotropical moist forests ranging from Belize to Brazil and Bolivia (Croat [9]). The small wind dispersed seeds ($< 2mg$) are produced in large woody capsules in the canopy of adult trees ($\geq 200mm$ DBH). Over the 16-year period for which data are available, 389 *Jacaranda* trees have been observed in the FDP, of which 236 achieved reproductive adult status. The population of *J. copaia* on the BCI FDP had 264 individuals $\geq 10mm$ DBH in the census of

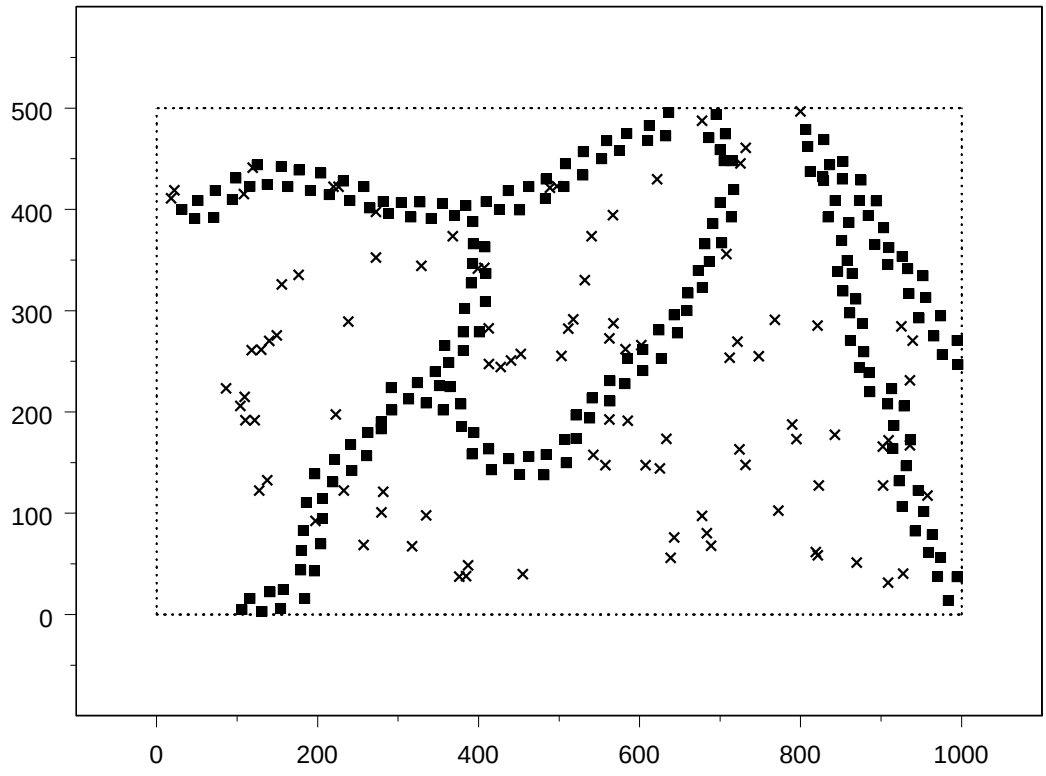


Figure 3.1: Plot of Traps on FDP (200 Network + 98 Gap Traps)

2000. *Jacaranda* shows a skewed size distribution with many large adults and a long lower tail containing a few small individuals. Such distributions typically occur in species that are shade-intolerant (Wright et al. [24]).

*J. copaia* trees thrive in the conditions present at FDP. On the $50ha$ site, according to the 2000 census, there were 264 *Jacaranda* trees, 199 of which were considered adult (DBH$>$ $200mm$). In 2000, there were another 91 adult *Jacaranda* trees that are not part of the official $50ha$ FDP site, but which are known to be located in a $100m$ 'buffer zone' surrounding the $50ha$ region. Figure 3.2 displays these trees with respect to the FDP and buffer zone. There appear to be fewer than 355 trees in the plot because of a number of high density clusters, which display on the plot as only 2-3 trees.

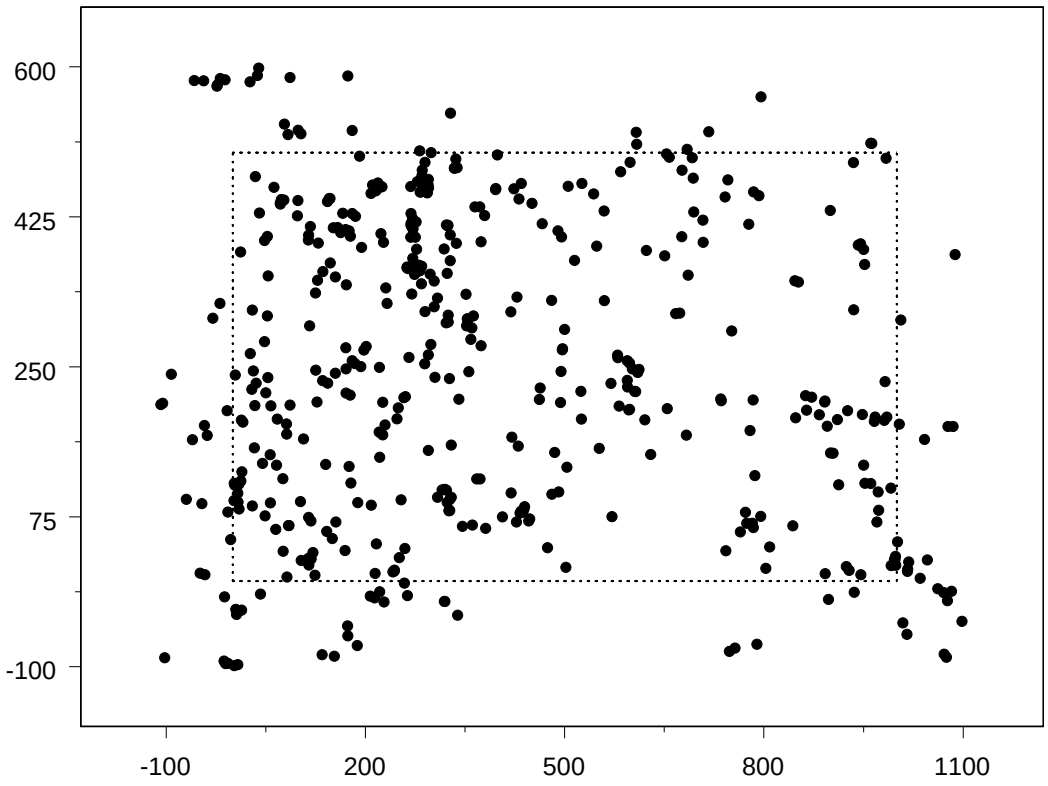


Figure 3.2: Plot of *Jacaranda* Trees on FDP and 100$m$-wide Buffer Zone

Of course, not all 290 adults in the FDP and buffer zone produce seeds every year, but many do so, as evidenced by the data shown in Table 3.1. This table displays the number of *Jacaranda* seeds collected by month in all 200 traps for the years 1987-2002. Several facts are immediately apparent. First, as is well known, *Jacaranda* seeds' dispersion occurs primarily from July to November, with the peak in September. This corresponds to the rainy season in Panama. Secondly, the total number of *Jacaranda* seeds collected varies tremendously from year to year. It seems almost as if there is a two-year cycle, consistent with behavior sometimes noted with tropical trees. In our analysis, we pooled the data into eight 2-year ('bi-year' periods) to smooth it out, although there is still much variability by bi-year as shown in Table 3.2. Finally, although obvious, it should be reiterated that the number of seeds

| month | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 00 | 01 | 02 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01 | . | 2 | 0 | 0 | 36 | 0 | 95 | 0 | 93 | 8 | 39 | 71 | 0 | 64 | 0 | 0 |
| 02 | . | 0 | 0 | 0 | 19 | 0 | 8 | 0 | 12 | 0 | 11 | 3 | 0 | 5 | 0 | 0 |
| 03 | . | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 0 | 9 | 0 | 4 | 0 | 1 | 0 |
| 04 | . | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| 05 | . | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 10 | 6 | 0 | 0 |
| 06 | 0 | 6 | 0 | 0 | 1 | 17 | 0 | 24 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 07 | 1 | 508 | 0 | 77 | 3 | 1786 | 0 | 1194 | 3 | 93 | 179 | 0 | 82 | 0 | 3 | 3 |
| 08 | 86 | 3245 | 7 | 1996 | 1 | 16004 | 0 | 7082 | 101 | 917 | 6751 | 32 | 1540 | 245 | 57 | 188 |
| 09 | 265 | 1488 | 145 | 3918 | 1 | 3984 | 0 | 4409 | 1697 | 3797 | 4186 | 28 | 7235 | 611 | 104 | 2994 |
| 10 | 2 | 51 | 114 | 781 | 0 | 1112 | 0 | 1125 | 384 | 2579 | 630 | 8 | 2330 | 214 | 18 | 661 |
| 11 | 5 | 5 | 0 | 224 | 10 | 669 | 0 | 629 | 12 | 356 | 183 | 0 | 595 | 42 | 1 | 116 |
| 12 | 1 | 1 | 0 | 74 | 0 | 124 | 0 | 115 | 8 | 112 | 272 | 0 | 98 | 3 | 0 | 15 |
| Total | 360 | 5307 | 266 | 7070 | 76 | 23696 | 103 | 14578 | 2311 | 7863 | 12264 | 143 | 11895 | 1190 | 184 | 3977 |

Table 3.1: Count of *Jacaranda* Seeds Collected on FDP by Month and Year(1987-2002)

collected per year is a small fraction of those produced on FDP, since the traps themselves cover only 0.2% of the entire FDP area.

| bi-year | Total Seeds |
|---|---|
| A 1987-1988 | 5667 |
| B 1989-1990 | 7396 |
| C 1991-1992 | 23815 |
| D 1993-1994 | 14684 |
| E 1995-1996 | 10130 |
| F 1997-1998 | 12359 |
| G 1999-2000 | 13072 |
| H 2001-2002 | 4161 |
| Total | 91286 |

Table 3.2: Total Number of *Jacaranda* Seeds Collected by Bi-year Period

The variation in seed numbers by traps is also large, even after pooling by bi-year. There are certain traps which tend to collect more seeds than others year after year, so one suspects that they are near trees which produce many seeds each bi-year. However, as shown in Table 3.3, the variation, both absolutely and relatively, is large. Table 3.3 displays the top 20 among the 200 traps over the 16-year period, where the traps are sorted by average rank over the 16-year period. Note that the rank by this sorting is not the same as the rank by total number of seeds collected, nor the ranking by sum of log(seeds), which correspond to

| Trap | 87-88 | 89-90 | 91-92 | 93-94 | 95-96 | 97-98 | 99-00 | 01-02 | Rank by Avg(Rank) | Rank by Avg(S) | Rank by Avg(log(S)) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 85 | 8.5 | 3 | 8 | 9 | 7 | 1 | 4 | 1 | 1 | 1 | 1 |
| 173 | 1 | 12 | 9 | 11 | 10 | 8 | 2 | 4 | 2 | 4 | 2 |
| 136 | 63.5 | 7 | 5 | 1 | 5 | 24 | 1 | 17 | 3 | 2 | 3 |
| 32 | 6 | 6 | 7 | 15 | 4 | 7 | 11 | 68 | 4 | 6 | 5 |
| 172 | 16 | 16 | 10 | 16 | 27.5 | 5 | 28 | 7 | 5 | 11 | 7 |
| 30 | 10 | 5 | 15 | 19 | 1 | 11 | 7 | 68 | 6 | 7 | 6 |
| 184 | 14 | 4 | 14 | 6 | 8 | 19.5 | 22 | 55.5 | 7 | 10 | 9 |
| 156 | 2 | 1 | 3 | 8 | 15 | 13 | 5 | 116 | 8 | 5 | 4 |
| 71 | 4 | 11 | 2 | 2 | 83.5 | 14 | 27 | 26 | 9 | 8 | 11 |
| 84 | 17.5 | 23 | 57 | 21 | 16 | 27 | 6 | 2 | 10 | 14 | 8 |
| 186 | 19.5 | 9 | 22 | 10 | 12 | 40 | 18 | 41 | 11 | 15 | 13 |
| 135 | 87.5 | 19 | 6 | 4 | 25 | 36.5 | 12 | 20.5 | 12 | 12 | 14 |
| 146 | 45 | 15 | 32 | 51 | 11 | 33 | 19 | 20.5 | 13 | 28 | 18 |
| 155 | 3 | 2 | 1 | 3 | 54 | 6 | 15 | 143.5 | 14 | 3 | 10 |
| 37 | 15 | 29.5 | 16 | 49 | 50 | 4 | 23 | 45.5 | 15 | 20 | 17 |
| 83 | 47.5 | 24 | 39 | 33 | 24 | 19.5 | 42 | 5 | 16 | 27 | 15 |
| 25 | 13 | 8 | 23 | 20 | 17.5 | 88 | 59 | 11 | 17 | 23 | 16 |
| 36 | 5 | 29.5 | 17 | 56 | 43 | 16 | 10 | 63 | 18 | 21 | 19 |
| 176 | 124.5 | 38 | 28 | 5 | 9 | 3 | 33 | 3 | 19 | 9 | 12 |
| 180 | 52 | 48.5 | 51 | 14 | 48 | 17 | 20 | 10 | 20 | 24 | 20 |

Table 3.3: Top Twenty Traps by Average Rank of Seeds per Bi-Year Period

the arithmetic and geometric means, respectively. Although there is some variation between the three ranking methods, all three contain approximately the same top 10 traps. However as one moves down the columns of Table 3.3, one observes less agreement between the three methods, and this disparity increases for the other traps not shown in the table.

In Section 1.3 we discuss how one might estimate the seed dispersal distribution from data collected over time from a seed-trap network with multiple sources. Some difficulties with these approaches are illustrated with the 1987-2002 FDP *Jacaranda* data set described above. A major difficulty, as will soon be seen, is that $S(j,t)$, the number of seeds caught in trap $j$ during period $t$, is observed, but $S(i,j,t)$, the number of such seeds which originated from tree $i$, is not. Botanists and ecologists have long wished that such data were available to them, but until recently, cost and technological limitations did not allow it. Wang and Smith [22] have written an interesting article discussing the use of genetic markers to definitively

identify the maternal source (which is the appropriate source for dioecious species such as *Jacaranda*) of seeds dispersed within a mapped area, thus allowing exact determination of distances traveled for all seeds which can be matched to a parental source. The first published study employing such methods appears to be Godoy and Jordano (2002) [10], in a study of an animal-dispersed tree species in Spain.

At UGA, the Statistical Consulting Center was contacted in August 2003 by F. Andy Jones, then a Ph.D. Student in Plant Biology, and his major professor, Dr. Steve Hubbell, about lending statistical assistance to part of Andy's Ph.D. dissertation dealing with estimation of the seed dispersal distribution of *Jacaranda* from genotyped data. In 2000, Andy had obtained leaf samples from all 199 adult *Jacaranda* trees on the FDP site, from 15 near-adult juveniles within the FDP, as well as from 91 adults in the 100 meter buffer zone surrounding the FDP site. The microsatellite DNA for these 305 trees at 11 micro-satellite loci were analyzed using di-nucleotide tandem repeats (DNTR) in a process described in Jones and Hubbell [13]. The idea was to find loci at which there were many different allele types (where, in this case, an allele type is a length of consecutive di-nucleotide tandem repeats) present in the population. Of the 11 loci examined, there appeared to be four which, when the allele patterns were examined jointly, determined almost uniquely the identity of the trees. For the 305 trees using alleles at these four loci, 280 unique patterns were found, with 262 of these occurring exactly once. Of the 18 allele patterns which occurred more than once, most belonged to pairs or triplets of trees which were very near one another.

Thus, in theory, one could collect *Jacaranda* seeds from the 200 traps on the FDP, subject them to genetic analysis at the four loci of interest to determine (almost surely) which parent had contributed the seed, and, thus, the exact distance traveled by all collected seeds. This greatly simplifies the estimation of the seed distribution function, in effect reducing the problem from the multiple unknown source problem discussed in Section 1.3 to a multiple version of the known source problem discussed in Section 1.2. Andy Jones indeed attempted to do just this in 2000-2002, but the analysis is not nearly as simple as the previous sentences

might suggest. The first difficulty is that some years do not produce many seeds, as noted in Table 3.1. For example, in 2001, there were so few seeds trapped relative to other years that Dr. Jones did not bother to genotype any seeds for that year. In the years 2000 and 2002, a sufficient number of seeds were obtained (1190 and 4138, respectively), but there was neither time nor physical resources to genotype them all, so sampling was used. The actual resolution of the genotyping process is shown in Table 3.4 below.

| | year | | |
|---|---|---|---|
| | 2000 | 2002 | Total |
| Seeds collected | 1190 | 4138 | 5228 |
| Seeds sub-sampled | 384 | 480 | 864 |
| -extract/PCR fail | -103 | -35 | -138 |
| Seeds genotyped | 281 | 445 | 726 |
| parent located | 243 | 373 | 616 |
| parent unknown | 38 | 72 | 110 |

Table 3.4: *Jacaranda* Seeds Collected and Genotyped in Traps by Year

First, a sample of seeds had to be selected from those available, since the maximum number of seeds which were able to be genotyped per year was about 400-500. Secondly, a certain proportion of seeds did not contain enough genetic material for the extraction process to succeed. Thus, as noted in Table 3.4, only 726 seeds of the 864 seeds sampled were successfully genotyped. Of these, 616 matched with one of the 306 parents in the FDP and buffer zone, so their true dispersal distances were known. (Except for the few cases where the genotyping yielded one of the 18 non-unique allele patterns, in which case the seed was assigned to the parent of that type nearest to the trap in which the seed was located.) Another 110 seeds (15% of the total genotyped) did not match any of the 280 allele patterns observed in the tree population and thus were inferred to have originated from a source outside the buffer zone. (The number of matches noted above is from a final careful analysis; earlier analyses yield as many as 119 unmatched seeds.)

These data were analyzed in Summer 2004, as part of a STAT 8000 consulting project undertaken by UGA Statistics graduate students JiEn Chen and Guo-Jing Weng, advised

by Jaxk Reeves and YiMei Cai. Some of their results as they apply to this dissertation are contained in Section 5.2. Much of this work has been incorporated into Jones et al. [14], published in The American Naturalist.

One possible criticism of the results found in the above paper is that the authors blindly assume that the assignment of seeds to trees is correct. There is certainly a possibility that a seed which is assigned to a parent on the FDP actually belongs to a parent with the same genotype elsewhere on the FDP (for the 18 multiple-tree allele patterns) or outside the buffer zone (for any type), but the probability of this occurring is small. A more severe error, whose occurrence was previously deemed remote, now appears to be more likely. That is, some of the 110 seeds classified as 'no match' may, in fact, have originated from parents in the FDP and buffer zone, but have been classified as 'no match' because of genotyping errors. Such errors may not be uncommon, as discussed in Bonin et al. [1]. Implications of corrections for genotyping errors in the *Jacaranda* data set are discussed at great length in Chapter 6 of this dissertation.

## 3.3 Estimation via Inverse Modeling Techniques

If one assumes a parametric density $f(r;\Theta)$ for the seed dispersal function, where $r$ is the distance and $\Theta$ represents the parameter(s), then the expected number of seeds caught in trap $j$ during year $t$ is:

$$\lambda_{jt} = E[S(j,t)] = c\sum_{i=0}^{I} A(i,t) * f(d_{ij},\Theta), \tag{3.2}$$

where $d_{ij}$ is the distance from tree $i$ to trap $j$, $A(i,t)$ is the total number of seeds dispersed by tree $i$ during year $t$, $c$ is a normalization constant adjusting for the trap size relative to the unit of measurement, and $i = 0$ is a generic notation to represent all trees not included within the $I$ which have been previously mapped.

While the above expectation is correct, it is not particularly useful as stated. Recall that the $S(j,t)$'s are observed and the $d_{ij}$'s are known for all $j$, for $i = 1, ..., I$. One might then be

tempted to use some sort of maximum likelihood technique to solve for the parameters $\Theta$ that maximize the likelihood of obtaining the observed sample. Unfortunately, even if one were to make simple assumptions about the functional form $f(r, \Theta)$, the model as stated above will not be identifiable, since there are no observations at the tree level. Various assumptions, some more realistic than others, have been made in order to make the model identifiable. A number of these are discussed below.

(i) Assume that $A(i, t)$ is constant.

While this makes the MLE solution relatively tractable, it does not agree well at all with current knowledge, since there are clearly huge variations in seed numbers from year to year as shown in Tables 3.1 and 3.2.

(ii) Assume that $A(i, t) = A(t)$.

This eliminates the problem noted in (i) by allowing the seed magnitude to vary by year, but it make the assumption that all trees in the mapped area are equally fecund. This also disagrees with past ecological research; larger mature trees are generally more fecund than smaller younger trees.

(iii) Assume that $A(i, t) = A(t) * B(i)$, where $B(i)$ is a function of the diameter at Breast Height for tree $i$, DBH($i$).

This is the most common assumption made in inverse modeling, and the most commonly assumed form for $B(i)$ is that it is proportional to the square of DBH($i$). Slightly more sophisticated models make $B(i)$ proportional to DBH$(i)^2$, conditional on DBH($i$) exceeding some minimum threshold of maturity. (This same result can be obtained by restricting the range $I$ of mapped trees to those which are 'adult', by some definition.) The assumption of (iii) has more credibility than those of (i) or (ii), although there is only marginal evidence to show that it is reasonable. A few researchers (Godoy and Jordan [10], F. A. Jones [13]) have collected seed capsules from beneath trees during

the seed dispersal season and fit regressions of these counts to DBH or DBH$^2$, rationalizing that total seeds produced by a tree during one year should be proportional to the number of capsules collected on the ground beneath the tree. Such fits offer limited support for the hypothesis that $B(i)$ is proportional to DBH$(i)$ or DBH$(i)^2$, or more generally to DBH$(i)^g$ for $1 \leq g \leq 2$.

(iv) Assume a log-linear model for $A(i,t)$.

