UNIVERSIDADE FEDERAL DO PARANÁ

DANIEL ZAGROBA BOJANOWSKI
GRACIANO ALCIDES LOLATTO

APLICAÇÃO DE REGRESSÃO LOGÍSTICA E MODELOS COM FRAÇÃO DE CURA EM UM ESTUDO SOBRE CLIENTES INADIMPLENTES DE UMA INSTITUIÇÃO FINANCEIRA

CURITIBA-PR
2018

DANIEL ZAGROBA BOJANOWSKI
GRACIANO ALCIDES LOLATTO

APLICAÇÃO DE REGRESSÃO LOGÍSTICA E MODELOS COM FRAÇÃO DE CURA EM UM ESTUDO SOBRE CLIENTES INADIMPLENTES DE UMA INSTITUIÇÃO FINANCEIRA

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Estatística do Setor de Ciências Exatas da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientadora: Profa. Dra. Suely Ruiz Giolo

CURITIBA-PR
2018

AGRADECIMENTOS

Gostaríamos, primeiramente, de agradecer a Deus pela vida, e por nela nos fornecer força para alcançarmos mais um objetivo. Que o Senhor continue iluminando nossas vidas.

À nossas famílias, pela paciência e apoio em muitos momentos, sempre que mais precisamos ao longo dessa jornada. Pela compreensão de muitas vezes não estar presente da forma que se era necessário. Seremos eternamente gratos e dividimos essa nossa conquista com vocês.

À nossa orientadora professora Dra. Suely Ruiz Giolo pela forma de ensinamento que nos estimulou com a ideia deste trabalho de conclusão de curso, bem como ao professor Dr. José Luiz Padilha da Silva por dedicar parte de seu tempo à avaliação deste conteúdo.

Gostaríamos também de estender nossos agradecimentos aos demais professores do Departamento de Estatística da UFPR. Mestres dedicados e compreensivos que nos passaram ensinamentos que levaremos em nossas vidas.

Agradecemos também a instituição financeira que gentilmente nos forneceu os dados para estudos, em especial aos colaboradores da área de Modelos de Cobranças.

E também de forma geral, sem citar nomes, aos nossos e nossas colegas que ingressaram no curso em 2013 ou que foram incorporados no decorrer do curso. Foram inúmeras as vezes que nos reunimos em horários mais diversos possíveis para estudarmos. Temos a certeza que cada um colaborou da forma que melhor podia. Nosso muito obrigado a todos e todas vocês.

“É preciso impor a si mesmo
algumas metas para se ter a coragem de alcançá-las”.

(BENITO MUSSOLINI)

RESUMO

Dado o cenário de instabilidade que a economia brasileira tem demostrado na década atual, carregando consigo o crescimento da inadimplência, neste presente estudo propomos estudar e estimar, com o auxílio de diversos fatores (797 covariáveis), o tempo que empresas (pessoas jurídicas), com atrasos entre 61 a 3600 dias, levam até sanarem suas dívidas. Foram utilizados dados reais (642.707 informações) fornecidos por uma grande instituição financeira. As técnicas de modelagem por regressão logística e análise de sobrevivência foram utilizadas. A primeira serviu como parâmetro de comparação, já que é atualmente utilizada pela instituição. Já a análise de sobrevivência, foi abordada considerando dois modelos: o modelo de mistura e o modelo tempo de promoção, ambos com fração de cura. Todos os modelos considerados mostraram-se bem ajustados aos dados e forneceram resultados satisfatórios. Entretanto, os modelos no contexto de análise de sobrevivência apresentaram a vantagem de fornecer mais informação do que o modelo de regressão logística. A variável “tempo” está presente nesses modelos, o que proporciona uma tomada de decisão diferenciada. Ou seja, dentro da gama de clientes “bons”, é possível identificar quais clientes se sobressaem por comparação dos tempos até o pagamento de suas dívidas. Dentre os modelos de sobrevivência considerados, o que se ajustou melhor aos dados foi o modelo de mistura logito-Cox. Com os resultados obtidos, a instituição poderá identificar características de inadimplentes quanto ao risco, ou suscetibilidades ao evento de recuperação durante o processo de cobrança e, assim, confirmar quais as ações são mais eficientes, baseadas nos perfis de clientes dentro de suas carteiras.

Palavras-chave: Modelagem estatística. Inadimplência. Regressão logística. Análise de sobrevivência. Modelo de mistura. Modelo tempo de promoção.

LISTA DE GRÁFICOS E FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O momento econômico brasileiro vivido na década atual, em especial no biênio 2015-2016, impôs severas dificuldades para empresas e consumidores, afetando a capacidade das empresas de honrarem todos os seus compromissos. Atualmente, ainda há efeitos da crise, mas também há sinais de retomada da economia. Em um curto espaço de tempo, espera-se que, à medida que os negócios se recuperem, a capacidade de pagamento das empresas que têm essas dificuldades também melhore.

Conforme relatórios divulgados pela SPC Brasil e CNDL (Confederação Nacional de Dirigentes Lojistas) (2018), o número de empresas registradas nos cadastros de devedores avançou 6,2% na comparação entre janeiro de 2018 e o mesmo mês do ano anterior. Conforme Tabela 1, a região Sudeste se destacou ao registrar crescimento de 9,47%, bem acima da média nacional (6,2%). No relatório publicado, é citado que parte do avanço na região Sudeste "deveu-se ao fim da obrigatoriedade de envio de carta com Aviso de Recebimento no Estado de São Paulo no processo de negativação, tal como dispunha a Lei Estadual nº 15.659". Ainda consta no relatório, que do total de empresas que estavam negativadas no país em janeiro de 2018, apenas 3,8% conseguiram deixar a lista de inadimplentes mediante pagamento ao longo do mês. Mesmo com essas baixas, o total de empresas negativadas cresceu.

TABELA 1 – Resumo da evolução de pessoas jurídicas inadimplentes na base do SPC Brasil

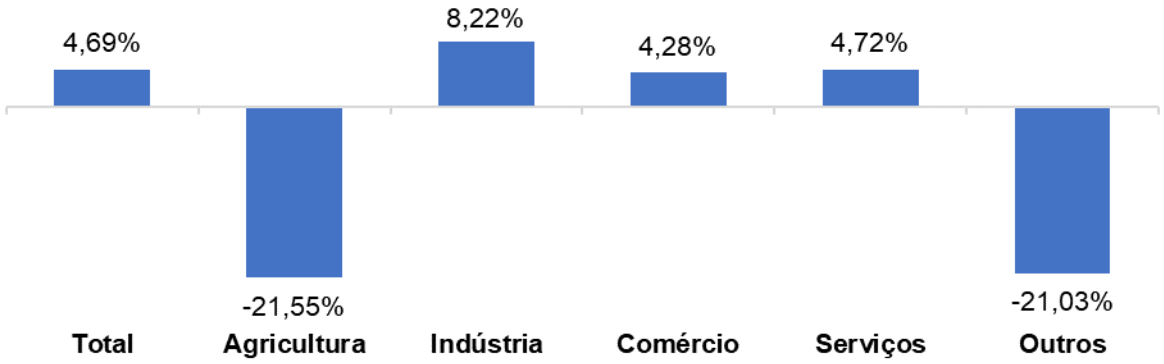
Região	Variação mensal (em relação ao mês anterior)		Variação anual (em relação ao mesmo mês do ano anterior)	
	Dez/16 a Jan/17	Dez/17 a Jan/18	Jan/16 a Jan/17	Jan/17 a Jan/18
Total Brasil	**0,65%**	**1,46%**	**5,28%**	**6,20%**
Centro-Oeste	0,37%	-0,46%	4,52%	2,13%
Nordeste	0,14%	-0,09%	6,84%	2,38%
Norte	-0,17%	-0,21%	5,48%	2,19%
Sudeste	1,02%	3,00%	5,38%	9,47%
Sul	0,66%	0,54%	3,27%	3,05%

FONTE: SPC Brasil (2018).

De acordo com a mesma publicação, representado no Gráfico 1, o número de pendências devidas por pessoas jurídicas apresentou crescimento de 4,69%. O destaque novamente se dá à região Sudeste, que apresentou variação anual de 8,04% (Tabela 2). Os dados setoriais mostram que a Indústria foi o setor credor a

registrar maior crescimento da inadimplência de empresas, com variação de 8,22%. Em seguida, aparecem o setor de Serviços (4,72%) e o do Comércio (4,28%).

GRÁFICO 1 – Setores das empresas com dívidas na base do SPC Brasil



FONTE: SPC Brasil (2018).

TABELA 2 – Resumo da evolução do número de dívidas de pessoas jurídicas inadimplentes na base do SPC Brasil

Região	Variação mensal (em relação ao mês anterior)		Variação anual (em relação ao mesmo mês do ano anterior)	
	Dez/16 a Jan/17	Dez/17 a Jan/18	Jan/16 a Jan/17	Jan/17 a Jan/18
Total Brasil	**0,57%**	**1,60%**	**3,45%**	**4,69%**
Centro-Oeste	0,51%	-0,55%	4,06%	0,59%
Nordeste	0,24%	-0,21%	6,69%	1,33%
Norte	-0,01%	-0,21%	5,67%	1,26%
Sudeste	0,76%	3,49%	1,75%	8,04%
Sul	0,72%	0,57%	2,71%	1,92%

FONTE: SPC Brasil (2018).

Nas demais regiões, como é possível observar na Tabela 2, a inadimplência entre empresas cresceu, porém bem menos do que no Sudeste. Ainda, segundo o relatório, essa “queda no ritmo do avanço da inadimplência reflete a redução do crédito ao longo da crise e o momento econômico mais favorável dos últimos trimestres”.

Assim sendo, em um cenário de melhores expectativas de juros e inflação, vigência de novas leis trabalhistas, entre outros fatores, talvez não seja exagerado dizer que o crédito às empresas poderá, possivelmente, retomar já no próximo ano. Isto influencia diretamente no pagamento de dívidas passadas, reduzindo assim o repasse (venda) de dívidas de clientes inadimplentes às empresas especializadas em cobrança. Estas empresas compradoras de créditos inadimplentes chegaram a pagar apenas 4% do valor da carteira, então é muito mais vantajoso ao banco recuperar este crédito.

1.1 JUSTIFICATIVA

Dada as constantes alterações nos cenários econômicos observados no Brasil nesta década, faz-se necessários estudos para que as instituições financeiras sejam capazes de especificar perfis de clientes endividados com potencial de recuperação em médio/longo prazo.

1.2 OBJETIVOS

1.2.1 Objetivo geral

O presente estudo teve como objeto geral estudar o tempo até a ocorrência do pagamento de dívidas de clientes pessoas jurídicas (PJ) que já se encontram em atraso, buscando a identificação de possíveis fatores (covariáveis) que afetam este tempo de pagamento.

1.2.2 Objetivos específicos

Dado o objetivo geral, o presente estudo teve como objetivos específicos:

a) Identificar os clientes com baixa propensão de pagamento, buscando antecipar ações capazes de reduzir a deterioração de portfólios, como possíveis venda de carteira, agregando valor ao mercado;
b) Auxiliar a instituição com indicadores capazes de identificar perfis de clientes mais propensos a pagamento em até 24 meses, possibilitando ações de cobranças diferenciadas;
c) Verificar entre os modelos considerados, qual o que melhor se ajustou ao banco de dados disponibilizado, bem como se os modelos no contexto de análise de sobrevivência apresentam ganho quando comparados ao modelo logístico, atualmente utilizado pela instituição financeira.

2 REVISÃO DE LITERATURA

2.1 ESTATÍSTICAS NO ÂMBITO DO CRÉDITO

Como comentado anteriormente, com as políticas de controle de inflação e uso da nova lei trabalhista, bem como o aumento no número de compras de bens de consumo, as empresas mostram sinais de melhoria, com isso aumentando a capacidade de pagamento de suas dívidas, tornando o ramo de recuperação de crédito atrativo aos interesses das instituições financeiras devido à rentabilidade esperada sobre o capital emprestado. Por outro lado, há também uma expansão do crédito, provocando maior exposição das instituições ao risco de inadimplência, ou seja, de não receberem - ou receberem de forma parcial - o capital previamente emprestado.

Nesse contexto, para garantir bons resultados financeiros, as empresas necessitam de métodos que auxiliem na gestão estratégica sobre os riscos envolvidos na contratação de crédito, desde a proposta de concessão até os processos de cobrança.

Segundo Thomas et al. (2002), até o início do século XX, todas as decisões relativas à concessão de crédito eram baseadas exclusivamente no julgamento subjetivo dos analistas. Somente a partir da publicação, em 1936, da técnica de Análise Linear Discriminante, desenvolvida por Fisher, é que a Estatística começou a ser pensada para identificar bons e maus pagadores. Assim, os primeiros modelos de *Credit Scoring* foram desenvolvidos por Durand (1941), com o objetivo de ordenar os proponentes quanto à probabilidade de pagar o capital emprestado. Diante da maior agilidade na decisão, menor custo, maior objetividade e até mesmo melhor poder preditivo, os modelos de *Credit Scoring* foram aos poucos se popularizando e atualmente são largamente utilizados (HAND; HENLEY, 1997).

Modelos de *Credit Scoring* utilizam-se de algoritmos matemáticos e técnicas estatísticas para calcular a probabilidade de que determinado evento aconteça. Aplicando fórmulas, o sistema atribui pontuação específica para cada característica do proponente/cliente para prever um resultado. Apenas as informações do *Credit Score* não garantem sucesso de um modelo de gestão de crédito na instituição financeira, devendo esta manter um acompanhamento continuo da posição dos clientes.

Este comportamento é chamado de *Behaviour Score* e baseia-se no conhecimento das operações dos clientes durante o relacionamento com a instituição. Estas informações podem ser: nível de utilização de crédito, hábitos de pagamentos, tempo de relacionamento, etc.

Estes dados estão sempre sendo gerados tornando esta modelagem bastante dinâmica e neste caso sendo constantemente revisado. Diferente do nosso caso em estudo, este modelo permite prever o risco de o cliente se tornar inadimplente em um horizonte específico.

Os dois modelos comentados classificam o risco da inadimplência; isto significa que se aplicam às populações que não são inadimplentes. Usualmente, é considerado inadimplente qualquer indivíduo que não conseguir honrar por total o pagamento de suas dívidas na data de vencimento, seja da parcela e/ou liquidação do contrato, independente da causa ou motivo.

Com vimos no capítulo anterior, este número está crescendo na população formada de pessoas jurídicas. Percebe-se que a inadimplência está bastante relacionada à economia do país; se a economia está em condições favoráveis podemos notar uma redução dos índices de inadimplência.

Na população de inadimplentes aplica-se o modelo de *Collection Score*. Assim, podemos classificar o risco em termos de pagamentos futuros das empresas que já se tornaram inadimplentes. Souza (2000) diz que: “é imprescindível que a empresa conheça aqueles clientes inadimplentes que têm alta probabilidade de não pagar o seu saldo devedor, para que seja possível estabelecer uma estratégia de atuação sobre eles”.

Nas instituições financeiras, esta é uma carteira estratégica do ponto de vista dos resultados financeiros. Seu gerenciamento é de grande importância e necessita de ferramentas que auxiliem na tomada de decisões a fim de aperfeiçoar o processo de cobrança maximizando os resultados das contas a receber.

Portanto, as empresas estão despendendo esforços e estudos para o desenvolvimento de novas técnicas que auxiliem os sistemas de *scoring*, sendo uma das mais recentes a Análise de Sobrevivência.

