DÂMOCLES AURÉLIO NASCIMENTO DA SILVA

UMA ABORDAGEM BAYESIANA PARA ANÁLISE DE SOBREVIVÊNCIA DE CLONES DE EUCALIPTOS NO PÓLO GESSEIRO DO ARARIPE-PE

Recife / 2006

DÂMOCLES AURÉLIO NASCIMENTO DA SILVA

UMA ABORDAGEM BAYESIANA PARA ANÁLISE DE SOBREVIVÊNCIA DE CLONES DE EUCALIPTOS NO PÓLO GESSEIRO DO ARARIPE-PE

Dissertação apresentada ao colegiado do Mestrado de Biometria da Universidade Federal Rural de Pernambuco, para obtenção do título de Mestre em Biometria.

Orientador: Prof. Dr. Eufrázio de Souza Santos.

Co-orientador: Prof. PhD José Antônio Aleixo da Silva

Recife / 2006

Ficha catalográfica
Setor de Processos Técnicos da Biblioteca Central – UFRPE

S586 u Silva, Dâmocles Aurélio Nascimento da
Uma abordagem Bayesiana para análise de sobrevi – vência de clones de eucalíptos no pólo gesseiro do Araripe – Pe / Dâmocles Aurélio Nascimento da Silva. -- 2006.
52 f. : il.

Orientador : Eufrázio de Souza Santos
Dissertação (Mestrado em Biometria) - Universidade Federal Rural de Pernambuco. Departamento de Estatística e Informática.
Inclui bibliografia.

CDD 581. 018 2

1. Análise de sobrevivência
2. Método de estimação Bayesiano
3. Regressão de Cox
4. Araripe (PE)
I. Santos, Eufrázio de Souza
II. Título

DÂMOCLES AURÉLIO NASCIMENTO DA SILVA

UMA ABORDAGEM BAYESIANA PARA ANÁLISE DE SOBREVIVÊNCIA DE CLONES DE EUCALIPTOS NO PÓLO GESSEIRO DO ARARIPE-PE

Dissertação apresentada ao colegiado do Mestrado de Biometria da Universidade Federal Rural de Pernambuco, para obtenção do título de Mestre em Biometria.

Aprovada em 24 de fevereiro de 2006

Orientador:

Prof. Dr. Eufrázio de Souza Santos
Universidade Federal Rural de Pernambuco – UFRPE
Orientador

Examinadores

Prof. Dr. Renato José de Sobral Cintra
Universidade Federal de Pernambuco - UFPE

Prof. PhD José Antônio Aleixo da Silva
Universidade Federal Rural de Pernambuco - UFRPE

Prof. Dra Claúdia Helena Dezotti
Universidade Federal Rural de Pernambuco - UFRPE

“ A sabedoria da vida não está em fazer aquilo que se gosta, mas em gostar daquilo que se faz.”

(Leonardo da Vinci)

AGRADECIMENTOS

A Deus, ser supremo, fonte de força inesgotável.

A meus pais, Damocles Aurélio da Silva e Maria do Carmo Nascimento, pelo esforço constante no objetivo de minha formação, e sempre me estimulando a ir mais além.

Ao meus irmãos Helfarne Aurélio e Sara Jacqueline pelo incentivo.

A minha namorada Claudejane Gomes pela paciência.

Ao prof. Dr. Eufrázio Santos, pela orientação na realização deste trabalho, pelo incentivo e cobranças e pela dedicação de tanto tempo de sua vida ao programa de Pós-graduação do mestrado em Biometria, como coordenador, professor e orientador

Aos colegas de Biometria, Arundo Nunes, Antônio Lopes, Antônio José, Heliovânio Torres, Sérgio de Sá, Carlos André, Leonardo Mendes, Fábio Cavalcanti e Ilzes Celi pelo apoio e companheirismo durante todo o curso.

Ao professor Wilson Rosa de Oliveira Júnior, pelo incentivo como professor.

Ao secretário da Biometria, Marco Antônio, pela amizade.

RESUMO

Visando contribuir, como alternativa para minimizar os impactos antrópicos de caracter negativo causado, principalmente, pela busca de material combustível para atender a demanda energética da região semi-árida brasileira, utilizamos as técnicas de análise de sobrevivência para compreensão do comportamento de uma floresta de eucaliptos ao longo do tempo, e com isto racionar o uso de madeira como combustível por cerâmicas, padarias, casas de farinha e calcinadoras de gesso existentes na região. Usaremos dados provenientes de um estudo transversal de 1500 células de eucaliptos, dividido em 4 estratos, tomando como base o período de 03/2002 a 09/2004.

Utilizou-se inicialmente o gráfico de probabilidade para, baseado no teste de Anderson-Darling, tomarmos a decisão de qual função de probabilidade utilizaríamos tanto no estudo clássico como na abordagem bayesiana. Uma vez tomada a decisão de escolha da distribuição de probabilidade, utilizamos o método de Kaplan-Meier e o método Atuarial (tábua de vida) para estimativa dos parâmetros e o teste não paramétrico log-rank para testar se as curvas da função de probabilidade diferiam entre categorias de uma mesma variável. Utilizamos esse teste ao nível de significância de 0,05. Para essas análises, foi utilizado o software estatístico Minitab versão 13 e o pacote estatístico SAS.

Na abordagem bayesiana utilizou-se a o método de Monte Carlo Cadeia de Markov (MCMC) para estimativa dos parâmetros, utilizando como priori a distribuição gamma, encontrada na literatura como a distribuição que melhor adequa-se para dados biológicos e como função de densidade, utilizou-se a da distribuição Weibull, escolhida como a de melhor ajuste as dados segundo o teste de Anderson-Darling. Para essa análise foi utilizado o Winbugs 1.4.

Os resultados quanto a análise dos parâmetros indicaram que as estimativas encontradas foram próxima, mesmo utilizando métodos de estimação distintos. Conclui-se que a melhor distribuição para analisar a população em questão é a Weibull, segundo o teste de Anderson-Darling e como método para estimação dos parâmetros da distribuição, tanto o método clássico, quanto o método bayesiano, mostram-se bons estimadores, verificado pela amplitude dos intervalos de confiança a 95%. Em face dos resultados, concluímos que deve-se ter um melhor controle dos eucaliptos, nos primeiros 6 meses de plantio.

ABSTRACT

Aiming at to contribute, as alternative to minimize the resources of impacts, mainly, for the search of combustible material to take care of the energy demand of the Brazilian half-barren region, we use the techniques of analysis of survival for understanding of a forest of eucalyptus to the long one of the time, and with this to ration the use wooden as combustible for ceramics, bakeries and existing calcinatory of plaster in region. The data given proceeding from a transversal study of 1500 cells of eucalyptus, divided in 4 stratus, taking as base the period of 03/2002 to 09/2004.

The graph of probability was used initially for, being based on the test of Anderson-Darling, takes the decision of which function of probability would use in such a way in the classic study as in the Bayesian boarding. A time taken to the decision of choice of the probability distribution, we use the method of Kaplan-Meier and the Actuarial method (life table) to determine the estimates of the parameters and the test distribution free log-rank to test if the curves of the function of probability differed between categories from one same one variable. We used this test to the level of significance of 5%. For these analyses, it was used statistical software Minitab 13 version and statistical package SAS.

In the Bayesian boarding was used method Carlo the Mount Chain of Markov (MCMC) for estimate of parameters, using as priori the distribution gamma, found in literature as the distribution that more good was adjusted for biological data and as function of density, used it of the Weibull distribution, chosen as of the better adjustment to the data according to test of Anderson-darling. For this analysis software Winbugs 1.4 was used.

The results how much to the analysis of the parameters they had indicated that the joined estimates had been closed, same using distinct methods of estimation. As much was concluded that the best distribution to analyze the population in question is the Weibull, according to test of Anderson-Darling and as method for estimation of the parameters of the distribution, the classic method, how much the Bayesian method, reveals good estimators, verified for the amplitude of the intervals reliable 95%. In face of the results, we conclude that if it must have one better control of the eucalyptus, in first the six months of the plantation.

LISTA DE TABELAS

		Página
TABELA 1	**EMV e valor da estatística de Anderson-Darling para as várias distribuições de probabilidade.**	**34**
TABELA 2	**Distribuição segundo estratos e tratamentos dos Eucaliptos em estudo em Araripina (PE). Brasil 2002 a 2004**	**37**
TABELA 3	**Teste de log-rank para comparação dos estratos**	**41**
TABELA 4	**Regressão de Cox**	**42**
TABELA 5	**Sobrevivência para estrato 1**	**43**
TABELA 6	**Sobrevivência para estrato 2**	**44**
TABELA 7	**Sobrevivência para estrato 3**	**44**
TABELA 8	**Sobrevivência para estrato 4**	**44**
TABELA 9	**Convergência do parâmetro com 1000 iterações**	**45**
TABELA 10	**Comparação dos métodos para o parâmetro de forma da Weibull.**	**45**
TABELA 11	**Valores medianos para os estratos**	**45**
TABELA 12	**Parâmetro de forma da Weibull**	**46**

LISTA DE FIGURAS

		Página
FIGURA 1	**Desenho da estrutura de um estudo transversal.**	**19**
FIGURA 2	**Gráfico de probabilidade considerando a distribuição de Weibull.**	**35**
FIGURA 3	**Gráfico de probabilidade considerando a distribuição de Normal**	**35**
FIGURA 4	**Gráfico de probabilidade considerando a distribuição Lognormal base e**	**36**
FIGURA 5	**Gráfico de probabilidade considerando a distribuição Lognormal de base 10**	**36**
FIGURA 6	**Gráfico de probabilidade considerando a distribuição Exponencial**	**37**
FIGURA 7	**Sobrevivência dos Eucaliptos, Araripina (PE), Brasil, 2002 a 2004**	**38**
FIGURA 8	**Sobrevivência dos Eucaliptos no Estrato 1**	**38**
FIGURA 9	**Sobrevivência dos Eucaliptos no Estrato 2**	**39**
FIGURA 10	**Sobrevivência dos Eucaliptos no Estrato 3**	**39**
FIGURA 11	**Sobrevivência dos Eucaliptos no Estrato 4**	**40**
FIGURA 12	**Sobrevivência dos Eucaliptos no Estrato 1,2,3 e 4**	**40**
FIGURA 13	**Função densidade de probabilidade**	**41**
FIGURA 14	**Função de risco**	**41**
FIGURA 15	**Função de risco**	**42**
FIGURA 16	**Gráfico de probabilidade considerando a distribuição Weibull com dados censurados.**	**43**

SUMÁRIO

1. Introdução..... 11
2. Objetivos..... 14
3. Revisão de literatura..... 15
4. Materiais e Métodos..... 19
    4.1. Origem do banco de dados utilizado no estudo e período de referência..... 19
    4.2. Desenho do estudo..... 19
    4.3. Métodos..... 20
        4.3.1. Método atuarial..... 20
        4.3.2. Método de Kaplan-Meier..... 22
        4.3.3. Modelo de Cox..... 24
        4.3.4. Método da Máxima Verossimilhança..... 26
            4.3.4.1. A função escore e a função de informação..... 27
        4.3.5. Gráfico de probabilidade..... 29
            4.3.5.1. O teste alternativo de Anderson-Darling..... 29
            4.3.5.2. Como realizar o teste de Anderson-Darling..... 30
        4.3.6. Distribuição de Weibull..... 30
        4.3.7. Método bayesiano..... 31
5. Resultados e Discussões..... 34
6. Conclusões..... 47
7. Bibliografia..... 48

1- INTRODUÇÃO

A microrregião de Araripina fica localizada no semi-árido pernambucano, formada por 10 municípios, tem área de 11792 Km^2. Porém, o Pólo Gesseiro do Araripe é constituído pelos Municípios de Ipubí, Trindade, Bodocó e Ouricurí que apresentam 95% das jazidas nacionais em atividade. Apresenta 332 empresas instaladas , produzindo 2302 mil toneladas/ano, operando a 23,77% da capacidade. Região onde predomina, em quase toda sua extensão, condições ecológicas desfavoráveis, com elevadas temperaturas, chuvas escassas e mal distribuídas, rios temporários e vegetação xerófila, tendo como atividades fundamentais às culturas de subsistência e a pecuária extensiva (SILVA, 2006).