Many statisticians, if they actually observed $A(i,t)$, would consider a linear model of the following type:

$$ln[A(i,t)] = \mu + \alpha_t + \beta_i + e_{it}, \tag{3.3}$$

where $\alpha_t$ and $\beta_i$ are main effects due to year$(t)$ and tree$(i)$, and where $e_{it}$ is a random error, typically assumed to be independently and identically distributed from some Normal$(0, \sigma^2)$ distribution. A slightly more sophisticated but similar approach is to fit a generalized linear model to the $A(i,t)$ using a log-link, the main-effects linear model shown in equation(3.3), and a Poisson distribution function. Of course, the linear or generalized linear model described above can not really be fit with the data at hand, since the $A(i,t)$ are not actually observed. One approach is to set the $e_{it} = 0$ and to solve for the $\mu, \alpha_t, \beta_i$ which yield the MLE estimate when equation (3.3) is substituted into equation (3.2). As stated, this procedure is not well-defined, since equation (3.2) gives an expected value, rather than a statistical model for $S(j,t)$ themselves. A logical approach is to assume that the $S(j,t)$ are, in fact, Poisson distributed with intensity $\lambda_{jt}$ given by equation (3.2). Assuming this and substituting (3.3) (with $e_{it} \equiv 0$) into (3.2) yields

$$ln(\lambda_{jt}) = \mu + \alpha_t + \ln[\sum_{i=0}^{I} e^{\beta_i} * f(d_{ij}, \Theta)]. \tag{3.4}$$

Without further information about either $\beta_i$ or $\Theta$, the model of equation 3.4 is not identifiable. From the raw $S(j,t)$ data, using either a linear model applied to $\ln(S(j,t))$ or a generalized linear model applied directly to the $S(j,t)$, it is easy to obtain estimates for $\mu$,

$\alpha_t$, or for the expression:

$$\hat{\gamma_j} = \ln[\sum_{i=0}^{I} e^{\beta_i} * f(d_{ij}, \Theta)]. \tag{3.5}$$

However, the estimates of $\hat{\beta}_i$ or $\hat{\Theta}$ are not separably estimable if both are unknown. Most typically, one assumes that the $\beta_i$'s are proportional to DBH$(i)^2$, and then proceeds to find the $\hat{\Theta}$ for given functional form, $f(r, \Theta)$ which minimizes the SSE for the $\gamma_j$'s . This is typically evaluated for several functional forms (which corresponds to changing the parameter, $c$, in Clark's hierarchy of dispersal functions, as given by equation (2.1)), and the $f(r, \hat{\Theta})$ which yields the best fit is declared to be the best-fitting model. If one tried to fix $f(r, \Theta)$ and solve for $\beta_i$'s by some sort of non-linear modeling, except in the rare case where $I << J$, one finds that the model is over-parameterized. Practical ramifications of this are discussed in Section 4.3 .

## 3.4 Estimation by Direct Modeling Techniques

In Section 3.3, we examined the approaches which one might use if one were in the classical seed trap situation where one knew the locations of all traps and all (within-site) sources, but did not know which seeds caught in a trap originated from which source. As demonstrated by the example of the previous section, seed dispersal density estimates can be obtained in such cases, but they are highly dependent upon assumptions which can not be checked in the classical case. We now explore improvements which can be made in the estimation if one is able to match seeds to sources.

If each seed in a trap could be correctly genotyped and uniquely matched to its source, one would then be able to observe $S(i, j, t)$, the number of seeds from source $i$ caught in trap $j$ during period $t$. Of course, in most practical situations, this will be a very sparse array containing mostly zeroes, since most seeds caught in a trap are from nearby trees. Nonetheless, one would likely assume that $S(i, j, t)$ followed a Poisson distribution with intensity parameter given by

$$\lambda_{ijt} = E[S(i, j, t)] = c * A(i, t) * f(d_{ij}, \Theta), \tag{3.6}$$

using the same notation as used in equation (3.2) above. The same problems with respect to observing/estimating the $A(i,t)$ as were noted in Section 3.3 are present here. Assuming they can be resolved, one could use generalized linear models with a Poisson distribution and log-link to obtain the MLE, $\Theta$, of the assumed parametric distribution, $f(r,\Theta)$. The key linear model equation is

$$ln(\lambda_{ijt}) = \mu + \alpha_t + ln(B_i) + ln[f(d_{ij},\Theta)]. \tag{3.7}$$

This approach seems straightforward, but is fraught with difficulties. Among these are the following:

(i) Generalized linear models when applied to data sets with many zero counts can behave very erratically. The best way to avoid such problems is to collect very large sample so that zero counts are rare. Unfortunately, seed genotyping was previously quite expensive and time-consuming, so that the observed $S(i,j,t)$'s of available data sets are not large. This difficulty will diminish somewhat in the future as genotyping becomes cheaper, although the general problem of sparseness, in the statistical sense, will still occur. The major difficulty with sparse data arises in estimating goodness of fit. If one naively performs a $G^2/df$ calculation and enumerates the degrees of freedom in the standard way, the fit will appear to be excellent, since one is gaining zero contribution to the fit statistic for the empty cells, while counting all such cells in the denominator. Restricting the denominator count to those cells with observed seeds causes $G^2$ to behave more like a Chi-squared statistic if the null hypothesis of a correct model is true, but that alone will not make the asymptotic Chi-squared approximation valid.

(ii) The assumption that $B(i)$ can be well-approximated is not currently warranted. (An analysis discussed later in this section effectively assumes that $B(i)$ is constant for all trees; not particularly realistic.) The most common assumption, as noted in Section 3.3, is that $B(i)$ is a function of DBH$(i)$, usually proportional to DBH$(i)^2$. In the example

examined in Chapter 5, the investigator actually obtained (for most sources), proxy counts for $B(i)$, but this would be unusual in most studies.

(iii) The linear model in equation (3.7) does not specify how to handle seeds which are not matched to any sources and are thus putatively assigned to sources outside the region of study. For these seeds, $d_{ij}$ is not known, or more precisely, $d_{ij}$ is censored, since it is known to be greater than the distance from trap $j$ to the nearest boundary. One could invent fictitious sources at the boundary for these seeds, but that would bias the estimates of the dispersal function in the low direction. Standard imputation procedures would also be difficult to employ, since the locations of the off-site adult *Jacaranda* trees are unknown. A better, but certainly more complicated, analysis method is to use techniques for dealing with censored data to handle these situations. This approach is examined in Section 5.2 of this dissertation.

(iv) The model assumes that the $S(i, j, t)$ is a complete correct count. In fact, since genotyping is expensive, one does not typically genotype all seeds caught within a trap during a time period. There is statistical variation due to seeds chosen to be genotyped, but that is generally not a problem given that the $S(i, j, t)$ themselves represent a random sample of possible observations. More crucial, however, is the assumption that the genotype assignment correctly matches the seed to its true source. Traditionally, those performing genetic analysis have been concerned about the Type I error, the probability that a seed will be classified as belonging to source $i$ when it really is from another source. This typically occurs when too few loci are sampled, so that two sources which are actually different display the same allele pattern. By increasing the number of loci sampled (and choosing loci with multiple alleles), geneticists have usually been able to make the Type I error in most experiments very small. Of course, as is well-known to statisticians, decreasing the probability of a Type I error, without making other changes to an experiment, will increase the probability of a Type II error. For many genetics experiments, this Type II error was believed to be of no real

consequence. However, in recent years, more concern has been focused on such errors, frequently called 'genotyping errors'. For the purposes of the current research, small probabilities of misclassification of alleles could have major implications, since if even one allele among the $2v$ genotyped at the $v$ loci is misgenotyped, the resulting allelic pattern will generally not match that of any of the known sources, thus resulting in an over-estimate of the proportion of seeds originating from off-site sources, and, of course, seriously affecting seed dispersal density estimates. This situation is investigated in detail in Chapter 6.

# Chapter 4

# Application Using Historical Trap Data

## 4.1 Combined Analysis

In this section, we will demonstrate the use of the techniques developed in Section 3.3 for non-genotyped data on the historical *Jacaranda* data collected at the FDP site from 1987-2002. As previously mentioned, we have collapsed this data into 8 'bi-years' to somewhat moderate the influence of year-to-year variability, although the effects are still large, as shown in Table 3.2. Thus, our data consist of the $200 * 8$ array of seed counts, $S(j,t)$, caught in trap $j$ during bi-year $t$. These range from a maximum of 1132 in trap 155 in 1991-92 to a minimum of 0 in 54 of the 1600 cells. If we fit a generalized linear model to these data, with Poisson distribution, log link, and additive relationship:

$$ln(\lambda_{jt}) = \mu + \alpha_t + \gamma_j \tag{4.1}$$

one finds that $\hat{\mu} = 3.1660$, and that the $\hat{\sigma}_\alpha = 0.5606$ over the 8 years and $\hat{\sigma}_\gamma = 1.3299$ over the 200 traps. This model is very over-dispersed relative to a Poisson (over-dispersion factor=21.3472), suggestting that even if one could disentangle the non-identifiability between the tree effects ($\beta(i)$'s) and the dispersion density, $f(r,\Theta)$, there is still extra variability (relative to Poisson variability) that can not be explained by either year or trap effects.

A similar analysis applied directly to the transformed data:

$$ln(S(j,t)) = \mu + \alpha_t + \gamma_j + e_{jt} \tag{4.2}$$

assuming a linear model yields: $\hat{\mu} = 2.9205$, $\hat{\sigma}_\alpha = 0.6595$, and $\hat{\sigma}_\gamma = 1.3274$, with $\hat{\sigma_e} = 0.8229$. These results are very similar to the Poisson model above, with the RMSE of this model

perhaps being slightly easier to interpret. For example, suppose that we were attempting to predict the number of seeds found in a 'typical' trap in a 'typical' bi-year, so that $\alpha_t$ and $\gamma_j$ are both approximately zero. Then the predicted amount (in log-scale) is 2.9205, with an approximate 95% prediction interval of $2.9205 \pm 1.96(0.8824) = [1.19, 4.65]$. Exponentiating, we obtain a point estimate of about 19 seeds for the trap, but the 95% prediction interval ranges from 3 seeds to 104 seeds! The analysis using the more sophisticated Poisson assumptions is equally bad, as shown by the dispersion factor of 21.35, meaning that the typical spread is 4.62 times what should be observed under Poisson assumptions. This serious uncertainty in predicting seed counts should not be under-appreciated. It is a problem which will be present no matter how correctly the seed dispersal function $f(d_{ij}, \Theta)$ or tree fecundity function $B_i$ are estimated.

Using $\gamma_j$'s as calculated by either equation 4.1 or equation 4.2, one can attempt to estimate the $f(r, \Theta)$ function that best fits, based on assumptions about the tree fecundities, $B_i$. The results in Table 4.1 below display these results using the estimated $\gamma_j$'s from the transformed linear model of equation 4.2 for four possible functions, $f(r, \Theta)$:

(a) Origin-mode two-dimensional gaussian ($c = 2$ in Eq. (2.1))

(b) Origin-mode two-dimensional exponential ($c = 1$ in Eq. (2.1))

(c) Origin-mode two-dimensional heavy-tail ($c = 0.5$ in Eq. (2.1))

(d) Two-dimensional log-normal ($r$ in log-scale)

Function (d) is not a member of Clark's class of models, since it does not assume that the mode occurs at the point of origination. Each of the four potential density functions is further estimated under four different assumptions on the tree fecundities:

(A) Null model

(B) All trees equally fecund

(C) Fecundity proportional to $\text{DBH}^2$

(D) Fecundity proportional to $\text{DBH}^g$ (with '$g$' estimated from data)

| Tree Fecundity | (a) OM2DG | (b) OM2DE | (c) OM2DH | (d) 2DLN |
|---|---|---|---|---|
| (A) Null Model | 358.1 | 358.1 | 358.1 | 358.1 |
| (B) Equally Fecund | 172.4 | 160.1 | 159.2 | 155.2 |
| (C) Fecundity $\sim \text{DBH}^2$ | 131.3 | 100.5 | 90.4 | 89.3 |
| (D) Fecundity $\sim \text{DBH}^g$ | 121.4 | 95.0 | 88.7 | 87.3 |
| MODEL D $\alpha$ | 35 | 15 | 1.61 | 0.98 ($\mu = 3.06$) |
| MODEL D g | 1.15 | 1.44 | 1.67 | 1.62 |
| MODEL D median dist. | $29m$ | $25m$ | $22m$ | $21m$ |
| MODEL D mean dist. | $31m$ | $30m$ | $32m$ | $35m$ |
| MODEL D $P(d > 100m)$ | .0003 | .0098 | .0459 | .0575 |

Table 4.1: Sum-of-Squared Error for Predicting $\gamma_j$

From Table 4.1, we can clearly see that assuming fecundity proportional to $\text{DBH}^2$ is much better than assuming all trees are equally fecund, which is, of course, much better than the null model assumption that all traps, on average, catch the same number of seeds, but not as good as the more general fecundity model in (D). In fact, the fecundity models (rows) of the table form a hierarchy of increasing generality going from A$\rightarrow$ D, so that, for a fixed distance function (column), the SSE decreases as one progresses down the table. The best (Row D) estimates of the distance scale parameter ($\alpha$) and fecundity parameter ($g$) for the best fitting models are shown at the bottom of Table 4.1. Obviously, the form of the distance function chosen has a large effect on the fit and on the median distance. Of the Clark dispersal models, the $c$=0.5 distribution appears to perform best, and there appears to be some advantage in allowing a more general form for fecundity than making it proportional to $\text{DBH}^2$. Indeed as Table 4.2 below shows, the Clark heavy-tail model with $c$=0.5, $\alpha$=1.61 and the 2-dimensional log-normal model with $\alpha$=0.98 and $\mu$=3.06 have very similar percentiles, at least over the middle range of the data. Both predict about the 5% of seeds to be dispersed greater than $100m$ (LDD), as opposed to about 1% for LDD for the Exponential and virtrully no LDD for the Gaussian. In the two tails, they are quite different, but the data set from which these

data were collected definitely do not contain many observations in the left-hand tail, since traps are rarely placed within $5m$ of a tree. There might be observations in the right-hand tail, but one does not know for sure unless one has genotyped data. For non-genotyped data, inverse modeling procedures tend to underestimate the occurrence of long-distance events.

| | ($c$=0.5, $\alpha$=1.61) | ($\mu$=3.06, $\sigma$=0.98) |
|---|---|---|
| Percentile | Heavy-Tail | Log-Normal |
| 99% | $163m$ | $209m$ |
| 95% | $97m$ | $107m$ |
| 90% | $72m$ | $75m$ |
| 75% | $42m$ | $41m$ |
| 50% | $22m$ | $21m$ |
| 25% | $10m$ | $11m$ |
| 10% | $5m$ | $6m$ |
| 5% | $3m$ | $4m$ |
| 1% | $1m$ | $2m$ |

Table 4.2: Comparison of Percentiles of Best-fitting Heavy-Tail and Log-Normal Distributions

Of course, none of the models in Table 4.1 fit very well, since even the best of them, with an SSE of approximately 87 units over 200 traps, yields a RMSE of 0.68 (in log-scale units), which means that answers could easily be incorrect by a factor of 2 in predicting the number of seeds in a particular trap for a given year. The point of Table 4.1 is not to provide a good model for the expected number of seeds per trap in a given year, but, rather, to give some intuition as to what sort of fecundity and distance functions are most likely to fit the data when it is not combined so crudely as was done here. Sections 4.2 and 4.3 provide more careful analyses of the complete data set, with Section 4.2 examining each bi-year separately and Section 4.3 pooling the data more carefully than was done here.

## 4.2 Separate Bi-year Analyses

The analysis provided in Section 4.1 suffers from the drawback that it attempts to fit the same model to each year, simply scaling each year's data to adjust for the seed intensity in that year. This might be too crude. In this section, we will attempt to analyze each bi-year separately. Of course, the intercepts of these models for different bi-years will be

quite different, reflecting the wide variation in the bi-year intensities, as noted in Table 3.2. However, we will be able to use the actual seed counts, $S(j,t)$ for trap $j$ in bi-year $t$ as our response variable, rather than the estimated $\gamma_j$'s used in Section 4.1. Similarly, rather than using an average DBH for each tree and selecting only the trees which were, on average, adults during the 16-year period, we can use those which were adult in a given year, along with a more precise estimate of each DBH. (Recall that DBH's are obtained from censuses only once every 5 years, so some interpolation is necessary, but this is still more precise than what was done in Section 4.1.) A more important benefit of modeling each bi-year separately is that we can obtain some idea of how robust our parameter estimates are over the 8 bi-year periods.

Table 4.1 displays results from fitting 13 different models (null model plus (3 fecundities * 4 distance functions)) for the combined data. Based on these results, we decided to fit models of the following form to our data (separately for each bi-year, $t$):

$$S(i,j,t) \sim Poisson(\lambda) \tag{4.3}$$

$$ln(\lambda) = \beta_0 + \beta_1 * q_{ij} + \beta_2 * ln(DBH_i) \tag{4.4}$$

where $q_{ij} = \sqrt{d_{ij}}$.

This is equivalent to (row D, column c) of Table 4.1, where $\beta_1 = -1/\alpha$ and $\beta_2 = g$. Thus, we might expect ($\beta_1$,$\beta_2$) for the 8 different bi-years to be near ($-0.62$, 1.67), if this model is consistent over time.

The major difficulty with fitting the above model, as noted previously, is that we do not actually observe the $S(i,j,t)$ (the number of seeds from tree $i$ which landed in trap $j$ during bi-year $t$), but, rather, $S(j,t)$, the total for trap $j$ during that time period. The easiest way around this difficulty is to use the E-M algorithm. In this case, this is equivalent to creating fictitious (not necessarily integer-valued) $S(i,j,t)$ values such that $\sum_i S(i,j,t) = S(j,t)$, finding the MLE's of ($\beta_0$,$\beta_1$,$\beta_2$) under that configuration (M-Step) and then using the expected values for $S(i,j,t)$ (for the given MLE's) subject to the summation constraint

(E-step), and to iterate until the process converges. The ($\hat{\beta}_0$,$\hat{\beta}_1$,$\hat{\beta}_2$) found by this procedure are the true MLE's, although one must be careful about interpreting the goodness-of-fit statistics ($G^2$), since the evaluation is over such a sparse subset of the entire possible tree-trap population. The values $G_I^2$ and $G_J^2$ might be a bit more interpretable as Chi-squared fit statistics. The former is the likelihood statistic calculated for each adult tree and summed over the trees, while the latter is the same calculated over the traps. That is,

$$G^2 = 2 \times \sum_{i=1}^{I} \sum_{j=1}^{J} S(i,j) \times ln\left(\frac{(S(i,j)}{E(S(i,j))}\right) \tag{4.5}$$

$$G_I^2 = 2 \times \sum_{i=1}^{I} \left( \sum_{j=1}^{J} S(i,j) \times ln\left(\frac{\sum_{j=1}^{J} S(i,j)}{\sum_{j=1}^{J} E(S(i,j))}\right)\right) \tag{4.6}$$

$$G_J^2 = 2 \times \sum_{j=1}^{J} \left( \sum_{i=1}^{I} S(i,j) \times ln\left(\frac{\sum_{i=1}^{I} S(i,j)}{\sum_{i=1}^{I} E(S(i,j))}\right)\right) \tag{4.7}$$

For the ungenotyped data case here, where the $S(i,j)$ are not actually observed, but are imputed by the E-M algorithm, it is easy to see that $G_J^2 = G^2$; this will not be true for the genotyped data case of Chapter 5. Under the E-M algorithm, the fictitious $S(i,j,t)$'s are best-possible-fit values subject to the trap constraint and are thus *more* consistent with a proposed model than could be obtained if the $S(i,j,t)$ were actually observed. This distinction is discussed more in Chapter 5, where the $S(i,j,t)$'s are observed.

The results of applying the E-M algorithm with the Poisson model of (4.1) and (4.2) to the separate bi-year data yields the results shown in Table 4.3. The left-hand side of the table displays observed characteristics of the data for the given bi-year. As noted previously, the number of seeds trapped in a bi-year varies considerably, from a low of 4161 in bi-year H to a high of 23815 in bi-year C. The third column, NZ Traps, displays the number (among 200 possible) of non-zero (i.e. non-empty) traps. Even in the least abundant bi-year, at least 180 of the 200 traps caught at least one seed. The fourth column displays the number of adult trees observed during the bi-year, where 'adult' is defined to mean DBH $\geq 200mm$. This number rises more-or-less monotonically from 251 to 290 over the 16-year period, although

the actual number of unique trees considered is over 300, since a few trees died while some younger trees matured.

| Bi-Year | Seeds | NZ Traps | Adult Trees | $\beta_0$ | $\beta_1$ | $\beta_2$ | $G^2 = G_J^2$ | $G_I^2$ | NZ trees |
|---|---|---|---|---|---|---|---|---|---|
| A | 5667 | 189 | 251 | 5.5685 | -0.7150 | 0.9620 | 2872 | 993 | 197 |
| B | 7396 | 192 | 252 | 5.3496 | -0.6544 | 1.4315 | 5126 | 1438 | 213 |
| C | 23815 | 200 | 274 | 6.6688 | -0.7008 | 1.5519 | 6635 | 6551 | 246 |
| D | 14684 | 191 | 267 | 6.4661 | -0.7265 | 1.1139 | 9488 | 3021 | 231 |
| E | 10130 | 196 | 282 | 6.1188 | -0.7407 | 0.9486 | 8739 | 4630 | 230 |
| F | 12359 | 199 | 290 | 5.8578 | -0.6559 | 0.7485 | 9852 | 4094 | 252 |
| G | 13072 | 199 | 289 | 6.3602 | -0.7472 | 1.0266 | 8651 | 4504 | 244 |
| H | 4161 | 180 | 290 | 5.8282 | -0.8521 | 0.0000 | 5353 | 3681 | 178 |

Table 4.3: Bi-year Models

The right-hand side of Table 4.3 displays some fit statistics for the model (4.4) fit to each bi-year. The intercepts, as noted previously, vary approximately as $ln(S(t))$, although the relationship is not perfect. The estimated $\beta_1$ coefficients range from $-0.65$ to $-0.82$, which is more negative than the $-0.62$ obtained from the crude combined analysis of Section 4.1. Except for bi-year H, which has some unusual aspects, the $\beta_2$ estimates are in the range $[0.75, 1.55]$, which is lower than the 1.67 predicted from the combined analysis. The $G^2$ value is the deviance statistic for the fit, although its interpretation is questionable given that the $S(i, j, t)$'s are fictitious. As can be noted, in every bi-year, the ratio of $G_J^2$ to the number of non-zero traps indicates over-dispersion of a major degree, from a factor of about 5.0 for bi-year A to over 26 for bi-year C. The last column, NZ trees, is the number of adult trees, which, under the model, are expected to have at least one seed caught in a trap during the bi-year. Of course, under the model, every tree-trap combination has a positive $E(S(i, j))$, but one can calculate the probability that the actual observed sum for a particular tree ($i$) summed over all traps ($j$) is zero and use this to calculate NZ trees. Note that the number of NZ trees range from about 61% of all adult trees (for Bi-year H) to 90% of adult trees (for Bi-year C), a range that may be larger than realisitc.

