FLÁVIA SÍLVIA CORRÊA TOMAZ

ANÁLISE DE AGRUPAMENTO PARA A AVALIAÇÃO DE IDENTIDADE DE MODELOS NÃO-LINEARES EM ANÁLISE DE SOBREVIVÊNCIA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

VIÇOSA

MINAS GERAIS – BRASIL

2009

Aos meus pais, Maria Helena e Expedito, e meus irmãos, Fernanda e Edevaldo, por sempre me apoiarem e me incentivarem a buscar meus ideais.

DEDICO!

AGRADECIMENTOS

Agradeço, primeiramente, a Deus por ter me dado forças ao longo dessa jornada.

Ao meu orientador, professor Luiz Alexandre Peternelli, pela paciência e dedicação durante o desenvolvimento deste trabalho.

Aos professores do Mestrado em Estatística Aplicada e Biometria.

Aos colegas de curso: Telma, Andréia, Bráulia, Alex, Danilo, Nilza, Thiago, Antônio, Willerson, Priscila, Ana Carolina e Moysés.

Ao Emmanuel Kennedy pelo auxílio nas pesquisas;

Aos funcionários do Departamento de Informática da Universidade Federal de Viçosa: Altino, Paulinho, Marisa e Eliana;

À FAPEMIG, pela bolsa concedida;

À professora Maria Cláudia pelo incentivo, carinho, pela amizade e por ter acreditado em mim.

À minha família e ao meu noivo pelo apoio e compreensão.

A todos que direta ou indiretamente contribuíram para a conclusão dessa etapa, os meus agradecimentos.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

RESUMO

TOMAZ, Flávia Sílvia Corrêa, M. Sc., Universidade Federal de Viçosa, fevereiro de 2009. **Análise de Agrupamento para a avaliação de identidade de modelos não-lineares em análise de sobrevivência.** Orientador: Luiz Alexandre Peternelli. Coorientadores: Fabyano Fonseca e Silva e Sebastião Martins Filho.

O objetivo desse trabalho foi comparar modelos não-lineares ajustados aos dados de sobrevivência de formigas submetidas a diferentes tratamentos através de metodologia alternativa. Essa metodologia consistiu no uso da técnica de análise de agrupamento, método de Ward, para a identidade de modelos usados em análise de sobrevivência. Os dados utilizados neste trabalho são referentes a um experimento realizado no laboratório de entomologia da Universidade Federal de Viçosa. Foi também utilizado um conjunto de dados simulado com base na distribuição de Weibull. Inicialmente aplicou-se técnica não paramétrica, estimador Kaplan-Meier, a fim de estimar as curvas de sobrevivência de cada tratamento e, em seguida, o teste *logrank* para a comparação dessas curvas. Para os dados reais foi ajustado o modelo logístico aos tempos de sobrevivência, enquanto que, para os dados simulados foi ajustado o modelo de Weibull. Para cada caso agrupou-se os parâmetros estimados de cada modelo utilizando-se as técnicas de análise de agrupamento. Os resultados encontrados pelo agrupamento foram equivalentes aos do teste *logrank*. Concluiu-se que a metodologia proposta mostrou ser eficiente e menos trabalhosa, quando várias curvas de sobrevivência precisam ser comparadas.

ABSTRACT

TOMAZ, Flávia Sílvia Corrêa, M. Sc., Universidade Federal de Viçosa, February, 2009. **Cluster analyses for evaluation of the identity of nonlinear models in survival analysis.** Adviser: Luiz Alexandre Peternelli. Co-advisers: Fabyano Fonseca e Silva and Sebastião Martins Filho.

The aim of this study was to compare non-linear models fitted to data on survival of ants under different treatments using alternative methodology. This methodology consists of using the technique of cluster analysis, Ward's method, to evaluate the identity of models used in survival analysis. The data used in this work are related to an experiment conducted in the entomology laboratory of the Federal University of Viçosa, Brazil. It also used a set of simulated data based on the Weibull distribution. Originally was applied a nonparametric technique, Kaplan-Meyer estimator, to estimate the survival curves, then the logrank test for comparison of these curves. For the real dataset it was fitted a logistic model of survival times, while for the simulated data it was fitted a Weibull model. The estimates of the parameters of each fitted model were grouped using the technique of cluster analysis. The results found by the grouping were equivalent to that by the logrank test. It is concluded that the proposed methodology showed to be efficient and less laborious, mainly when many survival curves need to be compared.

1. INTRODUÇÃO

A análise de sobrevivência é uma das áreas da Estatística que mais tem crescido nos últimos tempos. Essa análise é utilizada quando a variável resposta é o tempo decorrido até um evento de interesse, e tem como principal característica a presença de censura, sendo censura definida como observações incompletas ou parciais (KLEINBAUM; KLEIN, 2005).

Os conjuntos de dados envolvendo tempos de vida ou falha são representados por uma variável aleatória não negativa, usualmente contínua, e geralmente são especificados por sua função de sobrevivência (COLOSIMO; GIOLO, 2006).

A função de sobrevivência pode ser obtida através de estimadores não paramétricos, sendo o estimador Kaplan-Meier o mais conhecido. Pode-se também estimar a função de sobrevivência ajustando modelos paramétricos. Como geralmente os dados de sobrevivência tendem a ser assimétricos, a suposição de normalidade não é atendida e, assim sendo, as distribuições mais utilizadas para esses dados são a exponencial, a Weibull e a log-normal.

Na maioria dos experimentos planejados, o principal objetivo é comparar tratamentos. Na área de análise de sobrevivência, essa comparação é realizada através de procedimentos não paramétricos nos quais o grande destaque é dado ao teste *logrank*. Esse teste permite comparar duas ou mais curvas de sobrevivências relativas a diferentes tratamentos.

No caso de comparações múltiplas, isto é, de vários tratamentos cada qual representado por uma diferente curva de sobrevivência, atenção deve ser dada ao nível de significância. Como esse nível tende a aumentar com o número de pares de testes que são realizados simultaneamente, a utilização de correções é sugerida no nível de significância, a fim de evitar conclusões equivocadas. Todavia, quando o número de

pares de testes é elevado, as correções podem tornar o procedimento muito conservativo. Além disso, o número elevado de pares de tratamentos a serem testados torna o processo trabalhoso. Tais problemas poderiam afastar alguns pesquisadores do uso dessa técnica.

Portanto, o presente trabalho avaliou o uso da análise de agrupamento, método de Ward, na comparação de modelos paramétricos não lineares ajustados a certos dados de sobrevivência.

Utilizando dados reais e simulados sobre a sobrevivência de formigas submetidas a diferentes tratamentos, o objetivo do trabalho foi: 1) Aplicar o método de agrupamento de Ward para comparar os modelos ajustados aos dados de sobrevivência e 2) Avaliar a aplicabilidade do procedimento proposto para a comparação de modelos de sobrevivência.

2. REVISÃO DA LITERATURA

2.1. Análise de Sobrevivência

A análise de sobrevivência é uma coleção de procedimentos estatísticos para análise de dados cuja variável resposta é o tempo até a ocorrência de um evento de interesse. Esse tempo é denominado tempo de falha, de sobrevivência ou de vida. Esse tempo de falha pode ser o tempo até a morte de um paciente (indivíduo em estudo), cura ou recorrência da doença ou o tempo de falha de um componente mecânico ou eletrônico (COX; OAKES, 1984; LAWLESS, 1982).

Os dados de sobrevivência são caracterizados pela presença de censura. Nesse contexto, censura corresponde às observações incompletas ou parciais sobre os indivíduos. A ocorrência de dados censurados informa que o tempo de falha foi superior ao tempo observado no estudo ou experimento. Todos os dados, até mesmo os censurados, devem ser considerados na análise de sobrevivência, pois mesmo incompletas essas observações possuem informações sobre o tempo de vida do indivíduo e a não consideração das censuras no cálculo estatístico podem conduzir a conclusões viciadas (COLOSIMO; GIOLO, 2006).

Essas observações incompletas ocorrem por diversas razões. Dentre elas, a perda de acompanhamento do paciente no decorrer do estudo, a não ocorrência do evento de interesse até o final do experimento, a morte do indivíduo devido a uma causa diferente da estudada (LAWLESS, 1982).

As várias razões que geram a ocorrência de dados censurados conduzem a diferentes mecanismos de censura. Os tipos de censura são: a censura à direita, censura à esquerda e a censura intervalar. A censura à direita é assim chamada, pois o tempo de ocorrência do evento está à direita do último tempo de sobrevivência registrado. Esse tipo de censura engloba as censuras do tipo I, do tipo II e a aleatória. A censura do tipo I ocorre quando o estudo é terminado após um período pré-estabelecido de tempo. A censura do tipo II ocorre quando o estudo é terminado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos. A aleatória, também chamada de censura tipo III, é o tipo mais frequente e ocorre quando um indivíduo for retirado do estudo sem ter ocorrido o evento de interesse, ou vier a morrer por uma razão diferente da estudada. Na censura aleatória os tempos de entrada do indivíduo no estudo não são simultâneos. A censura à esquerda ocorre quando o tempo registrado é maior que o tempo de sobrevivência. A censura intervalar é a mais geral, e ocorre quando o acompanhamento do estudo é realizado a partir de visitas periódicas. Desse modo, o tempo de vida observado estará entre dois tempos sucessivos, não sendo conhecido o tempo exato do evento de interesse (LAWLESS; BABINEAU, 2006; COLLET, 1994).

Porém, poucos pacotes estatísticos acomodam dados com censura intervalar, assim, na prática é comum assumir que o evento de interesse tenha ocorrido no início, ponto médio ou final do intervalo; ou seja, ignorar a censura intervalar (COLLET, 1994). Desses mecanismos citados a censura à direita é a mais comum em dados de sobrevivência, e à esquerda ocorre com menor frequência (LAWLESS, 1982).

Uma análise estatística básica consiste em estudar os dados descritivamente, através das medidas de tendência central e de variabilidade. Contudo, a presença de dados incompletos traz problemas para essa análise estatística, inviabilizando esse tipo de análise. Logo, a análise descritiva envolvendo dados de vida consiste em estimar a função de sobrevivência e, a partir dela, estimar quantidades de interesse como o tempo médio, tempo mediano e alguns percentis (COLOSIMO; GIOLO, 2006).

Para a estimação da função de sobrevivência têm-se os estimadores não paramétricos de Kaplan-Meier, tabela de vida ou atuarial e o estimador de Nelson-Aalen (COLOSIMO; GIOLO, 2006).

O estimador de Kaplan-Meier é uma importante ferramenta para a análise de dados censurados, visto que é facilmente calculado e requer poucas suposições (EFRON, 1988). Ele é também conhecido como estimador produto-limite e é usado com frequência em estudos biométricos (COLOSIMO *et al*., 2000). As suposições para

o uso do estimador de Kaplan-Meier são: i) os participantes do estudo precisam ser independentes, ou seja, cada participante deve aparecer somente uma vez no seu grupo; ii) os grupos devem ser independentes, isto é, cada participante pertence a um único grupo, iii) a medida para o tempo deve ser precisa; iv) o ponto de início do evento deve ser definido claramente; v) a chance de sobrevivência dos participantes permanece constante. Assim, os indivíduos que iniciam o estudo em tempos diferentes têm a mesma chance de sobrevivência (PEAT; BARTON, 2005).

Frequentemente a escala para medida do tempo é o tempo de relógio ou real, porém outras possibilidades existem, tais como o tempo de operação de um sistema e a quilometragem de um carro (COX; OAKES, 1984).

No que diz respeito à definição da data de início do experimento, deseja-se que os indivíduos sejam comparáveis na origem do estudo. Em experimentos clínicos aleatorizados, a data de aleatorização é uma opção natural. Outras escolhas possíveis são a data do início do tratamento de uma doença ou do diagnóstico (COLOSIMO; GIOLO, 2006).

Além das suposições enumeradas acima, deve-se atentar para a clara especificação do evento de interesse ou falha. Em estudos médicos a falha pode significar morte a partir de uma causa específica, primeira recorrência de uma doença após o tratamento ou a incidência de uma nova doença. No contexto industrial, a falha é definida como o primeiro instante no qual o desempenho, medido de alguma maneira quantitativa, torna-se inferior a um nível aceitável definido pela especificação (COX; OAKES, 1984).

Conforme Colosimo e Giolo (2006) os estimadores se diferenciam no número de intervalos utilizados para a construção de cada um deles, sendo que o estimador de Kaplan-Meier e o de Nelson-Aalen baseiam-se em números de intervalos iguais ao número de falhas distintas. Para o estimador tabela de vida, os tempos são agrupados de forma arbitrária. Há estudos que mostram a superioridade do estimador de Kaplan-Meier, de modo que esse tem sido o mais utilizado em pesquisas clínicas e vem ganhando mais espaço na área de confiabilidade (COLOSIMO; GIOLO, 2006). Como exemplo desses estudos, pode-se citar o trabalho de Pereira e Pereira (2003) que compararam as estimativas dos estimadores de Kaplan-Meier e Nelson-Aalen em uma situação de introdução de censura do tipo I em um estudo entomológico, concluindo que o estimador Kaplan-Meier é mais preciso em relação ao de Nelson-Aalen.

A partir da curva de Kaplan-Meier é possível se obter estimativas de percentis. Pode-se também obter o valor de interesse diretamente das estimativas de Kaplan-Meier, ou por interpolação, caso o valor do tempo de interesse, por exemplo, o tempo mediano, estiver ao longo do degrau da curva de Kaplan-Meier (COLOSIMO; GIOLO, 2006). Segundo Collet (1994) o tempo mediano é definido como o menor tempo observado, para o qual o valor da função de sobrevivência estimada é menor ou igual a 0,5. Mas, como a curva de Kaplan-Meier é uma função escada, Colosimo e Giolo (2006) destacam que a estimativa mais adequada para a mediana é obtida por interpolação.

Outra medida de interesse nesses estudos é o tempo médio de vida. Essa quantidade é dada pela integral sob a curva de sobrevivência. Porém, quando o maior tempo observado é uma censura a estimativa do tempo médio fica subestimada e, neste caso, a mediana é preferida (COLOSIMO; GIOLO, 2006).

Além do estudo descritivo dos dados, pode-se ter como objetivo a comparação de vários tratamentos. Isto é, possivelmente existirá para cada tratamento uma curva de sobrevivência que, posteriormente, serão comparadas entre si. Para a comparação de duas ou mais curvas existem alguns testes não paramétricos, dos quais os mais difundidos são: o teste *logrank* e o teste de Wilcoxon. O teste *logrank* é o mais utilizado para a comparação de curvas de sobrevivência, quando a razão das funções de risco dos grupos é proporcional. Para as situações onde a funções de risco não são proporcionais, o teste de Wilcoxon é mais apropriado (COLLET, 1994).

Pereira e Vivanco (2002) avaliaram a aplicação do teste *logrank* à medida que se introduz porcentagens de censura, concluindo que a presença maior de censura pode fazer com que o teste *logrank* apresente resultados incorretos.

Além do tratamento não paramétrico, numerosos modelos são usados na análise dos tempos de sobrevivência (LAWLESS, 1982). Os dados de sobrevivência não são, geralmente, distribuídos de modo simétrico. Eles tendem a ser assimétricos positivamente. Por essa razão não é razoável assumir que os dados deste tipo sigam uma distribuição normal. As distribuições que ocupam lugar de destaque são a exponencial, a de Weibull e a log-normal (COLOSIMO; GIOLO, 2006). Lawless (1982) cita ainda a distribuição gama como pertencente a essa estreita categoria. Essas distribuições são consideradas importantes, pois se demonstram úteis na modelagem de várias situações (COLOSIMO; GIOLO, 2006; LAWLESS, 1982).

A distribuição exponencial é a mais simples para descrever o tempo de falha. Ela é caracterizada por uma taxa de risco constante e por apenas um parâmetro, que é a própria média da distribuição. A distribuição exponencial tem sido bastante usada para descrever tempos de falhas de sistemas eletrônicos e o tempo de vida de óleos isolantes e dielétricos (KLEINBAUM; KLEIN, 2005).

A distribuição de Weibull é talvez a mais utilizada para modelar tempos de vida. Ela é a generalização da distribuição exponencial, mas não assume uma taxa de risco constante. Sua taxa de risco é monótona. Essa distribuição tem ampla aplicabilidade, pois apresenta uma grande variabilidade de formas. A distribuição de Weibull tem sido usada em estudos de mortalidade causada por doenças humanas, em estudos biomédicos, industriais e entomológicos (COLOSIMO; GIOLO, 2006). Como exemplo da utilização do modelo de Weibull em estudos entomológicos pode-se citar: a análise de sobrevivência do ácaro predador *Iphiseiodes zuluagai Denmark & Muma* (*Acari: Phytoseiidae*) (REIS; HADDAD, 1997); o estudo de mortalidade de abelhas operárias alimentadas com pasta de cândi contendo diferentes teores de tanino (SANTORO *et al*., 2004); o estudo de sobrevivência de abelhas (Appis mellifera) tratadas com diferentes dietas (GUIMARÃES *et al*., 2004) e a análise de sobrevivência de operárias de Atta sexdens rubropilosa Forel isoladas do formigueiro e alimentadas com dietas artificiais (BUENO *et al*., 1997).

A distribuição log-normal é muito utilizada para caracterizar tempos de vida de produtos (fadiga de metal, semicondutores, diodos e isolação térmica) e de indivíduo (tempo de vida de pacientes com leucemia) (COLOSIMO; GIOLO, 2006). A distribuição log-normal generalizada foi proposta como alternativa aos modelos de Cox e Logístico para modelagem de dados grupados (SILVEIRA *et al*., 2003)

A distribuição gama, assim como a de Weibull, inclui a distribuição exponencial como caso particular. Ela é utilizada como modelo para os problemas de confiabilidade e sobrevivência humana na área médica (LEE; WANG, 2003).

A escolha do modelo mais adequado para o conjunto de dados é tópico importante, e pode ser realizada através de técnicas gráficas ou através de testes de hipóteses em modelos encaixados (COX; HINKLEY, 1974). A utilização de técnicas gráficas é a forma mais simples e eficiente para seleção de modelos (COLOSIMO; GIOLO, 2006).

