STATISTICAL SAMPLING AND MODELLING FOR CORK OAK AND EUCALYPTUS STANDS

Promotoren:	Prof. Dr. Dr. h.c. D.A.M.K. Rasch Hoogleraar in de Wiskundige Statistiek
	Prof. Dr. Ir. A. Stein Hoogleraar in de Wiskundige en Statistische Methoden, met speciale aandacht voor de ruimtelijke statistiek
Copromotor:	Drs. A. Otten Wetenschappelijk hoofdmedewerker Wiskundige en Statistische Methoden
Promotiecommissie:	Prof. Dr. Ir. J. Grasman (Wageningen Universiteit) Dr. M.N.M. van Lieshout (CWI, Amsterdam) Prof. Dr. Ir. R. Rabbinge (Wageningen Universiteit) Prof. Dr. Ir. M. Tomé (Universidade Técnica de Lisboa)

Maria João Paulo

STATISTICAL SAMPLING AND MODELLING FOR CORK OAK AND EUCALYPTUS STANDS

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
Prof. Dr. Ir. L. Speelman,
in het openbaar te verdedigen
op maandag 4 november 2002
des namiddags te half twee in de Aula.

Paulo, M.J., 2002

Statistical sampling and modelling for cork oak and eucalyptus stands.
PhD Thesis Wageningen University, Wageningen, the Netherlands.
Cover: Carlo Vromans

ISBN 90-5808-729-8

To my parents, who showed me the sea

To Carlo, who showed me the lighthouse

Abstract

This thesis focuses on the use of modern statistical methods to solve problems on sampling, optimal cutting time and agricultural modelling in Portuguese cork oak and eucalyptus stands. The results are contained in five chapters that have been submitted for publication as scientific manuscripts.

The thesis first addresses the decision of *when* to cut a rotation of eucalyptus production forest. The aim is to optimise the long term volume production, corrected for replant costs. On the long term the total financial yield divided by the total rotation time is an important economical asset. Successive rotations and their growth curves are considered as independent realisations of the same generating process. A Bayesian approach was taken, using Shumacher curves. Prior information on the curve parameters was based on a large number of observed growth curves. For known or accurately estimated curves, a 16 % gain in optimisation of cutting times could be achieved, as compared to using a common optimal cutting time. It is assumed that a farmer takes two volume measurements to decide upon the cutting time of a rotation, the first measurement at a fixed age, the second at an age that possibly depends upon the first measurement. Finding the optimal second measurement time is entangled with finding the optimal cutting time. In this thesis, simultaneous optimisation is carried out using numerical methods. The gain in using a variable optimised second measurement time, compared with an optimised fixed measurement time, however, was relatively small (up to 0.1 %), which is hardly above the numerical noise level.

A second problem addressed in this thesis concerns estimation of stem diameter growth curves in cork oaks. A data-set of 24 trees was used. A D-optimal experimental design has been compared with equidistant designs to measure trees at particular ages to allow for an optimal estimation of individual growth curves. An experimental design that is locally D-optimal for a central parameter is proposed. This fixed compromise design can be used for all trees. For individual growth curves and under certain conditions that are discussed in the thesis such a design provides better estimates than an equidistant design.

The third study concerns spatial modelling of quantitative cork oak characteristics. Spatial statistical methods are used to analyse cork oak stands, so-called montados. Spatial correlations between neighbouring trees of crown shapes, of crown sizes and of stem sizes are analysed using plots from two montados. A significant correlation is found between tree size and competition from neighbouring trees. In particular, larger trees have

a regular spatial distribution in a montado.

The final study in this thesis compares three sampling methods for use in cork oak farms. One method is currently in use by Portuguese farmer's associations to estimate cork value prior to stripping and the other two methods are compared to it. The three sampling methods are applied to two cork oak farms and to simulated stands. The latter are generated with spatial simulation methods on the basis of information obtained elsewhere. The current method has a 15-50 % larger bias. For a clustered pattern standard errors are lowest for the current method, but these are considerably higher for a regular or a random pattern.

In conclusion, this thesis shows that modern statistical methods are valuable to improve modelling and sampling of cork oak and eucalyptus forests. In particular, spatial relations among neighbouring trees should preferably be included into management of cork oak farms. Adequate sampling methods are basic to retrieve information of the highest quality.

Contents

Chapter 1

Introduction

Systems approaches are playing a prominent role in current agriculture and forest science. In forestry, stands of trees may serve as silvopastoral systems. Such systems are subject to environmental and weather conditions, as well as to management activities. Management has to decide when, where and how actions have to be taken. Many decisions have a quantitative background, and require a quantitative answer. They are preferably taken on the basis of measurements on trees in the current system, in other systems under similar conditions and on the same system in the past. They most likely could benefit therefore from well-interpreted statistical analyses.

This thesis is based on precisely this approach. As a demand-driven research it investigates the role that current mathematical statistical methods can play to answer relevant quantitative management questions. The thesis is focused on four typical research questions:

- What is the optimal cutting time of a rotation?
- What is an optimal experimental design to estimate stem diameter growth?
- How could one model spatial competition effects?
- What is the optimal sampling strategy to estimate tree characteristics?

Current methods from mathematical statistics are applied to answer these questions as good as possible from nowadays' perspective.

Several procedures have recently been developed in mathematical statistics. Since the early nineteen nineties, Bayesian methods, although dating back to the 18th century, have found a place between other common estimation and modelling procedures. This is mainly due to the increased flexibility and power of modern computer systems in handling the

increased amount of data and information required for such analysis. Bayesian procedures allow to make a better estimate of parameters on the basis of prior information. Using actual data, the prior is updated to give posterior estimates that may have a lower variance than the prior parameters. Another recent development has taken place in D-optimal designs, where issues of robustness are of an increasing concern. In spatial statistics and image analysis increasing attention has recently been given to statistics of shapes. Shapes are characterized by a low number (e.g. 4 – 8) of points to which possibly interpolating splines or polygons are fitted. So far, applications in agricultural and production forestry are lacking, however, although the benefits of these methods can be large.

This thesis develops these methods on Portuguese agricultural and forestry systems. As a production forest system it considers forest stands consisting of eucalyptus trees and of cork oaks. The eucalyptus was introduced in Portugal as an ornamental tree in 1829 and became economically important after the development of the paper industry in 1907. The most abundant eucalyptus species in Portugal is *Eucalyptus globulus* Labill., used mainly for pulpwood production and sawtimber. The eucalyptus is highly suited for pulpwood production due to its fast growth and excellent fiber qualities, yielding whiter fiber than any other tree species. Eucalyptus production forests are managed over very short rotations (10-15 years). Average yearly volume production is usually between 15-20 $m^3 ha^{-1}$.

Cork oak (*Quercus suber* L.) is grown in Portugal mainly in montados. A *montado* is a agroforestry system, where the farmer keeps one or more tree species in a low density and grows cows or sheep in the same area. Typically, montados occur in low populated areas where the soil is too poor for agriculture, particularly in the south-east part of the country. They cover large areas, often of a few hundreds of hectares in size. Cork oaks have a life span of 300–400 years, but are economically viable for less than 150 years. Cork is a thick and continuous layer of suberised cells, produced by the meristematic cork cambium (or *phellogen*). It makes up the external envelope of the stem and branches of the tree. The value of cork for industrial purposes highly depends on cork thickness. The highest value is associated with thicknesses between 29 and 40 mm. The growth of the trees and of the cork, as well as its quality, are determined both by genetic tree characteristics, site quality and management practices. Management of cork oak stands includes thinning, shape pruning, understorey clearing and soil fertility improvement. Production of cork is an important economic activity, as the world demand for cork keeps increasing and Portugal is its leading exporter.

Purpose

The purpose of this research is to answer some important quantitative management questions with current mathematical statistical methods. As such, it is a demand-driven approach and exemplary for the analysis and support of decision making at an agricultural and a forestry system. Attention is focused on current systems in Portugal, where these questions were raised. Some typical problems in cork oak farming and eucalyptus forestry are addressed. The use of Bayesian statistics is investigated. Attention is given to optimal experimental design, spatial statistics and modern simulation techniques. These allow a better insight into sampling and management of these systems. The statistical basis will be partly used in the initialization module of the SUBER model, a decision support system aimed to help cork oak farmers (see section 5.5).

Outline

The outline of the thesis is as follows.

Chapters 2 and 3 address the optimal cutting time of eucalyptus production forests. Eucalyptus is harvested for pulpwood when it attains approximately the biological rotation age, i.e. the age for which the average yearly production in one rotation is maximal. The economical aim is to maximise the long term yearly volume production, corrected for replant costs, as measured over several rotations by allowing flexible cutting time for each rotation by plantation age. The assumption is made replant costs are fixed in each rotation. The problem is divided into two parts. Chapter 2 determines the optimal cutting time. For maximisation of the long term volume production different growth curve are assumed to apply to each rotation. It is shown that the optimal cutting time depends both on the actual observed growth curve, and on all potential growth curves than may occur in future rotations. A prior distribution for the growth curve parameters is used. The general prior applied in this study covers volume growth curves observed in several stands in different parts of Portugal. Approximate optimal cutting time is obtained for the practical situation that a farmer measures trees at two different fixed ages at an early stage of their development and derives two total volume estimates for the forest. The actual growth curve can be estimated from the two volume estimates. Chapter 3 explores the last situation further. It is assumed that a farmer makes the first measurement at a fixed age. With that measurement and prior knowledge of volume growth, the time for the second measurement is chosen so that in conjunction with the cutting time choice, the long term production is exactly optimised. Both approximate and exact optimisation were reached by means of extensive use of numerical methods.

Chapters 4, 5 and 6 cover three separate problems in cork oak montados. Chapter 4 uses an optimal experimental design for estimation of stem diameter growth. Trees are commonly measured to estimate tree growth by means of a series of regular time points, a so-called equidistant designs. A locally D-optimal experimental design as an alternative consists of measuring trees at those moments in time, that the determinant of the asymptotic variance matrix of the (growth) parameters is minimized. This is equivalent to maximising the determinant of the information matrix of the parameters. This procedure yields measurements that are performed on specific ages of the trees. It allows a more precise estimation, but is sensitive to the parameter values. Advantages of using a hybrid experimental design, D-optimal for a central parameter, are analysed with respect to the estimation of individual tree growth on cork oaks. Its robustness is studied under parameter mis-specification.

Chapter 5 considers relations between neighbouring trees within single cork oak stands. It is analyzed to which degree competition between trees influences availability of nutrients and light and affects the shape and the size of tree crowns, which, in turn, are related to tree health and growth. Competition is measured by current competition indices. Their effects are studied by means of their correlation with tree diameter and tree height, crown size and crown shape. Competition at the crown level is assumed to depend on the distances to neighbouring trees and their sizes.

Chapter 6 compares three sampling procedures for estimation of density, basal area and cork volume in cork oak montados. Current management requires estimation of the value of cork, just before cork extraction. This value depends upon quantity and quality of cork. A yield estimate assists a farmer to set a price for his cork. A commonly used sampling procedure consists of a polygonal transept or *zigzag*, with a convenient starting point and covering the whole montado. Every tree located on the transept is sampled. The two other methods are cluster sampling with fixed plot radius and cluster sampling with a fixed number of trees and variable plot radius. Comparison is done on simulated montados with different point-and-diameter patterns. Bias and precision of the estimators, and sampling costs are considered.

Chapter 2

Optimal Bayesian design applied to volume yield and optimal cutting time prediction

Maria João Paulo and Albert Otten

In Ermakov, S.M.; Kashtanov, Yu.N. and Melas, V.B. editors. *Proceedings of the 4th St. Petersburg Workshop on Simulation*. 2001. St. Petersburg University. 370-378.

In this study we consider a typical problem where a forester wants to determine the optimal design in order to estimate the best cutting time for a eucalyptus stand. We combine a Bayesian prior with observational information to obtain estimates for the growth curve parameters and to determine the best cutting time, our optimisation criterion being long term net volume yield maximisation. Our preliminary results seem to indicate that the observation variance and choice of prior have a greater influence on yield than the choice of design.

2.1 Introduction

In a eucalyptus production forest the farmer is interested in cutting the trees at the age which maximises the long term volume production. For a particular (repeatedly used) growth curve, this age is called the biological rotation age and is determined by the line through the origin tangent to the volume growth curve. We take the Shumacher growth curve $V = A\mathrm{e}^{-k/t}$ to explain Volume as a function of time t. Under this model k is the biological rotation age, and the corresponding long term yearly yield is $(A/\mathrm{e})/k$. Traditionally, the farmer wants to be able to predict early in time the biological rotation age and the corresponding yield. However, the approach of cutting at the specific curve's rotation age is not optimal when a new curve (parameters) is 'drawn' from a known prior distribution after each rotation, and in this case finding the optimal cutting time is an optimisation problem in itself. In order to make the cutting time optimisation problem meaningful, we introduce costs for each harvest, corresponding to a constant volume loss V_0. First we describe the optimal cutting time strategy, and its numerical implications, when the actual curve is known to the farmer. Next we describe the optimal strategy when the actual curve is unknown but observations at 2 time points are made, and present a numerically feasible sub-optimal strategy for this situation. Using the latter strategy as definition of cutting time, we finally are able to address the design problem of optimising t_1 and t_2.

2.2 Materials and methods

One growth curve *versus* independent curves from same prior

In the forestry literature a growth period (i.e. time from when a stand of trees is re-generated until the time when it is harvested) is called *rotation*. Assuming we have one volume growth curve $V(t) = f(t)$ in all rotations, then it's well known that the cutting time that maximises the long term volume production is given by the line through the origin tangent to the volume growth curve, i.e. the point where $V(t)/t$ is maximum. This cutting time is called the biological rotation age. A similar result holds with costs, with $V(t) - V_0$ replacing $V(t)$. If we have one unknown growth curve in all rotations, on which we make observations, then we get to know the biological rotation age better after each rotation. Here we consider the situation where instead of having one common curve in every rotation for a given stand, we assume that we can have any curve $V(t) = f(\theta, t)$ each time again, where θ has a known prior distribution.

Data

The data set used is from plots belonging to a Portuguese pulp mill, Stora Celbi, Celulose Beira Industrial, S.A. It contains the evolution in time of volume per ha for 158 plots of *Eucalyptus globulus*. The plots have different areas, different tree densities and have different quality indices. Each plot has been measured between 4 and 25 times, at different ages. The volume has been estimated from measurements of diameter and height.
We used OLS to fit 11 growth functions, among others the exponential function (with 2 and 3 parameters), the logistic function and the Gompertz function, to the growth of volume per ha in time. We selected the Shumacher function $f(\theta, t) = A\mathrm{e}^{-k/t}$, where A is the curve asymptote, and k is the biological rotation age. This function gave the best fit in most cases.

Prior information

We briefly studied $\hat{A}$ and $\hat{k}$, the OLS-estimated parameters for the fitted volume growth curves. We observed their histograms, probability plots and bivariate plots, and repeated this for transformations of $\hat{A}$ and $\hat{k}$. We found that $\ln(\hat{A})$ and $\ln(\hat{k})$ for the pooled data followed approximately a bivariate normal distribution, with

$$\mu = \begin{pmatrix} 6.28 \\ 2.56 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.512 & 0.171 \\ 0.171 & 0.134 \end{pmatrix} \tag{2.1}$$

The within plot component is not negligible, but we proceed by taking $\pi_0 \sim \mathcal{N}(\mu, \Sigma)$ as the prior density for $\ln(A)$ and $\ln(k)$. So we generate parameters in the ln scale and when necessary we will transform them back into the original scale. We note that the elements of Σ in (2.1) are the total variance due to prior variation in $\ln(\theta)$, and in $\widehat{\ln(\theta)}$. We further note that the prior density defined above is a very 'broad' one, observed over all 158 plots. These plots differ not only in location but also have different quality classes and number of trees per hectare, for example. In a single plot we expect to observe differences in volume growth due to some external factors such as weather and soil fertilization, as well as genetic differences. Therefore we expect that the prior density observed for one single plot or for a given quality class will have, besides different mean values for the two parameters, a smaller covariance matrix. In this research it suffices to consider properties of one rotation. We assume that the farmer knows the family of Shumacher growth curves for the volume, and prior π_0 for $(\ln(A), \ln(k))$. We distinguish between (i) θ *known* and (ii) θ *unknown* to the farmer. In case of θ *unknown*, he makes two observations at times t_1 and t_2 with some error with known variance σ^2. Then he estimates the parameters, using information from the observations and from the prior distribution. The cutting time is then chosen. The fact that we use prior information has consequences to the resulting

cutting time, because if we know the underlying family of growth curves, we may want to reject a 'bad' growth curve by cutting it earlier than the biological rotation age, or even immediately at time t_2 (For θ *known*, we consider t_2 to be the minimally allowed cutting time, for comparison reasons). Similarly, if our observed curve seems to be a very good one, we may want to cut it a bit later than the biological rotation age. This seems to improve the long term production.

Strategy for the choice of the cutting time

With 'strategy' we mean the rule or function which assigns a cutting time $\mathcal{C}$ to a known curve (θ) or data (y_1, y_2). Formally, for a given strategy, the long term production is defined as the ratio $\phi = E(A\mathrm{e}^{-\frac{k}{\mathcal{C}}} - V_0)/E(\mathcal{C})$, where the expectations are taken over the prior distribution, and also over the observations (when the curve is unknown). The cutting time optimisation problem is the problem of finding the strategy which maximises ϕ. Analytical considerations show how to choose the cutting time once we know $\phi_{\max}$. *If θ is known* (i.e. the curve parameters are known), then the chosen cutting time $\mathcal{C}^*$ is such that

$$A\mathrm{e}^{-\frac{k}{\mathcal{C}}} - \phi_{\max}\mathcal{C} \tag{2.2}$$

is maximal, $\mathcal{C}^* \geq t_2$. This leads to either $\mathcal{C}^*$ being the latest time when the tangent to the growth curve equals $\phi_{\max}$, or to $\mathcal{C}^* = t_2$. *If θ is unknown*, then given the observations y_1, y_2 (in the ln scale), we would have to replace $A\mathrm{e}^{-\frac{k}{\mathcal{C}}}$ with $E_{A,k}(A\mathrm{e}^{-\frac{k}{\mathcal{C}}}|y_1, y_2)$, and search for $\mathcal{C}^*$ such that

$$E_{A,k}(A\mathrm{e}^{-\frac{k}{\mathcal{C}}}|y_1, y_2) - \phi_{\max}\mathcal{C} \tag{2.3}$$

is maximal. Since actually $\phi_{\max}$ is unknown, we take a guess for $\phi_{\max}$, say ρ, and use it in (2.2) and (2.3) instead. Then every ρ will define a cutting time strategy, and the performance can be measured by $\phi(\rho)$. The cutting time optimisation problem is reduced now to finding the ρ for which $\phi(\rho)$ is maximal. By construction, this optimal ρ also satisfies $\rho = \phi(\rho)$ and $\rho = \phi_{\max}$. Maximising $E_{A,k}(A\mathrm{e}^{-\frac{k}{\mathcal{C}}}|y_1, y_2) - \rho\mathcal{C}$ over $\mathcal{C}$ is actually too computer intensive, so we approximate this expression with

$$\hat{A}\mathrm{e}^{-\frac{\hat{k}}{\mathcal{C}}} - \rho\mathcal{C} \tag{2.4}$$

where $\hat{A}$ and $\hat{k}$ now are the highest posterior density estimates of A and k. The ρ for which $\phi(\rho)$ is maximal will be called suboptimal since it is expected that the approximation will slightly degrade performance.

The algorithm

For a given ρ, the evaluation of $\phi(\rho)$ requires two-fold (θ *known*) or four-fold (θ *unknown*) integrals. We only show the most complicated situation (θ *unknown*):

$$\frac{E_{\theta,y}\left[V(\theta,\mathcal{C})-V_0\right]}{E_{\theta,y}\left[\mathcal{C}\right]}=\frac{\int_k\int_A\int_{y_1}\int_{y_2}(A\mathrm{e}^{-\frac{k}{\mathcal{C}}}-V_0)\,h(y_1,y_2,\theta)\,dy_2dy_1dAdk}{\int_k\int_A\int_{y_1}\int_{y_2}\mathcal{C}\,h(y_1,y_2,\theta)\,dy_2dy_1dAdk} \tag{2.5}$$

Here $y_i(\theta,t)=\ln(A)-k/t+\varepsilon_i$, $i=1,2$, $\varepsilon_i\sim\mathcal{N}(0,\sigma^2)$; $h(y_1,y_2,\theta)$ is the joint density of Y_1,Y_2 and θ, $h(y_1,y_2,\theta)=g(y_1,y_2|\theta)\pi_0(\theta)$. Expression (2.5) for $\phi(\rho)$ is equivalent to

$$\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(A\mathrm{e}^{-\frac{k}{\mathcal{C}}}-V_0)\,\mathrm{e}^{-\frac{1}{2}(z_1^2+z_2^2+z_3^2+z_4^2)}dz_4dz_3dz_2dz_1}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathcal{C}\,\mathrm{e}^{-\frac{1}{2}(z_1^2+z_2^2+z_3^2+z_4^2)}dz_4dz_3dz_2dz_1} \tag{2.6}$$

obtained when (A,k,y_1,y_2) are generated from independent z, $z\sim\mathcal{N}(0,I)$.
Here $\begin{pmatrix}\ln(k)\\ \ln(A)\end{pmatrix}=\mu+U'\begin{pmatrix}z_1\\ z_2\end{pmatrix}$ where U is the Cholesky root of Σ, i.e. $U'U=\Sigma$. Observations at time points t_1 and t_2 are $y_1=\ln(A)-k/t_1+z_3\cdot\sigma$ and $y_2=\ln(A)-k/t_2+z_4\cdot\sigma$ respectively. $\hat{\theta}$ was obtained from y_1,y_2 using a *highest posterior density* estimate. The *posterior density* is that of θ given y: $\pi(\theta|y)=\frac{\pi(\theta)g(y|\theta)}{g(y)}$. We found for each pair of observations the $\hat{\theta}$ that maximised $\pi(\theta|y_1,y_2)$. The cutting time, $\mathcal{C}$, depends on $\hat{\theta}$, on t_2 and on ρ. For practical reasons, $\mathcal{C}$ was truncated at 30 (years). The ratio $\phi(\rho)$ in (2.6) was programmed in Fortran using calls to IMSL subroutines. We also implemented the routine to optimise $\phi(\rho)$ over ρ. Note that we thus have the algorithm to produce a (sub)optimal strategy for a given prior, $\sigma^2=0.5$, t_1,t_2 and costs V_0 (we took $V_0=50m^3$). Therefore we can ask for the sensitivity of the corresponding (sub)optimal ϕ (written as ϕ from now on) and for the (sub)optimal design.

Sensitivity of the (sub)optimal long term production ϕ

We calculated ϕ in the case that θ *is known* for the following situations:

1. for some values of the prior $\sigma_{\ln(A)}$ and $\sigma_{\ln(k)}$, maintaining the mean values of A and k constant. We note that if $\ln(X)\sim\mathcal{N}(\mu,\sigma^2)$ then $E(X)=\mathrm{e}^{\mu+\sigma^2/2}$. We modified the original standard deviation $\sigma_{\ln(A)}$ and $\sigma_{\ln(k)}$ in the prior distribution by a multiplier $\frac{1}{4}$, $\frac{1}{2}$, and 2 (and also 0 for $\sigma_{\ln(A)}$ and 4 for $\sigma_{\ln(k)}$). See Figure 2.1.

2. for some values of the prior $corr(\ln(A),\ln(k))$, maintaining $\mu_{\ln(A)}$, $\mu_{\ln(k)}$, $\sigma^2_{\ln(A)}$ and $\sigma^2_{\ln(k)}$ constant (see Figure 2.2).

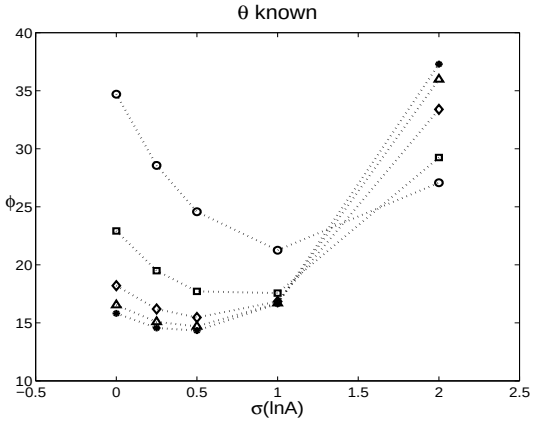

Figure 2.1: Dependence of ϕ on the standard deviation of $\ln(A)$ in the prior distribution ($E(\ln(A))$ fixed). Axis shows multiplier of $\sigma_{\ln(A)}$. Each curve corresponds to a different $\sigma_{\ln(k)}$ multiplier: $\frac{1}{4}$ ($*$); $\frac{1}{2}$ ($\triangle$); 1 ($\diamond$); 2 ($\square$) and 4 ($\circ$).

3. for some values of the time of the second measurement t_2. In the case where θ is known, no observations are generated, so t_2 is considered to be the earliest possible cutting time (see Figure 2.2).

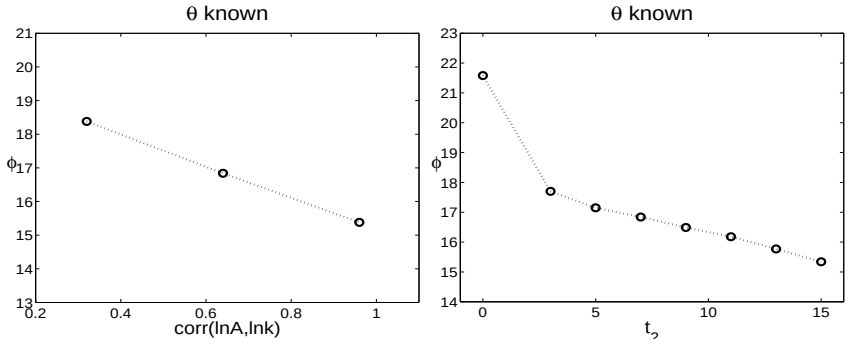

Figure 2.2: Dependence of ϕ on the prior correlation (left), and on the earliest possible cutting time in years(right).