## 4.3 Bi-year X Analysis

Since neither the combined data models nor the separate bi-years appear to fit the data too well, we investigated this more carefully, to see what aspects of the model could be improved upon to yield a better fit. To do this, we returned to the full data set to create a bi-year 'X', which is a fictitious average cohort of seeds. Thus, this analysis will be similar to that of Section 4.1, but with one cohort of a 'typical' size rather than the sample of >91000 seeds which one obtains from pooling all the data as was done in Section 4.1. The value of $S(j)$ used for bi-year X is the geometric mean of the 8 $S(j,t)$'s, provided that they are all $> 0$. Any $S(j,t)$'s which were equal to zero were replaced by $(1/e)$ before the geometric mean was calculated. The $S(j)$'s thus calculated were then rounded to the nearest integer to yield a plausible typical year's data set. This procedure yielded a total of 9087 seeds for bi-year X, with the 200 trap-sums varying from a minimum of 1 (for 16 of the traps) to a maximum of 451 (for trap #85). The trees chosen for the bi-year X analysis were those whose median DBH over the 16-year period was $\geq 200mm$, which yielded 291 trees, with each tree's median DBH over the 16-year period being used as the 'true' DBH for bi-year X. We then used the E-M algorithm described in Section 4.2 to fit four models to this bi-year X data, as shown in equations (4.8)-(4.11) below. The motivation for bi-year X is the creation of a sample which is typical with respect both to annual size of the seed sample, and to the average distance distribution of seeds from each tree. It does sacrifice a bit with respect to using average DBH for each tree, rather than using the yearly DBH(i)'s, but the behavior of the fecundity function, as shown by the $\beta_2$ estimates in Table 4.3, is not too stable, so this is not much of a loss. It is hoped that by pooling in this way, we can obtain more stable estimates for both the distance and fecundity functions, especially the former.

(sl :)

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 * s_{ij} + \beta_2 * LDBH_i \tag{4.8}$$

(dl:)

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 * d_{ij} + \beta_2 * LDBH_i \tag{4.9}$$

(ql:)

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 * q_{ij} + \beta_2 * LDBH_i \tag{4.10}$$

(el :)

$$ln(\lambda_{ij}) = \beta_0 + \beta_{ec} + \beta_2 * LDBH_i, \tag{4.11}$$

where $s_{ij}$=$d_{ij}^2$, $q_{ij}$=$d_{ij}^{1/2}$, and LDBH=$ln$(DBH/423). All four models express fecundity as a function of LDBH=$ln$(DBH/423), where '423' was chosen since $423mm$ is the median DBH of the 291 trees used in this analysis. The model notation used here and subsequently, is that the first letter is a short-hand for the type of distance function used ($s$, $d$, $q$, $e$) while the second is a short-hand for the type of fecundity fuction used, ('$l$' for LDBH, in this case).

The models differ in their distance functions, with the $s$, $d$, $q$ models of 4.8-4.10 referring to $c$=2, $c$=1, and $c$=0.5, respectively, from Clark's hierarchy. Equation 4.11 ('e' model) is a detailed 'empirical' distance model that divides the data into 24 (mostly equal) distance classes: 0-5$m$, 5-10$m$, ..... 95-100$m$, 100-150$m$, 150-200$m$, 200-400$m$, $>$ 400$m$. This allows us, assuming over-parameterization is not too severe, to see which parametric distance form really fits the data best for a 'typical' year (given ungenotyped data). The results for the four models are shown in Table 4.4.

| Model | $\beta_0$ | $\beta_1$ | $\beta_2$ | $G^2 = G_J^2$ | $G_I^2$ | NZTr | $E[n\|d = 25]$ | $P(d > 100)$ |
|---|---|---|---|---|---|---|---|---|
| sl | 2.8273 | -0.0005 | 0.9693 | 6774 | 1791 | 195 | 12.36 | .0068 |
| dl | 4.5622 | -0.0780 | 1.1185 | 4909 | 2556 | 183 | 13.63 | .0034 |
| ql | 6.1621 | -0.7665 | 0.8963 | 4664 | 2055 | 232 | 10.27 | .0530 |
| el | 4.6493 | | 0.9785 | 4305 | 1634 | 203 | 7.68 | .0000 |

Table 4.4: Best Models for Bi-year X

The $c = 2.0$ ($sl$ model) is woeful and will not be considered further. As noted previously, the $c = 0.5$ ($ql$ model) is the best of the Clark models and is not too much worse than the non-parametric distance ($el$) model ($G_J^2$ is 4664 vs. 4305), although certain characteristics

of these two models are somewhat dissimilar. For the $ql$ model, about 232 of the 291 trees are expected to have at least one seed caught in one of the traps during a typical year, with the expected number of seeds caught in a trap 25 meters from a tree being 10.27, and the proportion of seeds dispersed $> 100m$ (LDD) being about 5%. For the $el$ model, that are fewer non-zero trees expected, the expected number of seeds caught in a trap $25m$ away is smaller than for the $ql$ model, and no seeds are predicted to be LDD.

The $ln(\lambda)$ for a typical (DBH= $423mm$) tree under the four distance functions for bi-year X is shown in Figure 4.1. The discontinuous block function is for the $el$ model. It is actually even jumpier than displayed, since the '$-5$' shown for 85-90$m$ and $> 100m$ are really negative infinities caused by zero events being expected for these distance classes. Of course, the $el$ model is over-parameterized – we do not have enough observations in all of the classes to make good estimates. If we smooth this distance function a bit, as shown in Fig. 4.2, we see that it appears to be closer to to the $ql$ function than to the others, up to about $90m$. Beyond that distance, the $el$ function becomes extremely negative. However, we view this latter result somewhat skeptically, since we have no actual data for the distance classes. Although there are many tree-trap combinations yielding distances that are $> 90m$, the $el$ model, (and, in fact, all four models considered) maximize their likelihoods by making such long-distance dispersal events rare to non-existent. This phenomenon has been noted previously and occurs because the MLE criterion is most influenced by the commonly occurring events. Long-distance dispersal will not be detected by any model unless genotyped data determines definitively that LDD events occur with probabilities greater than implied by Fig. 4.2.

Model 4.11 ($el$) fits the distance function part of the intensity about as well as it can be fit, but the fit by the $G^2$ (or more appropriately, $G_J^2$) is still lacking. Since the distance fitting can not be improved upon, we considered adjusting the fecundity function to fit better. The models used heretofore parameterize fecundity as proportional to DBH$^g$. For a few trees, this causes extremely poor fits. This can be seen from examination of the individual (tree) components of the $G_I^2$ statistic. An obvious way to improve the fit would be to allow each

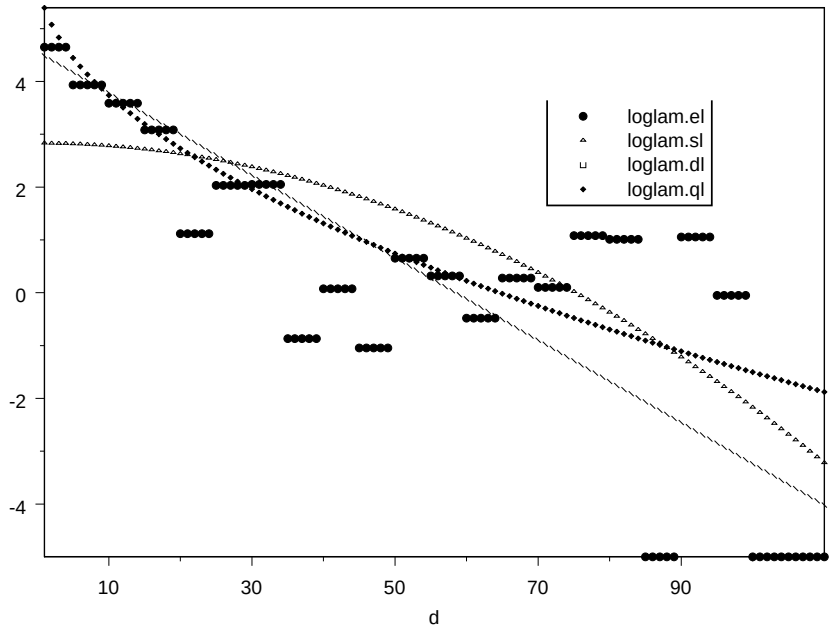


Figure 4.1: Log-intensities for Typical Tree for Bi-year X

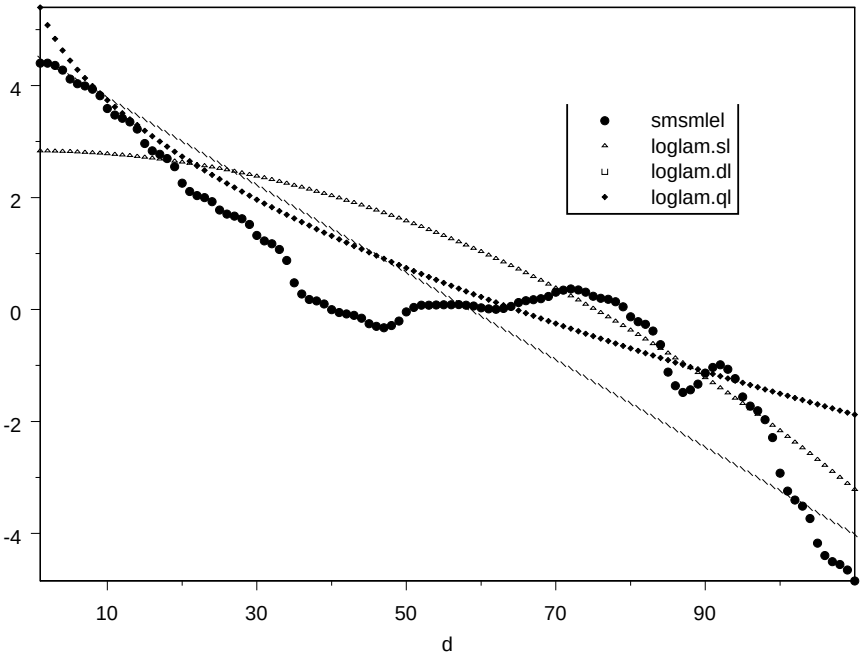


Figure 4.2: Log-intensities for Typical Tree for Bi-year X (Smoothed)

tree to have its own fecundity. Rather than doing this blindly, we did it sequentially, starting with the *dl*, *ql*, and *el* models of (4.9)-(4.11) and then replacing the worst-fitting tree (as given by $G_I^2$ components) by its own log(fecundity), and continuing sequentially. After 27 trees had been adjusted in this way, the $G_I^2$ value was near what one would expect if the fecundity function were correctly specified. That is, the $G_I^2$ statistic was approximatly equal to the expected number of non-zero trees minus the number of parameters estimated for each model.

The parameter and fit statistics for these fecundity-adjusted models are shown in Table 4.5. Note that $G^2$ has improved substantially, by about 3000 units for each of the three models, certainly worth the 27 parameters spent. Of course, forcing a model to fit in this way is not very satisfying, since it leaves unanswered such fundamental questions as what characteristics did these 27 trees have which caused them to be adjusted. (Generally, they are large trees whose fecundities were adjusted downward from the $\text{DBH}^g$ estimate, but that is not always true.) In any case, even if the adjustments make the data fit better, such adjustments are risky, because there is no verifiable evidence that these trees produce more or less seeds than expected, since the actual $S(i,j)$'s are not observed for this data set. Finally, even if we do feel justified in using these fecundity-adjusted models, they still are not very good, even in the best (*el*) case. The $G_I^2$ value is acceptable (and indeed can be made to approach zero by allowing more trees to have their own fecundity parameters), but $G^2$ ($= G_J^2$) will not become much smaller than shown in Table 4.5. If the model is adequate, $G_J^2$ should be approximately Chi-squared distributed with 148 *df* (200 traps $-52$ independent parameters estimated). As can be noted in Table 4.5, the $G_J^2$ value for this case is about 9 times larger than the *df*. It is difficult to evaluate $G^2$ precisely since we do not know the true number of non-zero tree-trap cell counts, but even using a liberal estimate, the $G^2$ value is about 4 times as large as it should be if the fit were adequate. So, even over-parameterizing drastically in both the distance and fecundity functions, we will never be able to obtain a classical (no over-dispersion) Poisson fit for a typical bi-year's non-genotyped data.

| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $G^2 = G_J^2$ | $G_I^2$ | NZTr | $E[n\|d = 25]$ | $P(d > 100)$ |
|---|---|---|---|---|---|---|---|---|
| d | 4.1024 | -0.0802 | 1.8240 | 1572 | 98 | 168 | 8.15 | .0030 |
| q | 5.7167 | -0.7968 | 2.1240 | 1623 | 106 | 209 | 5.66 | .0433 |
| e | 3.5810 | | 1.7380 | 1337 | 82 | 204 | 5.96 | .0000 |

Table 4.5: Best Models for Bi-year X (27 Trees Adjusted)

In summary, what we have learned about fitting inverse models to non-genotyped seed trap data sets are the following:

a) Different time periods (seasons of collection) can be modeled separately, assuming sufficient sample sizes, since year-to-year variation in seeds produced is great.

b) For *Jacaranda* trees, Clark's model with $c = 0.5$ appears to be a reasonable distance approximation, provided that there are not many long-distance-dispersal (LDD; $d > 100m$) events.

c) Modeling fecundity as being proportional to DBH$^2$ is reasonable, but more general models which make fecundity proportional to DBH$^g$ will typically estimate $g$ to be smaller than 2, usually $1 < g < 2$.

d) Even if one over-parameterizes by fitting certain influential trees separately, the dispersion of the overall model will be at least 9 times as great as that expected under Poisson conditions, when evaluation is at the trap level; i.e. using $G_J^2$. (Evaluation using $G^2$ is not appropriate because the value is biased low by the E-M fit. Even more importantly, because so many of the possible $291 * 200$ tree-trap combinations are expected to yield 0 seeds, a straight-forward evaluation of $G^2$ as a deviance statistic is not appropriate.)

e) If one were to obtain genotyped data (and, thus, observed $S(i, j)$ values, as discussed in Chapter 5), one would no longer need to use the E-M algorithm and $G_J^2 \neq G^2$. The maximum likelihood estimates of the parameters would minimize the $G^2$ but not

the $G_J^2$. Thus, the true $G_J^2$ for genotyped data would be worse than is suggested from non-genotyped data, since the observed $S(i,j)$'s can not fit as well as the fictitious $S(i,j)$'s obtained under the E-M algorithm. On the other hand, genotyped seed data allows one to see whether the assumptions made to obtain conclusions (b), (c), and (d) are reasonable.

# Chapter 5

# Application Using Genotyped Data

## 5.1 Background

The data set used for this analysis was described in Section 3.2. It is a small subset of the data analyzed in Section 3.4. For this analysis, the time variable $t$ is not included, since there were genotyped observations of seeds in only two years, (281 in 2000, 445 in 2002), and they were combined to achieve a pooled sample of 726 genotyped seeds, as shown in Table 3.4. Of these 726, 616 were matched to a known source within the FDP and buffer zone, while 110 (15%) did not match and were thus believed to have originated from off-site *Jacaranda* trees.

Although one might attempt an analysis along the lines described in Section 3.4, it will not work well. Recall that there are $I$=306 known sources (adult and semi-adult *Jacaranda* trees) which were genotyped prior to the seed collection. The number of traps used in the collection of the 726 genotyped seeds was $J$=298 (the established 200-trap network used in the historical analysis above was augmented by a special set of 98 traps which Andy Jones set up in 2000-2002 to more carefully sample source 'gap' areas in the FDP). So, clearly, with only 616 matched observations spread over $I * J = 306 * 298 = 91188$ tree-trap combinations, most observed $S(i, j)$ counts will be zero. This extreme sparseness, combined with the relatively high censoring rate (due to non-matched seeds) renders the approach of Section 3.4, without many further assumptions, to be worthless.

## 5.2 Censored Data Approach

A more straight-forward approach, similar to that utilized by JiEn Chen and Guo-Jing Weng in their Summer 2004 Statistical Consulting (STAT 8000) project, is to assume that the distances associated with the 726 genotyped seeds represent an independent and identically distributed sample of size $n$=726 from the seed dispersal distribution. Of course, this is not really correct, since the distribution of traps relative to trees yields a pattern of sample distances which is not equivalent to a simple random sample of all seeds dispersed from all trees in the region. Nonetheless, one can begin with this assumption to find an approximate density which can then be adjusted for sampling vagaries.

If one assumes that the distances $x_1, x_2, ..., x_n$ obtained from the $n$ genotyped seeds are a random sample from a population with pdf= $f(x, \phi)$ and cdf= $F(x, \phi)$, then the likelihood function is defined to be:

$$L(\phi, x) = \prod_{i=1}^{n} L_i(\phi|x_i) = \prod_{i=1}^{n} [f(x_i|\phi)]^{1-\delta_i} \times [1 - F(x_i|\phi)]^{\delta_i} \tag{5.1}$$

where $\delta_i = 0$ if $x_i$ is a matched seed distance, $\delta_i = 1$ if $x_i$ is an unmatched (right censored) seed distance.

Then, the log-likelihood has the form:

$$Log(L) = \sum_{i \in O} \ln f(x_i|\phi) + \sum_{i \in C} \ln[1 - F(x_i|\phi)] \tag{5.2}$$

where O is the set of indices for matched observations and C is the set of indices for censored observations. For the data in question, the cardinalities of these two sets are 610 and 116, respectively. This analysis assumed 116 censored observations, slighty different from the 110 now believed to be correct(see Chapter 6).

Solving the above equation for the maximum likelihood estimates for various parametric forms is not particularly difficult. Unfortunately, simple parametric models of the type discussed in Chapter 2 do not fit well at all. In cases like this, especially when one has no real idea about the true functional form of $f(x|\phi)$, it is always useful to plot the empirical distribution function, or more commonly, the Kaplan-Meier survival function. As is well-known,

this survival function incorporates the censored data (as much as possible) into the estimate. The K-M empirical survival function for the sample data is shown in Figure 5.1. Note that all of the censored observations fall between $100m$ and $350m$, since those are the minimum and maximum possible distances from a trap within the FDP to the edge of the buffer zone nearest to the trap. There are a few very long non-censored observations in the 350-$700m$ range, corresponding to genotyped seeds which traveled from a tree in one corner of the plot to a trap in the opposite corner. Of course, the censored observations might well be in the 350-$700m$ range if they were actually to be observed.

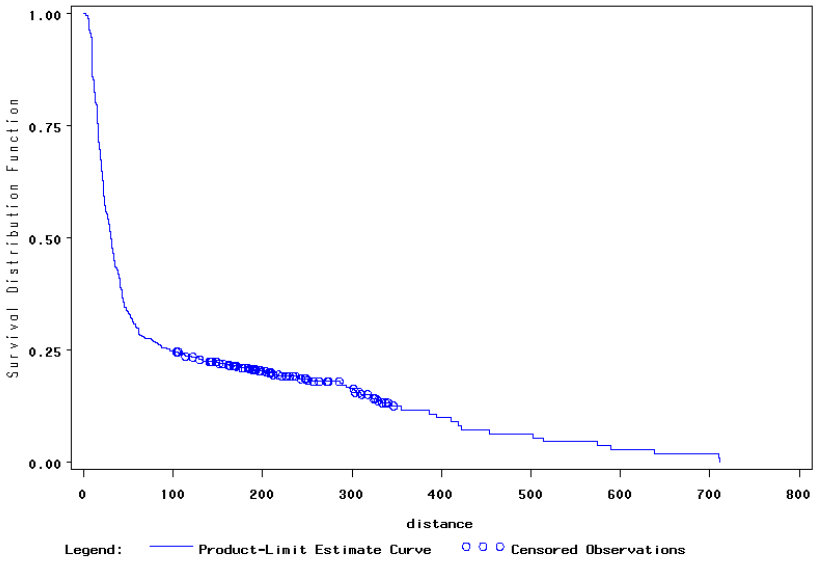


Figure 5.1: Empirical Survival Distribution Function.