2.2. Análise de Agrupamento

A análise de agrupamento teve origem nos campos da ciência, tais como taxonomia e psicologia, e tem recebido nas últimas décadas considerável atenção na literatura estatística (GNANADESIKAN, 1997)

A análise de agrupamento, também conhecida como análise de conglomerado ou cluster analysis tem como objetivo dividir um conjunto de observações (elementos, indivíduos, tratamentos, genótipos, etc.) em grupos homogêneos ou compactos, segundo algum critério conveniente de similaridade. Assim, os elementos pertencentes a um mesmo grupo serão homogêneos (similares) entre si, com respeito às certas características medidas, enquanto que os pertencentes a grupos diferentes deverão ser heterogêneos entre si em relação às mesmas características (MINGOTI, 2005; SHARMA, 1996).

Essa análise está presente em estudos de diversas áreas, por exemplo, em ecologia, psicologia, pesquisa de mercado, geoquímica, ergonomia, geografia, medicina, psiquiatria, sociologia, geologia, sensoriamento remoto, economia e engenharia (MINGOTI, 2005; RENCHER, 2002).

A análise de agrupamento foi utilizada por Souza *et al*. (2005) e Rodrigues *et al*. (2002) para estudos de divergência genética. Por Melo Júnior *et al*. (2005) e Lyra *et al*. (2006) em estudos de climatologia. Matos Júnior *et al*. (1999) compararam curvas de maturação de laranjas e Peternelli *et al*. (2005) compararam modelos logísticos através da análise de agrupamento.

Para iniciar o processo de agrupamento, é necessário decidir até que ponto dois elementos do conjunto de dados podem ser considerados como semelhantes (MINGOTI, 2005). Para isso é preciso determinar uma medida de proximidade. Essas medidas podem ser classificadas como medidas de similaridade ou dissimilaridade (KHATREE; NAIK, 2000). Nas medidas de dissimilaridade quanto maior o valor, menos parecidas são as observações; enquanto que nas medidas de similaridade um valor alto indica maior semelhança entre as observações. As medidas de similaridade são classificadas em coeficientes de correlação e coeficientes de associação (SHARMA,

1996). Ressalta-se, porém, que a maioria dos algoritmos trabalha com o conceito de dissimilaridade, ou seja, de distância (MARDIA *et al*., 1997).

As principais medidas de dissimilaridade são a distância euclidiana, a distância euclidiana média e a distância de Mahalanobis (KHATREE; NAIK, 2000; MINGOTI, 2005; CRUZ; CARNEIRO, 2006).

O número de estimativas de dissimilaridade é, de modo geral, relativamente grande, o que inviabiliza o reconhecimento de grupos homogêneos através de procedimentos gráficos. Assim, para a realização dessa tarefa, utilizam-se as técnicas de agrupamento (SHARMA, 1996).

Dos métodos de agrupamento, os mais utilizados são os hierárquicos e os de otimização. Dos métodos de otimização, o utilizado com maior frequência pelos melhoristas é o método de Tocher (CRUZ & CARNEIRO, 2006). O método alternativo derivado do método de Tocher é o método de Tocher sequencial, o qual foi proposto por Vasconcelos *et al*. (2007).

Frequentemente pode-se encontrar na literatura estatística a classificação dos métodos como hierárquico e não-hierárquico. Os métodos hierárquicos são classificados em aglomerativos e divisivos (MARDIA *et al*., 1997).

O agrupamento hierárquico talvez seja o método mais antigo. Nas técnicas hierárquicas, inicialmente, cada observação é considerada como um grupo isolado, e esses grupos são unidos em cada passo do processo, formando novos grupos até que se tenha um único grupo (KHATREE; NAIK, 2000; MINGOTI, 2005).

Os métodos hierárquicos mais comuns para o agrupamento de dados são: o método do vizinho mais próximo (single linkage), método do vizinho mais distante (complete linkage), método da média das distâncias (average linkage), método do centróide (centroid method) e o método de Ward (Ward's method) (MINGOTI, 2005; SHARMA, 1996).

O método de Ward foi proposto por Ward (1963) e é também chamado de "Mínima Variância" (MINGOTI, 2005). Nesse método a formação dos grupos se dá pela maximização da homogeneidade dentro dos grupos. A soma de quadrados dentro dos grupos é usada como medida de homogeneidade. Isto é, o método de Ward tenta minimizar a soma de quadrados dentro do grupo. Os grupos formados em cada passo são resultantes de grupo solução com a menor soma de quadrados (SHARMA, 1996).

Uma questão de importância é como se deve proceder para escolher o número de grupos que define a partição do conjunto de dados analisados. Existem algumas

medidas estatísticas que podem ajudar a responder a essa questão. Tais medidas são a soma de quadrados entre grupos, a correlação semiparcial, o desvio padrão conjunto de todas as variáveis que formam o grupo e a distância entre os grupos (KHATREE; NAIK, 2000; MINGOTI, 2005; SHARMA, 1996). Outros critérios como a análise do nível de similaridade, da estatística pseudo F, da estatística pseudo T^2 e do Cubic Clustering Criterium (CCC) podem auxiliar na definição do número final de grupos (MINGOTI, 2005). Chae *et al*. (2005) propõem um método para predizer o número de grupos utilizando a estatística de Rand.

Resumidamente a análise de agrupamento reúne observações dentro de grupos tal que cada grupo seja o mais homogêneo possível. Esse processo pode ser sintetizado em cinco etapas. A primeira é a escolha da medida de dissimilaridade, a seguinte é a escolha do método de agrupamento (hierárquico ou não-hierárquico). O terceiro passo é a escolha do tipo de agrupamento para o método escolhido seguido pela decisão sobre o número de grupos, e finalmente a interpretação do resultado do agrupamento (SHARMA, 1996; GNANADESIKAN, 1997).

3. MATERIAL E MÉTODOS

3.1. Dados

Para esse estudo foram usados dados reais (referenciado como conjunto de dados 1) e dados simulados (referenciado como conjunto de dados 2), que serão descritos a seguir.

3.1.1. Conjunto de dados 1

Os dados são provenientes de um experimento conduzido no laboratório de entomologia da Universidade Federal de Viçosa. Uma breve descrição se faz necessária para melhor entendimento do experimento.

Neste experimento foram utilizadas oito colônias de formigas cortadeiras: duas receberam folhas de acalifa como substrato para o crescimento do fungo, duas colônias receberam folhas de alfeneiro, duas receberam folhas de eucalipto, e as outras duas um substrato misto composto pelos três tipos de plantas citados acima. Desejou-se avaliar a influência do lixo da colônia na mortalidade de formigas. Para cada colônia foram utilizadas 20 placas de petri, sendo 10 placas contendo lixo do formigueiro e 10 sem lixo. Em cada placa foram colocadas 10 formigas operárias. As placas foram deixadas em local apropriado dentro do laboratório por cerca de 30 dias. A cada dia era observado e registrado o instante no qual ocorreu o evento de interesse. Neste caso, a morte da formiga. Os tratamentos foram designados da seguinte forma:

Tratamento 1: acalifa com a presença de lixo;
Tratamento 2: acalifa sem a presença de lixo;
Tratamento 3: alfeneiro com a presença de lixo;
Tratamento 4: alfeneiro sem a presença de lixo;
Tratamento 5: eucalipto com a presença de lixo;
Tratamento 6: eucalipto sem a presença de lixo;
Tratamento 7: substrato misto com a presença de lixo;
Tratamento 8: substrato misto sem a presença de lixo.

3.1.2. Conjunto de dados 2

Para comparação entre os resultados do teste *logrank* e o método de agrupamento de Ward, um segundo conjunto de dados foi simulado e analisado. Os dados representando tempos de vida foram simulados a partir de uma distribuição de Weibull. Para a simulação utilizou-se o software R® (R DEVELOPMENT CORE TEAM, 2008). Definiu-se como parâmetros para a simulação, os parâmetros do modelo de Weibull estimados a partir do conjunto 1. Utilizou-se o mecanismo de censura tipo I, estabelecendo o tempo do experimento de 30 dias. Formigas que não morreram neste tempo foram consideradas observações censuradas. Foram simulados seis tratamentos compostos por duzentas observações cada, conforme segue:

Tratamento 1: T ~ Weibull ($\gamma = 1{,}75;\ \alpha = 12{,}61$)

Tratamento 2: T ~ Weibull ($\gamma = 2{,}55;\ \alpha = 16{,}69$)

Tratamento 3: T ~ Weibull ($\gamma = 1{,}86;\ \alpha = 10{,}53$)

Tratamento 4: T ~ Weibull ($\gamma = 2{,}65;\ \alpha = 20{,}13$)

Tratamento 5: T ~ Weibull ($\gamma = 2{,}31;\ \alpha = 6{,}46$)

Tratamento 6: T ~ Weibull ($\gamma = 2{,}05;\ \alpha = 11{,}6$)

Onde γ é o parâmetro de forma, e α o de escala, ambos são positivos (COLOSIMO; GIOLO, 2006).

A função para a simulação dos dados pode ser encontrada no Apêndice A.

3.2. Métodos de análise

Serão apresentados os métodos estatísticos utilizados nas análises dos dados. As análises estatísticas foram feitas utilizando o software estatístico R® (R DEVELOPMENT CORE TEAM, 2008).

3.2.1. Estimador Kaplan-Meier

Para a estimação das curvas de sobrevivência para os tratamentos utilizou-se o estimador Kaplan-Meier, visto que, este método é o mais utilizado para estimação das curvas de sobrevivência.

O estimador de Kaplan-Meier é também chamado de estimador limite-produto (COX; OAKES, 1984). Conforme Colosimo e Giolo (2006) este estimador é uma adaptação da função de sobrevivência empírica que, na ausência de censura é dada por:

$$\hat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo t}}{\text{nº total de observações no estudo}}$$

Para a definição do estimador de Kaplan-Meier, consideremos, conforme descrito por Colosimo e Giolo (2006), que existam n pacientes no estudo e k ($\leq$ n) falhas nos tempos distintos e ordenados de falha, isto é, $t_1 < t_2 < \ldots < t_j$. Considerando S(t) uma função discreta com probabilidade maior que zero somente nos tempos de falha t_j, j = 1, 2, . . . , k, tem-se que:

$$S(t_j) = (1-q_1)(1-q_2)\ldots(1-q_j) \quad (1)$$

onde q_j é a probabilidade do indivíduo morrer no intervalo $[t_{j-1}, t_j)$ sabendo que ele não morreu até t_{j-1} e considerando $t_0 = 0$, isto é,

$$q_j = P(T \in [t_{j-1}, t_j) \mid T \geq t_{j-1})$$

Assim, conforme Collet (1994) e Colosimo e Giolo (2006) q_j é estimado por:

$$\hat{q}_j = \frac{d_j}{nj} \qquad (2)$$

onde d_j é o número de falhas (ou mortes) em t_j para j = 1,2, . . . , k e n_j é o número de indivíduos sob risco em t_j, isto é, aqueles que não falharam e não foram censurados até o instante anterior a t_j.

Assim de (1) e (2) temos que o estimador de Kaplan-Meier é dado por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right) \quad (3)$$

O estimador de Kaplan-Meier se reduz a função empírica quando os dados não apresentam censura (COLOSIMO e GIOLO, 2006). Ele também manterá a forma da função empírica quando os dados apresentarem censura do tipo I e II (COLOSIMO; GIOLO, 2006). Se o maior tempo de sobrevivência for uma observação censurada, a função de sobrevivência $\hat{S}(t)$, não atingirá o valor zero, ou seja, $\hat{S}(t) \neq 0$ (COLLET, 1994; COLOSIMO; GIOLO, 2006).

Para a construção de intervalos de confiança, a fim de testar hipóteses para S(t) é necessário avaliar a precisão do estimador de Kaplan-Meier (COLOSIMO, 2006). A variância assintótica do estimador é dada por:

$$\hat{\mathrm{Var}}\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j\left(n_j - d_j\right)} \qquad (4)$$

A expressão (4) é conhecida como fórmula de Greenwood (COLLET, 1994; COLOSIMO; GIOLO, 2006, COX; OAKES, 1984). Detalhes das derivações realizadas para obtenção da fórmula de Greenwood podem ser encontrados em Collet (1994, p.22-23).

Conforme Colosimo e Giolo (2006), $\hat{S}(t)$, para t fixo, segue uma distribuição assintótica normal, e assim, um intervalo aproximado de $100(1-\alpha)\%$ de confiança para S(t) é dado por:

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\hat{V}ar\left(\hat{S}(t)\right)}$$

onde $\frac{\alpha}{2}$ denota o $\frac{\alpha}{2}$ percentil da distribuição normal padrão.

Um problema surge para valores próximos de zero ou um; valores extremos. Para estes valores o intervalo de confiança pode apresentar limite inferior negativo ou superior maior que um (COLOSIMO; GIOLO, 2006). Nestes casos, um procedimento alternativo é utilizar a transformação para S(t), como por exemplo, $\hat{U}(t) = \log[-\log(\hat{S}(t))]$ sugerida por Collet (1996) e Kalbfleish e Prentice (1980) citado por Colosimo e Giolo (2006) que tem variância assintótica dada por:

$$\hat{V}ar\left(\hat{U}(t)\right) = \frac{\sum_{j:t_j<t} \frac{d_j}{n_j\left(n_j - d_j\right)}}{\left[\sum_{j:t_j<t} \log\left(\frac{n_j - d_j}{n_j}\right)\right]^2} = \frac{\sum_{j:t_j<t} \frac{d_j}{n_j\left(n_j - d_j\right)}}{\left[\log \hat{S}(t)\right]^2}$$

Assim, um intervalo de confiança de $100(1-\alpha)\%$ de confiança para S(t) é dado por:

$$\left[\hat{S}(t)\right]^{\exp\left\{\pm z_{\alpha/2}\sqrt{\hat{V}ar\left(\hat{U}(t)\right)}\right\}}$$

3.2.2. Estimação da mediana

Como o passo inicial de qualquer análise estatística consiste em uma descrição dos dados, estimou-se a mediana para cada tratamento. Essa medida descritiva foi obtida de duas maneiras: utilizando interpolação linear e conforme método descrito por Collet (1994). Optou-se pela mediana, pois a distribuição dos tempos de sobrevivência tende a ser assimétrica, logo a mediana é a medida de posição preferida (COLLET, 1994).

A mediana pode ser obtida diretamente da curva de Kaplan-Meier ou por interpolação linear, caso o tempo de interesse se encontre ao longo do degrau da curva

de Kaplan-Meier. A variância assintótica do estimador de percentis $\left(\hat{t}_p\right)$, conforme (COLOSIMO; GIOLO, 2004) é dada por:

$$\text{Var}\left(\hat{t}_p\right) = \frac{\text{Var}\left(\hat{S}\left(\hat{t}_p\right)\right)}{\left[f\left(\hat{t}_p\right)\right]^2} \quad (5)$$

Porém, conforme Colosimo e Giolo (2006) a utilização da equação (5) é inviabilizada devido à dificuldade de se obter uma estimativa para $f\left(\hat{t}_p\right)$.

Por sua vez, Collet (1994) define a mediana, t_{50}, como o menor tempo para o qual o valor da função de sobrevivência é menor ou igual à 0.5, isto é:

$$\hat{t}_{50} = \min\left\{ t_i \mid \hat{S}(t_i) \leq 0{,}5 \right\}$$

onde t_i é o tempo de sobrevivência para o i-ésimo indivíduo, i = 1, 2, ..., n. Assim, o intervalo de confiança aproximado para a mediana é dado:

$$\hat{t}_{50} \pm z_{\frac{\alpha}{2}} s.e\left(\hat{t}_{50}\right)$$

onde:

$\hat{t}_{50}$ - mediana estimada;

$s.e\left(\hat{t}_{50}\right)$ - erro padrão da mediana dada por:

$$\text{s.e}\left(\hat{t}_{50}\right) = \frac{1}{\hat{f}\left(\hat{t}_{50}\right)} \text{s.e.}\left[\hat{S}\left(\hat{t}_{50}\right)\right].$$

O erro padrão de $\hat{S}\left(\hat{t}_{50}\right)$, $\text{s.e.}\left[\hat{S}\left(\hat{t}_{50}\right)\right]$ é calculado usando a fórmula de Greenwood, dada pela equação (4), para o erro padrão da estimativa da função de sobrevivência de Kaplan-Meier. Uma estimativa para $\hat{f}\left(\hat{t}_{50}\right)$ é dada por:

$$\hat{f}\left(\hat{t}_{50}\right) = \frac{\hat{S}\left(\hat{u}_{50}\right) - \hat{S}\left(\hat{l}_{50}\right)}{\hat{l}_{50} - \hat{u}_{50}},$$

onde:

$$\hat{u}_{50} = \max\left\{t_j \mid \hat{S}\left(t_j\right) \geq 0{,}5 + \varepsilon\right\},$$

$$\hat{l}_{50} = \min\left\{t_j \mid \hat{S}\left(t_j\right) \leq 0{,}5 - \varepsilon\right\},$$

para i = 1, 2, ... n e valores pequenos de ε. Em muitos casos usa-se $\varepsilon = 0{,}05$, porém valores maiores de ε serão necessários se $\hat{u}_{50}$ e $\hat{l}_{50}$ se tornarem iguais (COLLET, 1994).

3.2.3. Teste *logrank*

Para comparação dos tratamentos foi utilizado o teste *logrank*. Inicialmente verificou-se a hipótese de igualdade de todos os tratamentos. Sendo constatada a presença de diferenças entre os tratamentos, identificaram-se quais tratamentos diferiam entre si. Para identificar os tratamentos que diferiam entre si foi realizada comparações entre tratamentos, dois a dois, com correção de Bonferroni para controlar o erro tipo I. Segue adiante definição do teste *logrank*.