As a first step towards the search for an optimal design, we also calculated ϕ in the case that θ *is unknown* for the following situations:

4. for some values of t_1 and t_2, maintaining the prior distribution and σ (see Figure 2.3).
5. for some values of the observation error standard deviation $\sigma = 0, 0.25, 0.5, 0.75, 1$

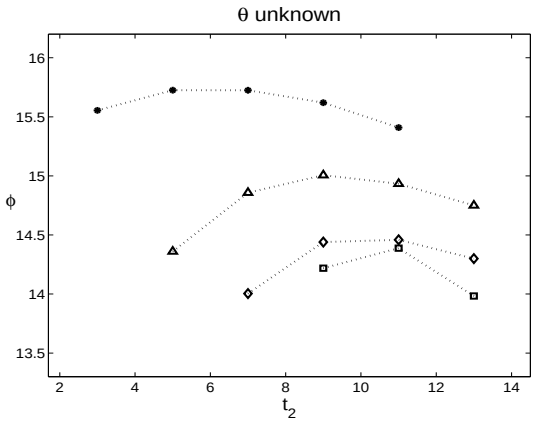

Figure 2.3: Dependence of ϕ on design choice. Each curve corresponds to a different t_1: 1 ($*$); 3 ($\triangle$); 5 ($\diamond$) and 7 ($\square$).

and ∞, maintaining the prior distribution constant and $t_1 = 3, t_2 = 7$ (see Figure 2.4).

2.3 Preliminary results

Case of known θ

In figures 1 to 3 we can see how changing some of the parameters can affect the resulting ratio ϕ. We know that ϕ will increase for higher $E(A)$ values and lower $E(k)$ values. The influence of the prior standard deviations on ϕ is less intuitive. When the curve parameters are known, ϕ first drops and then climbs, and we observe that the minimum shifts to larger $\sigma_{\ln(A)}$ values when $\sigma_{\ln(k)}$ increases. An explanation to the final increase could be that a high variation in the values of A will produce some bad volume curves with low A-values, which will be cut with little time loss, and some very profitable volume curves with high A-values, which will be cut at a very late time. The ratio also seems to decrease with a high correlation between the curve parameters, and of course with a higher t_2 (see Figure 2.2). Here t_2 is only a minimum cutting time, and is not taking part in the parameter estimation, therefore the higher t_2 is, the later the *bad curves* can be cut. We note that in the case where θ is known, measurement time optimisation does not make sense.

Case of unknown θ

In this case the execution of the algorithm to find the (sub)optimal strategy takes a minimum of 2 hours for a simple situation with fixed values for the parameters used. If

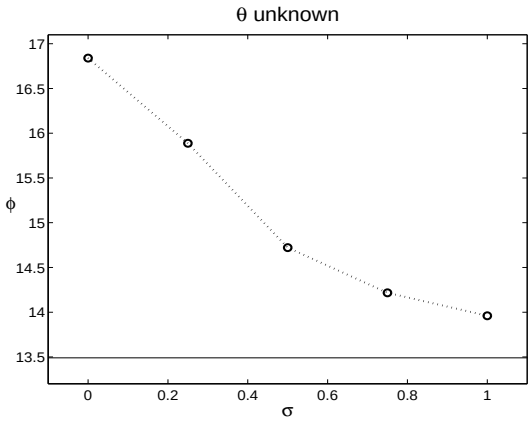

Figure 2.4: Dependence of ϕ on the standard deviation of the observations, σ, in the case where θ *is unknown*, $t_1 = 3$, $t_2 = 7$. The limit for ϕ when $\sigma \to \infty$ is also shown.

we want to run the algorithm for some different values, then the algorithm execution may take several days. Figure 2.3 shows that the effect of t_1 is much greater on the resulting ϕ than that of t_2. The fact that very early t_1 produces the best ϕ is caused by the chosen observation errors. These are constant in the ln scale, so the parameter estimation is considerably more accurate when t_1 is small. For fixed t_1 the curves show an optimum for t_2, as a result of two opposite effects: a higher t_2 improves parameter estimation, but it prevents early cutting. An increased standard deviation of the observations, σ, also produces a lower ϕ, as expected (see Figure 2.4).

2.4 Discussion

Although our goal is to determine the optimal design for a given prior distribution and observation error variance, we do not yet have results concerning optimal measurement times for the volume. The algorithm produced allowed us to check the dependence of ϕ on its components. We also conclude that ϕ is sensitive to the design choice. However, we have based our results on a very 'broad' prior distribution, corresponding to very heterogeneous plots, located in different points in Portugal. We expect a more homogeneous behaviour in the volume growth in plots confined to a particular region of the country. Therefore we recommend the use of a sharpened prior in a practical situation.

Acknowledgements

Research of the first author was sponsored by financial support from the Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE) through the program 'III Quadro Comunitário de Apoio' , to which we feel grateful. The authors wish to thank Stora Celbi, Celulose Beira Industrial, S.A., and M. Tomé for allowing the use of the eucalyptus data.

References

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

Berger, J. O., 1988: *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

Fedorov, V. V., 1972: *Theory of Optimal Experiments.* Academic Press, New York.

Paulo, M. J., Rasch, D. A. M. K. and Tomé, M., 1999: The robustness of locally optimal experimental designs in growth problems: a case study. *Bulletin of the International Statistical Institute* 52nd Session, Tome LVIII, Book 3, 105-106. Helsinki.

Chapter 3

A Bayesian approach for exact optimal measurement and cutting times

Maria João Paulo and Albert Otten

Submitted to Environmental and Ecological Statistics

In a eucalyptus production forest the owner is traditionally interested in cutting the trees at the age which maximises the yearly volume production. For a particular growth curve, this age is called the biological rotation age and is determined by the line through the origin tangent to the volume growth curve. In this study we consider a more general situation where a different growth curve, with a known prior distribution, can occur in each rotation. The goal now is to optimise the long term (volume) production, here defined as long term yearly volume yield reduced by costs of replanting. In this situation the optimal cutting time at each rotation depends both on the current growth curve and on the prior distribution. In this case we have two problems: estimating the growth curve for a particular rotation, and finding the optimal cutting time for that rotation. We assume that two volume measurements can be performed in each rotation, before deciding when to cut. The first measurement is always taken at a fixed age, but the age to perform the second measurement can be optimised, depending on the first observation. We compare some different priors and strategies with respect to the long term production. Volume is always assumed to grow according to a Shumacher curve. Despite the simple form of the curve, optimisation requires the use of numerical methods.

3.1 Introduction

The study is based on 158 eucalyptus plots belonging to a Portuguese pulp mill, Stora Celbi, Celulose Beira Industrial, S.A. The study plots are very diverse in tree density and in quality index. Each plot was regularly measured between 4 and 25 times during one rotation, at different ages (one rotation is the period of time between seeding or planting trees in a forest and their final cut). The volumes (m^3ha^{-1}) were subsequently derived from measurements of tree height and diameter at breast height, for each measurement time. The Shumacher growth curve $V = Ae^{-k/t}$ is used here to model the volume as a function of time t. For this model, and under the assumption that the same volume curve is drawn at each rotation, k is the biological rotation age, and the corresponding long term yearly yield is $(A/e)/k$. However, it is assumed here that a different curve for volume growth may be observed in each rotation. Here we use for short $\theta = (A, k)$. We assume that two observations in the ln-scale $y_i = \ln(A) - k/t_i + \varepsilon_i$, $i = 1, 2$ are to be made at times t_1 and t_2 for each rotation, with ε_i independent $\mathcal{N}(0, \sigma^2)$. The parameters (in the ln-scale, $\ln(\theta) = (\ln(A), \ln(k))$) are then fit by least squares, in the ln-scale. Based on the data and under the Shumacher model, a prior distribution was obtained for the curve parameters. It was found that for the pooled data, $\widehat{\ln(\theta)} = (\ln(\hat{A}), \ln(\hat{k}))$ followed approximately a bivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, with

$$\mu = \begin{pmatrix} 6.28 \\ 2.56 \end{pmatrix}, \ \Sigma = \begin{pmatrix} 0.512 & 0.171 \\ 0.171 & 0.134 \end{pmatrix} \tag{3.1}$$

Σ in (3.1) includes the variation both due to prior variation in $\ln(\theta)$, and due to variation in $\widehat{\ln(\theta)}$ given $\ln(\theta)$. The fact that we can use prior information has consequences for the choice of cutting time, because if the underlying family of growth curves is known, it becomes advantageous to reject a 'bad' growth curve by cutting it earlier than the biological rotation age. On the other hand if the observed curve is a very good one, then cutting it later than the biological rotation age improves the long term production. In order to make the problem more realistic we assume that there are fixed costs at each rotation (for replanting the stand).

In a previous study (Paulo and Otten, 2001) we discussed the optimisation of the cutting time when observations at fixed times t_1 and t_2 can be used. We studied the behaviour of the long term production for changes in fixed t_1 and t_2, in error standard deviation of the ln-observations and in parameters of the prior distribution (all changes were done one-at-a-time). Here we extend that previous study, by considering two strongly interfering optimisation problems: to find an optimal design (t_1, t_2) (with t_1 fixed for all curves, and t_2 variable) of times at which to measure volume before deciding when to cut, and to find the optimal cutting time at any rotation, based on the measurements performed at (t_1, t_2). Moreover, apart from approximate optimisation, exact optimisation

Table 3.1: Overview of the possible combinations of optimisation. Time t_1 is always fixed.

Situation	t_2	$\mathcal{C}$	function to maximise over $\mathcal{C}$	result
I	-	fixed optimal		$\mathcal{C}^{\mathrm{f}}$
II	fixed	sub-optimal	$\hat{A}\mathrm{e}^{-\hat{k}/\mathcal{C}} - \rho\mathcal{C}$, where $(\ln(\hat{A}), \ln(\hat{k}))$ are	
III	sub-optimal	(highest pos-	hpd-estimates of the curve parameters	$\mathcal{C}_\rho^{\mathrm{a}}(y_1, t_2, y_2)$
IV	optimal	terior density)	given y_1, t_2, y_2	
V	fixed	optimal		
VI	sub-optimal	(exact numeri-	$E_{A,k\mid y_1,t_2,y_2}(A\mathrm{e}^{-k/\mathcal{C}} - \rho\mathcal{C})$	$\mathcal{C}_\rho^{\mathrm{e}}(y_1, t_2, y_2)$
VII	optimal	cal integration)		

will be carried out now. We compare the situations where t_2 and the cutting time are fixed or optimised, and also when t_2 is suboptimal. We also compare the results obtained when the optimal cutting time is approximated or obtained with an exact procedure. We note that in this study it is not our aim to construct optimal designs for parameter estimation, such as D-optimal designs (as in Atkinson and Donev, 1992). Table 3.1 shows the possible combinations of optimisation.

3.2 Optimisation of the long term production

We assume from now on that a different volume growth curve can occur at each rotation, and that each curve is drawn from a Shumacher family of volume growth curves with the prior distribution π shown in the previous section. The long term volume production we want to maximise is defined by the ratio

$$\phi = E(V(\mathcal{C}) - V_0)/E(\mathcal{C})$$

where V is the total volume, V_0 is a constant representing a fixed cost of one rotation and $\mathcal{C}$ is the cutting time. The expectation E is taken over the joint distribution of A, k and $\mathcal{C}$. Practically, this expectation has to be expanded in iterated conditional expectations which may also involve conditioning with respect to the observations. If the curve is known and the maximum attainable value of ϕ, $\phi_{\max}$, is also known, analytical considerations show that the optimal cutting time is either the latest time when the tangent to the growth curve equals $\phi_{\max}$, or it is as early as possible (usually at time t_2). Since $\phi_{\max}$ is usually not known, we take a guess for for it, call it ρ, and use it instead. The cutting time optimisation problem is thus reduced to finding the ρ that maximises $\phi(\rho)$. By construction, this optimal ρ also satisfies $\rho = \phi(\rho)$ and $\rho = \phi_{\max}$. A similar idea is presented in Ribeiro and Betters (1995, eq.8) for a finite series of known growth curves. Obviously each value of ρ will define different cutting times for the same prior

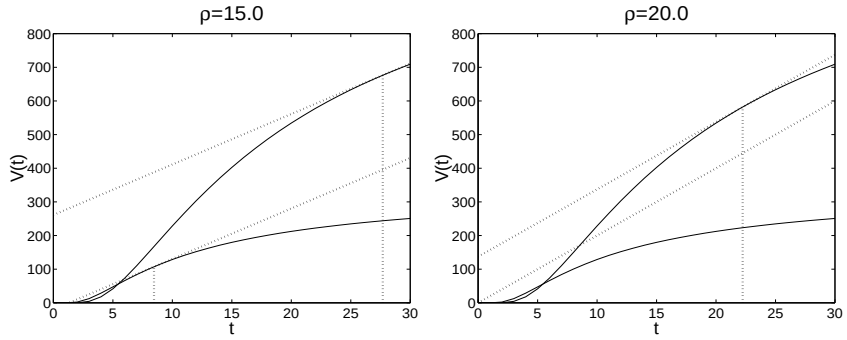

Figure 3.1: Example of two cutting times, optimal for ρ, for a good curve ($A_1 = 1250, k_1 = 17$) and for a bad one ($A_2 = 350, k_2 = 10$). Two different assumed ρ values (15 and 20) are use to show the different choices of cutting time. The curves $V(t)$ are shown in solid, and the tangents to the curves with a slope equal to ρ are shown with dashed lines. The corresponding optimal cutting time positions are also shown with dashed lines. For $\rho = 15$ (left) $\mathcal{C}_1 = 27.7$ and $\mathcal{C}_2 = 8.5$. For $\rho = 20$ (right) $\mathcal{C}_1 = 22.2$ and $\mathcal{C}_2 = t_2$ (taken 0 in the graph).

information, with different resulting $\phi(\rho)$. Although the cutting time (and later on also t_2) derived for a given ρ is not optimal in the final sense, we will use the term optimal here as well. Figure 3.1 shows the optimal cutting times for two different curves, and given two different ρ's. As seen in the figure, as ρ increases the optimal cutting times decrease. We note here that if the cost V_0 of one rotation was not used then there would be no solution for the optimal cutting time, assuming the curve is known. In this case for most curves the cutting times would become unrealistically low, and the more extreme the selection of 'good' curves (with the rejection of 'bad' ones), the better.

$\mathcal{C}$ optimal, t_2 fixed

If θ is unknown, we take two measurements y_1 and y_2 at fixed times t_1 and t_2 respectively. Given the observations, we need to find $\mathcal{C}$ for which

$$E_{A,k|y_1,y_2}(A\mathrm{e}^{-k/\mathcal{C}} - \phi_{\max}\mathcal{C}) \tag{3.2}$$

is maximal. As before, $\phi_{\max}$ is unknown so we use ρ instead, and maximise

$$E_{A,k|y_1,y_2}(A\mathrm{e}^{-k/\mathcal{C}} - \rho\mathcal{C})$$

In the notation of Table 3.1: $\mathcal{C}_\rho(y_1, y_2) = \underset{\mathcal{C},\, \mathcal{C}\geq t_2}{\arg\max} \left\{E_{A,k|y_1,y_2}(A\mathrm{e}^{-k/\mathcal{C}} - \rho\mathcal{C})\right\}$ Later, we search for optimal ρ, ρ_{opt}, such that $\rho_{\mathrm{opt}} = \phi_{\max}$. For a given ρ and $\mathcal{C} = \mathcal{C}_\rho(y_1, y_2)$, the

evaluation of $\phi(\rho)$ requires four-fold integrals:

$$\frac{E_{\theta,y}\left[V(\theta,\mathcal{C}_\rho(y_1,y_2))-V_0\right]}{E_{\theta,y}\left[\mathcal{C}_\rho(y_1,y_2)\right]}=\frac{\int_k\int_A\int_{y_1}\int_{y_2}(A\mathrm{e}^{-k/\mathcal{C}_\rho(y_1,y_2)}-V_0)\,h(y_1,y_2,\theta)\,dy_2dy_1dAdk}{\int_k\int_A\int_{y_1}\int_{y_2}\mathcal{C}_\rho(y_1,y_2)\,h(y_1,y_2,\theta)\,dy_2dy_1dAdk}$$

Here $h(y_1,y_2,\theta)$ is the joint density of y_1, y_2 and θ.

$\mathcal{C}$ optimal, t_2 optimal

When optimising t_2, t_2 will become a function of y_1, and the distribution of y_2 will depend on y_1 through t_2. For clarity we will write t_2 explicitly in funtion arguments and in (conditional) distributions. The natural evaluation order of the overall expectation is now

$$E_{y_1}(E_{y_2|y_1,t_2}(E_{A,k|y_1,t_2,y_2})) \tag{3.3}$$

In the inner expressions it does not matter wether t_2 is fixed or a function of y_1. In order to clarify the optimisation strategies some extra notation is used for $\mathcal{C}$ and t_2 as functions of conditions and of ρ (see Table 3.1 and the derivations in the Appendix). Once a functional form $\mathcal{C}_\rho(y_1,t_2,y_2)$ is obtained for $\mathcal{C}$, the next step is to optimise t_2, still for a given ρ, as a function of y_1. Thus $t_{2\,\rho}(y_1)$ is obtained by maximising over t_2:

$$E_{A,k,y_2|y_1,t_2}[A\mathrm{e}^{-k/\mathcal{C}_\rho(y_1,t_2,y_2)}-\rho\mathcal{C}_\rho(y_1,t_2,y_2)] \tag{3.4}$$

After carefully studying the optimal t_2 as a function of y_1 for different ρ and t_1 simple approximating functions for the optimal t_2 were constructed. To distinguish between the different situations we use double indices, the first referring to t_2 and the second to $\mathcal{C}$. The indices can be 'e', 'a' and 'f' meaning exact optimal, approximate optimal, and fixed optimal, respectively. For example $t^{\mathrm{a,e}}_{2\,\rho}(y_1)$ refers to the situation where we use $\mathcal{C}^{\mathrm{e}}_\rho(y_1,t_2,y_2)$ and an approximating function for t_2.

The last step is to optimise ϕ over ρ once functions $t_{2\,\rho}(y_1)$ and $\mathcal{C}_\rho(y_1,t_2,y_2)$ are chosen:

$$\phi(\rho)=\frac{E[A\mathrm{e}^{-k/\mathcal{C}_\rho(y_1,t_{2\,\rho}(y_1),y_2)}-V_0]}{E[\mathcal{C}_\rho(y_1,t_{2\,\rho}(y_1),y_2)]}$$

where the expectations are taken over the joint distribution of y_1,y_2,A,k. The resulting optimal ρ will be written as $\rho^{\mathrm{e,a}}_{\mathrm{opt}}$ if t_2^{e} and $\mathcal{C}^{\mathrm{a}}$ are used, and the corresponding $\phi(\rho)$ will be written accordingly as $\phi^{\mathrm{e,a}}_{\mathrm{max}}$, etc. In particular $\phi^{\mathrm{f,e}}_{\mathrm{max}}$ will be the maximal attainable ϕ when t_2 is fixed optimised and $\mathcal{C}$ exact optimised. Once ρ_{opt} is determined, the previous functions have to be used by the farmer to determine first t_2 from y_1 and then $\mathcal{C}$ from y_1,t_2,y_2.

$\mathcal{C}$ sub-optimal using highest posterior density approximation

In Paulo and Otten (2001) the hpd approximation was introduced in order to save cpu time. The technique consists in using hpd estimates (Berger, 1988) $\ln(\hat{A}), \ln(\hat{k})$ for $\ln(A), \ln(k)$, given observations (y_1, y_2), to approximate (3.2) by

$$\hat{A}\mathrm{e}^{-\hat{k}/\mathcal{C}} - \rho\mathcal{C}$$

and then optimise over $\mathcal{C}$. The advantage of the approximation is that the mixture $E_{A,k|y_1,y_2}A\mathrm{e}^{-k/t}$ of growth curves is reduced to one Shumacher curve $\hat{A}\mathrm{e}^{-\hat{k}/t}$. The technique produces a sub-optimal cutting time as a function of y_1 and y_2. No modification is needed in hpd estimation when t_2 is a function of y_1, $t_2 = t_2(y_1)$. Denoting parameter densities by π, observations' densities by g and joint densities of parameters and observations by h, we have

$$h(\theta, y_1, y_2) = \pi(\theta)g(y_1|\theta)g(y_2|y_1, t_2 = t_2(y_1)) = \pi(\theta)g(y_1, y_2|\theta, t_2 = t_2(y_1))$$

and posterior density $\pi(\theta|y_1, y_2) = h(\theta, y_1, y_2)/g(y_1, y_2)$. The marginal joint density $g(y_1, y_2)$ is $t_2(.)$ dependent, but is 'fixed' in the task of maximising the posterior density for given y_1, y_2. The sub-optimal $\mathcal{C}$, still for given ρ, will be denoted $\mathcal{C}^{\mathrm{a}}_{\rho}(y_1, t_2, y_2)$.

We recall that in the final optimisation over ρ, in general $\rho_{\mathrm{opt}} \neq \phi(\rho_{\mathrm{opt}})$. Although the joint density can be composed as a product of normal densities, the presence of k or $\mathrm{e}^{\ln(k)}$ in $\ln(A) - k/t$ prevents us from giving explicit expressions for hpd estimates.

Numerical optimisation and integration

The expectations $E_{A,k|y_1,t_2,y_2}$ can be conditioned one step further, yielding

$$E_{k|y_1,t_2,y_2}(E_{A|y_1,t_2,y_2,k})$$

For given y_1, t_2, y_2, k the volume asymptote A has a known log-normal distribution. Furthermore, A is either not present or is an isolated multiplier in the target functions of $E(\cdot)$, so we can always write $E_{A|y_1,t_2,y_2,k}(\cdot)$ explicitly as a function of y_1, t_2, y_2, k. The corresponding reduction in the number of iterated integrals needed makes numerical integration feasible, but still at the cost of hours in cpu-time for the hardest situations (situation where both t_2 and $\mathcal{C}$ are exactly optimised). The main reason for large cpu loads is that functions $\mathcal{C}_\rho$ and $t_{2\,\rho}$ are repeatedly called by the optimisation and integration routines, with new arguments each time (and results from previous calculations are not reused). The calls branch and nest at several stages. For example, a call to $\phi(\rho)$ generates an integral over y_1; the integrand generates a second integral over y_2, but only after $t_{2\,\rho}$ has been evaluated. This $t_{2\,\rho}$ call in turn generates several tries for candidate t_2 values.

For each t_2 an integral over y_2 is generated; the integrand again needs $\mathcal{C}_\rho(y_1, t_2, y_2)$, etc. An extra complication is that some normalising constants in conditional densities too have to be calculated by numerical integration.

Implementation

The optimisation of the long term production was implemented in Fortran, with use of IMSL-routines whenever possible. Exact $\mathcal{C}$-optimisation was done with mixtures of Newton-Raphson and/or interval halving techniques. The implicitly used assumption that the shapes of the needed mixture curves were still of the sigmoidal type (like a single growth curve) was never violated in the many test cases we generated.

Exact optimisation of t_2 was performed only over integer values (or a finer grid). Optimisation over ρ was done either using iterated substitution (when t_2 and $\mathcal{C}$ were exactly optimised) or using a IMSL-routine based on the Newton-Raphson algorithm. For numerical integration the IMSL-routines were used.

Hpd estimates were also calculated using Newton-Raphson techniques. For fixed t_1 and t_2 an efficient way was to fill in advance a fine grid of (y_1, y_2)-values with hpd estimates and later use it to interpolate from this grid. This approach did not work in the case of t_2 varying with y_1.

3.3 Results

Cutting time fixed but optimised

From now on times t_1 and t_2 are in y (years), and ϕ is in $\mathrm{m^3ha^{-1}y^{-1}}$. Furthermore, the costs of one rotation are set to $V_0 = 50$ $\mathrm{m^3ha^{-1}}$, and for practical reasons $\mathcal{C} \leq 30$. Having a fixed cutting time (situation I) in the present setting is not an optimal choice, but it is important to consider this situation to compare it with an optimal situation. Furthermore, if the gain of an optimal procedure is small compared to the investment of performing the two measurements on the forest, then this can be an attractive alternative. The optimal fixed cutting time is then $\mathcal{C} = 17.6$, and yields $\phi = 13.5$. This same value was obtained in an earlier study (Paulo and Otten, 2001) as a limiting situation when the standard deviation of the observations y_1 and y_2 was increased to infinity. Figure 3.2 shows how ϕ varies with fixed $\mathcal{C}$.

Cutting time optimised, t_1 and t_2 fixed

In this situation (V) the algorithm is run for a fixed (t_1, t_2) and guess values for ρ. We then need to find the optimal ρ, i.e., ρ for which ϕ is maximal for (t_1, t_2). Situation V

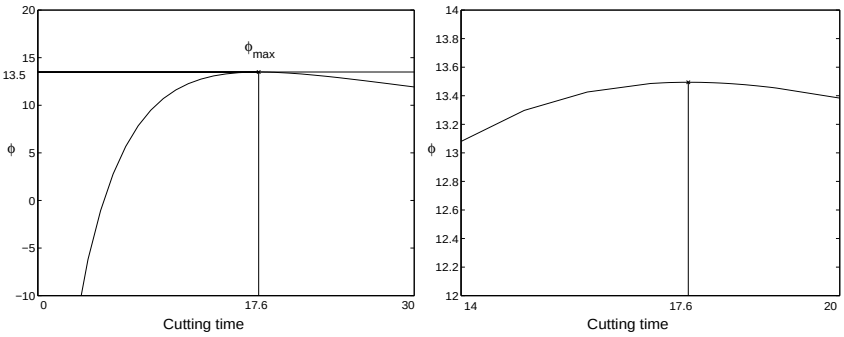

Figure 3.2: Situation where the cutting time is fixed, shown in detail on the right graph.

was run a number of times, for discrete (t_1, t_2) values: $t_1 = 3, 5$ and several t_2 values. We did not search in earlier t_1 times because they might be too early to use in practice. For the original prior distribution, and assuming that $\sigma = 0.5$, the fixed optimal design is $(t_1, t_2) = (3, 10)$ yielding $\phi = 15.07$. The result reported in our earlier study was $(t_1, t_2) = (3, 9)$ when using the hpd approximation (situation II). For the other choices of prior the fixed optimal t_2 varies substantially. Figures 3.3 and 3.4 show ϕ for fixed t_1 for different prior standard deviation values, $\sigma_{\ln(A)}$ and $\sigma_{\ln(k)}$. We also changed $\mu_{\ln(A)}$ and $\mu_{\ln(k)}$ accordingly in order to keep $E(A)$ and $E(k)$ at the original values. The standard deviation values were modified by factors $\frac{1}{2}$ and 2 ($\sigma_{\ln(A)}$) and by factors $\frac{1}{2}$ and 1.3 ($\sigma_{\ln(k)}$). The original value of $\sigma_{\ln(k)}$ is also shown. Time $t_1 = 3$ with optimal t_2 produces maximal ϕ in almost every situation, $t_1 = 5$ is better for halved $\sigma_{\ln(k)}$ combined with original $\sigma_{\ln(A)}$. For double $\sigma_{\ln(A)}$ time t_2 should be as early as possible to account for the greater variation in parameter A, but for halved $\sigma_{\ln(A)}$ t_2 around 12 or later are optimal. As seen in the graphs non-optimal designs can be considerably worse than the optimal design. Figure 3.5 shows ϕ for different error standard deviations of the observations σ (left), and for different correlations between $\ln(A)$ and $\ln(k)$ (right). σ was changed to 0.25 and to 0.75, and the original value 0.5 is also shown. The correlation between $\ln(A)$ and $\ln(k)$ was changed to 0.32 and 0.96, and and the original value 0.64 is also shown. The other parameters were left unchanged. An increase of σ leads to a fixed optimal design where t_1 and t_2 are further apart. In the case of changing the prior correlation we see that larger values produce larger fixed optimal t_2. Data in these graphs is not complete due to difficulties in some integral calculations.