In some cases, plotting the empirical survival function $s(t)$, or various transformations thereof, such as $\log[s(t)]$ or $\log[-\log(s(t))]$ allows one to determine if certain standard parametric survival distributions, such as exponential, Weibull, or log-normal are appropriate. For the plot shown in Figure 5.1, however, none of the standard distributions will work well. The large mass of censored data between 100-$350m$, followed by a few observations in the 350-$700m$ range causes this. If there were no observations beyond the censored mass, this would be the typical situation encountered with right-censoring, and one might have some

[0]Circles represent the censored data.

hope that a heavy-tailed failure-time distribution would fit the entire data. In this case, however, that can not possibly be the case, and the only feasible models would be those which are a mixture of two different densities. One such fitted density is shown in Figure 5.2, as is explained below. Slightly different density plots would be obtained under other assumptions, but the general pattern for all is the same as that shown in Figure 5.2. That is, the data appear to be a mixture, with a sizeable proportion of the data arising from a distribution with a mode near $25m$ and the remainder arising from a less common distribution with a mode near $350m$. Using this information, one returns to the log-likelihood approach of equation 5.2, but now attempts to maximize over the mixture of two densities:

$$Log(L) = \sum_{i \in O} \ln[pf_1(x) + (1-p)f_2(x)] + \sum_{i \in C} \ln[1 - (pF_1(x) + (1-p)F_2(x))] \qquad (5.3)$$

where $f_1$ and $f_2$ are the component densities for the left and right humps and $p$ is the proportion of the sample arising from density 1. The large hump in Figure 5.2, representing density $f_1(x|\phi)$, has more observations and can be more reliably estimated than the smaller hump, which contains most of the censored data. It appears that a log-normal distribution fits the left-tail portion of the data well. The second hump of the distribution is much more ambiguous, as is the parameter, $p$, of the mixing distribution. After fitting a number of models of the form 'lognormal + something else' for the data, Jones et al. [14] concluded that ($f_1 \equiv lognormal, f_2 \equiv normal$) fit reasonably well. The best parameters for this model for the combined (2000+2002) data set, as well as separately for the two years, are shown in Table 5.1. Note that the lognormal($f_1$) mean and variability parameters, as well as the mixing parameter, $p$, are relatively stable between the two years and pooled, while the normal($f_2$) mean and variability parameters are much less precisely estimated. Indeed, one obtains very similar log-likelihood scores with many different functional forms for $f_2$. Note that the best models shown here are not part of the Clark hierarchy of models shown in equation 2.1, since their polar coordiate adjusted modes are not at $r$=0. However, the combined estimate for $(\mu_1, \sigma_1)$ of the log-normal($\mu$=3.1060,$\sigma$=0.745) is relatively close to the ($\mu$=3.06,$\sigma$=0.98)

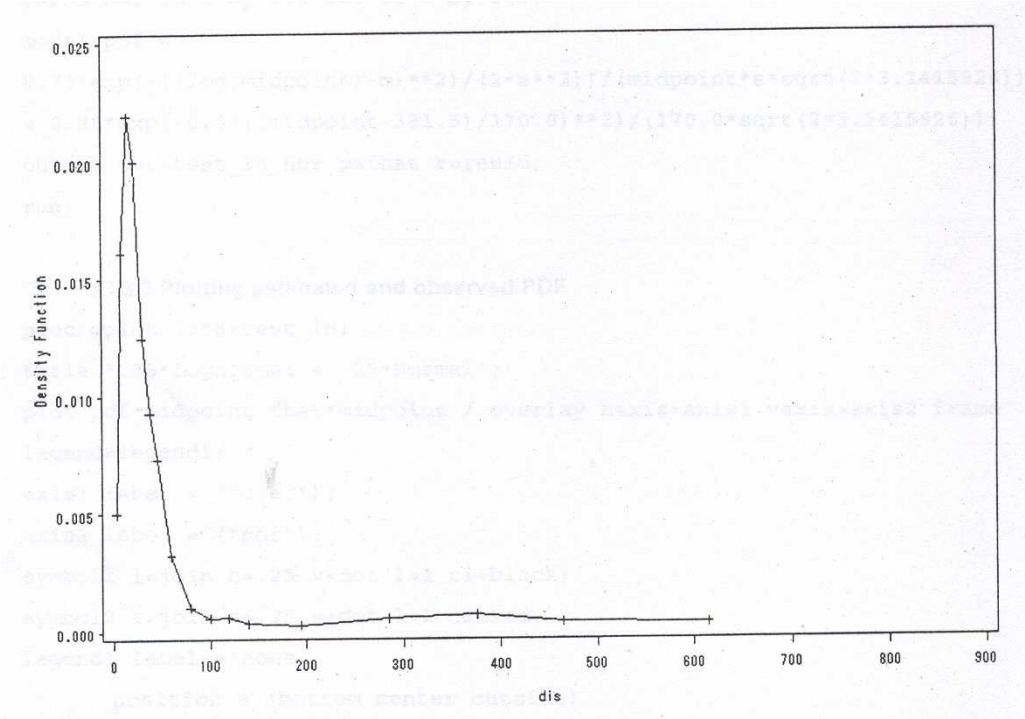


Figure 5.2: Estimated pdf vs. Distance

obtained when using the log-normal density for 'bi-year X' data in Section 4.3. So, for the lower tail of the distribution, a log-normal might be appropriate. The key point, that some sort of mixture model is necessary, is not too surprising to ecologists. A possible explanation for the undoubtedly complex true situation is that $f_1$ represents the typical dispersion that most ($p$) of the seeds experienced, but that a smaller proportion $(1-p)$ of the seeds are caught in wind up-drafts and thus dispersed at larger average distances, as given by $f_2$.

| Year | N* | Nobs** | Ncens*** | $\hat{p}$ | $\hat{\mu_1}$ | $\hat{\sigma_1}$ | $\hat{\mu_2}$ | $\hat{\sigma_2}$ |
|---|---|---|---|---|---|---|---|---|
| 2000 | 281 | 238 | 43 | 0.791 | 2.960 | 0.721 | 338.462 | 106.148 |
| 2002 | 445 | 372 | 73 | 0.756 | 3.210 | 0.735 | 393.241 | 164.389 |
| Combined | 726 | 610 | 116 | 0.766 | 3.106 | 0.735 | 375.862 | 153.332 |

*N is the total number of seeds genotyped
**Nobs is the number of observations matched to parent trees
***Ncens is the number of censored observations

Table 5.1: MLE's for the 'Lognormal + Normal' Model

## 5.3 Direct Estimation Approach

For the ungenotyped seed analyses performed in Chapter 4, it was not necessary for the models to consider whether an observed seed was 'matched' or 'unmatched', since seeds were observed only as totals found per trap. The E-M algorithm used to fit these models attempted to apportion each seed probabilistically to trees, with nearby trees and larger trees receiving more weight. These models did not attempt to assign any probability to beyond-the-buffer-zone trees (primarily because such trees' locations are unknown), but even if such trees had been included, given that every such tree is at least $100m$ from the nearest trap, and given the very low intensities associated with distances $> 100m$ for all models in the graphs of Figures 4.1 and 4.2, there is virtually no difference in the fitted models.

For the genotyped data set, however, one must decide how to handle the 110 unmatched seeds. A simple choice is to ignore these 110 seeds, using only the 616 matched seeds in the analyses. How 'wrong' this is depends on the analyst's viewpoint. Strictly speaking, for the direct analysis, it is not wrong at all. From that viewpoint, there is a set of $I$ genotyped adult trees whose locations have been mapped. In addition, a set of $J$ traps has been set out at known locations around these traps. This yields a total of $N{=}I * J$ possible tree-trap distances at which seed counts can be obtained and used to obtain MLE's of parameters for any model desired. There is no requirement that the $J$ traps be placed in any manner such that each tree has an equal chance of being sampled nor that the seeds collected be a 'random sample' from the population of seeds dispersed by the $I$ trees in the population. In fact, the only randomness necessary, given the design, is that the seeds be randomly sampled from the $J$ traps over the period of study (or that all seeds be used, if one is willing to believe that one season's worth of data is a 'random sample' from all that could be observed). It is true that some trees may tend to have more seeds collected, either because they are very large or near many traps (or both), but the model accounts for these factors. It might be that some distance classes are sampled more frequently than others, but that has little effect. For example, for the design used on the FDP, it is very rare for a trap to be placed within

$5m$ of any *Jacaranda* tree, so estimates of distance behavior at that short range may be problematic. Similarly, since LDD rates are expected to be low, it might be the case that a very large number of possible tree-trap combinations at long dispersal distances would need to be observed before reliable estimates of the tail intensity rate could be obtained, but there is no inherent bias in restricting the analysis to those seeds which originated from the $I$ genotyped trees and landed in the $J$ traps. The only bias would occur if there had been a mistake in genotyping, so that either some of the 110 'unmatched' seeds should have been matches or some of the 616 'matches' included seeds matched to the wrong source tree. The former occurrence (which is not that uncommon, as we shall see in Chapter 6) would still not bias the results, assuming that one believed that genotyping errors are made at random, not associated with any particular distance, $d_{ij}$, from tree to trap, nor with any particular tree or trap. The latter error, which we will treat as rare for the analyses of this section, could, as also discussed in Chapter 6, cause serious biases.

Thus, if we proceed to analyze the 616 'matched' genotyped seeds as in Chapter 4, but without resorting to the E-M algorithm (since each seed is uniquely matched to both trap and tree), our data set has the characteristics described next. There are $n$=616 seeds, theoretically distributed over $I$=292 adult trees and $J$=298 traps. However, only 114 of the 292 adult trees yielded at least one sample seed, with 96 seeds originating from one very fecund tree. Similarly, only 168 of the 298 traps yielded at least one matched seed, with the most being observed in any one trap being 34 seeds found in Trap #84. Finally, of the $N$=292 $\times$ 298=87016 possible tree-trap combinations, only 298 actually contained at least one observation. Hence, as noted previously, this data set is very sparse and will not yield very precise parameter estimates.

The first three models which we considered for this data set are the same three considered at the end of Section 4.3 (models 4.9-4.11), using LDBH (= $ln(DBH/423)$) as the fecundity measure, and with linear ($d_{ij}$), square-root ($q_{ij}$) or empirical distance class ($e_{ij}$) as the distance functions. The fit results for these three models are shown in the top panel

| Model | $\beta_0$ | $\beta_1$ | $\beta_2$ | $G^2$ | $G_I^2$ | $G_J^2$ | $G_{ec}^2$ | NZTr | $P$(LDD) |
|---|---|---|---|---|---|---|---|---|---|
| dl | -0.0753 | -0.0371 | 1.3666 | 3496 | 1340 | 973 | 619 | 183 | .1153 |
| ql | 2.2725 | -0.6419 | 1.3115 | 3049 | 1280 | 876 | 266 | 175 | .1173 |
| el | -0.4688 | | 1.5453 | 2775 | 1168 | 830 | 0 | 166 | .1023 |
| dn | -2.7316 | -0.0352 | 0.7459 | 2849 | 701 | 654 | 535 | 141 | .1338 |
| qn | -0.2739 | -0.6241 | 0.7118 | 2425 | 658 | 567 | 208 | 138 | .1310 |
| en | -1.9707 | | 0.6921 | 2215 | 620 | 558 | 0 | 135 | .1023 |

Table 5.2: Best Models for Genotyped Data (Top – LDBH, Bottom – NAFX)

of Table 5.2. Comparing these values to their analogues shown in Table 4.4 for the bi-year X analysis, we note several items of interest. As with Table 4.4, the fit statistics ($G^2$, $G_I^2$, $G_J^2$) improve as one progresses down the $(d)\longrightarrow(q)\longrightarrow(e)$ hierarchy of models. As with the non-genotyped analysis of Table 4.4, the $G_J^2$ and $G_I^2$ values are more legitimate to use for Chi-squared goodness-of-fit statistics than is $G^2$, but the $G_J^2$ values are no longer identical to the $G^2$ values. (If one were to use the E-M algorithm procedure and the same three models to analyze this data set, in effect ignoring the observed counts $S(i,j)$ of seeds from tree $i$ caught in trap $j$, and using only the trap-sum counts $S(j)$, one obtains values of $G^2$=$G_J^2$ of 833, 823, and 773 for the three models, respectively, substantially less than what is shown for $G^2$ or $G_J^2$ in the top panel of Table 5.2. This is not at all unexpected, and demonstrates that non-genotyped inverse-modeling approach typically used can give one a false sense of confidence in a model's fit to the data.) The intercepts ($\beta_0$) for three models are all much smaller than for the corresponding bi-year X models, reflecting the large change in magnitude in the total numbers of seeds observed ($n$=616 vs. $n$=9087). The more relevant differences are in the $\beta_1$ (distance) and $\beta_2$ (fecundity) multipliers, with the fecundity multipliers being about 30% to 50% larger for the genotyped data set, and the distance multiplier for the '$d$' model being much smaller in magnitude ($-.0371$ vs. $-.0780$) while that for the '$q$' model is somewhat smaller ($-.6419$ vs. $-.7665$). The $e_j$ values for the two non-parametric '$e$' models are not directly comparable, since they depend on their respective models' intercepts, but their

general pattern is similar. This can best be seen by comparing Figure 5.3 (Log-Intensities for a Typical Tree for Genotyped Data – Short Distances) with Figure 4.2 (Log-Intensities for a Typical Tree for Bi-year X). In both cases, the $e_j$ function is a bit jumpy, indicating possible over-parameterization. As with the non-genotyped data, the square-root distance function ($q$) appears to be reasonably close to the empirical class distance function ($e$) over the range from $0m$-$100m$. For distances beyond $100m$, as shown in the graph of Figure 5.4 (Log-Intensities for a Typical Tree for Genotyped Data - All Distances), the behavior for non-genotyped and genotyped seeds' data sets is quite different. From Figure 4.3, for non-genotyped seeds, the graphs do not extend beyond 100m since the expected event intensity is so low as to be negligible for the '$el$' model. For the genotyped data set, from Figure 5.4, however, the situation is quite different beyond $100m$, with the log-intensity for the '$el$' model appearing to level off at a much higher rate than predicted by the '$dl$' or '$ql$' models. Thus, we now have some definite statistical verification of the LDD effect. It is not strong (as noted by the fact that log-intensity estimates for $d > 100m$ are so low) and not nearly as pronounced as the bi-modal hump displayed by the density estimate approach of Section 5.2, but is present and statistically significant. An indication of the magnitude of this LDD effect can be seen by noting from Table 5.2 that all six models predict about 10%-13% LDD.

Let us now examine carefully what we have learned from our study of the non-genotyped seeds (91000 sampled over 8 bi-years) and the matched genotyped seeds (616 seeds sampled from 2000-2002). To facilitate matters, let us use the $ql$ models given by equation 4.4, so that $\beta_1$ refers to the slope for distance $q=\sqrt{d}$ parameters and $\beta_2$ refers to the slope for LDBH. Various characteristics of these data sets and fits for this model to these data sets are displayed in Table 5.3, and Figure 5.5 plots the point estimates of ($\beta_1$,$\beta_2$) for these data sets. The letters A-H in this table and figure refer to bi-years A-H, with the 2nd column of Table 5.3 using the (minimum. median, and maximum) of each statistic to summarize the information contained in Table 4.3 of the previous chapter. The 'X' refers to the synthetic cohort 'X', which is a type of average of bi-years A-H. The 'K' refers to the 616 matched

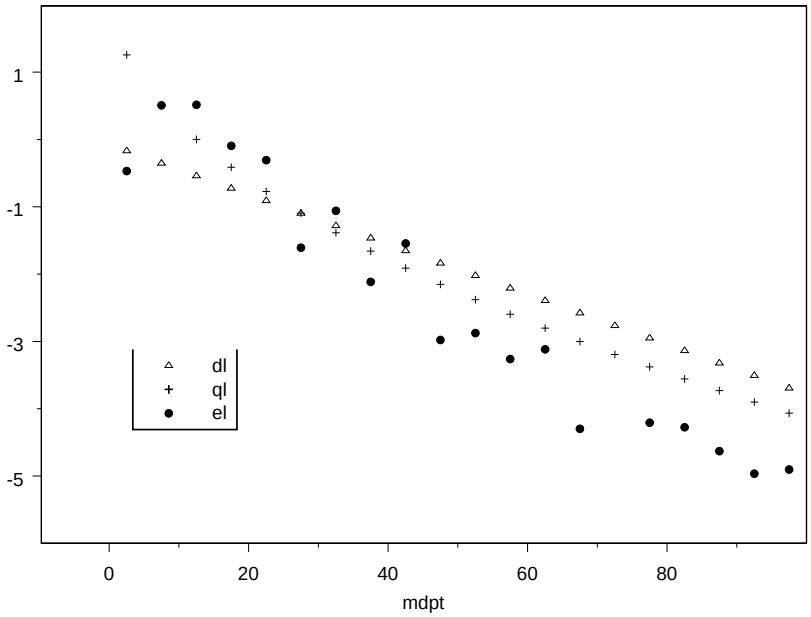


Figure 5.3: Log-intensities for Typical Tree for Genotyped Data (Short Distances)

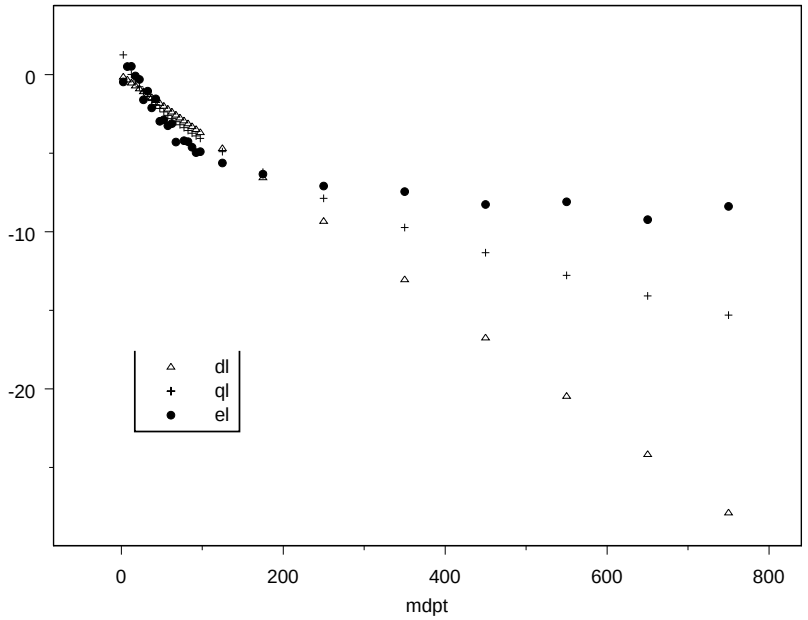


Figure 5.4: Log-intensities for Typical Tree for Genotyped Data (All Distances)

seeds analyzed by the direct ML method; i.e. where the distances are known. The 'J' refers to the same data set, but where the analysis is by the indirect E-M method which ignores the distance data and minimizes $G_J^2$, the deviance over the trap-sums. For data sets X, J, and K, approximate 95% confidence ellipsoids for ($\beta_1$,$\beta_2$) using the profile likelihood method (adjusting for over-dispersion) are also plotted in Figure 5.5 . The 'X' ellipsoid, although based on a synthetic seed cohort, is highlighted since it represents in some sense what might be found in a typical bi-year, if one had collected about 9100 non-genotyped seeds and used the inverse estimation methods of Chapter 4. Indeed, four of the eight bi-years (A, D, F, and G) have point estimates for ($\hat{\beta}_0$, $\hat{\beta}_1$) which fall in the bi-year X confidence ellipsoid. Under the *ql* model, each of the bi-year A-H data sets has its own confidence ellipse, of roughly the same size as that of bi-year X, but they are not displayed in order to maintain visual clarity. The 'J' analysis, although using the 616 matched genotyped seeds, does so in the manner which one would use if the genotyping were not known. The purpose for displaying it is to give some idea of how much confidence area variation is due to sample size, with the sizes of the two samples being 9087 for 'X' and 616 for 'J'. The 'K' ellipse is the one of primary interest, displaying the estimates and joint confidence interval for ($\beta_1$,$\beta_2$) under the *ql* model of Table 5.2. Although the 'K' ellipsoid is larger than that of 'X' due to the disparity in sample sizes, it is not nearly as large as the 'J' ellipsoid based on the same sample. This occurs because genotyping yields direct calculation of actual dispersion distances, contributing valuable information not available in the bi-year X data set. The extra information was not enough to completely offset the lack of sample size, since the 'K' ellipse still contains much more area than the 'X' ellipse. The most important point to garner from the two ellipses is that the genotyped data set (K) contains fairly convincing evidence that the $\beta_1$ coefficient is significantly less negative than is estimated from the ungenotyped data (X). That is, there is significantly more long-distance dispersion estimated to occur when the *ql* model is fit to the genotyped data than when it is fit to the ungenotyped data. Thus, even with a relatively small sample size (616 genotyped vs. 9087 ungenotyped seeds),

this effect can be detected. Another salient point which can be noted from all three ellipses is that it is difficult to estimate $\beta_2$ (the LDBH multiplier) very precisely. The bi-year X ellipsoid estimate for $\beta_2$ ranges from about 0.7 to 1.3, while that for the 'K' data set ranges from about 0.6 to 1.9 . Finally, note that all three ellipses are actually tilted slightly to the upper right, since there is a positive correlation between the ($\hat{\beta}_1$,$\hat{\beta}_2$) estimates in each case, but it is very small, on the order of $r$=+0.05.

| | Bi-Years A-H | Bi-Year X | Bi-Year X | Geno-J | Geno-K |
|---|---|---|---|---|---|
| # Seeds | [4161, 11189, 23815] | 9087 | 9087 | 616 | 616 |
| NZ Traps | [180, 194, 200] | 200 | 200 | 168 | 168 |
| Adult Trees | [251, 278, 290] | 291 | 291 | 292 | 292 |
| Method | Indirect(EM) | Indirect(EM) | Indirect(EM)(27) | Indirect(EM) | Direct |
| $\beta_0$ | [5.35, 5.89, -6.67] | 6.16 | 5.72 | 2.81 | 2.24 |
| $\beta_1$ | [-0.82, -0.72, -0.65] | -0.77 | -0.80 | -0.71 | -0.64 |
| $\beta_2$ | [-0.84, 1.02, 1.35] | 0.90 | 2.12 | 0.31 | 1.31 |
| $G^2$/NZ Tree-Traps | [1.48, 3.59, 6.59] | 2.69 | 0.95 | 1.82 | 6.50 |
| $G_J^2$/NZ Traps | [15.2, 38.9, 49.7] | 23.3 | 8.10 | 4.10 | 5.21 |
| $G_I^2$/NZ Trees | [5.0, 17.3, 26.6] | 8.90 | 0.46 | 1.91 | 7.31 |
| $P$(LDD) | [.030, .072, .109] | .053 | .043 | .076 | .118 |

Table 5.3: Parameter Estimates from *ql* Model for Various Data Sets

In addition to the improvement in distance function estimation which can be obtained from the genotyped data due to actual seed dispersal distances being observed, there is another advantage to this particular genotyped data set. The advantage is that there is more fecundity information than usual available for this set of genotyped trees. In the years (2000 and 2002) in which the the genotyped *Jacaranda* seeds were collected, Andy Jones also collected some fecundity data more informative than DBH. For 188 of the 292 adult *Jacaranda* trees in the FDP in both years, Andy collected data on the number of capsules found underneath each tree during a 1-month period. This is a better measure of fecundity than DBH, since the number of seed capsules found is expected to be approximately proportional to the number of seeds released. Of course, it will not be a perfect measure, since not all capsules contain the same number of seeds, not all seeds in a capsule disperse, the distribution of capsules over the one-month period may not be representative of the true seed production, and various other reasons. Nonetheless, it is the best proxy for actual seed

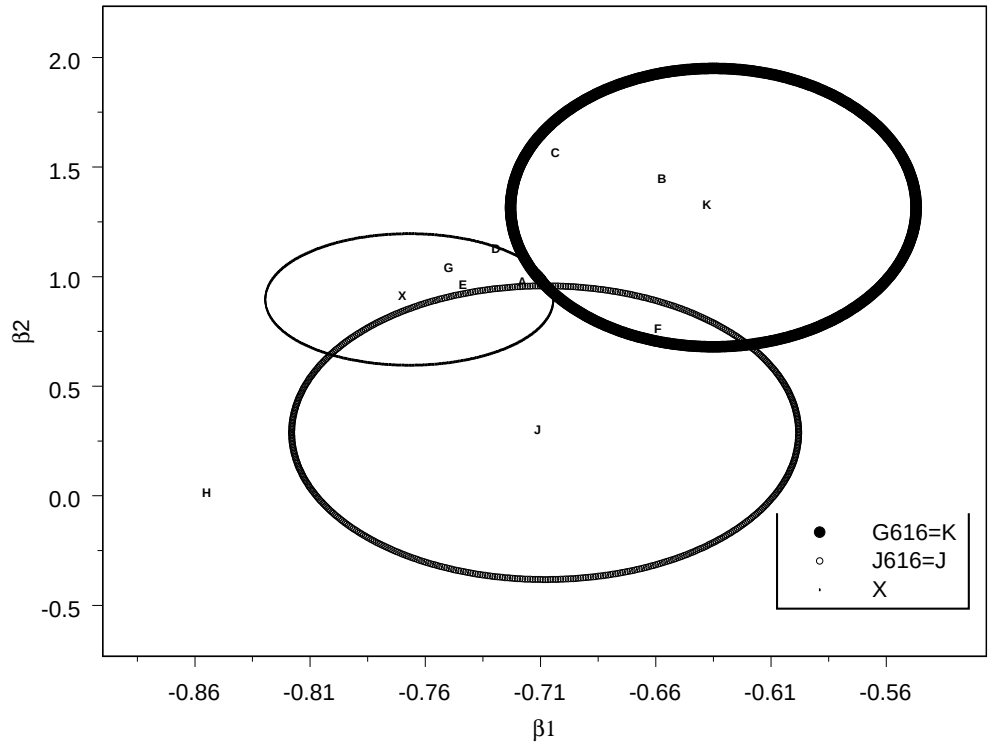


Figure 5.5: 95% Profile Likelihoods for ($\beta_1$,$\beta_2$) from $ql$ Models for Data Sets

counts, and can help determine the relative worth of LDBH as a fecundity measure. In particular, for each of the 188 trees whose capsules were counted, the new variable, NAFX$_i$, was created as:

$$NAFX_i = 0.5 * [ln(C2000_i) + ln(C2002_i)], \tag{5.4}$$

where C2000$_i$ and C2002$_i$ represent the capsule counts for tree $i$ in the years 2000 and 2002, respectively. In the rare cases where either of the capsule count values was zero, the log value was set to 0. For the 188 trees for which NAFX$_i$ is calculable, a simple linear regression was run to predict NAFX$_i$ from LDBH$_i$. The resulting equation was:

$$NAFX_i = 2.5356 + 2.3038 * LDBH_i + e \tag{5.5}$$

with $R^2$=0.25 and RMSE =1.45. Various other transformations were tried, but none were significantly better than above. A plot of the data (in the scales used) is shown in Figure 5.6. From this, if we believe that NAFX is a good measure of fecundity, then we have our first concrete evidence that assuming fecundity proportional to DBH$^2$ might be valid, since the 95% confidence interval for the LDBH coefficient ranges from 1.70-3.00, containing 2. Of course, as is evident from the large confidence interval, the relatively small R-squared, or simple examination of the data plotted in Figure 5.6, there is much individual variation in tree fecundity that will never be able to captured by anything as simple as a function of DBH. If one could afford to do what Andy Jones did (measure fecundities for individual trees), one will, of course, obtain much better fits than one can using DBH alone, as we shall see next, but one should realize that collection of such information is very unusual.