Segundo Collet (1994) a definição do teste *logrank* para comparação de duas ou mais curvas de sobrevivência é dada como segue: suponha que g $\left(g \geq 2\right)$ funções de sobrevivência devam ser comparadas. Seja U a estatística que compara o número de mortes (falhas) nos 1, 2, ... , g-1 grupos com seus respectivos valores esperados. Matematicamente temos:

$$U_{LK} = \sum_{j=1}^{r}\left(d_{kj} - n_{kj}d_j n_j^{-1}\right),$$

para k = 1, 2, ... g-1. A quantidade U_{LK} será expressa na forma de um vetor com $\left(g-1\right)$ componentes, que será denotada por U_L. A covariância entre U_{LK} e $U_{LK'}$ é dada por:

$$V_{LKK'} = \sum_{j=1}^{r}\frac{n_{kj}d_j\left(n_j - d_j\right)}{n_j\left(n_j - 1\right)}\left(\delta_{kk'} - \frac{n_{kj}}{n_j}\right)$$

Para k, k' = 1, 2, ..., g-1 onde $\delta_{kk'}$ é tal que:

$$\delta_{kk'} = \begin{cases} 1 \text{ se } K = K' \\ 0 \text{ se } K \neq K' \end{cases}$$

Esses termos são então colocados na forma de matriz de variância e covariância, V_L, que é simétrica.

$$V_L = \begin{bmatrix} V_{L11} & V_{L12} & \cdots & V_{L1(g-1)} \\ V_{L21} & V_{L22} & \cdots & V_{L2(g-1)} \\ \vdots & \vdots & \vdots & \vdots \\ V_{L(g-1)1} & \cdots & \cdots & V_{L(g-1)(g-1)} \end{bmatrix}$$

Assim, um teste aproximado para a igualdade das g funções de sobrevivência é baseado na estatística:

$$T = U_L' V_L^{-1} U_L ,$$

que sob hipótese nula tem uma distribuição qui-quadrado com $(g-1)$ graus de liberdade.

3.2.4. Modelos paramétricos

Além da análise não-paramétrica, ajustaram-se modelos não-paramétricos aos dados de sobrevivência para cada tratamento. Esses modelos foram: a distribuição exponencial, a distribuição de Weibull, a log-normal e logística.

3.2.4.1. Distribuição exponencial

A variável aleatória tempo de falha, T, tem distribuição exponencial, sua função densidade de probabilidade é dada, conforme Colosimo e Giolo (2006), por:

$$f(t) = \frac{1}{\alpha} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \quad t \geq 0$$

onde o parâmetro $\alpha > 0$ representa o tempo médio de vida. As funções de sobrevivência, S(t) e de taxa de falha $\lambda(t)$ são:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}$$

$$\lambda(t) = \frac{1}{\alpha}$$

A média e a variância são dadas por α e α^2 respectivamente.

3.2.4.2. Distribuição de Weibull

Uma variável T segue distribuição de Weibull a sua função densidade de probabilidade, a sua função de sobrevivência e a função taxa de risco são dadas, conforme Colosimo e Giolo (2006), por respectivamente:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$$

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$$

$$\lambda(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}$$

para $t \geq 0$ e $\gamma \succ 0$. O parâmetro α é o parâmetro de escala e γ o de forma, ambos são positivos (COLOSIMO; GIOLO, 2006). Se $\gamma = 1$ tem-se a taxa de risco constante, esse é o caso da distribuição exponencial, que é um caso particular da Weibull (COLOSIMO; GIOLO, 2006). Se $\gamma > 1$ a taxa de risco é estritamente crescente e no caso $\gamma < 1$ a taxa é estritamente decrescente (COLOSIMO; GIOLO, 2006; LEE; WANG, 2003).

A média e a variância para a distribuição de Weibull apresentam as seguintes expressões:

$$E(T) = \alpha\Gamma\left(1 + \left(\frac{1}{\gamma}\right)\right),$$

$$Var(T) = \alpha^2\left[\Gamma\left(1 + \left(\frac{2}{\gamma}\right)\right) - \Gamma\left(1 + \left(\frac{1}{\gamma}\right)\right)^2\right],$$

onde $\Gamma(k)$ é a função gama dada por $\Gamma(k) = \int_0^k x^k \exp\{-x\}dx$ (COLOSIMO; GIOLO, 2006).

3.2.4.3. Distribuição Log-normal

A distribuição log-normal, de uma forma mais simples, pode ser definida como a distribuição cujo logaritmo da variável aleatória T, segue a distribuição normal (COLOSIMO; GIOLO, 2006; LEE; WANG, 2003). Desse modo, uma variável que segue uma distribuição com parâmetros μ e σ com a função densidade de probabilidade dada, conforme encontrado em Colosimo e Giolo (2006), por:

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma}\exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}, \quad t > 0$$

pode ser analisada segundo uma distribuição normal; basta considerarmos o logaritmo dos dados no lugar dos valores originais (COLOSIMO; GIOLO, 2006). As funções de sobrevivência e taxa de risco, conforme Colosimo e Giolo (2006) são:

$$S(t) = \Phi\left(-\frac{\log(t)+\mu}{\sigma}\right),$$

$$\lambda(t) = \frac{f(t)}{S(t)},$$

onde $\Phi(\cdot)$ é a função de distribuição acumulada de uma normal padrão.

As taxas de falhas da distribuição log-normal não são monótonas, elas crescem, até um máximo, e depois decrescem (COLOSIMO; GIOLO, 2006). Logo, de acordo ainda com Colosimo e Giolo (2006) e Lee e Wang (2003) a média e a variância dessa distribuição são respectivamente:

$$E(T) = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}$$

$$Var(T) = \exp\{2\mu + \sigma^2\}\left(\exp\{\sigma^2\} - 1\right)$$

3.2.4.4. Distribuição Logística

A função densidade de probabilidade, função de sobrevivência e taxa de risco da distribuição logística, conforme apresentadas por Colosimo e Giolo (2006), são dadas a seguir.

$$f(y) = \frac{1}{\sigma}\exp\left\{\frac{t-\mu}{\sigma}\right\}\left(1 + \exp\left\{\frac{t-\mu}{\sigma}\right\}\right)^{-2}$$

$$S(y) = \frac{1}{1 + \exp\left\{\frac{t-\mu}{\sigma}\right\}}$$

$$h(y) = \frac{1}{\sigma}\exp\left\{\frac{t-\mu}{\sigma}\right\}\left(1 + \exp\left\{\frac{t-\mu}{\sigma}\right\}\right)^{-1},$$

Com $-\infty < \mu < \infty$ e $\sigma > 0$, os parâmetros de escala e locação respectivamente.

3.2.5. Métodos gráficos para escolha do melhor modelo

Conforme Colosimo e Giolo (2006) o primeiro método gráfico para a escolha do melhor modelo a ser usado consiste na comparação da função de sobrevivência do modelo proposto com o estimador de Kaplan-Meier, $\hat{S}(t)$. Nesse método ajusta-se o modelo proposto (exponencial, Weibull, log-normal) e estima-se as funções de sobrevivências a partir das estimativas dos parâmetros de cada modelo. Obtém-se também a estimativa de Kaplan-Meier para função de sobrevivência e, em seguida, comparam-se graficamente as funções de sobrevivências de Kaplan-Meier e do modelo proposto. O modelo que melhor se ajusta aos dados é aquele que mais se aproxima da curva de Kaplan-Meier. Colosimo e Giolo (2006) ressaltam que na prática isso é realizado através de gráficos de $\hat{S}(t)$ versus a sobrevivência do modelo proposto. Assim, o melhor modelo será aquele cujos pontos no gráfico estiverem mais próximos da reta y=x, onde x representa a sobrevivência estimada por Kaplan-Meier e y a sobrevivência estimada pelo modelo proposto.

Alternativamente pode-se colocar no mesmo gráfico as curva estimada por Kaplan-Meier versus o tempo e a estimada pelo modelo proposto versus o tempo. Dessa maneira, o modelo cuja curva mais se aproximar da curva de Kaplan-Meier indicará o modelo que melhor se ajusta aos dados (COLOSIMO; GIOLO, 2006).

O segundo método gráfico, apresentado por Colosimo e Giolo (2006) consiste na linearização da função de sobrevivência. Ao linearizar a função tem-se como resultado uma reta, caso o modelo seja adequado.

3.2.6. Comparação dos tratamentos pelo método de agrupamento de Ward

Como metodologia alternativa ao teste *logrank* foi proposta a utilização do método de agrupamento de Ward para comparação dos tratamentos. O método é descrito como segue:

O método de Ward, também chamado de método da variância mínima (MINGOTI, 2005; KHATREE; NAIK, 2000) é definido conforme Rencher (2002) como segue.

Se AB é um grupo obtido pela combinação dos grupos A e B, então a soma das distâncias dentro do grupo (dos elementos a partir do vetor de média dos grupos) são:

$$SSE_A = \sum_{i=1}^{n_A} (y - \bar{y}_A)(y - \bar{y}_A)' \quad (6)$$

$$SSE_B = \sum_{i=1}^{n_B} (y - \bar{y}_B)(y - \bar{y}_B)' \quad (7)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (y - \bar{y}_{AB})(y - \bar{y}_{AB})' \quad (8)$$

Onde $\bar{y}_{AB} = \frac{(n_A \bar{y}_A + n_B \bar{y}_B)}{n_A + n_B}$ e n_A, n_B e $n_{AB} = n_A + n_B$ são o número de pontos em A, B e AB respectivamente. Então, as somas das distâncias são equivalentes à soma de quadrados dentro dos grupos, elas são denotadas por SSE_A, SSE_B e SSE_{AB}.

O método de Ward une dois grupos A e B que minimizam o acréscimo em SSE, definido como:

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \quad (9)$$

Conforme Rencher (2002) pode-se demonstrar que I_{AB} em (9) possui duas formas equivalentes:

$$I_{AB} = n_A (\bar{y}_A - \bar{y}_{AB})'(\bar{y}_A - \bar{y}_{AB}) + n_B (\bar{y}_B - \bar{y}_{AB})'(\bar{y}_B - \bar{y}_{AB}) \quad (10)$$

$$I_{AB} = \frac{n_A n_B}{n_A + n_B} (\bar{y}_A - \bar{y}_B)'(\bar{y}_A - \bar{y}_B) \quad (11)$$

Assim, por (11) minimizar o aumento na SSE é equivalente a minimizar a distância entre grupos. Se A consiste somente de y_i e B de somente y_j, então $SSE_A = 0$ e $SSE_B = 0$, e (10) e (11) reduz a:

$$I_{AB} = SSE_A = \frac{1}{2}(\bar{y}_A - \bar{y}_B)'(\bar{y}_A - \bar{y}_B) = \frac{1}{2} d^2(y_i, y_j)$$

3.2.7. Estatísticas auxiliares para determinação do número de grupos

Um critério objetivo não existe para a determinação do número de grupo, porém algumas estatísticas podem nos auxiliar na partição final (KHATREE; NAIK, 2000; MINGOTI, 2005; SHARMA, 1996). Dentre elas pode-se citar a soma de quadrado entre grupos, a correlação semiparcial, a distância entre grupos e o desvio padrão do grupo. Foram utilizadas as seguintes medidas: soma de quadrado entre grupos, a correlação semiparcial, e a distância entre grupos. O cálculo dessas medidas foi implementado no software R® (R DEVELOPMENT CORE TEAM, 2008) (Apêndice B).

3.2.7.1. Soma de quadrado entre grupos – RSQ

Conforme descrito em Sharma (1996) a soma de quadrado entre grupos, RSQ, é a razão entre a soma de quadrado entre grupos (SSb) e a soma de quadrado total (SSt). Como SSt = SSb+SSw (onde SSw é a soma de quadrado dentro do grupos), quanto maior a SSb menor será a SSw. Consequentemente para certo conjunto de dados quanto, maior a diferença entre grupos mais homogêneos será cada grupo.

$$RSQ = \frac{SSb}{SSt}$$

3.2.7.2. Correlação semiparcial – SPRSQ

A correlação semiparcial é a medida de perda de homogeneidade pela união de dois grupos (KHATREE; NAIK, 2000). Se a perda de homogeneidade é zero, então o novo grupo é obtido pela fusão de dois grupos perfeitamente homogêneos. Por outro lado, se a perda de homogeneidade é grande então o novo grupo é obtido pela fusão de dois grupos heterogêneos (SHARMA, 1996). Um pequeno valor poderia implicar que

estamos unindo dois grupos homogêneos. Assim, para um bom grupo-solução SPR² deverá ser baixo (SHARMA, 1996).

A correlação semiparcial, segundo Mingoti (2005), é dada por:

$$\text{SPRSQ} = \frac{\text{BSS}}{\text{SSt}}$$

onde: BSS é a distância entre grupos usada no método de Ward.

3.2.8. Estudo de simulação: Análise da eficiência do método de agrupamento de Ward na comparação de modelos não-lineares.

Para avaliar a eficiência do método de agrupamento de Ward para testar a identidade de modelos não-lineares foi realizado um estudo baseado em simulação de dados. Neste caso a eficiência corresponde ao número de vezes que o método de agrupamento separa corretamente os tratamentos dentro de cada grupo.

O modelo estatístico usado foi $y_i = \mu_i + \varepsilon_i = E(y_i) + \varepsilon_i$, onde a parte sistemática é dada pelo modelo logístico $E(Y) = \frac{1}{1 + e^{\alpha(\beta - x)}}$. Nessa parametrização, α está relacionado com a taxa média de decrescimento da curva, e β corresponde ao ponto de inflexão, ou seja, ponto correspondente a Y=50% de sobrevivência (Apêndice C). O modelo logístico foi escolhido, pois os dados originais, dados de sobrevivência média, para os oito tratamentos apresentam tendência sigmoidal com assíntotas em zero e um. No modelo usado, considerou-se y_i como o valor gerado para uma observação que, conforme Hosmer e Lemeshow (1989) é dicotômica, isto é, assume o valor um ou zero (sobrevive ou morre) com probabilidade π_i e $1 - \pi_i$ respectivamente, em que π_i representa a probabilidade de sobrevivência no i-ésimo tempo e é calculada a partir de $E(y_i)$, isto é, $\pi_i = E(y_i)$.

Para definição dos cenários considerou-se o parâmetro $\alpha = -0{,}2456$, $\beta^*_1 = 10{,}28$ e β^*_2 de tal forma que $\beta^*_2 - \beta^*_1$ assuma os valores 1,028, 2,056, 3,084, 4,112, 5,14, 6,188, 7,196, 8,224, 9,252, 10,28, este valores correspondem a 10%, 20%,

30%, 40%, 50%, 60%, 70%, 80%, 90% e 100% de $\beta^*_1 = 10{,}28$. Foram considerados, também, diferentes tamanhos de amostra n = 10, 20, 30, 40, 50, 100. O parâmetro α foi considerado como fixo devido à pequena influência no cálculo da matriz de dissimilaridade.

Foram considerados 60 cenários que correspondem à combinação dos valores de α, β e n. Foram gerados oito tratamentos de forma que ao agrupá-los se tivesse dois grupos compostos por quatro tratamentos cada. Assim β_1 corresponde ao parâmetro para o primeiro grupo gerado e β_2 para o segundo grupo gerado. Ou seja, os tratamentos 1, 2, 3 e 4 foram gerados utilizando $\alpha = -0{,}2456$ e $\beta_1 = 10{,}28$ e os tratamentos 5, 6, 7 e 8 a partir de $\alpha = -0{,}2456$ e β_2 variando conforme a diferença especificada acima. Objetivou-se com este estudo verificar a relação entre a eficiência do agrupamento, tamanho da amostra e diferença entre os β de cada grupo de modelos. Para o estudo proposto foi desenvolvido uma rotina no software R® (R DEVELOPMENT CORE TEAM, 2008).

4. RESULTADOS E DISCUSSÕES

4.1. Análise dos dados – Conjunto de dados 1

Na Figura 1, estão representadas as curvas de sobrevivência estimadas por Kaplan-Meier para cada um dos oito tratamentos. Pode-se observar nesta figura, que a curva de sobrevivência do tratamento 5 ficou abaixo de todas as outras significando que as formigas alimentadas com este tratamento têm probabilidade menor de sobreviver a um determinado tempo "t" do que as formigas alimentadas pelos outros tipos de tratamentos. Verificou-se também que para os tratamentos 4 e 8 as curvas de sobrevivência são bem próximas e que formigas submetidas a esses tratamentos tiveram maiores chance de sobrevivências.

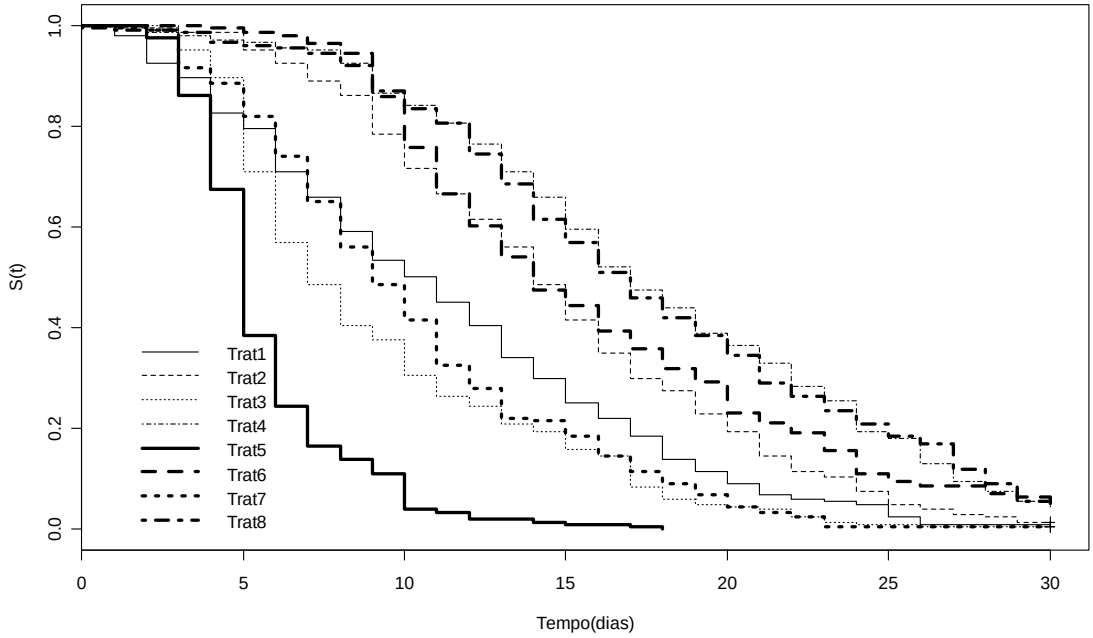

Figura 1 – Sobrevivência estimada por Kaplan-Meier para cada um dos 8 tratamentos.