The comparison of the exact optimal cutting time routine with the hpd approximation (V *versus* II) was done in terms of the optimal ϕ obtained in each situation and in terms of the computer time spent using each routine. We found that for fixed (t_1, t_2) the hpd

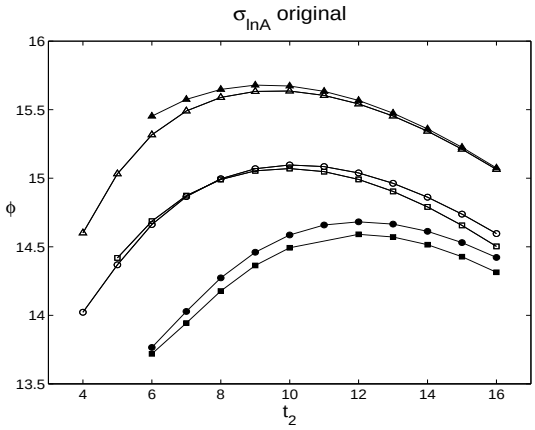

Figure 3.3: Fixed (t_1, t_2) and corresponding ϕ for the original $\sigma_{\ln(A)}$, and change in $\sigma_{\ln(k)}$ by factor 0.5 ($\triangle$), 1 (unchanged,$\circ$) and 1.3 ($\square$). Time t_1 is either 3 (empty symbol) or 5 (filled symbol).

routine always produced nearly the same optimal ϕ as exact numeric integration (the difference in optimal ϕ was up to 1%), and the time needed to run the exact computation of optimal cutting time is at least 100 times larger.

Cutting time optimised, t_1 fixed, t_2 optimised

The algorithm to find the exact optimal cutting time can still be further optimised with the simultaneous optimisation of t_2 (situation VII). This is a very computer intensive procedure, and is used here mainly to find an upper limit to ϕ. For a given prior distribution, optimal t_2 depends on y_1 and on ρ. The optimisation of a varying t_2 produced a negligible improvement (up to 0.1%) of optimal ϕ, compared with optimal fixed t_2. The approximating function for optimal t_2 (situation VI) allowed fast computations and at the same time produced an (even smaller) improvement of the resulting ϕ. Figure 3.6 shows some examples of optimal t_2 (restricted to discrete values only), as a function of y_1 (first measurement of volume, in the ln-scale), for fixed t_1 and for fixed ρ. As a result of our curve type and distributional assumptions, y_1 can attain very low values when t_1 is small. These low y_1 values correspond to volumes which would be too small to measure in practice. In order to preserve numerical accuracy all y_1 were accounted for during the integration process, but the optimal t_2 corresponding to very low y_1 values is very erratic and the final ϕ is insensitive for t_2 for such low y_1 values. These low y_1 values are therefore not shown in Figure 3.6, where y_1 is truncated to practical acceptable values.

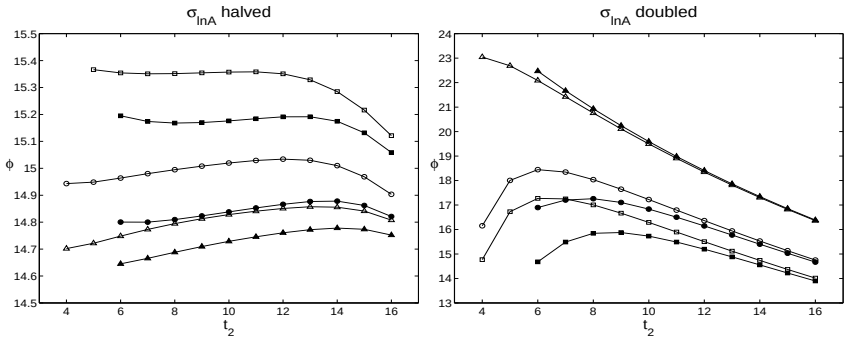

Figure 3.4: Fixed (t_1, t_2) and corresponding ϕ for change in the prior $\sigma_{\ln(A)}$ by factor $\frac{1}{2}$ (left) and 2 (right), and in $\sigma_{\ln(k)}$ by factor 0.5 ($\triangle$), 1 (unchanged,$\circ$) and 1.3 ($\square$). Time t_1 is either 3 (empty symbol) or 5 (filled symbol).

3.4 Discussion

In a previously published manuscript (Paulo and Otten, 2001) we found an approximate optimal cutting time and discussed the behaviour of the resulting long term production under changes in the prior distribution of the growth curve parameters and in the error standard deviation of the ln-volume observations. In this study we focus mainly on the optimisation of the volume production by optimising exactly the cutting time, and by choosing optimal times to measure volume in a stand for the second time, when guided by the first measurement. The exact optimisation of the cutting times was achieved through a change in the order of integration. The main difficulty here was that the integrand did not have an explicit form, and thus a lot of calls to other functions an integrals had to be made. As a result, computations became very heavy, and sometimes the numerical integration routine could not reach the specified numerical precision, and optimisation was not feasible.

The optimisation of the long term production depends obviously on the assumed type of growth curve, here it was the Shumacher type. We found that our objective function, the long term production, was unexpectedly insensitive to the use of non-optimal cutting times and non-optimal designs. We think that it could have been caused by the Shumacher curve properties. Even for one simple curve, the average yearly volume as a function of time is quite flat in the neighbourhood of time k. Other curve types might have produced sharper results. In this study we considered fixed replant costs, but the same results would have been obtained if we had considered random costs, or systematically changing costs as in the case of regeneration by coppice (Ribeiro and Betters, 1995). In fact only the mean of the costs influences the function we are maximising, so our procedure is immediately

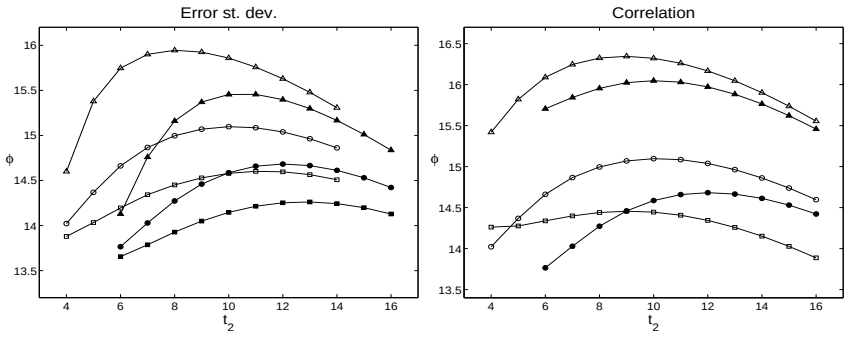

Figure 3.5: Fixed (t_1, t_2) and corresponding ϕ for different values of the error standard deviations of the observations (left): 0.25 ($\triangle$), 0.5 (original value, $\circ$) and 0.75 ($\square$), and for different values of the prior correlation (right): 0.32 ($\triangle$), 0.64 (original value, $\circ$) and 0.96 ($\square$). Time t_1 is either 3 (empty symbol) or 5 (filled symbol).

applicable to variable costs per rotation as long as that value is known.

3.5 Conclusions

In this study we assume a Shumacher curve type to describe volume growth, a prior distribution for the curve parameters' distribution and that the errors of the two observations in the ln-scale come from a $\mathcal{N}(0, \sigma^2)$ distribution. Under these assumptions, the optimisation of the cutting time (instead of cutting at a fixed optimal time) allows an improvement of 16% of the long term volume production, using a fixed optimal design. Optimising the second measurement time gave a very small extra improvement of the long term volume production, which is disappointing. This could be due to the growth curve choice (Shumacher curve), which appeared to be very robust for the effect of parameter mis-specification on volume/time ratios. The use of an approximating function to optimal t_2 worked well, in the sense that it did improve the objective function while substantially reducing the computation time, but the improvements in the long term production were even more disappointing. Under our model we think that a fixed optimal design is good enough in practice to estimate the optimal cutting time, and recommend its use instead of an arbitrary design. We did not find any practically useful explicit relation between the fixed optimal design and the parameters of the prior distribution. For practical applications of our routines the farmer needs to have a prior knowledge of the growth curve type, and its parameters' distribution, and he needs a computer program to obtain t_2 and $\mathcal{C}_{\text{opt}}$.

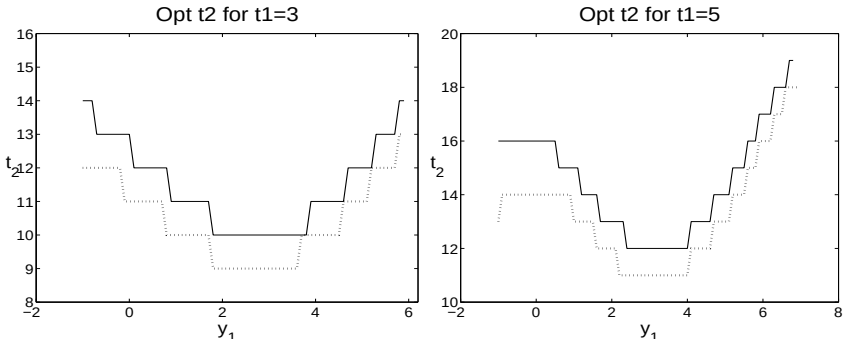

Figure 3.6: Optimal t_2 as a function of y_1, for $t_1 = 3$ (left) and for $t_1 = 5$ (right), for $\rho = 14$ (solid line) and for $\rho = 16$ (dashed line). Values of y_1 have been truncated. Piecewise linear functions were used to approximate optimal t_2.

Acknowledgments

Research of the first author was sponsored by financial support from the Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE) through the program 'III Quadro Comunitário de Apoio' , to which we feel grateful. Data used in this article was kindly made available by M. Tomé and Stora Celbi, Celulose Beira Industrial, S.A.

References

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

Berger, J. O., 1988: *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

Paulo, M. J. and Otten, A., 2001. Optimal Bayesian design applied to volume yield and optimal cutting time prediction. In Ermakov, S.M.; Kashtanov, Yu.N. and Melas, V.B. editors. *Proceedings of the 4th St. Petersburg Workshop on Simulation.* St. Petersburg University. 370-378.

Ribeiro, C.A.A.S. and Betters, D. R., 1995. Single rotation vs coppice systems for short-rotation intensive culture plantations - optimality conditions for volume production. *Biomass and Bioenergy* **8**, 395-400.

Appendix. Derivation of functions to be maximised

The need for maximisation of e.g. (3.4) over t_2 was stated without proof. A derivation of functions to be maximised is as follows. Suppose we choose specific functions $t_2(.)$ and $\mathcal{C}(.)$ which assign a second measurement time $t_2(y_1)$ to the first observation y_1 and a cutting time $\mathcal{C}(y_1, y_2)$ to the pair of observations (y_1, y_2), for all y_1 and y_2. Let ϕ be the attained long term production:

$$\phi = \frac{E[A\mathrm{e}^{-k/\mathcal{C}(y_1,y_2)} - V_0]}{E[\mathcal{C}(y_1, y_2)]} = \frac{E_1}{E_2}, \quad \text{say.}$$

Let $t_2'(.)$ combined with $\mathcal{C}'(.)$ be any other choice, leading to E_1', E_2' and ϕ'. Then the pair $\{t_2(.), \mathcal{C}(.)\}$ is optimal if and only if $\phi' \leq \phi$ for all $t_2'(.), \mathcal{C}'(.)$. Since the expected cutting time E_2' is positive, the latter condition is $E_1' - \phi E_2' \leq 0$ for all $t_2'(.), \mathcal{C}'(.)$. For $\{t_2(.), \mathcal{C}(.)\}$ we have $E_1 - \phi E_2 = 0$, so $\{t_2(.), \mathcal{C}(.)\}$ maximises $E_1' - \phi E_2'$ and the maximum is 0. Formulated in the finally used way: $\{t_2(.), \mathcal{C}(.)\}$ is optimal if and only if for some $\rho > 0$:

$$(i)\ E_1 - \rho E_2 = \max_{\{t_2'(.), \mathcal{C}'(.)\}} (E_1' - \rho E_2') \quad \text{and} \quad (ii)\ E_1/E_2 = \rho.$$

The solution to (i) and (ii) is found by substituting results of (i) for given ρ as a function of ρ in (ii), and by next solving (ii) as equation in ρ only. The maximisation of $E_1 - \rho E_2$ for given ρ can be solved entirely in successive steps by following the conditioning order of (3.3):

$$E_1 - \rho E_2 = E_{y_1}(E_{y_2|y_1,t_2(y_1)}(E_{A,k|y_1,t_2(y_1),y_2}[A\mathrm{e}^{-k/\mathcal{C}(y_1,y_2)} - V_0 - \rho\mathcal{C}(y_1, y_2)])).$$

For any values $y_1, t_2(y_1), y_2$ the inner expectation can be maximised separately. Let the function $\mathcal{C}_\rho(.)$ be defined such that $\mathcal{C}_\rho(y_1, t_2, y_2)$ maximises $E_{A,k|y_1,t_2,y_2}[A\mathrm{e}^{-k/\mathcal{C}} - V_0 - \rho\mathcal{C}]$ over $\mathcal{C}$ for any y_1, t_2, y_2. Then after substituting $\mathcal{C}_\rho$ in the inner expectation the middle expectation can be maximised separately for any y_1 by suitable choice of $t_2(y_1)$. Hence let $t_{2\,\rho}(.)$ be defined such that $t_{2\,\rho}(y_1)$ maximises

$$E_{y_2|y_1,t_2}(E_{A,k|y_1,t_2,y_2}[A\mathrm{e}^{-k/\mathcal{C}_\rho(y_1,t_2,y_2)} - V_0 - \rho\mathcal{C}_\rho(y_1, t_2, y_2)])$$

over t_2 for any y_1. Now for given ρ, requirement (i) leads to: for given y_1 measure again at time $t_{2\,\rho}(y_1)$, and for given y_1, y_2 cut at time $\mathcal{C}_\rho(y_1, t_{2\,\rho}(y_1), y_2)$. Note that V_0 can be skipped in (i) but not in (ii).

Chapter 4

Robustness and efficiency of D-optimal experimental designs in a growth problem

Maria João Paulo and Dieter A. M. K. Rasch
Biometrical Journal, **44** (2002) 5, 527-540

To assess tree growth, for example in diameter, a forester typically measures the trees at regular time points. We call such designs equidistant. In this study we look at the robustness and efficiency of several experimental designs, using the D-optimality criterion, in a case study of diameter growth in cork oaks. We compare D-optimal designs (unrestricted and replication-free) with equidistant designs. We further compare designs in different experimental regions. Results indicate that the experimental region should be adequate to the problem, and that D-optimal designs are substantially more efficient than equidistant designs, even under parameter mis-specification.

4.1 Introduction

In this study we used the D-optimality criterion to determine the best allocation of observations for the estimation of the unknown parameter vector θ of a given regression function $E(y_i) = f(x_i, \theta), i = 1, 2, \cdots, n$ with $\theta^T = (\theta_1, \theta_2, \cdots, \theta_p)$, the x_i from a given experimental region $\mathcal{X}$.

We used growth data of the diameter of 24 cork oaks from Portuguese forests. In order to compare the efficiency of several experimental designs, we first fitted a growth function to each of our 24 trees (24 empirical growth curves) and then we obtained the D-optimal design for each.

We looked at the robustness of a D-optimal unrestricted compromise design against parameter mis-specification. Further we looked at the relative efficiency of the equidistant design and designs in different experimental regions. In particular we wanted to compare D-optimal replication-free designs with the equidistant design. We propose the use of a compromise design (given by the average parameter vector) for all trees in the data set in further measurements.

The purpose of this study is to propose a design which is suitable for a high percentage of trees that farmers could encounter in practice. We think a farmer will be interested in and take action for a particular tree, where profit is not averaged over some prior distribution of tree parameters. Or, as far as a prior distribution is involved, it will vary between applications and will be narrowed in variability compared with that in our data set. Hence we do not follow the refinements in estimators and designs that would be offered by a random coefficient approach with known prior distribution as described for example in Fedorov et al.(1993).

4.2 Materials and methods

Data

Measurements of the annual diameter growth of each of 24 cork oaks with ages between 41 and 139 years (Tomé et al., 1999) were used. The experimental region $\mathcal{X}$ was set to the interval $[1, 144]$ to include all the ages of the trees and also because it was convenient as it will be seen later. We fitted five non-linear functions with 3 parameters by ordinary least squares on the measurements of each tree, and we used the residual variance criterion per tree to rank them. The Bertalanffy function provided a good fit to every tree, having for 22 trees the lowest or second lowest value of the residual variance. All the other functions fitted worse. In table 4.1 we show the geometric mean of the 24 residual variances for each function.

Table 4.1: Non-linear functions used to fit diameter growth.

Name	Expression	Geometric mean of residual variance
Bertalanffy	$f(x) = (\alpha + \beta e^{\gamma x})^3$	0.50
Gompertz	$f(x) = \alpha e^{\beta e^{\gamma x}}$	0.61
exponential	$f(x) = \alpha + \beta e^{\gamma x}$	0.80
logistic	$f(x) = \frac{\alpha}{1+\beta e^{\gamma x}}$	1.20
arc-tan	$f(x) = \frac{\alpha}{2}\{1 + \frac{2}{\pi}\arctan[\gamma(x-\beta)]\}$	2.46

We wanted to find one suitable family of curves to describe diameter growth for the cork oak and therefore chose the Bertalanffy function. For one tree, the Bertalanffy function with $p = 3$ parameters and parameter vector $\theta^T = (\alpha, \beta, \gamma)$ leads to the regression model

$$y_i = (\alpha + \beta e^{\gamma x_i})^3 + \epsilon_i \tag{4.1}$$

for the diameter growth. Here y_i is the measurement at time x_i and ϵ_i is the disturbance.

Twenty four individual parameter vectors $\theta^T = (\alpha, \beta, \gamma)$ were estimated for the Bertalanffy function (for convenience we denote these 24 estimates as θ_i instead of $\hat{\theta}_i$).

Model and least squares estimation

Suppose we have the model $y_i = f(x_i, \theta) + \epsilon_i$, $i = 1, 2, \cdots, n$ and θ p-dimensional, with ϵ_i i.i.d. and having $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$. The least squares estimator $\hat{\theta}$ minimizes $\sum_{i=1}^n (y_i - f(x_i, \theta))^2$. The well-known linearization of the least squares problem with iterative improvement leads to a series of normal equations. The coefficient matrices are of the type $F^T F = \sum_{i=1}^n \nabla f(x_i, \theta) \nabla f^T(x_i, \theta)$. In linear regression we would have $E(\hat{\theta}) = \theta$ and

$$\text{var}(\hat{\theta}) = \sigma^2 (F^T F)^{-1} \tag{4.2}$$

provided that θ is estimable. In non-linear regression these properties hold in an asymptotic sense when $n \to \infty$ and the Jenrich conditions (Jenrich, 1969) are fulfilled. Hence, further on, $\sigma^2 (F^T F)^{-1}$ will be called the asymptotic variance-covariance matrix of $\hat{\theta}$ and we write $V(\hat{\theta})$ for this matrix. If the ϵ_i's are normally distributed then $(F^T F)/\sigma^2$ is also the Fisher information matrix at θ. Sometimes it is convenient to call $F^T F$ the information matrix of the regression problem, and we will do so later in the text. We will base design optimality on $V(\hat{\theta}) = \sigma^2 (F^T F)^{-1}$, regardless of the quality of the asymptotics.

Design theory

In an exact design for estimating θ, the value of x has to be specified for each observation. Equivalently, a series of distinct x-values $(x_i, i = 1, 2, \cdots, q)$ is given together with the number of replicates $(n_i, i = 1, 2, \cdots, q)$. The x_i are called the support points of the design, and $n = \sum_{i=1}^{q} n_i$ is the size of the design. When developing a design, the x_i have to be chosen in a given experimental region $\mathcal{X}$. An exact design ξ can thus be represented in the form

$$\xi = \begin{pmatrix} x_1 & x_2 & \cdots & x_q \\ n_1 & n_2 & \cdots & n_q \end{pmatrix}. \tag{4.3}$$

In a continuous (and normalized) design one specifies a discrete distribution over support points with real 'weights' m_i and such a design is written as

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_q \\ m_1 & m_2 & \cdots & m_q \end{pmatrix}, \quad \sum_{i=1}^{q} m_i = 1, \quad m_i > 0 \quad \text{real.} \tag{4.4}$$

Each design of the form (2.2) can be specified by a continuous design and by its size n (by setting $m_i = n_i/n$).

Replication-free designs are exact designs that have the form

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_q \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{4.5}$$

i.e. they are exact designs with one single measurement at each support point. Depending on the problem, the support points may have to satisfy side conditions. In this case study we want to take one year as the practical unit and therefore we require support points to be integers in the experimental time region. This gives a finite set of candidate support points. The equidistant design is an example of a replication-free design. Replication-free designs are of interest to us because in most cases tree diameter is measured no more than once a year. Whenever confusion may arise, we will call designs with no restriction on the support points nor on the number of replications *unrestricted*.

D-optimal designs

We consider the D-optimality criterion based on the determinant of the asymptotic variance-covariance matrix $V(\hat{\theta})$ which takes the functional form $V(\theta, \xi)$ now. In the context of exact designs, a D-optimal design is defined as:

$$\xi^* = \arg\min |V(\theta, \xi)|, \quad \text{subject to} \quad \text{support}(\xi) \subset \mathcal{X}, \, \text{size}(\xi) = n \tag{4.6}$$

i.e. given n and $\mathcal{X}$, the D-optimal design ξ^* minimizes the determinant of the asymptotic variance-covariance matrix, or equivalently, maximizes $|F^T F|$. If f is intrinsically non-linear $V(\theta, \xi)$ depends not only on ξ but also on θ and thus the D-optimal design will also depend on θ. Therefore such designs are called locally D-optimal.

We know (Fedorov, 1972, p.120) that the minimal number of support points q needed to find a D-optimal continuous design in any of our regression situations is restricted to $p \leq q \leq p(p+1)/2$. It is also known that for continuous D-optimal designs where $q = p$ optimal weights are equal (Fedorov, 1972, p.85). In the case of exact D-optimal designs the n_i are as equal as possible.

The information matrix for the Bertalanffy function can be written as:

$$F^T F = 9 \sum_{j=1}^{n} (\alpha + \beta \mathrm{e}^{\gamma x_j})^4 \begin{pmatrix} 1 & \mathrm{e}^{\gamma x_j} & \beta x_j \mathrm{e}^{\gamma x_j} \\ \mathrm{e}^{\gamma x_j} & \mathrm{e}^{2\gamma x_j} & \beta x_j \mathrm{e}^{2\gamma x_j} \\ \beta x_j \mathrm{e}^{\gamma x_j} & \beta x_j \mathrm{e}^{2\gamma x_j} & \beta^2 x_j^2 \mathrm{e}^{2\gamma x_j} \end{pmatrix} \tag{4.7}$$

with $j = 1, 2, \cdots, n$ numbering the observations. The asymptotic variance-covariance matrix is $V(\theta, \xi) = \sigma^2 (F^T F)^{-1}$ and the D-optimality criterion for the Bertalanffy function is thus given by:

$$|V(\theta, \xi)| = \frac{\sigma^6}{|F^T F|}. \tag{4.8}$$

Schlettwein (1987) showed that for the Bertalanffy function, these asymptotic approximations are good for D-optimal designs, even for small n; see also the discussion in Rasch (1995a, p.631).

There is no analytical solution to find the D-optimal unrestricted design so the problem has to be solved numerically for each set of parameters. In our problem we take $\sigma^2 = 1$ w.l.o.g.

We used the program CADEMO to find a locally D-optimal exact design for each parameter vector θ_i. We took $n = 12$ measurements, which seemed a good design size to work with in practice. The experimental region was chosen to be $\mathcal{X}_{144} = [1, 144]$ in order to simplify the partition of the interval in twelve subintervals. Other choices for $\mathcal{X}$ and n were also possible.

Algorithm used to find replication-free designs

The algorithm in Rasch et al. (1995) finds the D-optimal replication-free design by full enumeration i.e. by evaluating all possible subsets of n integer points from the set $\mathcal{X}$. It was verified that the points x_i are allocated in the neighbourhood of the support points of the D-optimal unrestricted design. This algorithm performs very well for a small number of candidate points, that is, experimental regions $\mathcal{X}_H = [1, H]$ with H up to 40, but in the present case study we also have larger $\mathcal{X}$ so we could no longer use it. Our experience with fine grid replication-free D-optimal designs is that the design points appear in clusters around those of the unrestricted D-optimal designs. We used this knowledge in the following heuristic algorithm, based on the sequential construction of a D-optimal design (Atkinson and Donev, 1992) to obtain replication-free designs with integer points:

1. Let the D-optimal unrestricted exact design with n observations be given by $\xi^*(\theta) = \begin{pmatrix} x_1 & x_2 & \cdots & x_q \\ n_1 & n_2 & \cdots & n_q \end{pmatrix}$ and n be greater than q (i.e. at least some of the n_i's are greater than 1). The x_i's don't have to be integers but they have to be all different.

2. Start with the design $\xi^1 = \begin{pmatrix} x_1 & x_2 & \cdots & x_q \\ 1 & 1 & \cdots & 1 \end{pmatrix}$ and calculate $|V(\theta, \xi^1)|$ with expressions 4.7 and 4.8. Let $r = q$.

3. For $x\prime \in \{\langle x_1\rangle - 1, \langle x_1\rangle, \langle x_1\rangle + 1, \langle x_2\rangle - 1, \langle x_2\rangle, \langle x_2\rangle + 1, \cdots, \langle x_q\rangle - 1, \langle x_q\rangle, \langle x_q\rangle + 1\}$ in $\mathcal{X}$ but not already in ξ^r, calculate $|V(\theta, \xi^r \cup x\prime)|$. Choose $x\prime$ for which $|V(\theta, \xi^r \cup x\prime)|$ is minimal and let $\xi^{r+1} = \xi^r \cup x\prime$.

   If say x_d in ξ^{r+1} is non integer and is within distance 1 of $x\prime$ then delete x_d (obtaining thus a new ξ^r) and leave r unchanged. Otherwise set $r = r + 1$.

4. Repeat 3 until $r = n$.

For the smaller $\mathcal{X}$ it was verified that both algorithms found the same design. The replication-free designs $\xi^*_{rf}(\theta_i)$ were determined for θ_i, $i = 1, \ldots, 24$ and $\mathcal{X}_{144}$, and for θ_C (see next section), for some H values, both for $n = 12$ and for $n = 20$.