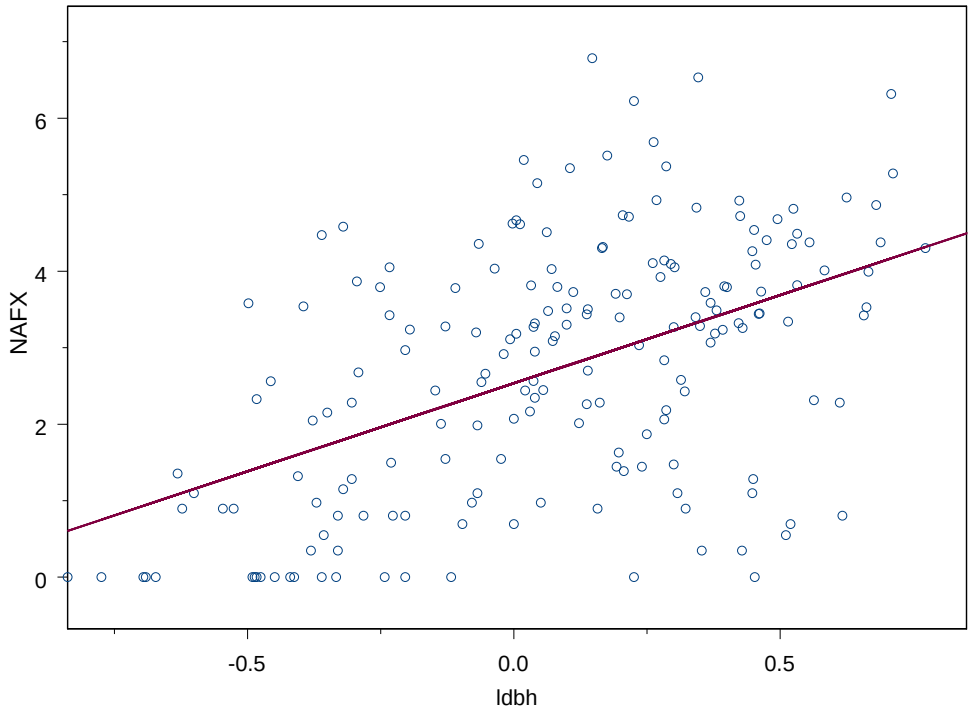


Figure 5.6: Plot of NAFX vs. LDBH for 188 Fecundized Trees

To see how much information is gained by using NAFX$_i$ rather than DBH$^g$ (since the latter is the best estimate that is typically available), we ran the three models of (4.9-4.11),

but with $\text{LDBH}_i$ replaced by $\text{NAFX}_i$ for the 188 trees for which $\text{NAFX}_i$ was available. For the other 104 trees, $\text{NAFX}_i$ was estimated from eq. (5.5). The $\text{NAFX}_i$'s were then used in Poisson models with these link fucntions:

(dn :)

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 * d_{ij} + \beta_2 * NAFX_i \tag{5.6}$$

(qn:)

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 * q_{ij} + \beta_2 * NAFX_i \tag{5.7}$$

(en:)

$$ln(\lambda_{ij}) = \beta_0 + e_{ij} + \beta_2 * NAFX_i \tag{5.8}$$

The results for these Poisson regression models as applied to the genotyped data are shown in the bottom panel of Table 5.2. One will note a tremendous improvement in the $G_I^2$ for all 3 models, with each being about one-half as large as found when using the LDBH fecundity measure. So, clearly, if one has special fecundity information for individual trees, as we do here, one should use it. On the other hand, such information does not cause $G_I^2$=0, so the fecundity adjustment of 27 trees used in the bi-year X analysis of Section 4.3 is overly optimistic. When such specialized information is not available, as is typically the case, our analyses show that assuming fecundity proportional to $\text{DBH}^g$, with $g$ estimated from the data is reasonable and that ( $1 < g < 2$) is a perhaps typical range. Note also that using the better fecundity estimates ($\text{NAFX}_i$) rather than the crude estimates ($\text{LDBH}_i$) has very little effect on the distance function estimates, as the $\beta_1$ estimates under the paired models of the two panels of Table 5.2 are very similar. This can also be seen by the fact that $G_e^2$, a fit statistic over the 24 distance classes, changes relatively little, if at all, between the model pairs of Table 5.2 – the bulk of the improvement occurs in fitting individual trees. Of course, even using the best distance function (e) and fecundity function (NAFX), as shown by the '*en*' model at the bottom of Table 5.2, still yields a $G_J^2$ value of 558, quite large for 168 non-zero traps, not even accounting for degrees of freedom lost for estimated parameters.

| dc | distance | mdpt | tree_trap | match | lec | len | lqn | ldn | lel | lql | ldl |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0-5m | 2.5 | 7 | 8 | 0.13 | -0.21 | 0.55 | -0.92 | -0.47 | 1.26 | -0.17 |
| 2 | 5-10m | 7.5 | 33 | 67 | 0.71 | -0.33 | -0.17 | -1.10 | 0.51 | 0.51 | -0.35 |
| 3 | 10-15m | 12.5 | 48 | 75 | 0.45 | -0.41 | -0.67 | -1.27 | 0.52 | 0.00 | -0.54 |
| 4 | 15-20m | 17.5 | 93 | 98 | 0.05 | -0.78 | -1.07 | -1.45 | -0.09 | -0.41 | -0.72 |
| 5 | 20-25m | 22.5 | 88 | 76 | -0.15 | -0.79 | -1.42 | -1.63 | -0.31 | -0.77 | -0.91 |
| 6 | 25-30m | 27.5 | 115 | 33 | -1.25 | -1.91 | -1.74 | -1.80 | -1.61 | -1.09 | -1.10 |
| 7 | 30-35m | 32.5 | 129 | 55 | -0.85 | -1.63 | -2.02 | -1.98 | -1.06 | -1.39 | -1.28 |
| 8 | 35-40m | 37.5 | 153 | 21 | -1.99 | -2.63 | -2.29 | -2.15 | -2.11 | -1.66 | -1.47 |
| 9 | 40-45m | 42.5 | 177 | 47 | -1.33 | -2.05 | -2.53 | -2.33 | -1.54 | -1.91 | -1.65 |
| 10 | 45-50m | 47.5 | 187 | 12 | -2.75 | -3.36 | -2.76 | -2.51 | -2.98 | -2.15 | -1.84 |
| 11 | 50-55m | 52.5 | 193 | 14 | -2.62 | -3.33 | -2.99 | -2.68 | -2.88 | -2.38 | -2.02 |
| 12 | 55-60m | 57.5 | 233 | 10 | -3.15 | -3.86 | -3.20 | -2.86 | -3.26 | -2.59 | -2.21 |
| 13 | 60-65m | 62.5 | 235 | 13 | -2.89 | -3.75 | -3.40 | -3.03 | -3.12 | -2.80 | -2.39 |
| 14 | 65-70m | 67.5 | 239 | 4 | -4.09 | -4.70 | -3.59 | -3.21 | -4.30 | -3.00 | -2.58 |
| 15 | 70-75m | 72.5 | 261 | 0 | . | . | -3.78 | -3.39 | . | -3.19 | -2.77 |
| 16 | 75-80m | 77.5 | 285 | 5 | -4.04 | -4.68 | -3.96 | -3.56 | -4.21 | -3.38 | -2.95 |
| 17 | 80-85m | 82.5 | 282 | 5 | -4.03 | -4.65 | -4.13 | -3.74 | -4.28 | -3.56 | -3.14 |
| 18 | 85-90m | 87.5 | 328 | 4 | -4.41 | -5.09 | -4.30 | -3.91 | -4.63 | -3.73 | -3.32 |
| 19 | 90-95m | 92.5 | 331 | 3 | -4.70 | -5.37 | -4.47 | -4.09 | -4.97 | -3.90 | -3.51 |
| 20 | 95-100m | 97.5 | 329 | 3 | -4.70 | -5.37 | -4.63 | -4.27 | -4.90 | -4.07 | -3.69 |
| 21 | 100-150m | 125 | 4062 | 18 | -5.42 | -6.03 | -5.44 | -5.23 | -5.63 | -4.90 | -4.71 |
| 22 | 150-200m | 175 | 5173 | 11 | -6.15 | -6.79 | -6.72 | -6.99 | -6.34 | -6.22 | -6.57 |
| 23 | 200-300m | 250 | 12740 | 13 | -6.89 | -7.55 | -8.33 | -9.63 | -7.10 | -7.88 | -9.35 |
| 24 | 300-400m | 350 | 13579 | 10 | -7.21 | -7.86 | -10.14 | -13.15 | -7.45 | -9.74 | -13.06 |
| 25 | 400-500m | 450 | 12381 | 4 | -8.04 | -8.61 | -11.70 | -16.67 | -8.28 | -11.34 | -16.77 |
| 26 | 500-600m | 550 | 10476 | 4 | -7.87 | -8.40 | -13.10 | -20.19 | -8.10 | -12.78 | -20.48 |
| 27 | 600-700m | 650 | 8529 | 1 | -9.05 | -9.47 | -14.38 | -23.71 | -9.23 | -14.09 | -24.19 |
| 28 | 700-800m | 750 | 6810 | 2 | -8.13 | -8.54 | -15.56 | -27.23 | -8.39 | -15.31 | -27.90 |
| 29 | 800-900m | 850 | 5155 | 0 | . | . | -16.66 | -30.75 | . | -16.44 | -31.61 |
| 30 | 900-1000m | 950 | 3211 | 0 | . | . | -17.70 | -34.27 | . | -17.51 | -35.32 |
| 31 | 1000-1100m | 1050 | 1037 | 0 | . | . | -18.69 | -37.79 | . | -18.53 | -39.03 |
| 32 | 1100-1200m | 1200 | 117 | 0 | . | . | -20.08 | -43.07 | . | -19.96 | -44.60 |
| 33 | 1200-1300m | 1250 | 0 | 0 | . | . | -20.53 | -44.83 | . | -20.42 | -46.45 |
| 34 | 1300-1400m | 1350 | 0 | 0 | . | . | -21.39 | -48.35 | . | -21.31 | -50.16 |
| 35 | 1400-1500m | 1450 | 0 | 0 | . | . | -22.23 | -51.87 | . | -22.17 | -53.87 |
| 36 | 1500-1600m | 1550 | 0 | 0 | . | . | -23.03 | -55.39 | . | -23.00 | -57.58 |
| 37 | >1600m | 1650 | 0 | 0 | . | . | -23.81 | -58.91 | . | -23.80 | -61.29 |
| Total | | | 87016 | 616 | | | | | | | |

Table 5.4: Log-intensity for Typical Tree by Distance Class for Genotyped Data Models

To examine the 6 different distance functions of Table 5.2 further, consider Table 5.4. The first 3 columns of the table define 37 different distance classes by boundary and midpoint. The first 23 of these are the same as used in the $el$ and $en$ models, with the $24th$ class expanded further for greater clarification. Column 4 ('tree_trap') of the table lists the number of tree-trap combinations which fall in each distance class (among the 292×298 tree-trap combinations considered for the genotyped dataset), while column 5 ('match') shows how many of the 616 matched seeds were assigned to trees in the distance class. The last 6 columns of Table 5.4 give the log-intensity by distance class as estimated by the 6 models for a 'typical' (DBH=$423mm$) tree. For the nonparametric distance ($el$ and $en$) models, these are the log-intensity estimates for the class, while for the parametric $q$ and $d$ models, these are the log-intensity functions evaluated at the mid-point of each interval (for a 'typical' tree). The column labeled $lec$ is simply $ln$(match/tree_trap), the empirical log-intensity unadjusted for tree sizes. From the previous discussion, we know that the three $n$ models will outperform the three $l$ models by the $G^2$ criterion, but that this will not manifest itself in the distance distribution, and indeed one does observe that $ldn{\sim}{=}ldl$, $lqn{\sim}{=}lql$, and $len{\sim}{=}lel$ over most of the range of the data. The closeness of the first two pairs can be seen from the similarity of the estimated $\beta_1$ coefficients of the respective paris in Table 5.2. The similarity between $len$ and $lel$ can be seen from the plot in Figure 5.7. Within a fecundity measure ($n$ or $l$), we know that $e$ and $q$ are fairly similar for the 0-100m distance range, but that for larger distances both $q$ and $d$ are too negative compared to $en$ or $el$ or $ec$ with respect to estimating events. That is, even for the genotyped data set (from which all the log-intensity functions shown in Table 5.4 are estimated), there is a tendency for parametric distance models to underestimate the probability of very long dispersal events. The tendency is not nearly as severe as that caused by using the indirect estimation methods for ungenotyped data, as discussed in the context of Figure 5.5, but it does occur to some extent. For the 6 models shown in Tables 5.2 and 5.4, the estimates of LDD (defined as dispersion greater than $100m$) are fairly similar between the models, but this occurs because the '$d$' and '$q$' models pile their LDD

density in the 100-150$m$ and 150-200$m$ categories, while the '$e$' models have longer tails. For example, of the 616 matched genotyped seeds, 63, in actuality, were dispersed from trees more than 100$m$ away from the trap in which they eventually settled. Both the $el$ and $en$ methods, by definition, yield 63/616=0.1022 as the expected proportion of LDD seeds, given the tree-trap network of distances and fecundities. The $ql$ and $qn$ models yield expected proportions of 0.1175 and 0.1310, respectively, for LDD seeds, but cluster more of these events in the $100-150m$ class and fewer in the $> 400m$ class than are actually observed for the 616 matched seeds.

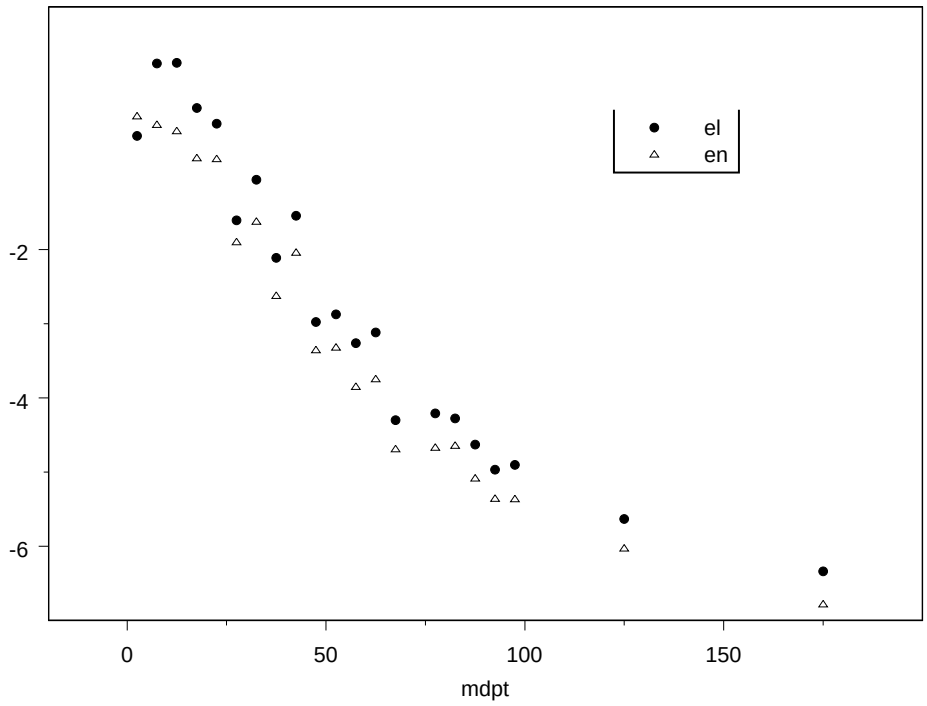


Figure 5.7: Comparison of Log-intensities for Typical Tree Under $el$ and $en$ Models

Thus, we can say in conclusion that both the genotyped and ungenotyped data sets give some general support to the belief that log-intensity function decays approximately at square-root of distance rate for $d_{ij} < 100m$, but that the genotyped data gives definite evidence of some ($\sim 10\% - 13\%$) long-distance dispersals, although certainly not nearly as strong as the censored data approach estimates of Jones, Chen, Weng and Hubbell [14] described in

Section 5.2. Both the ungenotyped and genotyped data sets lend some support to the idea that fecundity is proportional to DBH$^g$, with estimates of $g$ varying widely, but generally being less than the $g$=2 value often assumed in the inverse modeling literature. In any case, no matter what model is used, there is much variability in predicting an individual tree's fecundity, as displayed both by the necessity to separately adjust fecundity estimates for 27 trees (among 291) for the ungenotyped data set of Section 4.3 and by the relatively low $R^2$ of 0.25 from the regression equation of (5.5). Another way of stating all of this is that even if we over-parameterize to obtain approximately the best possible fit for distance (which is what the '$e$' models attempt to do) and to model individual trees' fecundities (using the NAFX$_i$ values), we can force $G_e^2$ and $G_I^2$ to become arbitrarily small, but neither the overall deviance, $G^2$, nor $G_J^2$, which is the best overall fit to the trap data, will fit adequately by any standard statistical convention. Thus, there must be at least some interaction between distance and fecundity; something very difficult to estimate.

# Chapter 6

## Misclassified data

### 6.1 Genotyping Errors

Chapter 5 of this dissertation discusses theoretically what new information one expects to gain from using genotyped seed data rather than the traditionally available non-genotyped seed-trap data. The findings are that the gains may be substantial, particularly in identifying the true probability of long-distance dispersal, since most inverse modeling schemes, as noted in Chapter 4, tend to under-estimate this upper tail of the dispersion distribution. For the FDP data to which we have devoted so much analysis in this dissertation, both the censored data approach of Section 5.2 and the direct estimation approach of Section 5.3 confirmed these results, with both yielding substantially higher probabilities of dispersal events greater than $100m$ than given by any of the indirect estimation ('inverse modeling') techniques of Chapter 4 when they were applied to the historical ungenotyped data collected from 1987-2002. Nonetheless, there is something unsettling about the results from Sections 5.2 and 5.3; although they both yield much higher estimates of LDD events than do the indirect methods, they do not come close to agreeing with each other. The censored data approach of Section 5.2 as published in Jones et. al. [14], has a much higher estimate of LDD ($>100m$) dispersion than does the direct estimation method of Section 5.3. Specifically, as shown in Table 5.1, the censored data approach (given the placement of trees and traps in the FDP) predicts about 23% of all seeds to be dispersed $>100m$, with the mean dispersal distance for these LDD seeds to be about $350m$. For the best fitting of the direct estimation models of Section 5.3, at most 13% of the seeds are expected to be LDD dispersed, and the mean dispersal

distance for these seeds is much less than $350m$, since the bulk of those so dispersed are in the $100-150m$ or $150-200m$ range, as can be seen from Table 5.4.

This is not a minor discrepancy that can be explained by sampling variability, so why does it occur? The immediately obvious answer is that the discrepancy occurs because the censored data approach of Section 5.2 used data from all 726 genotyped seeds, while the direct estimation approach of Section 5.3 used only the data for the 616 matched seeds, omitting the 110 unmatched seeds. However, as argued in Section 5.3, such omission is correct – if the 110 seeds arose from trees different than the 306 (292 adult) which were genotyped, then they are extraneous information, and discarding them is no different than discarding the many seeds from species other than *Jacaranda* which were obtained in the seed traps. One might argue that this is 'biasing' the results against LDD events, but it really is not, as there were plenty of tree-trap combinations at distances greater than $100m$ among the tree-traps in the network on the FDP (+buffer zone). Indeed, as shown in Table 5.4, of the $292*298=87016$ such tree-trap combinations, 96% occur at such long distances. It is also true that only 10% of the 616 matched seeds fell in such traps, with the other 90% falling in the 4% of the traps which were in the non-LDD ($<100m$) range.

So, perhaps the problem lies in the assumptions of the censored data approach. It was noted briefly in Section 5.2 that an approximation was made there which might not be quite valid. Let us re-examine this. For the censored data approach to be correct, the 726 seeds should be a random sample of all seeds in the area. That is not quite correct, since the seeds are a sample (probably not quite at random, either) from the 298 traps in the area, and the 200 network traps were not placed at random, but relatively near paths which were accessible. The 98 'gap-traps' that Andy Jones placed (see Figure 3.1) somewhat attempted to alleviate this, but also are not truly randomly placed. Nonetheless, this assumption of randomness for the tree-trap placement is not entirely baseless. If one assumed that the 292 genotyped adult trees within the FDP(+buffer zone) were fixed in location, and then randomly distributed the 298 traps within the $1000m*500m$ FDP area, the distribution of tree-traps by distance

class is not much different than what is shown in Table 5.4 for the 87016 tree-trap distances actually observed. The main difference is that the $0-5m$ class and the $5-10m$ class are under-represented, since the tree-trap network does not place traps directly beneath many trees. This might make some difference if one were really very interested in estimates at low distances, or if behavior of the function near the origin had a very large influence on the upper tail, but that does not seem to be the source of our problem. One might also argue that even if the trap placement was approximately random with respect to the 292 known genotyped trees, it might not be so with respect to the unknown off-site trees. This seems very unlikely to be the source of any discrepancy either, especially since there was no apparent trend to the location of unmatched seeds, such as being predominantly in the southwest corner of the FDP.