As estimativas do tempo em que 50% das formigas permaneceram vivas são apresentadas na Tabela 1.

Tabela 1- Tempos (dias) medianos para a sobrevivência de formigas submetidas a diferentes tratamentos

Tratamento	Mediana[1]	Mediana[2]	IC[3]	
1	10,00	10,00	8,51 ;	11,49
2	13,80	14,00	13,04 ;	14,96
3	6,82	7,00	6,16 ;	7,84
4	16,44	17,00	15,66 ;	18,34
5	4,60	5,00	4,77 ;	5,23
6	13,54	14,00	12,66 ;	15,33
7	8,80	9,00	8,04 ;	9,96
8	16,20	17,00	15,62 ;	18,38

[1]Mediana calculada por interpolação linear.
[2]Mediana calculada conforme Collet (1994).
[3]IC – Intervalo de confiança calculado conforme Collet (1994).

Como no estudo realizado o objetivo foi avaliar os tratamentos, foi realizada a comparação dos oito tratamentos através do teste *logrank*. O valor da estatística *logrank*, que sob a hipótese de igualdade das curvas de sobrevivência, tem uma distribuição qui-quadrado com sete graus de liberdade, resultou em T= 810, gerando um p-valor aproximadamente zero. Este resultado evidenciou a existência de diferenças entre os tratamentos.

Constatada a existência de diferenças entre os tratamentos, foi necessário, verificar quais curvas diferiam entre si. Para comparar pares de curvas foi realizado o teste *logrank* com correção de Bonferroni, conforme sugerido por Colosimo e Giolo (2006). De acordo com a correção de Bonferroni o nível de significância utilizado foi $\frac{\alpha}{c}$ onde α é o nível nominal de significância (isto é, 0,05) e c corresponde ao número de comparações a serem realizadas.

Os p-valores dos testes realizados são apresentados na Tabela 2.

Tabela 2 – p-valores referentes ao teste não paramétrico *logrank* com correção de Bonferroni.

Tratamentos	1	2	3	4	5	6	7
2	< 0,001*						
3	0,001 *	< 0,001*					
4	< 0,001*	< 0,001*	< 0,001*				
5	< 0,001*	< 0,001*	< 0,001*	< 0,001*			
6	< 0,001*	0,0692 ns	< 0,001*	0,014 ns	< 0,001*		
7	0,0348 ns	< 0,001*	0,107 ns	< 0,001*	< 0,001*	< 0,001*	
8	< 0,001*	< 0,001*	< 0,001*	0,926 ns	< 0,001*	0,0281 ns	< 0,001*

ns - não significativo a $\frac{\alpha}{c} = 0{,}0018$ de probabilidade.

* - significativo a $\frac{\alpha}{c} = 0{,}0018$ de probabilidade.

Dos resultados apresentados na Tabela 2 admitiu-se não haver diferenças significativas entre os tratamentos 1 e 7, 2 e 6, 3 e 7, 4 e 6, 4 e 8 e entre os tratamentos 6 e 8. Pode-se concluir também que o tratamento 5 difere de todos os outros tratamentos.

A análise das curvas de Kaplan-Meier e os resultados da Tabela 2 mostraram que os tratamentos sem a presença de lixo (tratamentos 2, 4, 6 e 8) apresentaram maior sobrevivência.

Além da análise não paramétrica, foram ajustados modelos paramétricos aos dados de sobrevivência para cada tratamento. Os modelos ajustados foram o exponencial, Weibull, log-normal e logístico. Na Tabela 3 são apresentadas as estimativas dos parâmetros dos modelos ajustados para cada tratamento.

Tabela 3 - Parâmetros estimados pela distribuição exponencial, Weibull, log-normal e logística, para a sobrevivência de formigas submetidas a diferentes tratamentos.

Tratamento	Exponencial	Weibull		Log-normal		Logística	
	$\hat{\alpha}$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
1	11,27	12,61	1,75	2,20	0,73	10,81	3,81
2	15	16,69	2,55	2,59	0,49	14,51	3,58
3	9,32	10,53	1,86	2,08	0,54	8,51	2,95
4	18,60	20,13	2,65	2,79	0,52	17,73	4,26
5	5,71	6,46	2,31	1,66	0,4	5,35	1,29
6	-	-	-	-	-	15,31	3,9
7	10,29	11,60	2,05	2,18	0,57	9,77	2,99
8	-	-	-	-	-	17,43	4,33

- O modelo não se ajustou aos dados

Com o objetivo de escolher o modelo que melhor se ajustou aos dados, utilizou-se, conforme sugerido por Colosimo e Giolo (2006) métodos gráficos. No método gráfico 1, foram construídos os gráficos das estimativas das sobrevivências obtidas pelo método de Kaplan-Meier versus as estimativas das sobrevivências obtidas através dos modelos exponencial, Weibull, log-normal e logística (Apêndice D). Os gráficos são apresentados na Figura 2.

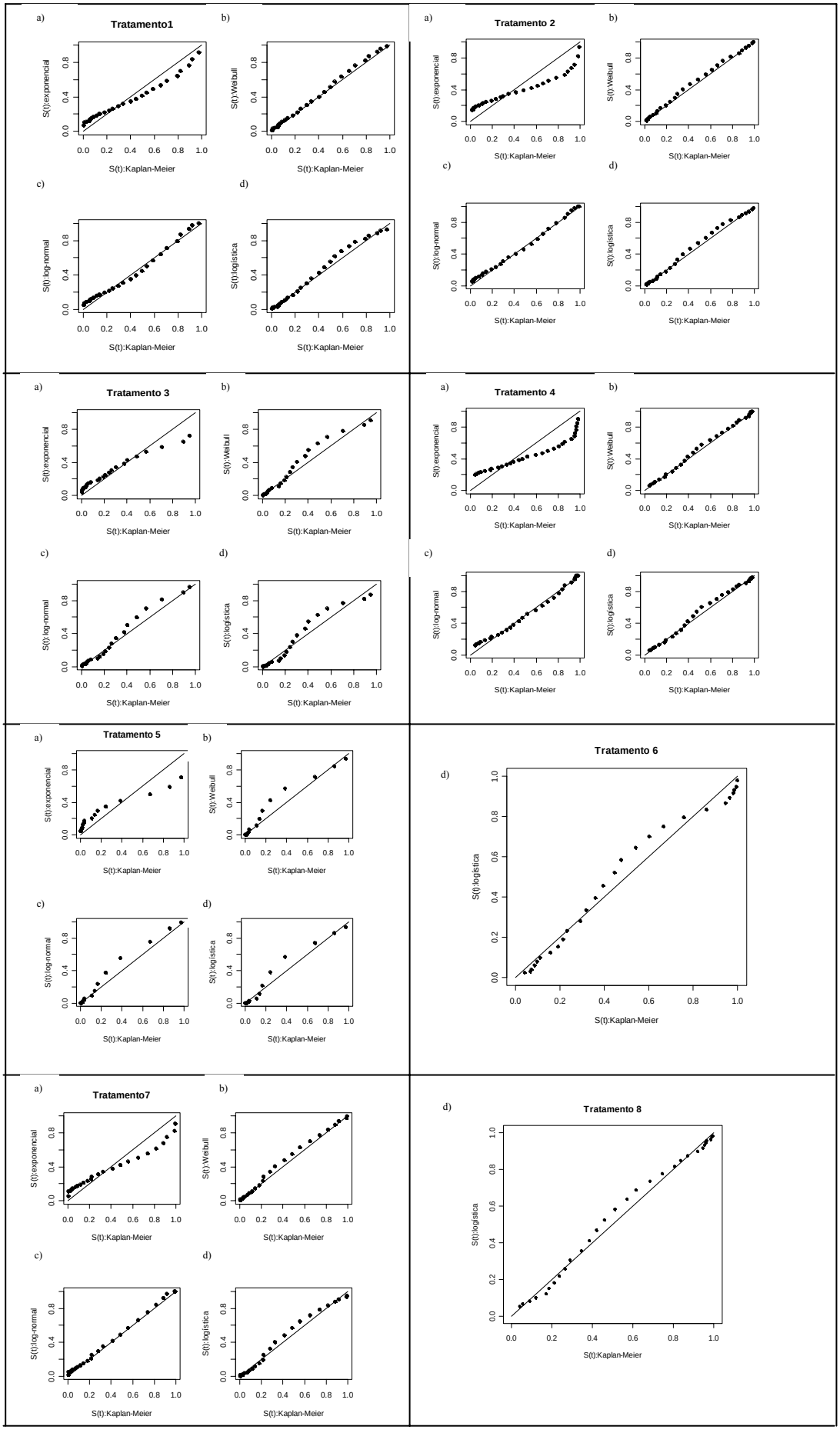

Figura 2 - Gráficos das sobrevivências estimadas por Kaplan-Meier *versus* sobrevivências estimadas pelos modelos a) exponencial, b) de Weibull, c) log-normal e d) logístico.

Na Figura 2, pode-se verificar que para os tratamentos 1, 2, 4 e 7 o modelo exponencial é o que mais se afasta da reta y = x, sugerindo assim, que esse modelo não é adequado para os conjuntos de dados dos respectivos tratamentos. Para os tratamentos 6 e 8, o modelo logístico é o único que se ajustou aos dados, conforme a Tabela 2. Os gráficos para esses modelos mostraram que as curvas para os tratamentos 6 e 8 não apresentaram grandes afastamentos em relação à reta, sendo assim, o modelo ajustado é candidato para os dados desses tratamentos. Em relação aos tratamentos 3 e 5, a distribuição exponencial apresentou considerável afastamento em relação à reta. Porém, para os modelos Weibull, log-normal e logístico, o padrão de afastamento, em relação aos demais tratamentos, mostrou um afastamento um pouco maior, o que dificultou a identificação do modelo adequado. Assim, foram considerados outros métodos gráficos, sugeridos por Colosimo e Giolo (2006) para confirmação dos resultados do primeiro método gráfico.

O segundo método gráfico consiste em construir os gráficos linearizados para os modelos considerados (exponencial, Weibull, log-normal, logístico). Na Figura 3 são apresentados esses gráficos.

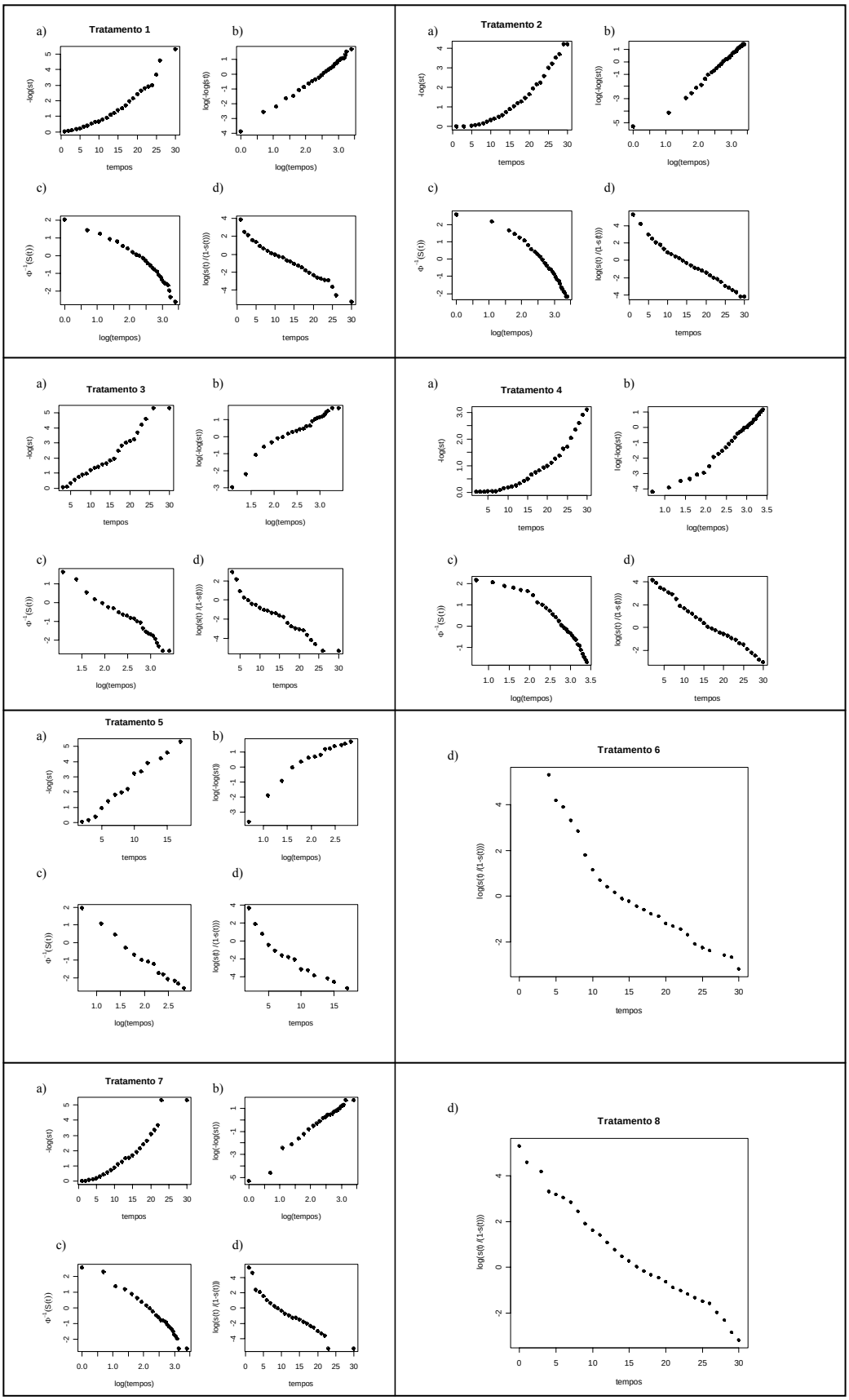

Figura 3 – Gráficos dos modelos a) exponencial, b) Weibull, c) log-normal e d) logístico, linearizados.

Conforme a Figura 3 tem-se que os modelos que não apresentaram afastamentos marcantes de uma reta para o tratamento 1, 2, 4 e 7 foram os modelos de Weibull e logístico. No caso do tratamento 3 todos os modelos não apresentaram excessivo afastamento, sugerindo, portanto, que qualquer um dos modelos poderia ser ajustado. Para os tratamentos 6 e 8 o único modelo ajustado pareceu adequado.

Através dessa análise tem-se que tanto o modelo de Weibull como o logístico apresentaram ajuste satisfatório para o estudo do conjunto de dados em questão. Essa afirmativa pode ser comprovada pela construção do gráfico contendo a curva de sobrevivência estimada por Kaplan-Meier, e a estimada pelos modelos julgados como adequados (Figuras 4 a 11), conforme sugerido por Colosimo e Giolo (2006).

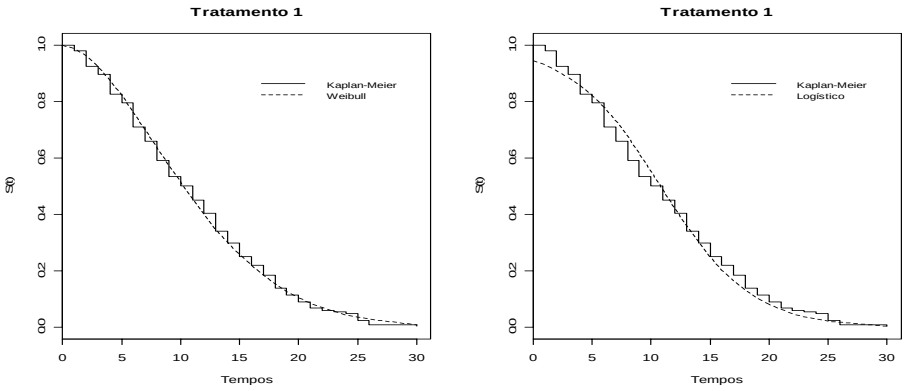

Figura 4 - Curvas de sobrevivência estimadas pelos modelos de Weibull e logístico versus a curva de sobrevivência estimada por Kaplan-Meier, para o tratamento 1.

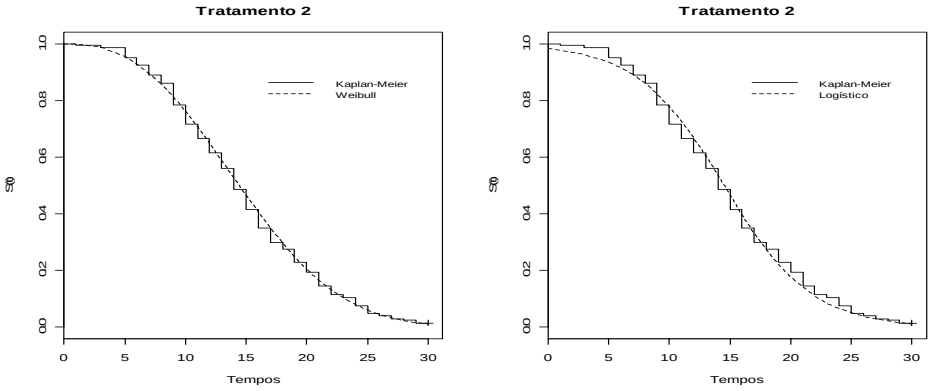

Figura 5 - Curvas de sobrevivência estimadas pelos modelos de Weibull e logístico versus a curva de sobrevivência estimada por Kaplan-Meier para o tratamento 2.