2.2 ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência é utilizada quando se deseja estimar a probabilidade de sobrevivência a um evento de interesse (denominado falha), associada a cada instante de tempo durante um período de observação (HANREJSZKOW; STROMBERG, 2013). Por exemplo, o tempo até a morte de um paciente ou o tempo até a recidiva de um tumor. Na literatura, esse tempo é geralmente denominado como tempo de vida. Com o desenvolvimento e aprimoramento de técnicas estatísticas, aliado ao avanço tecnológico, estudos que fazem uso de dados de sobrevivência têm sido mais frequentes (COLOSIMO; GIOLO, 2006). Devido ao crescimento ainda ser recente, começaram a aparecer trabalhos acadêmicos sob essa abordagem em instituições financeiras. Podemos citar os casos de Miola (2013) e Quidim (2005), em que o primeiro fez uso da metodologia de análise de sobrevivência com fração de cura para modelar os dados dos tempos de inadimplência, e o segundo fez uso da mesma metodologia para modelar o tempo até o cancelamento de cartões de crédito.

Mais recentemente, Tonegi (2017) também analisou as informações dos clientes Pessoa Física da instituição financeira que nos cedeu os dados dos clientes Pessoa Jurídica utilizados neste trabalho. A metodologia de análise utilizada pelo autor para modelar o tempo até a recuperação de clientes inadimplentes foi o modelo de mistura logito-Cox com fração de inadimplentes (apresentado na Seção 3.2.4).

Neste contexto, a análise de sobrevivência permite determinar quais variáveis afetam o risco de ocorrência de determinado fenômeno. A principal característica relacionada a dados de sobrevivência diz respeito à presença de censuras, que consistem na observação parcial da resposta e se dá, geralmente, pelo fato de alguns clientes abandonarem a carteira ou não experimentarem o evento de interesse em estudo. No nosso caso, o evento de interesse diz respeito ao cliente pagar a dívida, estudando assim o tempo até a ocorrência deste pagamento.

Este tempo de ocorrência pode ser definido como o tempo de falha, ou seja, quando o cliente pagou a dívida e se ausentou do estudo. Alguns elementos são necessários para a definição do tempo de falha, entre eles:

- Início do estudo: ser precisamente definido para que os indivíduos possam ser comparáveis no início do estudo, com exceção das diferenças medidas pelas covariáveis.

- Escala de medida: é quase sempre o tempo real (dias, meses, etc.).
- Evento de interesse: consiste, em nosso estudo, no pagamento da dívida.

Em relação à censura, pode ser definida como a observação parcial da resposta, geralmente devida ao abandono do cliente do estudo (exemplo: venda de carteira), antes que este experimente o evento de interesse.

Desta forma, introduz-se uma variável a mais na análise que indica se o cliente teve seu tempo até a ocorrência do evento de interesse observado ou não. Esta variável, conhecida como variável indicadora, é definida como:

$$\delta_i = \begin{cases} 1 \text{ se o tempo } t_i \text{ é um tempo de falha} \\ 0 \text{ se } t_i \text{ é um tempo de censura.} \end{cases}$$

Para o indivíduo *i* (*i* = 1, ..., *n*) tem-se o par (t_i, δ_i), sendo t_i = *tempo de falha ou censura.* Na presença de covariáveis tem-se (t_i, δ_i, x_i). Os tempos censurados devem ser sempre utilizados na análise, pois sua omissão certamente acarretará em conclusões viciadas. Dentre os tipos de censura podemos citar:

- Censura do tipo I ou à direita: ocorre, geralmente, quando após o fim do estudo alguns indivíduos não experimentaram o evento de interesse.
- Censura do tipo II: em vez do tempo final ser preestabelecido, o estudo será finalizado quando um número *k* de indivíduos experimentar o evento de interesse. Neste caso, os que deixaram de experimentar este evento terão seus tempos censurados.
- Censura aleatória: diferentemente das outras, esta não tem influência do pesquisador. Geralmente ocorre quando o indivíduo abandona o experimento antes da ocorrência do evento de interesse.
- Censura intervalar: ocorre quando não se conhece o tempo exato da ocorrência do evento de interesse.

Para estudarmos dados de sobrevivência, algumas funções são muito utilizadas. Uma delas, a função densidade de probabilidade, é definida como sendo o limite da probabilidade de um indivíduo experimentar o evento de interesse em um intervalo de tempo $(t,\ t + \Delta t)$ por unidade de Δt (comprimento do intervalo), ou simplesmente por unidade de tempo (LEE, 1992). É expressa por:

$$f(t) = \lim_{\Delta t \to 0} \frac{p(t\ \leq T\ < t + \Delta t)}{\Delta t},$$

em que $f(t) \geq 0$ para todo t, com a área abaixo da curva de $f(t)$ sendo igual a 1.

A função de sobrevivência é uma das principais funções utilizadas para descrever os tempos de sobrevivência. Dado que T é uma variável aleatória contínua e não negativa, podemos defini-la como sendo a probabilidade de um indivíduo não falhar (ou do evento não ocorrer) até um determinado tempo (LAWLESS, 1982). Denota-se como:

$$S(t) = P(T > t) \ = 1 - \int_0^t f(x)dx,$$

em que $f(x)$ é a função densidade de probabilidade.

Consequentemente, a função de distribuição acumulada é definida como a probabilidade de uma observação não sobreviver ao tempo t, isto é, $F(t) = 1 - S(t)$.

Podemos destacar algumas propriedades da função $S(t)$, como:

- É uma função monótona e decrescente;
- É contínua no tempo;
- $S(0) = 1$, isto é, a probabilidade de sobreviver ao tempo zero é um;
- $\lim_{t\to\infty} S(t) = 0$, isto é, a probabilidade de sobreviver ao tempo infinito é zero.

A função taxa de falha nos descreve a forma em que a taxa de falha muda com o tempo. Esta função fornece a probabilidade de o indivíduo experimentar o evento de interesse em um intervalo de tempo bem pequeno, dado que ele sobreviveu ao tempo t. É definida por:

$$\lambda\,(t) = \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t} \ge 0.$$

Além do interesse em estimar as funções especificadas anteriormente (densidade de probabilidade, sobrevivência e taxa de falha), tem-se o interesse em outras quantidades tais como o tempo médio de vida e a vida média residual.

O tempo médio de vida, como o próprio nome sugere, mede o tempo médio até a ocorrência do evento de interesse para um determinado perfil de clientes na carteira, sendo obtido pela área sob a curva obtida a partir da função de sobrevivência. Já a vida média residual ou tempo médio restante de permanência, $vmr(t)$, mede o tempo médio para experimentar o evento de interesse a partir de um tempo t, ou seja, o tempo que os clientes que não pagaram até o tempo t, podem ainda levar, em média, para quitar suas dívidas (COLOSIMO; GIOLO, 2006).

Para o cálculo do tempo médio e da vida média residual, tem-se suas respectivas expressões dadas por:

$$t_m = \int_0^{\infty} S(t)dt\,,$$

$$vmr(t) = \frac{\int_t^{\infty} S(u)du}{S(t)}\,.$$

Em 1958, Kaplan e Meier propuseram um estimador não paramétrico para estimar a função de sobrevivência, na presença de uma amostra com observações censuradas. Este estimador é denominado estimador de Kaplan-Meier ou estimador produto-limite.

Sejam $t_1 < t_2 < \cdots < t_k$, os k tempos distintos e ordenados de falhas, d_j o número de falhas em $t_j, j = 1, 2, \ldots, k$ e n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não apresentaram o evento de interesse e não foram censurados até o instante imediatamente anterior a t_j. O estimador produto-limite proposto por Kaplan-Meier é, então, definido por:

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

O Estimador de Kaplan-Meier (EKM) apresenta as seguintes propriedades estatísticas: é estimador não viciado para amostras grandes, é fracamente consistente, converge assintoticamente para um processo gaussiano e é o estimador de máxima verossimilhança de $S(t)$. As estimativas obtidas via o EKM são usualmente representada graficamente, mostrando o comportamento da curva de sobrevivência.

O EKM pode também ser utilizado para identificar o comportamento da função de sobrevivência de acordo com as categorias de covariáveis de interesse, produzindo, assim, evidências de possíveis fatores que possam afetar os tempos de sobrevida estudados.

3 MATERIAL E MÉTODOS

3.1 MATERIAL

3.1.1 Banco de dados

Os dados analisados neste trabalho foram disponibilizados por uma instituição financeira nacional de grande participação no mercado, que, por questão de sigilo, não terá seu nome informado. O sigilo com as informações dos clientes também foi mantido, de modo que somente a instituição tem como identificá-los.

O estudo ocorreu com base em uma amostra do portfólio total de clientes do segmento Empresas (Pessoa Jurídica – PJ). Trata-se de 642.707 informações de clientes que estão em inadimplência (tempo máximo de atraso de todos os contratos entre 61 e 3600 dias), que foram acompanhados por 24 meses, período aqui atribuído como “janela de performance”. Para analisar o tempo até o pagamento, foram avaliadas informações dos clientes fornecidas pela instituição (797 variáveis), assim como informações externas (negativações em *bureaus*).

A variável resposta considerada foi o tempo (em meses) desde o ponto de observação (em que o cliente deve estar com atraso máximo do escopo do modelo) até o retorno do cliente para a situação “em dia” e a permanência com atrasos inferiores a 30 dias nos próximos dois meses. A situação “em dia” pode expressar três circunstâncias: a) o cliente paga todas as parcelas em atraso e permanece com o contrato atual; (b) o cliente efetua uma renegociação e gera um novo contrato; e (c) o cliente efetua a liquidação de todos os contratos (pagamento total). A Tabela 3 mostra as informações de volumetria, em cada safra de coleta, bem como o número de clientes bons e maus em cada safra ao final da janela de performance.

TABELA 3 – Informações de volumetria e safras do banco de dados estudado

Safra	**Performance (24 meses)**				**Total**
	Maus	**(%)**	**Bons**	**(%)**	
Janeiro/2015	39.465	(83,4%)	7.828	(16,6%)	**47.293**
Março/2015	39.863	(83,1%)	8.125	(16,9%)	**47.988**
Maio/2015	40.512	(83,3%)	8.121	(16,7%)	**48.633**
Julho/2015	41.208	(83,4%)	8.226	(16,6%)	**49.434**
Setembro/2015	85.703	(82,8%)	17.814	(17,2%)	**103.517**
Novembro/2015	86.515	(82,8%)	17.947	(17,2%)	**104.462**
Junho/2017	-	-	-	-	**121.689**
Dezembro/2017	-	-	-	-	**119.691**
				Total	**642.707**

FONTE: Os autores (2018).

A pedido da instituição financeira, e com o intuito de manter alguns padrões por eles já adotados, serão utilizadas as primeiras quatro safras para o ajuste dos modelos (treinamento – DEV – e validação interna – VAL), duas para validação do modelo em safras externas (*Out-of-time* – OOT), e as duas últimas para averiguar a estabilidade em safras mais recentes (REC).

3.1.2 Recursos computacionais

Para realizar a análise exploratória e a seleção das variáveis foi utilizado o *software* SAS (*Statistical Analysis System*) Enterprise Guide 7.1. O mesmo *software* foi utilizado para o ajuste dos modelos de regressão logística e de mistura, este último com o auxílio de uma macro desenvolvida por Corbière e Joly (2007), denominada PSPMCM (*parametric and semiparametric mixture cure models*). Para o ajuste do modelo de promoção, fez-se uso do pacote GAMLSS no *software* R, versão 3.4.1. (R CORE TEAM, 2017).

3.2 MÉTODOS

3.2.1 Seleção de Covariáveis

Existem vários métodos que auxiliam no processo de seleção de covariáveis quando se deseja ajustar um modelo estatístico. As etapas do procedimento utilizado no presente trabalho estão descritas a seguir.

a) Etapa 1 – Análise univariada

i) Inicialmente, foram eliminadas as variáveis que apresentaram um volume igual ou superior a 80% de campos sem informação e/ou campos preenchidos com o valor zero (0). Também foram eliminadas as covariáveis com volume igual ou superior a 95% de campos com valores iguais;

ii) Foi calculada a estatística de Kolmogorov-Smirnov (KS – apresentado na Seção 3.2.3) associada à cada covariável, mantendo-se apenas as variáveis que apresentaram indicador igual ou superior a 5%.

Ao final dessa etapa, foram excluídas 546 covariáveis, dentre as 797 covariáveis disponíveis.

b) Etapa 2 – Análise multivariada: finalizada a etapa 1, foi realizada uma análise multivariada a fim de agrupar as covariáveis de acordo com suas similaridades e, então, selecionar as melhores dentro de cada grupo. Para esse propósito, os seguintes passos foram realizados:

i) Análise de conglomerados/agrupamentos: teve como objetivo agrupar covariáveis de acordo com as similaridades entre elas, sendo possível formar grupos com homogeneidade dentro do agrupamento e heterogeneidade entre eles. Para a realização dos agrupamentos de covariáveis, foi utilizado o *proc varclus* do SAS;

ii) Regressão logística: após obter os vários agrupamentos de covariáveis, foi ajustado um modelo de regressão logística dentro de cada agrupamento, com o auxílio do *proc logistic* do SAS, obtendo-se, assim, algumas estatísticas (por exemplo, a Estatística de Wald e seu respectivo *valor* p);

iii) Árvore de decisão: além da regressão logística, também foi feita uma árvore de decisão com as covariáveis de cada agrupamento. Foi utilizado o *proc split* do SAS, sendo que, a partir da árvore de decisão obtida, foi possível obter um escore para cada covariável preditora/explicativa. Com base nesse escore, foi identificada a importância de cada covariável e, em consequência, as que apresentaram impacto significativo na variável resposta; quanto maior o escore, maior a importância. Para mais informações sobre o assunto recomenda-se a leitura de Breiman (1984).

Com base nos passos da etapa 2, foi possível identificar quais covariáveis, dentro de cada agrupamento, poderiam ser consideradas candidatas ao modelo, seguindo um critério de seleção. Tal critério seguiu a seguinte ordem, ressaltando que mais de uma covariável pode ser selecionada dentro de cada agrupamento:

i) Maior Valor de Informação (IV – apresentado na etapa 3 desta seção) dentro de cada agrupamento (análise univariada);

ii) Maior KS dentro de cada agrupamento (análise univariada);

iii) Variável mais significativa, ou seja, com valor $p < 0{,}0001$, dentro de cada agrupamento (análise multivariada – regressão logística);

iv) Variável mais importante considerada pela árvore de decisão, dentro de cada agrupamento (análise multivariada – árvore de decisão);

v) Maior razão do R^2, ou seja, $1/R^2$ dentro de cada agrupamento (análise multivariada – análise de conglomerados/agrupamentos).

Ao final da etapa 2, foram excluídas 73 covariáveis.

c) Etapa 3 – Análise bivariada: nesta etapa, cada covariável foi examinada individualmente para determinar o seu poder preditivo. Uma variável com poder preditivo satisfatório é aquela que separa os clientes adimplentes (bons) dos inadimplentes (maus). Para determinar, em termos estatísticos, se uma covariável apresenta poder preditivo satisfatório, o seu valor de informação (IV) foi calculado.

i) Valor de Informação: o IV de uma variável foi calculado pela contagem dos bons e maus que caem em atributos (categoria, classes ou níveis), que abrange toda a gama de possíveis valores para cada covariável. Para cada atributo com contagem não nula, o peso de evidência (*Weights of Evidence* - WoE) foi calculado dividindo-se o percentual de bons pelo percentual de maus clientes, tomando-se o logaritmo natural deste quociente. O IV é definido como a soma, para todas os atributos de uma covariável, da diferença entre os percentuais de bons e maus clientes, multiplicado pelo peso de evidência, ou seja,

$$(IV) = \sum_{todas\ as\ categorias} (\%Bons - \%Maus) \times WoE.$$

Peso de Evidência (WoE): medida que apresenta de forma evidente a discriminação das categorias de uma covariável em relação ao critério de bons e maus, ou seja, é utilizada para medir se uma determinada classe está associada com um nível mais elevado ou mais baixo de risco. Se o WoE é positivo, isso significa que há uma proporção maior de bons do que maus caindo nesse mesmo atributo em particular (isto é, menor risco). Se o WoE é negativo, há uma maior proporção de maus do que bons (isto é, maior risco). Em geral, os atributos WoE inferiores recebem uma pontuação mais baixa do que aqueles com maior WoE.

$$(WoE) = ln\left(\frac{\%Bons}{\%Maus}\right).$$

Finalizada a etapa 3 (análise bivariada), outras 32 covariáveis explicativas foram excluídas. Ainda, houve 90 covariáveis que foram excluídas por informações de negócio, uma vez que elas apresentaram informações muito semelhantes à de alguma outra covariável, porém seu cálculo (busca de informação) demandava um tempo maior para processamento da informação. Dessa forma, pelo critério de otimização de tempo, optou-se em excluir essas covariáveis.