We then compared the different designs, obtained for different H. In particular we wanted to see how close the D-optimal replication-free designs were to the D-optimal unrestricted ones and whether increasing the number of points from $n = 12$ to $n = 20$ in the replication-free compromise designs would result in a significant improvement of the D-criterion value, and eventually compensate for the loss of efficiency due to a shorter $\mathcal{X}$. To compare the designs in a simple and informative way we plotted the D-criterion values, actually $|V(\theta, \xi)|^{1/3}$, for all $\mathcal{X}$ and all designs calculated for θ_C.

Efficiency of experimental designs

Given two designs ξ_1 and ξ_2, we can measure the efficiency of design ξ_1 with respect to design ξ_2, at θ, by $\left\{\frac{|F^T F(\xi_1)|}{|F^T F(\xi_2)|}\right\}^{1/p}$ (cf. Atkinson and Donev, 1992), or equivalently by

$$E = \frac{|V(\theta, \xi_2)|^{1/p}}{|V(\theta, \xi_1)|^{1/p}} \tag{4.9}$$

This measure is proportional to the design size of ξ_1 regardless of the dimension p of the model, so that for example two replicates of design ξ_1 for which $E = 0.5$ would be as efficient as one replicate of ξ_2. Usually we want to know the efficiency of some non-optimal design ξ_1 with respect to an optimal design ξ_2.

In practice an initial guess of θ may be quite bad, or we may want to work with one parameter vector for all trees instead of one for each tree. With a robustness measure we wanted to evaluate the performance of optimal designs in the case of a mis-specification of θ, that is, to see how much information about θ is preserved when an optimal design is used for another value of θ.

In this case we used a central value of θ, given by the average parameter $\theta_C = \frac{1}{24}\sum_{i=1}^{24}\theta_i$, as the mis-specified (but easily estimated) value. Then we checked how robust the replication-free compromise design $\xi^*_{rf}(\theta_C)$, D-optimal for θ_C, would be when used at $\theta_1, \theta_2, \cdots, \theta_{24}$. The robustness of the D-optimal replication-free compromise design against parameter mis-specification was calculated by expression (4.9), and setting $\xi_1 = \xi^*_{rf}(\theta_C)$ in $\mathcal{X}_{144}$ and $\xi_2 = \xi^*_{rf}(\theta_i)$ in $\mathcal{X}_{144}$, $i = 1, 2, \cdots, 24$.

Another measure we wanted to look at was the efficiency of the equidistant design ξ_{eq} with respect to the locally D-optimal replication-free designs $\xi^*_{rf}(\theta_i)$, also calculated by expression (4.9). The design ξ_{eq} has the support points {12, 24, 36, 48, 60, 72, 84, 96, 108, 120, 132, 144} and all $n_i \equiv 1$ (this is an interesting design because it is similar to those often used in practice).

The robustness of $\xi^*_{rf}(\theta_C)$ with respect to $\xi^*_{rf}(\theta_i)$, and the efficiency of ξ_{eq} with respect to $\xi^*_{rf}(\theta_i)$ are of interest in their own right. On the other hand, a comparison between these two measures should also provide information regarding how good ξ_{eq} is compared to $\xi^*_{rf}(\theta_C)$.

We also looked at the efficiency of $\xi^*_{rf}(\theta_k)$ with respect to $\xi^*_{rf}(\theta_i)$, $k, i = 1, \ldots, 24$ and $k \neq i$, to compare with the robustness of $\xi^*_{rf}(\theta_C)$.

The experimental region $\mathcal{X}_{144} = [1, 144]$ is too long in practice and it does not make much sense in the practical point of view to have measurements during 144 years. For comparison purposes D-optimal unrestricted designs for θ_C were re-calculated in shorter experimental regions $\mathcal{X}_H = [1, H]$, with $H \in \{24, 36, 48, 60, 72, 84, 96, 108, 120, 132\}$ and for $n = 12$. The efficiency of $\xi^*(\theta_C)$ for $\mathcal{X}_H$ with respect to $\xi^*(\theta_C)$ for $\mathcal{X}_{144}$ was evaluated

to see if by shortening the experimental region we can still get designs almost as good as those for $\mathcal{X}_{144}$.

Summary of designs used in the case study

A summary of designs used is given in table 4.2. The θ have meaning: $\theta_1, \theta_2, \cdots, \theta_{24}$: parameter vectors of the individual trees; θ_C: compromise, average value, $\theta_C = \frac{1}{24}\sum_{i=1}^{24}\theta_i$.

Table 4.2: Summary of designs in the case study.

Symbol	Indices and variables	Description
$\xi^*(\theta_i)$	$i = 1, 2, \cdots, 24$ $H = 144$ $q = 3$, $n = 12$	Locally D-optimal design for tree i
$\xi^*_{rf}(\theta_i)$	$i = 1, 2, \cdots, 24$ $H = 144$ $q = 3$, $n = 12$	Locally D-optimal replication-free design for tree i
$\xi^*(\theta_C)$	$H = 144$, 132, 120, 108, 96, 84, 72, 60, 48, 36, 24 $q = 3$, $n = 12$	Locally D-optimal for θ_C, or *compromise design.*
$\xi^*_{rf}(\theta_C)$	$H = 144$, 72, 60, 48, 36, 24 $q = n$, $n = 12$ and $n = 20$	Locally D-optimal replication-free for θ_C, or *replication-free compromise design.*
ξ_{eq}	$H = 144$, 72, 60, 48, 36, 24 $q = n$, $n = 12$	Equidistant design.

4.3 Results

The growth curves of the trees were very different as can be seen from the four examples in figure 4.1.

A brief study of the residuals showed presence of autocorrelation, but no sign of heteroscedasticity. We proceeded with OLS.

The locally D-optimal exact designs found were always designs with $p = q = 3$ support points. In every case one of the support points was equal to the maximum of the experimental region $\mathcal{X}$.

The average of the 24 parameter vectors was

$$\theta_C = (3.468, -2.416, -0.042)^T. \tag{4.10}$$

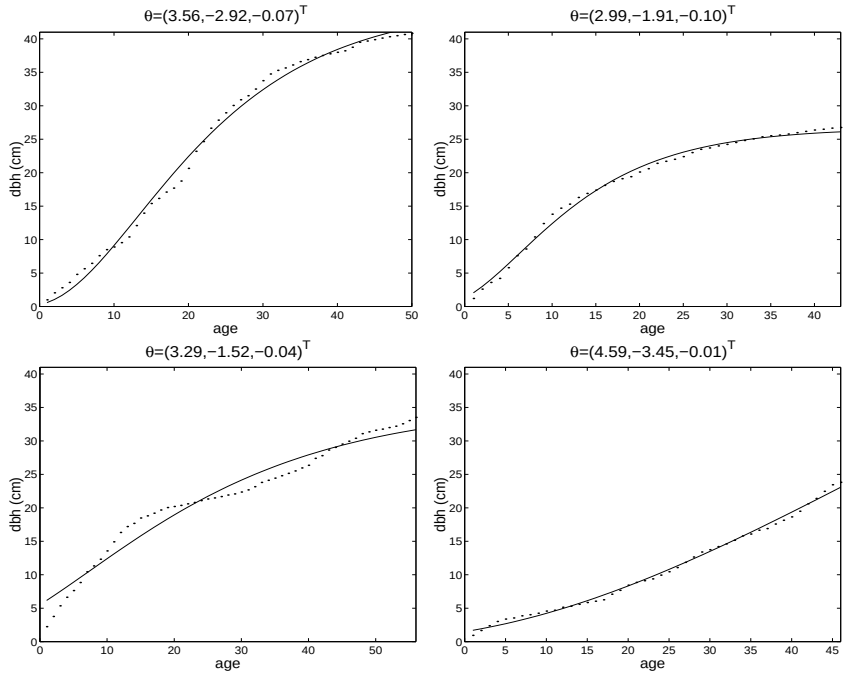

Figure 4.1: Empirical and fitted growth curves for diameter for 4 cork oaks, showing major differences in shape.

We verified that the efficiencies of the D-optimal replication-free $\xi^*_{rf}(\theta_i)$ relative to the D-optimal unrestricted designs $\xi^*(\theta_i)$ were nearly one, not surprising as the optimal replication-free design points were grouped around the optimal unrestricted design points. Therefore we think that the efficiencies in comparing two D-optimal designs will be nearly equal when using both designs unrestricted or both replication-free.

The values for the robustness of $\xi^*_{rf}(\theta_C)$ with respect to $\xi^*_{rf}(\theta_i), i = 1, 2, \cdots, 24$, calculated for $\mathcal{X}_{144}$, were 0.86 in average, being greater than 0.8 in 19 out of the 24 trees. The maximum value for the robustness was 0.997. The efficiency of the 12 point equidistant design however was 0.68 in average and its maximum value was 0.80 (see table 4.3). ξ_{eq} performed better than $\xi^*(\theta_C)$ only for two trees. From this we conclude that for the set of 24 trees the D-optimal unrestricted design for θ_C is, generally speaking, better than the equidistant design.

Table 4.4 displays the efficiencies of locally optimal designs when used with other trees, i.e, the efficiency of $\xi^*_{rf}(\theta_k)$ with respect to $\xi^*_{rf}(\theta_i)$, $k, i = 1, \ldots, 24$ and $k \neq i$. On the diagonal we show the robustness of $\xi^*_{rf}(\theta_C)$. We see that for any given tree the efficiency of $\xi^*_{rf}(\theta_C)$ is never inferior to the efficiency of a locally optimal design for another tree.

Table 4.3: $|V(\theta_i, \xi)|^{1/3} \cdot 10^5$ values and efficiencies of the (replication-free) optimal design, compromise design and equidistant design, for $\mathcal{X} = [1, 144]$.

θ_i	$\lvert V(\theta_i)\rvert^{1/3} \cdot 10^5$ for designs			Efficiency of ξ_1 relative to ξ_2		
	$\xi^*(\theta_i)$	$\xi^*(\theta_C)$	ξ_{eq}	$\xi^*(\theta_C), \xi^*(\theta_i)$	$\xi_{eq}, \xi^*(\theta_i)$	$\xi_{eq}, \xi^*(\theta_C)$
1	16.64	19.28	26.71	0.86	0.62	0.72
2	42.56	96.45	127.35	0.44	0.33	0.76
3	34.26	48.61	67.44	0.71	0.51	0.72
4	7.85	8.73	10.16	0.90	0.77	0.86
5	4.51	5.55	5.70	0.81	0.79	0.97
6	20.58	21.33	31.51	0.97	0.65	0.68
7	9.28	10.42	12.01	0.89	0.77	0.87
8	16.83	17.67	22.33	0.95	0.75	0.79
9	18.58	18.63	26.47	1.00	0.70	0.70
10	16.90	17.27	23.56	0.98	0.72	0.73
11	11.61	13.79	14.70	0.84	0.79	0.94
12	6.75	8.56	8.50	0.79	0.80	1.01
13	41.66	42.98	60.12	0.97	0.69	0.72
14	10.44	11.65	13.34	0.90	0.78	0.87
15	12.84	15.00	23.25	0.86	0.55	0.65
16	11.61	12.43	18.18	0.93	0.64	0.68
17	5.69	6.79	7.27	0.84	0.78	0.94
18	11.30	15.25	16.72	0.74	0.68	0.91
19	14.18	15.01	22.31	0.95	0.64	0.67
20	27.55	28.95	38.01	0.95	0.73	0.76
21	10.12	12.22	13.00	0.83	0.78	0.94
22	14.27	15.48	18.35	0.92	0.78	0.84
23	11.12	12.60	14.44	0.88	0.77	0.87
24	4.46	6.44	5.56	0.69	0.80	1.16

In fact, only three locally optimal designs seem to perform as well as $\xi^*_{rf}(\theta_C)$, namely the locally optimal designs for trees no. 9, 10 and 22.

The efficiency of the D-optimal designs for different $\mathcal{X}$'s can be seen in figure 4.2. The efficiency decreases very quickly to zero as the $\mathcal{X}$ shortens. By using shorter $\mathcal{X}$'s we have to accept estimators for θ with high variability.

The D-optimal unrestricted designs ($n = 12$) and D-optimal replication-free designs ($n = 12$, $n = 20$) for θ_C can be seen in table 4.5 for each $\mathcal{X}$. As mentioned before, $n = 20$ was used to find out whether a higher n would compensate for a smaller $\mathcal{X}$.

Figure 4.3 shows $|V(\theta, \xi)|^{1/3}$ for the designs in table 4.5. From that we see that for each experimental region the D-optimal and the D-optimal replication-free designs are clearly better than the equidistant design. Further, by changing the experimental regions we obtain different values of $|V(\theta, \xi)|^{1/3}$, and a larger experimental region seems essential to

Table 4.4: $E(\xi^*_{rf}(\theta_k), \xi^*_{rf}(\theta_i))^{1/3}$, $k \neq i$; in the diagonal $E(\xi^*_{rf}(\theta_C), \xi^*_{rf}(\theta_i))^{1/3}$ in $\mathcal{X} = [1, 144]$.

$\xi^*(\theta_k)$	θ_i											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0.86	0.72	0.91	0.65	0.62	0.90	0.66	0.79	0.90	0.81	0.67	0.67
2	0.75	0.44	0.94	0.42	0.38	0.71	0.43	0.40	0.63	0.58	0.34	0.34
3	0.94	0.92	0.70	0.55	0.51	0.84	0.55	0.60	0.79	0.72	0.50	0.50
4	0.50	0.28	0.37	0.90	0.99	0.81	1.00	0.91	0.87	0.95	0.97	0.98
5	0.38	0.21	0.28	1.00	0.81	0.69	0.99	0.83	0.76	0.87	0.95	0.96
6	0.83	0.58	0.77	0.82	0.78	0.96	0.82	0.79	0.97	0.96	0.74	0.74
7	0.49	0.30	0.37	1.00	0.99	0.81	0.89	0.88	0.87	0.95	0.95	0.96
8	0.74	0.33	0.54	0.88	0.87	0.90	0.88	0.95	0.97	0.95	0.94	0.96
9	0.87	0.51	0.72	0.85	0.82	0.98	0.85	0.91	1.00	0.97	0.84	0.84
10	0.72	0.43	0.59	0.92	0.89	0.97	0.92	0.88	0.98	0.98	0.86	0.86
11	0.46	0.18	0.30	0.97	0.98	0.75	0.97	0.94	0.84	0.91	0.84	1.00
12	0.37	0.14	0.24	0.97	0.98	0.68	0.97	0.88	0.77	0.86	0.98	0.79
13	0.93	0.53	0.76	0.77	0.75	0.94	0.77	0.92	0.97	0.90	0.81	0.82
14	0.50	0.24	0.35	0.99	0.99	0.80	0.99	0.94	0.87	0.94	0.99	1.00
15	0.72	0.69	0.81	0.72	0.67	0.97	0.73	0.58	0.88	0.89	0.56	0.56
16	0.97	0.69	0.89	0.72	0.69	0.96	0.73	0.82	0.95	0.88	0.71	0.71
17	0.49	0.19	0.32	0.96	0.96	0.77	0.95	0.96	0.86	0.91	1.00	1.00
18	0.21	0.28	0.22	0.89	0.84	0.64	0.89	0.42	0.60	0.80	0.56	0.56
19	0.74	0.59	0.75	0.80	0.75	0.99	0.80	0.67	0.93	0.94	0.65	0.65
20	0.86	0.43	0.66	0.81	0.79	0.91	0.81	0.97	0.96	0.91	0.87	0.88
21	0.40	0.27	0.32	1.00	0.99	0.71	1.00	0.81	0.77	0.89	0.92	0.93
22	0.57	0.25	0.39	0.97	0.96	0.84	0.97	0.98	0.91	0.95	0.99	1.00
23	0.48	0.30	0.36	1.00	0.99	0.80	1.00	0.87	0.86	0.95	0.94	0.95
24	0.25	0.11	0.17	0.93	0.95	0.52	0.93	0.70	0.59	0.74	0.87	0.89

$\xi^*(\theta_k)$	θ_i											
	13	14	15	16	17	18	19	20	21	22	23	24
1	0.94	0.76	0.85	0.99	0.60	0.59	0.89	0.88	0.66	0.71	0.66	0.45
2	0.58	0.46	0.76	0.81	0.27	0.44	0.73	0.44	0.43	0.40	0.43	0.27
3	0.79	0.62	0.85	0.96	0.42	0.53	0.85	0.67	0.55	0.56	0.56	0.36
4	0.78	0.99	0.71	0.58	0.94	0.89	0.78	0.80	1.00	0.98	1.00	0.88
5	0.65	0.94	0.58	0.44	0.94	0.90	0.65	0.68	1.00	0.94	0.99	0.95
6	0.90	0.87	0.96	0.93	0.64	0.77	0.99	0.79	0.83	0.81	0.83	0.59
7	0.76	0.98	0.71	0.58	0.91	0.91	0.78	0.77	1.00	0.96	1.00	0.88
8	0.96	0.98	0.79	0.79	0.91	0.74	0.86	0.98	0.89	0.96	0.88	0.69
9	0.98	0.93	0.91	0.93	0.77	0.76	0.96	0.93	0.85	0.88	0.85	0.62
10	0.90	0.96	0.90	0.82	0.78	0.85	0.95	0.84	0.93	0.91	0.93	0.71
11	0.76	0.99	0.63	0.53	1.00	0.83	0.71	0.81	0.97	0.99	0.96	0.87
12	0.67	0.95	0.57	0.44	0.99	0.83	0.64	0.73	0.97	0.96	0.96	0.92
13	0.97	0.88	0.85	0.94	0.76	0.67	0.91	0.98	0.78	0.85	0.78	0.56
14	0.79	0.90	0.68	0.57	0.97	0.86	0.76	0.82	0.99	0.99	0.99	0.87
15	0.74	0.74	0.86	0.90	0.45	0.73	0.99	0.55	0.73	0.65	0.74	0.50
16	0.96	0.82	0.91	0.93	0.64	0.66	0.94	0.88	0.73	0.76	0.73	0.50
17	0.80	0.99	0.65	0.57	0.84	0.81	0.73	0.85	0.96	0.99	0.95	0.84
18	0.38	0.70	0.64	0.34	0.46	0.74	0.65	0.29	0.89	0.64	0.90	0.84
19	0.80	0.81	0.99	0.89	0.54	0.78	0.94	0.64	0.81	0.74	0.81	0.57
20	0.99	0.92	0.81	0.88	0.83	0.68	0.88	0.95	0.81	0.89	0.81	0.60
21	0.65	0.94	0.62	0.47	0.89	0.93	0.68	0.67	0.83	0.93	0.99	0.94
22	0.85	1.00	0.72	0.64	0.98	0.83	0.80	0.89	0.97	0.92	0.96	0.82
23	0.75	0.98	0.70	0.56	0.91	0.91	0.77	0.75	1.00	0.96	0.88	0.89
24	0.48	0.83	0.43	0.29	0.89	0.84	0.49	0.52	0.93	0.85	0.92	0.69

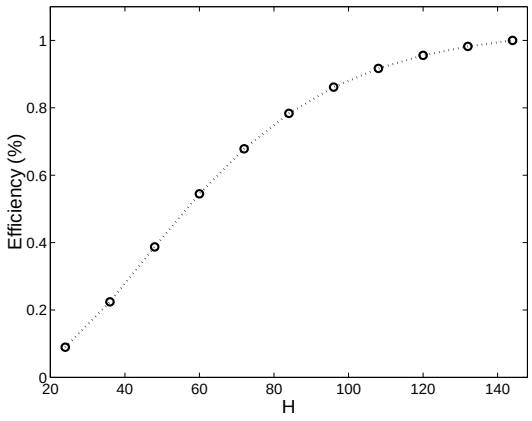

Figure 4.2: Efficiency of $\xi^*(\theta_C)$ for several experimental regions with respect to $\xi^*(\theta_C)$ for $\mathcal{X} = [1, 144]$.

minimize $|V(\theta, \xi)|^{1/3}$. By increasing the number of points in the replication-free design we manage to decrease the D-criterion value. In $\mathcal{X}_{48}$ and larger the 20-point replication-free design has a lower $|V(\theta, \xi)|^{1/3}$ value than the D-optimal design from the $\mathcal{X}$ immediately larger. For the smaller $\mathcal{X}$'s the increase by 8 points does not compensate the loss from shortening the $\mathcal{X}$.

4.4 Discussion

D-optimal designs provide an economic and efficient way to estimate unknown parameters of a growth curve. The trees of our sample had growth curves of the same family but with different parameters. We wanted to see if a common D-optimal design could be used to estimate the diameter growth parameters for all trees in a given forest since it would not be practical to use one design per tree. We took the average of the 24 parameters from the sample and found that under parameter mis-specification it provided a robust compromise design to use with all trees. This design performed better than the equidistant design, often used in practice. The result agrees with previously published work (Rasch et al, 1995b). Further, we saw that although replication-free designs are not as efficient as unrestricted designs they are better suited to the problem and are still better than equidistant designs. The experimental region should also be adequate to the curve. By shortening the experimental region we may loose too much information about the parameters and decrease the efficiency substantially. In general, increasing the replication-free design size in short experimental regions compensated the loss of

Table 4.5: Designs (D-optimal at θ_C, equidistant and D-optimal replication-free at θ_C) for each experimental region.

Experim. region	D-optimum ($n = 12$)[a] $\xi^*(\theta_C)$	Equidistant ($n = 12$) ξ_{eq}	Replication-free ($n = 12$) $\xi^*_{rf}(\theta_C)$	Replication-free ($n = 20$) $\xi^*_{rf}(\theta_C)$
$\mathcal{X}_{144}$	{9.52,41.36,144}	{12,24,$\cdots$,144}	{8-11,39-43,142-144}	{7-12,38-44,138-144}
$\mathcal{X}_{72}$	{7.79,34.39,72}	{6,12,$\cdots$,72}	{6-9,32-35,69-72}	{5-11,30-36,67-72}
$\mathcal{X}_{60}$	{6.76,30.39,60}	{5,10,$\cdots$,60}	{5-8,29-32,57-60}	{4-10,27-33,55-60}
$\mathcal{X}_{48}$	{5.59,26.11,48}	{4,8,$\cdots$,48}	{4-7,24-27,45-48}	{3-9,22-28,43-48}
$\mathcal{X}_{36}$	{3.97,20.79,36}	{3,6,$\cdots$,36}	{2-5,18-22,34-36}	{1-7,17-24,32-36}
$\mathcal{X}_{24}$	{1.74,14.32,24}	{2,4,$\cdots$,24}	{1-4,12-16,22-24}	{1-6,10-18,20-24}

[a] all with 4 replications per support point.

efficiency, except for our two shortest intervals. A few remarks should be made about the fitting of a theoretical curve to growth data. The assumptions made in section 2 might be non-realistic in some confounded aspects: the type of curve could be wrong, leading to lack of fit; the errors could be heteroscedastic and they could be substantially correlated at little time lag, degrading the quality of ordinary least squares (OLS) estimators and making the asymptotic variance formulae (4.2) for these estimators invalid when OLS is applied. An analysis of the residuals was performed visually and numerically, to see if these assumptions were violated in our case. We did not detect heteroscedasticity. The residuals were however highly autocorrelated. In order to check for the consequences of autocorrelated errors, we calculated the efficiencies presented in table 4.3 for first order autoregressive errors with serial correlations (using OLS-estimators as before). The resulting efficiencies showed a rapid degradation of the quality of the compromise design with respect to the equidistant design as the serial correlation coefficient ρ increased. To have a better impression of how the optimal compromise design would change when serial correlation is present, we recalculated the replication-free compromise design for several ρ values between 0.1 and 0.9, still using the OLS estimators. The design points obtained for $\rho > 0$ are still in the neighbourhood of the compromise design points obtained for $\rho = 0$. However, as the serial correlation is increased, the intervals between the resulting design points increases proportionally. For $\rho \leq 0.6$ we recommend to modify the unrestricted compromise design by spacing the replicate design points with 10ρ years in between. For higher values of ρ the equidistant design is a better option, since the space between consecutive design points becomes irregular.

We think that a good solution, not covered in this study, might require a model for the series of increments, instead of a model for growth curve measurements. In the latter

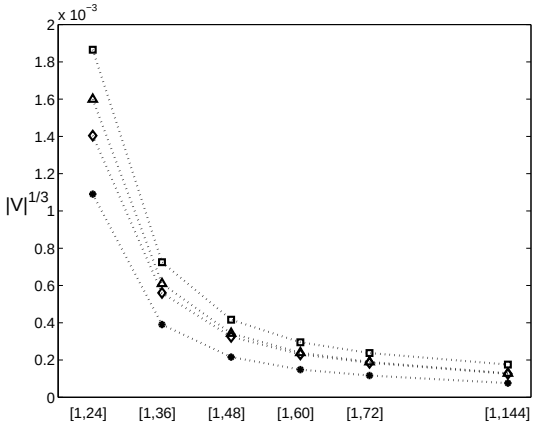

Figure 4.3: D-optimality criterion to the power 1/3 in different experimental regions, for designs: replication free 20 ($*$); replication free 12 ($\triangle$); D-optimal ($\diamond$) and equidistant ($\square$).

model one should also think carefully about what has to be estimated: background parameters, or function(s) of the realization of the stochastic process (cf. Cambanis, 1985; Fedorov, 1996).

Finally, the average parameter vector (θ_C) here merely serves as a tool to determine a compromise design; it is not intended to be an interesting population parameter to be estimated again later on. A type of problem not considered here would be to design estimation of a population parameter under constraints that the design is not too bad for individual trees. Constraint optimization is discussed e.g. in Cook and Fedorov (1995).

Acknowledgements

Research of the first author was sponsored by financial support from the Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE) through the program 'III Quadro Comunitário de Apoio' , to which we feel grateful. The authors wish to thank M. Tomé for letting us use the cork oak data. Further we thank A. Otten for reading and giving helpful suggestions on earlier versions of the manuscript. We would also like to thank E.P.J. Boer for the use of his algorithm.

References

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

Cambanis, S. 1985. Sampling designs for Time Series. *Handbook of statistics*, North-Holland, **5**, 337-362.

Cook, D. and Fedorov, V. 1995. Constrained optimization of experimental design. *Statistics* **26**: 129-178.

Fedorov, V. 1972. *Theory of Optimal Experiments.* Academic Press, New York.

Fedorov, V.; Hackl, P.; Mueller, W. 1993. Estimation and experimental design for second kind regression models. *Informatik, Biometrie und Epidemiologie in Medizin und Bioloogie.* **24**, 134-151.

Fedorov, V. 1996. Design of spatial experiments: model fitting and prediction. *Handbook of statistics*, North-Holland, **13**, 515-554.

Jenrich, R. 1969. Asymptotic properties of nonlinear least squares estimators. *The Annals of Mathematical Statistics* **40**, 633-643.

