

Análise de Sobrevivência: aplicação num estudo de podridão parda em pêssegos com o uso do ambiente





Governador do Estado
João Raimundo Colombo

Vice-Governador do Estado
Eduardo Pinho Moreira

**Secretário de Estado da
Agricultura e da Pesca**
Moacir Sopelsa

Presidente da Epagri
Luiz Ademir Hessmann

Diretores

Ivan Luiz Zilli Bacic
Desenvolvimento Institucional

Giovani Canola Teixeira
Administração e Finanças

Luiz Antonio Palladini
Ciência, Tecnologia e Inovação

Paulo Roberto Lisboa Arruda
Extensão Rural



BOLETIM TÉCNICO Nº 180

**Análise de Sobrevivência: aplicação num estudo de podridão
parda em pêsegos com o uso do ambiente R.**

**Cristiano Nunes Nesi
Giselda Alves
Silvia Emiko Shimakura
Louise Larissa May De Mio
Paulo Justiniano Ribeiro Júnior**



**EMPRESA DE PESQUISA AGROPECUÁRIA E EXTENSÃO RURAL DE
SANTA CATARINA
FLORIANÓPOLIS
2017**

Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (Epagri)
Rodovia Admar Gonzaga, 1347, Itacorubi, Caixa Postal 502
88034-901 Florianópolis, SC, Brasil
Fone: (48) 3665-5000, fax: (48) 3665-5010
Site: www.epagri.sc.gov.br

Editado pelo Departamento Estadual de Marketing e Comunicação (DEMC) / Epagri

Assessoria técnico-científica: Simone Silmara Werner

Editoração técnica: Lucia Morais Kinceler

Revisão textual: Abel Viana

Arte de capa: Victor Berretta

Foto de capa: Frutos de pêssego com sintomas de podridão parda (Foto: Giselda Alves).

Primeira edição *on line*: novembro 2017

Impressão: Gráfica CS

Tiragem: 500 exemplares

É permitida a reprodução parcial deste trabalho desde que a fonte seja citada.

Ficha catalográfica

NESI, C.N.; ALVES, G.A.; SHIMAKURA, S.E.; MAY DE MIO, L.L.; RIBEIRO JR, P.J. *Análise de Sobre-
vivência: aplicação num estudo de podridão parda em
pêssegos com o uso do ambiente R*. Florianópolis: Epagri,
2017. 30p. (Epagri. Boletim Técnico, 180)

Tempo ocorrência; Sobrevida; *Monilinia fructicola*

ISSN 0100-7416

AUTORES

CRISTIANO NUNES NESI

Engenheiro-agrônomo, Dr.
Epagri, Centro de Pesquisa para Agricultura Familiar
Servidão Ferdinando Tusset, s/n
89801 970 - Chapecó, SC
cristiano@epagri.sc.gov.br

GISELDA ALVES

Engenheira-agrônoma, Dra. em Agronomia
Universidade Federal do Paraná, Setor de Ciências Agrárias, Departamento de Fitotecnia
e Fitossanitarismo - Laboratório de Fitopatologia
Rua dos Funcionários, 1540 - Cabral
80035 050 - Curitiba, PR
alves.giselda@gmail.com

SILVIA EMIKO SHIMAKURA

Estatística, PhD in Statistics
Laboratório de Estatística e Geoinformação
Setor de Ciências Exatas
Universidade Federal do Paraná
Jardim das Américas, Caixa Postal 19081
81531 990 Curitiba, PR
silvia.shimakura@ufpr.br

LOUISE LARISSA MAY DE MIO

Engenheira-agrônoma, Dra. em Agronomia
Universidade Federal do Paraná, Setor de Ciências Agrárias, Departamento de Fitotecnia
e Fitossanitarismo - Laboratório de Fitopatologia
Rua dos Funcionários, 1540 - Cabral
80035 050 - Curitiba, PR
maydemio@ufpr.br

PAULO JUSTINIANO RIBEIRO JUNIOR

Laboratório de Estatística e Geoinformação
Setor de Ciências Exatas
Universidade Federal do Paraná
Jardim das Américas, Caixa Postal 19081
81531 990 Curitiba, PR
paulojus@ufpr.br

APRESENTAÇÃO

Este material aborda uma aplicação da técnica estatística denominada "Análise de Sobrevivência" no estudo da podridão parda em pêssegos. Apesar de bem conhecida e muito aplicada na área da medicina humana, esta técnica ainda tem poucas aplicações em estudos agronômicos.

No estudo de doenças de plantas, uma situação comum é aquela em que se avalia a ocorrência da doença e o tempo para sua ocorrência em uma população de indivíduos. Entretanto, o tempo de observação raramente tem distribuição normal e os dados podem ser censurados, ou seja, o estudo pode terminar antes que todos os indivíduos avaliados sofram o evento de interesse, observando-se parcialmente a resposta.

Frequentemente, dados desta natureza são submetidos à análise estatística convencional com exclusão de observações censuradas e transformação de dados. A análise de sobrevivência é aplicada em situações em que o tempo até a ocorrência de um evento é o objeto de interesse. Dados dessa natureza são rotineiramente coletados em epidemiologia de doenças de plantas, embora aplicações da técnica sejam pouco comuns.

Neste contexto, o objetivo deste boletim é descrever as principais técnicas em análise de sobrevivência e sua implementação em ambiente R, com aplicação num estudo sobre sintomas de podridão parda em pêssegos causada por *Monilinia fructicola*.

Espera-se que este material auxilie pesquisadores nas diversas áreas, utilizando uma técnica adequada à natureza dos dados de tempo-ocorrência, além de disponibilizar códigos em ambiente R para facilitar a análise e discussão que colabore na interpretação dos resultados.

SUMÁRIO

Introdução	9
1 Análise de sobrevivência	11
1.1 Funções de sobrevivência e de risco	11
1.2 Técnica não paramétrica	12
1.3 Modelo semiparamétrico de sobrevivência	14
1.4 Seleção de modelos e diagnóstico de ajuste	16
2 Aplicação	18
2.1 Descrição do estudo	18
2.2 Análise dos dados	18
2.2.1 Curvas de Kaplan-Meier	21
2.2.2 Teste <i>log-rank</i>	24
2.3 Modelo semiparamétrico	26
2.4 Verificação dos resíduos do modelo	33
3 Considerações finais	36
Referências	37

Introdução

A análise da ocorrência de um evento e do tempo para sua ocorrência em uma população de indivíduos é um problema estatístico comum. Neste contexto, um evento é definido como uma mudança qualitativa do indivíduo observado, que ocorre em um ponto particular no tempo (SCHERM & OJIAMBO, 2004). Na área médica, frequentemente o evento de interesse é o tempo até a cura ou morte do indivíduo, medido a partir de um particular tratamento ou do início da doença (McGILCHRIST & AISBETT, 1991; GOEL et al., 2010). Para situações como estas foi desenvolvida uma técnica de análise estatística denominada análise de sobrevivência, aplicada quando o tempo até a ocorrência de um evento (variável dependente) for o objeto de interesse (CARVALHO et al., 2011). Técnicas de análise estatística convencionais não são apropriadas para este tipo de dados, pois o tempo de observação raramente tem distribuição normal e os dados podem ser censurados, ou seja, o estudo pode terminar antes que todos os indivíduos avaliados sofram o evento de interesse (censura à direita), observando-se parcialmente a resposta (BEWICK et al., 2004). Geralmente, dados desta natureza são submetidos à análise estatística convencional, o que limita a capacidade de inferência. Descartar observações censuradas reduz o poder dos testes estatísticos devido à perda de graus de liberdade e introduz viés nas funções de sobrevivência (COLOSIMO & GIOLO, 2006), além de superestimar o risco, pois o tempo até a ocorrência do evento é desconhecido (CARVALHO et al., 2011). Neste sentido, a análise de sobrevivência utiliza o método da verossimilhança para estimação de parâmetros e efetivamente extrair informações relevantes e estimativas confiáveis mesmo em situações com censura (COLOSIMO & GIOLO, 2006). Apesar de o termo ter origem em estudos na área da saúde, a metodologia de análise de sobrevivência é aplicada em muitas áreas do conhecimento, como em demografia (OLIVEIRA et al., 2006), economia (OLIVEIRA & RIOS-NETO, 2007), entomologia (KRÜGER et al., 2008), agronomia (COUTO et al., 2009), educação (LIMA JUNIOR et al., 2012), fitopatologia (NESI et al., 2013; NESI et al., 2015), tecnologia de sementes (NESI et al., 2016), entre outros.