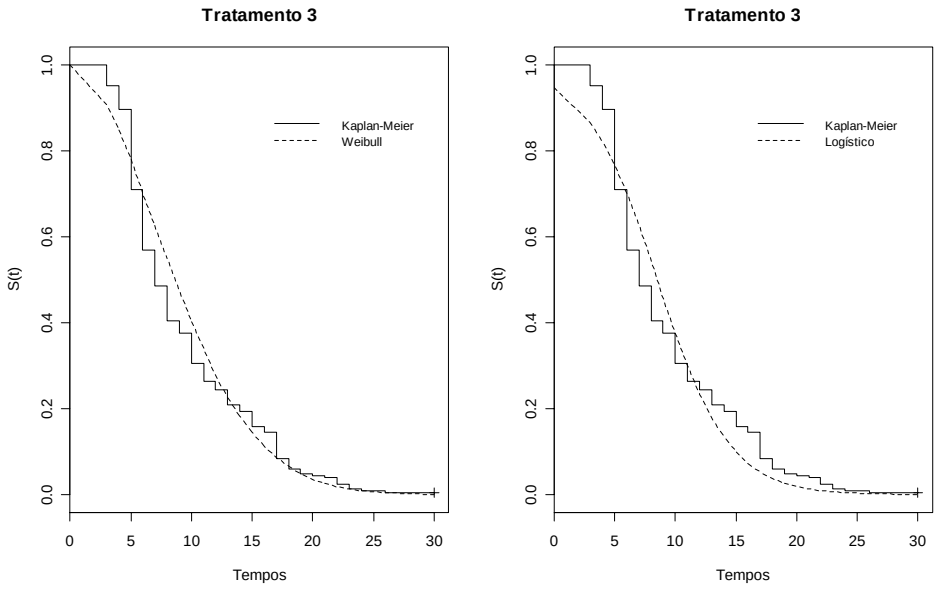

Figura 6 - Curvas de sobrevivência estimadas pelos modelos de Weibull e logístico versus a curva de sobrevivência estimada por Kaplan-Meier para o tratamento 3.

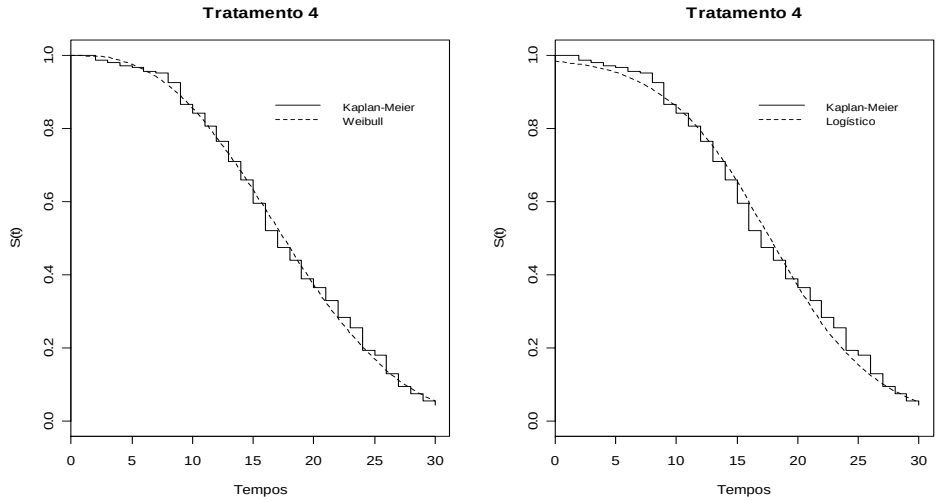

Figura 7 - Curvas de sobrevivência estimadas pelos modelos de Weibull e logístico versus a curva de sobrevivência estimada por Kaplan-Meier para o tratamento 4.

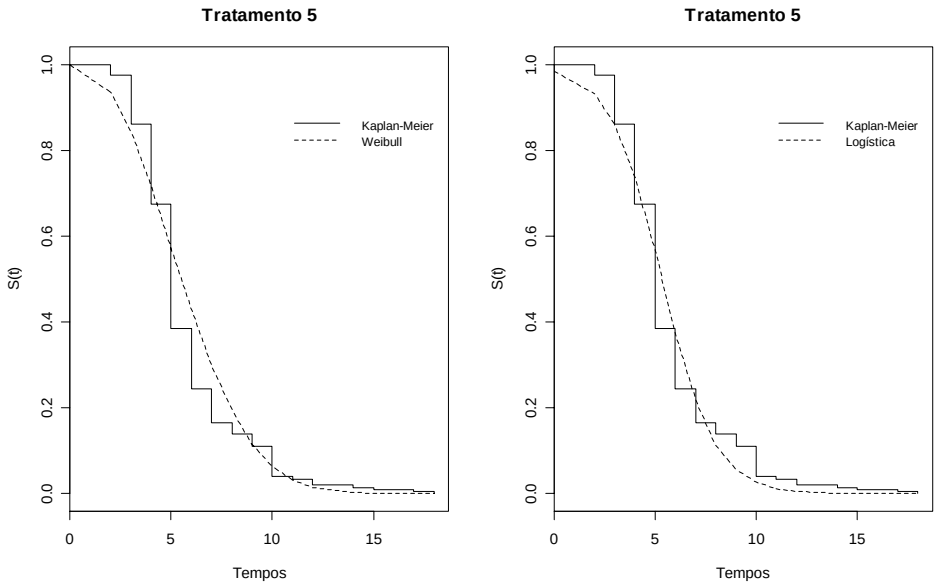

Figura 8 - Curvas de sobrevivência estimadas pelos modelos de Weibull e logístico versus a curva de sobrevivência estimada por Kaplan-Meier para o tratamento 5.

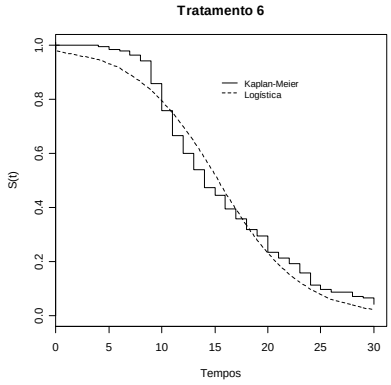

Figura 9 - Curva de sobrevivência estimada pelo modelo logístico versus a curva de sobrevivência estimada por Kaplan-Meier para o tratamento 6.

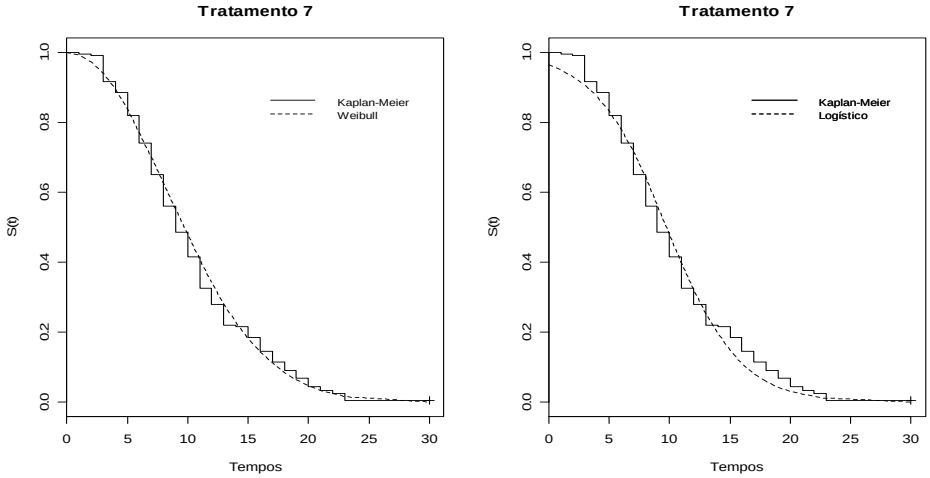

Figura 10 - Curvas de sobrevivência estimadas pelos modelos de Weibull e logístico versus a curva de sobrevivência estimada por Kaplan-Meier para o tratamento 7.

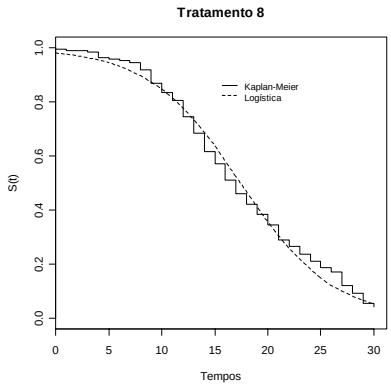

Figura 11 - Curvas de sobrevivência estimada pelo modelo logístico versus a curva de sobrevivência estimada por Kaplan-Meier para o tratamento 8.

Pode-se observar, portanto que o único modelo que se ajustou a todos os tratamentos é o modelo logístico.

Análise de agrupamento – Agrupamento dos coeficientes do modelo logístico para o conjunto de dados 1.

Uma vez que o modelo logístico se ajustou a todos os tratamentos, a avaliação do uso da análise de agrupamento para comparação de curvas foi realizada através do agrupamento das estimativas dos coeficientes desse modelo.

A Tabela 4 apresenta os parâmetros estimados para o modelo logístico para os oito tratamentos estudados. Esses parâmetros formaram a matriz de dados a ser utilizada no agrupamento dos tratamentos.

Tabela 4 - Matriz de dados - coeficientes do modelo logístico ajustado a cada tratamento.

Tratamentos	Parâmetros estimados	
	$\hat{\mu}$	$\hat{\sigma}$
1	10,81	3,81
2	14,51	3,58
3	8,51	2,95
4	17,73	4,26
5	5,35	1,29
6	15,31	3,90
7	9,77	2,99
8	17,43	4,33

Com base nos coeficientes do modelo ajustado para cada tratamento (Tabela 4) foi obtida a matriz de dissimilaridade (Tabela 5) baseada na soma de quadrados dos desvios entre tratamentos e aplicado o método de agrupamento de Ward.

Tabela 5 - Matriz de dissimilaridade dos coeficientes dos modelos ajustados.

	Trat 1	Trat 2	Trat 3	Trat 4	Trat 5	Trat 6	Trat 7
Trat 2	6,8714						
Trat 3	3,0148	18,1985					
Trat 4	24,0444	5,4154	43,3623				
Trat 5	18,0810	44,5748	6,3706	81,0427			
Trat 6	10,13	0,3712	23,5713	2,9930	53,0069		
Trat 7	10,1291	11,4078	0,7946	32,4873	11,2132	15,7599	
Trat 8	22,0474	4,5445	40,7354	0,0475	77,5840	2,3396	30,2356

O resultado do processo de agrupamento foi representado em um dendrograma (Figura 12).

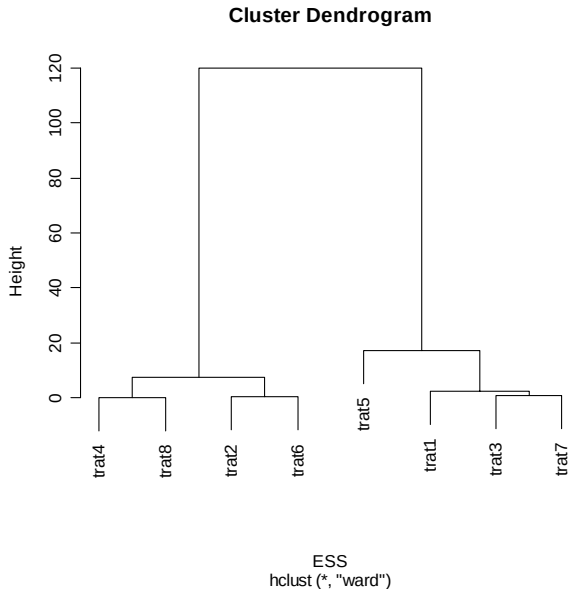

Figura 12 - Dendrograma obtido pelo Método de Ward aplicados aos dados da Tabela 4.

Os valores da estatística auxiliares para a determinação do número de grupos são apresentados na Tabela 6. Os gráficos obtidos para distância entre grupos (BSS), Soma de quadrados entre grupos (RSQ) e correlação semiparcial (SPRSQ) em função do número de grupos são apresentados na Figura 13.

Tabela 6 - Resumo das medidas estatísticas para o agrupamento dos tratamentos de acordo com as estimativas dos parâmetros do modelo ajustado.

Passo	Grupos	Número de grupos	BSS	SPRSQ	RSQ
1	{1},{2},{3},{4,8},{5},{6},{7}	7	0,0475	0,0003	1,000
2	{1},{3},{4,8},{5},{7},{2,6}	6	0,3712	0,0025	0,997
3	{1}, {2,6},{3,7},{4,8},{5}	5	0,7946	0,0054	0,992
4	{1,3,7},{2,6},{4,8},{5}	4	2,3297	0,0158	0,976
5	{1,3,7},{2,4,6,8},{5}	3	7,4369	0,0503	0,926
6	{1,3,5,7},{2,4,6,8}	2	17,0513	0,1154	0,810
7	{1,2,3,4,5,6,7,8}	1	119,7491	0,8103	0,000

BSS – Distância entre os grupos;

SPRSQ – Correlação Semiparcial;

RSQ – Soma de quadrados entre grupos.

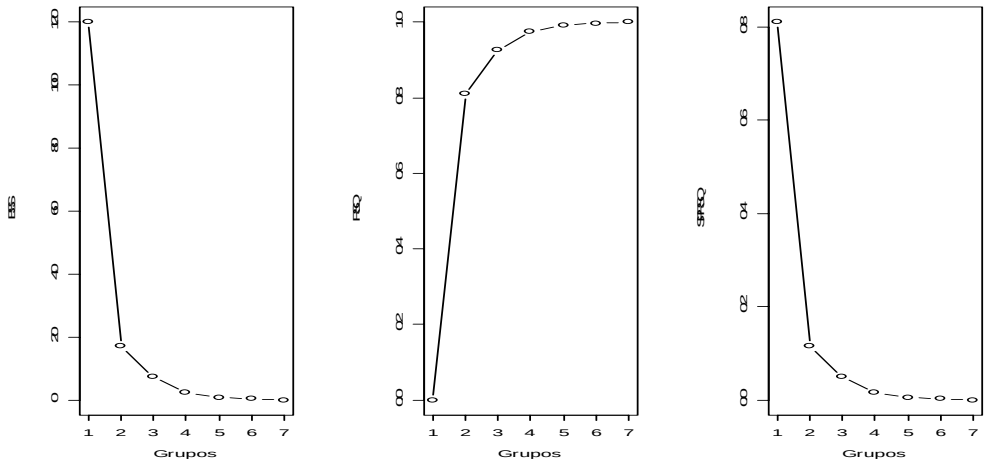

Figura 13 - Gráficos das estatísticas auxiliares (BSS, RSQ, SPRSQ) para a escolha do número de grupos.

Conforme Khatree e Naik (2000) a determinação do número de grupos não é um problema fácil, e é feita a partir de critérios heurísticos (BSS, SPRSQ, RMSSTD e RSQ). Foram utilizadas para determinação do número de grupos as estatísticas BSS, SPRSQ e RSQ. Como sugerem Mingoti (2005) e Sharma (1996) utilizou-se as estatísticas auxiliares BSS, SPRSQ e RSQ. Posteriormente foram construídos os gráficos dessas estatísticas em função do número de grupos (Figura 13) e, finalmente, foram detectados pontos de saltos relativamente grandes em relação aos demais valores de distância. Os valores das estatísticas BSS, SPRSQ devem ser pequenos e os da estatística RSQ altos (SHARMA, 1996). Sharma (1996) recomenda, ainda, que pesquisadores levem em consideração o objetivo do estudo para a avaliação do número de grupos. Assim, de acordo com a Tabela 6 e a Figura 13 optou-se por 3 grupos.

Os resultados obtidos pela análise de agrupamento mostraram equivalência entre os tratamentos 2,4, 6 e 8 (tratamentos sem a presença do lixo); equivalência entre os tratamentos 1, 3 e 7 (tratamentos com a presença do lixo) e indicaram a diferença do tratamento 5 em relação aos demais. Esses resultados corroboram com a hipótese inicial de que na ausência do lixo a sobrevivência é maior. Resultados equivalentes foram obtidos utilizando o teste *logrank* (Tabela 2).

Estudo similar foi realizado por Peternelli *et al*. (2005) que ajustaram um modelo logístico, para os dados percentuais de sobrevivência e compararam as curvas

por meio de análise de agrupamento. Os parâmetros estimados para o modelo usado, $E(Y) = \frac{1}{1+e^{\alpha(\beta-x)}}$, são apresentados na Tabela 7.

Tabela 7 - Estimativas dos parâmetros do modelo $E(Y) = \frac{1}{1+e^{\alpha(\beta-x)}}$

Tratamentos	Parâmetros estimados	
	$\hat{\alpha}$	$\hat{\beta}$
1	-0,2457	10,28
2	-0,2707	14,01
3	-0,3372	7,99
4	-0,2227	17,29
5	-0,8644	4,81
6	-0,2518	14,81
7	-0,3267	9,26
8	-0,2218	17,01

Fonte: Peternelli *et al*., 2005, p. 213.

Os resultados obtidos na comparação das curvas foram similares àqueles obtidos por meio de técnicas de análise de sobrevivência. Esse fato conduz à hipótese de que na presença de baixa taxa de censura do tipo I, isto é, quando quase a totalidade dos indivíduos experimentaram o evento de interesse, o modelo ajustado por Peternelli *et. al* (2005) é equivalente ao modelo ajustado aos tempos de sobrevivência via Kaplan-Meier (Figura 14).