Das 56 covariáveis restantes, o método de seleção *stepwise* foi empregado para selecionar as covariáveis mais significativas, fazendo-se uso do teste de razão

de verossimilhanças. Na sequência, foram ainda selecionadas as covariáveis explicativas de maior importância (pelo teste de Wald) e com correlações inferiores ao valor absoluto de 0,5.

Ao final de todas as etapas descritas, restaram 16 covariáveis. Essas covariáveis foram incluídas em um ajuste de modelo de regressão logística, permanecendo as 7 covariáveis mais significativas e que, posteriormente, poderão pertencer aos modelos que serão ajustados. A Tabela 4 apresenta essas covariáveis e suas respectivas categorias.

3.2.2 Critério e categorização das covariáveis

Como todos os modelos aqui apresentados possibilitam o uso de covariáveis categóricas, optou-se em utilizar o Peso de Evidência (WoE), apresentado anteriormente, e o Índice de Desvio das Variáveis (VDI) para definir as categorias das covariáveis, que foram escolhidas minimizando-se estas medidas. Segue informações sobre o VDI.

a) Índice de Desvio das Variáveis (*Variable Deviation Index* – VDI): medida que apresenta os desvios da covariável, total e em cada categoria. A população fora do período de desenvolvimento é comparada com a de desenvolvimento (*Devolopment* – *DEV*), permitindo, assim, avaliar se a variável é estável após o desenvolvimento (*Out-of-time* – *OOT*).

$$VDI = ln\left(\frac{\%Total\ DEV}{\%Total\ OOT}\right) \times (\%Total\ DEV - \%Total\ OOT).$$

A Tabela 4 apresenta as covariáveis que permaneceram no modelo logístico, e que serão também incluídas nos modelos de sobrevivência, bem como as classes de cada variável e indicadores como: volume de cada classe, taxa de maus (não pagadores), WoE ao longo do tempo, KS e VDI. É possível observar que os indicadores se apresentaram estáveis e com forte poder de discriminação entre as categorias de cada variável.

Também é importante observar que as classes das covariáveis estão fazendo sentido no contexto do estudo sob análise, de forma que:

a) Quantidade de restritivos de operação vencida LP (em lucros e perdas) ativo ou decursado: quanto mais apontamentos possuir, maior será a taxa de maus e mais negativa será o WoE.

b) Tempo de relacionamento em meses (data da abertura da conta) do cliente até entrar em atraso: quanto antes entrar em atraso, maior será a taxa de maus e mais negativo a WoE.
c) Atraso máximo nos contratos de renegociação nos últimos 6 meses: quanto maior o atraso, maior será a taxa de maus e mais negativo será a WoE. No caso de não possuir renegociações, a informação pertencerá à classe que representa os menores atrasos.
d) Quantidade total de restritivos ativo: quanto menos restritivos possuir, menor a taxa de maus e mais positivo a WoE.
e) Nível do grau máximo de restritivo decursados: quanto maior o grau, maior a taxa de maus e mais negativa a WoE.
f) Percentual máximo de baixa do restritivo de cheque sem fundo: quanto menor o percentual, maior a taxa de maus e mais negativo a WoE.
g) Percentual de utilização do limite de cartão de crédito: caso ainda possua cartão ativo, menor a taxa de maus e maior a WoE. Se possuir produto, mas não utiliza, maior a taxa de maus e menor a WoE. Caso não possua o produto, pertence a uma classe intermediária.

TABELA 4 – Informações das variáveis finais categorizadas, candidatas ao ajuste do modelo logístico e aos de sobrevivência.

Covariável	Total (%)	Taxa de Maus	WoE						KS	VDI			
			Jan/15	Mar/15	Mai/15	Jul/15	Set/15	Nov/15		Set/15	Nov/15	Jun/17	Dez/17
(Var A) *Quantidade de restritivos de operação vencida LP (em lucros e perdas) ativo ou decursado*									29,84%	0,000	0,002	0,048	0,046
A.1 Mais que 3	24,90%	94%	-1,17	-1,25	-1,09	-1,12	-1,20	-1,13		0,000	0,001	0,011	0,011
A.2 Com 2 ou 3	31,28%	89%	-0,46	-0,52	-0,46	-0,44	-0,50	-0,56		0,000	0,000	0,003	0,002
A.3 Apenas 1	32,69%	82%	0,09	0,07	0,04	0,04	0,03	-0,01		0,000	0,000	0,000	0,000
A.4 Sem restritivo	11,13%	46%	1,76	1,83	1,74	1,74	1,83	1,84		0,000	0,001	0,034	0,032
(Var B) *Tempo de relacionamento em meses (data abertura da conta) do cliente até entrar em atraso*									20,80%	0,001	0,001	0,013	0,021
B.1 Até 12 meses ou sem informação	34,87%	93%	-0,68	-0,69	-0,71	-0,68	-0,71	-0,70		0,000	0,000	0,001	0,000
B.2 De 13 até 23 meses	19,76%	85%	-0,08	-0,07	-0,10	-0,14	-0,16	-0,17		0,000	0,000	0,002	0,004
B.3 Superior a 23 meses	45,37%	75%	0,40	0,40	0,41	0,41	0,42	0,41		0,000	0,000	0,000	0,003
(Var C) *Atraso máximo nos contratos de renegociação nos últimos 6 meses*									12,28%	0,000	0,000	0,016	0,024
C.1 Atrasos superior a 1197 dias	16,87%	92%	-0,92	-0,82	-0,86	-0,83	-0,85	-0,86		0,000	0,000	0,003	0,003
C.2 De 447 a 1196 dias de atraso ou sem renegociações	70,00%	84%	-0,02	-0,03	-0,03	-0,02	0,00	0,00		0,000	0,000	0,000	0,001
C.3 Sem atraso ou com até 446 dias de atraso	13,13%	70%	0,79	0,72	0,78	0,71	0,65	0,63		0,000	0,000	0,013	0,020
(Var D) *Quantidade total de restritivos ativo*									17,79%	0,000	0,000	0,007	0,011
D.1 Superior a 7 restritivos	37,33%	89%	-0,54	-0,51	-0,47	-0,52	-0,54	-0,54		0,000	0,000	0,003	0,005
D.2 De 4 a 7 restritivos	28,12%	84%	-0,03	-0,09	-0,08	-0,05	-0,09	-0,11		0,000	0,000	0,005	0,006
D.3 Sem restritivos ou até 3	34,55%	76%	0,45	0,46	0,43	0,44	0,49	0,49		0,000	0,000	0,000	0,000
(Var E) *Nível do grau máximo de restritivo decursados*									17,67%	0,001	0,001	0,020	0,031
E.1 Grave ou muito grave	23,24%	92%	-0,96	-0,87	-0,87	-0,93	-0,86	-0,87		0,000	0,000	0,011	0,019
E.2 Baixo ou médio	8,65%	87%	-0,32	-0,27	-0,31	-0,30	-0,27	-0,24		0,001	0,001	0,002	0,001
E.3 Sem restritivo ou grau remoto	68,11%	80%	0,25	0,23	0,24	0,24	0,23	0,22		0,000	0,000	0,007	0,011
(Var F) *Percentual máximo de baixa do restritivo de cheque sem fundo*									11,92%	0,001	0,003	0,004	0,005
F.1 Sem nenhuma baixa	23,71%	90%	-0,62	-0,64	-0,62	-0,57	-0,63	-0,62		0,001	0,001	0,000	0,000
F.2 Baixa parcial	37,96%	86%	-0,22	-0,19	-0,20	-0,28	-0,29	-0,31		0,000	0,000	0,001	0,001
F.3 Baixa total ou Sem nenhum restritivo	38,33%	74%	0,58	0,57	0,55	0,56	0,57	0,55		0,000	0,001	0,001	0,002
(Var G) *Percentual de utilização do limite de cartão de crédito*									12,84%	0,001	0,003	0,009	0,009
G.1 Sem utilização	82,79%	86%	-0,21	-0,19	-0,21	-0,22	-0,24	-0,26		0,000	0,001	0,000	0,000
G.2 Sem produto	6,07%	77%	0,45	0,36	0,49	0,44	0,53	0,57		0,000	0,001	0,004	0,003
G.3 Com utilização (0,01% a 100%)	11,15%	67%	0,91	0,88	0,85	0,89	0,89	0,89		0,001	0,002	0,005	0,006

Fonte: Os autores (2018).

3.2.3 Regressão Logística

A regressão logística nos permite estimar a probabilidade associada à ocorrência de um determinado evento dado um conjunto de variáveis exploratórias disponíveis. Por exemplo, a probabilidade de o cliente pagar todas as suas dívidas em um horizonte de n meses observando suas informações de cadastros e/ou histórias. Por ser de fácil compreensão, é muito utilizada nas instituições financeiras, seja para liberação de crédito ou ações de cobranças (GONÇALVES; GOUVÊA; MANTOVANI, 2013).

Seja Y a variavel resposta que pode assumir somente dois valores representados por sucesso $(Y = 1)$ e fracasso $(Y = 0)$. O valor esperado de Y é dado por:

$$E(Y) = P(Y = 1) = \pi$$

em que $P(Y = 1)$ denota a probabilidade de ocorrência do evento $(Y = 1)$.

A distribuição condicional da variável resposta Y segue uma binomial com probabilidade dada pela média condicional $\pi(\boldsymbol{x}) = E(Y \mid \boldsymbol{x})$. Assim, a probabilidade de sucesso $(Y = 1)$, dado o vetor de variáveis independentes $\boldsymbol{x} = (x_1, x_2, \ldots x_p)'$, é representado por $P(Y = 1 \mid \boldsymbol{x}) = \pi(\boldsymbol{x})$ e, consequentemente, $P(Y = 0|\boldsymbol{x}) = 1 - \pi(\boldsymbol{x})$ é a probabilidade de fracasso.

A forma tradicional de expressar o modelo de regressão logística é

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}$$

sendo $g(\boldsymbol{x})$ uma função contínua e linear nos parâmetros, podendo variar de $-\infty$ a $+\infty$, dada pela equação:

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

em que $\boldsymbol{\beta}$ denota o vetor de parâmetros do modelo, estimados pelo método da máxima verossimilhança.

A maneira mais usual de interpretar os coeficientes do modelo logístico é por meio da razão de chances, em inglês, *odds ratio.* Em um modelo com variável resposta binária e uma única covariável binária, a chance da resposta estar presente entre indivíduos com $x = 1$ é definida como $\pi(1) / [1 - \pi(1)]$. Então, a razão de chances denotada por ψ, é definida como segue:

$$\psi = \frac{\pi(1) \,/\, [1 - \pi(1)]}{\pi(0) \,/\, [1 - \pi(0)]}$$

e, consequentemente,

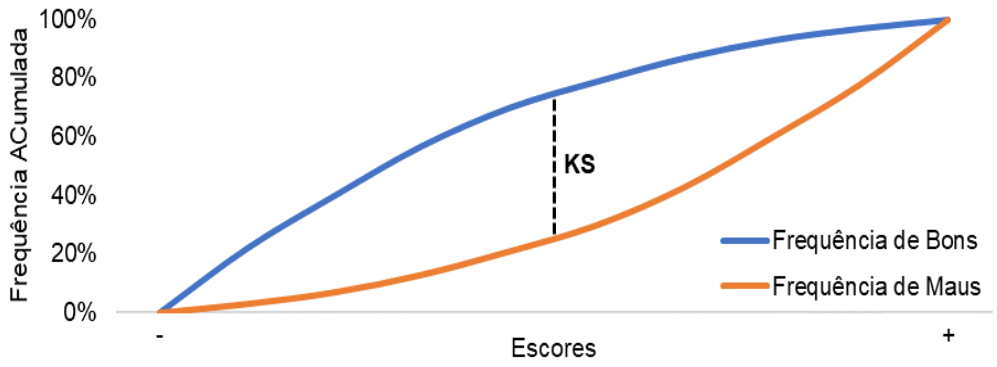
$$\psi = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

Um teste bastante utilizado para testar a igualdade entre funções de distribuição não paramétricas é o teste de Kolmogorov-Smirnov (KS). Em *scores*, para *Collections Scoring*, ele é utilizado para comparar a distribuição do escore entre os clientes bons e maus. Em modelos com boa capacidade de discriminação, espera-se que os clientes bons estejam concentrados nos escores mais altos e os clientes maus nos escores baixos. Assim, calculando a frequência acumulada de bons e maus por classes de escore, define-se a estatística de KS como:

$$(KS) = \max_i \, |\text{Frequência acumulada de pagadores na categoria i} - \text{Frequência acumulada de não pagadores na categoria i}|$$

de modo que quanto maior o valor de KS, melhor performance tem o modelo. A Figura 1 apresenta um exemplo teórico da estatística KS.

FIGURA 1 – Funções de distribuições empíricas para cálculo da estatística KS



Também utilizado para comprar ajuste de modelos, o Critério de Informação de Akaike (AIC) desenvolvido por Hirotugu Akaike em 1974, é uma informação quantitativa que representa a distância entre o modelo estimado e o modelo real de distribuição dos dados observados. O AIC é obtido por: $AIC = -2l(\theta) + 2d$, em que l denota o log da função de verossimilhança e d é a dimensão do vetor de parâmetros do modelo. O critério penaliza os modelos em função do número de parâmetros adicionados e é tomado para a escolha do modelo de regressão (o modelo com o menor AIC é indicado como o melhor modelo).

O Critério Bayesiano de Schwarz (SBC ou BIC), proposto por Schwarz (1978), tem como pressuposto a existência de um “modelo verdadeiro” que descreve a relação entre a variável dependente e as diversas variáveis explanatórias entre os diversos modelos sob seleção. Assim o critério é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O SBC é obtido por: $SBC = -2\log f(x_n|\theta) + p\log n$, em que $f(x_n|\theta)$ é o modelo escolhido, p é o número de parâmetros a serem estimados e n é o número de observações da amostra. O modelo com menor BIC é considerado o de melhor ajuste.

Um teste bastante utilizado para mensurar a importância das covariáveis em um dado modelo é denominado teste da razão de verossimilhanças. É obtido por meio da comparação entre o modelo sob a hipótese nula $H_0: \theta = \theta_0$ e o modelo irrestrito. A estatística deste teste, sob H_0, segue distribuição aproximada qui-quadrado com número de graus de liberdade igual a diferença de parâmetros dos dois modelos que estão sendo comparados, sendo expressa por:

$$TRV = -2log\left[\frac{L(\theta_0)}{L(\hat{\theta})}\right] = 2\left[logL\left(\hat{\theta}\right) - logL(\theta_0)\right] \sim \chi_p^2.$$

Outro teste, também utilizado para mensurar a importância das covariáveis em um modelo, é o teste de Wald. Este teste é uma generalização do teste t *de Stude*nt (WALD, 1943). Sob a hipótese nula $H_0: \theta = \theta_0$, sua estatística é dada por:

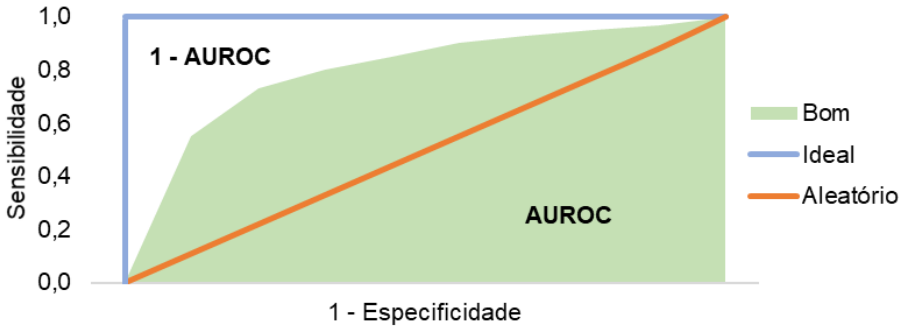
$$W = \left(\hat{\theta} - \theta_0\right)^T [-IF(\theta_0)]\left(\hat{\theta} - \theta_0\right) \sim \chi_p^2,$$

em que IF é a matriz de informação de Fisher avaliada em θ. Sob H_0, W segue distribuição aproximada qui-quadrado com graus de liberdade igual ao número de parâmetros testados.