Rasch, D. A. M. K. 1995a. *Mathematische Statistik.* Joh. Ambrosius Barth, Heidelberg, Leipzig.

Rasch, D. A. M. K. 1995b. The robustness against parameter variation of exact locally optimum designs in nonlinear regression - a case study. *Computational Statistics & Data Analysis* **20**, 441-453.

Rasch, D. A. M. K., Hendrix, E. M. T. and Boer, E. P. J., 1997. Replication-free optimal designs in regression analysis. *Computational Statistics* **12**, 19-52.

Schlettwein, K. 1987. *Beiträge zur Analyse von vier Speziellen Wachstumsfunktionen.* Diplomarbeit Sekt. Math. WPU Rostock.

Tomé, M.; Coelho, M. B.; Pereira, H. and Lopes, F. 1999. A management oriented growth and yield model for cork oak stands in Portugal. A. Amaro e M. Tomé (ed.). *Empirical and Process-Based Models for Forest Tree and Stand Growth Simulation.* Edições Salamandra, Lisboa, Portugal, 271 -289.

Chapter 5

A spatial statistical analysis of cork oak competition in two Portuguese silvopastoral systems.

Maria João Paulo, Alfred Stein and Margarida Tomé
Accepted by the Canadian Journal of Forest Research

This study considers competition between cork oaks at three plots in two representative Portuguese stands. It uses spatial point pattern functions to describe densities and quantify differences between stands. Relations between cork oak characteristics and indices measuring inter-tree competition are modelled. Tree competition has a significant effect on tree crown characteristics. In particular, cork oaks with much competition have smaller and more elongated crowns. A standard model to relate crown diameter with diameter at breast height was improved. R^2 increased from 0.53 to 0.63 by including a crown shape parameter and competition indices.

5.1 Introduction

The object of this study is the cork oak (*Quercus suber* L.) in two Portuguese stands (montados). Worldwide, cork oak forests cover approximately 2.5 million ha, mainly in seven countries: Portugal (which contains 30% of the world's cork oaks), Algeria (21%), Spain (20%), Morocco (16%), France (5%), Italy (4%) and Tunisia (4%). In these stands the main product is cork, a thick and continuous layer of suberised cells, produced by the meristematic cork cambium (or phellogen), which makes up the external envelope of the stem and branches.

In Portugal, cork oaks are grown in silvopastoral agroforestry systems, called montados. In a montado cork oak trees grow in a low density and are sometimes inter-mixed with a small number of other tree species. Cattle or sheep graze in the same area. Tree density in montados is usually below 100 trees ha^{-1}.

Competition between trees influences the availability of nutrients and light and affects shape and size of crowns (Deleuze et al., 1996). On the other hand, crown condition and shape are obviously related to tree health and growth (Dawkins, 1958; Ottorini et al., 1996; Moravie et al., 1999; Gill et al., 2000). Most literature refers to relationships between tree growth and crown or tree growth and competition.

The aim of this study is to explore relations between crown size, tree size, crown shape and inter-tree competition for cork oaks. Crown diameter is strongly related to diameter at breast height. We explore the use of competition indices and crown shape parameters to explain differences in crown diameter. Such relationships allow us to estimate crown size using diameter at breast height and spatial information.

5.2 Data description

Cork and cork oaks

A cork oak has a life span of 300–400 years. Cork oak trees are economically viable for less than 150 years however, as cork growth intensity decreases with age, leading to cork that is too thin. The cork of the first harvest has a hard and irregular structure. The cork from the second harvest is more even, but only mature cork obtained at the third and following debarking on trees of 40 years of age or older reaches a perfect quality. A mature cork oak tree can produce more than 50 kg of cork in a single stripping.

During the first 40 years, a farmer has to make several investments before having any profits. Any decision during this period may have consequences on production in later years. To allow cork harvest, the management of cork oak stands includes thinning, shape pruning, understorey clearing and soil fertility improvement. Cork production is the main driving force of this system, whereas other products are efficiently used as well.

In Portuguese cork oak stands, much attention focuses on maintenance of cork quality. Production of cork is an important economic activity. Cork quality depends on the number and size of pores, the absence of defects such as insect galleries and the absence of great wood inclusions (Ferreira et al., 2000). The value of cork for industrial purposes highly depends on cork thickness. The highest value is associated with thicknesses between 29 and 40 mm. Cork quality is likely to be affected by environmental and local characteristics of the stand, such as tree density and competition. As producing large amounts of high quality cork is a lengthy and uncertain process, competition is an important topic to study.

In Portugal mainly two types of montados occur: adult montados that were regenerated in the past by natural regeneration or seeding, and new plantations with cork oak usually planted along lines. The first type is at the moment the most important as concerns cork production. Most of the new plantations are not yet ready for debarking. The adult stands are greatly variable in terms of stand structure and stand density and go from more or less regularly distributed to aggregated stands.

Study sites

Two montados are analyzed in this study. They are located approximately 60 km and 90 km east of Lisbon, respectively. The first, M_I, is located in Herdade do Vale Mouro, near the village of Coruche. The second, M_{II}, is located in Herdade Os Ruivos, near the village of Mora. They cover different spatial structures as occurring in montados that were selected by the local Association of landowners as representative in the Coruche region, which is important for cork production.

In M_I we measured 1 plot of a 200×200 m^2 size. It contains 389 cork oaks, of which 353 occur at production age and 36 are debarked for the first time. This plot is located in a flat terrain at an altitude of approximately 100 m. In M_{II} we measured 2 plots of a 140×150 m^2, plot $M_{II,A}$ and plot $M_{II,B}$, respectively. Plot $M_{II,A}$ contains 141 cork oaks and 9 trees of other species, whereas plot $M_{II,B}$ contains 145 cork oaks and 3 trees of other species. The plots at M_{II} are located at approximately 130 m of altitude, on a slightly uneven terrain. Difference in altitude within the plots is smaller than 1.5 m. The two montados were originally seeded with accorns and there has been grazing since the trees were large enough. The initial tree density is unknown. Age of these trees is hard to assess, as no written records are available. M_I has an uneven-aged structure, and the older trees are approximately 140 years of age. The M_{II} montado is closer to an even-aged stand, and the older trees are between 90 and 100 years. The soil is fertilized every 4 to 5 years and seeded to allow grazing.

Trees were measured shortly after cork extraction, during the month of July. Measured

variables were coordinates of tree location, diameter at breast height (d) without cork, total height (h), crown radius (c_α), mean crown diameter (d_c) and basal area (g)(Table 5.1). Crown radius was determined visually by stretching a tape from the tree bark to the edge of the projection of the crown on the horizontal plane, and using a compass to determine each direction. The crown was measured in 4 directions in M_I, a procedure commonly applied in sampling practices, and in 8 directions in M_{II} to test the effect of sample size on crown modelling.

Table 5.1: Variables measured in cork oak plots.

Variable	Description	Units
x	Horizontal coordinate of tree (azimuth 30° for M_I and 221° for $M_{II,A}$ and $M_{II,B}$)	m
y	Vertical coordinate of tree (azimuth 120° for M_I and 131° for $M_{II,A}$ and $M_{II,B}$)	m
d	Diameter at breast height	cm
h	Total height	m
c_α	Crown radius in direction α, $\alpha = k \cdot \pi/4$ ($M_{II,A}$ and $M_{II,B}$) $\alpha = \pi/6 + k \cdot \pi/2$ (M_I)	m
d_c	Mean crown diameter (obtained from c_α)	m
g	Individual-tree basal area (obtained from d)	m^2

5.3 Methods

Point patterns

A key factor governing tree competition is the frequency of small inter-tree distances for the same overall density. Competition is stronger with many small inter-tree distances, that occur more frequently in aggregated point processes than in random or regular point processes. Point processes are stochastic processes, whose realisations consist of point events in time or space called point patterns. To identify the point process underlying tree positions, a window W is defined for each plot, given by the plot boundaries. Let $N(\mathrm{d}\omega)$ denote the number of trees at an area of size $\mathrm{d}\omega$. Then the intensity $\lambda(\omega)$ at ω is defined as

$$\lambda(\omega) = \lim_{|\mathrm{d}\omega| \to 0} \{E\left[N(\mathrm{d}\omega)\right] / |\mathrm{d}\omega|\} \tag{5.1}$$

(Diggle, 1983), i.e. the number of trees in each window divided by the area of that window.

To compare the point pattern with a completely random spatial pattern (CSR), second order characteristics are applied. The nearest-neighbour distance distribution function $G(r)$ is defined as $G(r)$=P[distance from an arbitrary tree to the nearest other tree is at most r]. For any distance r the empirical $\hat{G}(r)$ uncorrected function is the number of trees with at least one neighbour within distance r, divided by the total number of trees. Similarly, the empty space function is given by $F(r)$=P[distance from an arbitrary point to the nearest tree is at most r]. The uncorrected $\hat{F}(r)$ function is the ratio of the total area of the window which is covered by circles of radius r centered in each tree, and the area of the window. In this study we focus on the $J(r)$-function based on the uncorrected $G(r)$ and $F(r)$ functions (Van Lieshout and Baddeley, 1996; Baddeley et al, 2000), defined as

$$J(r) = \frac{1 - G(r)}{1 - F(r)} \tag{5.2}$$

for which edge correction is not necessary. For the CSR process, $J(r) = 1$, whereas $J(r) < 1$ suggests clustering, and $J(r) > 1$ suggests regularity. To compare the actual point pattern with CSR, for $\mathrm{M_{II,A}}$ 100 simulations are made of CSR processes with the same intensity as in $\mathrm{M_{II,A}}$ and $\hat{J}^{(s)}(r)$ are calculated for $s = 1, \cdots, 100$, using maximum and minimum of $\hat{J}^{(s)}(r)$ as envelopes. These were plotted together with the estimated $\hat{J}(r)$ and the average $\hat{J}(r)$ of the simulations. The same analysis was done for the two plots $\mathrm{M_I}$ and $\mathrm{M_{II,B}}$.

Crown shape

Tree crown shape is largely determined by its vegetative growth characteristics and by competition (Biging and Gill, 1997). Ellipses are usually applied to graphically represent the cross-sections of tree crowns, usually based on 4 measured crown radii. To improve upon this, we measured 8 radii in $\mathrm{M_{II,A}}$ and in $\mathrm{M_{II,B}}$. In $\mathrm{M_I}$, measurements were made into the 4 directions $\pi/6 + k \cdot \pi/2$ for $k = 0, \ldots, 3$, and in $\mathrm{M_{II,A}}$ and $\mathrm{M_{II,B}}$ into the 8 directions $k \cdot \pi/4$, $k = 0, \ldots, 7$ clockwise from the north. For analytical purposes we described tree crowns as a polygon Π with 120 vertices, obtained from the original 4 or 8 crown radii. For every α_j with $0 \leq \alpha_{k-1} \leq \alpha_j \leq \alpha_k \leq 2\pi$ the radius is estimated by weighted linear interpolation

$$\hat{c}(\alpha_j) = \frac{(\alpha_j - \alpha_{k-1})}{(\alpha_k - \alpha_{k-1})} \cdot c_{\alpha_k} + \frac{(\alpha_k - \alpha_j)}{(\alpha_k - \alpha_{k-1})} \cdot c_{\alpha_{k-1}} \tag{5.3}$$

where $k = 1, \ldots, 8$ (with $\alpha_0 = \alpha_8$), $j = 1, \ldots, 120$ and c_{α_k} are the measured crown radii. The estimated radii $\hat{c}(\alpha_j)$ equal the weighted average of the two closest measured radii,

with weights inversely proportional to the absolute difference between angles. Such a crown representation is exact on the measured radii and can be applied on any number of measurements. In addition, no parametric shape is forced to the crown, whereas the final shape is smooth and has the same number of vertices regardless of the initial number of measured radii.

Shape parameters were calculated on the approximating polygons Π. A shape parameter is a function $S(\Pi) \rightarrow R^1$ that is invariant to any translation, rotation or re-sizing of polygon Π (Glasbey and Horgan, 1995, p.170). In this study, the area, perimeter, maximum diameter $d_{\text{c max}}$ and minimum diameter $d_{\text{c min}}$ of each polygon Π were calculated, as well as length (l) and breadth (b) as defined in Glasbey and Horgan (1995, p.153). The following shape parameters were used:

- compactness $ct = 4\pi \cdot \text{area}/(\text{perimeter})^2$. The compactness parameter compares the area of Π with the area of a circle with the same perimeter. Values for ct vary between 0 for a line segment and 1 for a circle.

- elongation $el = l/b$. The elongation parameter measures the length of Π as compared to its breadth. As el corresponds to fitting the vertical projection of Π into a rectangle with the same length and breadth, it varies between 1 and $+\infty$.

- eccentricity $ec = d_{\text{c max}}/d_{\text{c min}}$. The eccentricity parameter also measures the elongation of Π, comparing largest with smallest Π diameter. The ec parameter varies between 1 (when all diameters are equal) and ∞.

Competition indices

In this study competition effects between trees are modelled in terms of crown size and shape. Competition at the crown level is assumed to depend on the distance to neighbouring trees, as well as on their number and size. Therefore ten competition indices were selected from the literature (Moravie et al., 1999), and were adapted to properly measure aspects of competition. We used all trees in the plots to calculate the competition indices, therefore also trees from other species and border trees, i.e. trees within 10 m of the plot border.

Most indices involved local tree density, inter-tree distances and size of neighbours (Table 5.2). Distance independent indices were computed for search radii of 10, 15, 20 and 30 m around each tree. Correlation coefficients between crown shape and tree size parameters, d, h and d_{c} were computed, and bivariate plots were made to check for non-linear relationships.

To check for the influence of autocorrelation in the significance of correlation coefficients a size permutation test was performed. Observed tree sizes were randomly allo-

Table 5.2: Competition indices used in this study (Moravie et al., 1999). The index i refers to the subject tree, j refers to a competitor, ϱ is a size measure, such as d, h or g, and r_{ij} is the distance between tree i and tree j. Distances used were r=10, 15, 20 and 30 m (also ∞ for CI_{10}).

Index	Expression	Reference	$\hat{\rho}(CI, d_c)$
Distance independent indices			
CI_1	Number of trees (competitors) within r meters, (N_c)		-0.34
CI_2	Number of competitors within r meters such that $\varrho_j > \varrho_i$		-0.55
CI_3	Sum of size of trees within r meters, $\sum_{j=1}^{N_c} \varrho_j$	Steneker and Jarvis (1963)	-0.25
CI_4	Sum of basal area of bigger trees within r meters $\sum_{j=1}^{N_c} g_j \mathbf{1}_{(\varrho_j > \varrho_i)}$		-0.38
CI_5	Size ratio, $\frac{\varrho_i}{\varrho_i + \sum_{j=1}^{N_c} \varrho_j}$	Daniels et al. (1986)	0.55
Distance dependent indices			
CI_6	Distance to nearest tree (NN)		0.28
CI_7	Distance to NN such that $\varrho_j > \varrho_i$		0.51
CI_8	Difference in size with nearest tree $\varrho_{NN} - \varrho_i$		-0.46
CI_9	Size ratio proportional to distance $\sum_{j=1}^{N_c} \frac{\varrho_j}{\varrho_i} \frac{1}{r_{ij}}$	Daniels et al. (1986), Tomé and Burkhart (1989)	-0.54
CI_{10}	Size difference proportional to distance $\sum_{j=1}^{N_c} \frac{\varrho_j - \varrho_i}{r_{ij}}$		-0.68

cated to the observed tree locations and correlation coefficients between tree size and the competition indices were re-calculated. This was repeated 100 times, and the simulated correlations were compared with the observed one. The observed correlation coefficient was significant if its absolute value exceeded 95% of the simulated absolute correlations.

Directional crown parameters

Preferential growth direction may influence crown shape, as for example, isolated trees may have a preferential southern growth direction, where the crown intercepts most sunlight. For trees in a stand, crown competition from a preferential growth direction may affect trees more than competition from another direction. Analysis of the crown shape parameters alone is unlikely to reveal a preferential growth direction. A preferential growth direction could be found by analyzing summary statistics of crown radii of isolated trees. A tree with crown $c_{\alpha,i}$ is isolated from trees with crowns $c_{\alpha,j}$ at distances r_{ij} if $r_{ij} \geq \max_{\alpha} c_{\alpha,i} + \max_{\alpha} c_{\alpha,j}$ for all j. According to this definition, 9 isolated cork oaks occur in $\mathrm{M_{II,A}}$.

Crowns well exposed to the south are expected to have larger crowns as compared to crowns that are poorly exposed to the south direction. To test this for $\mathrm{M_{II,A}}$, let $\frac{c_{5\pi/4}}{c_{\pi/4}}$ and $\frac{c_{\pi}}{c_0}$ represent the ratio between crown radius directions south-west and north-east, and between crown radius directions south and north, respectively, and $\frac{c_{5\pi/4}-c_{\pi/4}}{c_{5\pi/4}+c_{\pi/4}}$, the relative difference between crown radius directions SW and NE . Scatter plots and correlations are used to study the relations between these parameters and crown size or tree size.

Modelling the crown diameter

For the relationship between d_c and d, Dawkins (1963) used a linear relation in tropical high forest trees. This linear relation is reportedly weak for trees from other forests (see for example De Gier, 1989). However the Portuguese National Forest Inventory currently uses a linear regression equation to estimate crown cover in cork oak montados (DGF, 1990). In this study the following relationships are explored:

- $d_\mathrm{c} = b_0 + b_1 \cdot d$
- $d_\mathrm{c} = b_0 + b_1 \cdot d + b_2 \cdot h$
- $d_\mathrm{c} = b_0 + b_1 \cdot d + b_2 \cdot d^2$
- $d_\mathrm{c} = b_0 + b_1 / d$

To improve upon that basic relationship we introduced spatial information such as competition indices into this model. Linear regression models for d_c with crown shape

measures, crown directional parameters and competition parameters as explanatory variables were determined. A first selection was made to eliminate indices strongly correlated with d. Stepwise regression with forward and backward elimination was then applied to remove non-significant contributors, using the S-Plus software. The procedure calculates the Cp statistic for the current model, as well as for reduced and augmented models. It adds or drops the term that mostly reduces Cp (MathSoft, 1997).

Data reduction

To investigate effects of crown measurement intensity, the crown data in $M_{II,A}$ were reduced from 8 to 4 and a 120-vertex polygon was fitted to both the full 8-radii data (Π_8) and the reduced 4-radii data(Π_4). Two sets of Π_4 were obtained for each crown, corresponding to measurements on orthogonal directions. Shape parameters $S(\Pi_8)$ were compared with shape parameters $S(\Pi_4)$. The average of the ratios for every tree in $M_{II,A}$ between $S(\Pi_4)$ and $S(\Pi_8)$ was used to measure its similarity, i.e. its logarithmic transformation should be close to zero. Let $\Delta = \log[S(\Pi_8)/S(\Pi_4)]$. The hypothesis $H_0 : E(\Delta) = 0$ was tested using Wilcoxon signed rank test ($\alpha = 0.05$).

5.4 Results

Description of the Plots

Summary statistics for the three plots in the two montados are given in table 5.3. Average d equals 32 cm at M_I, 40 cm at $M_{II,A}$ and 37 cm at $M_{II,B}$. The tallest cork oaks occur at $M_{II,A}$ ($\overline{h} = 10.9$ m), where trees are on average 1.6 m higher than at M_I ($\overline{h} = 9.3$ m) and 1.4 m higher than at $M_{II,B}$ ($\overline{h} = 9.5$ m). In $M_{II,A}$ tree height is more variable than in the two other plots, as the standard deviation is 2.8, whereas in M_I it is 2.0 and in $M_{II,B}$ it is 2.2. Average d_c is 7.0 m at M_I, 8.1 m at $M_{II,A}$ and 7.4 m at $M_{II,B}$. This indicates that cork oaks are largest in stem diameter, tallest and with the largest crown diameter in plot $M_{II,A}$ and are smallest in stem diameter, shortest and with the smallest crown diameter in plot M_I. Cork oaks in M_I are more variable in d and d_c than in $M_{II,A}$ and $M_{II,B}$. Plot $M_{II,B}$ is similar to plot $M_{II,A}$ in terms of tree sizes.

Point patterns

The windows W_I, $W_{II,A}$ and $W_{II,B}$ for plots M_I, $M_{II,A}$ and $M_{II,B}$ are given in Figure 5.1. Clearly, cork oaks are unequally spaced in $M_{II,A}$ and $M_{II,B}$, and more regularly spaced in M_I. Numbers of trees equal $|W_{II,A}| = 145$, $|W_{II,B}| = 146$ and $|W_I| = 380$, leading to process intensities equal to $\lambda_{II,A} = \lambda_{II,B} = 69$ trees ha^{-1} and $\lambda_I = 95$ trees ha^{-1}.

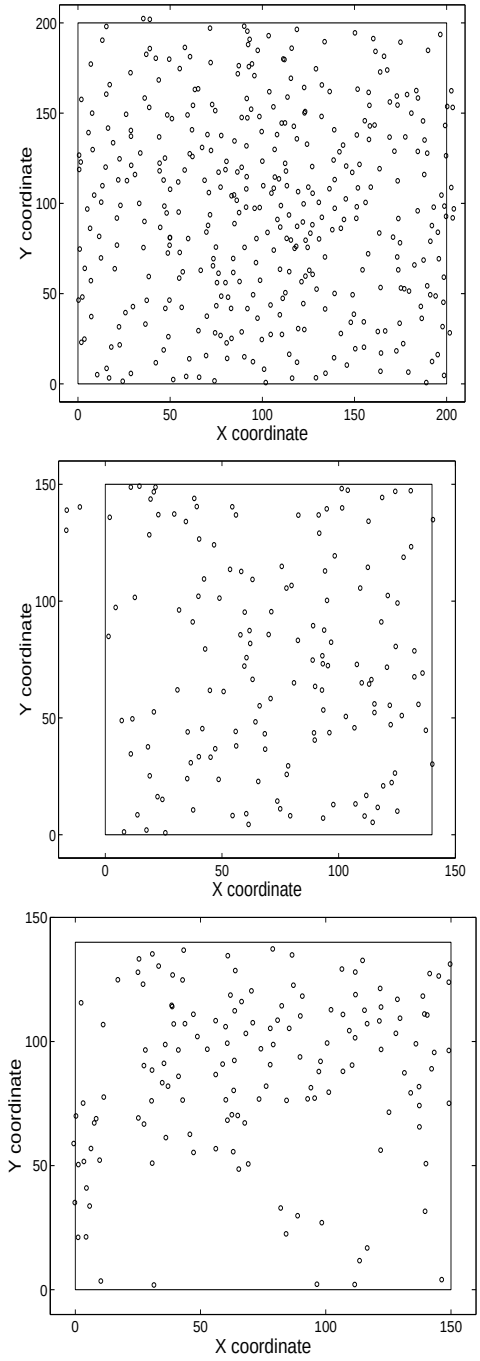

Figure 5.1: Tree locations in M_I, $M_{II,A}$ and $M_{II,B}$ (from top to bottom). The X-axis values increase in the 30^o azimuth in M_I, and in the 221^o azimuth in $M_{II,A}$ and $M_{II,B}$. The window (inner frame) defines the process area.

Table 5.3: Summary statistics of variables measured in the cork oak plots.

	x	y	d	h	d_c
M_I					
Min	0.3	0.7	13.5	4.4	2.1
Mean			32.0	9.3	7.0
Max	203.9	202.4	78.0	14.9	15.9
Std Dev.			13.0	2.0	2.8
$M_{II,A}$					
Min	-16.7	0.8	18.7	4.3	2.8
Mean			40.4	10.9	8.1
Max	140.1	149.2	71.3	19.8	14.0
Std Dev.			11.5	2.8	2.3
$M_{II,B}$					
Min	-0.7	1.9	17.8	3.8	2.2
Mean			37.1	9.5	7.4
Max	149.7	137.3	82.8	17.7	14.7
Std Dev.			11.5	2.2	2.1

The median inter-tree distances are approximately 6 m for all three plots, and they are larger than 4 m for 75% of the trees in the three plots. The $\hat{J}(r)$ function for the three plots, as well as the CSR envelopes and average, is shown in Figure 5.2. The $\hat{J}(r)$ function calculated for plot M_I has values greater than 1 for $r \leq 10$ m falling outside the upper CSR envelope. It shows that M_I has a more regular pattern. For $M_{II,A}$ $\hat{J}(r)$ is approximately equal to 1 for $r \leq 5$ m, and decreases for $r > 5$ m. Both for small and large values of r, $\hat{J}(r)$ values are inside the CSR envelopes, showing no significant deviation from CSR. Plot $M_{II,B}$ has more pronounced tree aggregation than plot $M_{II,A}$. The observed $\hat{J}(r)$ values are greater than 1 for $r \leq 5$ m and for $r \geq 7$ m they are smaller than 1. Some values are outside the CSR envelopes, suggesting that the underlying spatial process is aggregated. This result is likely to be related to the two large open areas observed in Figure 5.1. According to the farmer, initially the seeds did not develop in those areas, and the exposure to grazing on a later stage made natural regeneration impossible (*pers. comm.*). This type of open area is very common in Portuguese montados.

Crown shape

Figures 5.3 to 5.5 show the tree crowns in the plots, described as polygons. Clearly visible overlaps of crowns on these maps have been observed in the field. Differences in tree crown shape range from almost circular crowns to highly elongated crowns. Elongated crowns occur in trees that are close to other trees, whereas isolated trees display a more regular, round crown. Summary statistics of crown shape parameters from $M_{II,A}$ are given in

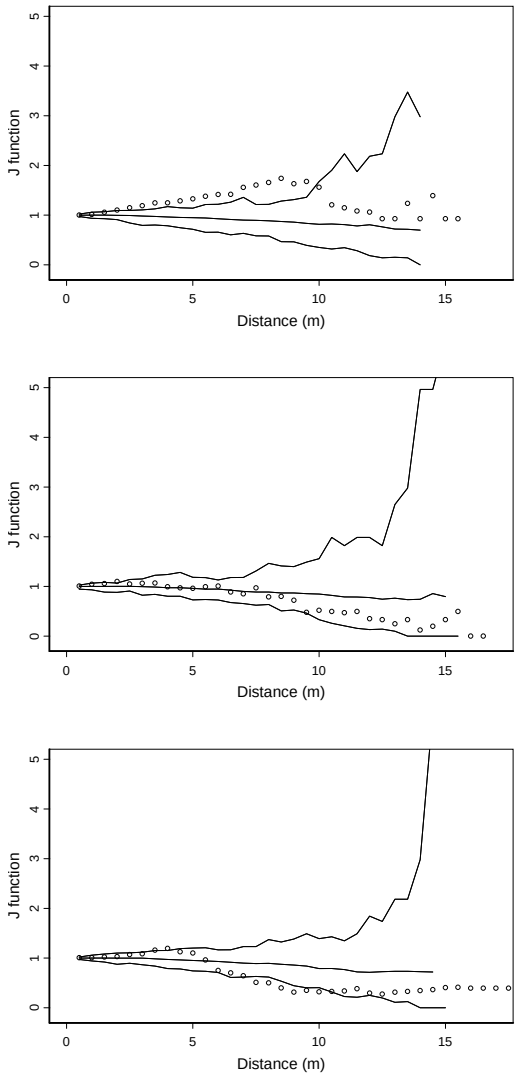

Figure 5.2: Empirical uncorrected J-function for M_I, $M_{II,A}$ and $M_{II,B}$ (from top to bottom).