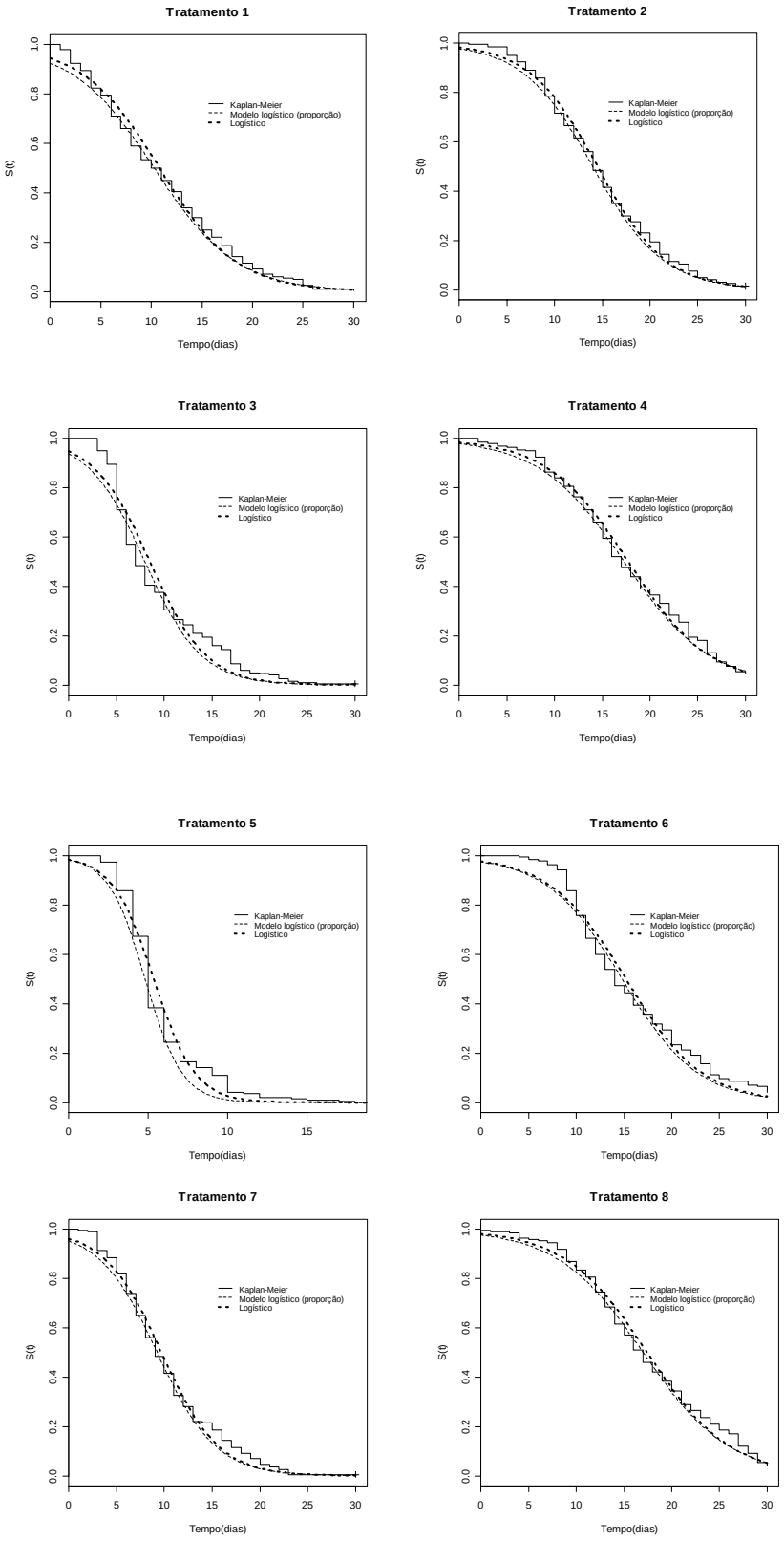

Figura 14 - Curvas de sobrevivência estimadas pelo modelo logístico considerando os dados percentuais de sobrevivência média (Modelo logístico (proporção)) e o modelo logístico ajustado aos tempos de falha (logístico) versus a curva de sobrevivência estimada por Kaplan-Meier.

De acordo com a parametrização apresentada por Peternelli *et al*. (2005) tem-se que o parâmetro β corresponde à mediana. Ao se comparar os valores medianos estimados para cada um dos tratamentos através do modelo utilizado pelos referidos autores, observou-se que esses valores se encontram próximos aos tempos medianos estimados por interpolação também dos tempos medianos estimados segundo Collet (1994) (ver Tabela 1).

A fim de avaliar a influência da taxa de censura na adequacidade do modelo aumentaram-se as taxas de censuras dos dados do tratamento 1. Para o acréscimo de censura tipo I determinou-se tempos de censura de 25, 20, 15 e 10 dias. Após o acréscimo na taxa de censura construiu-se o diagrama de dispersão dos dados para verificar o comportamento dos mesmos (Figura 15).

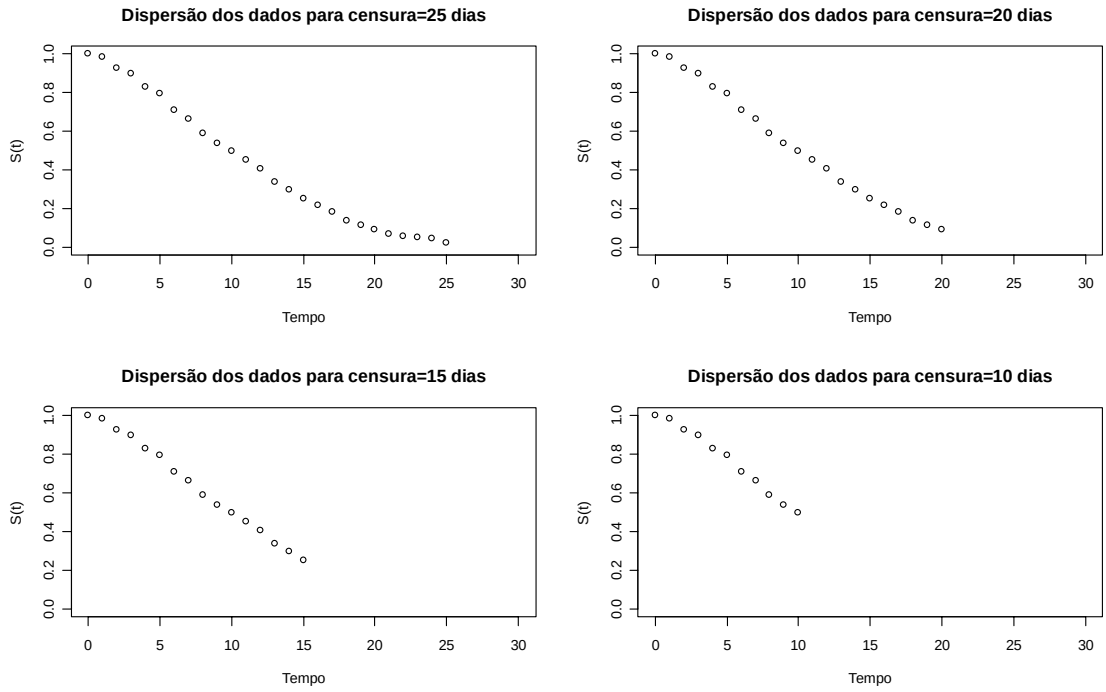

Figura 15 - Gráfico de dispersão dos dados do tratamento 1 para tempos de censura igual a 25, 20, 15, 10.

De acordo com a Figura 15, o primeiro gráfico apresenta comportamento não linear e forma de S sugerindo que o modelo logístico possa ser ajustado. Porém, à medida que se aumenta a censura esse comportamento não é visualizado.

Ao utilizar tempo de censura igual a 25 dias tinha-se ainda um grande número de indivíduos que morreram fazendo com que a sobrevivência estimada se aproximasse de zero, porém quando foi aumentada gradativamente essa taxa de censura as sobrevivências estimadas se afastaram cada vez mais do valor zero. Ao usar o tempo de

censura igual a 20, 15 e 10 dias pode-se notar que o modelo logístico se afastou da curva de Kaplan-Meier. Esse afastamento se dá a partir dos tempos determinados como tempos de censura (Figura 16).

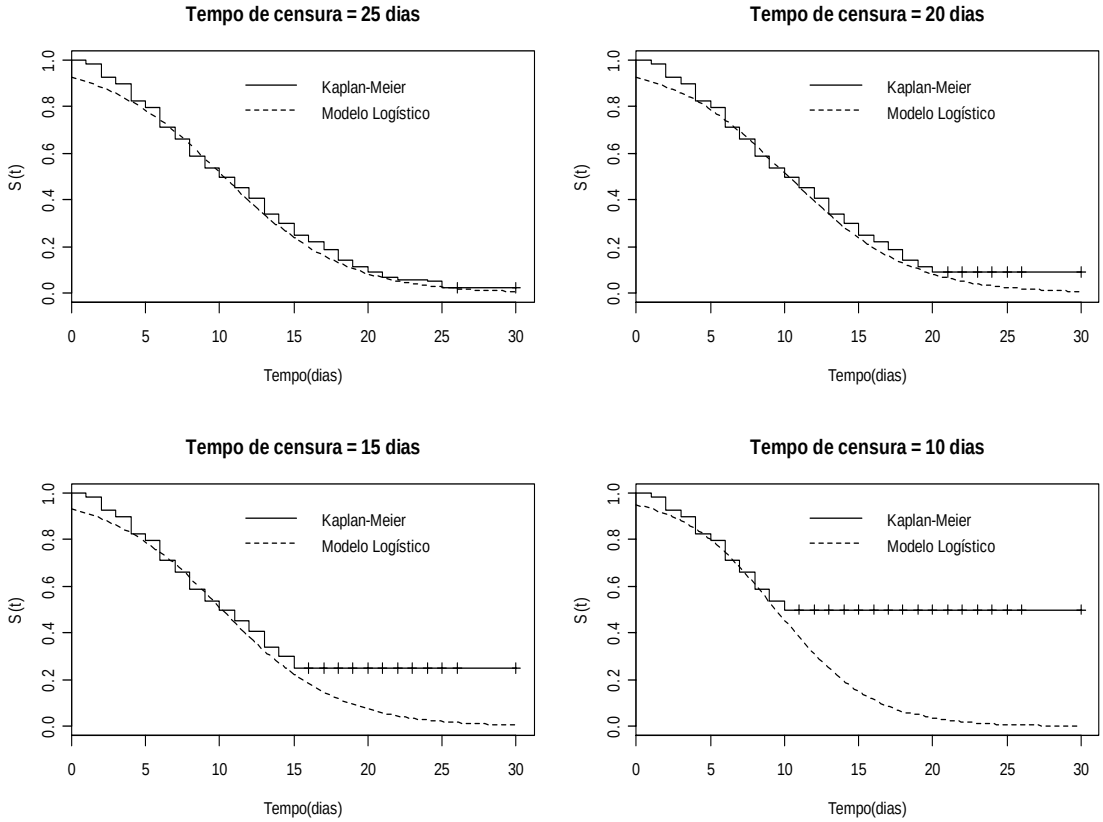

Figura 16 - Curvas de sobrevivência estimadas pelo modelo logístico versus a curva de sobrevivência estimada por Kaplan-Meier, para diferentes níveis de censura.

4.2. Análise de dados – Conjunto 2

As curvas estimadas a partir do estimador de Kaplan-Meier para o conjunto de dados simulados a partir da distribuição de Weibull são mostradas na Figura 17.

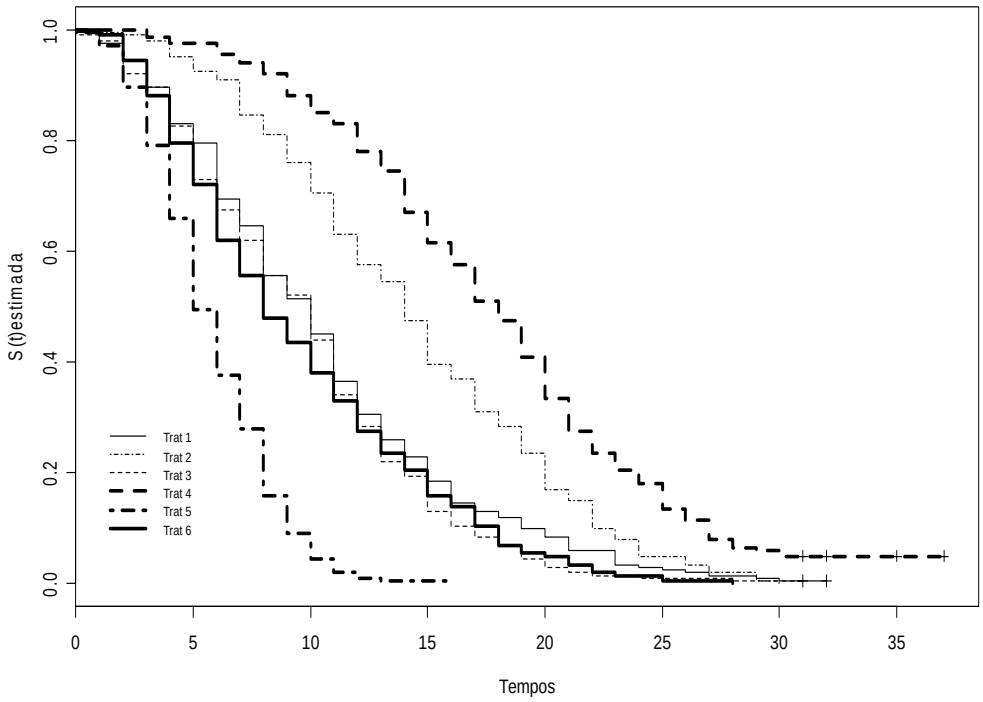

Figura 17 - Sobrevivência estimada por Kaplan-Meier para os dados simulados.

O teste de igualdade das seis curvas de sobrevivência realizado pelo teste *logrank*, forneceu o valor da estatística T = 498 e um p-valor muito pequeno (< 0,0001), o que indica a existência de diferenças entre os tratamentos.

A fim de verificar quais tratamentos diferiam entre si, foi realizado o teste *logrank* aos pares e utilizada a correção de Bonferroni, conforme sugerido por Colosimo e Giolo (2006). Assim, o nível de significância utilizado foi 0,05/15 = 0,0033. A tabela 8 mostra os resultados do teste *logrank* realizados para as comparações dos pares de tratamentos.

Tabela 8 - p-valores referentes ao teste não-paramétrico *logrank* com correção de Bonferroni.

	Tratamentos				
	1	2	3	4	5
2	$< 0{,}003^{*}$				
3	$0{,}158^{ns}$	$< 0{,}003^{*}$			
4	$< 0{,}003^{*}$	$< 0{,}003^{*}$	$< 0{,}003^{*}$		
5	$< 0{,}003^{*}$	$< 0{,}003^{*}$	$< 0{,}003^{*}$	$< 0{,}003^{*}$	
6	$0{,}0785^{ns}$	$< 0{,}003^{*}$	$0{,}843^{ns}$	$< 0{,}003^{*}$	$< 0{,}003^{*}$

^{ns} - não-significativo a $\frac{\alpha}{c} = 0{,}0033$ de probabilidade.

* - significativo a $\frac{\alpha}{c} = 0{,}0033$ de probabilidade

Conforme os resultados apresentados na tabela 8 pode-se verificar que não há evidências de diferenças entre os tratamentos 1 e 3, 1 e 6 e entre os tratamentos 3 e 6. Concluiu-se ainda que o tratamento 5 difere de todos os outros tratamentos.

Análise de agrupamento – Agrupamento dos coeficientes do modelo de Weibull

Para a realização do agrupamento utilizou-se como matriz de dados a matriz formada pelos parâmetros estimados para modelo Weibull ajustado aos dados simulados. As estimativas desses parâmetros são apresentadas na tabela 9.

Tabela 9 - Matriz de dados - coeficientes do modelo Weibull ajustado a cada tratamento.

Tratamento	Parâmetros estimados	
	$\hat{\alpha}$	$\hat{\gamma}$
1	11,80	1,76
2	16,21	2,47
3	10,97	1,85
4	20,29	2,66
5	6,58	2,36
6	10,74	1,81

Com base nos coeficientes dos modelos ajustados para cada tratamento (Tabela 9) foi obtida a matriz de dissimilaridade (Tabela 10) baseada na soma de quadrados dos desvios entre tratamentos e aplicado o método de agrupamento de Ward.

Tabela 10 - Matriz de dissimilaridade dos coeficientes do modelo ajustado.

	Trat 1	Trat 2	Trat 3	Trat 4	Trat 5
Trat 2	9,9767				
Trat 3	0,3464	13,9002			
Trat 4	36,5168	8,3756	43,8021		
Trat 5	13,8020	46,3646	9,7754	94,1281	
Trat 6	0,5573	15,1471	0,0269	45,9885	8,8249

O resultado do processo de agrupamento está representado em um dendrograma (Figura 18).

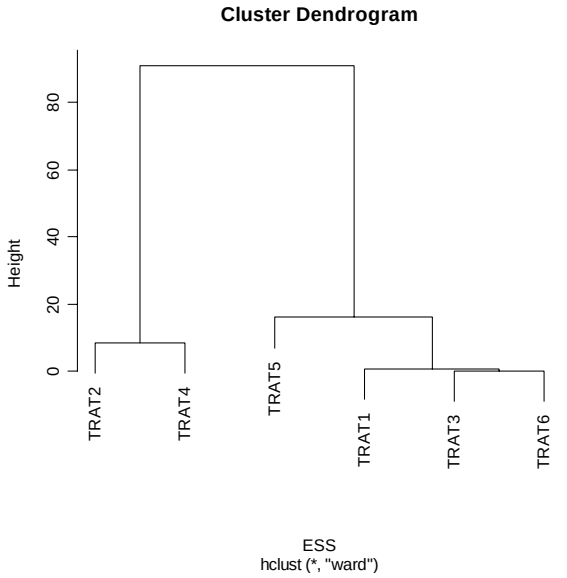

Figura 18 - Dendrograma do Método de Ward aplicados aos dados da Tabela 8.

Os valores das estatísticas auxiliares para a determinação do número de grupos são apresentados na Tabela 11. Os gráficos obtidos para a distância entre grupos (BSS), soma de quadrados entre grupos (RSQ) e correlação semiparcial (SPRSQ) em função do número de grupos são apresentados na Figura 19.

Tabela 11 - Resumo das medidas estatísticas para o agrupamento dos tratamentos de acordo com as estimativas dos parâmetros do modelo ajustado.