O semi-árido brasileiro é uma região carente de informações sobre o comportamento de essências florestais nativas e exóticas que podem se adaptar à região sem provocar alterações nas condições ambientais, capazes de minimizar ao máximo os impactos antrópicos de caráter negativos, causados, principalmente, pela busca de material combustível para atender a demanda energética da região. A utilização de madeira retirada da caatinga, por exemplo, agrava a desertificação e o desequilíbrio ambiental no semi-árido nordestino.

O setor de produção secundária, no qual se enquadra as calcinadoras de gesso do Araripe tem um consumo energético em que predomina o uso de biomassa e esse consumo oscila de acordo com os preços dos derivados de petróleo, notadamente o óleo BPF, que também é usado pelas calcinadoras do Araripe. O alto preço do óleo BPF faz com que as empresas migrem para consumo de lenha, aliado a outros fatores incentivadores como a precária fiscalização e a aplicação, por conseguinte, de penalidades irrisórias (ALBUQUERQUE, 2002).

Segundo Albuquerque (2002), o consumo de energéticos florestais em Pernambuco foi estimado em 12.117.151 st/ano, sendo o setor residencial responsável por 73,5% e o industrial por 26,5% deste total. Os produtos florestais, quando integrados a outras atividades produtivas, representam a segunda fonte

de energia do estado. As atividades que apresentam maior utilização de energéticos por ordem de consumo são: cerâmicas, padarias, casas de farinha e calcinadoras de gesso, com 24,78%; 18,15%; 13,57%; 13,40% do consumo total, respectivamente.

Das 72 calcinadoras da região, 32 reverteram o processo de calcinação de gipsita com óleo BPF para lenha. E desse total apenas 13 estão autorizadas pela Companhia Pernambucana do Meio Ambiente (CPRH) para usar a madeira como combustível. Essas 32 empresas produzem cerca de 500 mil toneladas por ano, que corresponde ao consumo de 500 mil st/ano de lenha, (INOJOSA, 2005).

As técnicas estatísticas conhecidas como análise de sobrevivência são utilizadas quando se pretende analisar um fenômeno em relação a um período de tempo, isto é, ao tempo transcorrido entre um evento inicial, no qual o sujeito ou um objeto entra em um estado particular e um evento final, que modifica este estado.

Somente nas décadas de 1950 e de 1960 apareceram às primeiras propostas de estimadores das probabilidades de sobrevivência que incorporavam a censura, vale dizer , modelos para observações incompletas. As principais técnicas são o método atuarial e o método do produto-limite de Kaplan-Meier

No estudo de análise de sobrevivência com abordagem Bayesiana, utilizam-se em geral como prioris, alguns modelos de distribuições como a de Weibull, a Exponencial e a Gama. Mas novos estudos estão sendo realizados através de método iterativos e já para algumas situações se encontra na literatura que essa distribuições não são boas para prioris, em que nesse caso é utilizado duas ou mais dessas distribuições em conjunto.

Nesse trabalho, comparamos inicialmente dados de sobrevivência de floresta de eucalipto, situada em Araripina. Foram realizadas medições em intervalos de seis meses, iniciando-se em março de 2002 e finalizando em setembro de 2004, observando 30 meses de vida dessas árvores.

Com todos os dados foi feito uma análise para saber qual dentre algumas distribuições encontradas na literatura seria a melhor para ser utilizada na análise de sobrevivência das árvores.

2. OBJETIVOS

2.1 Objetivo Geral

- Através de uma análise de sobrevivência estudar o desenvolvimento, quanto as perda , de uma floresta de Eucalipto no período de 03/2002 à 09/2004.

.

2.2 Objetivos Específicos:

- Comparar os modelos que fazem suposições na sobrevivência das árvores

- Comparar os métodos clássicos e bayesiano, através das estimativas dos parâmetros.

3. REVISÃO DE LITERATURA

Qualquer tentativa de descrição matemática de fenômenos biológicos observados deve envolver certa idealização dos fatos observados. As fórmulas matemáticas só podem proporcionar um modelo matemático simplificado da realidade, uma espécie de retrato idealizado dos aspectos característicos do fenômeno em investigação (MARTINS, 1998).

Segundo MARTINS (1998), a construção de um modelo é baseada em informações obtidas da realidade através das observações ou medidas. Em geral, é difícil afirmar, com certeza, se um modelo matemático idealizado é ou não adequado, antes que algum teste de observação seja realizado.

A análise de sobrevivência é um conjunto de técnicas e modelos estatísticos que analisa dados ao longo do tempo (a variável aleatória contínua $T \geq 0$), buscando, entre outras informações, o tempo de ocorrência de um dado evento (LOUSADA NETO *et al*.,2002). Considera-se sobrevivência, o tempo desde a entrada do indivíduo no estudo até a ocorrência do evento de interesse(falha) ou até a censura(perda por tempo de observação incompleta) na observação (KLEINBAUM,1995).

O Método atuarial e o Método de Kaplan-Meier, assumem inicialmente que as observações censuradas têm a mesma probabilidade de sofrerem o evento que aquelas que permanecem em observação. A distinção essencial entre o método atuarial e o método de Kaplan-Meier é que este último elimina a necessidade de assumir que as censuras das observações ocorram uniformemente durante este intervalo. Assume-se apenas que as observações censuradas teriam a mesma experiência futura do que aquelas que continuam sendo observadas (KAHN & SEMPOS, 1989). O método de Kaplan-Meier pode ser utilizado para qualquer tamanho de amostra em estudo, mas é especialmente, útil naqueles estudos com um número pequeno de observações, enquanto o método atuarial é mais apropriado para grandes amostras (LEE, 1992).

Na análise de sobrevivência, os parâmetros mais importantes são a probabilidade de sobrevivência no curso de cada um dos intervalos considerados e a probabilidade de sobrevivência acumulada (tratada correntemente como taxa de sobreviver), isto é, a probabilidade de sobreviver do tempo zero até o tempo final considerado. Esta última equivale à probabilidade de sobreviver em todos os intervalos anteriores ao momento considerado e, usualmente, é denominado s(t). Geralmente há uma variável de interesse, também chamada de variável dependente ou resposta. No entanto, a variável dependente de interesse é o tempo decorrido até o aparecimento de algum evento. Há, ainda, uma ou mais variáveis, denominadas independentes, preditoras ou covariáveis, cujo relacionamento com a variável dependente, em geral, é imprescindível, pois os modelos estatísticos expressam a variável dependente como uma função matemática conhecida das variáveis independentes.

Na construção da tábua de vida, CHIANG(1968) pressupõe que a força de mortalidade, definida como coeficiente de mortalidade instantâneo, seja constante em cada grupo etário. Admite ainda que os vários riscos de morte atuam simultaneamente em cada indivíduo, havendo para cada risco uma correspondente força de mortalidade (teoria dos riscos competitivos). A soma dessas é igual à força de mortalidade total, existindo uma razão constante entre as forças de mortalidade de uma causa e a total, em cada idade.

As tábuas de vida, segundo CHIANG(1968), são mais coerentes com a realidade, pois consideram a interdependência dos vários riscos e seus efeitos ao se eliminar uma causa específica (ou grupo de causas); servem para avaliar não só o impacto das causas de óbito em populações com sistemas confiáveis de estatísticas vitais, como, também, naquelas com cobertura parcial de óbitos, admitindo-se que a sub-notificação não seja diferenciada por causas.

A função de risco descreve como a probabilidade instantânea de falha (taxa de falha) se modifica com o passar do tempo, e também é conhecida como taxa de falha instantânea (COX E OAKES,1989, e LOUSADA NETO *et al*.2002). Ela pode ser utilizada para caracterizar classes especiais de distribuições de tempo de sobrevivência de acordo com o seu comportamento como função do tempo

(constante, crescente, decrescente ou mesmo não-monótona). Torna-se portanto necessária alguma metodologia para selecionar o modelo mais apropriado antes de se proceder o ajuste. Informações estruturais acerca do fenômeno, vinculadas ao conhecimento do pesquisador sobre o mesmo, podem servir de indicações para a determinação da forma da função de risco. A escolha do modelo estatístico mais apropriado dependerá do tipo de delineamento do estudo proposto, de seus objetivos, das variáveis estudadas e da maneira pela qual foram coletados e categorizados os dados. Uma vez escolhido o modelo, as principais técnicas são o método atuarial e o método do produto-limite de Kaplan-Meier.

Na abordagem bayesiana, os principais conceitos envolvidos são a probabilidade a priori e a probabilidade a posterior(ANTELMAN, 1977).

ANTELMAN(1977) e SANTOS (2001) esclarecem que a inferência bayesiana é constituída por três estruturas básicas: a distribuição a priori, que indica o estado atual de informação do pesquisador, ou seja representa o que é conhecido adicionalmente ao experimento antes da observação dos dados; a função de verossimilhança, que expressa todo o conhecimento do experimento contido nos dados, ou seja, codifica toda a informação relevante contida nos dados sobre o parâmetro; e a distribuição a posteriori, que representa o conhecimento sobre o experimento atualizado pelos dados, especificando o estado da informação sobre o parâmetro de interesse, após a observação dos dados.

SILVA E SUÁREZ (2000) e SANTANA ET. AL. (2002) destacam que a abordagem clássica é principalmente empírica utilizando somente a informação amostral como base para estimar e testar hipóteses a respeito de parâmetros populacionais, enquanto que a abordagem bayesiana utiliza toda informação amostral além do julgamento pessoal para escolha de uma distribuição de probabilidade a priori adequada para o cálculo de estimativas.

BOX E TIAO (1973) discutiram as idéias de JEFFREYS(1961), sobre a distribuição a priori para representar o estado de ausência de informação ou ignorância a respeito do comportamento probabilístico dos parâmetros. O estudo abrangeu os casos uniparamétricos e multiparamétricos. É, talvez, a discussão

mais difundida sobre priori não-informativa. Segundo BRASIL (1991), a crítica mais freqüente à análise Bayesiana é que diferentes prioris conduzem a diferentes respostas. Contudo, querendo encontrar objetividade se pode usar prioris não-informativas.

Se houver amostra pequena de dados, então é necessário fazer sérias considerações para a informação a priori. Escolher uma priori não-informativa simplesmente porque ela é conveniente será insatisfatório. Conseqüentemente, quando a amostra é grande, intervalos de confiança clássicos e intervalos de predição Bayesiano serão quase idênticos numericamente (POLLARD, 1986).

Segundo LEANDRO (2001), o intensivo desenvolvimento computacional aliado ao uso de métodos de Monte Carlo Cadeia de Markov (MCMC) têm tornado as técnicas bayesianas mais acessíveis. O uso de MCMC é provavelmente o fator mais importante no desenvolvimento deste campo e tem tornado computacionalmente viável a aplicação de Métodos bayesianos.

Segundo SANTANA (2002) a abordagem bayesiana pode ser considerada uma extensão da clássica, pois ambas as abordagens podem levar aos mesmos resultados, entretanto, estes resultados são diferentes em termos de interpretação. Ou melhor, é um enfoque alternativo para os métodos frequentista da estatística clássica (BERGERUD E REED, 1998 e SILVA e SUÄREZ, 2000)

4. MATERIAL E MÉTODOS

4.1 Origem do banco de dados utilizado no estudo e período de referência.

Para este estudo, serão utilizados dados obtidos em 30 meses de monitoramento do Ensaio de Floresta composta por 15 clones de Eucaliptos, onde são feitas 2 medições por ano, no período de março de 2002 a setembro de 2004.

4.2 Desenho do estudo

Este é um estudo quantitativo e transversal com base nos dados referentes a floresta de eucaliptos, representado na figura 1.