Para avaliar a qualidade de predição do modelo, é comum o uso da curva ROC, que faz uso do conceito de “sensibilidade” e “especificidade” para descrever quantitativamente o desempenho de um modelo. A sensibilidade representa a probabilidade de o modelo apresentar um resultado positivo para um cliente bom, sendo calculada como a razão entre o número de clientes bons e o total de clientes. Quanto à especificidade, representa a probabilidade de o modelo apresentar resultado negativo para um cliente mau, sendo calculada como a razão entre o número de clientes maus e o total de clientes. Assim sendo, quanto maior o poder do modelo em discriminar os indivíduos bons e maus, mais a curva ROC se aproxima do canto

superior esquerdo, no ponto que representa a sensibilidade e especificidade do melhor valor de corte. Quanto melhor o modelo, mais a área sob a curva ROC (AUROC) se aproxima de 1, conforme ilustrado na Figura 2.

FIGURA 2 – Exemplo de curva ROC e indicador AUROC



Como alternativas ao modelo logístico em *Collections Scoring*, surgiram mais recentemente modelos no contexto de análise de sobrevivência, sendo possível não apenas visualizar a situação dos clientes ao final da janela de performance, mas também seu desempenho ao longo da mesma.

Nos modelos usuais de análise de sobrevivência, supõe-se que o evento de interesse pode ser observado em todos os indivíduos, desde que o tempo de acompanhamento seja suficientemente grande. No entanto, existem situações em que o evento de interesse pode não ocorrer para todos os indivíduos. Por exemplo, no caso financeiro, clientes inadimplentes que não conseguirão pagar suas dívidas. Estes indivíduos são considerados imunes ao evento de interesse e dizemos que o conjunto de dados referente a eles possui uma fração de cura ou fração de imunes.

Um indicativo da presença de indivíduos imunes na população é a ocorrência de um alto percentual de censura no final do estudo. Apesar disso, mesmo quando esse número representa proporções elevadas, é necessário avaliar se o tempo de acompanhamento foi grande o suficiente para que a suspeita da existência de uma fração de curados seja mantida (MALLER; ZHOU,1996). Uma forma de detectar a presença de imunes nos dados é observar o gráfico de Kaplan-Meier. Na presença de imunes, este gráfico tende a se estabilizar em um valor estritamente positivo durante um intervalo de tempo significativo, caracterizando uma função de sobrevivência imprópria (função que não tende a zero à medida que o tempo cresce). A presença desse comportamento indica a existência de indivíduos imunes na população.

Os modelos mais conhecidos propostos para esse tipo de situação são: o modelo de mistura e o modelo tempo de promoção.

3.2.4 Modelo de Mistura

Os modelos utilizados para analisar dados de sobrevivência geralmente assumem que todos na população estudada são suscetíveis ao evento de interesse e, eventualmente, experimentarão esse evento se o acompanhamento for suficientemente longo. Entretanto, tem surgido, em anos mais recentes, um interesse maior em modelos para a análise de dados de sobrevivência com fração de cura, o que se deve ao fato de que, por mais longo que seja o período de acompanhamento, pode existir um grupo de indivíduos que não experimentará o evento. Esses modelos com fração de cura, assumem que a população estudada é uma mistura de indivíduos suscetíveis (não curados), que podem experimentar o evento de interesse, e indivíduos não suscetíveis (curados), que nunca o experimentarão. Tais modelos permitem estimar, simultaneamente, se o evento de interesse tem probabilidade elevada de ocorrer, que é chamado de incidência, e quando ele ocorrerá, dado que isso pode acontecer, que é chamado de latência.

Seja U o indicador que denota se um indivíduo é susceptível $(U = 1)$ ou não suscetível $(U = 0)$ ao evento de interesse e T uma variável aleatória não negativa que indica o tempo até que o evento de interesse ocorra, definido apenas quando $U = 1$. O modelo de mistura com fração de cura é dado por:

$$S_p(t|\boldsymbol{x},\boldsymbol{z}) = \{[1 - \pi(\boldsymbol{z})] \times S(t \mid U = 0,\boldsymbol{x})\} + \{\pi(\boldsymbol{z})\, S(t \mid U = 1,\boldsymbol{x})\}$$
$$= \{[1 - \pi(\boldsymbol{z})]\} + \{\pi(\boldsymbol{z}) S(t \mid U = 1,\boldsymbol{x})\},$$

em que $S_p(t|\boldsymbol{x},\boldsymbol{z})$ é a função de sobrevivência incondicional associada à variável T para toda a população, $S(t \mid U = 1,\boldsymbol{x}) = P(T > t \mid U = 1,\boldsymbol{x})$ é a função de sobrevivência associada aos indivíduos suscetíveis com vetor de covariáveis $\boldsymbol{x} = (x_1, \dots, x_p)'$, e $\pi(\boldsymbol{z}) = P(U = 1 \mid \boldsymbol{z})$ é a probabilidade de ser suscetível dado o vetor de covariáveis $\boldsymbol{z} = (z_1, \dots, z_p)'$, que pode ou não incluir as mesmas covariáveis em $\boldsymbol{x}$.

Conhecedores da teoria de modelos de mistura com fração de cura, os autores Corbière e Joly (2007) desenvolveram e disponibilizaram uma macro SAS

capaz de ajustar tais modelos no contexto paramétrico e semiparamétrico, na presença de covariáveis.

Para o componente de incidência, $\pi(\boldsymbol{z})$, as funções de ligação disponíveis na macro citada são: a) logito ($\text{logit}\big(\pi(\boldsymbol{z})\big) = \beta_0 + \beta_1 z_1 + \cdots + \beta_q z_q = \beta'\boldsymbol{z}$), b) probito ($\Phi^{-1}\big(\pi(\boldsymbol{z})\big) = \beta'\boldsymbol{z}$), e c) complemento log-log ($\log\Big(-\log\big(1-\pi(\boldsymbol{z})\big)\Big) = \beta'\boldsymbol{z}$).

No caso do componente de latência, $S(t\,|U\,=\,1,\,\boldsymbol{x})$, é possível, no contexto paramétrico, considerar as distribuições:

a) Exponencial: a distribuição exponencial é uma das mais simples e importantes distribuições de probabilidade utilizada na modelagem de dados que representam o tempo até a ocorrência de algum evento de interesse. A distribuição exponencial se caracteriza por ser a única distribuição que apresenta uma função de taxa de falha constante, ou seja, a função de risco independe do tempo (LEE; WANG, 2003):

$$S(t \mid U\,=\,1,\,\boldsymbol{x}) = \exp[-\exp(\log(\,t) - \mu(\boldsymbol{x}))].$$

b) Weibull: a distribuição de Weibull foi proposta originalmente 1939. Esta distribuição é muito usada para descrever o tempo de vida de produtos industriais. Além disso, é muito importante na prática, pois apresenta uma grande variedade de formas para a função de risco:

$$S(t \mid U\,=\,1,\,\boldsymbol{x}) = \exp\left[-\exp\left(\frac{\log(\,t) - \mu(\boldsymbol{x})}{\sigma}\right)\right].$$

c) Log-normal: como o próprio nome sugere, existe uma relação entre a distribuição log-normal e a distribuição normal, o que facilita a apresentação e análise de dados provenientes da distribuição log-normal:

$$S(t \mid U\,=\,1,\,\boldsymbol{x}) = 1 - \emptyset\left(\frac{\log t - \mu(\boldsymbol{x})}{\sigma}\right).$$

O logaritmo de uma variável com distribuição log-normal com parâmetros μ e σ tem uma distribuição normal com média μ e desvio padrão σ, ou variância igual a σ^2. Portanto, dados provenientes de uma distribuição log-normal podem ser analisados segundo uma distribuição normal, se for considerado o logaritmo dos dados em vez dos valores originais (KLEIN; MOESCHBERGER, 1997).

d) Log-logística: esta distribuição tem se apresentado como uma alternativa à distribuição de Weibull e à log-normal:

$$S(t|U\ =\ 1,\ \boldsymbol{x}) = \left[1 + \exp\left(\frac{\log t - \mu\,(\boldsymbol{x})}{\sigma}\right)\right]^{-1}$$

em que $\emptyset$ denota a função de distribuição da N(0,1) e $\mu(\boldsymbol{x}) = \boldsymbol{\gamma}'\boldsymbol{x}$, com $\boldsymbol{\gamma}$ o vetor de coeficientes de regressão associados às covariáveis em $\boldsymbol{x}$.

No contexto semiparamétrico, está disponível na macro o modelo de risco proporcionais de Cox. Em termos do componente de latência, a função de sobrevivência associada a esse modelo fica expressa por:

$$S(t|U = 1, \boldsymbol{x}) = S_0(t \mid U = 1)^{\exp(\mu(\boldsymbol{x}))} = S_0(t \mid U = 1)^{\exp(\gamma'\boldsymbol{x})}.$$

O modelo de regressão de Cox (Cox, 1972) permite a análise de dados provenientes de estudos de tempo de vida em que a resposta é o tempo até a ocorrência de um evento de interesse, ajustando por covariáveis; o mesmo começa a ser utilizado extensivamente em estudos financeiros atualmente. A principal razão desta popularidade é a presença do componente não-paramétrico, que o torna bastante flexível. Aos interessados em se aprofundar mais em relação ao modelo de Cox sugerimos consultar Cox e Hinkley (1974).

No presente trabalho, fez-se uso da macro SAS mencionada, ajustando-se todos os modelos que ela possibilita, a fim de identificar o que melhor se ajusta aos dados.

3.2.5 Modelo Tempo de Promoção

O modelo de mistura, citado anteriormente, é um caso particular dos modelos tempo de promoção. Esses, apresentados por Yakovlev et al. (1993) e, posteriormente, estudados por Yakovlev e Tsodikov (1996) e aprofundado por Chen et al. (1999), assumem que o evento de interesse pode ocorrer por um número M não observável de causas, seguindo uma distribuição Poisson, tal que:

$$P(M = m) = \frac{\theta^m\, e^{-\theta}}{m!}, \quad m = 0,1,2, \ldots$$

Dado $M = m$, sejam Z_j variáveis aleatórias i.i.d. contínuas não negativas com função de distribuição $F(t)$ e função de sobrevivência $S(t)$, representando o tempo até o evento devido à *j*-ésima causa, $j = 1, \ldots, m$. Assumindo que M e Z_j são variáveis não observáveis independentes, e que o tempo até o evento é dado por:

$T = min\{Z_1, \ldots, Z_M\}$ se $M \geq 1$ e $T = \infty$ se $M = 0$, com $P(T = \infty | M = 0) = 1$, tem-se que o modelo tempo de promoção fica expresso, em termos da função de sobrevivência populacional, $Sp(t)$, por:

$$S_p(t) = \sum_{m=0}^{\infty} S(t)^m \, P[M = m],$$

e, em termos da função de densidade, por:

$$f_p(t) = -S_p{}'(t) = \sum_{m=0}^{\infty} m \, S(t)^{m-1} \, f(t) \, P[M = m].$$

Para estimação dos parâmetros é, em geral, utilizado o método da máxima verossimilhança, com a função de verossimilhança dada por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f_p(t_i | \boldsymbol{x}_i, \boldsymbol{z}_i)]^{\delta_i} [S_p(t_i | \boldsymbol{x}_i, \boldsymbol{z}_i)]^{1-\delta_i}$$

com δ_i a variável indicadora de falha e $\boldsymbol{\theta}$ o vetor de parâmetros.

No contexto do modelo tempo de promoção, Castro et al. (2010) propuseram uma outra distribuição para o número de causas, a distribuição binomial negativa, isto é, M ~ BN. Para viabilizar o ajuste desse modelo, desenvolveram algumas ferramentas com o pacote GAMLSS no *software R*. Para comparar o modelo que assume M ~ BN com outros modelos, utilizaram ferramentas de análise como: a *deviance* global (função de verossimilhança em seu máximo), os valores AIC e SBC (Schwarz Bayesian criterion), correlação de Filliben e gráficos dos resíduos.

O coeficiente de correlação de Filliben (FILLIBEN, 1975) calcula a correlação entre os dados ordenados e a estatística mediana ordenada da distribuição normal com média zero e variância um. Quanto mais próximo o coeficiente é de 1, mais "normais" são os dados.

De acordo com o modelo proposto por Castro et al. (2010), tem-se:

$$P[M = m] = \frac{\Gamma(\alpha^{-1} + m)}{m! \, \Gamma(\alpha^{-1})} \left(\frac{\alpha\theta}{1 + \alpha\theta}\right)^m (1 + \alpha\theta)^{\frac{-1}{\alpha}}$$

em que $m = 0,1,2,\ldots, \theta > 0, \alpha \geq -1$ e $1 - (1 + \alpha\theta) > 0$.

A esperança e variância de M são iguais a $E(M) = \theta$ e $Var(M) = \theta(1 + \alpha\theta)$. Desse modo, a função $S_p(t)$ para esse modelo fica expressa por:

$$S_p(t) = \sum_{m=0}^{\infty} S(t)^m \, P[m = m] = \begin{cases} \left[1 + \alpha\theta F(t)^{\frac{-1}{\theta}}\right] & \rightarrow \alpha > \dfrac{-1}{\theta}, \alpha \neq 0 \\ \exp[-\theta f(t)] & \rightarrow \alpha = 0 \end{cases}$$

e a função densidade de probabilidade por:

$$f_p(t) = -S_p^n(t) = \begin{cases} \left[1 + \alpha\theta F(t)^{\frac{-1}{\theta}}\right] & \rightarrow \alpha > \dfrac{-1}{\theta}, \alpha \neq 0 \\ \exp[-\theta f(t)] & \rightarrow \alpha = 0 \end{cases}$$

Para $t \rightarrow \infty$, $F(t) \rightarrow 1$. A fração de imunes resulta em:

$$p_0 = \begin{cases} [1 + \alpha\theta]^{\frac{-1}{\alpha}} & para \quad \alpha > \dfrac{-1}{\theta}, \alpha \neq 0 \\ \exp(-\theta) & para \quad \alpha = 0. \end{cases}$$

Note que quando $\alpha = 0$, $S_p(t) = \exp[-\theta F(t)]$, o que dá origem ao modelo de promoção, em que a proporção de imunes é $p_0 = \exp(-\theta)$. Já quando $\alpha = -1$, tem-se $S_p(t) = (1 - \theta) + \theta S(t)$, que corresponde ao modelo de mistura padrão em que a proporção de imunes é dada por $p_0 = 1 - \theta$.