No estudo de doenças de plantas, apesar de dados sobre ocorrência e tempo serem rotineiramente coletados em laboratórios ou campos experimentais, a análise de sobrevivência é ainda pouco comum. Aplicações desta técnica em doenças de plantas podem ser observadas em Dallot et al. (2004), que identificaram fatores de risco na infecção

de pessegueiros por *Plum pox virus* no tempo. Como resultado, afetaram a persistência da doença. Ojiambo e Scherm (2005) empregaram a técnica no estudo do tempo para abscisão de folhas de mirtilo em função da severidade da mancha foliar causada por *Septoria albopunctata*, considerando conjuntamente a idade e a localização da folha na copa. Copes e Thomson (2008) utilizaram essa análise para determinar o período de incubação da ferrugem causada por *Colletotrichum gloesporioides* em ramos de camélia. O período latente de *Mycosphaerella pinodes* em ervilhas foi estimado utilizando análise de sobrevivência, considerando-se como variáveis explicativas a agressividade do isolado, a duração do molhamento foliar, a concentração do inóculo, a idade da planta e a susceptibilidade do hospedeiro no trabalho de Setti et al. (2010). Nos trabalhos expostos, a utilização de análise de sobrevivência para dados em doenças de plantas não abordam situações comuns como a possibilidade de acomodar a variação entre safras numa análise conjunta ou a existência de agrupamentos naturais como frutos de uma mesma árvore, o que pode influenciar o tempo para ocorrência do evento (GORFINE et al., 2006) e invalidar a suposição de independência dos tempos entre os indivíduos (COLOSIMO & GIOLO, 2006).

Em pessegueiros, a podridão parda, causada por *Monilinia fructicola*, é uma doença mundialmente importante, pois em frutos maduros e aparentemente sadios a doença pode aparecer alguns dias depois devido à infecção latente. Conforme abordado, técnicas de análise estatística convencional não são apropriadas para dados de tempo para expressar sintomas, pois pode ocorrer censura, ou seja, o tempo de observação termina antes de que todos os frutos apresentem o sintoma da podridão. Além disso, tempo de sobrevivência frequentemente não segue distribuição normal.

O objetivo deste trabalho foi descrever as principais técnicas em análise de sobrevivência e sua implementação em ambiente R, com aplicação num estudo para avaliar o efeito de diferentes momentos de proteção física de frutos, em pessegueiro, sobre os sintomas de podridão parda causada por *Monilinia fructicola*.

Neste estudo, o tempo para expressar os sintomas em frutos, que é a variável dependente, é interpretado como um tempo de sobrevivência, e o conjunto de dados contém observações censuradas, pois no final do estudo ainda há frutos sem sintomas de podridão parda.

1 Análise de sobrevivência

O modelo de sobrevivência, como qualquer modelo de regressão, é composto por uma variável resposta, covariáveis explicativas, a função de ligação e a estrutura de erro (CARVALHO et al., 2011). A variável resposta pode ser expressa de três formas interconversíveis: probabilidade de sobrevivência, taxa de incidência e taxa de incidência acumulada. A probabilidade de sobrevivência é definida como a probabilidade de não ocorrer o evento de interesse dentro do intervalo de tempo $[t_0, t]$. Por definição, essa probabilidade é 1 no início da observação (em t_0). Normalmente, costuma-se utilizar uma transformação em escala tal que a probabilidade de sobrevivência seja interpretada como frequência de sobreviventes no tempo t em relação ao total de indivíduos sob estudo, ao invés de uma probabilidade. A taxa de incidência é definida como a taxa instantânea de ocorrência do evento de interesse no tempo t , condicionada à sua não ocorrência anterior a t . A incidência acumulada difere do conceito de risco, embora sejam aproximadamente iguais quando a incidência é baixa.

1.1 Funções de sobrevivência e de risco

A função de sobrevivência $S(t)$ descreve a probabilidade de um indivíduo sobreviver por mais de um determinado tempo t , ou seja, a probabilidade de que, decorrido um tempo t , o indivíduo não apresente o evento. Isto é definido como $S(t) = Pr(T > t)$, em que T é o tempo até a ocorrência de um evento. De maneira semelhante, a distribuição acumulada do tempo de vida $F(t)$ é a probabilidade de que o indivíduo expresse o evento antes do tempo t . A partir dessa definição, $S(t)$ é simplesmente o complemento da função de distribuição acumulada:

$$S(t) = P(T > t) = 1 - Pr(T \leq t) = 1 - F(t).$$

A função $\lambda(t)$ expressa o risco instantâneo de um indivíduo expressar o evento no tempo t , condicional à permanecer sem o evento até o tempo t :

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{Pr[(t \leq T < t + \Delta t) \mid T \geq t]}{\Delta t}.$$

A função de risco é também chamada de função ou taxa de incidência, força de infecção, taxa de falha, força de mortalidade ou força de mortalidade condicional. Apesar de se utilizar o nome risco, $\lambda(t)$ é uma taxa e não uma probabilidade. Sua unidade é tempo⁻¹ e pode assumir qualquer valor real maior que zero, não estando restrita ao intervalo $[0, 1]$. A função de risco pode ser relacionada com a função de sobrevivência de várias formas, entre elas

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

em que $f(t)$ é a função densidade de probabilidade interpretada como a probabilidade de um indivíduo sofrer um evento em um intervalo instantâneo de tempo (COLOSIMO & GIOLO, 2006; CARVALHO et al., 2011).

De maneira geral, a análise de sobrevivência visa obter estimadores das funções usadas para especificar o tempo de vida, testando a dependência dessas funções com covariáveis.

1.2 Técnica não paramétrica

A principal técnica não paramétrica em análise de sobrevivência é o estimador de Kaplan-Meier, proposto por Kaplan e Meier (1958), que traz uma solução para a estimação de $S(t)$ quando há indivíduos censurados. Neste método, a função de sobrevivência é recalculada após cada fruto observado expressar sintomas. Ele utiliza os conceitos de independência de eventos e de probabilidade condicional para desdobrar a condição de sobreviver até o tempo t em uma sequência de elementos independentes que caracterizam a sobrevivência em cada intervalo de tempo anterior a t , cuja probabilidade é condicional aos indivíduos que estão em risco no período (AKBAR et al., 2009; GOEL et al., 2010; CARVALHO et al., 2011). Assume-se que as ocorrências dos eventos são independentes entre os indivíduos. Conseqüentemente, a função de sobrevivência é estimada pelo produto das probabilidades de permanecer sem o evento até o tempo t . A função de sobrevivência

$S(t)$ é estimada empiricamente por

$$\widehat{S}(t) = \prod_{i=1}^n \left(1 - \frac{d_i}{n_i}\right),$$

em que d_i é o número de indivíduos que sofreram o evento em um dado tempo t_i ($i = 1, \dots, k$) e n_i é o número de indivíduos sob risco no tempo t_i , ou seja, os indivíduos que não sofreram o evento e não foram removidos do estudo (censurados) até o instante imediatamente anterior a t_i .

O gráfico de $S(t)$ em função do tempo (t) é chamado curva de sobrevivência. O método de Kaplan-Meier estima esta curva a partir dos tempos de sobrevivência sem que seja necessário pressupor uma distribuição de probabilidades mesmo quando há censura no conjunto de observações (AKBAR et al., 2009). Por convenção, os gráficos de Kaplan-Meier são representados com degraus para indicar os instantes do tempo em que ocorrem eventos terminais (sintomas) e sinais (+) para indicar observações censuradas (CARVALHO et al., 2011). A abordagem não paramétrica da análise de sobrevivência, utilizando o estimador de Kaplan-Meier, permite realizar testes de significância estatística para comparar tratamentos (AKBAR et al., 2009; GOEL et al., 2010). Neste contexto, o teste log-rank (LR) é o mais usado, mas deve ser aplicado apenas para comparar grupos definidos por variáveis categóricas (AKBAR et al., 2009) e quando a razão das funções de risco dos tratamentos a serem comparados é aproximadamente constante, característica denominada de riscos proporcionais (COLOSIMO & GIOLO, 2006). Com o teste log-rank, testa-se a hipótese nula de que não há diferença entre as curvas de sobrevivência de cada estrato. O teste log-rank (LR) compara a distribuição da ocorrência dos eventos observados em cada estrato com a distribuição que seria esperada se a incidência fosse igual em todos os estratos. A estatística do teste é calculada por

$$LR = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i},$$

em que O e E são, respectivamente, o total de eventos observados e esperados em cada estrato, em que os esperados são obtidos supondo que a hipótese nula é verdadeira (CARVALHO et al., 2011). Para comparação de apenas dois tratamentos, o valor da estatística log-rank tem distribuição de χ^2 com um grau de liberdade (AKBAR et al., 2009). O teste log-rank pode ser estendido para mais de dois tratamentos, embora em várias situações

seja necessário estudar, simultaneamente, a relação entre o tempo de vida e diversas variáveis explicativas, além de categorizar variáveis contínuas para aplicação do teste. Para contornar estes inconvenientes e tornar a análise mais explicativa, utilizam-se os modelos de regressão paramétricos ou semiparamétricos.

1.3 Modelo semiparamétrico de sobrevivência

A regressão paramétrica exige que se suponha uma distribuição estatística para o tempo de sobrevivência. Esta suposição, caso não seja adequada, pode fazer com que as estimativas dos parâmetros sejam incorretas. Entretanto, muitas vezes o interesse do estudo não é estimar os parâmetros da distribuição do tempo de sobrevivência, mas estimar o efeito de covariáveis. Foi partindo dessas questões que COX (1972) propôs um modelo denominado modelo semiparamétrico de riscos proporcionais, que passou a ser o mais utilizado na análise de dados de sobrevivência (CARVALHO et al., 2011).