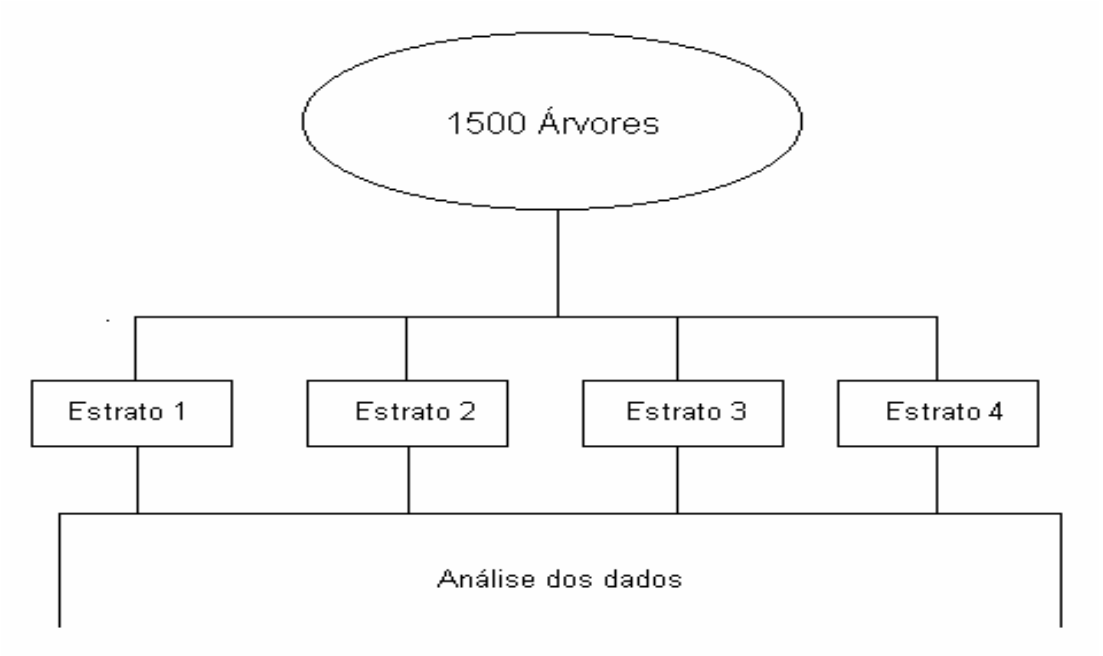


Figura 1. Estrutura de um estudo transversal.

4.3 Métodos

O ideal para uma análise de sobrevivência seria uma situação na qual tivéssemos no mínimo uma causa de morte que pudéssemos estudá-la. Esta não é a situação que se coloca a um investigador que dispõe apenas de informações acerca da sobrevivência das árvores, sem saber qual a causa da morte. Pois, considerando que a morte não é um evento repetitivo e nem atribuível a um único fator, isso pode causar interpretações equivocadas quanto ao resultado. Logo diante desta situação apenas foi possível analisar a sobrevivência das árvores sem uma causa de morte especifica e aparente.

Portanto, torna-se necessária uma metodologia para selecionar o modelo mais apropriado antes de se proceder o ajuste. Informações acerca do fenômeno, vinculadas ao conhecimento do pesquisador sobre o mesmo, podem servir de indicações para a determinação da forma da função de risco. Utilizou-se o teste de Anderson-Darling, para verificarmos qual o melhor modelo estatístico. Uma vez escolhido o modelo, as principais técnicas são o método atuarial e o método do produto-limite de Kaplan-Meier.

4.3.1 Método atuarial ou tábua de vida

O método atuarial para dados incompletos (LEE, 1992) calcula as probabilidades de sobrevivência em intervalos fixados previamente, e o número dos expostos a risco corresponde aos pacientes vivos ao início de cada intervalo **X**. O número de expostos ($\mathbf{l_x}$) é ajustado de acordo com o número de censuras que ocorreram neste período, na suposição de que as censuras ocorreram uniformemente durante o período **x** e que a experiência subseqüente dos casos censurados é a mesma daqueles que permanecem em observação (KAHN & SEMPOS, 1989). Assim, na presença de censura, é feito um ajuste no número de pessoas expostas ao risco no início do período **x**, subtraindo-se metade das censuras do total de expostos ao risco no início do período, supondo-se que estes

indivíduos estiveram, em média, expostos ao risco apenas metade do intervalo de seguimento (SZKLO & NIETO, 2000). Nessa tábua de vida, o tempo também é dividido em intervalos fixos e a probabilidade de óbito (q_x) e a de sobrevida (p_j) são calculadas para cada um dos intervalos. Tem-se então:

$$q_x = d_x / l_x^*, \qquad p_x = 1 - q_x \qquad \text{e} \qquad l_x^* = l_x - \frac{w_x}{2}$$

no qual,

q_x = probabilidade condicional de morte no intervalo x ;

p_x = probabilidade condicional de sobrevida no intervalo x ;

d_x = número de mortos no intervalo x;

q^*_x = complemento de p^*_x ;

l^*_x = o número de expostos ao risco, no intervalo x, corrigidos de acordo com a censura;

p^*_x = probabilidade condicional de sobrevida no intervalo x corrigida pela censura;

d^*_x = número de mortes corrigidas pela censura.;

l_x = número de pessoas expostas ao risco no início do período;

w_x = número de pessoas perdidas de observação no intervalo **X;**

Uma das fórmulas das funções de sobrevida é a da probabilidade de sobrevida acumulada até o tempo **x**, ou **S(t)** é dada por:

$$S(t) = \prod_{x=0}^{x-1} p_x^* = \prod_{x=0}^{x-1} 1 - q_x^*$$

Há ainda a fórmula da função de riscos (*hazard function*), ou **h(t)**, também conhecida como força instantânea de mortalidade ou taxa instantânea de óbito em um período curto de tempo, dado que um indivíduo estava vivo até o instante **t-1**:

$$h(t) = \frac{d_x}{l_x^* - \frac{d_x}{2}}$$

dado que $T > t_{x-1}$.

4.3.2 Método de Kaplan-Meier

Na análise de sobrevida pelo método de Kaplan-Meier (KAPLAN & MEIER, 1958; LEE, 1992; KLEINBAUM, 1995) os intervalos de tempo não são fixos, mas determinados pelo aparecimento de uma falha (por exemplo, o óbito). Nessa situação, o número de óbitos em cada intervalo deve ser um. Esse é um método não paramétrico, ou seja, que independe da distribuição de probabilidade (KAPLAN & MEIER,1958 e COLTON, 1979), e para calcular os estimadores, primeiramente, deve-se ordenar os tempos de sobrevida em ordem crescente (t_1 , t_2 , ..., t_n). Os sobreviventes ao tempo t (l_t) são ajustados pela censura, ou seja, os pacientes censurados entram no cálculo da função de probabilidade de sobrevida acumulada até o momento de serem considerados como perda. Isto propicia o uso mais eficiente das informações disponíveis .

Define-se a função S(t) por um estimador conhecido como *estimador produto limite de Kaplan-Meier*, pois é o limite do produto dos termos até o tempo **t**:

$$S(t) = \prod_{t=0}^{j} \frac{l_j - i}{l_j}, \text{ em que } \begin{cases} \text{i = 1, se for falha} \\ \text{i = 0, se for censura} \end{cases}$$

e l_j = número de expostos ao risco no início do período.

FALHA : é a morte da árvore e CENSURA : é a perda de informação sobre a sobrevivência da árvore, ou seja perdemos informação da árvore se saber se ela morreu ou não.;

No caso de haver empate, utiliza-se na fórmula o maior valor de **i**. Por exemplo:

$$\text{se } t_2 = t_3 = t_4 \Rightarrow p(t_2) = p(t_3) = p(t_4) = \frac{l_2 - 4}{l_2}$$

ou seja, dentre aqueles que apresentaram o mesmo número de falhas no intervalo t, utiliza-se apenas o de maior índice no cálculo da probabilidade de t.

Tanto o Método Atuarial como o Método de Kaplan-Meier assumem como premissa que as observações censuradas têm a mesma probabilidade de sofrerem o evento que aquelas que permanecem em observação, isto é, as censuras devem ser independentes da sobrevida. Nos estudos que contemplam períodos extensos de observação é necessário assegurar que não tenham existido mudanças importantes nas características destes indivíduos e no diagnóstico, exposição ou tratamento da doença em estudo ao longo deste período. Tais mudanças poderiam introduzir viés nas estimativas de sobrevida, cuja direção depende das características da coorte e do período estudado (KAHN & SEMPOS, 1989; SZKLO & NIETO, 2000).

A aplicação desses modelos permite comparar o conjunto de curvas de sobrevida das diversas categorias de uma única variável independente. Para comparar as curvas de sobrevida acumulada entre diferentes categorias de uma mesma variável, recomenda-se utilizar o teste *log-rank* (COX & OAKES, 1984; KLEIBAUM, 1995), que se baseia no confronto entre os óbitos observados nos dois grupos e aqueles esperados. O teste log-rank é aplicado para testar se as curvas diferiam entre categorias de uma mesma variável. A diferença entre óbitos observados e esperados é avaliada por meio do teste do qui-quadrado.

4.3.3 Modelo de Cox

A análise de regressão múltipla também pode ser feita na análise de sobrevida, quando se deseja avaliar o efeito conjunto de algumas variáveis independentes, sejam as observações incompletas ou não. Os primeiros modelos de regressão para análise de sobrevida foram desenvolvidos na década de 1960 e eram totalmente paramétricos, ou seja, baseados nas premissas de validade da estatística tradicional. Em 1972, Cox desenvolveu um modelo de regressão semi-paramétrico, também conhecido como *modelo de riscos proporcionais de Cox*, *modelo de Cox*, ou *regressão de Cox* (COX, 1972). Essa técnica é indicada quando se deseja estudar sobrevivência sob o prisma de causalidade ou da predição, pois fornece as estimativas das razões de risco dos fatores estudados, podendo-se avaliar o impacto que alguns fatores de risco ou fatores prognósticos têm no tempo até a ocorrência do evento de interesse. A função de riscos (*hazard function* - h(t)), no modelo de Cox, é considerada como variável dependente, e os riscos de morte por uma determinada causa são o produto de uma função não especificada de tempo (que é comum a todos os indivíduos) e uma função conhecida (que é a combinação linear das covariáveis X_i, sendo i = 1, 2, ..., k). Nele, a função de riscos h(t) é escrita em termos das covariáveis:

$$h\left(t / X, X_2, \ldots, X_k\right) = h_0(t) \cdot \exp\left(\beta_1 X_1 + \ldots + \beta_k X_k\right)$$

em que $h_0(t)$ é a parte não paramétrica do modelo, e, em estudos em que o objetivo é estimar fatores prognósticos, não há interesse em definÍ-la (pois é comum a todos os indivíduos). Os coeficientes de regressão (β_i) são estimados pelo método da máxima verossimilhança parcial.

Ao se fazer a divisão dos dois lados da equação por $h_0(t)$, obtém-se:

$$\frac{h(t / X_1, X_2, ..., X_k)}{h_0(t)} = \exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$$

O quociente $\frac{h(t / X_1, X_2, ..., X_k)}{h_0(t)}$

é chamado de *função das razões de riscos*

$$HR(i) = HR_1 = \exp[\beta_1 X_{i1} + ... + \beta_1 X_{i1}]$$

Esta fórmula também é útil para estimar a razão entre as funções de riscos (HR) para cada uma das variáveis independentes (Xi), supondo todas as outras $X_{j\neq i}$ como constantes.

$$HR(X_i) = \exp(\beta_i)$$

As suposições feitas são as de que diferentes indivíduos têm funções de riscos proporcionais entre si, e que a razão entre essas funções de risco não varia no tempo. Quando, durante o período de seguimento, a probabilidade de sobrevida de um grupo de indivíduos expostos a determinado fator não for proporcional à dos não expostos, isto é, os riscos não são constantes e proporcionais durante o período, deve-se fazer modificação no modelo que Cox propôs inicialmente. Esse último modelo é conhecido como *modelo de Cox com variável tempo-dependente*

Para a regressão de Cox, a equação tomada foi:

Predição = constante + coeficiente(preditor) + ... + coeficiente(preditor) + escala(error), ou seja , em que $Y_p = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \sigma \varepsilon_p$

Predição(Y_p): È o log do tempo de falha para as distribuições Weibull, Exponecial e Lognomal ou é o tempo de falha para as distribuição Normal, Valor extremo e Logístico.

Preditores (X_1, X_2, ..., X_p): As variáveis preditoras podem ser continuas ou categóricas.

Constante (β_0)O valor de Y, onde as variáveis são iguais a zero.

Escala (σ): O parâmetro de escala é para as distribuições Weibull e Exponencial igual ao inverso da forma, ou seja escala = 1.0 / forma.

Erro(ε_p): Neste modelo, supõe-se que o erro segue a distribuição do valor extremo

4.3.4 Método de Máxima Verossimilhança

O método da máxima verossimilhança é um dos melhores métodos para obter estimadores pontuais de um parâmetro. O estimador de máxima verossimilhança (EMV) de um parâmetro θ é o valor que maximiza a função de verossimilhança $L(\theta)$, onde $L(\theta) = \prod_{i=1}^{n} f(y_i, \theta)$, em que $f(y_i, \theta)$ é a função densidade de probabilidade discreta ou contínua.