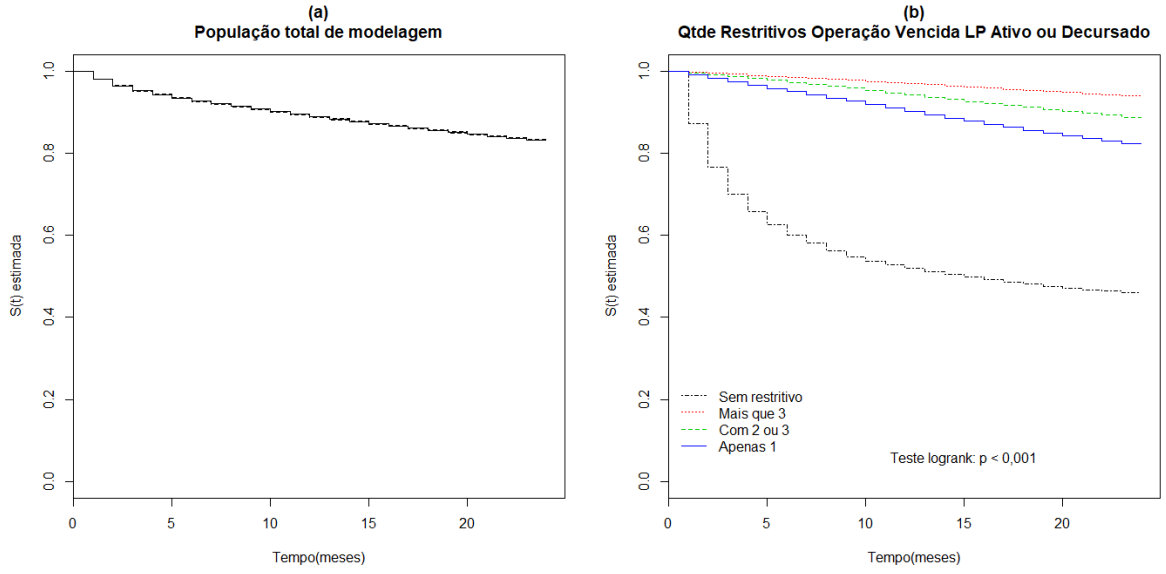
4 APRESENTAÇÃO DOS RESULTADOS E DISCUSSÃO

4.1 ANÁLISE DESCRITIVA

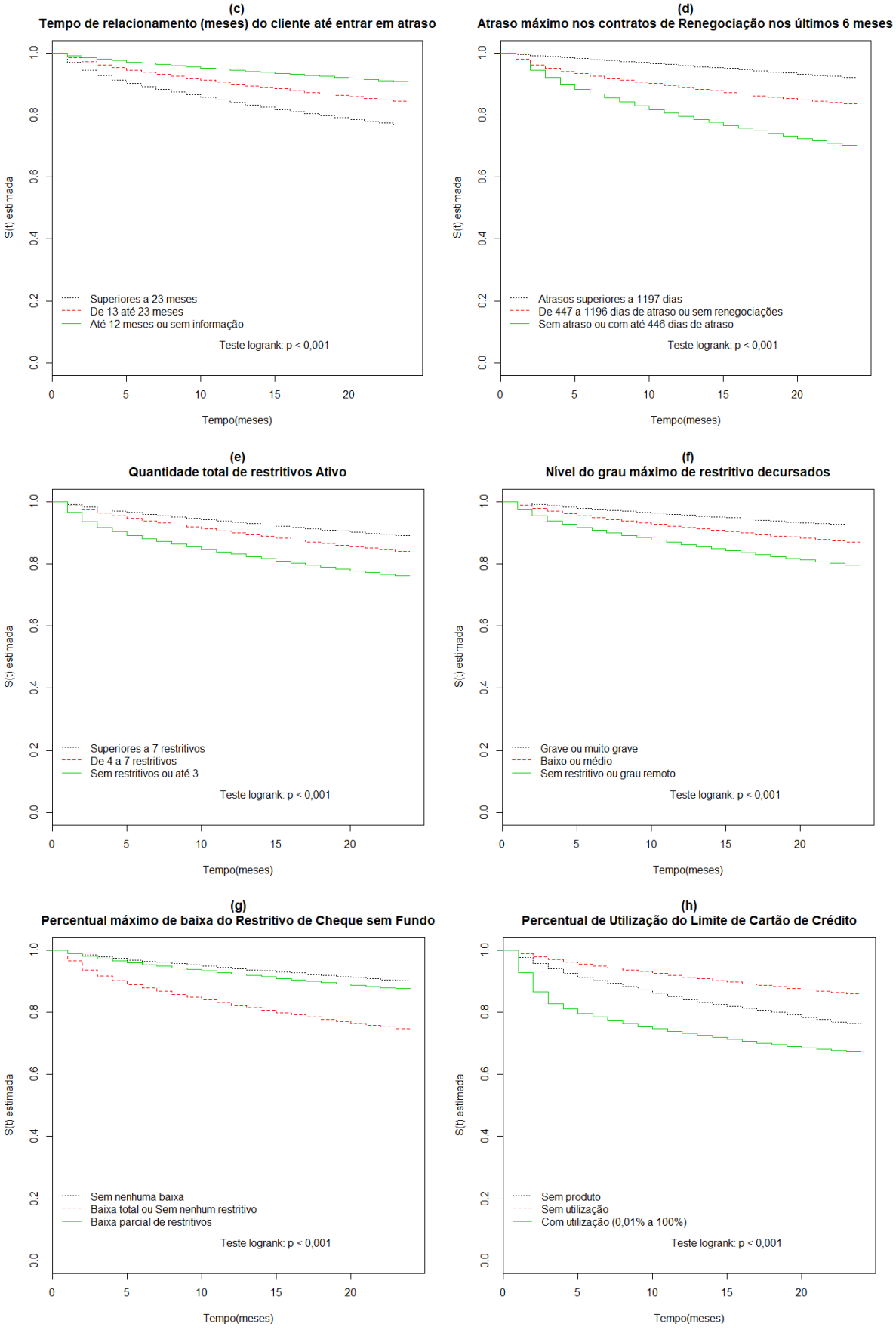
A partir das principais informações do banco de dados, conforme descrito na Seção 3.1.1, foram comparadas as curvas de Kaplan-Maier estimadas para cada categoria de cada covariável. Para isso, foi aplicado o teste *logrank* com o intuito de verificar se existem similaridades entre as curvas de sobrevivência sujeitas a dados censurados. Em todos os casos, o resultado foi pela rejeição da hipótese nula, evidenciando, assim, a existência de diferenças entre as curvas associadas a pelo menos duas categorias de cada covariável. Como o teste *logrank* indicou que a curva de sobrevivência de pelo menos uma das categorias difere das demais (para cada covariável), também foram efetuados testes dois a dois para cada covariável, que também evidenciaram a rejeição da hipótese nula. Assim sendo, as categorias definidas para cada covariável foram mantidas e utilizadas tanto para o modelo logístico, quanto para os modelos de sobrevivência.

As curvas de sobrevivência mostradas de (b) a (h) na Figura 3 representam as probabilidades de sobreviver ao tempo t (isto é, de o cliente não liquidar a dívida em atraso) estimadas via o EKM para as sete covariáveis descritas na Tabela 4. Já a curva mostrada em (a) na Figura 3 representa a curva de sobrevivência da população sob estudo, estimada por Kaplan-Meier na ausência de covariáveis.

FIGURA 3 – Curvas de Kaplan-Meier da população sob estudo e para cada covariável



Continuação FIGURA 3 – Curva de Kaplan-Maier da população sob estudo e para cada covariável



FONTE: Os autores (2018).

4.2 MODELO LOGÍSTICO

Para ajustar um modelo logístico aos padrões que a instituição atual possui, foram inicialmente definidas as safras que compuseram cada etapa do modelo:

- Safras de desenvolvimento do modelo (TOT): composta pelas safras de janeiro, março, maio e julho de 2015 (70% treinamento e 30% validação);
- Safras de validação fora do tempo de desenvolvimento (OOT): composta pelas safras de setembro e novembro de 2015. Etapa essa capaz de averiguar se o ajuste se faz eficaz em safras fora do desenvolvimento;
- Safras de validação recente (REC): composta pelas safras de junho e dezembro de 2017. Estas safras são utilizadas com o intuito de observar se os perfis elencados no desenvolvimento ainda estão apresentando a mesma distribuição (identificar mudanças na distribuição de perfis no portfólio).

Após as etapas de seleção de variáveis descritas na Seção 3.2.1, das quais permaneceram sete, foi utilizado o método de seleção *forward*, fazendo todas as combinações possíveis. Como todas as covariáveis finais já apresentavam valor p significativo, foi utilizado o critério de Akaike (AIC) e a curva ROC para selecionar a melhor combinação de covariáveis. Ao final, a combinação que apresentou o melhor ajuste foi a composta das covariáveis: quantidade de restritivos de operação vencida LP ativo ou decursado (*Var A*), tempo de relacionamento em meses do cliente até entrar em atraso (*Var B*) e quantidade total de restritivos ativo (*Var D*).

A Tabela 5 mostra que todas as covariáveis apresentaram correlações inferiores a 0,5, o que satisfaz o critério usualmente definido pela instituição para manutenção de covariáveis no modelo.

TABELA 5 – Correlações entre as variáveis que permaneceram no modelo logístico ajustado

Covariáveis	Quantidade de restritivos de operação vencida LP ativo ou decursado	Tempo de relacionamento em meses do cliente até entrar em atraso	Quantidade total de restritivos ativo
Quantidade de restritivos de operação vencida LP ativo ou decursado	1,0000	0,2140	0,3844
Tempo de relacionamento em meses do cliente até entrar em atraso	0,2140	1,0000	0,0665
Quantidade total de restritivos ativo	0,3844	0,0665	1,0000

FONTE: Os autores (2018).

A Tabela 6 apresenta as estimativas associadas a cada categoria das covariáveis que permaneceram no modelo de regressão logística ajustado, bem como outros indicadores importantes. Para facilitar a interpretação, foi utilizada a categoria com menor WoE como sendo a de referência.

TABELA 6 – Estimativas e valores estatísticos associados às variáveis no modelo logístico

Parâmetro	Categoria	Estimativa	Erro padrão	Valor p	IC Estimativas (95%)
Intercepto	-	-3,2674	0,0231	<0,0001	(-3,3127 :-3,2221)
Quantidade Restritivo LP Ativo ou Decursado	Com 2 ou 3	0,4962	0,0242	< 0,0001	(0,4487 : 0,5436)
	Apenas 1	0,8866	0,0239	< 0,0001	(0,8397 : 0,9335)
	Sem restritivo	2,5493	0,0255	< 0,0001	(2,4993 : 2,5993)
Tempo Relacionamento (em meses)	De 13 até 23 meses	0,4343	0,0206	< 0,0001	(0,3939 : 0,4746)
	Superiores a 23 meses	0,7919	0,0166	< 0,0001	(0,7593 : 0,8245)
Quantidade total de Restritivo Ativo	De 4 a 7 restritivos	0,3121	0,0182	< 0,0001	(0,2764 : 0,3478)
	Sem restritivo ou até 3	0,3936	0,0174	< 0,0001	(0,3594 : 0,4277)

FONTE: Os autores (2018).

Com o intuito de verificar a estabilidade das informações nas covariáveis ao longo do tempo, foi feito uso do VDI entre as etapas do modelo, tendo sempre como referência as informações coletadas na etapa TOT. Por convenção da instituição, variações superiores a 0,10 representam ponto de atenção e não são aceitas no modelo. Como é possível observar na Tabela 7, e como foi direcionado na etapa de seleção das covariáveis, todas que permaneceram no modelo se apresentaram estáveis e bem inferiores ao limite superior permitido.

TABELA 7 – Estabilidade (VDI) nas variáveis para safras pós desenvolvimento do modelo

Etapas	Quantidade de restritivos de operação vencida LP ativo ou decursado	Tempo de relacionamento em meses do cliente até entrar em atraso	Quantidade total de restritivos ativo
OOT	0,0008	0,0007	0,0003
REC	0,0470	0,0166	0,0089

FONTE: Os autores (2018).

A partir da Tabela 8, em que são apresentados os valores de AIC (Critério de Informação de Akaike), SBC (Critério Bayesiano de Schwarz) e também de $-2 * (log\ L)$ ($log\ L\ =$ logaritmo da função de verossimilhança), pode-se constatar valores inferiores quando da inclusão das covariáveis, o que sugere que essas variáveis ajudam a explicar a variável resposta.

TABELA 8 – Estatísticas associadas ao modelo logístico selecionado

Distribuição	AIC	SBC	-2 * Log L
Somente intercepto	122.070,02	122.079,83	122.068,02
Intercepto e covariáveis	104.656,97	104.696,23	104.648,97

FONTE: Os autores (2018).

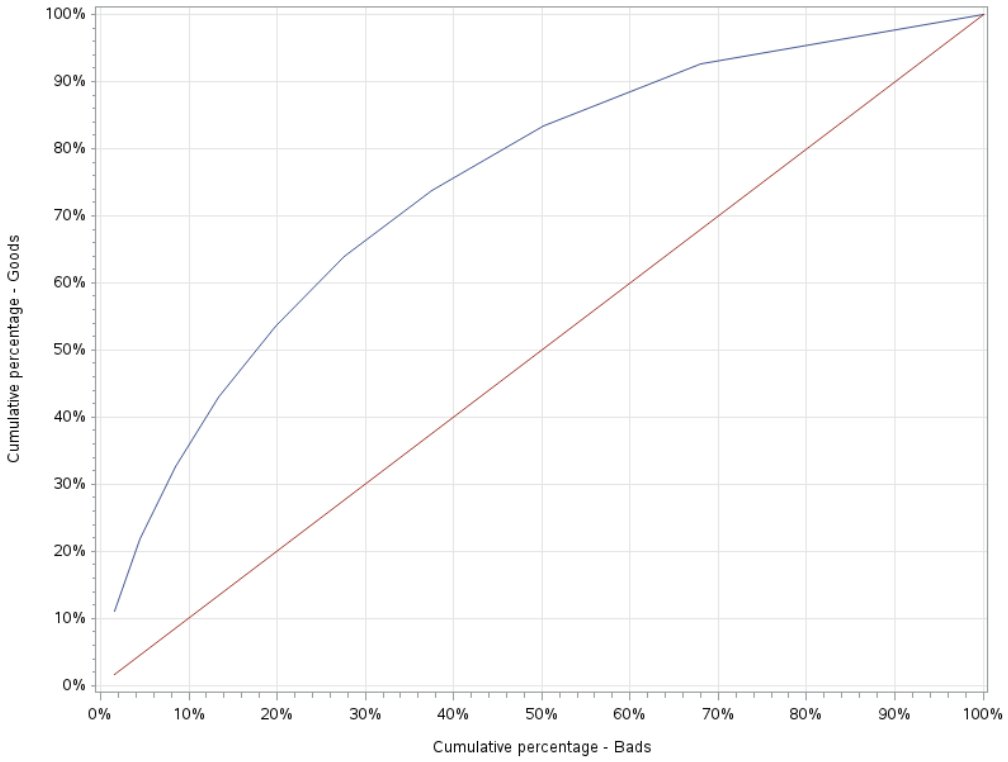
Como as covariáveis auxiliaram a explicar a variável resposta (Tabela 8) e mostraram-se estáveis ao longo do tempo (Tabela 7), foi possível apresentar a expressão do modelo logístico. Por se tratar de covariáveis categóricas, nota-se que estas foram incluídas no modelo por meio de variáveis *dummy*. Como categoria de referência para a covariável “quantidade de restritivos de operação vencida LP ativo ou decursado” foi utilizado “sem restritivo”. Já para a covariável “tempo de relacionamento em meses do cliente até entrar em atraso” foi utilizada a categoria “até 12 meses ou sem informação” e, finalmente, para a covariável “quantidade total de restritivos ativo”, a categoria “superior a 7 restritivos”. Assim sendo, segue a expressão do modelo logístico ajustado aos dados:

$$logit\big(\hat{\pi}(\boldsymbol{z})\big) = -3{,}2674 + 0{,}4962z_{A.2} + 0{,}8866z_{A.3} + 2{,}5493z_{A.4} + 0{,}4343z_{B.2} + 0{,}7919z_{B.3} + 0{,}3121z_{D.2} + 0{,}3936z_{D.3}$$

em que as categorias A.2, A.3, ..., D.2 foram definidas na Tabela 4.