O modelo de Cox de riscos proporcionais é definido em MCGILCHRIST e AIS-BETT (1991) pelo produto dos componentes não paramétrico $\lambda_0(t)$ e paramétrico $g(x^\top \theta)$

$$\lambda(t) = \lambda_0(t)g(x^\top \theta),$$

em que $\lambda(t)$ descreve o risco do indivíduo apresentar o evento de interesse no tempo t , $\lambda_0(t)$ é o risco basal no tempo t e $g(x^\top \theta)$ é o efeito multiplicativo das variáveis explicativas combinadas na função. O componente não paramétrico $\lambda_0(t)$ não é especificado e é uma função não-negativa do tempo, usualmente chamado função de risco de base ou basal, pois $\lambda(t) = \lambda_0(t)$ na ausência de covariáveis ($x = 0$). O componente paramétrico é frequentemente expresso na forma

$$g(x^\top \theta) = \exp(x^\top \theta) = \exp(\theta_1 x_1 + \dots + \theta_p x_p),$$

em que x é o vetor de variáveis explicativas e θ é o vetor de parâmetros a ser estimado (FOX & WEISBERG, 2011).

Em algumas situações não se pode assumir que o risco basal $\lambda_0(t)$ seja o mesmo para todos os indivíduos do estudo, como nos casos em que ocorre algum tipo de estratifi-

cação. Nesse caso, a variável que os separaria em diferentes grupos, com $\lambda_{0a}(t) \neq \lambda_{0b}(t) \neq \lambda_{0c}(t)$ define estes diferentes estratos.

O modelo de Cox é dito de riscos proporcionais porque as razões entre as taxas de risco dos indivíduos são assumidas constantes no tempo de acompanhamento do estudo (CARVALHO et al., 2011). Assim, a razão entre o risco de ocorrência do evento para dois indivíduos k e l , com covariáveis $x_k^\top = (x_{k1}, x_{k2}, \dots, x_{kp})$ e $x_l^\top = (x_{l1}, x_{l2}, \dots, x_{lp})$ é

$$\frac{\lambda_k(tx_k)}{\lambda_l(tx_l)} = \frac{\exp(x_k^\top \theta)}{\exp(x_l^\top \theta)},$$

ou seja, é constante no tempo.

O modelo de regressão de Cox é caracterizado pelos coeficientes θ 's, que medem os efeitos das covariáveis sobre a função de risco. Estas quantidades devem ser estimadas a partir das observações amostrais para que o modelo fique determinado.

O procedimento de regressão usado para ajustar o modelo de Cox consiste em maximizar a função de verossimilhança parcial $L(\theta)$ para o vetor de parâmetros (θ) ,

$$L(\theta) = \prod_{i=1}^n \left[\frac{\exp(x_i^\top \theta)}{\sum \exp(x_i^\top \theta)} \right]^{\delta_i},$$

em que δ_i é uma variável aleatória que indica a ocorrência ($\delta = 1$) ou censura ($\delta = 0$) do evento (COLOSIMO & GIOLO, 2006; CARVALHO et al., 2011).

Eventualmente, os indivíduos sob estudo podem apresentar grande heterogeneidade ou diferentes fragilidades não atribuíveis a qualquer característica medida. Este problema pode ser tratado com a introdução de um efeito aleatório para cada indivíduo, o que torna as estimativas dos efeitos das covariáveis mais consistentes (COLOSIMO & GIOLO, 2006; CARVALHO et al., 2011). Partindo do modelo de Cox clássico, a inclusão de efeitos aleatórios é feita a partir de uma variável aleatória desconhecida (Z), que reflete a heterogeneidade ou fragilidade do indivíduo ou grupo (McGILCHRIST e AISBETT, 1991; SARGENT, 1998) atuando multiplicativamente sobre o risco de base. Assim, para um indivíduo com covariáveis representadas pelo vetor x e efeito aleatório $Z = z$, o modelo de riscos proporcionais passa a ser

$$\lambda(t) = z\lambda_0(t)g(x^\top \beta).$$

Assume-se que os valores das fragilidades são uma amostra independente da va-

riável aleatória Z com distribuição de probabilidade conhecida de média 1 e variância desconhecida (KOSOROK et al., 2004; GORFINE et al., 2006). Para estimar a variância do efeito aleatório é necessário escolher uma distribuição estatística para a variável aleatória Z .

1.4 Seleção de modelos e diagnóstico de ajuste

Quando se estima um modelo de fragilidade, duas opções são usuais. A primeira é assumir a distribuição gama para a fragilidade, e estima-se a variância da fragilidade utilizando-se a verossimilhança perfilhada (EM). Como alternativa, assume-se fragilidade com distribuição log-normal, em que o método padrão de estimação da variância é o da máxima verossimilhança restrita aproximada (REML). Nos dois casos, a minimização do critério AIC pode ser utilizado como algoritmo de maximização (CARVALHO et al., 2011). Não há método que permita identificar se a distribuição gama é melhor que a normal para ajustar o efeito aleatório, ou se uma particular abordagem para estimação da variância da fragilidade é melhor que outra. O que se espera é que as diferentes abordagens sejam razoavelmente consistentes, apontando as mesmas unidades com risco diferenciado.

Para comparar diferentes modelos de regressão de Cox, utiliza-se o teste de Wald e teste da razão de verossimilhanças. O teste da razão de verossimilhanças compara modelos aninhados, avaliando se a inclusão de uma ou mais covariáveis no modelo aumenta de modo significativo a verossimilhança de um modelo em relação ao modelo mais parcimonioso.

Após a seleção do modelo, ainda restam algumas questões, como: o modelo se ajusta bem aos dados? Qual o poder explicativo do modelo?

A suposição básica para o modelo de Cox é a dos riscos proporcionais (CARVALHO et al., 2011), que pode ser avaliada pela visualização gráfica dos resíduos de Schoenfeld e pela significância do coeficiente de correlação linear simples entre os resíduos padronizados de Schoenfeld e o tempo para cada uma das covariáveis.

Os resíduos de Schoenfeld são a diferença entre o valor da covariável x_j para um indivíduo para o qual se verificou o acontecimento de interesse num determinado instante t_i , e uma média ponderada dos valores para essa covariável para todos os indivíduos em risco em t_i . Nos indivíduos cujo tempo de vida é censurado, estes resíduos são sempre

nulos, indicados como valores omissos. Os resíduos de Schoenfeld são úteis para aferir a premissa de riscos proporcionais subjacente ao modelo de Cox. Para além da análise gráfica, é igualmente possível averiguar a presença de uma correlação linear entre os resíduos e o tempo (ρ), através de um teste estatístico. Neste teste são consideradas as seguintes hipóteses:

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho \neq 0$$

Assim, admite-se a proporcionalidade dos riscos quando não se rejeita a hipótese nula de que a correlação é igual a zero.

2 Aplicação

2.1 Descrição do estudo

Em um pomar comercial da região metropolitana de Curitiba com pessegueiros de uma mesma cultivar, dez frutos verdes por planta foram inoculados (24 plantas) com suspensão de *M. fructicola* em cinco diferentes momentos: 17, 24, 49, 64, e 67 dias após a plena floração.

Imediatamente após a inoculação, os frutos foram envolvidos com sacos de papel parafinado branco e assim permaneceram até o momento da colheita. Na colheita, o saquinho foi retirado e os frutos permaneceram a 25°C com luminosidade contínua durante 12 dias, procurando reproduzir um ambiente propício à expressão dos sintomas de *M. fructicola*. A variável resposta foi o tempo decorrido entre a retirada do invólucro e a expressão do sintoma em cada fruto. O trabalho foi executado com a mesma metodologia nos anos 2008 e 2009. Os frutos que apresentavam sintomas da doença no momento em que o saquinho foi aberto foram descartados, pois não se sabia o momento exato em que ocorreu a expressão do sintoma. Os dados de incidência de podridão parda foram registrados diariamente.

O objetivo da análise de sobrevivência foi avaliar o efeito do momento da inoculação (número de dias após a plena floração), na sobrevivência e no risco de expressar sintomas de podridão parda em pós-colheita. Considerou-se que o risco de base pode variar entre as safras e avaliou-se a existência de um efeito comum compartilhado por frutos de uma mesma árvore.

2.2 Análise dos dados

Inicialmente as análises estatísticas foram conduzidas usando as estimativas de Kaplan-Meier estratificadas por momento de inoculação e ensacamento dos frutos para funções de sobrevivência, descrevendo o tempo para expressar os sintomas de podridão parda. Partindo do pressuposto da proporcionalidade, estimou-se o efeito dos momentos, sem fazer a suposição sobre a distribuição do tempo de sobrevivência, utilizando-se o mo-

delo semiparamétrico de Cox, com diferentes distribuições para a fragilidade e diferentes métodos de estimação.

Todas as análises foram realizadas com o ambiente R (R CORE TEAM, 2017), utilizando-se os procedimentos do pacote `survival` (THERNEAU, 2015).