Although not mentioned in Section 5.2 or [14], there is a technical problem with the censored data likelihood of equation (5.1). That formulation is correct in one dimension, as if one observes patients from a known time zero until an event happens, but the patients are occasionally censored at some time $t$ such that we know the event has not occurred up to time $t$, but we do not know what happens afterwards. If one thinks of starting with the seed in a trap and tracing it back to its source, with the wall at the edge of the buffer zone corresponding to distance censoring, then the analogy seems to make sense. But we do not really know that the source of the unmatched seed is exactly in the direction of the nearest boundary, and we actually have more two-dimensional survival information, since we know that the seed did not match any of the 292 trees in the FDP(+ buffer zone). Trying to correctly express this 2-dimensional survival function for the censored data points is extremely difficult. What was actually done in Jones [14] imposes a conservative bound that causes the $(1 - F(x_i|\phi))$ terms in the censored part of the likelihood of equation (5.1) to be larger than is correct. The overall effect is that the censored distances estimated by the procedure used in [14] are, if anything, less than the true values! Thus, while there are technical deficiencies in the censored data approach of [14], it is very unlikely that these

are responsible for the discrenpancy between their results and those found from the direct estimation method of Section 5.3.

So, what, then, might be happening? Why is the number of unmatched seeds so incongruent with what we observe among the matched seeds? One possible answer is that at least some of the 110 unmatched seeds did not really originate from outside the buffer zone. Instead, they came from a known source, but either they (or the tree of origin) were genotyped incorrectly. Recall from Section 3.2 that all seeds which did not exactly match one of the genotyped trees were considered to have originated from outside the buffer zone. This is a reasonable assumption if one is sure that genotyping errors are very rare. However, if they occurred only 2.03% of the time, that alone explains all the discrepancy, since an error rate of 2.03%, if all 8 alleles are genotyped independently, yields an 84.8% of correct genotyping per seed, in accord with 616 matches out of 726 seeds genotyped. We investigated the literature on genotyping errors and found it to be somewhat extensive, but not particularly specific, in that we can not find any source which says "an error rate of $X$% per allele genotyped is common". There are many factors which affect error rates [1]. Two of the most important are the skill of the person doing the genotyping, and the size of the fragment being used to obtain the microsatellite DNA. For example, in our case, where both trees and seeds were genotyped, one suspects that it is much easier to obtain good results from the trees' leaf samples, since many are available, than it is from the small amount of material available from a $2mg$ seed. Indeed, this seems to be borne out by the data shown in Table 6.1 concerning 'missing alleles'. A missing allele occurs when neither allele shows up on the gel - a condition sometimes known as 'total dropout' in the genotyping literature. Of the 299 adult trees that were genotyped, 286 gave results at all 4 loci. In our analyses we used only the 291 which had either no missing alleles or at most one locus (2 alleles) completely missing. For the 726 seeds, notice that the standards were much lower, with 78 of the 726 seeds having one locus completely missing. The number of seeds with more alleles missing is not shown explicitly, although we know from Table 3.4 that 864 seeds were originally examined and 138 of them

failed to be genotyped, either because of lack of genetic material or because the number of 'total dropouts' was deemed too large. So, clearly, seeds are much harder to genotype than trees. From Table 3.4, we can also infer that the technicians analyzing the seed data improved in proficiency between 2000 (when 27% of the seeds could not be genotyped) and 2002 (when only 7% failed). This sort of dramatic improvement is usually indicative of a novice technician and might be an indication that s/he is prone to make other genotyping errors, as discussed next.

| Missing | Trees | Seeds |
|---|---|---|
| 0 | 286 | 648 |
| 2 | 5 | 78 |
| 4 | 1 | - |
| 6 | 6 | - |
| 8 | 1 | - |
| Total | 299 | 726 |

Table 6.1: Missing Alleles for Genotyped Trees and Seeds

Total dropout is not common (from Table 6.1, one can infer that it occurred for only 0.4% of all tree loci and 2.7% of all seed loci used in the analyses), and it should not cause any particular error, since the information for that locus is simply missing. So, by itself, total dropout is more of a nuisance that causes loss of power than it is a source of bias. However, total dropout rate is a warning of a more severe bias-causing error, 'allelic dropout', also known as 'false homozygosity'. What occurs there is that only one of the two alleles shows up as a band on the gel, so the analyst scores this as a 'double band'. A 'double band' actually occurs if the locus is homozygous for the allele of interest; i.e. if the two alleles are the same. However, if the alleles are different and one drops out, the analyst will falsely record this as a homozygous pair. Unlike 'total dropout', which is easy to identify, one does not know for sure whether any particular homozygous result is a true homozygote or due to 'allelic dropout'. Of course, if the proportion of homozygotes among the genotyped loci is higher than expected theoretically, that is an indication that 'allelic dropout' is occurring. To see whether there is any evidence of this for our data set, consider the data in Table 6.2.

The rows represent the names of the four loci used in obtaining the microsatellite DNA from the *Jacaranda* trees and seeds, chosen as described in Jones and Hubbell [13] from among 11 possible loci. These loci, at positions 9, 18, 21, and 31, have different numbers of alleles ever observed, ranging from a low of 8 levels for locus 21 to a high of 11 for locus 18. In general, a good locus is one that has many possible alleles, so that there is less chance for ambiguity in making matches. The first of the four right-hand columns in Table 6.2 displays the results for the 280 unique trees. (There were 292 genotyped adult trees used in the analyses of Chapter 5, but only unique patterns are considered here.) The 'DZ' column stands for 'double zero', which is the code for total dropout, and, as noted before, we see that it is rare for trees, with only 4 of the 280∗4=1120 tree loci experiencing the event. The column labeled 'HM' contains the number of homozygotes observed among the 280 trees at the given locus. It is not immediately obvious from the summary given whether this number is more or less than expected, or, in fact, what is expected. This is discussed more below.

The other three column-pairs on the right-hand side of Table 6.2 refer to various groups of seeds. One could list all 726, or the 616 matched and 110 unmatched, but just as we do not list all 292 trees, but rather only the 280 unique trees, we desire to consider unique seed patterns. From left to right, the groups are 251*us* (153 unique matched seeds + 98 unique non-matched seeds), 153*mus* (153 unique matched seeds) and 98*umn* (98 unique unmatched seeds).

| | | 280trees | | 251us | | 153mus | | 98umn | |
|---|---|---|---|---|---|---|---|---|---|
| Locus | Levels | DZ | HM | DZ | HM | DZ | HM | DZ | HM |
| 9 | 9 | 1 | 36 | 7 | 34 | 6 | 16 | 1 | 18 |
| 18 | 11 | 1 | 30 | 34 | 45 | 26 | 11 | 8 | 34 |
| 21 | 8 | 0 | 81 | 6 | 76 | 6 | 42 | 0 | 34 |
| 31 | 9 | 2 | 37 | 11 | 40 | 6 | 19 | 5 | 21 |
| Total | | 4 | 184 | 58 | 195 | 44 | 88 | 14 | 107 |

Table 6.2: Total and Allelic Dropout for Trees and Seeds

The first of the 3 seed columns (251*us*) contains all 251 unique seeds patterns, with 153 being from the 616 matched seeds and 98 being from the 110 unmatched seeds (obviously,

there are many more 'repeats' for the matched seeds). To see how the joint distribution of the alleles for 280 unique trees and 251 seeds compare, one must examine Tables 6.3-6.6, for loci 9, 18, 21, and 31, respectively. In each table, the top panel contains the tree data and the bottom contains the seed data. The total in the bottom right hand corner of each table is usually slightly less than 280 (251) because the 'DZ' loci are not included. All the matrices shown are upper-right triangular since the genotyping procedure can not differentiate between (A,B) and (B,A), and arbitrarily stores all such heterozygote counts in the $A<B$ cell. The values shown as the row and column headings are the allele values and represent the lengths of the di-nucleotide tandem repeats (DNTR) found at that microsatellite locus. At loci 9, 21, and 31 the values are (almost) always even integers, whereas at locus 18 they are (almost) always odd integers. The skip between levels is usually two because these are 'di-nucleotide' repeats, so the smallest level of increase is 2 units.

| JACC9 | 172 | 178 | 180 | 182 | 184 | 186 | 188 | 190 | 192 | 194 | 196 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 172 | 5 | 3 | 24 | 12 | 6 | 11 | 14 | 10 | 0 | 1 | 0 | 86 |
| 178 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 4 |
| 180 | 0 | 0 | 9 | 19 | 12 | 7 | 14 | 12 | 4 | 4 | 0 | 81 |
| 182 | 0 | 0 | 0 | 6 | 9 | 2 | 16 | 7 | 1 | 4 | 1 | 46 |
| 184 | 0 | 0 | 0 | 0 | 4 | 5 | 11 | 3 | 0 | 3 | 0 | 26 |
| 186 | 0 | 0 | 0 | 0 | 0 | 4 | 8 | 2 | 0 | 0 | 0 | 14 |
| 188 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 9 | 1 | 4 | 0 | 19 |
| 190 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 |
| 194 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| Total | 5 | 3 | 34 | 38 | 32 | 29 | 69 | 45 | 6 | 17 | 1 | 279 |
| JACC9 | 172 | 178 | 180 | 182 | 184 | 186 | 188 | 190 | 192 | 194 | 196 | Total |
| 170 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 172 | 5 | 1 | 11 | 11 | 5 | 4 | 16 | 10 | 0 | 0 | 0 | 63 |
| 178 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 180 | 0 | 0 | 8 | 12 | 9 | 3 | 16 | 7 | 6 | 4 | 0 | 65 |
| 182 | 0 | 0 | 0 | 2 | 12 | 2 | 21 | 3 | 0 | 5 | 0 | 45 |
| 184 | 0 | 0 | 0 | 0 | 2 | 7 | 10 | 5 | 0 | 1 | 0 | 25 |
| 186 | 0 | 0 | 0 | 0 | 0 | 3 | 12 | 2 | 0 | 1 | 0 | 18 |
| 188 | 0 | 0 | 0 | 0 | 0 | 0 | 13 | 8 | 0 | 4 | 0 | 25 |
| 190 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Total | 5 | 1 | 20 | 25 | 28 | 19 | 89 | 36 | 6 | 15 | 0 | 244 |

Table 6.3: Locus 9 Distribution of Allele Pairs (Top – Trees, Bottom – Seeds)

| JACC18 | 269 | 271 | 275 | 277 | 279 | 281 | 283 | 285 | 287 | 289 | 293 | 297 | 299 | 303 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 269 | 6 | 6 | 8 | 10 | 8 | 3 | 7 | 1 | 1 | 0 | 0 | 21 | 0 | 0 | 71 |
| 271 | 0 | 2 | 2 | 8 | 3 | 1 | 7 | 2 | 2 | 0 | 2 | 17 | 2 | 0 | 48 |
| 275 | 0 | 0 | 1 | 4 | 4 | 1 | 4 | 1 | 0 | 0 | 1 | 9 | 0 | 0 | 25 |
| 277 | 0 | 0 | 0 | 2 | 2 | 4 | 6 | 2 | 2 | 0 | 0 | 16 | 1 | 0 | 35 |
| 279 | 0 | 0 | 0 | 0 | 3 | 0 | 3 | 2 | 4 | 0 | 0 | 18 | 1 | 0 | 31 |
| 281 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 8 | 0 | 0 | 12 |
| 283 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 12 | 3 | 0 | 20 |
| 285 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 8 | 0 | 0 | 10 |
| 287 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 293 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 2 | 0 | 6 |
| 297 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 2 | 1 | 19 |
| Total | 6 | 8 | 11 | 24 | 20 | 9 | 27 | 13 | 13 | 1 | 4 | 131 | 11 | 1 | 279 |

| JACC18 | 269 | 271 | 275 | 277 | 279 | 281 | 283 | 285 | 287 | 289 | 293 | 297 | 299 | 303 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 269 | 13 | 8 | 5 | 19 | 4 | 1 | 2 | 0 | 0 | 1 | 0 | 16 | 0 | 0 | 69 |
| 271 | 0 | 6 | 1 | 5 | 2 | 1 | 5 | 3 | 1 | 0 | 0 | 15 | 1 | 0 | 40 |
| 275 | 0 | 0 | 2 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 13 |
| 277 | 0 | 0 | 0 | 4 | 0 | 2 | 4 | 1 | 0 | 0 | 0 | 15 | 1 | 0 | 27 |
| 279 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 2 | 2 | 0 | 0 | 8 | 2 | 0 | 17 |
| 281 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 0 | 4 |
| 283 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 0 | 9 | 1 | 0 | 15 |
| 285 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 4 | 0 | 0 | 7 |
| 287 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 288 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 293 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| 294 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| 295 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 297 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 1 | 0 | 17 |
| Total | 13 | 14 | 8 | 28 | 10 | 4 | 14 | 10 | 6 | 1 | 1 | 99 | 9 | 0 | 217 |

Table 6.4: Locus 18 Distribution of Allele Pairs (Top – Trees, Bottom – Seeds)

From Table 6.1 we had noted that the seeds had a much higher rate of total dropout than the trees, and from Table 6.2, we can see that locus 18 seems to be the main offender, followed by locus 31, with loci 9 and 21 occurring the least (but all occurring at rates significantly higher than for the 280 trees). Since total dropout is a warning sign for 'allelic dropout', we might expect this same sort of pattern by locus. What we hope to glean from the data in Table 6.2 and Tables 6.3-6.6 is whether there is evidence of excess homozygosity in the seeds relative to the trees, especially for the unmatched seeds. The standard procedure used in genetics to measure excess homogeneity is the deviation from Hardy-Weinberg equilibrium. In statistical terms, the Hardy-Weinberg model simply states that the alleles pair with one another independently in the population. We're not particularly interested in testing that hypothesis – we're more interested in seeing if there is a trend toward homozygosity for the

| JACC21 | 230 | 232 | 234 | 236 | 240 | 242 | 244 | 246 | 248 | 250 | 252 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 230 | 50 | 4 | 6 | 102 | 15 | 13 | 0 | 3 | 4 | 10 | 0 | 207 |
| 232 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 3 |
| 234 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 4 |
| 236 | 0 | 0 | 0 | 25 | 7 | 10 | 0 | 0 | 4 | 9 | 0 | 55 |
| 240 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 3 |
| 242 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 3 |
| 248 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| 250 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 3 |
| Total | 50 | 4 | 6 | 130 | 23 | 26 | 1 | 3 | 11 | 26 | 0 | 280 |
| JACC21 | 230 | 232 | 234 | 236 | 240 | 242 | 244 | 246 | 248 | 250 | 252 | Total |
| 230 | 39 | 7 | 5 | 90 | 12 | 8 | 0 | 0 | 2 | 11 | 1 | 175 |
| 232 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| 233 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 236 | 0 | 0 | 0 | 33 | 5 | 8 | 0 | 0 | 1 | 12 | 0 | 59 |
| 242 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 3 |
| 248 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 250 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 4 |
| Total | 39 | 7 | 5 | 123 | 19 | 16 | 1 | 0 | 4 | 29 | 2 | 245 |

Table 6.5: Locus 21 Distribution of Allele Pairs (Top – Trees, Bottom – Seeds)

seeds as opposed to the trees, especially for the ungenotyped seeds. To do this, we calculate the expected number of homozygotes at a locus under H-W equilibrium, as:

$$E(HM) = (\sum_{i=1}^{k} N_i^2)/(4N), \tag{6.1}$$

where $k$ is the number of allele classes for that locus, $N_i$ is the number of occurrences of allele $i$ among the $2N$ alleles, and $N$ is the number of non-dropout individuals observed at that locus. This expected number of homozygotes is then compared with the observed number of homozygotes (the values listed as 'HZ' in Table 6.2; the sum of the respective diagonals in Tables 6.3-6.6) and a proportion of excess homozygosity, PEH (which may be negative), is calculated as:

$$PEH = \frac{Obs - E(HM)}{E(HZ)} = \frac{Obs - E(HM)}{N - E(HM)} \tag{6.2}$$

| JACC31 | 137 | 138 | 140 | 142 | 144 | 146 | 148 | 150 | 152 | 154 | 156 | 169 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 138 | 0 | 0 | 3 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 7 |
| 140 | 0 | 0 | 16 | 33 | 0 | 15 | 43 | 8 | 6 | 18 | 4 | 0 | 143 |
| 142 | 0 | 0 | 0 | 7 | 3 | 15 | 16 | 4 | 2 | 8 | 2 | 0 | 57 |
| 144 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 1 | 0 | 0 | 5 |
| 146 | 0 | 0 | 0 | 0 | 0 | 1 | 9 | 3 | 4 | 5 | 3 | 0 | 25 |
| 148 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 10 | 2 | 10 | 3 | 1 | 36 |
| 150 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 152 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 2 |
| 154 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| Total | | | 19 | 41 | 3 | 31 | 82 | 26 | 16 | 45 | 14 | 1 | 278 |
| JACC31 | 137 | 138 | 140 | 142 | 144 | 146 | 148 | 150 | 152 | 154 | 156 | 169 | Total |
| 136 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 137 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 138 | 0 | 1 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | 11 |
| 140 | 0 | 0 | 20 | 28 | 0 | 18 | 44 | 9 | 2 | 14 | 2 | 0 | 137 |
| 142 | 0 | 0 | 0 | 7 | 2 | 13 | 9 | 5 | 3 | 7 | 1 | 0 | 47 |
| 144 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 3 |
| 146 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 0 | 7 | 1 | 0 | 12 |
| 148 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 10 | 0 | 6 | 2 | 0 | 23 |
| 150 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 154 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 4 |
| Total | 1 | 1 | 23 | 38 | 2 | 33 | 62 | 25 | 5 | 43 | 7 | 0 | 240 |

Table 6.6: Locus 31 Distribution of Allele Pairs (Top – Trees, Bottom – Seeds)

Note that this is not a statistic for calculating significance from the hypothesis of Hardy-Weinberg equilibrium, but simply a measure of excess. The Observed, Expected, and PEH values for the trees, unique seeds, unique matched seeds, and unique unmatched seeds are shown in Table 6.7. Note that for the unique tree set, there is a tendency for excess heterozygozity (expressed as negative excess homozygosity, PEH). This is not uncommon in natural populations, where greater diversity in alleles may increase the odds of survival. For the 251 unique seeds, all the PEH's are more positive than they were for the trees (although only that for Locus 18 is itself positive) indicating the expected trend toward homozygosity. If one decomposes the 251 seeds into those from the 153 unique matched seeds and those from the 98 unique unmatched seeds, the difference becomes very striking. The unique matched seeds

have a PEH pattern similar (or perhaps slightly more negative) than the original trees. This is not too surprising, since these seeds match with some subset of the trees. However, the 98 unmatched seeds have a much higher homozygosity proportion, with all loci being positive. Roughly speaking, the increase in homozygosity rate for the unmatched seeds relative to the trees appears to be 4%, 30%, 8%, and 12% for the four loci, respectively. Of course, these are rough estimates and far above the true rate of excess homozygosity (since we've selectively chosen the non-matched seeds), but it is fairly convincing evidence that allelic dropout, especially at loci 18 and 31, may be the cause of some of the unmatched seeds.

| | 280 Unique Trees | | | 251 Unique Seeds | | | 153 Unique Matched Seeds | | | 98 Unique Unmatched Seeds | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LOCUS | HM | E(HM) | PEH | HM | E(HM) | PEH | HM | E(HM) | PEH | HM | E(HM) | PEH |
| 9 | 36 | 39.55 | -.015 | 34 | 36.43 | -.012 | 16 | 21.10 | -.012 | 18 | 15.80 | +.027 |
| 18 | 30 | 38.15 | -.034 | 45 | 33.00 | +.065 | 11 | 19.82 | -.082 | 34 | 14.26 | +.261 |
| 21 | 81 | 91.93 | -.058 | 76 | 82.52 | -.040 | 42 | 51.05 | -.094 | 34 | 32.36 | +.025 |
| 31 | 37 | 50.79 | -.061 | 40 | 47.09 | -.037 | 19 | 30.79 | -.101 | 21 | 16.97 | +.053 |

Table 6.7: Excess Homozygosity for Unique Trees and Seeds

A third kind of error that can occur in genotyping is 'binning error'. The actual number of DNTR's is an integer, but the distance slid on the gel is continuous, so there is some sort of binning convention to assign alleles to the closest even (or odd, for locus 18) integer. Sometimes the drift is too large and the allele is assigned to the wrong bin. Di-nucleotide tandem repeats seem more vulnerable to this error than others, which is why much modern genotyping, if tandem repeats are utilized to measure alleles, uses tri-nucleotide or tetra-nucleotide tandem repeats. Unfortunately, DNTR was the state of the art in 2000-2002, so we must do the best we can with it. There are no good literature estimates for how often such binning errors occur, although all models for such have a much higher probability for a shift of $\pm 2$ units than for $\pm 4$ units, for example. There is also some evidence that binning errors, unlike allelic dropout, may not occur independently for different loci from the same individual.

Finally, there is always the possibility of human recording error. Although the genotyping process was somewhat automated, it is clear that some human intervention occurred. For example, for locus 21, '236' is a very common allele value, but one seed had '326', obviously

a transcription error, recorded. Similarly, one allele for locus 18 was recorded as '27', when it was obviously something in the range of $270-279$. There were also some judgment calls made by Andy Jones concerning when a match occurred if six of the alleles matched perfectly and the other two were allelic dropouts. Careful examinations of this kind were what allowed the number of unmatched seeds to drop from 119 at the time of the original analysis to 116 at the time of the publications of the Jones et. al [14] article to 110 at the current time. One of the largest changes from unmatched to matched (7 seeds) is worth mentioning. It arose from our examination of the unmatched seeds to see if any of these uniquely matched each other. If so, that might be an indication that there was some large off-site tree whose seeds were drifting onto the FDP. For the most part, the answer to that question was 'no'; the 110 seeds currently classified as unmatched contain 98 unique patterns. However, there was one group of 7 seeds in 3 nearby traps which matched perfectly (or perfectly with one allelic dropout) with each other but not with any genotyped tree. These three traps (#35, #36, and #37) are all very near the boundary of the FDP. We have examined all nearby trees' alleles and can not find any which seem like they are matches with minor recording errors. We are sure that there must be a tree with the exact pattern of these 7 seeds very near the traps, most likely at the very edge of the buffer zone near the FDP. We are not sure why this tree was not genotyped, but feel sure that it exists and have added it as a $292nd$ adult genotyped tree in our analyses. We do not think that it is at all credible that these 7 seeds come from some super-tree over $100m$ away, out of the buffer zone. Although there is certainly evidence in the data set of seeds dispersing more than $100m$ from their source, there is no evidence at all of one tree dispersing so many seeds over $100m$. Even the most fecund tree which we observed, which dispersed 96 seeds in total, did not have 7 caught seeds dispersed more than $30m$ away. In any case, we were able to detect some human errors and correct them, but there may be others which are still lurking in the data and not amenable to statistical detection.