A função $f(y_i,\theta)$ pode ter mais de um parâmetro, geralmente representados pelo vetor $\tilde{\theta}$. Nesse caso, a função de verossimilhança pode ser escrita da seguinte maneira:

$$L(\tilde{\theta}) = \prod_{i=1}^{n} f(y_i, \tilde{\theta})$$

O estimador de máxima verossimilhança de θ é geralmente denotado por $\hat{\theta}$, e baseado em uma amostra aleatória usualmente representada por $x_1, x_2, ..., x_n$. Assim o estimador $\hat{\theta}$ que melhor explica os dados da amostra é o valor de θ que maximiza a probabilidade dos dados sob o modelo.

A função de máxima verossimilhança $L(\theta)$ é um produto de termos, o que facilita para trabalhar com logaritmos, pois o logaritmo do produto é a soma do logaritmo dos fatores. Assim, o logaritmo da função de verossimilhança é naturalmente o logaritmo de $L(\theta)$, ou seja:

$$l(\theta) = \ln[L(\theta)]$$

Logo, o valor de θ que maximiza $L(\theta)$, maximiza também $l(\theta)$. Na prática, geralmente é mais fácil trabalhar com o logaritmo da função de verossimilhança. Então o EMV $\hat{\theta}$ é o valor de θ que maximiza o logaritmo da função de verossimilhança (MARTINEZ – ESPINOSA et.al., 2000)..

4.3.4.1 A função escore e a função de informação

Para calcular o valor de $\hat{\theta}$, faz-se necessário maximizar para todos os possíveis valores de $l(\theta)$. Isto é realizado geralmente pela diferenciação do $l(\theta)$

em relação a θ. Igualando-se a derivada a zero, encontramos $\hat{\theta}$. Deriva-se novamente (ou seja, efetua-se a derivada segunda) e se verificarmos que o resultado foi negativo, então o valor máximo foi encontrado.

Com a derivada primeira do logaritmo da função de verossimilhança em relação a θ, define-se a função escore $s(\theta)$, dada por:

$$s(\theta) = l'(\theta) = d_\theta[l(\theta)] = \frac{dl(\theta)}{d\theta}$$

A função de informação , $f(\theta)$, é a derivada segunda do logaritmo da função de verossimilhança em relação a θ, multiplicada por (-1), dada por:

$$f(\theta) = -l''(\theta) = -s'(\theta) = -\frac{d^2 l(\theta)}{d\theta^2}$$

O espaço paramétrico Ω é o espaço dos possíveis valores de θ. Neste caso, Ω é um intervalo de valores reais, em que a primeira e segunda derivada de $l(\theta)$ em relação a θ existem para todo ponto interno de Ω. Se $\hat{\theta}$ é um ponto interior de Ω, a primeira derivada será zero e a segunda derivada será negativa para $\theta = \hat{\theta}$. Logo, sob estas condições, temos:

$$s(\hat{\theta}) = 0$$

com $f(\hat{\theta}) > 0$.

Os valores de $\hat{\theta}$ são encontrados através da determinação das raízes da função escore $s(\hat{\theta}) = 0$. Dependendo das distribuições, esse cálculo pode ser

simples ou pode necessitar de métodos numéricos para sua solução. Um dos métodos iterativos muito utilizado, para esse cálculo é o método iterativo de Newton-Raphson (KALBFLEISCH, 1985).

4.3.5 Gráfico de probabilidade

O gráfico de probabilidade é um gráfico de probabilidades acumuladas estimadas. Onde as percentagens (probabilidades associadas aos dados) são transformadas e usadas como a variável Y, contra as dados x ou contra o logaritmo dos dados ln(x).

O gráfico de probabilidades é formado por pontos e por uma reta estimada. Os pontos deste gráfico representam percentagens dos dados e são calculados utilizando uma combinação dos métodos não paramétricos e paramétricos. A reta estimada é uma representação gráfica dos percentis, os quais são obtidos utilizando estatística de ordem, estimadores de máxima verossimilhança de uma distribuição de probabilidades selecionada e a função inversa da função de distribuição acumulada desta distribuição selecionada. Considerando que a reta estimada é uma representação dos percentis, primeiro é preciso calcular os percentis para distintas percentagens, com base na distribuição selecionada. Portanto, a transformação de escala, escolhida para linearizar a reta estimada, depende da distribuição dos parâmetros escolhidos. Assim, quanto mais próximos da linha estimada, melhor a distribuição de probabilidade estima os parâmetros(MARTINEZ-ESPINOSA e CALL JÚNIOR, 2000).

4.3.5.1 O teste Alternativo de Anderson-Darling

Para confirmar o ajuste gráfico, alguns teste de hipóteses não paramétricos podem ser utilizados. Estes testes consideram a forma da distribuição da população em lugar dos parâmetros (ROMEU, 2003). Por este motivo são chamados de testes não-paramétricos. As medidas de ajuste dependem do método de estimação utilizado, sendo o teste de Anderson-Darling, usado para os

métodos de máxima verossimilhança e de mínimos quadrados. E uma medida da proximidade dos pontos e da reta estimada no gráfico de probabilidade. O teste de Anderson-Darling é um teste alternativo aos teste de aderência de Chi-quadrado e Kolmogorov-Sminov, o qual tem a vantagem de ser mais sensível que os dois mencionados, pois dá mais peso aos pontos das caudas de distribuição. Assim, valores pequenos da estatística de Anderson-Darling indicam que a distribuição estima melhor os dados (STEPHENS, 1974).

4.3.5.2 Como realizar o teste de Anderson-Darling

Para estabelecer um critério de rejeição ou não rejeição do modelo (distribuição de probabilidade), é formulada o seguinte teste de hipótese:

$$\begin{cases} H_0 : Y \;\; segue \;\; uma \;\; \det er\min ada \;\; distribuição \;\; de \;\; probabilidade \\ H_1 : Y \;\; não \;\; segue \;\; uma \;\; ter\min ada \;\; distribuição \;\; de \;\; probabilidade \;\; proposta \end{cases}$$

A estatística do teste para tomar a decisão é dada por:

$$A^2 = -n - \sum_{i=1}^{n} \frac{(2i-1)}{n} \ln\left[F(x_i) + \ln\left(1 - F(x_{n+1-i})\right)\right]$$

em que F é a função de distribuição acumulada da distribuição especifica. Observe que x_i são os dados ordenados (NIST, 2002).

Os valores críticos ou de rejeição para o teste de Anderson-Darling dependem da distribuição específica que está sendo testada. O teste de Anderson-Darling é um teste unicaudal e a hipótese nula H_0 é rejeitada se o teste estatístico fornecer valor superior ao crítico. Cabe observar que este teste pode ser ajustado, multiplicando-no por uma constante que depende do tamanho da amostra. Estas constantes podem ser encontradas em livro como o NIST (2002).

4.3.6 Distribução Weibull

A função de sobrevivência para a distribuição Weibull é:

$$f(t_i, \mathbf{z}_i) = re^{\mathbf{bz}_i} t_i^{r-1} \exp(-e^{\mathbf{bz}_i} t_i^r)$$

onde t_i é o tempo de falha de um indivíduo com o vetor de covariávies $\mathbf{z}_i$ e $\mathbf{b}$ é o vetor de parâmetro da distribuição.

A função de risco para a distribuição apresenta-se da seguinte forma:

$$\lambda_0(t_i) = rt_i^{r-1}$$

A parametrização de μ é dado por $\mu_i = e^{bz_i}$ e $t_i \sim$ Weibull(t, m_i). O valor mediano para a sobrevivência para indivíduos com covariáveis do vetor z_i é dado por $m_i = (\log 2e^{-bz_i})^{1/r}$

4.3.7 Método Bayesiano

A análise Bayesiana de dados de sobrevivência tem tido uma grande gama de aplicações devido ao grande avanço nas técnicas computacionais e de modelagem, ocasionando uma grande quantidade de aplicação na área biológica.

A abordagem Bayesiana para análise de sobrevivência inclui o estudo de vários tipos de modelos, modelos paramétricos e semi-paramétricos, modelos de riscos proporcionais e não proporcionais, modelos de sobrevivência, modelos de tempo de falha acelerada e etc.

Na abordagem Bayesiana, podemos trabalhar sem informação a priori e nesse caso o método frequentista passa a ser um caso particular do método bayesiano, ou podemos trabalhar com informação a priori onde torna-se necessária alguma metodologia para selecionar o modelo mais apropriado para ser aplicado no estudo. Informações estruturais acerca do objeto de estudo e o conhecimento do pesquisador sobre o assunto, ajudam na escolha do modelo de distribuição. Em análise de sobrevivência com abordagem bayesiana são usadas geralmente como priori as distribuições Weibull, Exponencial e Gama (Ibrahim,2001).

O paradigma Bayesiano é baseado ao se especificar um modelo de probabilidade para os dados observados D, dado um vetor de parâmetros

desconhecidos θ, levando a função de verossimilhança $L(\theta \mid D)$. Então, presumimos que θ é aleatório e tem uma distribuição a priori denotada por $\pi(\theta)$. A inferência concernente a θ é então baseada na distribuição a posteriori, que é obtida pelo teorema de Bayes. A distribuição a posteriori de θ é dada por:

$$\pi(\theta \mid D) = \frac{L(\theta \mid D)\pi(\theta)}{\sum L(\theta \mid D)\pi(\theta)}$$

A função de densidade da distribuição gamma é dada por:

$$f(\lambda, \gamma) = \frac{\lambda^{\gamma}}{\Gamma(\gamma)} t_i^{\lambda-1} \exp(-\lambda t_i)$$

A função de densidade da distribuição Weibull é :

$$f(\lambda, \gamma) = \lambda\gamma t_i^{\lambda-1} \exp(-\lambda t_i^{\gamma})$$

em que $t \in [0, \infty)$.

Então a função de verossimilhança $L(\lambda, \gamma)$ será:

$$L(\lambda, \gamma) = \prod_{i=1}^{n} f(\lambda, \gamma) = \prod_{i=1}^{n} \lambda\gamma t_i^{\lambda-1} \exp(-\lambda t_i^{\gamma})$$

em que $\log_e L(\lambda, \gamma) = n \log_e \lambda + (\gamma - 1) \sum_{i=1}^{n} \log_e t_i - \lambda \sum_{i=1}^{n} t_i^{\gamma}$

Derivando $\log_e L(\lambda, \gamma)$ em relação a λ e γ, temos

$$\frac{\partial \log_e L(\lambda, \gamma)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i^{\lambda} \quad \text{e} \quad \frac{\partial \log_e L(\lambda, \gamma)}{\partial \gamma} = \frac{n}{\gamma} + \sum_{i=1}^{n} \log_e t_i - \lambda \sum_{i=1}^{n} t_i^{\gamma} \log_e t_i$$

Logo, nos temos que:

$$\hat{\lambda} = \left(\frac{\sum_{i=1}^{n} t_i^{\hat{\gamma}}}{n} \right)^{-1} \qquad \text{e} \qquad \hat{\lambda} = \frac{\frac{n}{\hat{\gamma}} + \sum_{i=1}^{n} \log_e t_i}{\sum_{i=1}^{n} t_i^{\hat{\gamma}} \log_e t_i}$$

em que $\hat{\lambda}$ e $\hat{\gamma}$ são os estimadores de máxima verossimilhança. Se igualarmos as equações anteriores, obtemos uma equação mais simples para $\hat{\gamma}$, dada por:

$$\left(\frac{\sum_{i=1}^{n} t_i^{\hat{\gamma}}}{n} \right)^{-1} = \frac{\frac{n}{\hat{\gamma}} + \sum_{i=1}^{n} \log_e t_i}{\sum_{i=1}^{n} t_i^{\hat{\gamma}} \log_e t_i} \Rightarrow \left(\frac{n}{\sum_{i=1}^{n} t_i^{\hat{\gamma}}} \right) \sum_{i=1}^{n} t_i^{\hat{\gamma}} \log_e t_i = \frac{n}{\hat{\gamma}} + \sum_{i=1}^{n} \log_e t_i$$

Logo a função de distribuição a porteriori $f(x/\lambda,\gamma)$será:

$$f(x/\lambda,\gamma) \;\; \alpha \;\; \frac{\lambda^{\gamma}}{\Gamma(\gamma)} t_i^{\lambda-1} \exp(-\lambda t_i) \;\; \cdot \;\; \prod_{i=1}^{n} \lambda \gamma t_i^{\lambda-1} \exp(-\lambda t_i^{\gamma})$$

em que o símbolo α indica proporcionalidade e pode ser expressa por:

$$f(x/\lambda,\gamma) \;\; \alpha \;\; \frac{\lambda^{\gamma}}{\Gamma(\gamma)} t_i^{\lambda-1} \exp(-\lambda t_i) \;\; \cdot \;\; \left[n \log_e \lambda + (\gamma - 1) \sum_{i=1}^{n} \log_e t_i - \lambda \sum_{i=1}^{n} t_i^{\gamma} \right]$$

Para realizar estas análises serão utilizados os programas computacionais SAS, MINITAB 13, Winbugs 1.4.