Os comandos apresentados a seguir fazem a importação do conjunto de dados e sua apresentação inicial. Para o conjunto de dados deste estudo, a variável "tipo" refere-se ao "status" do fruto no momento da observação e indica ocorrência (1) ou não (0) do evento de interesse, neste caso, o sintoma de podridão parda.

```
> # pacote para análise de dados de sobrevivência
> require(survival)
> # importação dos dados
> dados=read.csv('ensacamentoBT.csv',header=T)
> # visualização das primeiras seis linhas de dados (padrão)
> head(dados)
```

```
  id Ano Tratamento Arvore tempo tipo
1  1  1          17      1     7     1
2  2  1          17      1    11     0
3  3  1          17      1    11     0
4  4  1          17      1    11     0
5  5  1          17      1    11     0
6  6  1          17      1    11     0
```

A análise de sobrevivência necessita que um tratamento seja considerado como referência para as comparações. Isto pode ser feito por critérios como o próprio desenho do estudo em que algum tratamento já é naturalmente referência (padrão ou testemunha) ou escolhido por critérios estatísticos como aquele com o maior número de observações. Neste estudo considerou-se como referência o tratamento 17 daf, em que os frutos foram inoculados e protegidos aos 17 dias após a plena floração.

```
> # definindo Tratamento, Ano e Arvore como fatores
> dados$Tratamento <- as.factor(dados$Tratamento)
> dados$Ano <- as.factor(dados$Ano)
> dados$Arvore <- as.factor(dados$Arvore)
> # definindo o tratamento 7 como "tratamento referência"
> dados$Tratamento <- relevel(dados$Tratamento, "17")
```

O tempo mediano para expressão de sintomas não pode ser estimado em nenhum tratamento ou safra, pois em todos os casos menos de 50% dos frutos apresentaram sintomas.

```

> # 'Surv' = comando para combinar, em um único objeto, a informação referente ao
> #tempo de sobrevivência e do status
> # 'survfit' = comando para estimar a tabela e a curva de sobrevivência
> #estratificados por Tratamento
> # Ano = 1: safra 2008
> (s1=survfit(Surv(tempo, tipo)~ Tratamento,data=subset(dados, Ano=="1")))

```

```

Call: survfit(formula = Surv(tempo, tipo) ~ Tratamento, data = subset(dados,
  Ano == "1"))

```

	n	events	median	0.95LCL	0.95UCL
Tratamento=17	122	3	NA	NA	NA
Tratamento=24	194	6	NA	NA	NA
Tratamento=49	199	10	NA	NA	NA
Tratamento=64	219	14	NA	NA	NA
Tratamento=67	129	18	NA	NA	NA

```

> # Ano = 2: safra 2009
> (s2=survfit(Surv(tempo, tipo)~ Tratamento,data=subset(dados, Ano=="2")))

```

```

Call: survfit(formula = Surv(tempo, tipo) ~ Tratamento, data = subset(dados,
  Ano == "2"))

```

	n	events	median	0.95LCL	0.95UCL
Tratamento=17	62	1	NA	NA	NA
Tratamento=24	74	4	NA	NA	NA
Tratamento=49	232	20	NA	NA	NA
Tratamento=64	240	43	NA	NA	NA
Tratamento=67	225	64	NA	NA	NA

```

> # Soma do número de frutos nas duas safras
> (s=survfit(Surv(tempo, tipo)~ Tratamento,data=dados))

```

```

Call: survfit(formula = Surv(tempo, tipo) ~ Tratamento, data = dados)

```

	n	events	median	0.95LCL	0.95UCL
Tratamento=17	184	4	NA	NA	NA
Tratamento=24	268	10	NA	NA	NA
Tratamento=49	431	30	NA	NA	NA
Tratamento=64	459	57	NA	NA	NA
Tratamento=67	354	82	NA	NA	NA

As saídas do programa R com o número total de frutos por tratamento e o número de frutos com sintomas nas duas safras foram organizadas e apresentados na Tabela 1.

Tabela 1: Número total de frutos e número de frutos que expressaram sintomas de *Monilinia fructicola* em função da safra e do número de dias após o florescimento (daf) em que os frutos foram inoculados e protegidos.

Tratamento (daf)	Safra 2008		Safra 2009		Soma	
	Total	Com sintomas	Total	Com sintomas	Total	Com sintomas
17	122	3	62	1	184	4
24	194	6	74	4	268	10
49	199	10	232	20	431	30
64	219	14	240	43	459	57
67	129	18	225	64	354	82

Observa-se que o número total de frutos por tratamentos varia também entre safras, pois ocorre abortamento de frutos, o que pode ser natural ou causado pelo ato de ensacá-los ainda em estágio muito frágil.

2.2.1 Curvas de Kaplan-Meier

Estimaram-se as curvas de sobrevida pelo método de Kaplan-Meier para verificar, visualmente, o risco de expressão dos sintomas em função do tempo, além do pressuposto de proporcionalidade do risco entre tratamentos. As curvas para cada tratamento foram construídas por ano e agrupando-se os anos como pode ser observado nos comandos a seguir. Os objetos `s1`, `s2` e `s` são resultados da função `survfit`, apresentados anteriormente.

```
> par(mfrow = c(2, 2))# divide a janela gráfica em duas linhas e duas coluna
> # ANO 1
> plot(s1, conf.int=F, lty=1:7, col=c("red", "green4", "blue", "tomato1", "purple4"),
+      xlab="dias", ylab="S(t)", ylim=c(0.6, 1), main="2008", lwd=2)
> # ANO 2
> plot(s2, conf.int=F, lty=1:7, col=c("red", "green4", "blue", "tomato1", "purple4"),
+      xlab="dias", ylab="S(t)", ylim=c(0.6, 1), main="2009", lwd=2)
> # 2 anos
> plot(s, conf.int=F, lty=1:7, col=c("red", "green4", "blue", "tomato1", "purple4"),
+      xlab="dias", ylab="S(t)", ylim=c(0.6, 1), main="2008 e 2009", lwd=2)
> plot(0, pch=NA_integer_, main="", axes=F, col.lab="white")
> legend("center", bty="n", seg.len=2,
+       col=c("red", "green4", "blue", "tomato1", "purple4"),
```

```
+ lty=1:7, cex=1.5, lwd=2, legend=c('17 daf', '24 daf',  
+ '49 daf', '64 daf', '67 daf'),  
+ merge=TRUE, trace=FALSE)
```

As curvas de sobrevivência estimadas por Kaplan-Meier são apresentadas por safra e conjuntamente na Figura 1. Nas duas safras, a probabilidade de os frutos permanecerem sem sintomas de podridão parda reduz com o tempo em todos os tratamentos, mas o comportamento é diferente entre os anos. A distinção no efeito de tratamentos não é clara na safra de 2008, em que a probabilidade de observar fruto sadio se mantém acima de 80%. Já na safra 2009 e conjuntamente, esta probabilidade cai para alguns tratamentos.

O método de Kaplan-Meier gera uma função escada que salta em cada tempo onde ocorreu um evento (fruto com sintoma). O tamanho desse salto depende do número de eventos observados nesse tempo ou seja, do número de frutos com sintomas em cada momento de avaliação. Essa função é uma escada porque o risco se mantém constante até a ocorrência do próximo evento.