Passo	Grupos	Número de grupos	BSS	SPRSQ	RSQ
1	{1},{2},{3,6},{4},{5}	5	0,0269	0,0002	0,999
2	{1,3,6},{2},{4},{5}	4	0,5395	0,0051	0,995
3	{1,3,6},{2,4},{5}	3	8,3756	0,0723	0,922
4	{1,3,5,6},{2,4}	2	16,046	0,1385	0,784
5	{1,2,3,4,5,6}	1	90,8022	0,7838	0,000

BSS – distância entre grupos

SPRSQ – correlação semiparcial

RSQ – soma de quadrados entre grupos

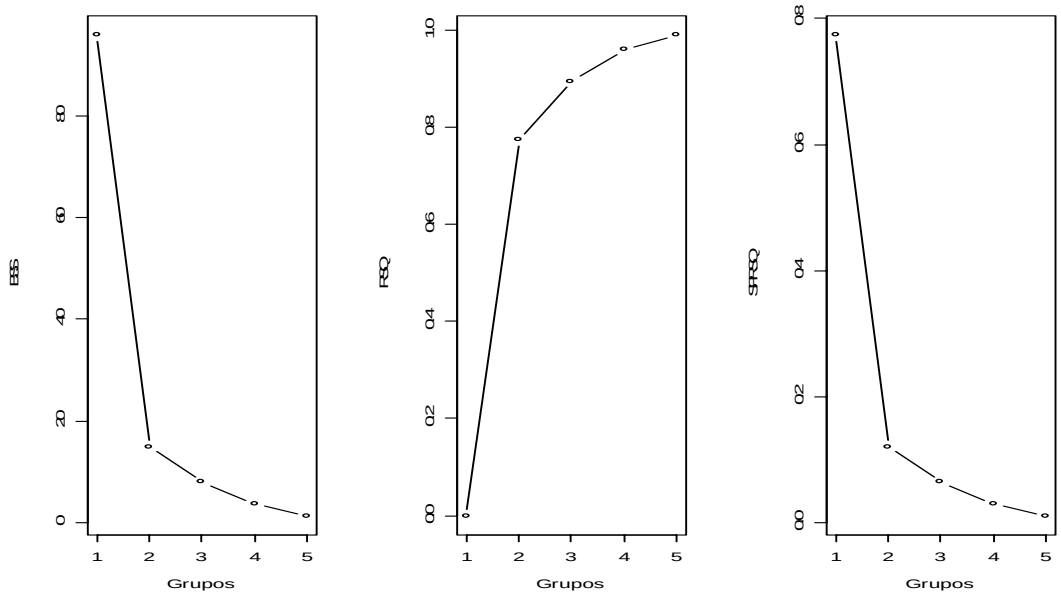

Figura 19 - Gráficos das estatísticas auxiliares (BSS, RSQ, SPRSQ) para a escolha do número de grupos.

Baseado nos gráficos de algumas estatísticas auxiliares sugeridas por Khatree e Naik (2000) na determinação do número de grupos (Tabela 10) construiu-se os gráficos dessas estatísticas em função do número de grupos (Figura 19) conforme recomendado por Mingoti (2005) e Sharma (1996). Em seguida, procurou-se detectar pontos de saltos relativamente grandes em relação aos demais valores de distância. Os valores das estatísticas BSS, SPRSQ devem ser pequenos e os da estatística RSQ altos (SHARMA, 1996). Sharma (1996) recomenda, ainda, que pesquisadores levem em consideração o objetivo dos estudos para a avaliação do número de grupos. Assim, de acordo com a Tabela 11 e a Figura 19, optou-se por 3 grupos.

Os resultados obtidos pela análise de agrupamento mostraram equivalência entre os tratamentos 1,3 e 6; equivalência entre os tratamentos 2 e 4 e indicaram a diferença do tratamento 5 em relação aos demais. Se for considerado um nível de dissimilaridade menor, os tratamentos 2 e 4 não seriam idênticos, resultado equivalente ao teste *logrank*.

4.3. Análise da eficiência do método de agrupamento de Ward na comparação de modelos não-lineares.

Foram geradas um total de 8 amostras (tratamentos) em dois grupos, isto é, $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = -0{,}2456$ e $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_1^*$ e $\beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_2^*$. Gerados os tratamentos, esses foram submetidos ao ajuste do modelo proposto $\left(Y = \frac{1}{1 + \exp(\alpha(\beta - x)} \right)$ e os parâmetros estimados foram submetidos à análise de agrupamento método de Ward. Como foi estabelecido o número de grupos a ser formado (2 grupos); o passo seguinte foi cortar o dendrograma gerado em dois grupos e avaliar se o agrupamento foi realizado corretamente. Isto é, se os tratamentos que foram gerados de um mesmo α e β pertenciam a um mesmo grupo. Foram realizadas 1000 simulações e registrado o número de vezes que o método de Ward conseguiu identificar corretamente os elementos dentro de cada grupo.

A Tabela 12 mostra a proporção de vezes que o método de Ward agrupou corretamente os tratamentos. A Figura 20 apresenta a proporção de acertos no agrupamento dos tratamentos para amostras aleatórias de tamanhos 10, 20, 30, 40, 50, 100 em função da diferença entre os betas.

Tabela 12 - Proporção de acertos no agrupamento dos tratamentos para diferentes tamanhos de amostra e diferenças entre os betas ($\beta_2^* - \beta_1^*$).

	Tamanho da Amostra					
$\beta_2^* - \beta_1^*$	10	20	30	40	50	100
1,028	0,086	0,192	0,327	0,444	0,551	0,839
2,056	0,445	0,775	0,91	0,959	0,986	0,998
3,084	0,821	0,978	0,999	1	1	1
4,112	0,966	0,997	1	1	1	1
5,14	0,996	1	1	1	1	1
6,188	1	1	1	1	1	1
7,196	1	1	1	1	1	1
8,224	1	1	1	1	1	1
9,252	1	1	1	1	1	1
10,28	1	1	1	1	1	1

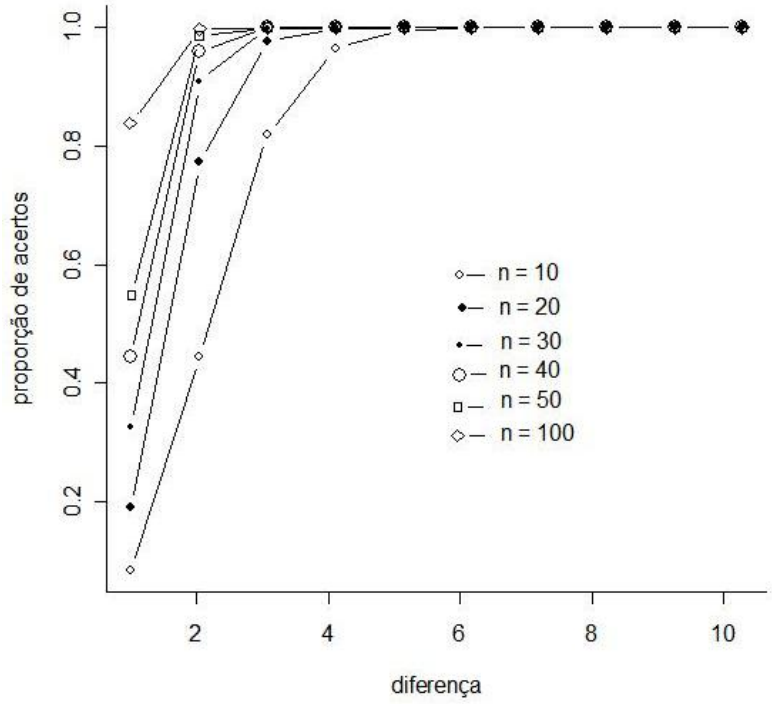

Figura 20 - Proporção de acertos em função do tamanho da amostra para uma dada diferença entre os valores dos betas.

Pela tabela 12 e Figura 20 pode-se observar que: i) para pequenas diferenças entre β_2^* e β_1^*, isto é, $\beta_2^* - \beta_1^* \cong 1$ é necessário que o tamanho da amostra seja grande $(n \cong 100)$ para que o método de agrupamento de Ward classifique corretamente os modelos dentro de cada grupo; ii) quando a diferença entre β_2^* e β_1^* é aproximadamente 4 ($\beta_2^* - \beta_1^* \cong 4$), o método de agrupamento apresenta alta eficiência, ou seja, proporção de acerto próximo a 1, mesmo quando o tamanho da amostra é pequeno; iii) em geral, a medida que o tamanho da amostra aumenta eficiência do agrupamento para a identidade de modelos também aumenta; iv) quando $\beta_2^* - \beta_1^* > 5{,}14$ a eficiência do método independe do tamanho da amostra.

5. CONCLUSÕES

O modelo logístico ajustado mostrou-se eficiente para descrever os tempos de falha dos conjuntos de dados 1.

A aplicação da metodologia tradicional para comparação de curvas de sobrevivência (teste *logrank*) permitiu avaliar os resultados obtidos pela metodologia proposta, ou seja, o uso do método de agrupamento de Ward para comparação de curvas de sobrevivência.

Os resultados apresentados pela análise de agrupamento são similares aos do teste *logrank*. Assim, o método de agrupamento de Ward mostrou-se como uma técnica potencial para a comparação de modelos não lineares aplicados à análise de sobrevivência.

O método de agrupamento de Ward é uma técnica simples de ser aplicada e exige somente que uma matriz de dissimilaridade seja criada e que o processo de agrupamento seja realizado. Não exige maior atenção acerca do nível de significância quando comparações múltiplas são realizadas, e se mostra menos trabalhoso do que o teste *logrank* quando vários tratamentos devem ser comparados.

O método de comparação de modelos via análise de agrupamento se aplica quando o mesmo modelo puder ser indicado para os diversos tratamentos. Portanto, será necessário avaliar/pesquisar maneiras de agrupar tratamentos de forma não-paramétrica, isto é, que não se baseie nos parâmetros estimados dos modelos ajustados.

6. REFERÊNCIAS BIBLIOGRÁFICAS

BUENO, O. C.; MORINI, M. S. C.; PAGNOCCA, F. C.; HEBLING, M. J. A.; SILVA, O. A. Sobrevivência de operárias de Atta sexdens rubropilosa Forel (Hymenoptera: Formicidae) isoladas do formigueiro e alimentadas com dietas artificiais. **An. Soc. Entomol. Brasil**, v. 26, n. 1, p. 107-113, Abril, 1997.

CHAE, S. S.; DUBIEN, J. L.; WARDE, W. D. A method of predicting the number of cluster using Rand's statistic. **Computacional Statistics & Data Analysis**, v. 50, p. 3531 – 3546, 2006.

COLLET, D. **Modelling survival data in medical research**. 1. ed. London: Chapman & Hall, 1994. 347 p.

COLOSIMO, E. A.; FERREIRA, F. F.; OLIVEIRA, M. D.; SOUSA, C. B. Empirical comparasions between Kaplan-Meier and Nelson-Aalen Survival Function Estimators. Belo Horizonte: Departamento de Estatística da UFMG, 2000. 18 p. Relatório técnico – RTP - 01/2000.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. 1. ed. São Paulo: Edgar Blücher, 2006. 370 p.

COX, D. R., HINKLEY, D.V. **Theoretical Statistics**. Chapman and Hall, London.

COX, D. R.; OAKES, D. **Analysis of Survival Data**. 1. ed. London: Chapman & Hall, 1984. 201 p.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético.** 2. ed. Viçosa: Ed. UFV, 2006.

EFRON, B. Logistic Regression, Survival Analysis and Kaplan-Meier curve. **American Statistical Association**, v. 93, n.402, p. 414 – 425, Jun. 1988.

GNANADESIKAN, R. **Methods for statistical data analysis of multivariate observations.** 2. ed. New York, John Wiley an Sons, 1997.

GUIMARÃES, C. R.; CIRILLO, M. A.; BRIGHENTI, D. M. Modelos de sobrevivência para a avaliação do tempo de vida de operárias de Apis Mellifera tratadas com diferentes dietas. **In:** RBRAS, 9, 2004, Uberlândia. p. 590 – 594.

HOSMER, D.W.; LEMESHOW, S. **Applied logistic Regression**. New York: John Wiley & Sons, 1989.

KHATREE, R.; NAIK, D. N. **Multivariate data reduction and discrimination with SAS software.** New York: John Wiley and Sons, 2000.

KLEINBAUM, D. G.; KLEIN, M. **Survival Analysis:** A self-Learning Text. 2. ed. New York: Springer, 2005. 590 p.

LAWLESS, J. F.; BABINEAU, D. Models for interval censoring and simulation-based inference for lifetime distributions. **Biometrika**,v. 93, n. 3, p. 671-686, 2006.

LAWLESS, J.F. **Statistical models and methods for time data**. New York: John Wiley and Sons, 1982. 580 p.

LEE, E. T.; WANG, J. W. **Statistical methods for survival data analysis.** 3. ed. New York: John Wiley and Sons, 2003. 513 p.

LYRA, G. B., GARCIA, B. I. L.; PIEDADE, S. M. de S., SEDIYAMA, G. C.; SENTELHAS, P. C. Regiões homogêneas e funções de distribuição de probabilidade da precipitação pluvial no estado de Táchira, Venezuela. **Pesq. agropec. bras.**, v. 41, n. 2, p. 205 - 215. Fev. 2006.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. New York: Academic Press, 1997.

MATOS JÚNIOR, D.; GONZALES, A. F.; POMPEU JÚNIOR, J.; PARAZZI, C. Avaliação de curvas de maturação de laranjas por análise de agrupamento. **Pesq. agropec. bras**., v. 34, n. 12, p. 2203 – 2209. Dez. 1999.

MELO JÚNIOR, J. C. F. de; SEDIYAMA, G. C.; FERREIRA, P. A.; LEAL, B. G. Determinação de regiões homogêneas quanto à distribuição de freqüência de chuvas no leste do estado de Minas Gerais. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 10, n.2, p. 408 – 416, 2006.

MICHAUD, P. Clustering techniques. **Future Generation Computer System**, v. 13, p.135 – 147. Abril 1997.

MINGOTI, S. A.; **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**, Editora UFMG, 2005.

PEAT, J. K.; BARTON, B. **Medical statistics:** a guide to data analysis and critical appraisal. 1. ed. BMJ books, 2005. 324 p.

PEREIRA, P. J.; PEREIRA, M. A. T. Comparação dos estimadores de Kaplan-Meier e de Nelson-Aalen para um estudo entomológico, com introdução de censura tipo I. **In:** RBRAS, 48, 2003, Lavras. p. 668 – 672.

PEREIRA, P. J.; VIVANCO, M. J. F. Avaliação da aplicação do teste log-rank em dados entomológicos quando são introduzidos mecanismos de censura. **In:** RBRAS, 47, 2002, Lavras. p. 2 – 22.

PETERNELLI, L. A.; CECON, P. R.; PETERNELLI, E. F. O; SOUZA, E. F. M; LEITE, M. S. O. Heurística do uso da análise de agrupamento para comparação de modelos de regressão. **In:** RBRAS, 50, 2005, Londrina.

R DEVELOPMENT CORE TEAM. 2008. **R**: A language and environment for statistical computing**.** Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: 2008.

REIS, P. R.; HADDAD, M. L. Distribuição de Weibull como modelo de sobrevivência de Iphiseiodes zuluagai Denmark & Muma (Acari: Phytoseiidae). **An. Soc. Entomolol. Brasil**, v.26, n. 3, p. 441- 444. Dez. 1997.

RENCHER, A. C. **Methods of multivariate analysis**. 2. ed. New York: John Wiley & Sons, 2002. 708p.

RODRIGUES, L. S.; ANTUNES, I. F.; TEIXEIRA, M. G.; SILVA, J. B. da. Divergência genética entre cultivares locais e cultivares melhoradas de feijão. **Pesq. Agropec. bras.**, v. 37, n. 9, p. 1275 – 1284. Set. 2002.

SANTORO, K. R.; VIEIRA, M. E. Q; QUEIROZ, M. L.; QUEIROZ, M. C.; BARBOSA, R. P. Efeito do tanino de stryphodendron SPP. Sobre a longevidade de abelhas Appis Mellifera L. (abelhas africanizadas). **Archivos de zootecnia**, v. 53, n. 203, p. 282 – 291. Set. 2004.

SHARMA, S. **Applied multivariate techniques**. New York: John Wiley & Sons, 1996.

SILVEIRA, L. V. de A.; PASSOS, J. R. de S.; COLOSIMO, E. A. Distribuição lognormal generalizada em dados de sobrevivência grupados. **In:** RBRAS, 48, 2003, Lavras. p. 657 – 662.

SOUZA, F.F.; QUEIRÓZ, M.A. ; DIAS, R.S.C. Divergência genética em linhagens de melancia. **Horticultura Brasileira**, v.23, p.179-183, 2005.

VASCONCELOS, E. S. de; CRUZ, C. D.; BHERING, L. L.; RESENDE JÚNIOR, M. F. R. . Método alternativo para análise de agrupamento. **Pesq. agropec. bras.**, v. 42, n.10, p. 1421 – 1428. Out. 2007.

WARD, J. H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, p. 236 – 244. Mar. 1963.