5. RESULTADOS E DISCUSSÕES

As informações sobre a morte das árvores foram obtidas dos dados referentes a Eucaliptos. Foram consideramos como censurados todas as árvores cujo acompanhamento foi interrompido após um período de seguimento de no mínimo 30 meses.

A análise foi realizada através da estimativa de curvas de sobrevivência (para os 4 extratos) utilizando o método de Kaplan Meier. Possíveis diferenças nas curvas de sobrevivência foram testadas através do teste de logaritmo de escores (log rank), adotando-se como limite de significância um α = 0,05. Finalmente, as variáveis estudadas foram modeladas utilizando regressão de Cox, com vista à determinação de riscos (hazard ration) e definição das covariáveis com valor na predição do tempo de sobrevivência.

Na tabela (1) apresenta-se o valor de Anderson-Darling , considerando todas as árvores.

Tabela 1: EMV e valor da estatística de Anderson-Darling para as várias distribuições de probabilidade

Distribuição	Coeficientes		Estatistica de Anderson-Darling
Weibull	30 (escala)	7,55 (forma)	16,68
Normal	14,25 (locação)	10,96 (forma)	16,87
Lognormal base *e*	2,37 (locação)	0,74 (escala)	17,89
Lognormal base 10	1,03 (locação)	0,32 (escala)	17,89
Exponencial	53,44 (escala)	1.0000 (forma)	17,21

Nas Figuras 2 a 6 são apresentados os gráficos de probabilidade para os dados considerando as distribuições da Tabela 1.

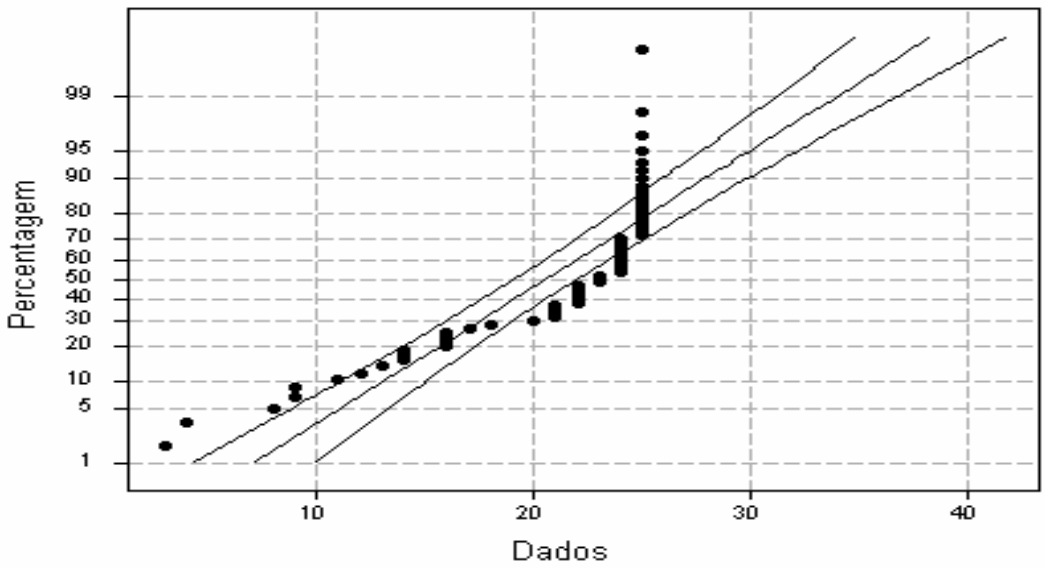


Figura 2: Gráfico de probabilidade considerando a distribuição de Weibull.

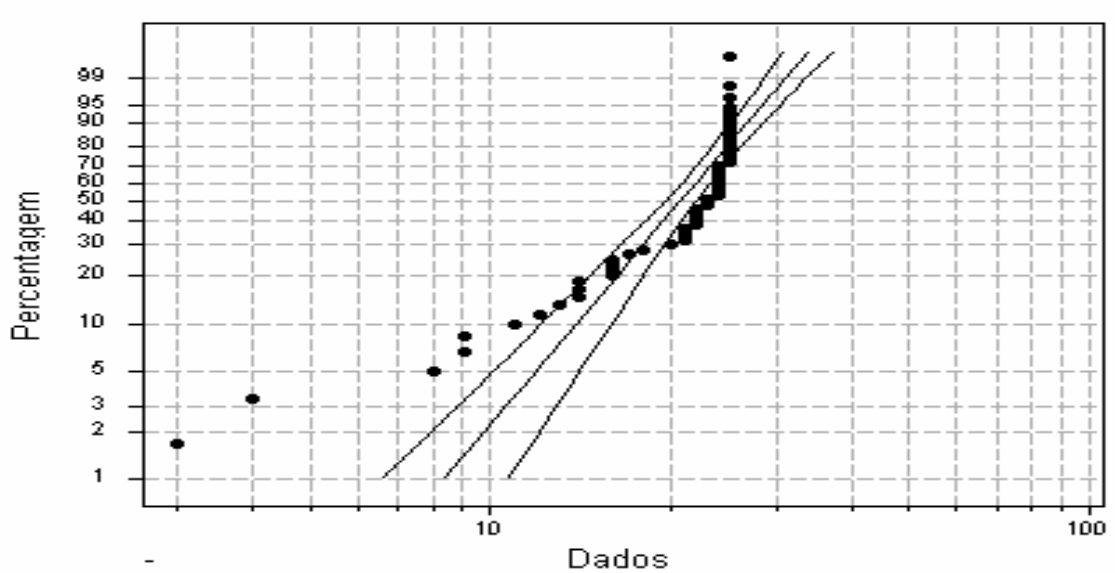


Figura 3: Gráfico de probabilidade considerando a distribuição de Normal

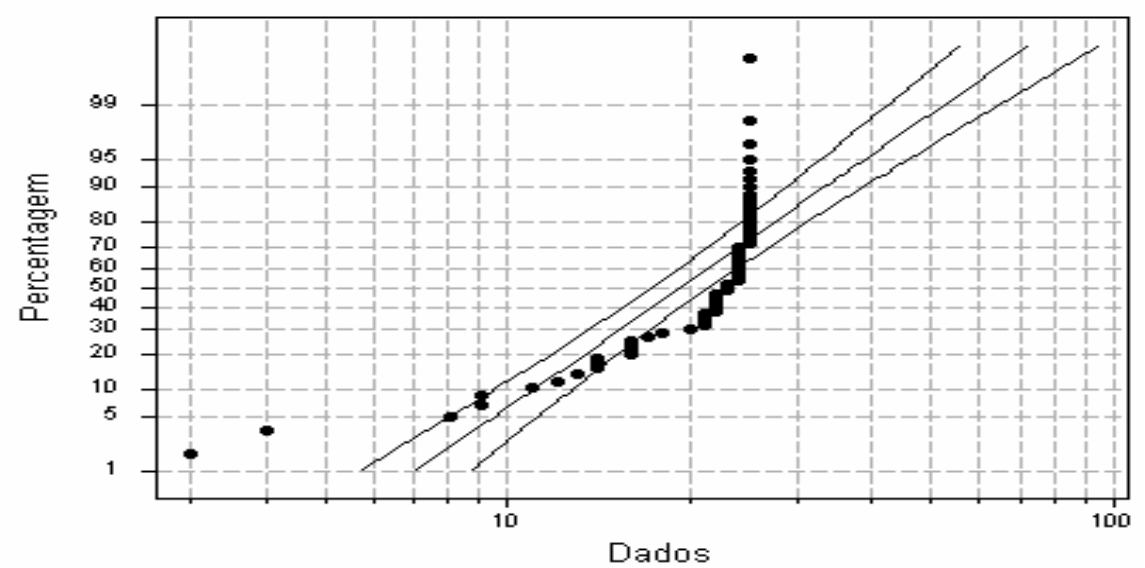


Figura 4: Gráfico de probabilidade considerando a distribuição Lognormal base e.

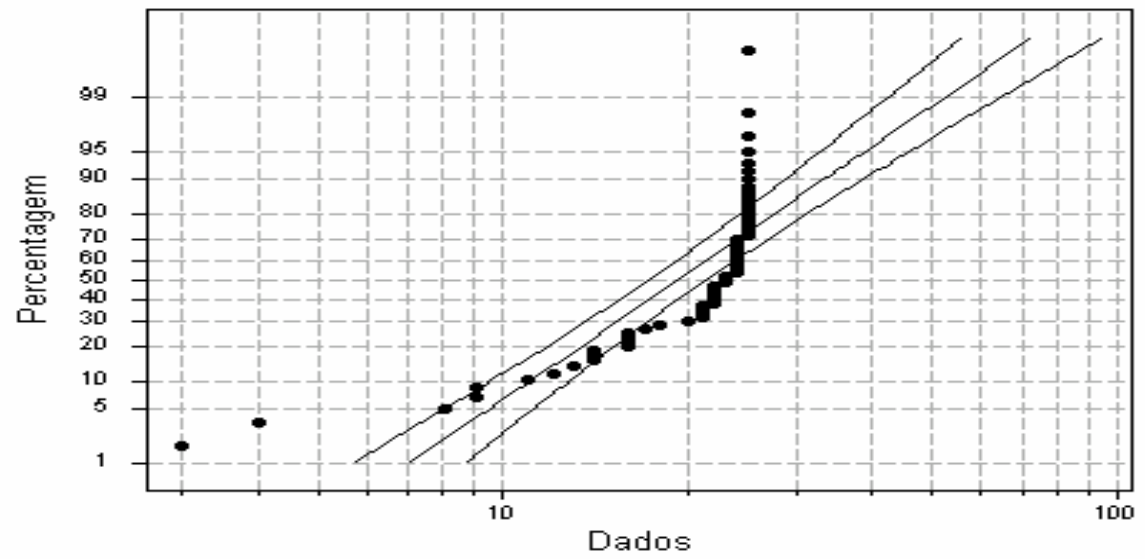


Figura 5: Gráfico de probabilidade considerando a distribuição Lognormal base 10.

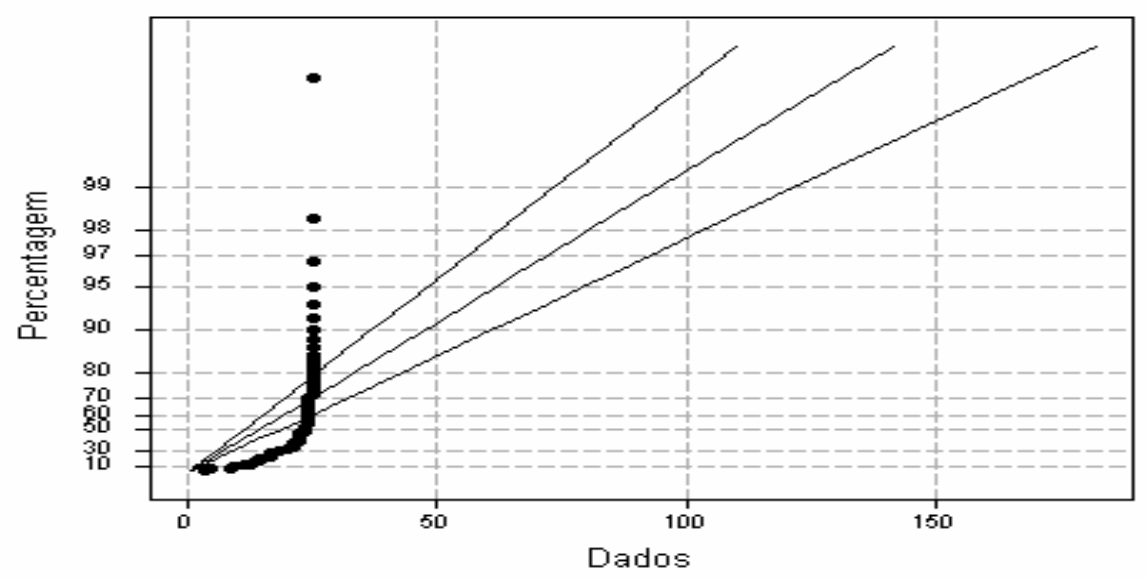


Figura 6: Gráfico de probabilidade considerando a distribuição Exponencial

Da tabela 1 e das Figuras 2 a 6 pode-se concluir que a distribuição Weibull é, dentre as distribuições apresentadas, a mais adequada para os dados em estudo.