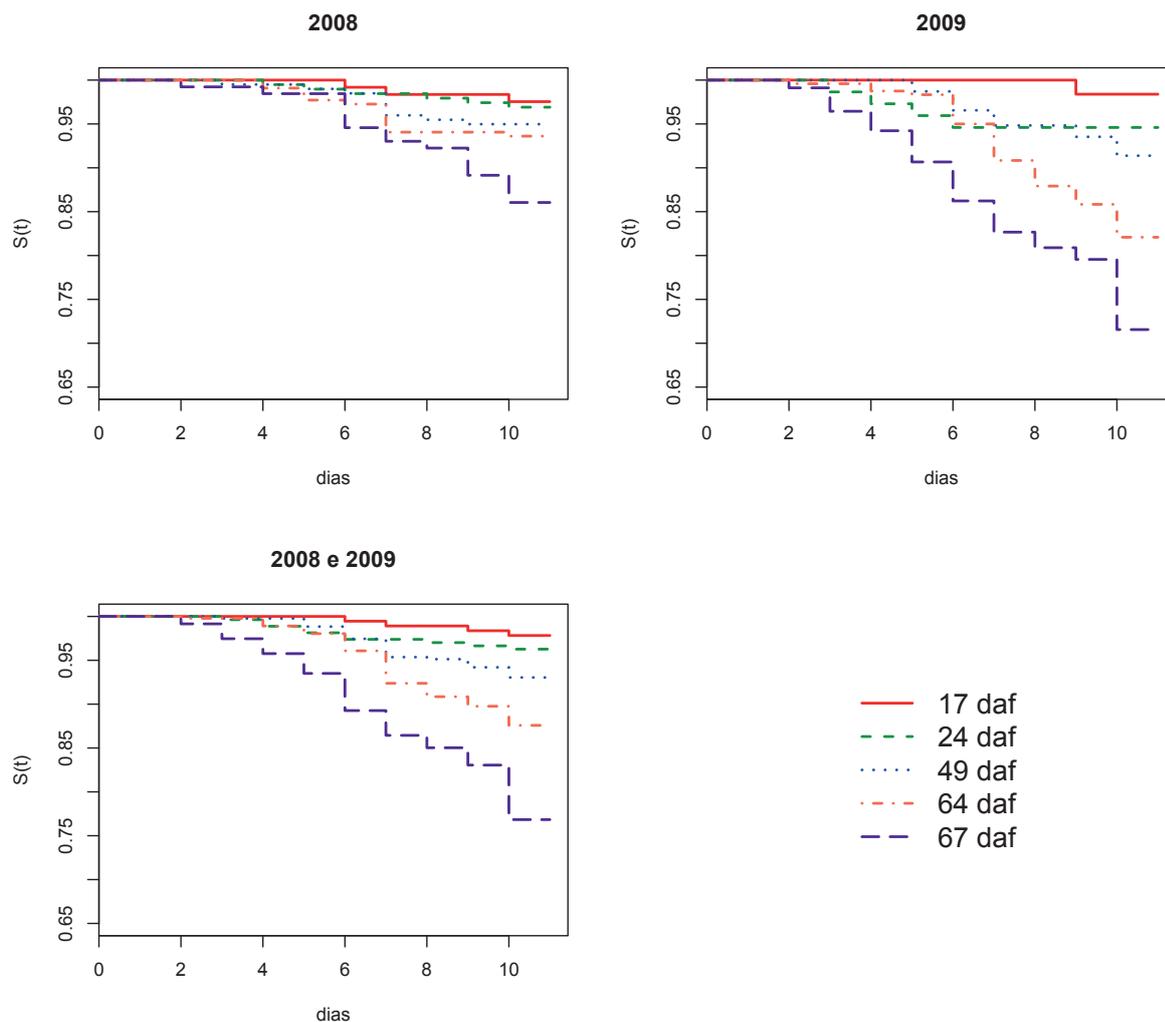


Figura 1: Estimativas das curvas de Kaplan-Meier para as funções de sobrevivência descrevendo o tempo para expressar os sintomas de *M. fructicola* em frutos de pêssgo inoculados e protegidos em diferentes momentos após a floração (safras 2008, 2009 e conjuntamente)

Em todas as curvas o pressuposto de proporcionalidade de riscos ao longo do tempo parece razoável, com a probabilidade de se observar fruto sadio diminuindo em função do tempo em todas as classes de diâmetros de ensacamento (Figura 1). Os tratamentos em que os frutos foram inoculados e ensacados mais tarde após a plena floração destacam-se pela menor probabilidade de se observar fruto sadio nos dois anos e com os anos agrupados.

2.2.2 Teste *log-rank*

Para comparar as curvas formalmente, utilizou-se o teste *log-rank*, que compara os valores observados e esperados de cada tratamento sob a hipótese de que o risco é o mesmo em todos os tratamentos. Isto equivale a testar se o tempo para incidência de podridão é semelhante em cada tratamento.

Os valores da estatística *log-rank* foram significativos ($p < 0,05$) em todas as safras e conjuntamente (Tabela 2), indicando diferenças na função de sobrevivência entre tratamentos. Embora a técnica não paramétrica de Kaplan-Meier e o teste *log-rank* não façam quase nenhuma restrição sobre a distribuição de probabilidades do tempo até a expressão dos sintomas), os métodos são limitados pois não permitem testar o efeito de diferentes covariáveis ao mesmo tempo (COLOSIMO & GIOLO, 2006; CARVALHO et al., 2011), o que neste estudo compõe o efeito dos tratamentos.

```
> # 'survdifff' = comando para calcular o teste log-rank
> # Ano = 1: safra 2008
> survdiff(Surv(tempo, tipo) ~ Tratamento, rho=0, data=subset(dados, Ano=="1"))

Call:
survdifff(formula = Surv(tempo, tipo) ~ Tratamento, data = subset(dados,
  Ano == "1"), rho = 0)

      N Observed Expected (O-E)^2/E (O-E)^2/V
Tratamento=17 122      3   7.34   2.563   3.021
Tratamento=24 194      6  11.59   2.698   3.525
Tratamento=49 199     10  11.79   0.273   0.358
Tratamento=64 219     14  12.85   0.103   0.140
Tratamento=67 129     18   7.43  15.041  17.769

Chisq= 20.9 on 4 degrees of freedom, p= 0.000336

> # Ano = 2: safra 2009
> survdiff(Surv(tempo, tipo) ~ Tratamento, rho=0, data=subset(dados, Ano=="2"))

Call:
survdifff(formula = Surv(tempo, tipo) ~ Tratamento, data = subset(dados,
  Ano == "2"), rho = 0)

      N Observed Expected (O-E)^2/E (O-E)^2/V
Tratamento=17  62      1   10.5   8.605   9.592
Tratamento=24  74      4   12.1   5.382   6.076
Tratamento=49 232     20  38.2   8.637  12.461
Tratamento=64 240     43  38.1   0.642   0.925
Tratamento=67 225     64  33.2  28.506  39.071

Chisq= 53.1 on 4 degrees of freedom, p= 8.05e-11

> # Soma do número de frutos nas duas safras
> survdiff(Surv(tempo, tipo) ~ Tratamento, rho=0, data=dados)
```

```
Call:
survdif(formula = Surv(tempo, tipo) ~ Tratamento, data = dados,
rho = 0)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
Tratamento=17	184	4	20.7	13.46	15.43
Tratamento=24	268	10	29.7	13.03	15.81
Tratamento=49	431	30	47.4	6.37	8.74
Tratamento=64	459	57	49.3	1.19	1.66
Tratamento=67	354	82	36.0	58.93	74.59

Chisq= 94.6 on 4 degrees of freedom, p= 0

Os resultados da aplicação da função `survdif` são apresentados na Tabela 2 apenas para as safras consideradas conjuntamente (soma dos frutos). Nos argumentos da função `survdif`, a expressão `rho=0` indica a realização do teste *log-rank*.

Tabela 2: Número de frutos total (N), números observados e esperados de frutos com sintomas de *Monilinia fructicola* em cada momento após o florescimento (Chisq= 94, 6 gl, p= 0)

Tratamento	N	Observado	Esperado	(O-E) ² /E	(O-E) ² /V
17 daf	184	4	20,7	13,5	15,4
24 daf	268	10	29,7	13,0	15,8
49 daf	431	30	47,4	6,4	8,7
64 daf	459	57	49,3	1,2	1,7
67 daf	354	82	36,0	58,9	74,6

Nesta situação, como o teste de qui-quadrado foi significativo ($p = 0$), rejeita-se a hipótese da nulidade, concluindo que pelo menos um tratamento difere significativamente dos outros em algum momento do tempo. Assim, há efeito significativo do momento em que os frutos são inoculados na expressão de sintomas de podridão parda.

2.3 Modelo semiparamétrico

Como o comportamento entre os tratamentos foi semelhante nos dois anos, os dados foram agrupados e as análises semiparamétricas realizadas com os anos agrupados. Considerou-se "ano" como fator de estratificação, o que equivale afirmar que o risco de base não é o mesmo para os dois anos, o que se justifica pelo aumento do inóculo inicial da doença no segundo ano.

Utilizou-se o modelo de riscos proporcionais de Cox para avaliar o efeito do momento da inoculação e proteção (daf). As árvores utilizadas no estudo apresentam variabilidade entre si e, pela falta de covariáveis que a expliquem, será considerado que as árvores apresentam diferentes fragilidades. Desta forma, será introduzida uma variável que representa um efeito aleatório para cada árvore, o que torna as estimativas dos efeitos de tratamentos mais consistentes.

● Modelo 1

Inicialmente ajustou-se um modelo que considera apenas o efeito do momento da inoculação e proteção:

```
> # o comando coxph() ajusta os parâmetros do modelo de regressão de Cox
> m1 <- coxph(Surv(tempo, tipo) ~ Tratamento, data=dados)
> summary(m1)
```

Call:

```
coxph(formula = Surv(tempo, tipo) ~ Tratamento, data = dados)
```

n= 1696, number of events= 183

	coef	exp(coef)	se(coef)	z	Pr(> z)
Tratamento24	0.5564	1.7444	0.5916	0.940	0.346976
Tratamento49	1.1906	3.2889	0.5323	2.237	0.025309 *
Tratamento64	1.7966	6.0290	0.5173	3.473	0.000514 ***
Tratamento67	2.4852	12.0041	0.5121	4.853	1.21e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
Tratamento24	1.744	0.57327	0.5471	5.562
Tratamento49	3.289	0.30405	1.1587	9.336
Tratamento64	6.029	0.16587	2.1875	16.616
Tratamento67	12.004	0.08331	4.4000	32.749

Concordance= 0.694 (se = 0.021)

Rsquare= 0.052 (max possible= 0.797)

Likelihood ratio test= 90.95 on 4 df, p=0

Wald test = 76.26 on 4 df, p=1.11e-15

Score (logrank) test = 95 on 4 df, p=0

- Modelo 2

Este modelo considera o efeito do momento da inoculação e proteção e a estratificação por ano:

```
> # o comando coxph() ajusta os parâmetros do modelo de regressão de Cox
> m2 <- coxph(Surv(tempo, tipo) ~ Tratamento + strata(Ano), data=dados)
> summary(m2)
```

Call:

```
coxph(formula = Surv(tempo, tipo) ~ Tratamento + strata(Ano),
      data = dados)
```

n= 1696, number of events= 183

	coef	exp(coef)	se(coef)	z	Pr(> z)
Tratamento24	0.6173	1.8539	0.5918	1.043	0.29689
Tratamento49	1.0323	2.8076	0.5332	1.936	0.05286 .
Tratamento64	1.6572	5.2446	0.5180	3.199	0.00138 **
Tratamento67	2.2720	9.6985	0.5137	4.423	9.75e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower .95	upper .95
Tratamento24	1.854	0.5394	0.5813	5.913
Tratamento49	2.808	0.3562	0.9873	7.984
Tratamento64	5.245	0.1907	1.9001	14.476
Tratamento67	9.699	0.1031	3.5435	26.544

Concordance= 0.672 (se = 0.029)

Rsquare= 0.041 (max possible= 0.763)

Likelihood ratio test= 71.61 on 4 df, p=1.033e-14

Wald test = 62 on 4 df, p=1.1e-12

Score (logrank) test = 72.96 on 4 df, p=5.329e-15

- Modelo 3

O modelo agora acomoda o efeito do momento da inoculação e da proteção, estratificado por ano, e considera cada árvore como efeito de fragilidade. Como *default* da função, assumiu-se a distribuição Gama para a fragilidade e estima-se a variância da fragilidade utilizando-se a verossimilhança perfilhada (EM):

```
> m3 <- coxph(Surv(tempo, tipo) ~ Tratamento + strata(Ano) + frailty(Arvore), data=dados)
> summary(m3)
```

Call:

```
coxph(formula = Surv(tempo, tipo) ~ Tratamento + strata(Ano) +
      frailty(Arvore), data = dados)
```

n= 1696, number of events= 183

	coef	se(coef)	se2	Chisq	DF	p
Tratamento24	0.619	0.5922	0.5919	1.09	1.00	3.0e-01

```

Tratamento49  1.018 0.5339  0.5335  3.63 1.00 5.7e-02
Tratamento64  1.649 0.5186  0.5182 10.11 1.00 1.5e-03
Tratamento67  2.262 0.5145  0.5140 19.32 1.00 1.1e-05
frailty(Arvore)                13.10 8.69 1.4e-01

```

```

          exp(coef) exp(-coef) lower .95 upper .95
Tratamento24    1.857    0.5385    0.5818    5.929
Tratamento49    2.766    0.3615    0.9716    7.877
Tratamento64    5.202    0.1922    1.8827   14.374
Tratamento67    9.598    0.1042    3.5011   26.310

```

```

Iterations: 8 outer, 29 Newton-Raphson
Variance of random effect= 0.07800985  I-likelihood = -1182.8
Degrees of freedom for terms= 4.0 8.7
Concordance= 0.703 (se = 0.03 )
Likelihood ratio test= 93.99 on 12.68 df,  p=1.71e-14

```

Para a escolha entre os três modelos utilizou-se a probabilidade de concordância (concordande). Concordância significa que, ao selecionar aleatoriamente duas observações, a que possuir menor tempo de sobrevivência é também aquela que possui maior risco estimado pelo modelo de Cox. Assim, o modelo 3 com concordância 0,703 (Tabela 3) tem um resultado discriminatório muito bom, de acordo com a classificação proposta por Carvalho et al. (2011). A inclusão do efeito de árvore, leva a melhoria significativa no modelo.

- Modelo 4

A partir do modelo 3, em que se assumiu distribuição Gama, agora estimou-se a variância da fragilidade utilizando-se a minimização do critério de informação de Akaike (AIC) como algoritmo de maximização:

```

> m4 <- coxph(Surv(tempo, tipo) ~ Tratamento + strata(Ano) + frailty(Arvore, method="aic"), data=dados)
> summary(m4)

```

Call:

```

coxph(formula = Surv(tempo, tipo) ~ Tratamento + strata(Ano) +
      frailty(Arvore, method = "aic"), data = dados)

```

n= 1696, number of events= 183

```

          coef  se(coef) se2  Chisq DF  p
Tratamento24  0.6191 0.5923  0.5920  1.09 1.00 3.0e-01
Tratamento49  1.0165 0.5339  0.5335  3.62 1.00 5.7e-02
Tratamento64  1.6485 0.5186  0.5183 10.10 1.00 1.5e-03
Tratamento67  2.2607 0.5146  0.5141 19.30 1.00 1.1e-05
frailty(Arvore, method =                13.94 9.28 1.4e-01

```

```

          exp(coef) exp(-coef) lower .95 upper .95
Tratamento24    1.857    0.5384    0.5818    5.929
Tratamento49    2.764    0.3619    0.9705    7.869

```

```

Tratamento64      5.199      0.1923      1.8814      14.367
Tratamento67      9.590      0.1043      3.4980      26.292

Iterations: 9 outer, 29 Newton-Raphson
Variance of random effect= 0.08667691 I-likelihood = -1182.9
Degrees of freedom for terms= 4.0 9.3
Concordance= 0.703 (se = 0.03 )
Likelihood ratio test= 95.15 on 13.27 df, p=1.898e-14

```

● Modelo 5

Como alternativa, assumiu-se fragilidade com distribuição lognormal (Gauss), em que o método padrão de estimação da variância é o da máxima verossimilhança restrita aproximada (REML):

```

> m5 <- coxph(Surv(tempo, tipo) ~ Tratamento + strata(Ano) + frailty(Arvore, dist="gauss"), data=dados)
> summary(m5)

```

Call:

```

coxph(formula = Surv(tempo, tipo) ~ Tratamento + strata(Ano) +
      frailty(Arvore, dist = "gauss"), data = dados)

```

n= 1696, number of events= 183

	coef	se(coef)	se2	Chisq	DF	p
Tratamento24	0.6189	0.5922	0.5919	1.09	1.00	3.0e-01
Tratamento49	1.0172	0.5338	0.5335	3.63	1.00	5.7e-02
Tratamento64	1.6486	0.5185	0.5182	10.11	1.00	1.5e-03
Tratamento67	2.2606	0.5145	0.5140	19.31	1.00	1.1e-05
frailty(Arvore, dist = "g				12.78	8.33	1.4e-01

	exp(coef)	exp(-coef)	lower .95	upper .95
Tratamento24	1.857	0.5385	0.5817	5.928
Tratamento49	2.766	0.3616	0.9714	7.874
Tratamento64	5.200	0.1923	1.8819	14.366
Tratamento67	9.589	0.1043	3.4984	26.284

```

Iterations: 8 outer, 24 Newton-Raphson
Variance of random effect= 0.0738499
Degrees of freedom for terms= 4.0 8.3
Concordance= 0.703 (se = 0.03 )
Likelihood ratio test= 92.97 on 12.32 df, p=1.843e-14

```

● Modelo 6

A partir do modelo 5, em que se assumiu distribuição lognormal, agora estimou-se a variância da fragilidade utilizando-se a minimização do critério de informação de Akaike (AIC) como algoritmo de maximização:

```

> m6 <- coxph(Surv(tempo, tipo) ~ Tratamento + strata(Ano) + frailty(Arvore, dist="gauss", method="aic"), data=dados)
> summary(m6)

```

```

Call:
coxph(formula = Surv(tempo, tipo) ~ Tratamento + strata(Ano) +
      frailty(Arvore, dist = "gauss", method = "aic"), data = dados)

n= 1696, number of events= 183

              coef se(coef) se2   Chisq DF  p
Tratamento24  0.619 0.5922  0.5920  1.09 1.00 3.0e-01
Tratamento49  1.016 0.5339  0.5335  3.62 1.00 5.7e-02
Tratamento64  1.648 0.5186  0.5182 10.10 1.00 1.5e-03
Tratamento67  2.259 0.5145  0.5140 19.28 1.00 1.1e-05
frailty(Arvore, dist = "g
              13.93 9.09 1.3e-01

              exp(coef) exp(-coef) lower .95 upper .95
Tratamento24  1.857  0.5385  0.5817  5.928
Tratamento49  2.761  0.3621  0.9698  7.863
Tratamento64  5.195  0.1925  1.8801 14.355
Tratamento67  9.578  0.1044  3.4938 26.257

Iterations: 9 outer, 28 Newton-Raphson
Variance of random effect= 0.08503443
Degrees of freedom for terms= 4.0 9.1
Concordance= 0.703 (se = 0.03 )
Likelihood ratio test= 94.48 on 13.08 df,  p=2.098e-14

```

As medidas de ajuste entre os modelos 3, 4, 5 e 6 diferem pouco entre as distribuições e métodos de estimação da variância (Tabela 3), o que também se espera dos parâmetros estimados em cada modelo. A hipótese nula de que os efeitos aleatórios são iguais a zero foi rejeitada nas quatro configurações de distribuição para fragilidade e método de estimação.

Tabela 3: Medidas de ajuste dos modelos com diferentes distribuições para fragilidade e métodos para estimar a variância. O teste de Wald para a fragilidade e o teste da razão de verossimilhanças são seguidos pelos graus de liberdade entre parênteses e são significativos a 1% ($p < 0,01$)

Medidas	Cox		Gama		Lognormal	
	Clássico	EM	AIC	REML	AIC	
	(m1)	(m3)	(m4)	(m5)	(m6)	
WT	76,26 (4)	-	-	-	-	
VEA	-	0,0780	0,0867	0,0738	0,0850	
I-l	-	-1182,8	-1182,9	-	-	
RV	90,95 (4)	93,99 (12,68)	95,15 (13,27)	92,97 (12,32)	94,98 (13,08)	
Conc.	0,694 (4)	0,703	0,703	0,703	0,703	

WT = Wald test; VEA = Variância do Efeito Aleatório; I-l = I-lilelikelihood; RV = Razão de Verossimilhança; Conc.= Concordância

A Figura 2 apresenta de forma gráfica a distribuição do risco para cada árvore. As árvores foram ordenadas pela estimativa pontual da fragilidade, não havendo alteração na ordem das fragilidades entre os modelos. Os comandos em R para construção da Figura 2 são apresentados a seguir:

```
> par(mfrow = c(2, 2))
> source("Rfun.r")
> plot.frail(dados$Arvore, m3, xlab="Árvore", main="(a)")
> plot.frail(dados$Arvore, m4, xlab="Árvore", main="(b)")
> plot.frail(dados$Arvore, m5, xlab="Árvore", main="(c)")
> plot.frail(dados$Arvore, m6, xlab="Árvore", main="(d)")
```

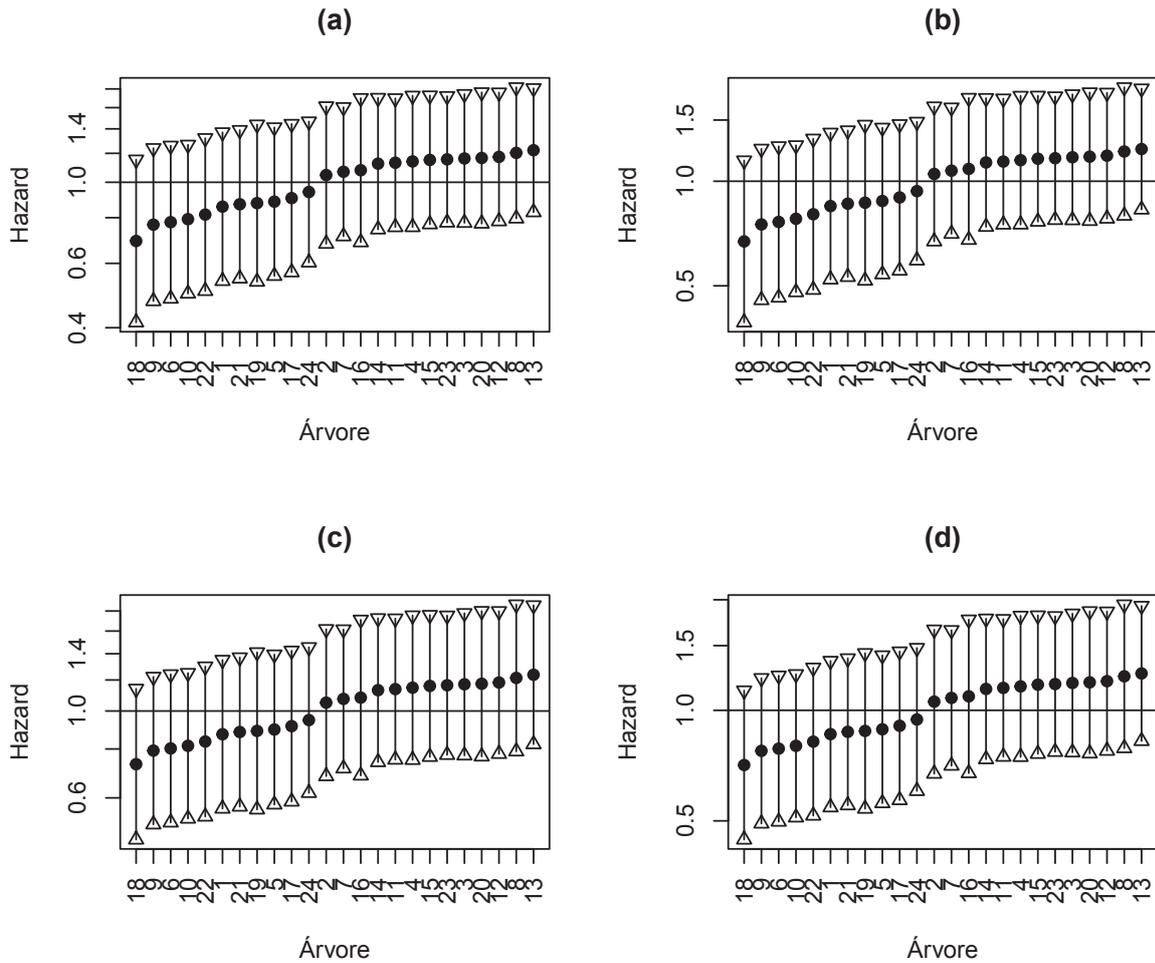


Figura 2: Estimativas pontuais de fragilidade e respectivos intervalos com 95% de confiança, utilizando diferentes distribuições para o efeito de fragilidade (Gama e Lognormal) e algoritmos de estimação (EM, AIC e REML): (a) Gama-EM; (b) Gama-AIC; (c) Gauss-REML; (d) Gauss-AIC

A variação encontrada na estimativa da variância dos efeitos aleatórios entre os modelos é pequena e pode ser desconsiderada. Além disso, as árvores podem ser consideradas equivalentes em relação ao risco de expressão de sintomas dos seus frutos. Em função disso, os modelos podem ser considerados equivalentes na estimativa do efeito da proteção no controle da incidência de *M. fructicola*.

2.4 Verificação dos resíduos do modelo

Na Tabela 4 apresenta-se o teste para a correlação linear entre o tempo de sobrevivência e o resíduo. Observa-se que não se rejeita a hipótese nula de risco proporcional entre os tratamentos. Assim, não há tendência linear decrescente ao longo do tempo.

```
> # 'cox.zph' retorna os resíduos de Schoenfeld padronizados
> #  $rh_0$  é a hipótese nula de risco proporcional
> (residuo.sch <- cox.zph(m4))