O modelo ajustado apresentou um bom poder de discriminação, com a área abaixo da curva ROC, mostrada na Figura 4, igual a $0{,}747$.

FIGURA 4 – Curva ROC associada ao modelo de regressão logística ajustado aos dados



FONTE: Os autores (2018).

Outros indicadores de força medidos estão informados na Tabela 9, em que, além do KS e IV já descritos anteriormente, se tem:

- Taxa de inadimplência: quantos clientes permaneceram inadimplentes ao final dos 24 meses em relação aos que foram observados no início.
- *Odds ratio*: a razão entre a chance de o cliente ser ‘bom’ e a do cliente ser ‘mau’.
- Coeficiente de Gini: consiste em um número entre 0 e 1, em que 0 corresponde à completa igualdade e 1 corresponde à completa desigualdade.

TABELA 9 – Indicadores para avaliar estabilidade na performance do modelo

Etapas	Taxa de Inadimplência	KS	Odds Ratio	IV	Coeficiente de Gini
TOT	0,833	0,3558	0,2006	1,0684	0,534
OOT	0,828	0,3739	0,2076	1,1379	0,545

FONTE: Os autores (2018).

É possível observar, a partir da Tabela 9, que todos os indicadores apresentaram informações muito semelhante nas etapas, o que direciona para a compreensão de que o modelo logístico se apresenta bem ajustado, com indicadores estáveis e com bom poder de discriminação.

4.3 MODELO DE MISTURA

Para o ajuste do modelo de mistura com fração de inadimplentes, foi feito uso da macro SAS disponibilizada por Corbière e Joly (2007). Como já discutido anteriormente, esta macro possibilita ajustar modelos de mistura nos contextos paramétrico e semiparamétrico com covariáveis.

Foram feitas todas as combinações que a macro possibilita para os componentes “incidência” e “latência”, isto é,

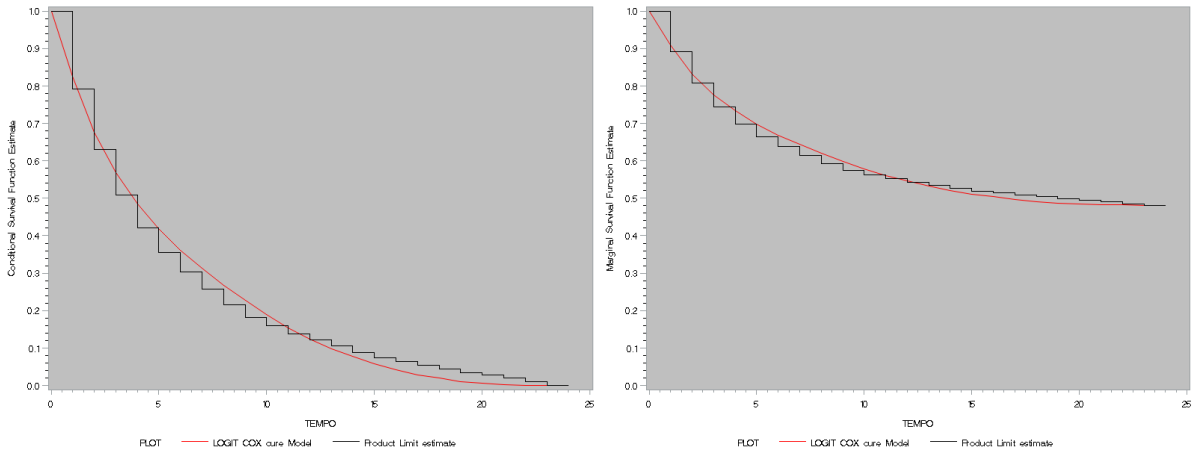
- Incidência: as funções de ligação probito, logito e complemento log-log.
- Latência: no caso semiparamétrico utilizou-se a função de sobrevivência associada ao modelo de Cox e, para os casos paramétricos, as funções: exponencial, Weibull, log-normal e log-logística.

Todas as sete covariáveis finais foram testadas e excluídas manualmente uma a uma, respeitando-se o grau de importância de acordo com a estatística de Wald, do menor ao maior, até se atingir o melhor ajuste. Entre todas as possibilidades, as combinações de covariáveis que se mostraram mais eficientes foram: quantidade de restritivos de operação vencida LP ativo ou decursado (*Var A*), tempo de relacionamento em meses do cliente até entrar em atraso (*Var B*) e quantidade total de restritivos ativo (*Var D*).

Para efeito de comparação e escolha do melhor ajuste, se fez uso de análises gráficas, das informações de iterações até a convergência e, também, do coeficiente de correlação de Pearson e do R^2, que auxiliaram na compreensão de quanto a curva estimada pelo modelo se aproxima da observada. Como os modelos apresentaram as mesmas covariáveis (logístico e de mistura), tem-se para ambos 36 perfis de clientes (combinações das categorias das três covariáveis nos modelos).

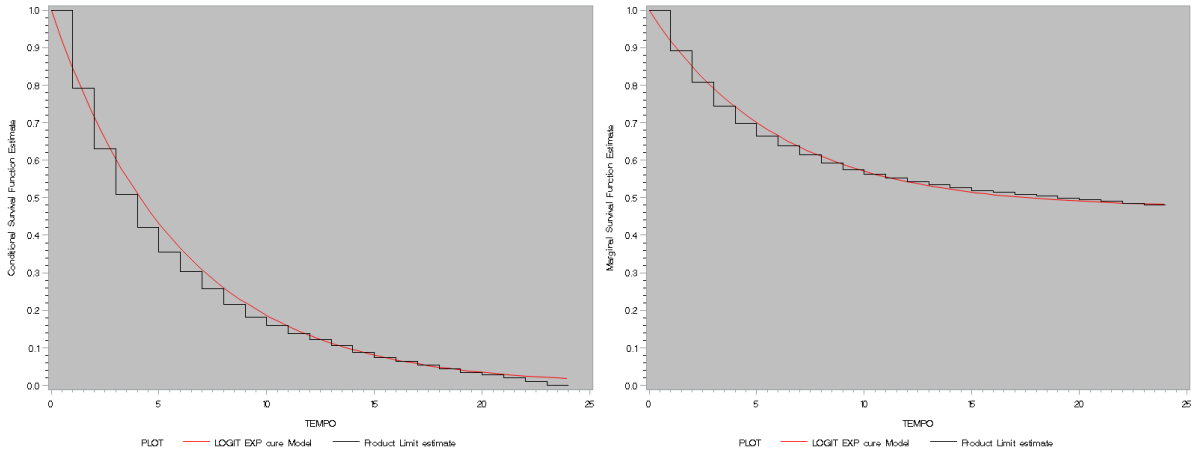
As Figuras 5 a 19 apresentam as curvas estimadas para $S_p(t|\boldsymbol{x},\boldsymbol{z})$ e $S(t\,|\boldsymbol{x})$, com $\boldsymbol{x}$ *e* $\boldsymbol{z}$ os vetores associados a cada modelo ajustado, em que podemos observar que os modelos com função de ligação logito ou probito e com $S(t\,|\boldsymbol{x})$ sob os modelos de Cox, Exponencial ou Weibull são os que apresentaram os melhores ajustes gráficos.

FIGURA 5 – Curva estimada para $S(t\,|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste logito + Cox



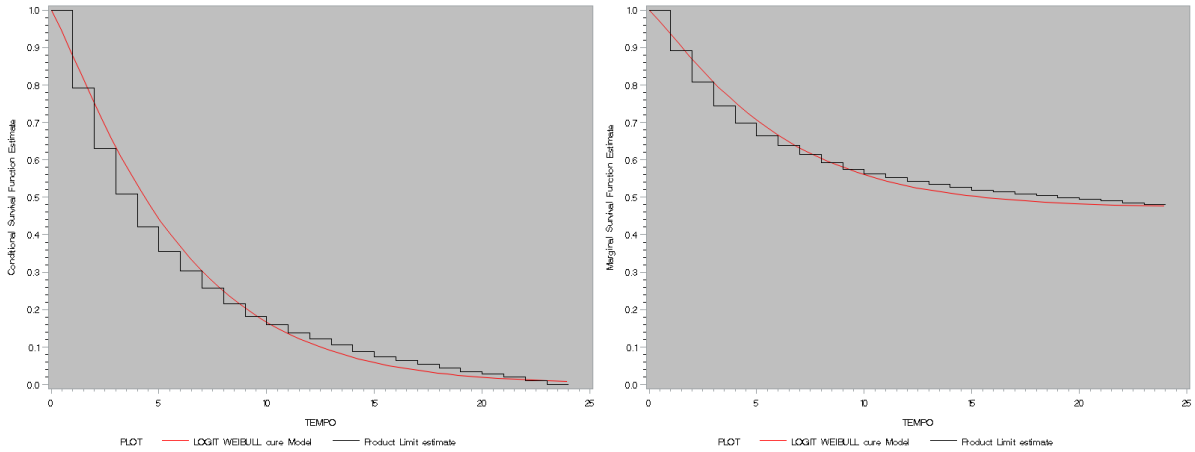
FONTE: Os autores (2018).

FIGURA 6 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste logito + exponencial



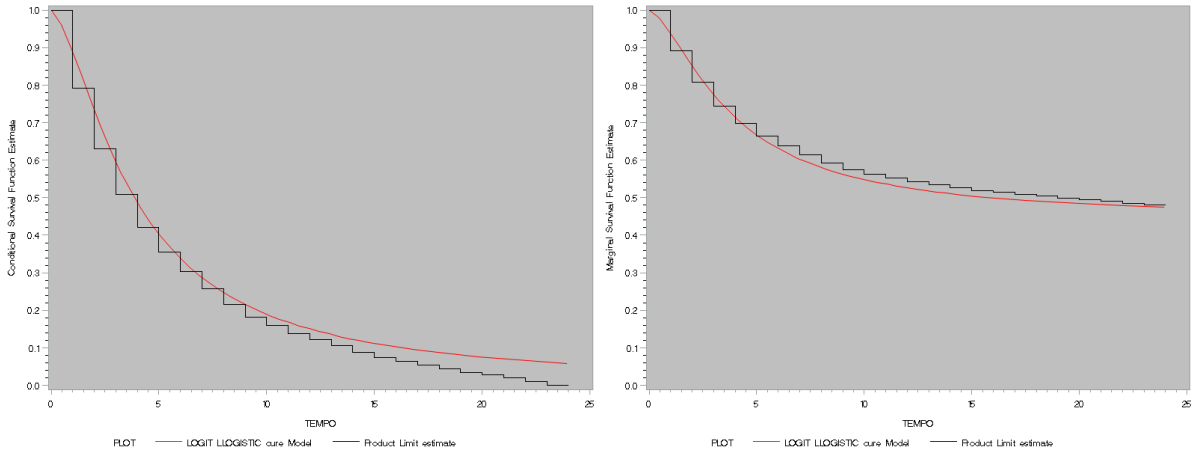
FONTE: Os autores (2018).

FIGURA 7 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste logito + Weibull



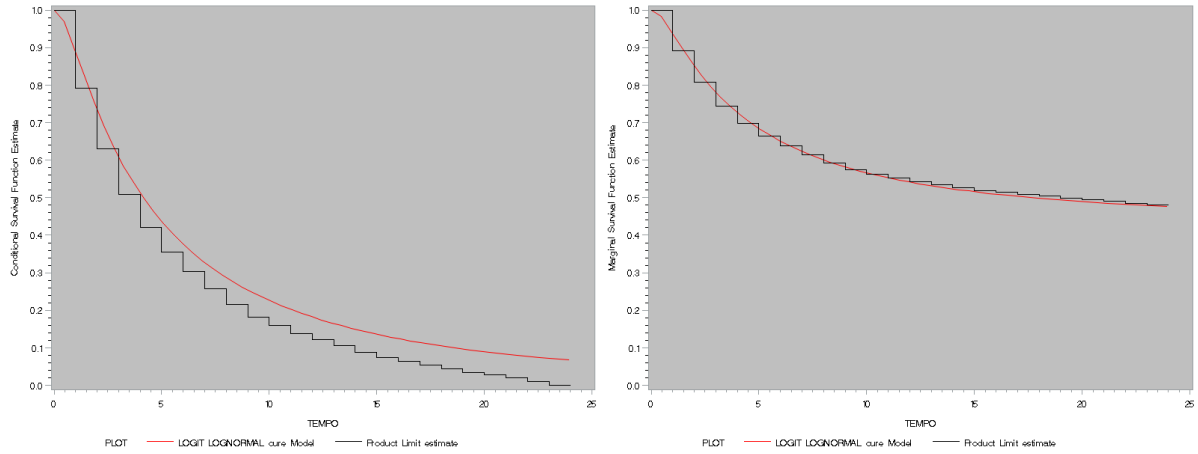
FONTE: Os autores (2018).

FIGURA 8 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste logito + log-logística



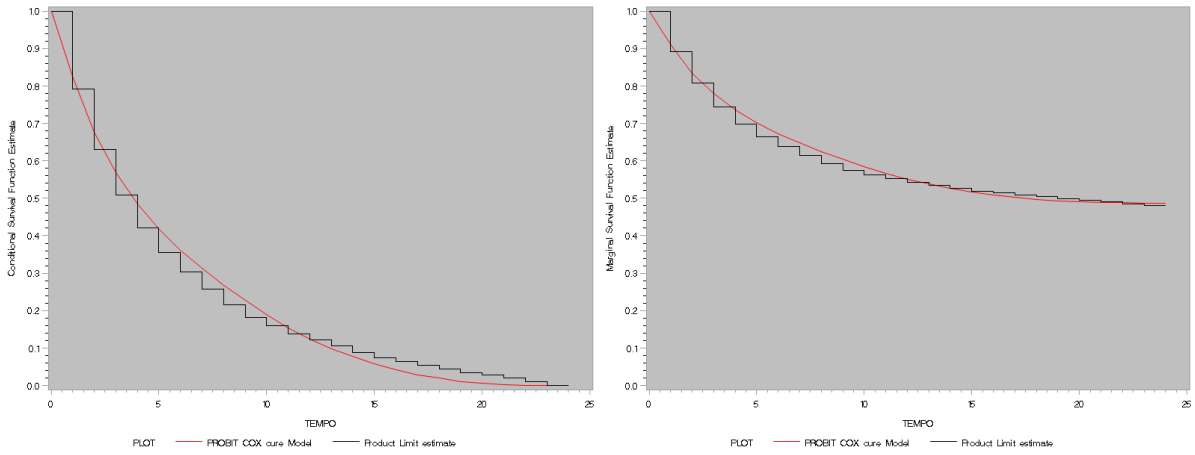
FONTE: Os autores (2018).