Na tabela 2, apresenta-se o número de Eucaliptos em cada estrato com 6 meses de plantio.

Tabela 2: Distribuição segundo estratos e tratamentos dos Eucaliptos em estudo em Araripina(PE), Brasil, 2002 a 2004

	Extratos								TOTAL	
	1		2		3		4			
Tratamentos	No.	(%)	No.	(%)	No.	(%)	No.	(%)	No.	(%)
1	25	2.01	21	1.69	24	1.93	24	1.93	94	7.55
2	25	2.01	24	1.93	25	2.01	25	2.01	99	7.95
3	24	1.93	25	2.01	21	1.69	16	1.29	86	6.91
4	25	2.01	25	2.01	25	2.01	24	1.93	99	7.95
5	24	1.93	22	1.77	16	1.29	10	0.8	72	5.78
6	25	2.01	25	2.01	24	1.93	14	1.12	88	7.07
7	25	2.01	25	2.01	22	1.77	17	1.37	89	7.15
8	22	1.77	24	1.93	22	1.77	14	1.12	82	6.59
9	25	2.01	25	2.01	25	2.01	24	1.93	99	7.95
10	25	2.01	25	2.01	25	2.01	20	1.61	95	7.63
11	25	2.01	22	1.77	16	1.29	9	0.72	72	5.78
12	21	1.69	9	0.72	14	1.12	4	0.32	48	3.86
13	25	2.01	23	1.85	19	1.53	13	1.04	80	6.43
14	25	2.01	24	1.93	24	1.93	16	1.29	89	7.15
15	24	1.93	11	0.88	14	1.12	4	0.32	53	4.26
	365	29.3	330	26.5	316	25.4	234	18.8	1245	100

Na análise de Sobrevivência, encontramos a curva de sobrevivência para o grupo de árvores como um todo, mostrada na Figura 7.

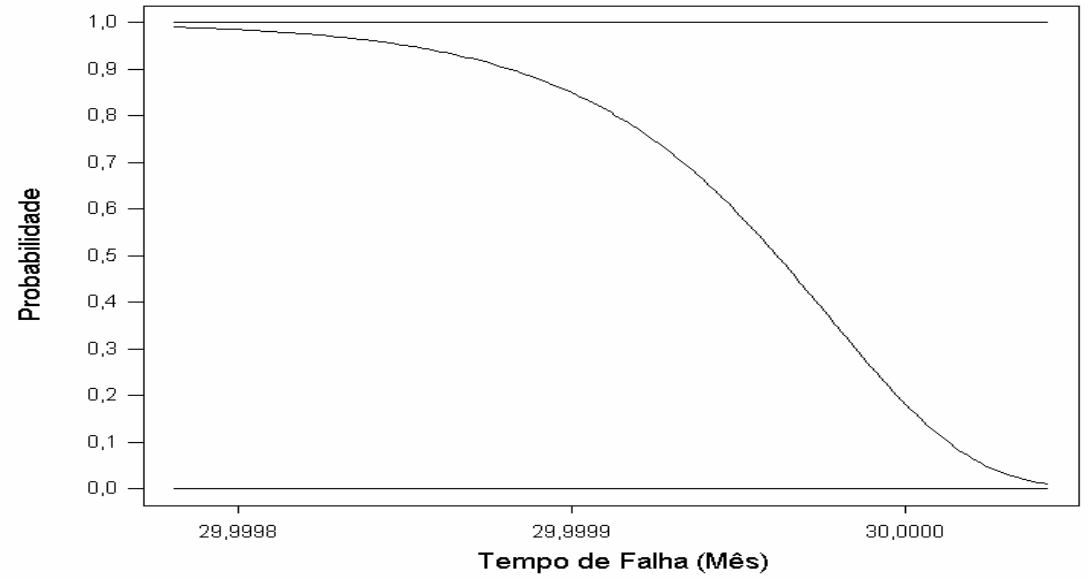


Figura 7: Sobrevivência dos Eucaliptos, Araripina (PE), Brasil, 2002 a 2004

O valor mediano de sobrevida para o grupo foi de 30 meses. Graficamente observa-se que os dados tem uma alta probabilidade de sobrevivência ao longo do tempo, tendo o fim do experimento como fator determinante a ocorrência da censura.

A Figura 8 mostra a curva de sobrevivência para as árvores do estrato 1.

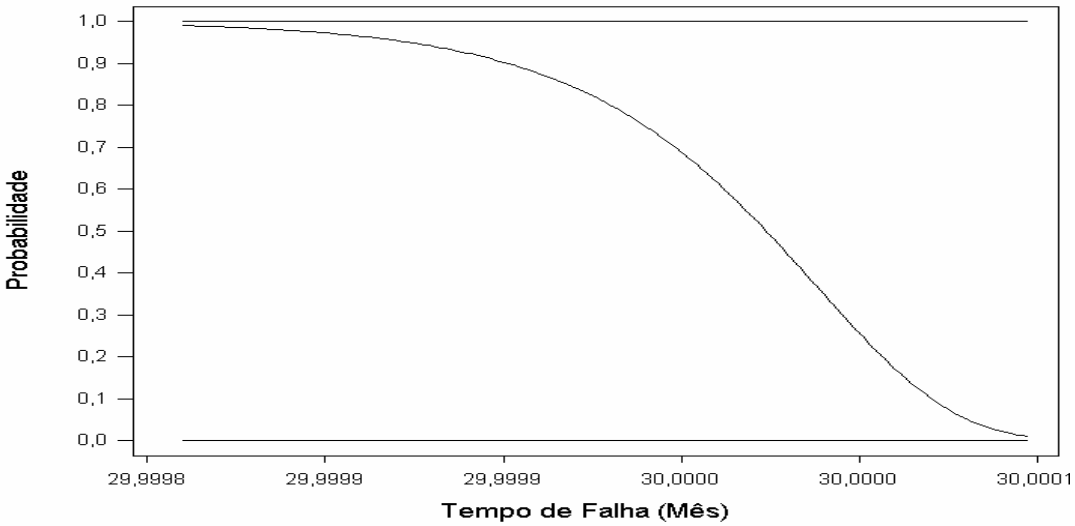


Figura 8: Sobrevivência dos Eucaliptos no Estrato 1

O valor mediano de sobrevida para o estrato 1 foi de 30 meses.

A Figura 9 mostra a curva de sobrevivência para as árvores do estrato 2

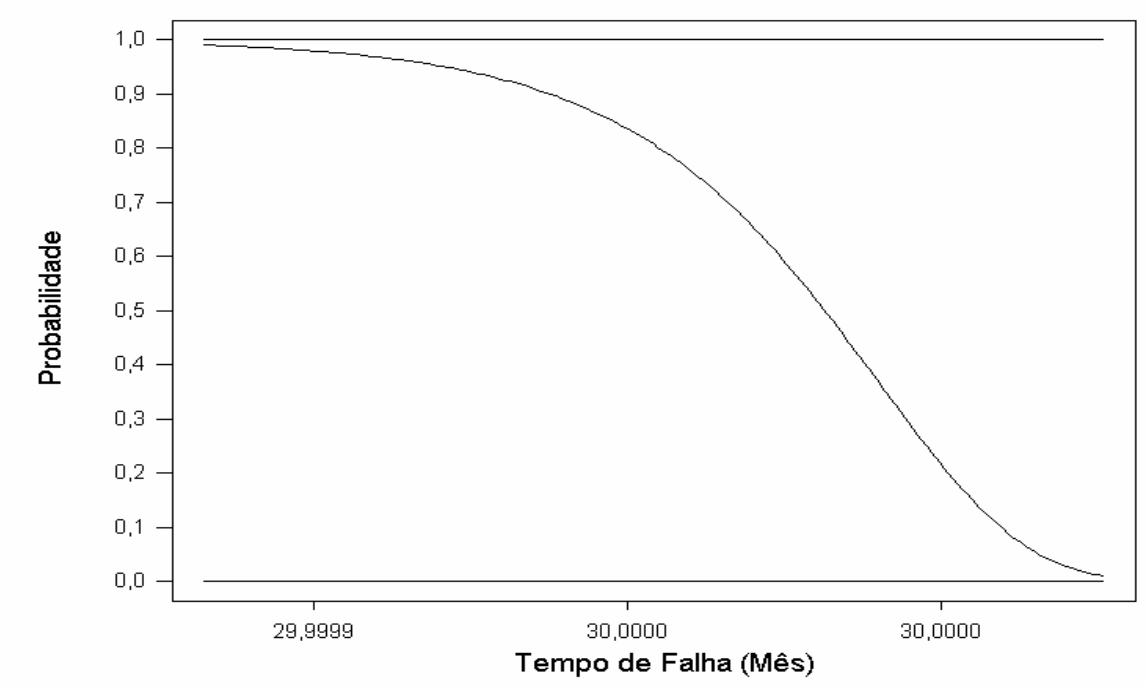


Figura 9: Sobrevivência dos Eucaliptos no Estrato 2

O valor mediano de sobrevida para o extrato 2 foi de 30 meses.

A Figura 10 mostra a curva de sobrevivência para as árvores do estrato 3

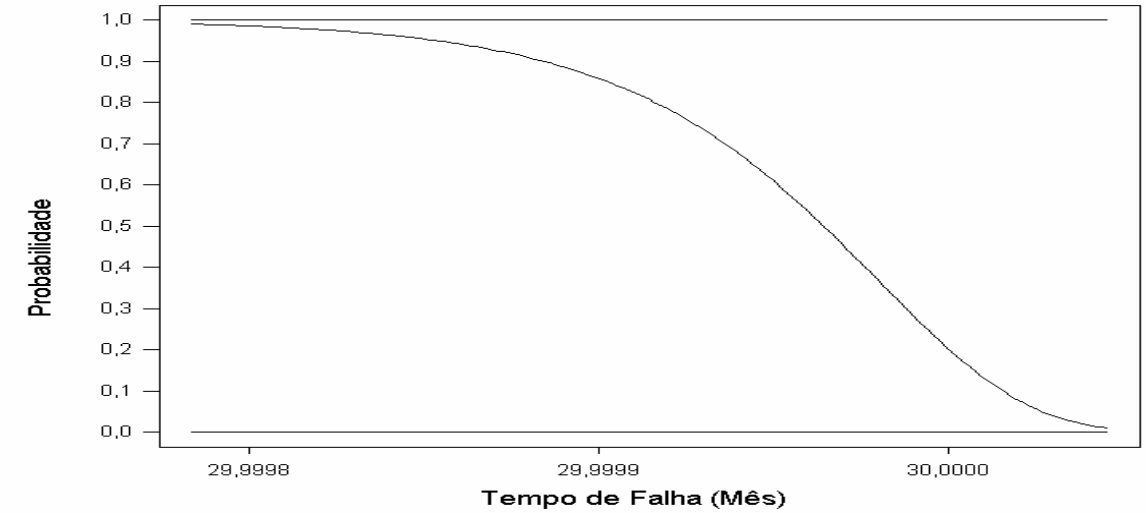


Figura 10: Sobrevivência dos Eucaliptos no Estrato 3

O valor mediano de sobrevida para o estrato 3 foi de 30 meses.