## 6.2 Global Estimation of Misclassification Proportion

The previous section discussed genotyping errors and gave some evidence that they have occurred in the set of 726 genotyped seeds, but with little attempt to quantify exactly how often they occurred. This section attempts to produce estimates and confidence bounds for the proportion of misclassified seeds by two very different methods, the Hypergeometric-Poisson method and the Density-Matching method, with both demonstrating that a substantial proportion of the 110 unmatched seeds are likely misclassified. Both methods, while developed for the FDP data set, are easily modified to estimate misclassification for similarly collected genotyped data sets.

Both methods were developed because the distribution of unmatched seeds did not "look right". One expects such seeds, if they truly were from off-site trees, to be more concentrated in the traps near the edges of the FDP, but there was no evidence of this at all. As a matter of fact, the best predictor of presence of an unmatched seed in a trap is not the trap's location, but how many matched seeds are in the trap. This does not make much sense if the seed is really from off-site, but seems quite reasonable if one believes that a certain proportion of these seeds are, in fact, genotyped seeds that have been misclassified. A display that partially demonstrates this is Table 6.8. There, each of the 298 traps is classified according to how many non-matched seeds were present in the trap. Each category was then sub-divided into those traps that contained at least 1 matched seed and those with no matched seeds. From the last column, we see that there were 113 traps among the 298 which never had any seeds sampled from them. (Or, possibly seeds were sampled, but they were among the 138 for which genotyping failed.) Of the remaining 185 traps, there were 17 with only non-matched seeds, 106 with only matched seeds, and 62 with both matched and non-matched seeds. A simple 2×2 odds-ratio statistic for association of matched and unmatched presence for the 298 traps yields a value of 3.89, very strongly indicating ($p<0.0001$) that the two types of seeds are not independent, but positively associated with one another. This makes no sense at all, since the off-site seeds, if the unmatched seeds are indeed from off-site, should behave

approximately independently of the matched seeds. Indeed, if anything, one might expect a negative association, since traps near the center of the FDP are expected to receive seeds from on-site trees and rarely from off-site, while the converse would be true for traps near the boundaries.

| Non-Matched Seeds | Total Traps | MatchSeed>0 Traps | MatchSeed=0 Traps |
|---:|---:|---:|---:|
| 4 | 1 | 1 | 0 |
| 3 | 6 | 5 | 1 |
| 2 | 16 | 15 | 1 |
| 1 | 56 | 41 | 15 |
| 0 | 219 | 106 | 113 |
| Total | 298 | 168 | 130 |

Table 6.8: Distribution of Traps by Number of Non-Matched Seeds

The fact that independence was so strongly rejected caused us to formulate a model for the number of unmatched seeds in a trap. The simplest version of this, called the Hypergeometric-Poisson model, can be written as:

$$Pr(X = x|T, n) = \sum_{k=x}^{T-n+x} \frac{\binom{k}{x}\binom{T-k}{n-x}}{\binom{T}{n}} \frac{\lambda^k}{k!} exp(-\lambda) \tag{6.3}$$

where

$x$ = observed number of unmatched seeds in a trap

$T$ = Total number of seeds (genotyped & ungenotyped) caught in trap

$n$ = Total number of seeds (matched and unmatched) sampled from trap

$\lambda$ = Poisson parameter.

What this model means is that there is an unknown Poisson parameter, $\lambda$, which governs the unobserved number $k$ of off-site seeds that land in a trap. Assuming no genotyping errors, this number $k$ must be at least as large as the observed number of non-matched seeds $x$ which are observed when a sample of $n$ seeds is chosen at random from the $T$ seeds which were caught in the trap. The values for $x$ and $n$ are observed for all 298 traps, although, of

course, for the 113 traps where $n$=0, there is no likelihood to be calculated. The value of $T$ for the 200 network traps is known (or taken) to be the sum of the number of seeds observed in that trap in the years 2000 & 2002. For the gap-traps, since they are not part of the official network, we do not know what $T$ is, since Andy Jones simply reported the number of seeds he genotyped from these traps, not the total collected for the bi-year period. For the network traps, the ratio of T/n for most traps was in the range $5<(T/n)<10$, so we tried using both $T$=10×$n$ and $T$=5×$n$ for the gap-traps, with $T$=5×$n$ yielding slightly better fits. The maximum likelihood estimate of $\lambda$, if one maximizes the likelihood from equation (6.3) over all 185 traps for which any of the 726 genotyped seeds were found is, $\hat{\lambda}$=1.90. Goodness of fit according to the G-squared statistic is poor ($G^2/df$=478/184=2.60), indicating that this model completely fails to model the situation.

The fit above is the best which one can achieve if one holds steadfastly to the belief that there are no genotyping errors - every unmatched seed truly originated from a tree beyond the buffer zone. An improved generalization of the model allows introduction of another parameter, $P$, which is the probability that an on-site seed is incorrectly categorized as a non-matched seed. In that case, the model of (6.3) is generalized to:

$$Pr(X = x|T, n) = \sum_{k=x}^{T-n+x} \sum_{x_f=0}^{x} \frac{\binom{k}{x_t}\binom{T-k}{n-x_t}}{\binom{T}{n}} \frac{\lambda^k}{k!} exp(-\lambda) \times \binom{n - x_t}{x_f} P^{x_f}(1 - P)^{n-x} \qquad (6.4)$$

where

$x$ = observed number of unmatched seeds in a trap

$T$ = Total number of seeds (genotyped & ungenotyped) caught in trap

$n$ = Total number of seeds (matched and unmatched) sampled from trap

$\lambda$ = Poisson parameter,

$P$ = Binomial misclassification parameter.

In this formulation, the observed number of unmatched seeds, $x$, is decomposed into two unobserved parts, $x$=$x_f$+$x_t$, representing the false (i.e. misclassified) and true (i.e. from off-site) unmatched seeds. As with the unobserved $k$, the possible values of $x_f$ are summed over

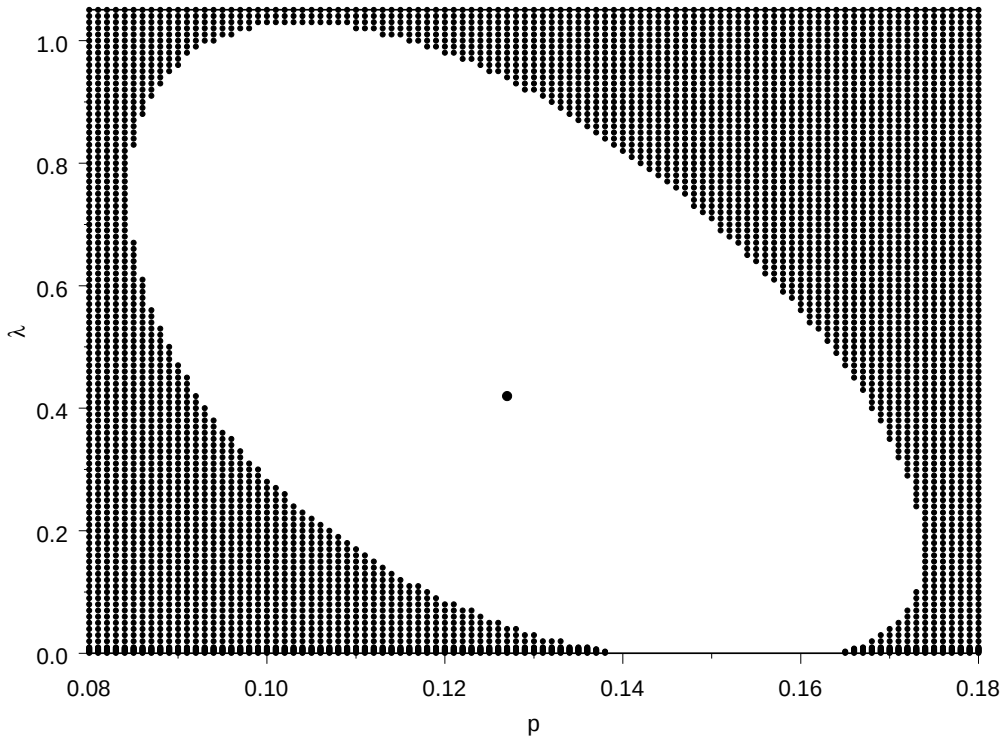


Figure 6.1: 95% Joint Confidence Ellipsoid for ($p$, $\lambda$)

when calculating the likelihood. When this likelihood was evaluated over all 185 traps containing genotyped seeds, the MLE's were ($\hat{P}$, $\hat{\lambda}$)= (0.127, 0.42), with ($G^2/df$=343/183=1.87). This fit is still suspect, but vastly superior to that from assuming that $P$=0, as above. Figure 6.1 uses the profile likelihood method (adjusted for over-dispersion) to find a joint 95% confidence ellipsoid for ($P$, $\lambda$). Note that ($P$, $\lambda$)=(0, 1.90), the original solution obtained above, is far from the ellipse, indicating that the assumption of no genotyping error is very unrealistic. The MLE point estimate at the center of the ellipse corresponds to the situation where only 20 of the 110 unmatched seeds are real; the other 90 are mistakes! Of course, this is just a point estimate, and there is a fair amount of variability, as indicated by the profile. The further to the left one moves in the ellipse, the lower $P$ is, and, hence, the fewer of the 110 unmatched seeds are considered erroneous. The upper left bound of the ellipse,

near the point ($P$, $\lambda$)=(.085, 0.68) corresponds to an expectation of 53 real off-site seeds and 57 misgenotyped seeds. So, by the Hypergeometric-Poisson approximation method, there is fairly convincing evidence that at least half of the 110 unmatched seeds may be due to genotyping mistakes.

| DC | Distance | MDPT | Tree_trap | Match616 | Fictreetrap | R616o | E616o | L616f | R616f | E616f |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0-5m | 2.5 | 7 | 8 | 0 | 1.1429 | 0.00 | 1.51 | 4.5174 | 0.00 |
| 2 | 5-10m | 7.5 | 33 | 67 | 0 | 2.0303 | 0.00 | 0.66 | 1.9285 | 0.00 |
| 3 | 10-15m | 12.5 | 48 | 75 | 0 | 1.5625 | 0.00 | 0.07 | 1.0732 | 0.00 |
| 4 | 15-20m | 17.5 | 93 | 98 | 0 | 1.0538 | 0.00 | -0.41 | 0.6665 | 0.00 |
| 5 | 20-25m | 22.5 | 88 | 76 | 0 | 0.8636 | 0.00 | -0.82 | 0.4415 | 0.00 |
| 6 | 25-30m | 27.5 | 115 | 33 | 0 | 0.2870 | 0.00 | -1.19 | 0.3055 | 0.00 |
| 7 | 30-35m | 32.5 | 129 | 55 | 0 | 0.4264 | 0.00 | -1.52 | 0.2183 | 0.00 |
| 8 | 35-40m | 37.5 | 153 | 21 | 0 | 0.1373 | 0.00 | -1.83 | 0.1600 | 0.00 |
| 9 | 40-45m | 42.5 | 177 | 47 | 0 | 0.2655 | 0.00 | -2.12 | 0.1196 | 0.00 |
| 10 | 45-50m | 47.5 | 187 | 12 | 0 | 0.0642 | 0.00 | -2.40 | 0.0909 | 0.00 |
| 11 | 50-55m | 52.5 | 193 | 14 | 0 | 0.0725 | 0.00 | -2.66 | 0.0701 | 0.00 |
| 12 | 55-60m | 57.5 | 233 | 10 | 0 | 0.0429 | 0.00 | -2.91 | 0.0547 | 0.00 |
| 13 | 60-65m | 62.5 | 235 | 13 | 0 | 0.0553 | 0.00 | -3.14 | 0.0431 | 0.00 |
| 14 | 65-70m | 67.5 | 239 | 4 | 0 | 0.0167 | 0.00 | -3.37 | 0.0343 | 0.00 |
| 15 | 70-75m | 72.5 | 261 | 0 | 0 | 0.0000 | 0.00 | -3.59 | 0.0276 | 0.00 |
| 16 | 75-80m | 77.5 | 285 | 5 | 0 | 0.0175 | 0.00 | -3.80 | 0.0223 | 0.00 |
| 17 | 80-85m | 82.5 | 282 | 5 | 0 | 0.0177 | 0.00 | -4.01 | 0.0182 | 0.00 |
| 18 | 85-90m | 87.5 | 328 | 4 | 0 | 0.0122 | 0.00 | -4.21 | 0.0149 | 0.00 |
| 19 | 90-95m | 92.5 | 331 | 3 | 0 | 0.0091 | 0.00 | -4.40 | 0.0123 | 0.00 |
| 20 | 95-100m | 97.5 | 329 | 3 | 0 | 0.0091 | 0.00 | -4.59 | 0.0101 | 0.00 |
| 21 | 100-150m | 125 | 4062 | 18 | 41 | 0.0044 | 0.18 | -5.66 | 0.0035 | 0.14 |
| 22 | 150-200m | 175 | 5173 | 11 | 360 | 0.0021 | 0.77 | -6.26 | 0.0019 | 0.69 |
| 23 | 200-300m | 250 | 12740 | 13 | 3188 | 0.0010 | 3.25 | -6.84 | 0.0011 | 3.40 |
| 24 | 300-400m | 350 | 13579 | 10 | 8342 | 0.0007 | 6.14 | -7.43 | 0.0006 | 4.96 |
| 25 | 400-500m | 450 | 12381 | 4 | 15412 | 0.0003 | 4.98 | -7.90 | 0.0004 | 5.70 |
| 26 | 500-600m | 550 | 10476 | 4 | 24600 | 0.0004 | 9.39 | -8.31 | 0.0002 | 6.04 |
| 27 | 600-700m | 650 | 8529 | 1 | 33807 | 0.0001 | 3.96 | -8.68 | 0.0002 | 5.75 |
| 28 | 700-800m | 750 | 6810 | 2 | 42376 | 0.0003 | 12.45 | -9.01 | 0.0001 | 5.16 |
| 29 | 800-900m | 850 | 5155 | 0 | 50473 | 0.0000 | 0.00 | -9.32 | 0.0001 | 4.51 |
| 30 | 900-1000m | 950 | 3211 | 0 | 58445 | 0.0000 | 0.00 | -9.61 | 0.0001 | 3.91 |
| 31 | 1000-1100m | 1050 | 1037 | 0 | 65663 | 0.0000 | 0.00 | -9.89 | 0.0001 | 3.34 |
| 32 | 1100-1200m | 1200 | 117 | 0 | 67285 | 0.0000 | 0.00 | -10.27 | 0.0000 | 2.33 |
| 33 | 1200-1300m | 1250 | 0 | 0 | 64595 | 0.0000 | 0.00 | -10.39 | 0.0000 | 1.98 |
| 34 | 1300-1400m | 1350 | 0 | 0 | 58880 | 0.0000 | 0.00 | -10.63 | 0.0000 | 1.43 |
| 35 | 1400-1500m | 1450 | 0 | 0 | 50345 | 0.0000 | 0.00 | -10.86 | 0.0000 | 0.97 |
| 36 | 1500-1600m | 1550 | 0 | 0 | 40991 | 0.0000 | 0.00 | -11.08 | 0.0000 | 0.64 |
| 37 | >1600m | 1650 | 0 | 0 | 97982 | 0.0000 | 0.00 | -11.29 | 0.0000 | 1.23 |
| Total | | | 87016 | 616 | 682782 | | 41.12 | | | 52.16 |

Table 6.9: Estimated Off-site Seeds by Density Matching Method

One objection to the Hypergeometric-Poisson method is that it does not take distance into account, since it assumes the same Poisson intensity ($\lambda$) for each trap to experience an off-site

seed landing. A very straight-forward method which does take distance into account is the Density-Matching method. This method can best be illustrated by examination of Table 6.9. The first 5 columns of this table are identical to those of Table 5.4, showing the distribution of the 616 matched seeds ($Match616$) relative to the distribution of the 87016 tree-trap distances. This allows calculation of the crude observed rate ($R616o$=$Match616/Tree\_trap$) shown in the '$R616o$' column of Table 6.9. The column labeled 'Fictreetrap' is of much relevance here. It is calculated by assuming that the adult *Jacaranda* trees beyond the buffer zone (the source from which true unmatched seeds must have arisen) have the same spatial density as in the FDP+buffer zone (292 trees/$840000m^2$), and then arranging these fictitious trees in a grid pattern with this density up to $900m$ away from the outer edge of the buffer zone in all directions. Next, for each of these fictitious off-site trees, the distance between it and each of the 292 traps on the FDP was calculated and binned, as shown in Table 6.9. Of course, the minimum possible distance class for these trees is $100-150m$, and this event would be rare (only 41 expected occurrences), since it would require both the existence of a tree near the outer buffer zone boundary and a trap in the FDP near the inner buffer zone boundary. The most typical distance classes are very large, such as 1200-1300$m$. The Density-Matching method now simply says that the off-site trees will have the same seed catching rate as was observed for the matched seeds in those same classes; i.e. the column of the table given by $E616o$=Fictreetrap$\times R616o$. For the first 20 distance classes, the value of $R616o$ is irrelevant, since there are no off-site trees within those ranges. However, as the distance increases beyond $100m$, more tree-trap combinations become eligible. The empirical rates estimated by $R616o$ are very low beyond $100m$, but the number of possible fictitious tree-trap combinations is high, so a non-trivial expected value accrues in the $E616o$ column. It sums to 41.12, so if the empirical rate were exactly correct for the off-site trees, we expect about 41 real off-site seeds to appear in our sample (and, thus, the other 69 unmatched seeds must be due to genotyping error). This point-estimate is certainly consistent with the values found by the previous method, but a confidence interval calculation remains elusive.

This estimate is clearly rather variable, since the empirical $R616o$ estimate is so erratic. For example, one does not really believe that the true rate for 70-75$m$ is zero just because no 70$m$ distance seeds were observed for the 616 matched seeds. Similarly, although there were no observed matched seed distances at greater than 800$m$ does not mean that this could never happen. It is quite possible that the true rate for this category is so low that it was never observed in the 5155 network tree-traps in that category, but with 10 times as many observations in that category for the fictitious trees, it might occur. Another point to consider here is the large influence played by the two observations in the 700-800$m$ range. If either one of these seeds had not been present, the estimate in $E616o$ would decrease by 6 seeds. As usual when making inferences about extremes in the tails, there is little data to go upon and results depend heavily upon what one wants to assume about tail behavior. The '$L616f$' column represence a parametric fit to the log-intensity for the 616 seed dataset which is very similar to '$lqn$' of Table 5.4 for $d < 100m$, but closer to '$en$' for $d > 100m$. The corresponding '$R616f$' and '$E616f$' columns display the rate and number of off-site seeds which would be expected under this log-intensity, assuming that the off-site adult *Jacaranda* trees have the same spatial density and fecundity distribution as the 292 on-site adult trees. This provides slightly smoother log-intensity estimates and a slightly larger estimate (52.16) of the expected number of real non-matched seeds which should be present. So, in summary, the point estimate based on the empirical 616 seed rate, the fitted 616 seed rate (and indeed estimates based on any of the six distance functions examined in Section 5.3), all lead to the conclusion that at least half of the non-matched seeds are misgenotyped.

## 6.3 Corrections for Misclassification and Revised Estimates

Section 6.1 of this chapter discussed the possibility of genotyping errors occurring and produced some evidence that both 'total dropout' and 'allelic dropout' had occurred during the seed genotyping. Section 6.2 produced two methods to estimate the global proportion of misclassification, and both agreed that at least half of the 110 non-matched seeds might

indeed be seeds which should have been matched to an on-site source. These are nice results from a theoretical perspective, but from a practical viewpoint, one desires to know precisely which seeds were the 'mistakes', so that these can be corrected, enabling the methods introduced in Chapter 5 to be applied correctly. Unfortunately, estimating global misclassification rates is much easier than specifying exactly which unmatched seeds are 'wrong'. We initially attempted to do this by eye, believing that it would be easy to find non-matches which differed by one allele from known tree sources. While it is true that some non-matched seeds immediately became apparent as matched seeds with an allelic dropout, the process was not nearly as easy as one might think, since many genotyped trees (which we believe to be correct and unique) have very similar profiles. We must ensure that we do not go overboard in switching an assignment from 'unmatched' to 'matched' based on unsound reasoning. Eventually we decided to evaluate every seed(k) in a trap(j) by calculating a score that measures its probabilistic distance to each source tree(i) (including a fictitious perfectly matched off-site source), with the seed then being matched to the tree which gave the best score. This is what Andy Jones did initially, too, although his scoring algorithm was a very simple one that matched a seed to the nearest tree which yielded a perfect match on all 8 alleles and to an off-site 'non-match' if this did not occur. This yielded 552 perfect matched seeds. He modified this slightly later to allow matches to also occur if there were an agreement on all six observed alleles and total dropout for another pair. This added 64 more matches, for the current total of 552+64=616 matches. Our method includes Dr. Jones' method as a special case, but allows more flexibility in making other assignments.

From Chapters 4 and 5, we already have some measures of how likely tree(i) is to deposit a seed in trap(j), such as:

$$ln(\lambda_{ij}) = \beta_0 + \beta_1 * q_{ij} + \beta_2 * NAFX(i),$$

so this is a good start in measuring the log(likelihood) that a seed from tree(i) ends up in trap(j). If the seed(k) in trap(j) is a perfect match to tree(i), the above is a good score to use. However, if the seed is not a perfect match to tree(i), it is still possible that tree(i)

is the true source, but that various genotyping errors have occurred. Such errors are rare (probabilistically unlikely), so they must be penalized in some way. Nevertheless, if after appropriate penalization, a certain tree has a higher score than a theoretical perfect match far away, the seed should be assigned to the source which yields the better score. Thus, we considered score functions of the form:

$$Z_{ik(j)} = \beta_0 + \beta_1 * q_{ij} + \beta_2 * NAFX(i) - cz * nz(k) - cd * nd(k,i) - cs * ns(k,i) \qquad (6.5)$$

where $i$, $j$, and $k$ refer to the tree, trap, and seed respectively, $cz$, $cd$, and $cs$ are some positive penalty coefficients to be determined, and

$nz(k)$ = Number of double zero ('total dropout') loci for seed $k$,

$nd(k,i)$ = Number of loci in seed $k$ which would match to tree $i$ if an allelic dropout event had occurred, and

$ns(k,i)$ = Total number of absolute bin shift units between the alleles of seed(k) and tree(i), excluding loci with $nz$ or $nd$ events.