Table 5.4. Values for compactness were on average 0.7, and the average elongation ($\overline{el}$) is 1.2. Average eccentricity ($\overline{ec}$) is 1.5, and in general $ec < 2$. Two cork oaks were removed from the data-set because their crown diameter in one direction was close to 0, thus yielding very large ec values.

Table 5.4: Descriptive statistics of shape parameters for the crown in $M_{II,A}$.

Variable	min	mean	max	variance
ct	0.4	0.7	0.9	0.01
el	1.0	1.2	1.9	0.03
ec	1.1	1.5	4.0	0.23

Crown shape parameters are uncorrelated to d ($|\hat{\rho}| < 0.15$) and h ($|\hat{\rho}| < 0.10$), but d_c was correlated to ec ($\hat{\rho} = -0.4$) and to a lesser extent also to ct and el ($|\hat{\rho}| < 0.3$). Crowns with a round shape may have a larger size than those that are elliptically shaped.

Competition indices

Competition indices in table 5.2 were computed using d, h or basal area to compare tree sizes, for different fixed values of r. Tree size (d and h) had a high correlation with indices CI_2, CI_4, CI_5 and CI_7 to CI_{10}, which account for the relative size of neighbours. For example, d was uncorrelated with the distance to the closest tree (CI_6, $\hat{\rho} \leq 0.1$), but it was highly correlated with the distance to the closest bigger tree (CI_7, $\hat{\rho} = 0.62$). This was also observed for h. Correlations with d_c were between $|\hat{\rho}| = 0.25$ for CI_3 and $|\hat{\rho}| = 0.68$ for CI_{10}. In general, indices with the number and/or distance to bigger neighbours had higher correlations with d_c (Table 5.2).

The size-permutation test showed that the correlation coefficients between h and the competition indices were all non-significant. Indices correlated with d were CI_2, CI_5, CI_7 and CI_9 ($\alpha = 0.05$). All indices were correlated with d_c, the observed correlation values largely exceeding the simulated ones. Correlations at the highest significance level occur for values of r up to 20 m.

Correlation is also present between competition indices and crown shape ($|\hat{\rho}| \approx 0.4$) in $M_{II,A}$. Compactness had the highest correlation with CI_2, CI_3, CI_4, CI_5, CI_6 and CI_9 ($0.3 \leq |\hat{\rho}| \leq 0.4$). Elongation was poorly correlated with all competition indices but eccentricity was highly correlated with CI_9 ($|\hat{\rho}| = 0.46$). Trees subject to competition are generally less compact, more elongated and have eccentric crowns.

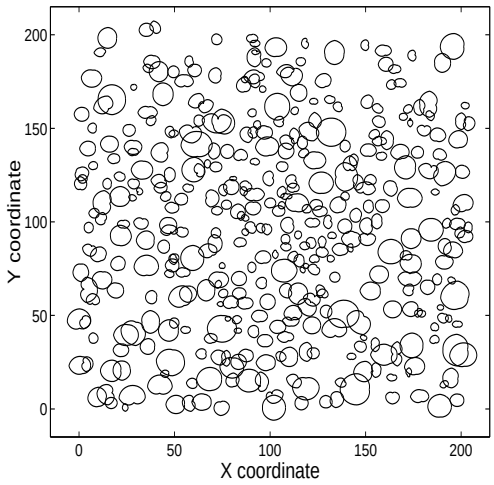

Figure 5.3: Map of tree crowns for M_I based on 4 crown measurements.

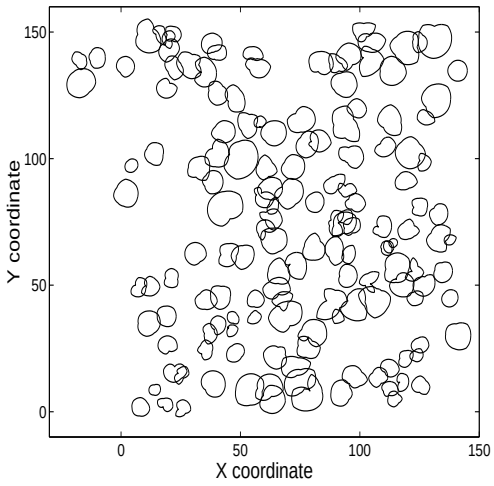

Figure 5.4: Map of tree crowns for $M_{II,A}$ based on 8 crown measurements.

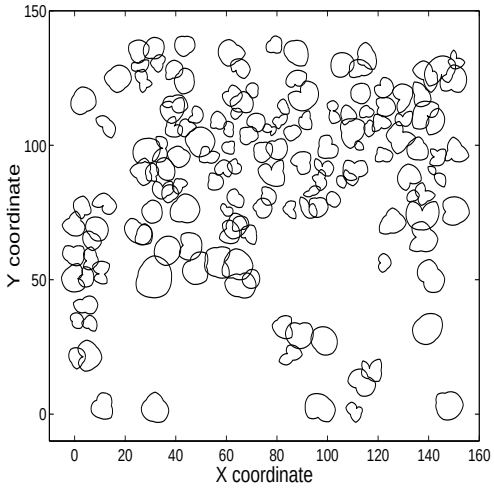

Figure 5.5: Map of tree crowns for $M_{II,B}$ based on 8 crown measurements.

Directional crown growth

The comparison of minimum, average and maximum crown radii of isolated trees and non-isolated trees is shown on figure 5.6. The size of the crown radii is more variable in the case of non-isolated trees. The summary statistics were calculated for each crown radius separately. Both groups of trees display some elongation towards the north-south direction. Seventy per cent of the trees in $M_{II,A}$ have a larger crown radius into the south direction than into the north direction. Also, the crown radius into the south is 25% larger than the radius into the north for half of the trees. Correlations between the directional crown parameters and d_c were all very low, and bi-variate plots showed no structural relations.

Models for crown diameter

Figure 5.7 shows the relationship between d_c and d for the three plots. A linear dependency is present for the observed values of d_c and d. Model $d_c = b_0 + b_1 \cdot d$ ($R^2 = 0.53$) fitted the data from all three plots better than $d_c = b_0 + b_1/d$ ($R^2 = 0.47$). Model $d_c = b_0 + b_1 \cdot d + b_2 \cdot d^2$ brought an improvement of at most 0.002 to the R^2 obtained for the simple linear model.

The variable h also did not improve the linear model $d_c = b_0 + b_1 \cdot d$ much, as it resulted in an increase of R^2 with only 0.02. Therefore, the linear model $d_c = b_0 + b_1 \cdot d$ was selected. The slope is larger for M_I than for $M_{II,A}$ and $M_{II,B}$. The predicted values

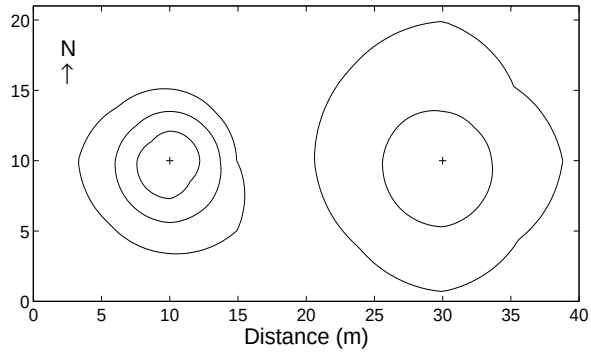

Figure 5.6: Comparison of average, minimum and maximum crown radii for isolated trees, and average and maximum crown radii for non-isolated trees, in $M_{II,A}$. The minimum radii of non-isolated trees are zero.

of d_c are larger in $M_{II,A}$ than in $M_{II,B}$, for all d. M_I has larger d_c than the two plots in M_{II} for $d \geq 32$ cm (Table 5.5 and Figure 5.7). A similar comparison was made between isolated and non isolated trees, by fitting the same model to each of the two groups of trees. The 9 isolated cork oaks in plot $M_{II,A}$ have larger predicted crown values for the same d than non-isolated trees in the same plot. However, the small sample size of isolated trees does not allow us to conclude that there is any difference in crown size between the two groups.

Table 5.5: Estimated parameters for a linear relationship $d_c = b_0 + b_1 \cdot d$ in the three plots.

Parameter	M_I	$M_{II,A}$	$M_{II,B}$
b_0	0.97	2.25	2.15
b_1	0.19	0.15	0.14
R^2	0.73	0.53	0.58

Bi-variate plots of d_c against correlated competition indices and against correlated crown shape parameters (not shown), indicate that the relationships are approximately linear. We obtained one improved model for mean crown diameter. Table 5.6 shows the two linear models for d_c. The first is the same model as in table 5.5. The second model equals

$$d_c = 4.24 + 0.14 \cdot d - 1.58 \cdot ec + 0.12 \cdot CI_6 \quad (R^2 = 0.63) \tag{5.4}$$

It adds information on competition (CI_6) and on crown shape (ec) to the first model. This model is significantly better than the first model. It predicts larger crowns in trees with

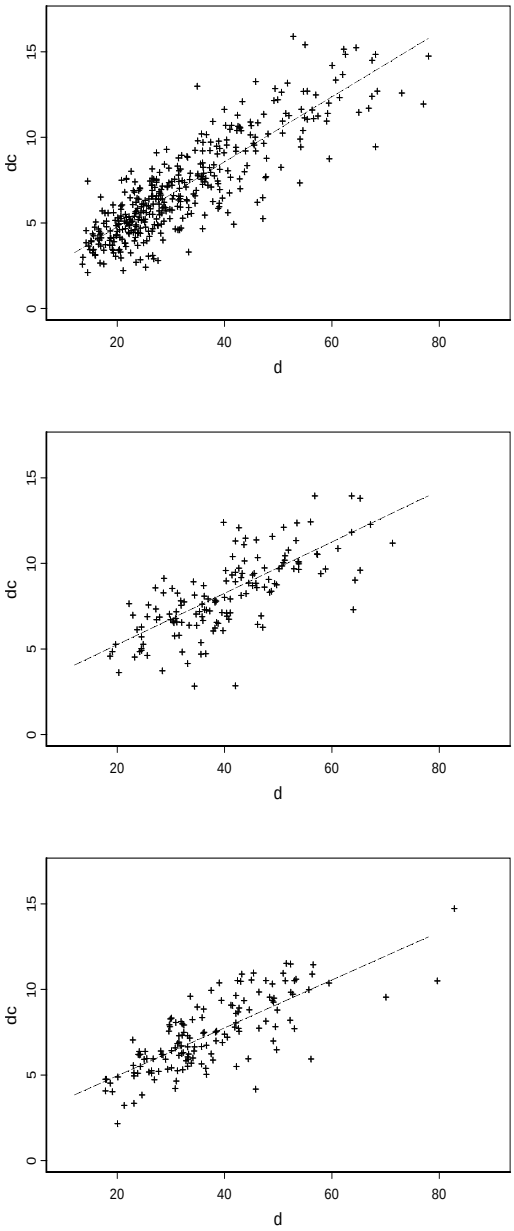

Figure 5.7: Relationship between diameter at breast height (d) and crown diameter (d_c), for the three plots M_I (top), $M_{II,A}$ (middle) and $M_{II,B}$ (bottom). The fitted regression lines are those specified in Table 5.5.

a larger distance to their nearest neighbour. The predicted values and 0.95 confidence intervals for d_c can be found in Table 5.7. Here we used the approximately minimum, median and maximum observed values of the explanatory variables. The table shows that for the same d and ec we expect more than 1 m increase in d_c if the distance to the nearest neighbour increases from the minimum (2 m) to the maximum (14 m) observed value in $M_{II,A}$. This model also predicts more eccentric crowns to be smaller in size (negative sign in the relationship). Very eccentric crowns ($ec = 2$) are expected to be 1.5 m smaller in diameter than non-eccentric crowns ($ec = 1$).

Table 5.6: Regression models for d_c according to a stepwise regression procedure, for $M_{II,A}$.

Variable	Coeficient	Std. Error	t value	$Pr(>\|t\|)$	R^2
intercept	2.25	0.49	4.60	0.00	0.53
d	0.15	0.01	12.51	0.00	
intercept	4.24	0.82	5.16	0.00	0.63
d	0.14	0.01	12.88	0.00	
ec	-1.58	0.33	-4.79	0.00	
CI_6	0.12	0.05	2.30	0.02	

Table 5.7: Confidence intervals (95%) for $d_c = f(d, ec, CI_6)$, for $M_{II,A}$.

		Confidence interval for d_c		
ec	CI_6	d=18	d=40	d=70
1	2	(4.66, 6.02)	(7.78, 8.87)	(11.66, 13.15)
1	6	(5.28, 6.35)	(8.47, 9.14)	(12.29, 13.48)
1	14	(5.97, 7.57)	(9.09, 10.42)	(13.03, 14.65)
1.4	2	(4.14, 5.27)	(7.27, 8.11)	(11.10, 12.45)
1.4	6	(4.73, 5.63)	(7.97, 8.38)	(11.70, 12.80)
1.4	14	(5.32, 6.94)	(8.43, 9.82)	(12.36, 14.05)
2	2	(3.21, 4.30)	(6.33, 7.16)	(10.12, 11.52)
2	6	(3.73, 4.73)	(6.88, 7.57)	(10.65, 11.94)
2	14	(4.25, 6.11)	(7.33, 9.02)	(11.26, 13.24)

Analysis of plots M_I and $M_{II,B}$

Comparison of the crown shape parameters obtained for the two plots reveal similar values of ct and el to those found in $M_{II,A}$. In plot $M_{II,B}$ we find crowns slightly more eccentric than in plot $M_{II,A}$, $\overline{ec} = 1.7$, whereas plot M_I has less eccentric crowns ($\overline{ec} = 1.3$). We found higher correlation between ct and d_c in $M_{II,B}$ ($\hat{\rho} = 0.36$) and M_I ($\hat{\rho} = 0.41$) than those found in $M_{II,A}$. Correlations with d and h were low, as for $M_{II,A}$.

Correlations between the competition indices and tree size were higher in $M_{II,B}$ than in $M_{II,A}$, whereas M_I had similar correlations values as in $M_{II,A}$.

The model $d_c = b_0 + b_1 \cdot d$ for plot M_I has an R^2 of 0.73, whereas for plot $M_{II,B}$ R^2 is 0.58. The addition of variables CI_6 and ec improved the initial model in both plots. Results were $R^2 = 0.75$ for M_I and $R^2 = 0.64$ for $M_{II,B}$.

Data reduction

Six tests were performed in total, two tests each for compactness, elongation and eccentricity. The average ratios between the shapes obtained with the reduced data and the shapes obtained with the 8 measurements were equal to 0.94 for compactness, 1.05 for elongation and 1.2 for eccentricity. The Wilcoxon signed rank test rejected the null hypothesis for compactness and eccentricity, but not for elongation ($\alpha = 0.05$).

5.5 Discussion

Three parameters were used to analyse crowns in terms of their compactness, elongation and eccentricity, using crown radii measurements. Shape parameters are applied on images of objects. Crown radii were interpolated towards 120 points of a polygon using a linear interpolation procedure weighted by angular differences. Other interpolation methods might have been applied as well. The resulting shapes however were more realistic than if we had joined the measured radii for example with straight lines. Also, parametric functions such as splines and trigonometric linear regression functions force a particular shape to the crown. Functions that fit a larger variety of crown shapes need a larger number of crown measurements, and add random noise. These were therefore avoided. When the 8 crown measurements were reduced to 4, two of the three calculated shape parameters were significantly different from the previously obtained.

Trees under competition had more elongated, less round crowns than isolated trees. This agrees with findings of Brisson (2001), that in forests of sugar maple isolated trees have the most symmetrical crown, whereas trees under competition are more asymmetrical and display crowns more developed away from the main competitive pressure of neighbouring trees. We found correlations indicating that eccentric crowns tend to be smaller, whereas round crowns tend to be larger. But since we had a small number of isolated trees, we could not find a significant difference in the sizes of the two groups of trees.

Ledermann and Stage (2001) hypothesize that stand-average competition indices represent the underground situation, while distance-dependent indices represent the aboveground environment. We found large correlations between crown diameter and indices

involving the number and size of neighbours, also in distance-independent indices. Indices weighting the size of competitors with their distance to the subject tree have a confounding effect with the subject tree's size, and result in increased estimated correlation values. The random allocation of tree sizes to the observed tree locations yields correlation values under the independence of tree size and tree location. The correlations observed in the montados can be compared with percentage points in the simulated distribution. All correlations between d_c and the competition indices were significant at the $\alpha = 0.05$ level.

Larger competition effects are found in aggregated and random point patterns, for the same overall density, because inter-tree distances can be very small. Competition might be reduced by reducing tree density and by planting trees according to a regular pattern, since both result in larger minimum tree distances (Smith et al., 1997). Larger crowns for the same d are observed in the plot with a regular pattern, M_I, than in the plots with random or aggregated patterns ($M_{II,A}$ and $M_{II,B}$). A more extensive study should be performed to research the effect of point patterns on crown size, and to see if decreasing tree competition would increase cork production.

The linear model $d_c = b_0 + b_1 \cdot d$ fitted the data at least as well as the other tested models. A quadratic function or more complex functions might better explain variation in crown diameter for a different range of d_c and d. However, the chosen model is more appealing because of its simplicity and the good fit for the observed values of crown diameter.

This study should be envisaged as a preliminary analysis aiming at defining the methodologies to be used in data collection in the future, and for the characterization of the structure of adult montados to be used in the initialization module of the SUBER model (Tomé et al., 1999). The SUBER model is to provide the landowners with a forecast of the consequences of different silvicultural practices - thinnings, fertilisation, debarking levels, grazing, etc - in the future yield of the stands, based on spatial characteristics and tree size distribution of their stands.

5.6 Conclusions

In this study we explored relations for cork oaks. Competition indices accounting for the relative size of neighbouring trees were the most correlated to crown size. The crown of a cork oak has a different shape and size when it is under competition, in particular if it is close to larger trees. It is more elongated and eccentric, and less round. Ultimately this may have an effect on crown size, given by its mean diameter. A model for crown diameter was obtained using d, crown shape and distance to the nearest neighbour as explanatory variables. The resulting model explains 63% of the variation in crown size, and is an

improvement on the model currently used by the Portuguese National Forest Inventory. Increasing inter-tree distances and decreasing density is likely to result in larger trees. In particular, regular patterns help increase minimum inter-tree distances for a given density.

Acknowledgments

Research of the first author was sponsored by financial support from the Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE) through the program 'III Quadro Comunitário de Apoio' , to which we feel grateful. Data used in this research was collected under the project CORKASSESS (Projecto CE FAIR CT97 1438 CORKASSESS - Field assessment and modelling of cork production and quality). The authors are also grateful to APFC (Associação dos Produtores Florestais do Concelho de Coruche e Limítrofes) for facilitating the field work in the two montados.

References

Baddeley, A.J. and Gill, R.D. 1997. Kaplan-Meier estimators of distance distributions for spatial point processes, Ann. Statist. **25**: 263-292.

Baddeley, A.J., Kerscher, M., Schladitz ,K. and Scott, B.T. 2000. Estimating the $J-$function without edge correction, Statist. Neerlandica, **54**: 315-328.

Biging, G.S. and Gill, S.J. 1997. Stochastic models for conifer tree crown profiles. For. Sci. **43**: 25-34.

Brisson, J. 2001. Neighbourhood competition and crown asymmetry in *Acer saccharum*. Can. J. For. Res. **31**: 2151-2159.

Dawkins, H.C. 1963. Crown diameter: their relation to bole diameter in tropical forest trees. Commonwealth For. Rev. **42**: 318-333.

Dawkins, H.C. 1958. The management of natural tropical high-forest with special reference to Uganda. Institute paper. Imperial forestry institute. University of Oxford, Oxford.

DGF, 1990. Inventrio florestal do sobreiro. Direco Geral das Florestas, Estudos e Informao **300**.

Deleuze, C., Herve, J.C., Colin, F. and Ribeyrolles, L. 1996. Modelling crown shape of Picea abies: Spacing effects. Can. J. For. Res. **26**: 1957-1966.

Diggle, P.J. 1983. Statistical analysis of spatial point patterns. Academic Press, New York.

Ferreira, A., Lopes, F. and Pereira, H. 2000. Caractérisation de la croissance et de la qualité du liège dans une région de production. Ann. For. Sci. **57**: 187-193.

De Gier, A. 1989. Woody biomass for fuel: estimating the supply in natural woodlands and shrublands. PhD thesis, Albert-Ludwigs-University, Freiburg.

Gill, S.J., Biging, G.S. and Murphy, E.C. 2000. Modelling conifer tree crown radius and estimating canopy cover. For. Ecol. Manage. **126**: 405-416.

Glasbey, C.A. and Horgan, G.W. 1995. Image Analyis for the Biological Sciences. Wiley, Chichester.

Green, P.J. and Silverman, B.W. 1994. Nonparametric regression and generalized linear models. Chapman and Hall, London.

Ledermann, T. and Stage, A.R. 2001. Effects of competitor spacing in individual-tree indices of competition. Can. J. For. Res. **31**: 2143-2150.

MathSoft 1997. S-Plus 4 Guide to Statistics. MathSoft, Seattle.

The MathWorks 2000. Matlab function reference, Natick, MA, USA.

Moravie, M.A., Durand, M., and Houllier, F. 1999. Ecological meaning and predictive ability of social status, vigour and competition indices in a tropical rain forest (India). For. Ecol. Manage. **117**: 221-240.

Ottorini, J.M., Le Goff, N. and Cluzeau, C. 1996. Relationships between crown dimensions and stem development in Fraxinus excelsior. Can. J. For. Res. **26**: 394-401.

Smith, D.M., Larson, B.C. and Kelty, M.J. 1997. The practice of silviculture: applied forest ecology. Wiley, New York.

Tomé , M., Coelho, M., Pereira, H. and Lopes, F. 1999. A management oriented growth and yield model for cork oak stands in Portugal. *In* A. Amaro e M. Tomé (edit.), Empirical and Process-Based Models for Forest Tree and Stand Growth Simulation, Edições Salamandra, Lisboa, Portugal, pp. 271 -289.

Van Lieshout, M.N.M. and Baddeley, A.J. 1996. A nonparametric measure of spatial interaction in point patterns, Statist. Neerlandica **50**: 344-361.

Chapter 6

Comparison of three sampling methods in the management of cork oak stands

Maria João Paulo, Margarida Tomé and Albert Otten
Submitted to Forest Ecology and Management

In this study we compare three sampling methods to estimate several variables in cork oak stands. The first method is to sample circular plots with fixed area. In the second method we sample circular plots with fixed number of trees. The third method consists in sampling *zigzags* each consisting of trees close to fixed points in a pre-defined path. This latter method, commonly used by Portuguese farmers, lead in most situations to estimators with larger biases and standard errors than the other two methods.

6.1 Introduction

Cork oak is Portugal's second most important forest species. It occupies an area of 640 000 ha and is the second most exported forest product. Portugal contributes to approximately 52% of the world's cork supply.

Cork oaks are grown specifically for the production of cork in cork oak stands known as *montados*, silvopastoral systems where cork production is associated with cattle and sheep breeding and grazing. The cork oak grows in poor soils and adapts to difficult conditions, such as high temperatures and lack of rain for lengthy periods. They are often grown in areas threatened by desertification. Their economic value plays an important role in the ecological protection of large areas.

Extraction of cork takes place every 9 to 11 years in adult trees. Before extraction, farmers sample the montado to estimate the value of cork. This depends upon quantity and quality of cork. The quality of cork is defined by its thickness, the number and size of pores, and several other characteristics. Each cork segment is rated, based on visual assessment, into one of 7 quality classes, where class 1 is the best quality and class 7 (called *refugo*) is the worst. The yield estimate(s) helps farmers to set a price for their cork, and in the choice of management alternatives. A commonly used sampling procedure followed by farmers is to define a polygonal transept (*zigzag*, see Figure 6.1) with a convenient starting point and covering the whole montado, and to sample every tree that crosses the transept.

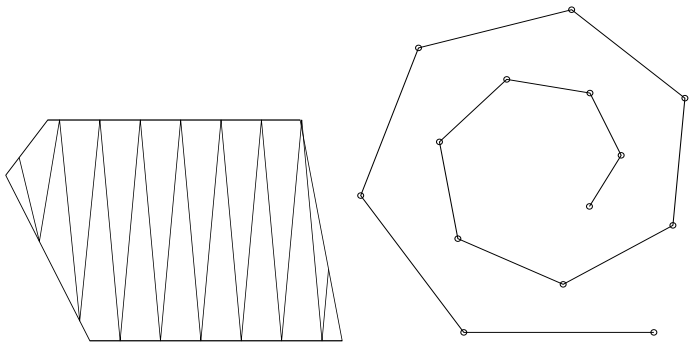

Figure 6.1: Example of zigzag sampling in montados (left) and in the circular plots (right).

In this study we compare the results of zigzag sampling with two other sampling methods - cluster sampling with fixed plot radius, cluster sampling with a fixed number of trees (and variable plot radius). Cluster sampling with fixed area is a widely used sampling method in extensive inventories to estimate stand variables such as tree density

and basal area. A number of plots are randomly selected, and all trees in each plot are measured. The trees in the stand have an equal probability of being selected and the usual estimator of the population total is unbiased, if boundary effects are negligible.

In their article from 1992, Jonsson et al. recommend the use of an alternative method for forest inventory, namely to measure a fixed number of trees nearest to the center of the plot. They claim the method is more efficient than the fixed circular plot size method, and is more accurate than other low-cost alternative methods. Furthermore, the authors show that for simulated forests the estimators they propose have a bias smaller than 10%, under the condition that the variables of interest are independent of the underlying point process.

The objective of this study is to extend the simulation to forests where the diameters at breast height are not spatially independent, and to extend the sampling methods to the three types mentioned above. Furthermore, we consider the estimation of some typical cork oak characteristics. The sampling methods are compared in terms of the bias and precision of the estimators and sampling costs.

6.2 Sampling in montados

In cluster sampling, a simple random sample of n primary units over an area is selected, followed by taking actual samples at a number of m_k, $k = 1, \ldots, n$ secondary units in each of the n primary units. In this study, the primary units are circular plots, and the secondary units are the trees in each plot. Although we actually measure the secondary units, it is the primary units that are selected. We assume throughout this text that the primary units are randomly sampled.