A Figura 11 mostra a curva de sobrevivência para as árvores do estrato 4

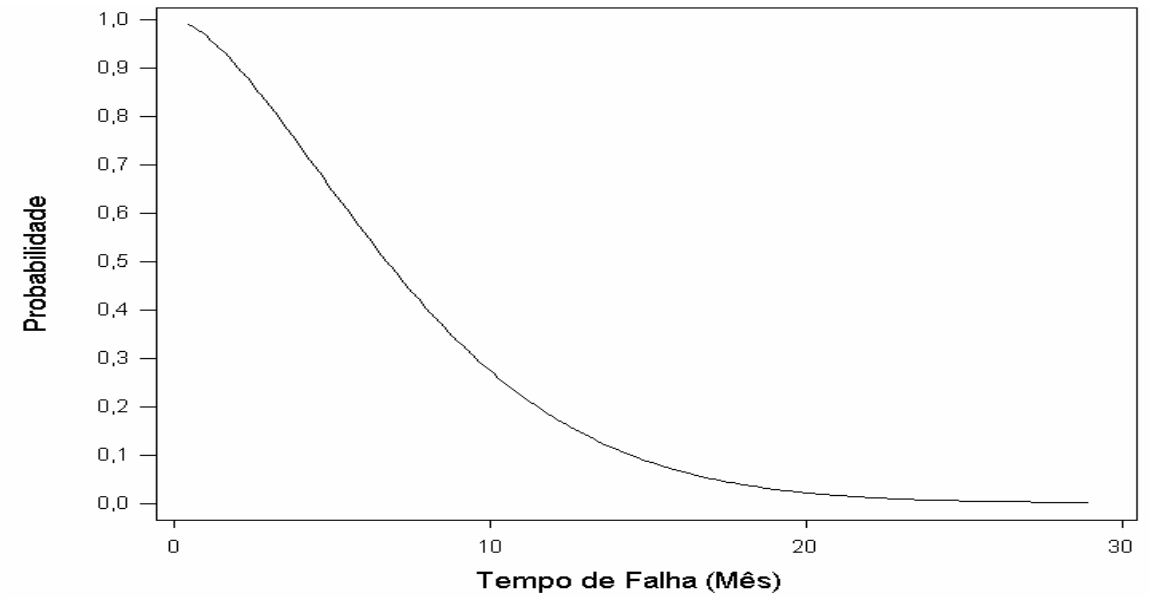


Figura 11: Sobrevivência dos Eucaliptos no Estrato 4

O valor mediano de sobrevida para o estrato 4 foi de 6, 72 meses.

Logo o estrato 4 apresentou valores mais baixos de sobrevivência. Apontando-se um cuidado maior nos primeiros 6 meses de plantio, pois apresentou uma mortalidade maior confrontada com os outros estratos nesse período.

A Figura 12 mostra o estrato 4 apresentando valores de sobrevivência abaixo dos demais estratos. Estrato 1(preto), estrato 2 (vermelho), estrato 3 (azul) e estrato 4 (verde).

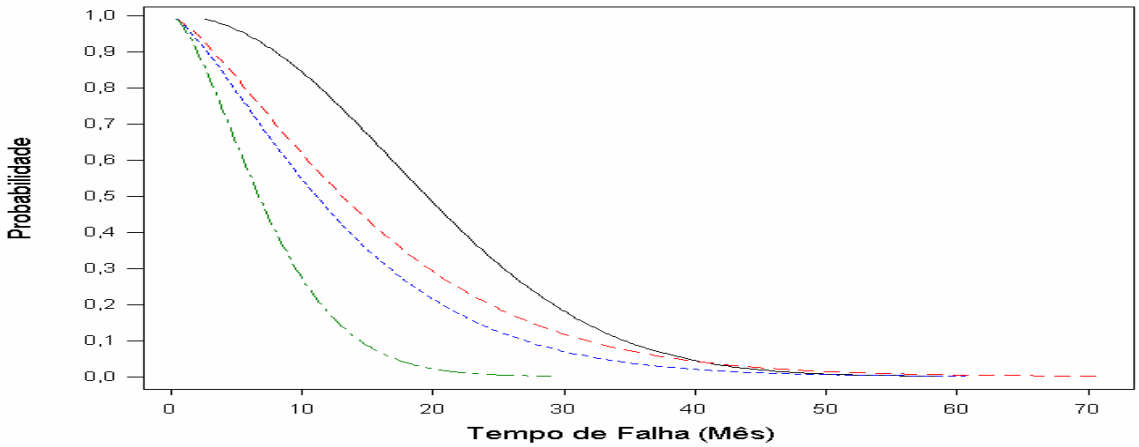


Figura 12: Sobrevivência dos Eucaliptos no Estrato 1,2,3 e 4

Nas figuras 13 e 14 mostra a função de probabilidade e a função de risco respectivamente. Na figura 14 observa-se o estrato 4 com uma curva mais inclinada, demonstrando um alto risco nos primeiros meses de plantio.

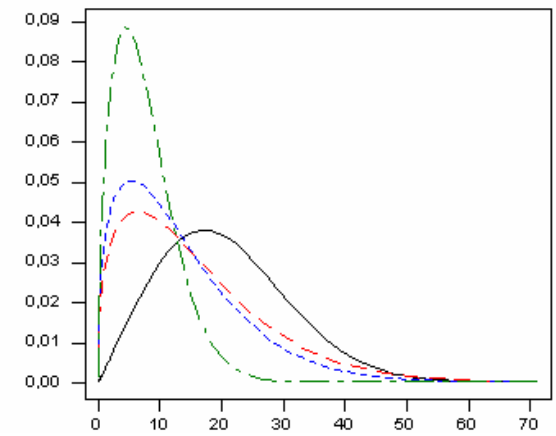


Figura 13: Função densidade de probabilidade

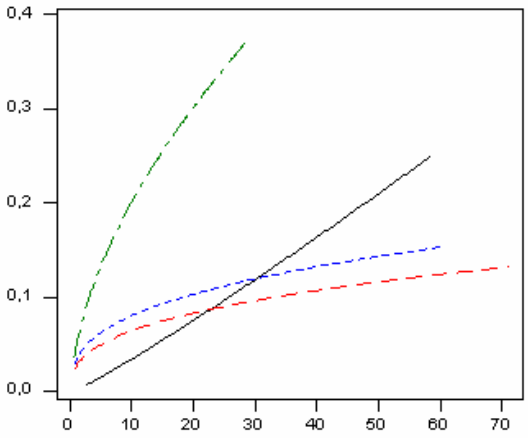


Figura 14: Função de risco

O teste de log rank, na tabela 3, apresentou o seguinte resultado para Comparação das Curvas de Sobrevivência dos estratos 1,2,3 e 4.

Tabela 3: Teste de log-rank para comparação dos estratos

Teste Estatístico			
Método	Qui-Quadrado	Graus de liberdade	P-Value
Log-Rank	10,84	3	0,0126

O teste *log-rank* foi aplicado para testar se as curvas diferiam entre categorias de uma mesma variável.

Considerando alfa igual a 0.05, o teste evidencia que há forte diferença significativa entre os estratos.

Para a construção de um modelo multivariado, utilizando a regressão de Cox, foram considerada os estratos.

Tabela 4: Regressão de Cox

Variável Resposta : Mês
Método de Estimação: Máxima Verossimilhança

Tabela de Regressão de Cox

Preditores	Coeficientes	Erro Padrão	Z	P	Intervalo de confiança a 95%	
					Lim. Inferior	Lim. Superior
Intercepto	-3.75	3.444	-1.09	0.276	-10.5	3
Estrato 1	0.2559	0.1313	1.95	0.051	-0.0015	0.5134
Estrato 2	-0.02129	0.04105	-0.52	0.604	-0.1017	0.05917
Estrato 3	0.0898	0.1018	0.88	0.378	-0.1098	0.2894
Estrato 4	-0.10663	0.06573	-1.62	0.105	-0.2355	0.02221
Forma	2.307	0.4865			1.526	3.488

Gráfico da função de risco para os estratos 1,2,3 e 4, figura 15.

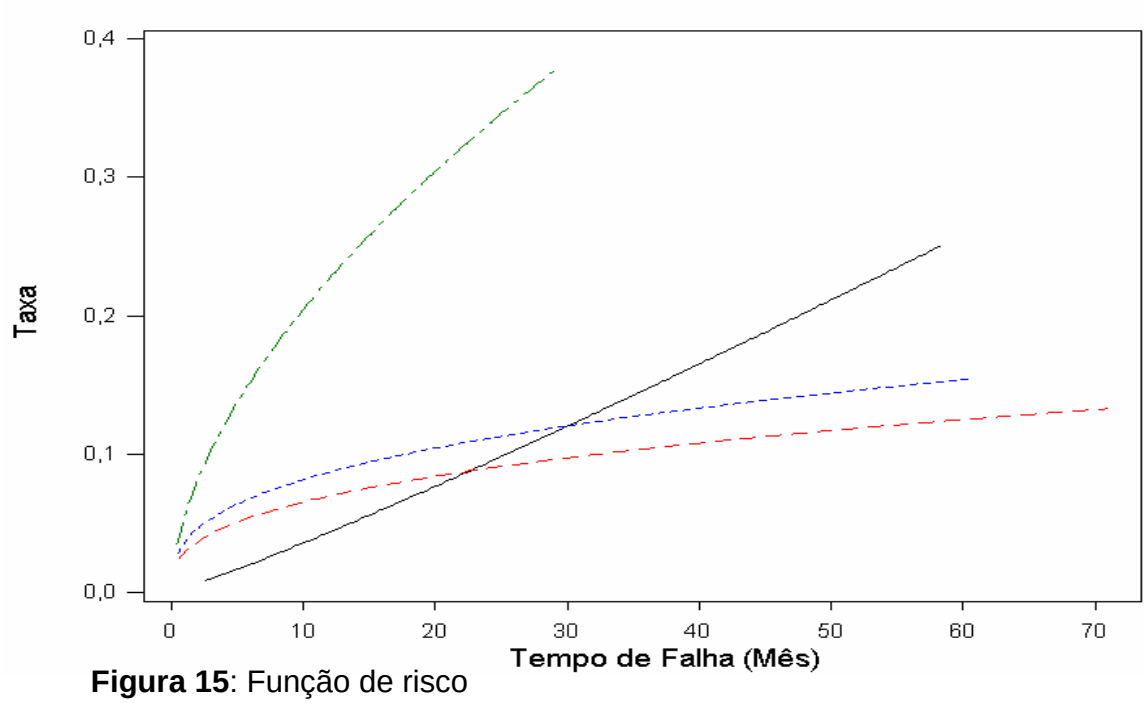


Figura 15: Função de risco

Gráfico da probabilidade considerando a distribuição Weibull dos estratos 1,2,3 e 4, na figura 16.

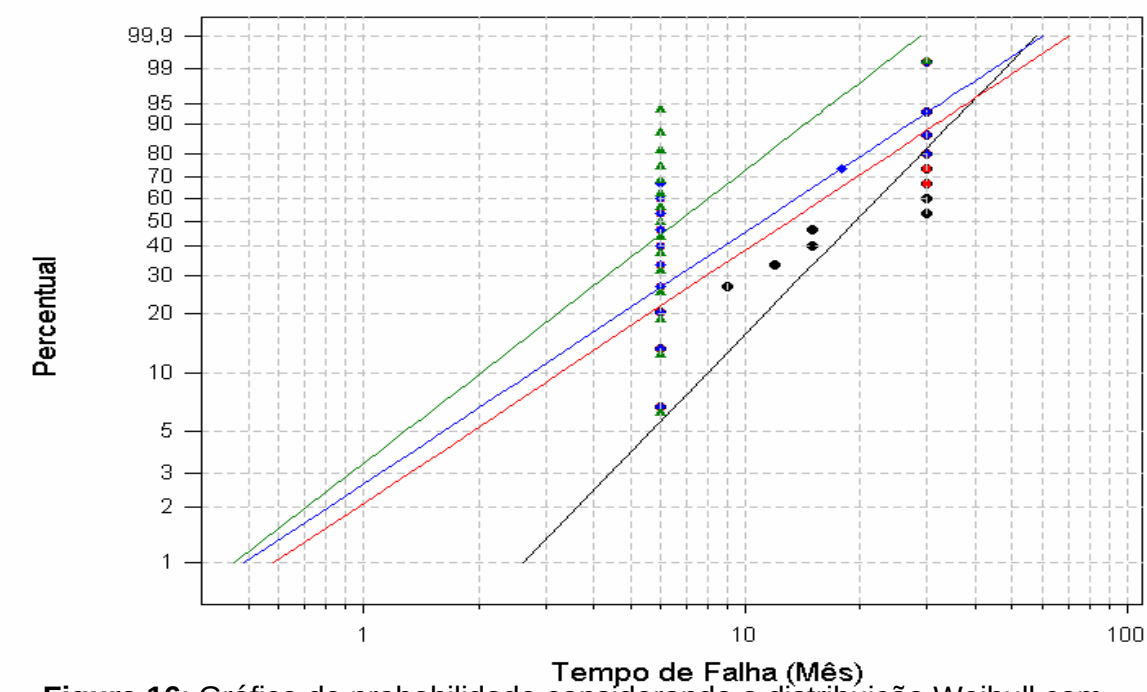


Figura 16: Gráfico de probabilidade considerando a distribuição Weibull com dados censurados.