An example illustrating this is shown in Table 6.10. Suppose that Tree(i) and Seed(k) alleles are as shown there. In this case, $nz(k)$=1 because of the allelic dropout at Locus 18. The value of $nd(k,i)$=1 also, because Seed(k)'s Locus 31 homozygote pair (148, 148) could be a match to Tree(i)'s Locus 31 pair (148, 154) if allelic dropout had occurred. The 4 alleles at the other two loci are then compared and summed to obtain $ns(i,k)$=0+0+|−2| + |4| = 6, since there is a perfect match at Locus 9, but discrepancies of −2 and +4 units at Locus 21.

| | Locus 9 | | Locus 18 | | Locus 21 | | Locus 31 | |
|---|---|---|---|---|---|---|---|---|
| Tree(i) | 180, | 182 | 283, | 297 | 236, | 236 | 148, | 154 |
| Seed(k) | 180, | 182 | 0, | 0 | 234, | 240 | 148, | 148 |

Table 6.10: Sample Tree-Seed Allele Comparison

There are several matters to consider in developing this score function and algorithm:

(a) What are the appropriate penalties ($cz$, $cd$, $cs$) for errors?

(b) What is the location of the hypothetical off-site perfect-match tree?

(c) How does one handle suggested switches among matched trees?

To be consistent with the results of section 6.2, the appropriate penalties must be determined by trial and error so as to leave only about 40-50 seeds as unmatched. A value of $cz$ twice as large as $cd$ may be reasonable since a total dropout can be thought of as a double allelic dropout. Probably $cs$ should be set higher than $cd$, since allelic dropout is generally believed to be more common that bin shift. Care must be taken not to put the hypothetical off-site 'perfect match' tree too close to the buffer zone, or most unmatched seeds (and some matched seeds) will choose the off-site tree as the best match. In our algorithm, we did not allow a matched seed to switch to an off-site 'match', even if the off-site match score was higher. We had not considered the possibility that the algorithm would recommend switching a matched tree to another genotyped (but not perfectly matched) tree, but this did happen if we made our penalties too small. At first, we over-rode these suggestions, but after more careful analysis, we decided to keep some of these switches.

Initially, we set all the penalties very high ($cz$=$cd$=$cs$=10) to see if the algorithm recreated the original assignment of trees. It did so in that any seed which was not a perfect match on all 8 alleles was matched to an off-site tree. That is, the procedure yielded 552 matches and 174 non-matches. Some seeds which should have been matched to long distance sources were instead matched to fictitious 'nearer' off-site trees, but the over-ride provision kept this from being implemented. If $cd$=$cs$=10, but $cz$=0, we recreated Andy Jones' initial categorization of 616 matched and 110 unmatched seeds. As the $cz$ value was lowered in tandem with the $cd$ value, in a 2:1 ratio, with $cs$ kept high, we found that the number of originally non-matched seeds which remained unmatched was about 70-80. To lower this value to the expected 40-50 or so, we needed to reduce $cs$, but doing this led to more internal switches, which we wished to avoid. Table 6.11 displays results for some of the

parameter values we tried. In this table, the row values are *cs*, the column values are *cd*, and *cz*=0 in all cases. There are two values reported in each cell. The number above the diagonal is the number (out of 174) of non-matched seeds which stay non-matched under this parameter configuration. The number beneath the diagonal is the number (out of 552) originally matched seeds which are now paired to a different (not perfectly matched) tree after implementaton of the algorithm. The four corner values illustrate the point. If both *cd* and *cs* are set high, as shown in the lower right-hand corner of Table 6.11, most (82) of the 174 unmatched seeds remain unmatched and very few (9) of the 552 matched seeds change their tree affiliation. (Some of those that do change are revealing – they are apparent perfect matches to trees that are hundreds of meters away, but could, with a simple allelic dropout, be perfect matches to a very nearby tree which has several other perfect match seeds in the same trap as the seed in question. There is little doubt in our mind that this non-perfect assignment is the correct one.) If *cs* and *cd* are both set low, as in the upper left of the table, we have the undesirable situation where every unmatched seed has found a match, but a great many of the 552 perfectly matched seeds have also changed their affiliations. This means that we are almost totally ignoring genetic information and matching seeds to the nearest large tree. Somewhat surprisingly, this seems to achieve the correct answer for 446 of the 552 perfectly matched seeds, since only 106 are switched. Although it is easy to find regions of the table which are not good, it is hard to say exactly what is right. The MLE point-estimate from the Hypergeometric-Poisson method predicted that only 20 of the non-matched seeds are correct, so this corresponds to a point near (*cs*=0.225, *cd*=0.225), but this seems to have far too much switching of matched seeds. If we try for the point estimate from the Density-Matching approach (which is in the feasible region for the Hypergeometric-Poisson approach), we want about 52 of the non-matched seeds to remain unmatched. This appears to occur near the point (*cs*=0.50, *cd*=0.40), although this does lead to 12 switches among the 552 original perfectly matched seeds. Our examination of these 12 switches leads us to believe that over half are surely correct, whereas others are questionable, depending on

exactly how strongly one feels about the relative probability of a long distance event versus the probability of a bin shift.

| cs\cd | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.40 | 0.50 | 1.00 | 1.50 |
|---|---|---|---|---|---|---|---|---|---|
| 0.05 | 0/106 | | 0/106 | 0/106 | | | 0/104 | 1/97 | 1/94 |
| 0.10 | | | | | | | | | |
| 0.15 | | | | | | | | | |
| 0.20 | | | 15/31 | 15/31 | | | | | |
| 0.25 | | | 23/20 | 23/20 | | | | | |
| 0.30 | | | | | | | | | |
| 0.35 | | | | | | | | | |
| 0.40 | | 42/15 | 42/15 | 42/15 | | | 43/14 | 49/14 | 52/14 |
| 0.45 | | 46/12 | 46/12 | 47/12 | 48/12 | 49/12 | 49/12 | | |
| 0.50 | | 50/12 | 51/12 | 51/12 | 51/12 | **52/12** | 52/12 | | |
| 0.55 | | | | | 56/12 | 57/12 | 58/12 | | |
| 0.60 | | | | | | | | | |
| 0.65 | | | | | | | | | |
| 0.70 | | | | | | | | | |
| 0.75 | | | | | | | | | |
| 0.80 | | | | | | | | | |
| 0.85 | | | | | | | | | |
| 0.90 | | | | | | | | | |
| 0.95 | | | | | | | | | |
| 1.00 | 74/10 | | | 76/10 | | | 77/10 | 80/10 | 82/9 |

Table 6.11: Non-Match Holdouts/Match Switches for Selected *cs* and *cd*

Once the general form of the score function was found, we tweaked it in various ways, using slight refinements of the distance ($q$) and fecundity ($NAFX$) functions displayed in equation (6.5) to make the fit better. For example, as noted in Section 5.3, both the $qn$ and $ql$ log-intensity functions are more steep than the empirical $el$ log-intensity function beyond $100m$. To make sure that use of $qn$ did not bias the results away from LDD results, we estimated a log-intensity function which behaved the same as $qn$ for distances $<100m$, but more like $en$ for distances $>100m$. In the end, using this function and ($cz$=0, $cd$=0.40, $cs$=0.50) in the score function, our best correction for the data is that 52 of the 110 non-matched seeds belong to off-site trees, while 58 are matched to on-site trees, acknowledging genotyping errors. In addition, 23 of the 616 matched seeds are switched to non-perfectly-matching, but

significantly closer trees, with this happening 12 times for the 552 original perfectly matched trees and 11 times for the 64 trees which were 6-matches. This is summarized in Table 6.12.

| | Original Match | New Match | Off-site | Total |
|---|---|---|---|---|
| Perfect Match at 4 Loci | 540 | 12 | 0 | 552 |
| Perfect Match at 3 Loci + '00' | 53 | 11 | 0 | 64 |
| Unmatched | 0 | 58 | 52 | 110 |
| Total | 593 | 81 | 52 | 726 |

Table 6.12: Comparison of Original and Final Genotyped Seed Resolutions

The entire procedure of using the score function of equation 6.5 to produce the results in Table 6.12 seems reasonable, but is not statistically rigorous. It would be preferable to specify prior distributions for the three types of genotypic error – total dropout, allelic dropout, and bin shift, and then to use Bayesian methods to allow the data to select the optimal values of the paramters $cz$, $cd$, and $cs$. Unfortunately, the previous research on genotyping errors is not particularly precise concerning the rates of these errors, since they depend so heavily upon the amount of genetic material available for genotying and upon the skill of the technician. For the 81 seeds in Table 6.12 for which the score function of equation 6.5 found switches in tree assignment to be appropriate, there were 19 total dropouts (+59 for other seeds), 74 allelic dropouts, and 82 bin shifts. Over the entire data set of 726 genotyped seeds, this corresponds to a locus dropout rate of .0268, an allelic dropout rate of .0127, and an allelic bin shift rate of .0141. All of these values are within the realm of possibility for genotypic errors of these types, but are higher than one expects when genotyping is performed by experts using modern (tetra-nucleotide) tandem repeat procedures. The total dropout and allelic dropouts found in the 81 seeds appear to be justified, but there is some doubt about the relatively high rate of bin shift errors. Of the 82 claimed bin shift errors, 32 require a shift of two units, 18 require a shift of four units, and the other 32 require larger shifts. Most experts [1] believe that 4-shift errors are less common than 2-shift errors, and that 6-shift and higher errors almost never occur under ordinary circumstances.

Table 6.13 compares the dispersal distribution of the original 616 matched seeds with the revised set of 674 'matched' seeds. The first 6 columns are identical to those of Table 6.9, except for 'Match674', which is the newly estimated distribution. Of course, it is very similar to 'Match616', since there has been no change at all to 593 of the 674 seeds. However, one will note that there are fewer very extreme distances observed, and that three of the longest-dispersing seeds from the original 616 matched seeds (at distances $453m$, $513m$, and $710m$) have now been shifted to shorter distances. The 'L674f' column gives the estimates of the log-intensity based on the *ql* model applied to the 674 seed match, but with higher intensity used at distances greater than $100m$ to account for the previously noted deficiency of the *ql* model at long distances. The penultimate two columns in the table, 'E616f' and 'E674f', represent the number of off-site seeds which would be expected to land in the 298-trap network if the fitted 616 seed or 674 seed log-intensities, respectively, were correct, assuming the off-site adult *Jacaranda* trees had the same spatial density and fecundity distribution as the 292 on-site adults. From the 'Match674' column, one observes that the estimate of the probability of LDD dispersal from the revised data set is now 67/674=.10, which is slightly less than the original 11%-13% estimate, and much less than the 23% estimated by the censored data approach of Section 5.2. The 'Match674' distribution is more tightly packed in the 100-300$m$ region than the 'Match616' distribution, with dispersal over $400m$ being very rare. If one desires a parametric function which best estimates the log-intensity of the revised data over the entire range of the data, we recommend the 'L674f' funtion shown as the last column of Table 6.13.

| DC | Distance | MDPT | Tree_Trap | Match616 | Match674 | Fictreetrap | E616o | E616f | E674f | L674f |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0-5m | 2.5 | 7 | 8 | 9 | 0 | 0.00 | 0.00 | 0.00 | 1.59 |
| 2 | 5-10m | 7.5 | 33 | 67 | 77 | 0 | 0.00 | 0.00 | 0.00 | 0.74 |
| 3 | 10-15m | 12.5 | 48 | 75 | 82 | 0 | 0.00 | 0.00 | 0.00 | 0.14 |
| 4 | 15-20m | 17.5 | 93 | 98 | 108 | 0 | 0.00 | 0.00 | 0.00 | -0.34 |
| 5 | 20-25m | 22.5 | 88 | 76 | 81 | 0 | 0.00 | 0.00 | 0.00 | -0.75 |
| 6 | 25-30m | 27.5 | 115 | 33 | 36 | 0 | 0.00 | 0.00 | 0.00 | -1.12 |
| 7 | 30-35m | 32.5 | 129 | 55 | 59 | 0 | 0.00 | 0.00 | 0.00 | -1.46 |
| 8 | 35-40m | 37.5 | 153 | 21 | 27 | 0 | 0.00 | 0.00 | 0.00 | -1.78 |
| 9 | 40-45m | 42.5 | 177 | 47 | 50 | 0 | 0.00 | 0.00 | 0.00 | -2.07 |
| 10 | 45-50m | 47.5 | 187 | 12 | 15 | 0 | 0.00 | 0.00 | 0.00 | -2.35 |
| 11 | 50-55m | 52.5 | 193 | 14 | 16 | 0 | 0.00 | 0.00 | 0.00 | -2.61 |
| 12 | 55-60m | 57.5 | 233 | 10 | 8 | 0 | 0.00 | 0.00 | 0.00 | -2.86 |
| 13 | 60-65m | 62.5 | 235 | 13 | 14 | 0 | 0.00 | 0.00 | 0.00 | -3.10 |
| 14 | 65-70m | 67.5 | 239 | 4 | 5 | 0 | 0.00 | 0.00 | 0.00 | -3.33 |
| 15 | 70-75m | 72.5 | 261 | 0 | 1 | 0 | 0.00 | 0.00 | 0.00 | -3.55 |
| 16 | 75-80m | 77.5 | 285 | 5 | 5 | 0 | 0.00 | 0.00 | 0.00 | -3.76 |
| 17 | 80-85m | 82.5 | 282 | 5 | 5 | 0 | 0.00 | 0.00 | 0.00 | -3.97 |
| 18 | 85-90m | 87.5 | 328 | 4 | 4 | 0 | 0.00 | 0.00 | 0.00 | -4.17 |
| 19 | 90-95m | 92.5 | 331 | 3 | 3 | 0 | 0.00 | 0.00 | 0.00 | -4.37 |
| 20 | 95-100m | 97.5 | 329 | 3 | 3 | 0 | 0.00 | 0.00 | 0.00 | -4.56 |
| 21 | 100-150m | 125 | 4062 | 18 | 19 | 41 | 0.18 | 0.14 | 0.16 | -5.50 |
| 22 | 150-200m | 175 | 5173 | 11 | 15 | 360 | 0.77 | 0.69 | 0.74 | -6.18 |
| 23 | 200-300m | 250 | 12740 | 13 | 15 | 3188 | 3.25 | 3.40 | 3.37 | -6.85 |
| 24 | 300-400m | 350 | 13579 | 10 | 9 | 8342 | 6.14 | 4.96 | 4.55 | -7.51 |
| 25 | 400-500m | 450 | 12381 | 4 | 3 | 15412 | 4.98 | 5.70 | 4.91 | -8.05 |
| 26 | 500-600m | 550 | 10476 | 4 | 3 | 24600 | 9.39 | 6.04 | 4.92 | -8.52 |
| 27 | 600-700m | 650 | 8529 | 1 | 1 | 33807 | 3.96 | 5.75 | 4.46 | -8.93 |
| 28 | 700-800m | 750 | 6810 | 2 | 1 | 42376 | 12.45 | 5.16 | 3.83 | -9.31 |
| 29 | 800-900m | 850 | 5155 | 0 | 0 | 50473 | 0.00 | 4.51 | 3.21 | -9.66 |
| 30 | 900-1000m | 950 | 3211 | 0 | 0 | 58445 | 0.00 | 3.91 | 2.67 | -9.99 |
| 31 | 1000-1100m | 1050 | 1037 | 0 | 0 | 65663 | 0.00 | 3.34 | 2.20 | -10.30 |
| 32 | 1100-1200m | 1200 | 117 | 0 | 0 | 67285 | 0.00 | 2.33 | 1.46 | -10.74 |
| 33 | 1200-1300m | 1250 | 0 | 0 | 0 | 64595 | 0.00 | 1.98 | 1.22 | -10.88 |
| 34 | 1300-1400m | 1350 | 0 | 0 | 0 | 58880 | 0.00 | 1.43 | 0.85 | -11.14 |
| 35 | 1400-1500m | 1450 | 0 | 0 | 0 | 50345 | 0.00 | 0.97 | 0.56 | -11.40 |
| 36 | 1500-1600m | 1550 | 0 | 0 | 0 | 40991 | 0.00 | 0.64 | 0.36 | -11.65 |
| 37 | >1600m | 1650 | 0 | 0 | 0 | 97982 | 0.00 | 1.23 | 0.67 | -11.89 |
| Total | | | 87016 | 616 | 674 | 682782 | 41.12 | 52.16 | 40.15 | |

Table 6.13: Revised Estimates by Distance Class

# Chapter 7

# Conclusion

In Chapter 4, we demonstrated inverse modeling techniques for non-genotyped data. By examining combined bi-year data, separate bi-year data and pooled bi-year 'X' data, we have learned that different time periods should be modeled separately. And for *Jacaranda* trees, Clark's model with $c = 0.5$, with fecundity modeled as being proportional to $\text{DBH}^2$ is reasonable. But more general models which make fecundity proportional to $\text{DBH}^g$ will typically estimate $g$ to be smaller than 2, usually $1 < g < 2$. Even if one over-parameterizes by fitting certain influential trees separately, the dispersion of the overall model will be at least nine times as great as that expected under Poisson conditions when evaluation is at the trap level.

In Chapter 5, from the genotyped data due to actual seed dispersal distances being observed, the distance function estimation improved substantially in the upper tail. Both the genotyped and ungenotyped data sets give some general support to the belief that log-intensity function decays approximately as square-root of distance rate for $d_{ij} < 100m$. However, the genotyped data gives definite evidence of some ($\sim 11\% - 13\%$) long-distance dispersals, although certainly not nearly as strong as the censored data approach estimates of Jones, Chen, Weng and Hubbell [14] described in Section 5.2. Both the ungenotyped and genotyped data sets lend some support to the idea that fecundity is proportional to $\text{DBH}^g$, with estimates of $g$ varying widely, but generally being less than the $g$=2 value often assumed in the inverse modeling literature. If we over-parameterize to obtain approximately the best possible fit for distance and to model individual trees' fecundities, we can force $G_e^2$ and $G_I^2$

to become arbitrarily small, but neither the overall deviance, $G^2$, nor $G^2_J$, which is the best overall fit to the trap data, will fit adequately by any standard statistical convention.

In Chapter 6, we discussed the possibility of genotyping errors occurring and produced some evidence that both 'total dropout' and 'allelic dropout' had occurred during the seed genotyping. We then used two methods to estimate the global proportion of misclassification, and both agreed that at least half of the 110 non-matched seeds might indeed be seeds which should have been matched to an on-site source. Eventually, we decided to evaluate every seed(k) in a trap(j) by calculating a score that measures its probabilistic distance to each source tree(i) (including a fictitious perfectly matched off-site source), with the seed then being matched to the tree which gave the best score. After doing this, our estimate of the probability of LDD dispersal from the revised data set is 10%, slightly less than the original 11%-13% estimate, but much greater than the estimates from non-genotyped data (5%) and much smaller than the LDD estimates (23%) found by Jones et. al. [14].

Overall, we find that the use of direct estimation on genotyped seed data will allow for better estimation of the dispersal function than will indirect estimation on non-genotyped seed data. This is especially true in the upper tail of the distribution. However, we do offer the following caveats:

(a) There is much fecundity variability not measurable by DBH.

(b) Large numbers of seeds must still be collected for good dispersal estimates to be obtained, even with genotyped data.

(c) It is *very* important to reduce genotyping errors. We recommend using tetra-NTR's rather than di-NTR's in the future, if possible.

# Bibliography


[1] Bonin, A., E. Bellemain, P.B. Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet (2004). How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, 13: 3261-3273.

[2] Cain, M.L., B.G. Milligan, and A.E. Strand (2000). Long-distance seed dispersal in plant populations. *American Journal of Botany*, 87: 1217-1227.

[3] Clark, J.S. (1998). Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. *The American Naturalist*, 152: 204-224.

[4] Clark, J.S., M. Silman, R. Kern, E. Mackin, and J.H. RisLambers (1999). Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology*, 80: 1475-1494.

[5] Clark, J.S., E. Mackin, and L. Wood (1998). Stages and spatial scales of recruitment limitation in southern Appalachian forests. *Ecological Monographs*, 68: 213-235.

[6] Clark, J.S., M. Lewis, and L. Horvath (2001). Invasion by extremes: population spread with variation in dispersal and reproduction. *The American Naturalist*, 157: 537-554.

[7] Constable, J. L., M. V. Ashley, J. Goodall and A. E. Pusey (2001). Noninvasive paternity assignment in Gombe chimpanzees. *Molecular Ecology*, 10: 1279-1300.

[8] Creel S., G. Spong, J.L. Sands, J. Rotella, J. Zeigle, L. Joe, K.M. Murphy and D. Smith (2003). Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, 12:2003C2009.

[9] Croat, T. (1978). The flora of Barro Colorado Island. Stanford University Press. Stanford, California.

[10] Godoy, J.A., and P. Jordano (2001). Seed dispersal by animals: exact identification of source trees with endocarp DNA microsatellites. *Molecular Ecology*, 10: 2275-2283.

[11] Higgins, S.I., and D.M. Richardson (1999). Predicting plant migration rates in a changing world: the role of long-distance dispersal. *The American Naturalist*, 153: 464-475.

[12] Jeffery, K.J., L.F. Keller, P. Arcese, and M.W. Bruford (2000). The development of microsatellite loci in the song sparrow, Melospiza melodia and genotyping errors associated with good quality DNA. *Molecular Ecology Notes*, 1: 11-13.

[13] Jones, F.A., and S.P. Hubbell (2003). Isolation and characterization of microsatellite loci in the tropical tree Jacaranda copaia (Bignoniaceae). *Molecular Ecology Notes*, 3: 403-405.

[14] Jones, F.A., J. Chen, G-J Weng, and S.P. Hubbell (2005). A genetic evaluation of seed dispersal in the Neotropical tree, Jacranda copaia (Bignoniaceae). *The American Naturalist*, 166: 543-555.

[15] 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.

[16] Nathan, R., and H.C. Muller-Landau (2000). Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology and Evolution*, 15, 278-285.

[17] Ouborg N.J., Y. Piquot, J.M. Van Groenendael (1999). Population genetics, molecular markers and the study of dispersal in plants. *Journal of Ecology*, 87: 551-568.

[18] Portnoy, S., and M. F. Willson (1993). Seed dispersal curves: behavior of the tail of the distribution. *Evolutionary Ecology*, 7: 25C44.

[19] Ribbens, E., J.A. Silander, and S.W. Pacala (1994). Seedling recruitment in forests: calibrating models to predict patterns of tree seedling dispersion. *Ecology*, 75: 1794-1806.

[20] Stoyan, D., and S. Wagner (2001). Estimating the fruit dispersion of anemochorous forset trees. *Ecological Modelling*, 145, 35-47.

[21] Turchin, P. (1998). Quantitative analysis of movement: measuring and modeling population redistribution in animals and plants. Sinauer, Sunderland, Massachusetts.

[22] Wang, B.C., and T.B. Smith (2002). Closing the seed dispersal loop. *Trends in Ecology and Evolution*, 17: 379-385.

[23] Willson, M.F. (1993). Dispersal mode, seed shadows, and colonization patterns. *Vegetatio*, 107/108: 261-280.

[24] Wright S.J., H.C. Muller-Landau, R. Condit, and S.P. Hubbell (2003). Gap dependent recruitment, realized vital rates, and size distributions of tropical trees. *Ecology*, 84: 3174-3185.