The trees can be sampled around each plot centre in two different ways: 1) sample all trees within a fixed distance r from each plot centre, or 2) sample a fixed number m of closest trees to each plot centre. Both methods can be described as cluster sampling, but in method 1 trees are selected into the sample with equal probabilities, and in method 2 they are selected with unequal probability. Table 6.1 lists the variables used in the remainder of this chapter.

Cluster Sampling with equal inclusion probabilities

For cluster sampling with equal selection probabilities we use a fixed radius r from the center of the circular plots. A sample taken with this method will be referred to as $\mathrm{R}_{k,r}$ with $k = 1, \ldots, n$. For each primary unit (plot) in the sample we can determine the area (A), the number of cork oaks, and for each tree, the tree diameter at breast height without cork (d), the cork quality (c_q) and thickness (c_t). These statistics can be

Table 6.1: List with variables.

Variable or index	Description
n	Sample size, number of primary units in the sample
k	Index for plots, $k = 1, \ldots, n$
m	Number of trees in plot
M	Total number of trees in a stand
r	Plot radius
R, R_r	Plot with fixed radius
T, T_m	Plot with fixed number of trees
Z	Zigzag plot
A	Area
i, j	Indices
x_1, x_2	Spatial coordinates
d	Diameter at breast height (1.3 m) without cork
ct	Cork thickness
c_{q}	Cork quality
h_{stem}	Stem height
h_{cork}	Maximum cork stripping height
N	Stand tree density (number of trees divided by area)
G	Stand basal area (total basal area in stand divided by area)
g	Tree basal area (without cork, at 1.3 m)
V_l	Stand cork volume in quality class l, divided by area
v_l	Tree cork volume in quality class l
CI	Competition index
Y	A population total divided by area
z	A standard normal deviate
μ	Mean value
σ^2	Variance
s^2	Sample variance
se	Standard error
Δ	Distance
L	Path length (b=between plots, w=within a plot)
c_1, c_2, c_3	costs
α, β, γ	Constants

combined to estimate the population density (N, number of trees divided by the area), basal area (G, the total cross-sectional area at breast height divided by the stand area, expressed here in m^2ha^{-1}), and cork volume in each quality class (V_l in class l divided by area). Cork volume was used instead of its weight, the latter being the usual quantity associated with cork value (price). Tree density (N) is estimated as $\hat{N} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{A}m_k$, where m_k is the number of cork oaks in plot k, and A is the (fixed) plot area. The estimator for the stand basal area is $\hat{G} = \frac{1}{n}\sum_{k=1}^{n}(\frac{1}{A}\sum_{i=1}^{m_k} g_{ki})$, where g_{ki} is the basal area of tree i in plot k: $g = \pi(d/2)^2$. For the cork volume in class l we use the estimator $\hat{V}_l = \frac{1}{n}\sum_{k=1}^{n}(\frac{1}{A}\sum_{i=1}^{m_k} v_{ki,l})$; since the cork sample from each tree is assigned one single quality class, the contribution $v_{ki,l}$ from one tree is either its whole cork volume v_{ki}, or zero.

Cluster Sampling with unequal inclusion probabilities

When a fixed number m of trees is sampled at each primary sampling unit each tree is associated with a different probability of being selected (unequal probability sampling). The Horvitz-Thompson approach to obtain the unbiased estimator for the population total is to divide the measurements performed on the observed trees by their inclusion probabilities (c.f.r. Thompson, 1992). In practice it is impossible to calculate these inclusion probabilities since they depend on the unobserved locations of all trees in the surrounding area, and even if the locations of all trees were known, the calculation of the inclusion probabilities would be cumbersome. This is because for a given tree, a joint area of overlapping convex simplices is needed. The calculations are far more complicated than those for the nearest neighbour Dirichlet cell (for the latter c.f.r. Ripley, 1981). For locally random patterns the probability of inclusion in a single plot can be approximated by m divided by the local tree density. The resulting estimators are approximately unbiased if the variable of interest takes (spatially) independent values for different trees.

Samples (plots) with m trees are denoted by $\mathrm{T}_{k,m}$, $k = 1, \ldots, n$. For the observed local tree density in plot k we take $(m-1)/A_k$, for the same reason as in Jonsson, i.e., in Poisson processes this is an unbiased estimator of the intensity. Then the estimator for the population tree density becomes $\hat{N} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{A_k}(m-1)$, where A_k is the plot area defined by the distance to tree m. The estimators for the stand basal area and for cork volume are $\hat{G} = \frac{1}{n}\sum_{k=1}^{n}(\frac{m-1}{m}\frac{1}{A_k}\sum_{i=1}^{m} g_{ki})$, and $\hat{V}_l = \frac{1}{n}\sum_{k=1}^{n}(\frac{m-1}{m}\frac{1}{A_k}\sum_{i=1}^{m} v_{ki,l})$, with g and v_l defined as before.

Zigzag sampling

The sampling method followed by some farmers consists in defining a zigzag transept covering the whole montado, and sampling every tree that crosses that path. Here we

adapt the method so that it can be applied to smaller areas. Starting from a randomly selected point, we define a single path (primary unit) with a constant number of vertices covering the whole plot, and at each vertex we sample the closest tree (secondary units). The total area A corresponding to the sample unit is determined by the sum of the areas of the circles around each vertex in the path, with radius given by the distance to the closest tree. Some trees and area parts could be counted more than once here. Sample size is set at n zigzag paths by choosing n starting points in the stand area. We name these plots Z. The estimators for tree density, basal area and cork volume are then defined as in T plots, with m equal to the total number of trees measured in each zigzag.

Assumptions in sampling primary units

In many practical situations the proportion of total sampled area (or sampled number of trees) is very small. In that case sampling of plots can be considered as sampling with replacement. Here we consider only this situation. Thus no attempt is made here to construct sampling of primary units without overlap, and the variance formulae for sampling with replacement are assumed to be satisfactory. Let n be the number of primary units in the sample, and y a variable of interest. Then the variance of the above estimators using cluster sampling (for both methods) and zigzag sampling can be estimated with $\hat{\text{var}}(\hat{Y}) = s^2/n$. Here $\hat{Y} = \frac{1}{n}\sum_{k=1}^{n}\hat{Y}_k$, and $\hat{Y}_k$ are the estimators for the primary units, and s^2 is the sample variance of the $\hat{Y}_k$, $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(\hat{Y}_k - \hat{Y})^2$. In our simulation experiment we do not sample many plots in one simulated stand. Instead, we simulate the same type of stand repeatedly and sample only one plot per stand. The variability between plots and hence s^2 obtained from these plots reflects then the variability in large scale stands. By large scale we mean that dependencies of characteristics of different plots in the same stand become negligible at plots distances which are still small compared to the stand size. For the simulated stand types we estimate the standard error at an actual sample size of n plots from one large stand as $s/\sqrt{n}$. The value of n will be chosen to meet certain requirements, such as fixed costs.

6.3 Simulated stands

To test the three sampling methods we simulated cork oak stands using information obtained in the analysis of a $200 \times 200\,\text{m}^2$ plot in Herdade do Vale Mouro (M_I). The spatial characteristics of M_I were analysed in Paulo et al. (2002). In M_I the trees had a regular point pattern and tree density was equal to $95\,\text{ha}^{-1}$. The diameter at breast height, d, was not randomly distributed with respect to tree positions; a negative correlation was present between the sizes of neighbouring trees. For c_q or c_t we found no evidence against

complete randomness.

Since the performance of the estimators derived from a fixed number of sampled trees is likely to depend on the spatial tree distribution as well as the spatial distribution of the measured variables we simulated stands from different underlying point processes, and different degrees of spatial correlation for d. Tree coordinates were generated either as random patterns, clustered or regular. Tree density was set to 100 ha^{-1} to be similar to the tree density observed in M_I. Further, diameters were generated according to the marginal distribution of d found in M_I, which was approximately a shifted lognormal with parameters $\mu = 3.3$ and $\sigma = 0.4$, and $d_{min} = 3$ cm (observed mean is 32 cm and observed standard deviation is 13 cm). The marginal distributons observed for c_q and c_t in M_I were used similarly. For c_t this distribution was approximately normal with mean $\mu = 31$ mm and standard deviation $\sigma = 7.5$ mm. In M_I the cork samples were classified into classes 3 to 7. The probabilities for c_q used in the simulations were derived from observed frequencies of cork in each quality class in M_I. The probabilities are 0.04, 0.12, 0.28, 0.34, and 0.22 for classes $c_q = 3$, 4, 5, 6 and 7, respectively. Generated d values were either randomly assigned to trees, or according to a penalty function in order to keep trees at some distance from neighbouring trees, this distance increasing with tree size. We call such a distribution of d values, among tree locations, regular, because large trees tend to exhibit a regular point pattern. Cork characteristics c_q and c_t were always assigned independently of tree locations and diameters.

Toroidal edge correction was performed to reduce edge effects. With this method the study area is regarded as a torus, so that points on opposite edges are considered to be close (Ripley, 1981). In a rectangular area of size P_1 by P_2 the distance between tree i, with coordinates $(x_1^{(i)}, x_2^{(i)})$, and tree j, with coordinates $(x_1^{(j)}, x_2^{(j)})$, becomes $\Delta_{ij} = \sqrt{(\Delta x_1)^2 + (\Delta x_2)^2}$, with $\Delta x_1 = \min\left\{|x_1^{(i)} - x_1^{(j)}|, P_1 - |x_1^{(i)} - x_1^{(j)}|\right\}$ and similarly for Δx_2.

In each simulated stand circular plots were sampled from the stand centre according to the three methods. R plots were sampled with fixed radius $r \in \{20, 25, 30, 35, 40\}$, for T plots we used a fixed number of trees $m \in \{13, 20, 28, 38, 50\}$. The m values correspond to the expected number of trees of the R plots (for the fixed tree density of 100 ha^{-1}). Z samples of trees in each generated stand were obtained for a fixed number of trees ($m = 14$) by creating a spiral transept, with a maximum radius of 56 m, where the closest tree to each vertex is sampled, the distance between vertices increasing with their distance to the center of the stand. The path was pre-defined and equal for every simulated stand.

Generation of tree location and size distribution

Six types of stands were simulated (see Table 6.2). For types I-V 250 stands with a size

Table 6.2: Simulated stand types.

Type	Point pattern spatial distribution	Diameter (d)
I	random	random
II	moderately regular	regular
III	very regular	very regular
IV	regular	random
V	random	regular
VI	clustered	regular

of $160 \times 160\,\mathrm{m}^2$ were generated independently, with a fixed number of trees, $M = 256$, corresponding to a tree density of 100 ha^{-1}. For type VI 250 stands were generated with a fixed number of trees and with a larger area that was posteriorly reduced by removing the stand borders. Tree coordinates (x_1, x_2), and diameters d were generated for each tree. The simulation details for each type of stand are as follows:

I. Tree positions were generated from a uniform distribution, and for all trees diameters d were generated independently of tree positions, $d_i = d_{\min} + \mathrm{e}^{\mu+\sigma z_i}$, $i = 1, 2, \ldots, M$, where the z_i are i.i.d. standard normal deviates.

II. For this stand type the joint distribution of positions and diameters was based on the minimization of a competition index depending on diameter and distance of tree pairs, inspired in the Metropolis-Hastings algorithm with a (Gibbs-type) penalty function:

   (a) Generate coordinates $x_1^{(0)}$ and $x_2^{(0)}$ from a uniform distribution for M trees in a fixed size area.

   (b) Generate tree diameters as for stand type I.

   (c) For each tree i in turn, $i = 1, 2, \ldots, M$, generate new candidate coordinates $x_1^{(1)}$ and $x_2^{(1)}$, keeping d. Calculate index $CI = \alpha \sum_{j=1, j\neq i}^{M} d_i d_j / \Delta_{ij}^2$, both at the original location of i, $(x_1^{(0)}, x_2^{(0)})$, and at the new location for i, $(x_1^{(1)}, x_2^{(1)})$. If $CI^{(1)} < CI^{(0)}$ then accept the new location for tree i. Otherwise accept the new location with probability $p = \mathrm{e}^{-[CI^{(1)} - CI^{(0)}]}$.

   (d) repeat step (c) 1000 times.

   Probability p of accepting the new location is partly determined by α, a constant controlling the scale of CI.

   After many cycles through step (c), the pattern obtained resembles the pattern observed in $\mathrm{M_I}$: the spatial tree pattern and nearest-neighbour distance distributions

are quite similar. The similarity of nearest-neighbour distributions was judged with QQ-plots. The d values displayed a similar correlation value with CI as the observed in M_I. The resulting point pattern is more regular than in I, and so is the combination of the tree locations (point pattern) and the diameters (marks of the point pattern).

III. Stands as in II, but with a more regular pattern were simulated, by increasing the value of α (and so increasing the rejection probability of the new point). The resulting pattern tends to be more regular than in II, both for the tree pattern and for the combination of tree locations (point patterns) and diameters (marks of the point patterns).

IV. Stands as in II were produced by using $CI = \beta \sum_{j=1, j\neq i}^{m} 1/\Delta_{ij}$ which does not depend on the diameters. Here β is again a constant determining the rejection probability of the new tree location. In this case the generated trees tend to be regularly spaced, but tree diameter is independent of the diameters of other trees.

V. The following procedure was used to simulate stands with a random point pattern for tree locations and a regular diameter distribution:

   (a) Generate tree coordinates x_1 and x_2, from a uniform distribution.

   (b) Generate tree diameters as for stand type I.

   (c) Perform 200 random permutations of d. Each permutation corresponds to one assignment of the d values to trees in the stand.

   (d) For each permutation calculate $CI = \sum_{i=1}^{m} \sum_{j=1}^{m} d_i d_j / \Delta_{ij}^2$.

   (e) Choose the permutation with minimum CI.

VI. Clustered patterns of trees, with a regular diameter distribution. The point pattern was obtained by generating a parent process uniformly distributed over an area of 500×500 m^2, and then generating children clustered around each parent. Each cluster had 20 children and a 60 m radius. The number of parents was fixed to 97, to obtain a tree density of 100 ha^{-1} as before. To avoid an edge effect we discarded the stand border (of a width equal to the cluster radius) after which the point density at the inner area was in average 100 ha^{-1} as wished. To obtain a regular distribution of diameters over tree positions we proceeded as in (V) for the inner area of the generated stand.

6.4 Data

We applied the three sampling methods on available data from two cork oak montados, one in Herdade da Machoqueira do Grou (HG), and the other in Herdade do Vale Mouro

(M_I). Both farms are located in Central Portugal, close to the village of Coruche. In montado M_I one plot of a $200 \times 200 m^2$ size was measured in July 1998, shortly after cork extraction. Measured variables were coordinates of tree location, d, h_{stem}, h_{cork}, c_t and c_q. It is a very homogeneous plot, but small in size to test the three sampling methods. To overcome this, we can regard the total sampling area as a grid of rectangles, all identical to the measured plot. The initial plot forms a border with mirrored copies of itself, thus extending the total area available. Sample R plots with a radius $r \in 20, 30, 40$ m, T plots with $m \in 12, 27, 48$ (corresponding to a tree density of 95 ha^{-1}) and Z plots are defined in this extended area. This is equivalent to assuming a toroidal surface for M_I. Figure 6.2 shows the plot in M_I, with an example of 14 circular sample plots. Sampling was done repeatedly (250 times each plot, with replacement) in order to obtain precise estimates of biases and standard deviations.

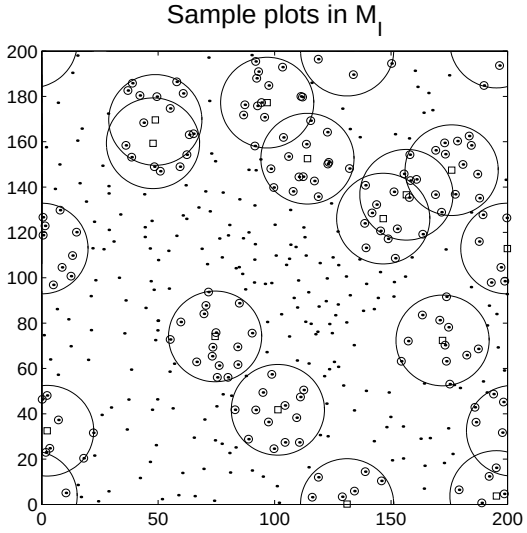

Figure 6.2: Circular sample plots in M_I. Trees are marked as points, sampled trees have small circles around them, plot centres are marked with squares. A toroidal-type of surface was assumed, hence the high frequency of overlapping plots.

Montado HG is one of several management units with cork oaks in Herdade da Machoqueira do Grou. HG has 308 ha in total and is a mixed stand with cork oak, Pinus pinea and occasionally also Pinus pinaster. It is very heterogeneous in terms of tree density. The sampling took place after cork extraction, in 1998. The stand was divided in 7 distinct strata according to species composition and tree density, the year of harrowing, and c_q. Two random circular plots with a radius greater or equal than 40 m were sampled from each stratum. Table 6.3 summarizes the plots (restricted to r=40 m) within

the seven strata in the montado.

Table 6.3: Main characteristics of the plots in HG.

Stratum	Year of harrowing	cork oak quantity	size (ha)	Plot number	Cork oaks in plot[1]	cork oak density (ha^{-1})
1	93/94	high	47.4	1	27	54
				2	79	157
2	93/94	high	28.1	3	15	30
				4	35	70
3	93/94	medium	48.5	5	21	42
				6	40	80
4	94/95	medium	48.7	7	40	80
				8	28	56
5	94/95	high	78.6	9	26	52
				10	31	62
6	94/95	medium	34.7	11	35	70
				12	32	64
7	94/95	low	21.3	13	17	34
				14	21	42

[1] Circular plots with 40 m radii.

Every cork oak inside each plot was measured for tree coordinates, d, h_{stem} and h_{cork}. A smaller number of trees was also sampled for c_{t} and c_{q}. As in M_{I}, R plots, T plots with $m \in 8, 19, 33$ (corresponding to a tree density of 65 ha^{-1}) and Z plots were posteriorly re-sampled from the measured plots in HG.

Stratified sampling

To obtain the final estimate of a population characteristic Y for HG we weighted each estimate, $\hat{Y}_k$ from plot k, with the area fraction of the corresponding stratum (h). If we have a total number q of strata in the montado, then the weight corresponding to each stratum h is $w_h = A_h / \sum_{h=1}^{q} A_h$. Clearly $\sum_{h=1}^{q} w_h = 1$, and thus $\hat{Y} = \sum_{h=1}^{q} w_h \bar{Y}_h$ (with $\bar{Y}_h$ the mean of the $\hat{Y}_k$'s in stratum h). Similarly, we have $\text{var}(\hat{Y}) = \sum_{h=1}^{q} w_h^2 \sigma_h^2 / n_h$, with n_h as the number of plots in stratum h. Since in the case of HG we have $n_h = 2$, $h = 1, \ldots, 7$, the σ_h^2 become very imprecise to estimate. We therefore simplify offering unbiasedness by estimating one σ^2 for all strata, and thus in HG the estimated $\text{var}(\hat{Y})$ becomes $\hat{\text{var}}(\hat{Y}) = \sum_{h=1}^{7} w_h^2 s^2 / 2$, where s^2 is pooled over strata.

6.5 Model for cork volume

In the simulated stands cork volume of a tree was generated from d and c_t (generated) values, because cork quantity is correlated with tree diameter. First a model was obtained for plot M_I, and then used to generate volumes in the simulated stands. In M_I we had direct measurements of d and c_t, and accurate estimates of cork volume per tree, obtained from c_t and the measured debarked surface area of each tree. The model obtained in the ln-scale for M_I was

$$v/\mathrm{m}^3 = \mathrm{e}^{\gamma_1}(d/\mathrm{cm})^{\gamma_2}(c_t/\mathrm{mm})^{\gamma_3} \tag{6.1}$$

Here $\gamma_1 = -13.97$, $\gamma_2 = 2.20$ and $\gamma_3 = 1.09$. An $R^2 = 0.84$ and $\hat{\sigma}^2 = 0.134$ were obtained for this model (in the ln-scale). Therefore this expression was used to generate cork volume from d and c_t in the simulated stands (in M_I the volumes have been obtained directly from tree measurements).

Since measurements of cork quantity (volume or weight) are very difficult to obtain, most of the time these have to be estimated from other tree measurements. In HG there were no measurements of cork quantity (volume or debarked area) for individual trees. Therefore, we cannot calibrate model (6.1) to use in HG. The measurements performed in HG include stripping height (maximum height of the stem which was debarked) and height of fork (height at which the stem divides in two or more main branches). The latter should be taken into account because branched trees have a larger surface and thus produce more cork. An alternative model to estimate individual volumes was used instead of (6.1):

$$v = \pi c_t d(h_1 + \sqrt{2}h_2) \tag{6.2}$$

where $h_1 = \min(h_{\mathrm{stem}}, h_{\mathrm{cork}})$ and $h_2 = \max(0, h_{\mathrm{cork}} - h_{\mathrm{stem}})$. This expression derives from the fact that the diameter of two branches above the fork (d_1 and d_2) relate to the diameter below the fork (d_0) as $d_0 \approx \sqrt{d_1^2 + d_2^2}$ (the stem volume is more or less equally distributed by the 2 branches). If we further assume that $d_1 \approx d_2$ then $d_1 + d_2 \approx \sqrt{2}d_0$, which can be used when d_1 and d_2 are unknown. In M_I, where volume measurements are available, model (6.1) fitted the data better than model (6.2).

6.6 Comparison of the methods

From here on we refer to the estimated single plot standard deviation as s, and to the estimated standard error ($s/\sqrt{n}$) of an estimator as se. Means and standard deviations were calculated for $\hat{Y}_k$, $k = 1, 2, \ldots, 250$, for all characteristics, stand types and sample plot sizes considered. We compared the bias and s of the plots obtained with the three

sampling methods from all simulated stand types. Bias in T-estimators was estimated through paired comparison with the R-estimators for similar plot size - the R-estimators being unbiased, and highly correlated with the T-estimators. The standard error of the estimated bias was also calculated. Usually the F-test is used to compare standard deviations of two independent samples (normality assumed). But since the samples obtained from the simulated stands with the three different methods are in principle dependent, the F-test is slightly conservative, so the Pitman test was also used. The test uses the fact that for two variables Y_1 and Y_2 (dependent or not), $\mathrm{cov}(Y_1+Y_2, Y_1-Y_2) = \sigma_1^2 - \sigma_2^2$, so that under $H_0 : \sigma_2^2 = \sigma_1^2$ we have $\rho(Y_1 + Y_2, Y_1 - Y_2) = 0$ for the correlation. Since $t = \hat{\rho}\sqrt{n-2}/\sqrt{1-\hat{\rho}^2} \overset{H_0}{\sim} t_{n-2}$ (normality assumed), for the two-sided alternative hypothesis we reject H_0 if $|t| \geq t_{\alpha/2,n-2}$. The standard error is a usual measure to describe the precision of estimators, and is associated with sample size. The standard deviations can be compared immediately when procedures are used at the same sample size of n plots. To present a fair comparison of the three methods, we also compare them at the same fixed total costs, thus at different n values. Then in the (estimated) standard error $se = s/\sqrt{n}$ both s and n are varying with the method.

Determination of sampling costs

To define the cost of sampling, we divide the costs into two components: travelling costs corresponding to moving inside and between plots, and costs of measuring the variables of interest for each tree. All costs are expressed as time units. To determine the time of sampling in cork oak stands under the usual conditions we will assume that distances between plots R or plots T are covered by car, at a speed of 4 km·h^{-1}. This low speed is chosen to reflect the poor accessibility in montados. We also assume that the distances inside each plot are covered by foot, at a speed of 2 km·h^{-1}. We consider a further 3 minutes to measure each tree. To calculate the travelled path length inside the simulated R and T plots, L_w, we assume the forester follows a number of ring shaped paths inside the circular plot until he has measured all trees in the plot. The minimum path length, for a different number of equal width rings, is then used. In Poisson forests the within plot path length is approximately proportional to m, the number of trees in the plot. In Z plots the total path length between nodes can be calculated precisely. We further add to this the distance travelled by the forester from each node to the sampled tree (and back). In HG the path length travelled inside each plot can be easily determined because the trees were sequentially numbered. To estimate the average distance between visited plots, L_b, we assume that the plots are randomly located in a stand with fixed area. Then with a reasonable travelling strategy, L_b depends on area and sample size, approximately as $L_b = \sqrt{A_{\mathrm{stand}}/n}$. In HG $A_{\mathrm{stand}} = 308$ ha, and for $\mathrm{M_I}$ and for the simulated stands we

consider $A_{\text{stand}} = 370$ ha, as found in another montado. After the simulations, average values of L_w (all plots) and m (R-plots) are known. The total time needed for a sample size of n plots takes the form

$$t_{\text{Tot}} = c_1\sqrt{A_{\text{stand}}}\sqrt{n} + n(c_2 L_w + c_3 m) \tag{6.3}$$

with $c_1 = 1/4$ h·km^{-1}, $c_2 = 1/2$ h·km^{-1}, and $c_3 = 0.05$ h. We note that time is not proportional to sample size. Since a fixed total sampling time is considered for each method, the corresponding number of plots, not rounded, can be found by solving equation (6.3).

6.7 Results for the simulated stands

Standard deviation and standard error

In the following text we refer to the standard deviation of R, T and Z plots respectively as s_1, s_2 and s_3. The R and T estimators obtained for samples with similar plot size were highly correlated, with correlations between 0.85 and 0.95, for all variables. Z and R estimators, on the other hand, had correlations often lower than 0.1. In the comparison of R and T samples the Pitman test was therefore used. In the case of the dependent samples, relative differences of less than 6% between s values were already significant. Standard deviations for R and T plots were very similar, with s_2 up to 6% larger than s_1, except in the exceptional situation of volume estimation in the least frequent quality classes, for the smallest T plots. We observed that s_1 and s_2 were the highest for the clustered stands, and minimal for the most regular point patterns. As expected, they decreased with plot size, as Figure 6.3 shows. In general however, the ratio s_2/s_1 did not change with plot size, nor with point pattern nor with diameter distribution. In the comparison between R and Z samples, relative differences between s values were significant when greater or equal to 10%. Standard deviation s_3 was larger than s_1 for more than 10% in most cases, with a larger difference in regular patterns and for regular d distributions. The ratio s_3/s_1 was maximal (2.50) in the regular (type III) stands, and minimal in the clustered stands with value approximately 0.80. The standard errors were compared for the different sampling methods, and for three different forest types, for a variable number of plots, corresponding to a constant time. The sampling time was set to 30 hours, and a stand of 370 ha was assumed. Sample size (not rounded, see Table 6.4) was determined as described in the previous section. The resulting standard errors are also displayed in Figure 6.4 for variables N and G, and for the most extreme types of stand considered.