FIGURA 9 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste logito + log-normal



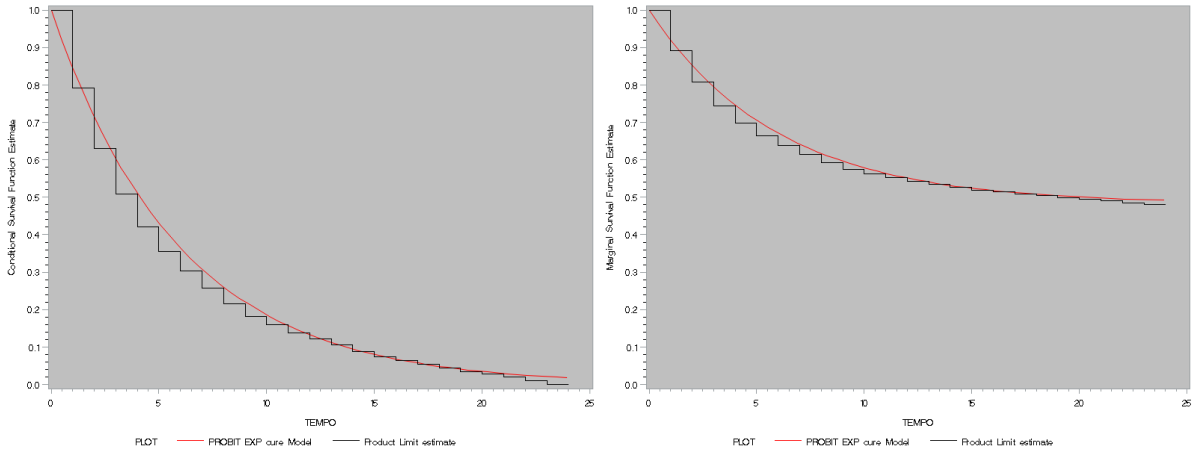
FONTE: Os autores (2018).

FIGURA 10 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste probito + Cox



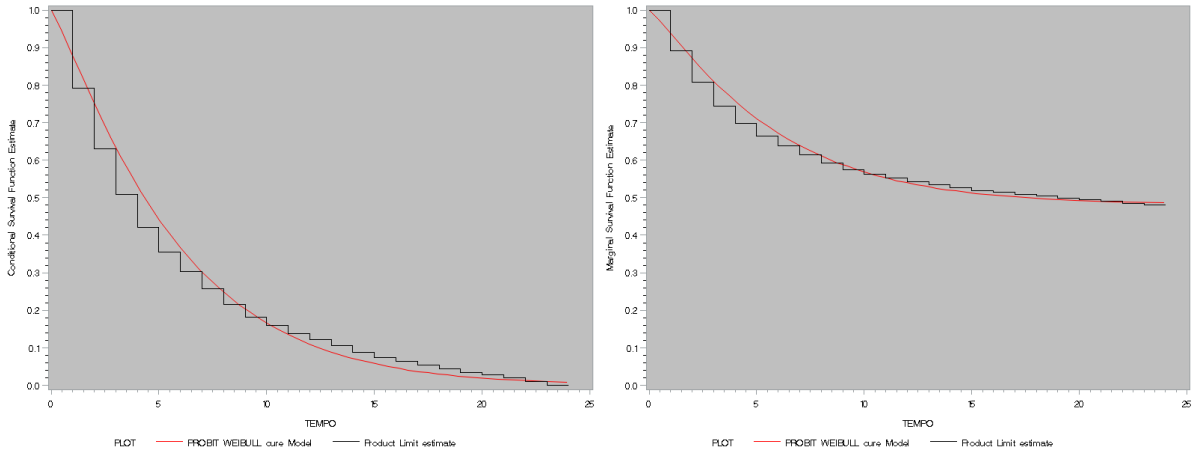
FONTE: Os autores (2018).

FIGURA 11 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste probito + exponencial



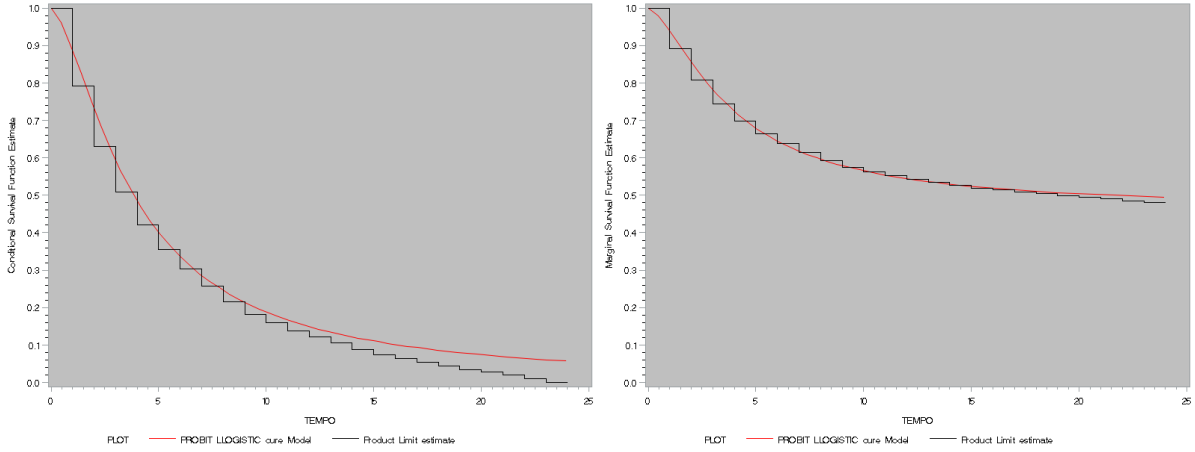
FONTE: Os autores (2018)

FIGURA 12 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste probito + Weibull



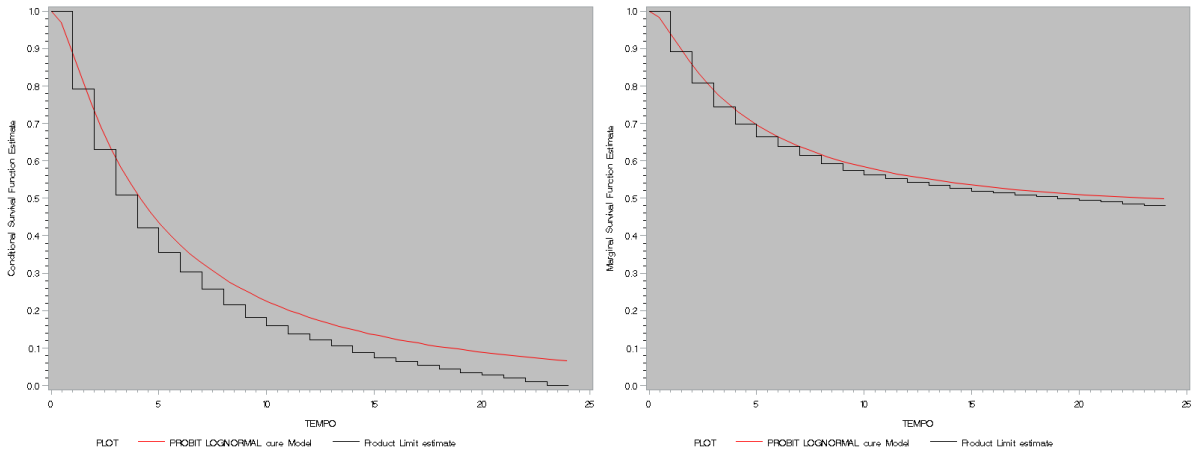
FONTE: Os autores (2018).

FIGURA 13 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste probito + log-logística



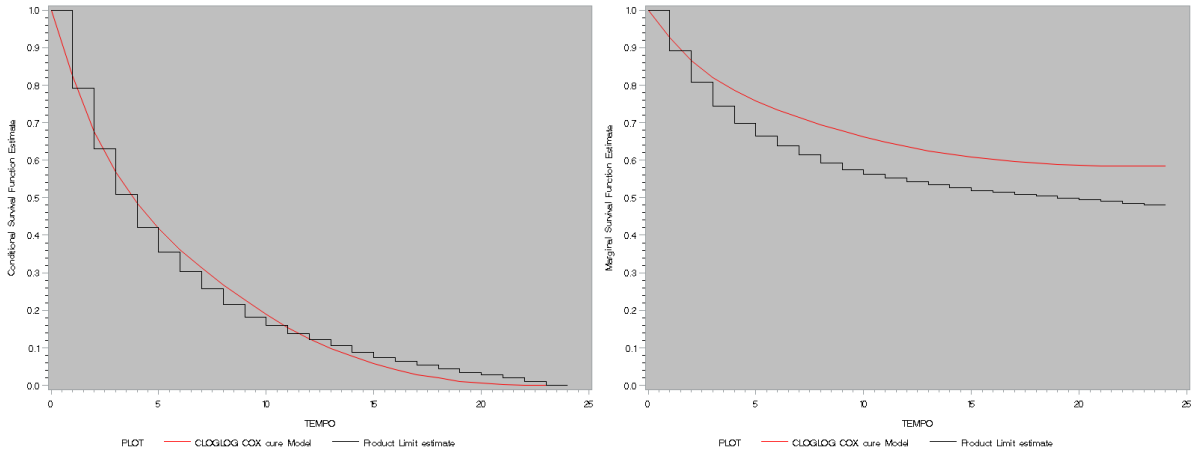
FONTE: Os autores (2018).

FIGURA 14 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste probito + log-normal



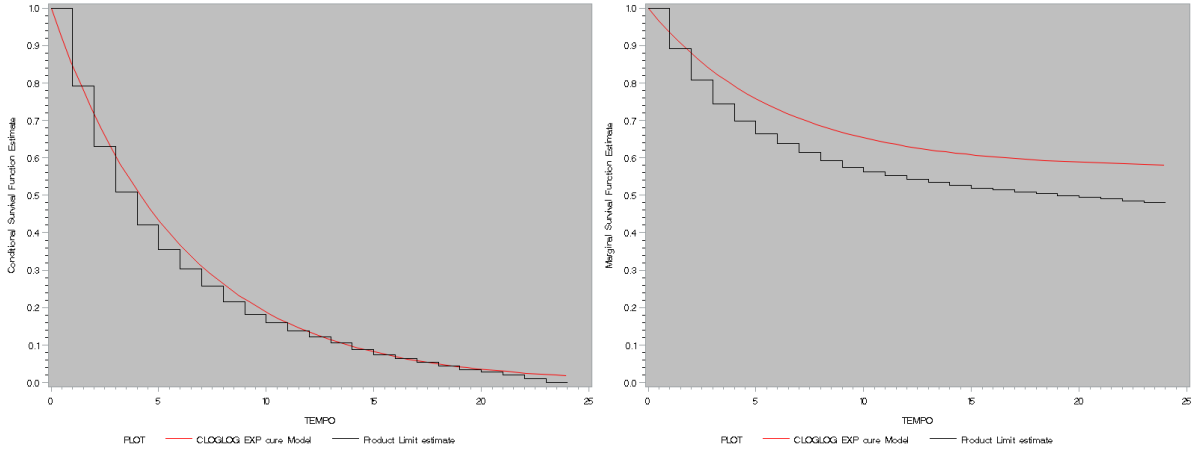
FONTE: Os autores (2018).

FIGURA 15 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste complemento log-log + Cox



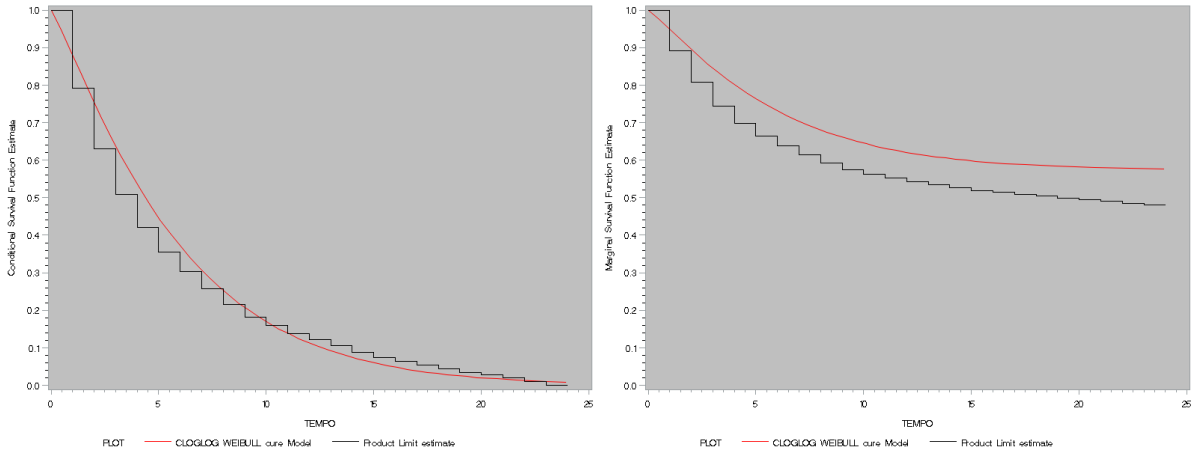
FONTE: Os autores (2018).

FIGURA 16 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste complemento log-log + exponencial



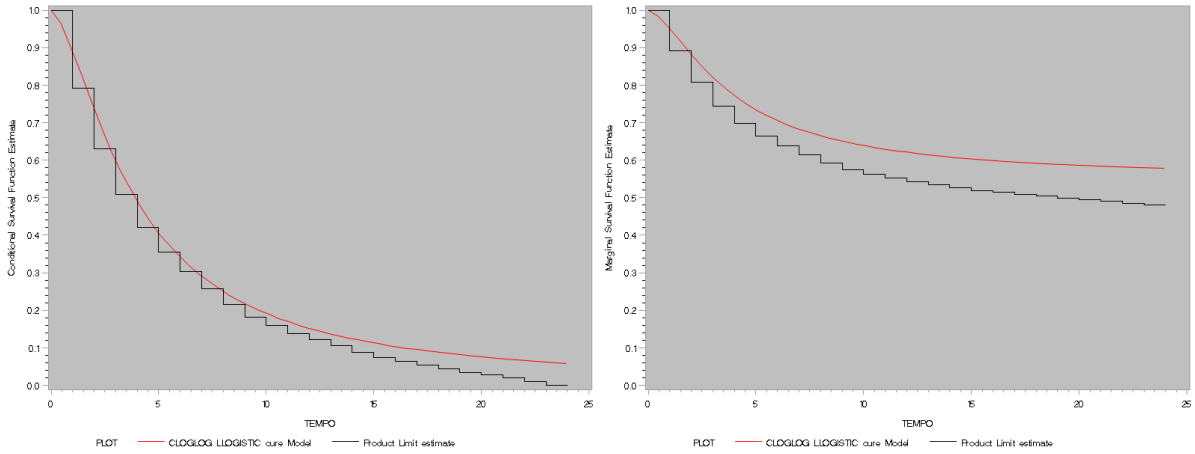
FONTE: Os autores (2018).

FIGURA 17 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste complemento log-log + Weibull



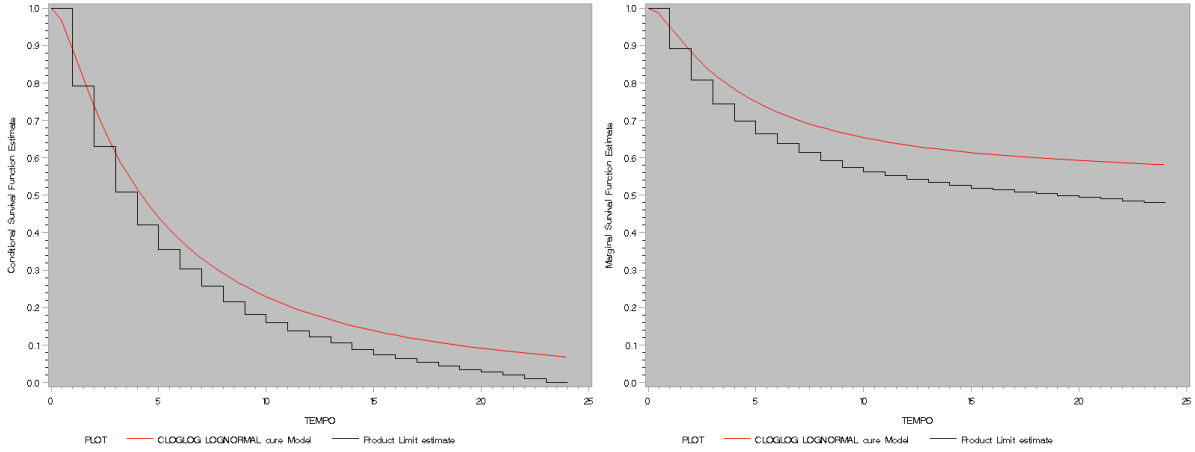
FONTE: Os autores (2018).