Na Tabela 5, apresenta-se as probabilidades para o estrato 1.

Tabela 5: Sobrevivência para estrato 1

Estimador Kaplan-Meier

Tempo	Número de Falha	Probabilidade de Sobrevivência	95.0% Normal CI Inferior	Superior
6	3	0,8000	0,5976	1,0000
9	1	0,7333	0,5095	0,9571
12	1	0,6667	0,4281	0,9052
15	2	0,5333	0,2809	0,7858
30	8	0,0000	0,0000	0,0000

Na Tabela 6, apresenta-se as probabilidades para o estrato 2.

Tabela 6: Sobrevivência para estrato 2

Estimador Kaplan-Meier

Tempo	Número de Falha	Probabilidade de Sobrevivência	95.0% Normal CI Inferior	Superior
6	9	0,4000	0,1521	0,6479
30	6	0,0000	0,0000	0,0000

.

Na Tabela 7, apresenta-se as probabilidades para o estrato 3.

Tabela 7: Sobrevivência para estrato 3

Estimador Kaplan-Meier

Tempo	Número de Falha	Probabilidade de Sobrevivência	95.0% Normal CI Inferior	Superior
6	10	0,3333	0,0948	0,5719
18	1	0,2667	0,0429	0,4905
30	4	0,0000	0,0000	0,0000

.

Na Tabela 8, apresenta-se as probabilidades para o estrato 4.

Tabela 8: Sobrevivência para estrato 4

Estimador Kaplan-Meier

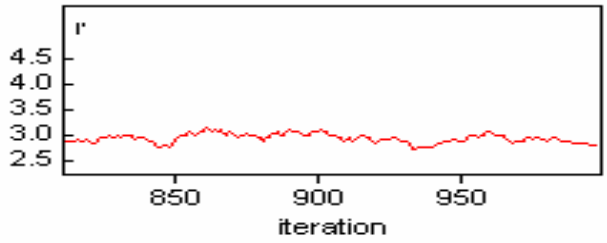
Tempo	Número de Falha	Probabilidade de Sobrevivência	95.0% Normal CI Inferior	Superior
6	15	0,0625	0,0625	0,1811
30	1	0,0000	0,0000	0,0000

.

O resultado do modelo inicial da regressão de Cox, expresso como função de risco (hazard fuction) , probabilidade de sobrevivência e respectivos intervalos de confiança, revela o estrato 1 como o que apresenta maior probabilidade de sobrevida em relação aos outros estratos.

Na análise bayesiana , verificou-se que a estimativa do parâmetro de forma da Weibull, convergiu com 1000 iterações, como mostra a tabela 9.

Tabela 9: Convergência do parâmetro com 1000 iterações



Na tabela 10, apresenta-se as estimativas do parâmetro de forma da Weibull, tanto pelo método clássico com pelo método bayesiano.

Tabela 10: Comparação dos métodos para o parâmetro de forma da Weibull.

Parâmetro	IC 95% para o parâmetro de forma							
	Bayesiana				Clássica			
	estimativa	Inferior	Superior	amplitude	estimativa	Inferior	Superior	amplitude
forma	3.513	2.856	4.146	1.290	2.307	1.526	3.4878	1.9618

Estimando o valor mediano para os extrato com abordagem bayesiana , após 1000 interações, encontramos os resultados apresentados na tabela 11:

Tabela 11: Valores medianos para os estratos

variável	mediana	inicio	interações
estrato1	29	1	1000
estrato2	29	1	1000
estrato3	30	1	1000
estrato4	5	1	1000

Utilizou-se como priori para estimação do parâmetro de forma da distribuição de sobrevivência uma gamma (1, 0.001) encontrando o valor de 3.515 como resultado, apresentado na tabela 12.

Tabela 12: Parâmetro de forma da Weibull

parâmetro	forma	início	interações
r	3.513	1	1000

6. CONCLUSÕES:

Concluí-se que, a melhor distribuição para analisar a população em questão é a Weibull, segundo o teste de Anderson-Darling e como método para estimação dos parâmetros da distribuição, tanto o método clássico como o método bayesiano, mostram-se bons estimadores, verificado pela amplitude dos intervalos de confiança a 95%. Quanto à função de risco e a probabilidade de sobrevivência verificou-se que dentre os estratos analisados o que teve maior probabilidade de sobrevivência foi o estrato 1 e o menor o estrato 4. Através da probabilidade de sobrevivência, o valor mediano aponta que deve-se ter um melhor controle nos primeiros 6 meses de plantio. O estrato 4 apresentou maior mortalidade nesse período confrontado com os outros estratos nesse período. A regressão de Cox, através da função de risco, apresenta uma curva mais inclinada, demonstrando um alto risco de morte nos primeiros meses de plantio.

Uma sugestão para próximas medições em relação aos Eucaliptos, seria a preocupação em coletar informações sobre causas de morte dessas árvores,como temperatura, clima, alguns componentes químicos que são adicionados ao solo, umidade do solo, e etc., para assim podermos efetuar uma competitividade entre os riscos de morte, e assim termos dados para estudarmos melhor a censura nesses casos.

7. BIBLIOGRAFIA:

ALBUQUERQU, P. A., **Diagnóstico Ambiental e Questões Estratégicas: Uma Análise considerando o Pólo Gesseuro do Sertão do Araripe – Estado de Pernambuco.** Tese (Doutorado em Ciências Florestais). Curitiba. UFPR, 2002. 185p.

ANDERSEN, P. K.. Survival analysis 1982-1991: The second decade of the proportional harzards regression model. *Statistics in Medicine*, 10:1931-1944., 1991.

ANDERSEN, P. K.; BORGAN, O.; GILL, R. D. & KEIDING, N.. *Statistical Models Based on Counting Process*. New York: Springer-Verlag**,** 1993

ANDERSON-DARLING, Disponível em: http://www.itl.nist.gov/div898/hanbook/eda/section3/eda35e.htm Acesso em: 11 de Fevereiro 2006.

ANTELMAN, GORDON. Elementary Bayesian Statistics, Edward Elgar Publishing Limited & Landsdown Place, Cheltenham Glas, UK, 1977.

BERGERUD, W. A.; REED, W.J. Bayesian statistical metholds. Chapter 7. in V. Sit, anb. Taylos (editor). 1998 Statistical methods for adaptive mana gement studies. Res. Br. B. C. Min. For., Res. Br. Victoria, BC, land Manager. Hamdb. No.1998.

BOX, G.E.P.; TIAO, G.C.Bayesian inference in statistical analysis. New York: Addison Wesley,. 588p., 1973.

BRASIL, G.H. Estatística Bayesiana: Alguns aspectos Teóricos e Práticos, in: Reunião Regional de Estatística, 22., 1991, Recife. Mini curso...Recife, UFPE, 1991. Apostila.

CHIANG, CL, Introduction to stochastic processes in biostatistics. New York: John Wiley 1968.

COLTON, T., *Statistica in Medicine*. Padova: Piccin Editore. 1979.

COX, D. R. Regression models and life-tables (with discussion). J. Royal Statistic Society Series B. n. 74, p. 187 – 220, 1972.

COX, D. R. & OAKES, D., *Analysis of Survival Data*. London: Chapman & Hall. 1984.

CROWLEY, J. & BRESLOW, N., Statistical analysis of survival data. *Annual Review of Public Health*, 5:385-411. 1994.

HARRIS, E. K. & ALBERT, A., *Survivorship Analysis for Clinical Studies.* New York: Marcel Dekker Incorporation. 1991

IBRAHIM, J. G.; CHEN, M. H.; SINHA, D. Bayesian Survival Analysis, Spring-erlag. New York. (2001).

INOJOSA, Josias, **Auto Custo do Óleo BPF Ressuscita o Uso da Lenha,** [ON LINE] Disponível na Internet via, URL http://www.sindusgesso.org.br. Arquivo capturado em 14 de fevereiro de 2006.

JEFFREYS, H> Theory of probability. 3ed., Oxford: Clarendon Press, 578p. 1961.

KAHN, H. A. & SEMPOS, C. T., 1989. *Statistical Methods in Epidemiology*. New York/Oxford: Oxford University Press.

KALBFLEISCH, J.G. Probability and Statistical Inference. 2ed. New York:Springer-Verlag. V.2: Statistical Inference, 1985.

KAPLAN, E. L. & MEIER, P., Non parametric estimation from incomplete observation. *Journal of the American Statistics Association*, 53:457-481. 1958.

KLEINBAUM, D. G., *Survival Analysis: A Self-Learning Text.* New York: Springer. 1995.

LATORRE, M. R. D. O., *Comparação entre Alguns Métodos Estatísticos em Análise de Sobrevivência: Aplicação em uma Coorte de Pacientes com Câncer de Pênis*. Tese de Doutorado, São Paulo: Faculdade de Saúde Pública, Universidade de São Paulo. 1996.

LEANDRO, R. A. Introdução à Estatística Bayesiana. Departamento de Ciências Exatas, ESALQ/USP Piracicaba, SP – 7 a 13/07/2001.

LEE, E. T., *Statistical Methods for Survival Data Analysis*. 2nd Ed. New York: John Wiley & Sons. 1992.

LOUZADA-NETO. F, MAZUCHELI J., ACHCAR J.A. Análise de Sobrevivência e Confiabilidade. III Jornada Regional de Estatística e II Semana da Estatística, Maringá, 2002.

MARTÍNEZ E ESPINOSA, M; CALIL JÚNIOR,C. Determinação do valor característico da resistência da madeira: Distribuições de probabilidades simétricas e assimétricas. Revista madeira: Arquitetura e Engenharia, V.1, n.2, p.25-30, 2000.

MARTINS, C. A. C. Análise de regressão logística. 53f. (Dissertação de Mestrado) – Centro de Ciências Exats e da Natureza, Universidade Federal de Pernambuco, Recife, 1998.

MILLER Jr., R. G., *Survival Analysis*. New York: John Wiley & Sons. 1981.

NIST – NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. Engineering statistics handbook and handbook of statistical methods. Sematech: NIST, 2002. Disponível em http://www.itl.nist.gov/div898/hanbook/date . Acesso em: 18 março 2005

POLLARD,William E., Baeysin statistics for evalution research: an introduction, Berverly Hills: sage publications, 1986

PONTES, Lúcia, **Pólo Gesseiro do Araripe recebe novos investimentos,** [ON LINE] Disponível na internet via, URL http://www.pe.gov/jornal/jor23/np01.htm. Arquivo capturado em 10 de fevereiro de 2006.

ROMEU, Jorge Luis, Anderson-Darling: A Goodness of fit test for Small Sample Assumptions. New York: RAC START, v. 10, n.5. 2003.

SANTOS, J.N.M. Abordagem Bayesiana do equilíbrio de Hardy-weinberg .52f. Dissertação (Mestrado) – Biometria, Universidade Federal de Pernambuco, Recife. 2001

SILVA, A. L. C. e SUÁREZ, G. P. Qué es la inferência bayesiana? JANO EMC. Viernes, v.58 n, 1338, p.65-66, mar.2000.

SILVA, João, **Microrregiões do Sertão,** [ON LINE] Disponível na internet via, URL http://www.pe.-az.com.br/regioes/sertao.htm. Arquivo capturado em 14 de fevereiro de 2006.

STEPHENS, M.A. EDF: statistics for goodness of fit and some comparisons. Journal of the American Statistical Association. V.69, p.730-737 . 1974.

SZKLO, M. & NIETO, F. J., *Epidemiology: Beyond the Basics*. Annapolis: Aspen Publishers. 2000.