These figures show that, for the given sampling time, the standard errors of R and T estimators are approximately the same for random and regular point patterns, regardless

Table 6.4: Results from simulated stands: standard error of estimators for the different methods with a fixed cost of 30 hours, and based on a stand area of 370 ha.

Stand type I Method	n	E(m)	Total length within plot	average distance between plots	St. dev. $\hat{N}$	St. dev. $\hat{G}$	St. error $\hat{N}$	St. error $\hat{G}$
R_{20}	38.3	12.9	120	310	27.3	3.7	4.4	0.60
R_{30}	18.0	28.4	270	450	18.0	2.3	4.2	0.54
R_{40}	10.3	50.3	510	600	12.8	1.8	4.0	0.57
T_{13}	37.5	13	150	310	29.0	3.8	4.7	0.62
T_{28}	18.0	28	300	450	19.0	2.4	4.5	0.57
T_{50}	10.3	50	520	600	13.4	1.8	4.2	0.57
Z	27.4	14	600	370	30.9	4.3	5.9	0.83
Stand type III Method	n	E(m)	Total length within plot	average distance between plots	St. dev. $\hat{N}$	St. dev. $\hat{G}$	St. error $\hat{N}$	St. error $\hat{G}$
R_{20}	38.8	12.6	130	310	18.4	3.0	2.3	0.40
R_{30}	17.9	28.4	280	460	10.6	1.8	2.0	0.38
R_{40}	10.2	50.3	530	600	8.0	1.3	1.9	0.37
T_{13}	37.2	13	160	320	17.2	2.8	2.2	0.40
T_{28}	18.0	28	300	450	10.5	1.9	2.2	0.36
T_{50}	10.3	50	530	600	8.2	1.3	1.9	0.36
Z	27.7	14	590	370	35.6	5.3	7.1	1.09
Stand type VI Method	n	E(m)	Total length within plot	average distance between plots	St. dev. $\hat{N}$	St. dev. $\hat{G}$	St. error $\hat{N}$	St. error $\hat{G}$
R_{20}	39.9	12.4	110	310	51.9	5.4	8.2	0.86
R_{30}	18.4	27.8	260	450	42.3	4.1	9.9	0.95
R_{40}	10.4	49.6	500	600	37.5	3.7	11.6	1.15
T_{13}	37.2	13	160	320	53.8	5.8	8.8	0.94
T_{28}	18.0	28	310	450	43.3	4.3	10.2	1.02
T_{50}	10.3	50	530	600	37.0	3.5	11.5	1.10
Z	27.1	14	630	370	39.5	4.5	7.6	0.87

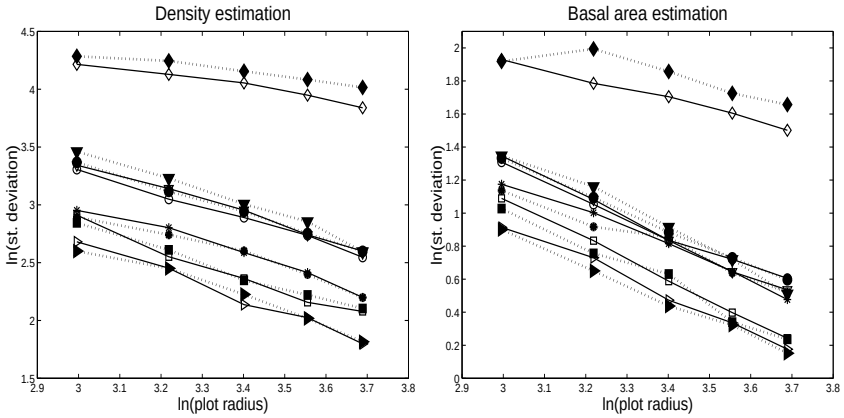

Figure 6.3: $\ln(s(\hat{N}))$ vs $\ln(r)$ (left) and $\ln(s(\hat{G}))$ vs $\ln(r)$ (right), obtained from simulations for R_r plots (solid lines) and for T_m plots (dotted). Different stand types are shown with different symbols: type I ($\circ$); type II ($\square$); type III ($\triangleright$); type IV ($*$); type V ($\triangledown$); type VI ($\diamond$).

of plot size. For these patterns Z plots yield comparatively too large standard errors. For clustered patterns Z plots and smallest R and T plots have the lowest standard errors.

Bias

Standard error of the bias estimates in T estimators, obtained by paired comparison with the R estimators, was always lower than 6% of the $\hat{Y}$ value. The estimated precision of bias estimates was therefore satisfactorily low. In the T plots no significant bias was found for the N and G estimators, for the simulated stand types. For the V estimators, a bias of up to 15% was observed, in the smaller plots. The bias was negligible in the larger plots. The magnitude of the bias was not noticeably affected by the point pattern or d distribution.

Bias estimates of Z estimators had an estimated precision of less than 20% of μ. A negligible bias was found for random patterns (stand types I and V). In the clustered patterns (stand type VI) biases were present of a magnitude up to 15% for all estimators. In regular patterns (stand types II, III and IV) bias was highest: between 30% and 50% in moderately regular patterns (II), between 20% and 30% in regular patterns with a random d distribution (IV), and between 45% and 80% in the very regular patterns (III). Biases were positive in sign for the regular patterns and negative for the clustered patterns. The bias in Z plots was affected both by the point pattern and by the combined distribution of d and tree locations.

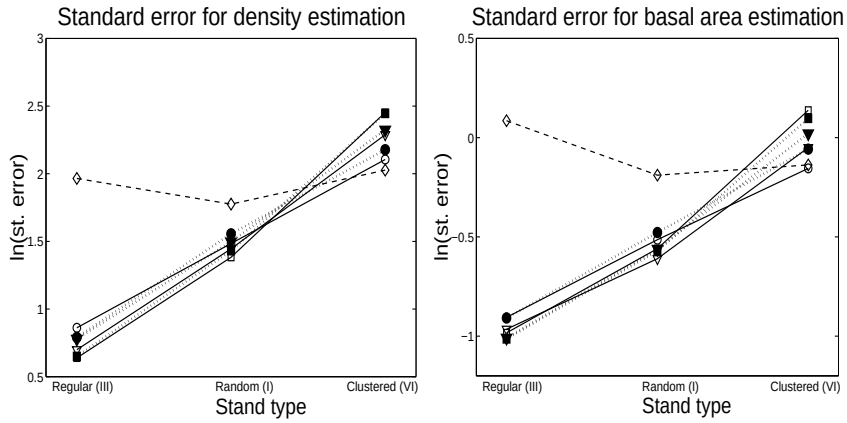

Figure 6.4: $\ln(se(\hat{N}))$ and $\ln(se(\hat{G}))$ obtained from simulations for three different point patterns (horizontal axis). Results are shown for different sampling methods, R plots are shown with solid lines, T plots are shown with dotted lines and black filled symbols: R_{20} and T_{13} ($\circ$), R_{30} and T_{28} ($\triangledown$), R_{40} and T_{50} ($\square$), and Z ($\diamond$). The number of plots in all situations was chosen for fixed costs, equal to 30 hours.

6.8 Results for the montados

The standard errors of the estimators obtained in HG and in M_I, for constant times of 30 hours and based on a stand area of 308 ha and 370 ha respectively, are displayed in Figure 6.5. In M_I estimators of N and G have lower standard errors in smaller R and T plots, and largest standard errors in Z plots. On the other hand V_3 and V_4 (with small frequencies), have lowest standard errors in the largest plots. In HG standard deviations were estimated from 14 plots, thus $\hat{se}$ provides a very crude impression. In HG the lowest standard errors for N and for G are obtained with the Z plots. The Z plots were not used to estimate cork volume for this data set because the number of cork samples in these plots was too low.

Since the values of the variables under study were not completely known in HG, bias could not be evaluated for this data set. In the M_I data set larger bias was observed in T_{12} estimators, about 15% for the N and G estimators, and 20% for the V_3 estimators, which had a low observed frequency. For Z plots yet larger bias was observed (up to 65% for the N and G estimators, and 30% for the V estimators).

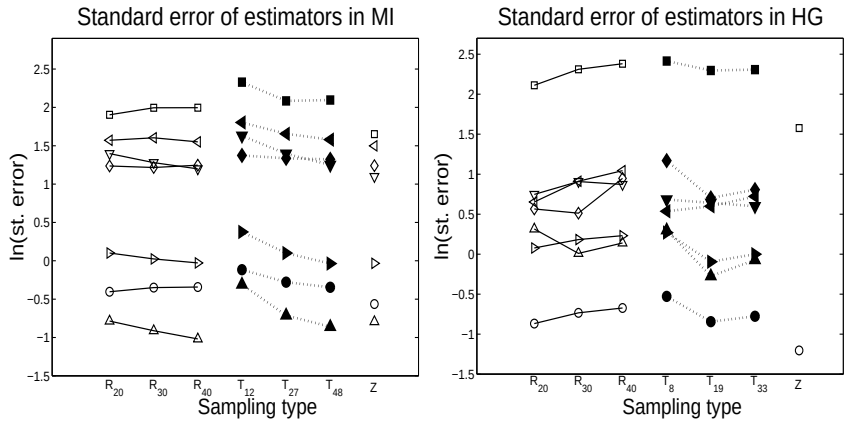

Figure 6.5: $\ln(se)$ obtained in data-sets M_I (left) and HG (right), by each method (horizontal axis). The estimators are for mean tree density (□,ha^{-1}), mean basal area (○,m^2ha^{-1}), and mean volume (△, ▷, ▽, ◁, and ◇, by increasing order of quality, $0.1m^3$). The number of plots in all situations was chosen for fixed costs, equal to 30 hours, and for a stand area of 370 ha in M_I and 308 ha in HG. Standard deviations were estimated from a sample of 250 (re-sampled) plots in M_I and a sample of 14 plots in HG.

6.9 Discussion

The results in this study agree with those in Jonsson et al (1992). In their article they report biases smaller than 10% in T plots sampled from simulated forests. The main purpose of this study was to compare the sampling methods for a fair number of different point patterns and spatial distributions of diameters. In order to obtain good estimates for the standard deviations and biases of the estimators, we focused our analysis on simulated stands. Underlying stationary isotropic processes were assumed. We did not try however to exhaust all patterns that are likely to be found in practice. In fact there may be montados where point pattern, tree density, and diameter size distribution vary considerably in space, but those more complex situations were not simulated. Data was used to illustrate the diversity of situations arising in reality: the plot in M_I illustrates a homogenous stand, whereas HG exemplifies the difficulties that may arise in stands with an inhomogeneous or non-stationary point pattern.

The total number of simulations (250) was limited by computer time, we think however that for practical purposes the resulting precision of the simulations is satisfactory. Further, stands were simulated with a fixed number of trees to facilitate the storage of the simulations in a matrix format, thus allowing very fast computations. This results in an (unintentional) loss of randomness. This loss can be neglected in the present situ-

ation because the sampling plots are small compared to the stand size. In fact the loss of variability can be illustrated in the case of the tree density estimator in type I stands. The standard deviation of the tree density in the R_{40} plots is expected to be 14.1 (from a Poisson count with parameter 50.3 for an area of 0.503 ha). The observed value is 12.8 (Table 6.4), corresponding to a variability loss of less than 10%, but in close agreement with the theoretical conditional value, which is 12.7. Since the simulated stands of type II had some similarities with M_I, the sampling results both for the simulations and data were compared. Lower variability and lower biases were found for the estimators obtained from sampling the simulated stands.

The comparison between standard errors obtained with different sampling methods depends of course on sample size. It seemed reasonable then to use variable sample sizes for each method, corresponding to a fixed amount of costs. The sampling times (used as sampling costs) used here were merely illustrative. Other costs could have been considered, such as extra time for setting R plots, or for taking cork samples. For example, the path length within plots (when walking from tree to tree) was calculated based on a hypothetical rule according to which the forester samples the trees following a path inside rings in the circular plots, which yield the values shown in Table 6.4. For trees in a regular square grid, the distance between trees in an optimal path would be $\sqrt{A/m}$, which is approximately equal to the distances calculated with our path in the simulations. For random tree positions the expected nearest neighbour distances are $0.5\sqrt{A/m}$. A path along nearest neighbours is nearly never possible. If the optimal path is used to visit trees (solution of the travelling salesman problem), then the expected average distance between trees in the path is asymptotically (for $A \to \infty$ and A/m fixed) converging to $0.7124\sqrt{A/m}$ (Jonhson and McGeoch, 1997), that is, about 70% of the distance obtained with our non-optimal path. For the travelling between plots in the stand we considered an average distance equal to $\sqrt{A/n}$. Here also a reduction of up to 70% could be obtained. If different costs were to be assumed then different sample sizes would have been obtained, with consequences for the standard errors of estimators and for the choice of the most cost-effective method.

6.10 Conclusions

Methods R and T produce very similar estimates for the types of forest considered in this study. The single plot standard deviations (s) obtained with the two methods differed very little, for equivalent plot sizes, and the biases observed in T estimators were in most cases lower than 10%. Method Z produced estimators with a large bias in all non-random point patterns. For the Z method s was considerably larger than the s obtained by the other two methods, except in clustered forests and in the HG data set, which is very

heterogeneous.

Although s decreases with an increase in plot size, this is not necessarily true for the standard errors (se) of estimators when the number of plots depends on a fixed amount of costs. For a fixed amount of costs, in regular and random patterns the standard errors were lowest for R and T estimators obtained with large plot sizes, and highest for Z estimators. In M_I (which has a moderately regular point pattern) unexpectedly standard errors were not always lowest for the large plot sizes. In clustered forests standard errors were lower for R and T estimators obtained with small plot sizes and lowest for Z estimators. This was also observed to some degree in the HG data set.

The choice between R and T can safely be based on practical convenience. The Z method is clearly disadvantageous since it produces estimators with large biases and in all but clustered patterns also large standard errors.

Acknowledgments

Research of the first author was sponsored by financial support from the Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE) through the program 'III Quadro Comunitário de Apoio' , to which we feel grateful. Data used in this research was collected under the project CORKASSESS (Project CE FAIR CT97 1438 CORKASSESS - Field assessment and modelling of cork production and quality). The authors are also grateful to APFC (Associação dos Produtores Florestais do Concelho de Coruche e Limítrofes) for facilitating the field work in the farms.

References

Cochran, W.G. 1977. Sampling Techniques. Third Edition. Wiley. New York.

Jonhson, D.S., and McGeoch, L.A. 1997. The travelling salesman problem: a case study. *In* Arts, E., Lenstra, J.K. (edit.), Local search in combinatorial optimisation. Wiley. Chichester.

Jonsson, B., Holm, S., Kallur, H. 1992. A forest inventory method based on density-adapted circular plot size. Scand. J. For. Res. 7:405-421.

Paulo, M. J., Stein, A., Tomé , M. 2002. A spatial statistical analysis of cork oak competition in two Portuguese silvopastoral systems. Accepted by the Canadian Journal of Forest Research in June 2002.

Ripley, B. D. 1981. Spatial statistics. Wiley. New York.

Thompson, S.K. 1992. Sampling. Wiley. New York.

Chapter 7

General conclusions

In this thesis it has been shown how the use of current mathematical statistical methods can help to improve the modelling and estimation of cork oak and eucalyptus stand and trees' characteristics. The improvement of information in stand management of cork oak and eucalyptus stands is valuable for decision making and may help to increase production.

Optimisation of long term volume yield in eucalyptus stands depends upon a prior distribution of the volume growth parameters, on the ages of measurement and on the error distribution of the volume observations. For the considered prior and error distribution, the long term volume yield significantly increases if individual optimised cutting time is used instead of a common optimised cutting time. The main gain is obtained from optimising fixed measurement times and individual cutting times. Only a small additional increase is obtained by optimising individually the second measurement time as well.

D-optimal designs are more economical and efficient for estimation of individual diameter growth of cork oaks. A replication-free compromise design, D-optimal for the average of the sample's growth parameters, performed better than the equidistant design. Since in practical situations the residuals obtained with fitting parametric curves to empirical data are often autocorrelated, this situation was also analysed. For an autocorrelation up to 0.6 a compromise design is still recommended, obtained by spacing with 10ρ years the replicate design points from the unrestricted design. For higher autocorrelation values the equidistant design is a better option.

Spatial relations between cork oak trees were explored in relation with competition indices based on size and distance of neighbouring trees. Those accounting for the relative size of neighbours were significantly correlated to crown size of subject trees. The crown of a cork oak differs in shape and size when it is under competition. A crown of a tree

close to larger trees is more elongated. This may have an effect on crown size in the end. Crown diameter was modelled using stem diameter, crown shape and distance to the nearest neighbour as explanatory variables. An increase in inter-tree distances and a decrease density is likely to result in larger trees. In particular, regular patterns help increase minimum inter-tree distances for a given density.

Three sampling methods were compared to estimate tree density, basal area and cork volume: cluster plots with fixed radius (method R), cluster plots with a fixed number of trees (method T), and sampling trees standing in a zigzag path (method Z). Methods R and T produced similar estimates. Bias of T estimators was negligible, and their standard errors were equivalent to those produced by R estimators. A choice between R and T can therefore be based on practical convenience. Method Z, often used in Portuguese cork oak farms, yields estimators with a larger bias and larger standard errors. Bias and standard error depended strongly on the spatial pattern of the trees and on independence. Largest bias and largest standard errors occurred for regular point patterns with a conditional size distribution. In clustered patterns the observed standard errors were smaller than those obtained by R and T estimators, but biases were larger than for T estimators.

Samenvatting

In dit proefschrift worden moderne wiskundig statistische methoden toegepast op problemen binnen hedendaagse Portugese bosbouw systemen. Hier bestaat behoefte om via betere beslissingen ten aanzien van het bosbeheer de productie te optimaliseren. Er wordt achtereenvolgens aandacht besteed aan het gebruik van Bayesiaanse methoden, groeikrommen, optimale proefopzetten, ruimtelijke analyse van patronen van kronen van kurkeiken en aan steekproefmethoden. Vier onafhankelijke vraagstellingen staan centraal aangaande opstanden van kurkeik en eucalyptus.

De eerste vraagstelling richt zich op optimalisering van de omlooptijd van eucalyptusbossen die dienen ten behoeve van pulpproductie. Opvolgende rotaties en meer in het bijzonder hun groeikrommen worden beschouwd als onafhankelijke realisaties van hetzelfde genererende proces. Het doel is om de lange-termijn volumeproduktie, gecorrigeerd voor kosten van herplanten, te optimaliseren. Op lange termijn is de totale financiële opbrengst gedeeld door de totale tijd een economisch belangrijk gegeven. Een Bayesiaanse aanpak is gevolgd, onder de veronderstelling van Shumacher groeikrommen en met gebruik making van prior informatie t.a.v. de parameters. Deze is gebaseerd op een groot aantal waargenomen groeikrommen. In het geval van bekende of adequaat geschatte groeikrommen was de winst aanzienlijk in het optimaliseren van individuele kaptijdstippen, vergeleken met het kiezen van één vast kaptijdstip. In deze studie is uitgegaan van de veronderstelling dat twee volumemetingen worden verricht ter ondersteuning van de keuze van het kaptijdstip van een rotatie. De eerste meting op een vast tijdstip, tweede echter op een tijdstip dat afhankelijk is van het resultaat van de eerste meting. Een belangrijk probleem is het vinden van een optimale strategie voor het kiezen van dat tweede meettijdstip. Dit proefopzetprobleem is volledig verstrengeld met het kaptijdstip-probleem. Naar beide werd tegelijk geoptimaliseerd, met behulp van numerieke methoden. De winst van een geoptimaliseerd variabel tweede tijdstip was gering, vergeleken met een optimaal vast tijdstip.

Het tweede onderzoek behandelt het schatten van groeikrommen van de stamdiameter van individuele bomen in kurkeik opstanden. Aanbevolen wordt om een lokaal D-optimale proefopzet te gebruiken in de keuze van tijdstippen waarop de diameter moet worden gemeten. De keuze is dan zo dat de parameters van de groeikrommen zo goed mogelijk geschat worden. Om praktische redenen wordt bij een groep bomen het gebruik van een gemeenschappelijk compromis van meettijdstippen aanbevolen. In de beschikbare testgegevens gaf een dergelijke aanpak betere resultaten bij individuele groeicurves dan het gebruik van een equidistante proefopzet.

Het derde onderzoek betreft ruimtelijke modellering en het gebruik van ruimtelijke statistische methoden bij kurkeik-opstanden. De analyse betrof de ruimtelijke correlatie tussen kroonvorm, kroonomvang en stamdiameter in paren naburige bomen. Er werd een significante correlatie gevonden tussen de omvang van een boom en de competitiedruk van naburige bomen. Vooral grotere bomen bleken binnen de opstand een regelmatige ruimtelijke distributie te hebben. De ruimtelijk invloeden tussen naburige bomen zijn van belang en de opstand kan er aanzienlijk voordeel bij hebben als hier ten aanzien van het beheer rekening mee wordt gehouden.

Het vierde onderzoek richt zich op drie steekproefmethoden die men kan gebruiken in kurkeik montado's (agroforestry systeem) voor het schatten van dichtheid, grondvlak en het kurkvolume. De schattingen zijn voor de producenten zowel van economisch belang ls om keuzes te maken ten aanzien van het bosbeheer. De veelgebruikte zig-zag methode is vergeleken met twee andere bemonsteringsmethoden op gesimuleerde gegevens. Vanuit diverse beginpunten is een beperktere zigzag bemonsterd om een eerlijke vergelijking te kunnen maken. Deze methoden bemonsteren rond startpunten ofwel alle bomen binnen een bepaalde straal, ofwel een vast aantal van de meest naburige bomen. De simulaties zijn gebaseerd op moderne ruimtelijke simulatiemethoden en vertegenwoordigen een breed scala van in de praktijk optredende ruimtelijke positie- en diameterpatronen. Onder de meeste omstandigheden bleek de zigzag-methode te moeten worden afgeraden, omdat het onzuivere schatters oplevert die tevens een grotere variantie hebben dan de schattingen die met alternatieve methoden worden bepaald.

Het gebruik van moderne statistische methoden blijkt waardevol te zijn voor het verbeteren van schattingsmethoden en steekproefmethoden zoals die gebruikt worden in kurkeik en eucalyptus bossen. Adequate steekproefmethoden zijn essentieel om kwalitatief goede informatie te verkrijgen.

Acknowledgements

I started working on statistical applications to production forestry in 1994, while doing an internship at the National Forestry Station in Lisbon (EFN/INIA). The work under the supervision of Dr. Mário Tavares gave me the enthusiasm to continue doing research in the same field. After that I joined an MSc course at the Department of Forestry at the Instituto Superior de Agronomia (ISA) in Lisbon. Prof. Margarida Tomé was my MSc supervisor and in 1997 we put some ideas together for my PhD research. In May 1997 I visited Prof. Dieter Rasch at the department of mathematics in Wageningen with a draft proposal for a PhD research on modelling and sampling cork oak stands. Some time later he accepted to become my supervisor, and in 1998 I started the PhD research at Wageningen University. I am very grateful to my promotors Prof. Dieter Rasch and Prof. Alfred Stein and to my supervisor drs. Albert Otten. Dieter guided my work during the first two and a half years of this thesis, and introduced me to the methodology of optimal designs. Alfred guided my research in the last two years, specially on spatial statistics models, and gave me many good ideas. Albert was my direct supervisor during the most important stage of my research. He helped me plan my work and outline my research. I particularly enjoyed discussing and testing my ideas with him. His advice, criticism and patience were invaluable for the conclusion of this thesis. Margarida was the supplier of the research problems and of the data used in this thesis. She gave me many good ideas, supervised the field work and reviewed my articles, but most of all she inspired my research from the beginning.

The local farmers' association of Coruche district, APFC, facilitated all the field work, and the colleagues from ISA and friends from Wageningen helped carrying it: Alice Almeida, Nuno Manano, Sónia Lopes, João Oliveira, Carlo Vromans, Robbert van der Steeg, Bernd Horsman and Nico ten Heggeler. Marta B. Coelho from ISA helped planning the field work, and promptly replied to all my questions and requests of data.

And because one needs more than air to survive, the financial support from Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE), through the program 'III Quadro Comunitário de Apoio', was also greatly appreciated.

At the department of mathematics of Wageningen University I was lucky to have very friendly colleagues. I was surprised by the relaxed atmosphere at the department, and the continuous involvement of everyone. I have knocked at almost everyone's door at one time or another with a question or request. I always felt welcomed and always left with

an answer. Apart from that the coffee breaks attended by most were very *gezellig* (cozy). The secretaries in particular, Ria, Ditsie, Heleen and Hanneke, provided precious help with all non-scientific issues, and encouraged me to address them in Dutch from the very beginning of my stay in Wageningen.

I am also grateful to my AIO mates, for the relaxing lunch breaks, either at the canteen, in the swimming pool or accross the river Rijn on sunny days. Annemarie produced many funny situations with her spontaneity and sense of humour. Mark supplied my imagination with all the interesting discussions over science, literature and philosophy. Eric gave me help and good advices on several occasions. He proved to have an unshakeable good mood even after being provoked a few times. Well done, Eric! The last months before finishing my thesis were the most difficult. Anna's patience, positive thinking and calmness were truly appreciated.

Koen, Hank and Els kindly provided me with a room in their houses and helped me make the contact with Wageningen University back in 1997.

I also want to thank all my friends for the good times but also for their support in less good times. In particular to my dear friend Esther, for the cosy lunches, the laughs we had together, her support during less easy times, and above all for always being there. And also to the mates from Dijkgraaf 4-1B for being so friendly and helpful, and Jasper for that as well as for the Dutch lessons.

I am so grateful to my dear parents, for all the love, care and attention they always gave me, for motivating me to learn, letting me dream, and for being such perfect parents. And to my lovely sister, for always being such a great friend and for understanding me so well.

Finally I thank Carlo for the invaluable friendship, care and support during the last five years. And, among other things, for guiding me in Wageningen, for the bike lessons, for the many translations, and for free-willingly doing all the cooking and cleaning at home during the last months.

About the author

Maria João Paulo was born on 24 January 1969 in Cascais, Portugal. She fell in love with science at an early age. While at school she chose a study direction in science and health, with extra courses in biology. Her dream at the time was to study wild animals in Africa. In 1987 she abandoned her dream and started her studies in Geographic Engineering. In 1995 she graduated from the Faculty of Science, University of Lisbon, in Probability and Mathematical Statistics. Two high points of this graduation were the 6 months ERASMUS scholarship at the University of Sheffield and the internship at the National Forestry Station in Lisbon, where she worked on statistical applications to forestry. In 1995 she enrolled on an MSc course on Natural Resources Management at the Department of Forestry, Technical University of Lisbon. The work with Prof. Margarida Tomé brought her to Wageningen University, Department of Mathematics, where she worked on her PhD research from 1998 to 2002, under the supervision of Prof. Dieter Rasch, Prof. Alfred Stein and drs. Albert Otten. She presently works on a postdoc project at the Biometris unit, Wageningen University and Research Centre.