      rho chisq    p
Tratamento24 -0.1060 2.066 0.151
Tratamento49 -0.0512 0.476 0.490
Tratamento64 -0.0401 0.291 0.590
Tratamento67 -0.0530 0.507 0.477
GLOBAL          NA 3.272 0.513
```

Tabela 4: Resultados do teste da correlação linear dos resíduos dos tratamentos com o tempo de sobrevivência; rh_0 é o coeficiente de correlação; $chisq$, o valor da estatística qui-quadrado; e p corresponde ao p-valor da estatística qui-quadrado

Tratamento	rh_0	chisq	p-valor
24 daf	-0,1060	2,066	0,1151
49 daf	-0,0512	0,476	0,490
64 daf	-0,0401	0,291	0,590
67 daf	-0,0530	0,507	0,477
Global	-	3,272	0,513

Observa-se que em todos os tratamentos não há tendência sistemática no gráfico de resíduos (Figura 3), confirmando que o risco relativo dos tratamentos é constante durante o tempo de observação.

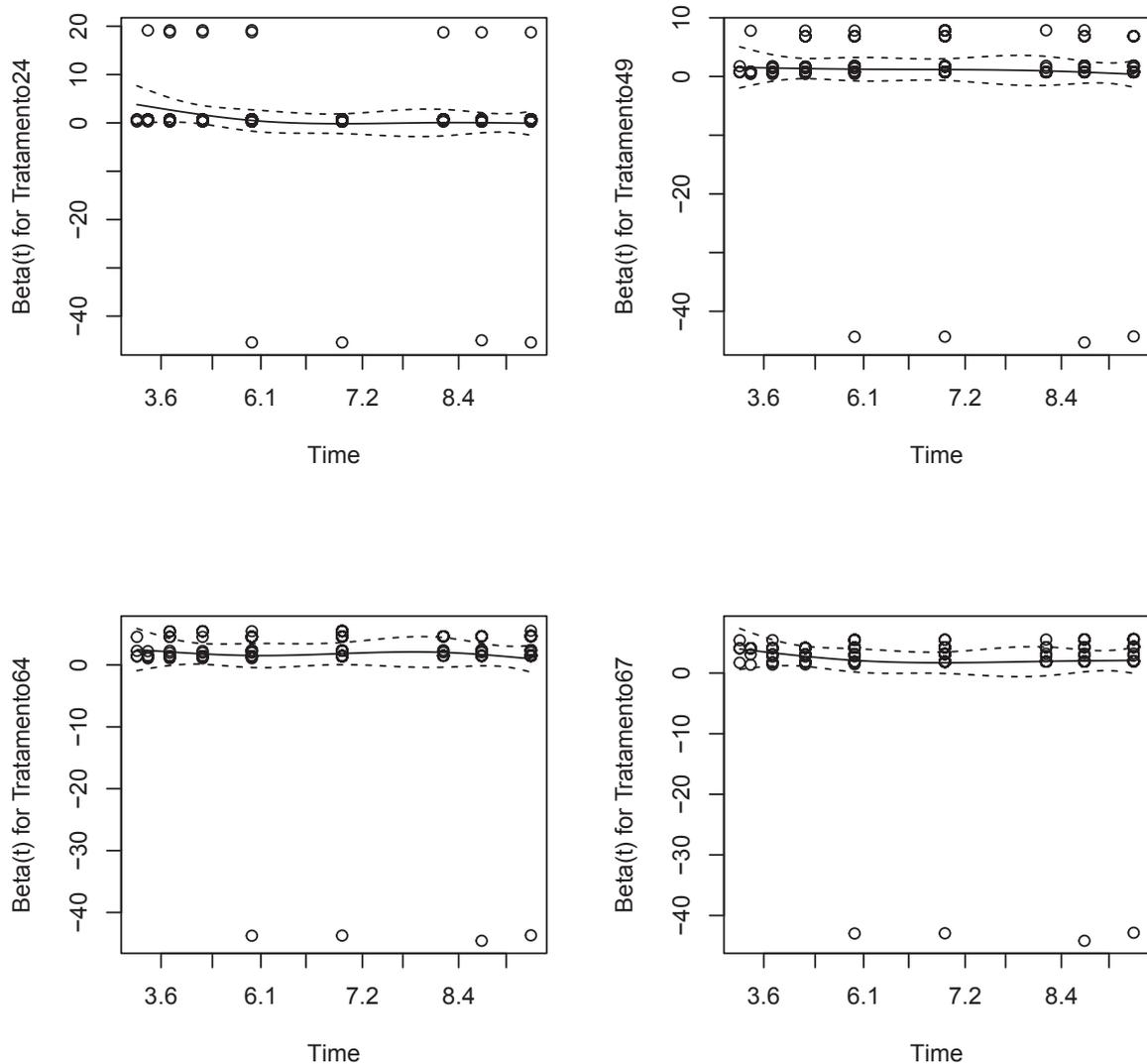


Figura 3: Resíduos de Schoenfeld para o modelo de riscos proporcionais de Cox ajustado aos dados de incidência de *M. fructicola* em pêsego em função do momento em que os frutos foram inoculados e protegidos (tratamentos).

Como as pressuposições foram atendidas, as inferências a partir do modelo semi-paramétrico de Cox são válidas e apresentadas na Tabela 4.

Tabela 5: Estimativas do risco relativo de expressão de sintomas para cada tratamento em cada distribuição para a fragilidade e método para estimar a variância (* indica efeito significativo e *ns* efeito não significativo)

Tratamento	Gama				lognormal			
	(EM)	IC95%(LI; LS)	(AIC)		(REML)	(AIC)		
24 daf	1,857	ns (0,582; 5,929)	1,857	ns	1,857	ns	1,857	ns
49 daf	2,766	ns (0,972; 7,877)	2,764	ns	2,766	ns	2,761	ns
64 daf	5,202	* (1,883; 14,374)	5,199	*	5,200	*	5,195	*
67 daf	9,598	* (3,501; 26,310)	9,590	*	9,589	*	9,578	*

As alterações nos efeitos fixos entre os modelos com diferentes distribuições e métodos para estimar as fragilidades não são importantes. Além disso, observa-se que as diferentes possibilidades são consistentes, apontando como significativas as diferenças entre todos os tratamentos em relação ao momento 17 daf.

Frutos inoculados e ensacados após 49 daf aumentam significativamente o risco de apresentar sintomas, variando de 5,2 aos 64 daf até 9,6 vezes a chance de ocorrência aos 67 daf.

3 Considerações finais

A apresentação das técnicas mais usuais em análise de sobrevivência aliada à aplicação neste problema real permitiu ilustrar uma primeira abordagem para dados de tempo-ocorrência. No exemplo apresentado, a análise conjunta com efeito basal variando entre safras, a verificação de um fator de agrupamento (árvores) como efeito aleatório além do pressuposto de riscos proporcionais para o modelo de Cox, são procedimentos padrões em problemas convencionais que permitem análise respeitando a natureza da variável, além de qualificar as análises em problemas convencionais.

As técnicas apresentadas têm extensões ou diferentes abordagens para problemas mais complexos. A ocorrência do evento antes da observação no indivíduo (dados com censura à esquerda) ou a ocorrência em um intervalo de tempo (censura intervalar) necessitam de abordagens diferentes das apresentadas. Além disso, quando o pressuposto de riscos proporcionais no modelo de Cox não é atendido, pode-se utilizar modelos cujas covariáveis são tempo-dependentes, denominados de modelos de tempo de vida acelerado. Em casos de covariáveis contínuas a relação entre covariável e os riscos associados podem ser estimados respeitando a forma funcional da covariável com funções não paramétricas, como regressão linear local (*lowess*) e as funções *spline*.

Referências

- AKBAR, A.; PASHA, G.R.; NAQVI, S.F.H. Properties of Kaplan-Meier estimator: group comparison of survival curves. **European Journal of Scientific Research**, v.32, n.3, p.391-397, 2009.
- BEWICK, V.; CHEEK, L.; BALL, J. Statistics review 12: Survival analysis. **Critical Care**, v.8, n.5, p.389-394, 2004.
- CARVALHO, M.S.; ANDREOZZI, V.L.; CODEÇO, C.T.; CAMPOS, D.P.; BARBOSA, M.T.S.; SHIMAKURA, S.E. **Análise de Sobrevivência: teoria e aplicações em saúde**, 2.ed. Rio de Janeiro: Fiocruz, 2011, 432p.
- COLOSIMO, E.A.; GIOLO, S.R. **Análise de sobrevivência aplicada**. São Paulo: Editora Edgard Blücher, 2006.
- COPES, W.E.; THOMSON, J.L. Survival analysis to determine the length of the incubation period of *Camellia twig blight* caused by *Colletotrichum gloeosporioides*. **Plant Disease**, n.92, p.1177-1182, 2008.
- COUTO, M.R.M.; JACOBI, L.F., DAL'COL LÚCIO, A.; LOPES, S.J.; MEDEIROS, S.L.P. Análise de sobrevida da área foliar de meloeiros em sistema hidropônico. **Ciência e Natura**, v.31, n.1, p.7-16, 2009.
- COX, D.R. Regression models and life tables (with discussion). **Journal Royal Statistical Society**, v.34, p.187-220, 1972.
- DALLOT, S.; GOTTWALD, T.; LABONNE, G.; QUIOT, J.B. Factors affecting the spread of *Plum pox virus* strain M in peach orchards subjected to roguing in France. **Phytopathology**, v.94, p.1390-1398, 2004.
- GOEL, M.K.; KHANNA, P.; KISHORE, J. Understanding survival analysis: Kaplan-Meier estimate. **International Journal of Ayurveda Research**, v.1, n.4, p.274-278, 2010.
- GORFINE, M.; ZUCKER, D.M.; HSU, L. Prospective survival analysis with a general semiparametric shared frailty model: a pseudo full likelihood approach. **Biometrika**, v.93, n.3, p.735-741, 2006.
- KAPLAN, E.L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v.53, n.282, p.457-481, 1958.
- KOSOROK, M.R.; LEE, B.L.; PINE, J.P. Robust inference for univariate proportional hazards frailty regression models. **The Annals of Statistics**, v.32, n.4, p.1448-1491, 2004.

KRÜGER, R.F.; KROLOW, T.K.; AZEVEDO, R.R.; DUARTE, J.L.P.; RIBEIRO, P.B. Sobrevivência e reprodução de *Synthesiomia nudiseta* (Diptera, Muscidae). **Iheringia**, Série Zoologia, v.98, n.1, p.45-49, 2008.

LIMA JUNIOR, P.; SILVEIRA, F.L.; OSTERMANN, F. Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira. **Revista Brasileira de Ensino de Física**, v.34, n.1, p.1403.1-1403.10, 2012.

McGILCHRIST, C.A.; AISBETT, C.W. Regression with frailty in survival analysis. **Biometrics**, v.47, p.461-466, 1991.

NESI, C.N.; SHIMAKURA, S.E., RIBEIRO JUNIOR, P.J., MAY DE MIO, L.L. Survival analysis: a tool in the study of post-harvest diseases in peaches. **Revista Ceres**, v.62, n.1, p.52-61, 2015.

NESI, C.N.; ARRUDA, G.O.S.F.; MENEGATTI, A. Superação de dormência em sementes de Jatobá avaliadas por análise de sobrevivência. **Revista de Ciências Agroveterinárias**, v.15, n.1, p.42-49, 2016.

NESI, C.N.; SHIMAKURA, S.E.; RIBEIRO JUNIOR, P.J.; MAY DE MIO, L.L. Survival analysis in plant pathology. **Idesia** [online]. v.31, p.107-110, 2013.

OJIAMBO, P.S.; SCHERM, H. Survival analysis of time to abscission of blueberry leaves affected by Septoria leaf spot. **Phytopathology**, v.95, p.108-113, 2005.

OLIVEIRA, E.L.; RIOS-NETO, E.G.; OLIVEIRA, A.M.H.C. Transições dos jovens para o mercado de trabalho, primeiro filho e saída da escola: o caso brasileiro. **Revista Brasileira de Estudo de População**, v.23, n.1, p.109-127, 2006.

OLIVEIRA, A.M.H.C.; RIOS-NETO, E.G. Uma avaliação experimental dos impactos da política de qualificação profissional no Brasil. **Revista Brasileira de Economia**, v.61, n.3, p.353-378, 2007.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**, R Foundation for Statistical Computing, Vienna, Austria, 2015.

SCHERM, H.; OJIAMBO, P.S. Applications of survival analysis in botanical epidemiology. **Phytopathology** v.94, p.1022-1026, 2004.

SETTI, B.; BENECHAIKH, M.; HENNI, J.E.; CLAIRE, N. Survival analysis to determine the length of latent period of *Mycosphaerella pinodes* on peas (*Pisum sativum* L.). **African Journal of Microbiology Research**, v.4, n.18, p.1897-1903, 2010.

THERNEAU, T. **A Package for Survival Analysis in S. version 2.38**. 2015. Disponível em <<http://CRAN.R-project.org/package=survival>>.