APÊNDICES

APÊNDICE A – Funções utilizadas na simulação de tratamentos segundo a distribuição de Weibull.

```
lifetime<-function(alfa,gama,tamanho.amostra)
{
n<-tamanho.amostra
tempo<-alfa*(-log(1-runif(n)))^(1/gama)
media<-alfa*gamma(1+(1/gama))
var<-(alfa^2)*((gamma(1+(2/gama)))- gamma(1+(1/gama))^2)
censtime<-c(rep(30,n))
ztimes<-pmin(tempo,censtime)
status<-as.numeric(censtime>tempo)
dados<-data.frame(tempo,status)
return (dados)
}

############################################################
#Simulação dos tratamentos
############################################################

trat1<-lifetime(12.61,1.75,200)
trat2<-lifetime(16.69,2.55,200)
trat3<-lifetime(10.53,1.86,200)
trat4<-lifetime(20.13,2.65,200)
trat5<-lifetime(6.46,2.31,200)
trat6<-lifetime(11.60,2.05,200)
```

APÊNDICE B: Estatísticas auxiliares para a determinação do número de grupos.

```
agrupa.ward<-function(dados)
{
dist<-dist(dados, method= "euclidian")
distancia<-dist^2
ESS<-0.5*distancia
saida<-hclust(ESS,method= “ward”)
plot(saida)
return(saida)
}

sqtc<-function(dados)
{
sq<-0
for (i in 1:ncol(dados))
{quad.x<- (dados[,i])^2
sum.x<- sum(dados[,i])
sq<-sq+(sum(quad.x)-((sum.x)^2)/nrow(dados))}
return(sq)
}

R2<-function(dados)
{
ward<-agrupa.ward(dados)
n<-length(ward$height)
a<-matrix(-99,-99,nrow=n,ncol=n)
for (i in 1:n)
{j<-n-i
a[i,]<-c(rep(1,i),rep(0,j))}
b<-ward$height
SSW<-a%*%b
R2<-1-SSW/sqtc(dados)
return(R2)
}


#Calcula o spr2
spr2<-function(saida.dendrograma, dados)
{ cor.semiparcial<-NULL
cor.semiparcial<-saida.dendrograma$height/sqtc(dados)
return(cor.semiparcial[])
}
```

APÊNDICE C: Demonstrações.

Seja $E(Y) = \frac{1}{1+e^{\alpha(\beta-x)}}$. Denomina-se E(Y) = f(x) = y. Desse modo, a taxa de variação média de y por unidade de variação de x, quando x varia de x_1 a $x_1 + \Delta x$ é dada por:

$$\frac{f(x_1+\Delta x)-f(x_1)}{\Delta x} = \frac{\Delta y}{\Delta x}.$$

Da definição acima tem-se que a taxa de variação de E(Y) por unidade de variação em x, denominada taxa de crescimento/decrescimento médio, α_m, será dada por:

$$\alpha_m = \frac{\frac{1}{1+e^{\alpha[\beta-(x_1+\Delta x)]}} - \frac{1}{1+e^{\alpha(\beta-x_1)}}}{\Delta x} = \frac{\frac{1}{1+e^{\alpha(\beta-x_1)}e^{-\alpha\Delta x}} - \frac{1}{1+e^{\alpha(\beta-x_1)}}}{\Delta x} =$$

Fazendo $e^{\alpha(\beta-x_1)} = k$, vem:

$$\alpha_m = \frac{\frac{1}{1+ke^{-\alpha\Delta x}} - \frac{1}{1+k}}{\Delta x} = \frac{1}{\Delta x}\left[\frac{1}{1+ke^{-\alpha\Delta x}} - \frac{1}{1+k}\right]$$

Tem-se que Δx é uma quantidade positiva muito pequena $(\Delta x > 0)$.

Assim, se:

$$\frac{1}{1+ke^{-\alpha\Delta x}} > \frac{1}{1+k} \Rightarrow \alpha_m > 0$$

Para isso:

$$1+ke^{-\alpha\Delta x} < 1+k \Rightarrow e^{-\alpha\Delta x} < 1 \Rightarrow -\alpha\Delta x < 0 \Rightarrow \alpha > 0$$

Por outro lado, se:

$$\frac{1}{1+ke^{-\alpha\Delta x}} < \frac{1}{1+k} \Rightarrow \alpha_m < 0$$

Para isso:

$$1+ke^{-\alpha\Delta x} > 1+k \Rightarrow e^{-\alpha\Delta x} > 1 \Rightarrow -\alpha\Delta x > 0 \Rightarrow \alpha < 0$$

Logo tem-se que:

$$\alpha_m \begin{cases} > 0 \Leftrightarrow \alpha > 0, \text{função crescente} \\ < 0 \Leftrightarrow \alpha < 0, \text{função decrescente} \end{cases}$$

Portanto, α relaciona-se com a taxa média de crescimento/decrescimento da curva.

Para demonstrar que β relaciona-se com o ponto de inflexão da curva, E(Y) será denominado $\pi(x)$, ou seja, $E(Y) = \pi(x) = \frac{1}{1+e^{\alpha(\beta-x)}}$. Define-se como ponto de inflexão, aquele cuja derivada de segunda ordem é igual a zero.

Seja $\pi'(x)$ a derivada primeira da função $\pi(x)$.

$$\pi'(x) = \frac{-\left(-\alpha \cdot e^{\alpha(\beta-x)}\right)}{\left[1+e^{\alpha(\beta-x)}\right]^2} = \alpha \frac{1}{1+e^{\alpha(\beta-x)}} \frac{e^{\alpha(\beta-x)}}{1+e^{\alpha(\beta-x)}}$$

Observação: $1-\pi(x) = 1 - \frac{1}{1+e^{\alpha(\beta-x)}} = \frac{1+e^{\alpha(\beta-x)}-1}{1+e^{\alpha(\beta-x)}} = \frac{e^{\alpha(\beta-x)}}{1+e^{\alpha(\beta-x)}}$

Assim, tem-se que:

$$\pi'(x) = \frac{-\left(-\alpha \cdot e^{\alpha(\beta-x)}\right)}{\left[1+e^{\alpha(\beta-x)}\right]^2} = \alpha \frac{1}{1+e^{\alpha(\beta-x)}} \frac{e^{\alpha(\beta-x)}}{1+e^{\alpha(\beta-x)}} = \alpha\pi(x)[1-\pi(x)]$$

Obtida a derivada de primeira ordem, determina-se a derivada segunda.

$$\pi''(x) = \frac{d}{dx}[\alpha\pi(x)(1-\pi(x))] = \alpha\left\{\pi(x)\left(\frac{d}{dx}(1-\pi(x))\right) + (1-\pi(x))\left(\frac{d}{dx}(\pi(x))\right)\right\} =$$

$$\alpha\left\{\pi(x)\left(-\frac{d}{dx}(\pi(x))\right) + (1-\pi(x))(\alpha\pi(x)(1-\pi(x)))\right\} =$$

$$\alpha\left\{-\pi(x)\cdot(\alpha\cdot\pi(x)\cdot(1-\pi(x)))+(\alpha\cdot\pi(x)\cdot(1-\pi(x)))-\left(\alpha\cdot\pi^2(x)\cdot(1-\pi(x))\right)\right\}=$$

$$\alpha\left\{-\alpha\pi^2(x)(1-\pi(x))+\alpha\pi(x)(1-\pi(x))-\alpha\pi^2(x)(1-\pi(x))\right\}=$$

$$\alpha\left\{\alpha\pi(x)(1-\pi(x))-2\alpha\pi^2(x)(1-\pi(x))\right\}=\alpha\{\alpha\pi(x)(1-\pi(x))[1-2\pi(x)]\}=$$

$$\alpha^2\pi(x)(1-\pi(x))[1-2\pi(x)]$$

$$\therefore \pi''(x)=\alpha^2\pi(x)(1-\pi(x))[1-2\pi(x)]$$

Ponto de inflexão:

$$\pi''(x)=0\Rightarrow \quad \pi(x)(1-\pi(x))[1-2\pi(x)]=0$$

Como $0<\pi(x)(1-\pi(x))<1$, deve-se ter:

$$1-2\pi(x)=0\Rightarrow\ \pi(x)=\frac{1}{2}\Rightarrow\frac{1}{1+e^{\alpha(\beta-x)}}=\frac{1}{2}\Rightarrow$$

$$1+e^{\alpha(\beta-x)}=2\Rightarrow \quad e^{\alpha(\beta-x)}=1\Rightarrow\alpha(\beta-x)=0\Rightarrow x=\beta$$

Portanto β relaciona-se com o ponto de inflexão da curva.

APÊNDICE D – Estimativas da sobrevivência para os tempos de sobrevivência usando o estimador Kaplan-Meier, modelos exponencial, de Weibull, log-normal e logístico.

TRATAMENTO 1

Tempos	Kaplan-Meier	Exponencial	Weibull	log-normal	Logístico
1	0.980	0.915	0.988	0.999	0.929
2	0.925	0.837	0.961	0.980	0.910
3	0.895	0.766	0.922	0.934	0.886
4	0.825	0.701	0.875	0.868	0.857
5	0.795	0.642	0.820	0.791	0.821
6	0.710	0.587	0.761	0.712	0.779
7	0.660	0.537	0.700	0.636	0.731
8	0.590	0.492	0.637	0.566	0.676
9	0.535	0.450	0.575	0.502	0.617
10	0.500	0.412	0.514	0.444	0.553
11	0.450	0.376	0.455	0.393	0.488
12	0.405	0.345	0.400	0.348	0.423
13	0.340	0.316	0.348	0.309	0.360
14	0.300	0.289	0.301	0.274	0.302
15	0.250	0.264	0.258	0.243	0.250
16	0.220	0.242	0.219	0.216	0.204
17	0.185	0.221	0.185	0.193	0.165
18	0.140	0.202	0.155	0.172	0.132
19	0.115	0.185	0.129	0.154	0.104
20	0.090	0.170	0.106	0.139	0.082
21	0.070	0.155	0.087	0.124	0.064
22	0.060	0.142	0.071	0.111	0.050
23	0.055	0.130	0.057	0.100	0.039
24	0.050	0.119	0.046	0.090	0.030
25	0.025	0.109	0.036	0.081	0.024
26	0.010	0.100	0.029	0.074	0.018
30	0.005	0.070	0.010	0.050	0.006

TRATAMENTO 2

Tempos	Kaplan-Meier	Exponencial	Weibull	log-normal	Logístico
1	0.995	0.936	0.999	0.999	0.978
3	0.985	0.819	0.986	0.999	0.961
5	0.950	0.717	0.955	0.977	0.934
6	0.925	0.670	0.929	0.948	0.915
7	0.890	0.627	0.897	0.906	0.891
8	0.860	0.587	0.858	0.851	0.860
9	0.785	0.549	0.813	0.789	0.823
10	0.715	0.513	0.763	0.721	0.779
11	0.665	0.480	0.708	0.652	0.728
12	0.615	0.449	0.651	0.585	0.668
13	0.560	0.420	0.589	0.520	0.604
14	0.485	0.393	0.528	0.460	0.536
15	0.415	0.369	0.467	0.405	0.466
16	0.350	0.344	0.407	0.355	0.397
17	0.300	0.322	0.351	0.310	0.333
18	0.275	0.301	0.297	0.270	0.274
19	0.230	0.282	0.249	0.235	0.222
20	0.195	0.264	0.205	0.204	0.177
21	0.145	0.247	0.166	0.177	0.140
22	0.115	0.231	0.132	0.153	0.110
23	0.105	0.216	0.104	0.133	0.085
24	0.075	0.202	0.080	0.115	0.066
25	0.050	0.189	0.061	0.100	0.051
26	0.040	0.177	0.045	0.086	0.039
27	0.030	0.165	0.034	0.075	0.030
28	0.025	0.155	0.024	0.065	0.023
29	0.015	0.145	0.017	0.056	0.017
30	0.015	0.135	0.012	0.049	0.013

TRATAMENTO 3

Tempos	Kaplan-Meier	Exponencial	Weibull	log-normal	Logístico
3	0.950	0.723	0.908	0.965	0.866
4	0.895	0.651	0.848	0.901	0.822
5	0.710	0.585	0.779	0.808	0.767
6	0.570	0.525	0.704	0.703	0.701
7	0.485	0.472	0.626	0.598	0.625
8	0.405	0.424	0.549	0.500	0.543
9	0.375	0.381	0.474	0.414	0.459
10	0.305	0.342	0.403	0.340	0.376
11	0.265	0.307	0.338	0.278	0.301
12	0.245	0.276	0.279	0.227	0.235
13	0.210	0.248	0.228	0.185	0.179
14	0.195	0.223	0.183	0.150	0.135
15	0.160	0.200	0.145	0.122	0.100
16	0.145	0.180	0.113	0.100	0.073
17	0.085	0.161	0.087	0.082	0.053
18	0.060	0.145	0.066	0.067	0.039
19	0.050	0.130	0.050	0.055	0.028
20	0.045	0.117	0.037	0.045	0.020
21	0.040	0.105	0.027	0.037	0.014
22	0.025	0.094	0.020	0.031	0.010
23	0.015	0.085	0.013	0.025	0.007
24	0.010	0.076	0.010	0.021	0.005
26	0.005	0.061	0.005	0.015	0.003
30	0.005	0.040	0.001	0.007	0.001

TRATAMENTO 4

Tempos	Kaplan-Meier	Exponencial	Weibull	log-normal	Logístico
2	0.985	0.898	0.998	0.999	0.976
3	0.980	0.851	0.994	0.999	0.969
4	0.970	0.806	0.986	0.997	0.962
5	0.965	0.764	0.975	0.988	0.952
6	0.955	0.724	0.960	0.973	0.940
7	0.950	0.686	0.941	0.948	0.925
8	0.925	0.650	0.917	0.914	0.908
9	0.865	0.616	0.888	0.873	0.889
10	0.840	0.584	0.855	0.826	0.860
11	0.805	0.554	0.817	0.775	0.829
12	0.765	0.525	0.776	0.721	0.793
13	0.710	0.497	0.731	0.667	0.752
14	0.660	0.471	0.683	0.614	0.706
15	0.595	0.446	0.632	0.563	0.655
16	0.520	0.423	0.580	0.513	0.600
17	0.475	0.401	0.528	0.467	0.543
18	0.440	0.380	0.475	0.423	0.484
19	0.390	0.360	0.424	0.383	0.426
20	0.365	0.341	0.374	0.346	0.370
21	0.330	0.323	0.327	0.312	0.317
22	0.285	0.306	0.282	0.281	0.268
23	0.255	0.290	0.241	0.253	0.225
24	0.195	0.275	0.203	0.228	0.187
25	0.180	0.261	0.169	0.205	0.154
26	0.130	0.247	0.139	0.184	0.126
27	0.095	0.234	0.113	0.165	0.102
28	0.075	0.222	0.091	0.149	0.082
29	0.055	0.210	0.072	0.133	0.066
30	0.045	0.199	0.056	0.120	0.053

TRATAMENTO 5

Tempos	Kaplan-Meier	Exponencial	Weibull	log-normal	Logístico
2	0.975	0.705	9.353e-01	0.992	9.308e-01
3	0.860	0.591	8.431e-01	0.920	8.608e-01
4	0.675	0.496	7.177e-01	0.753	7.401e-01
5	0.385	0.417	5.739e-01	0.550	5.674e-01
6	0.245	0.350	4.290e-01	0.371	3.766e-01
7	0.165	0.293	2.988e-01	0.237	2.177e-01
8	0.140	0.246	1.931e-01	0.147	1.136e-01
9	0.110	0.207	1.155e-01	0.090	5.575e-02
10	0.040	0.174	6.370e-02	0.054	2.648e-02
11	0.035	0.146	3.233e-02	0.033	1.237e-02
12	0.020	0.122	1.506e-02	0.020	5.737e-03
14	0.015	0.086	2.502e-03	0.007	1.223e-03
15	0.010	0.073	8.888e-04	0.004	5.636e-04
17	0.005	0.051	8.430e-05	0.002	1.196e-04
18	0.000	0.042	2.243e-05	0.001	5.511e-05

TRATAMENTO 6

Tempos	Kaplan-Meier	Logístico
0	1.000	0.981
4	0.995	0.948
5	0.986	0.934
6	0.980	0.916
7	0.965	0.894
8	0.944	0.867
9	0.859	0.835
10	0.758	0.796
11	0.667	0.751
12	0.601	0.700
13	0.540	0.644
14	0.475	0.583
15	0.444	0.520
16	0.394	0.456
17	0.359	0.393
18	0.318	0.334
19	0.293	0.280
20	0.232	0.231
21	0.212	0.189
22	0.192	0.153
23	0.157	0.122
24	0.111	0.097
25	0.096	0.077
26	0.086	0.061
28	0.071	0.037
29	0.066	0.029
30	0.040	0.023

TRATAMENTO 7

Tempos	Kaplan-Meier	Exponencial	Weibull	Log-normal	Logístico
1	0.995	0.907	0.993	0.999	0.950
2	0.990	0.823	0.973	0.995	0.931
3	0.915	0.748	0.939	0.971	0.906
4	0.885	0.678	0.893	0.918	0.873
5	0.820	0.615	0.837	0.842	0.831
6	0.740	0.558	0.772	0.752	0.779
7	0.650	0.506	0.701	0.659	0.716
8	0.560	0.460	0.627	0.570	0.644
9	0.485	0.417	0.552	0.488	0.564
10	0.415	0.378	0.478	0.415	0.481
11	0.325	0.343	0.408	0.351	0.399
12	0.280	0.312	0.342	0.297	0.322
13	0.220	0.283	0.283	0.250	0.253
14	0.215	0.257	0.230	0.210	0.195
15	0.185	0.233	0.184	0.177	0.149
16	0.145	0.211	0.145	0.149	0.111
17	0.115	0.192	0.112	0.126	0.082
18	0.090	0.174	0.085	0.106	0.060
19	0.070	0.158	0.064	0.090	0.044
20	0.045	0.143	0.047	0.076	0.032
21	0.035	0.130	0.034	0.065	0.023
22	0.025	0.118	0.024	0.055	0.016
23	0.005	0.107	0.017	0.047	0.012
30	0.005	0.054	0.001	0.016	0.001

TRATAMENTO 8

Tempos	Kaplan-Meier	Logístico
0	0.995	0.982
1	0.990	0.978
3	0.985	0.966
4	0.965	0.957
5	0.960	0.946
6	0.955	0.933
7	0.945	0.918
8	0.920	0.898
9	0.870	0.875
10	0.835	0.848
11	0.805	0.815
12	0.745	0.778
13	0.685	0.736
14	0.615	0.688
15	0.570	0.637
16	0.510	0.582
17	0.460	0.525
18	0.420	0.467
19	0.385	0.410
20	0.345	0.356
21	0.290	0.305
22	0.265	0.258
23	0.235	0.216
24	0.210	0.180
25	0.185	0.148
26	0.170	0.121
27	0.120	0.099
28	0.090	0.080
29	0.055	0.065
30	0.040	0.052