FIGURA 18 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste complemento log-log + log-logística



FONTE: Os autores (2018).

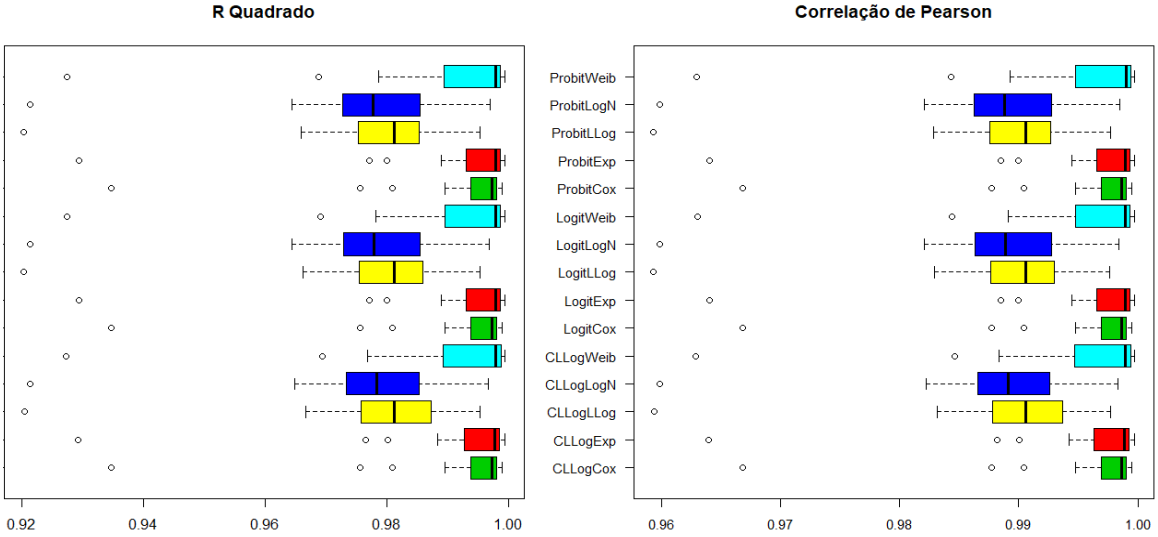
FIGURA 19 – Curva estimada para $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$, respectivamente, com $\boldsymbol{x}$ e $\boldsymbol{z}$ os vetores associados ao ajuste complemento log-log + log-normal



FONTE: Os autores (2018).

Seguindo com a comparação dos modelos, a fim de definir o que melhor se ajustou aos dados, tem-se na Figura 20 a representação (*boxplots*) dos valores de R^2 e da correlação de Pearson (obtidos para os 36 perfis de clientes) para cada ajuste. Além disso, a Tabela 10 apresenta um resumo das principais medidas.

FIGURA 20 – *Boxplots* do R^2 e da correlação de Pearson para os modelos de misturas ajustados



FONTE: Os autores (2018).

TABELA 10 – Resumo dos principais indicadores, R^2 e correlação de Pearson para os modelos de misturas ajustados

Incidência	Latência	Nº Interações	R^2			Correlação de Pearson			AIC	-2 * Log L
			Mínimo	Média	Máxima	Mínimo	Média	Máxima		
Logito	Cox	2	0,9348	0,9936	0,9990	0,9668	0,9968	0,9995	174474	174472
Logito	Exponencial	51	0,9294	0,9937	0,9994	0,9641	0,9968	0,9997	341882	341850
Logito	Weibull	47	0,9275	0,9924	0,9995	0,9631	0,9962	0,9997	341406	341372
Logito	LogLogística	35	0,9204	0,9802	0,9953	0,9594	0,9900	0,9976	219253	219219
Logito	LogNormal	32	0,9214	0,9784	0,9969	0,9599	0,9891	0,9984	339484	339450
Probito	Cox	2	0,9348	0,9936	0,9990	0,9668	0,9968	0,9995	174474	174472
Probito	Exponencial	57	0,9294	0,9937	0,9994	0,9641	0,9968	0,9997	341882	341850
Probito	Weibull	52	0,9274	0,9924	0,9995	0,9630	0,9962	0,9997	341398	341364
Probito	LogLogística	38	0,9204	0,9801	0,9954	0,9594	0,9900	0,9977	219251	219217
Probito	LogNormal	26	0,9214	0,9783	0,9970	0,9599	0,9891	0,9985	339480	339446
CLogLog	Cox	2	0,9348	0,9936	0,9990	0,9668	0,9968	0,9995	174474	174472
CLogLog	Exponencial	47	0,9293	0,9935	0,9994	0,9640	0,9967	0,9997	341885	341853
CLogLog	Weibull	46	0,9272	0,9923	0,9994	0,9629	0,9961	0,9997	341430	341396
CLogLog	LogLogística	34	0,9204	0,9805	0,9953	0,9594	0,9902	0,9977	219256	219222
CLogLog	LogNormal	30	0,9214	0,9786	0,9966	0,9599	0,9892	0,9983	339496	339462

FONTE: Os autores (2018).

A partir da Tabela 10, nota-se, para as situações em que se fez uso do modelo de Cox no componente de latência, que os modelos convergiram muito mais rápido do que os demais, além de apresentarem os maiores valores de R^2 e correlação de Pearson. Nota-se, ainda, que os valores obtidos para as funções de ligação logito e probito foram muito semelhantes. Assim, para embasar a tomada de decisão sobre qual modelo selecionar, optou-se em avaliar as estatísticas mostradas na Tabela 11 associadas aos modelos na ausência e presença das covariáveis. Com base nelas, optou-se pelo modelo logito + Cox pelo fato deste modelo ter apresentado valores

levemente menores na presença das covariáveis, quando comparado ao modelo probito + Cox, bem como por apresentar interpretações mais fáceis dos parâmetros.

TABELA 11 – Estatísticas associadas ao modelo de mistura com o modelo de Cox no componente de latência e função de ligação logito ou probito no componente de incidência

Critério	Somente Intercepto	Com Covariáveis	
		Logito	Probito
AIC	174.474,49	149.483,12	149.483,99
SC	174.484,67	149.564,50	149.565,37
-2 Log L	174.472,49	149.467,12	149.467,99

FONTE: Os autores (2018).

A Tabela 12 apresenta as estimativas dos parâmetros associado ao componente de incidência do modelo de mistura logito-Cox com fração de inadimplentes. Todas as categorias apresentam erros-padrão e valores p baixos.

TABELA 12 – Estimativas e testes associados ao componente $\pi(\boldsymbol{z})$ do modelo de mistura com fração de inadimplentes selecionado

Parâmetro	Categoria	Estimativa	Erro padrão	p-valor	IC Estimativas (95%)
Intercepto	-	-3,2674	0,0231	<0,0001	(-3,3127 : -3,221)
Quantidade restritivo de operação vencida LP ativo ou decursado	Com 2 ou 3	0,4962	0,0242	<0,0001	(0,4487 : 0,5436)
	Apenas 1	0,8866	0,0239	<0,0001	(0,8397 : 0,9335)
	Sem restritivos	2,5493	0,0255	<0,0001	(2,4993 : 2,5993)
Tempo relacionamento (em meses)	De 13 a 23	0,4343	0,0206	<0,0001	(0,3939 : 0,4746)
	Superior a 23	0,7919	0,0166	<0,0001	(0,7593 : 0,8245)
Quantidade total de restritivo ativo	De 4 a 7	0,3121	0,0182	<0,0001	(0,2764 : 0,3478)
	Sem restritivo ou até 3	0,3936	0,0174	<0,0001	(0,3594 : 0,4277)

FONTE: Os autores (2018).

Assim como no modelo de regressão logística, foram utilizadas, para facilitar a interpretação, as categorias com menor WoE como sendo a categoria de referência para cada covariável. Assim sendo, para a covariável “quantidade de restritivos de operação vencida LP ativo ou decursado” a categoria de referência foi “mais do que 3”, enquanto para a covariável “tempo de relacionamento em meses do cliente até entrar em atraso” foi “até 12 meses ou sem informação”, e para a covariável “quantidade total de restritivos ativo” foi “superior a 7 restritivos”.

A curva ROC, assim como observado no modelo logístico, apresentou sua área abaixo da curva igual a 0,747, o que indica que o componente de incidência, representado pelo logito no modelo de mistura, se ajustou bem e que apresenta um bom poder de discriminação entre clientes “bons” e “maus”.

A Tabela 13 apresenta as estimativas associadas às covariáveis no componente de latência do modelo de mistura, sendo possível constatar que as covariáveis que permaneceram foram as mesmas do componente de incidência. Além disso, é possível observar erros-padrão pequenos e efeito significativo de todas as covariáveis (ao menos uma das categorias com valor p significativo), o que sustenta a permanência da covariável no componente de latência do modelo.

TABELA 13 – Estimativas e testes associados ao componente $S(t \mid \boldsymbol{x})$ do modelo de mistura com fração de inadimplentes selecionado

Parâmetro	Categoria	Estimativa	Erro padrão	p-valor	IC Estimativas (95%)
Quantidade restritivo de operação vencida LP ativo ou decursado	Com 2 ou 3	-0,0105	0,0229	0,6467	(-0,0554 : 0,0344)
	Apenas 1	0,0350	0,0224	0,1177	(-0,0089 : 0,0789)
	Sem restritivo	0,9392	0,0225	<0,0001	(0,8951 : 0,9833)
Tempo relacionamento (em meses)	De 13 até 23	0,0527	0,0183	0,0038	(0,0168 : 0,0886)
	Superior a 23	0,0832	0,0148	<0,0001	(0,0542 : 0,1122)
Quantidade total de restritivo ativo	De 4 a 7	-0,0039	0,0158	0,8068	(-0,0349 : 0,0271)
	Sem restritivo ou até 3	0,0939	0,0148	<0,0001	(0,0649 : 0,1229)

FONTE: Os autores (2018).

No Anexo 1 podem ser visualizados os gráficos das curvas $S(t|\boldsymbol{x})$ estimadas e observadas para todos os 36 perfis de clientes presentes no estudo. Pequenas distorções para os perfis 29, 32 e 35 se devem ao baixo volume de clientes pertencentes a eles (volumetria presente na Tabela do Anexo 2). Para os demais perfis, as curvas se apresentaram bem semelhantes, constatando que o modelo ajustado foi sensível e bastante capaz de captar os mais variados tipos de clientes pertencentes no estudo.

Conforme critério já discutido para definição do modelo de mistura que melhor se ajustou aos dados (Figura 20 e Tabela 10), tem-se disponível no Anexo 2 uma tabela detalhada dos coeficientes de correlação de Pearson e dos valores de R^2 para cada um dos 36 perfis (combinação das categorias das três covariáveis presentes no modelo final).

4.4 MODELO TEMPO DE PROMOÇÃO

Para o ajuste de modelo de promoção, foi necessário reduzir o tamanho da população utilizada para modelar os dados, devido ao tempo de processamento e limitações. Para isso, foi selecionado as quatro safras de desenvolvimento (TOT) do

modelo logístico e, devido à limitação sistêmica de processamento, foi extraída uma amostra aleatória de 20 mil registros com o auxílio da função *sample* no R.

Três distribuições foram consideradas para a variável M, que corresponde ao número não observável de causas. São elas: a binomial negativa, a Bernoulli e a Poisson. Para a variável T foi considerada a distribuição de Weibull e para $\mathrm{p_0}$ o modelo logito. Assim como no modelo de mistura, as sete covariáveis (Tabela 4) foram incluídas nos modelos e, então, excluídas uma a uma as com efeito não significativo ao nível de significância de 5%. Entre todas as possibilidades, a combinação de covariáveis que se mostrou mais eficiente foi a com as mesmas que permaneceram no modelo de mistura, mostrando a força/importância dessas covariáveis aos dados modelados. São elas: quantidade de restritivos de operação vencida LP ativo ou decursado (*Var A*), tempo de relacionamento em meses do cliente até entrar em atraso (*Var B*) e quantidade total de restritivos ativo (*Var D*).

Foram consideradas as mesmas categorias de referência utilizadas no modelo de mistura, ou seja, “sem restritivos”, “até 12 meses” e “sem restritivos ou até 3” para as covariáveis “quantidade de restritivos de operação vencida LP ativo ou decursado”, “tempo de relacionamento em meses do cliente até entrar em atraso” e “quantidade total de restritivos ativo”, respectivamente.

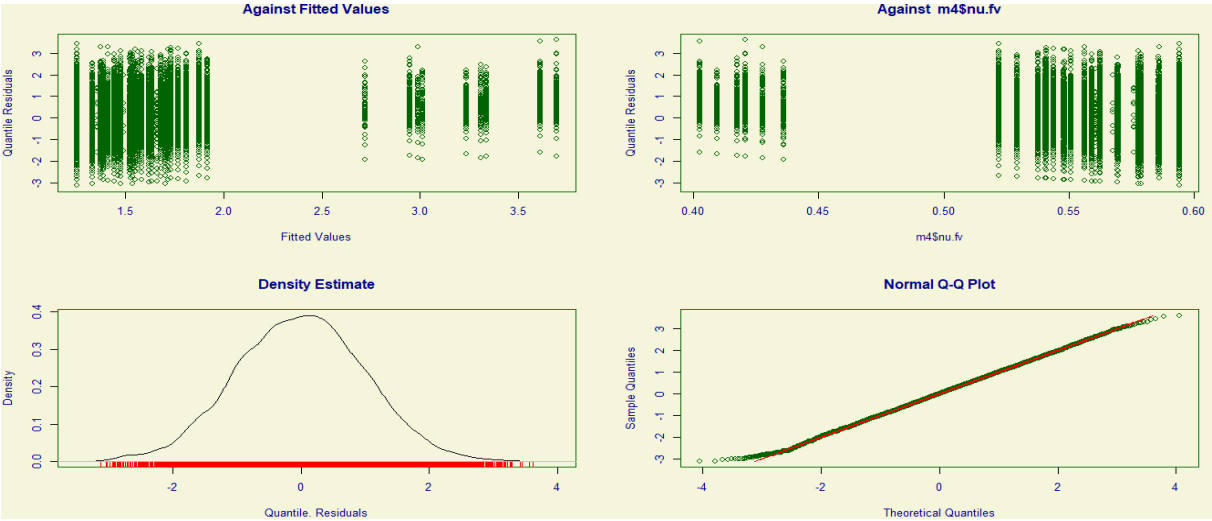
A Tabela 14 e os gráficos nas Figuras 21, 22 e 23 apresentam os resultados das análises dos resíduos quantílicos para cada umas das três distribuições. Observa-se leve vantagem da distribuição Binomial Negativa quando observados os coeficientes de assimetria (que se apresentaram mais centrado na média) e curtose (apresentaram caudas mais pesadas). O gráfico QQ-plot associado ao modelo com $M \sim BN$ também apresentou desvios mais leves nas caudas quando comparado com os QQ-plots dos modelos com as distribuições Bernoulli e Poisson, o que também é observado no coeficiente de correlação de Filliben.

TABELA 14 – Resumo dos resíduos quantílicos para os modelos com diferentes distribuições para M

Estatísticas	**Binomial Negativa**	**Bernoulli**	**Poisson**
Média	-0,006402	-0,002311	-0,005536
Variância	0,989875	0,984527	0,975064
Coeficiente de assimetria	-0,002578	0,107515	0,091087
Coeficiente de curtose	2,930777	2,734577	2,711994
Coeficiente de correlação de Filliben	0,999842	0,998506	0,998782

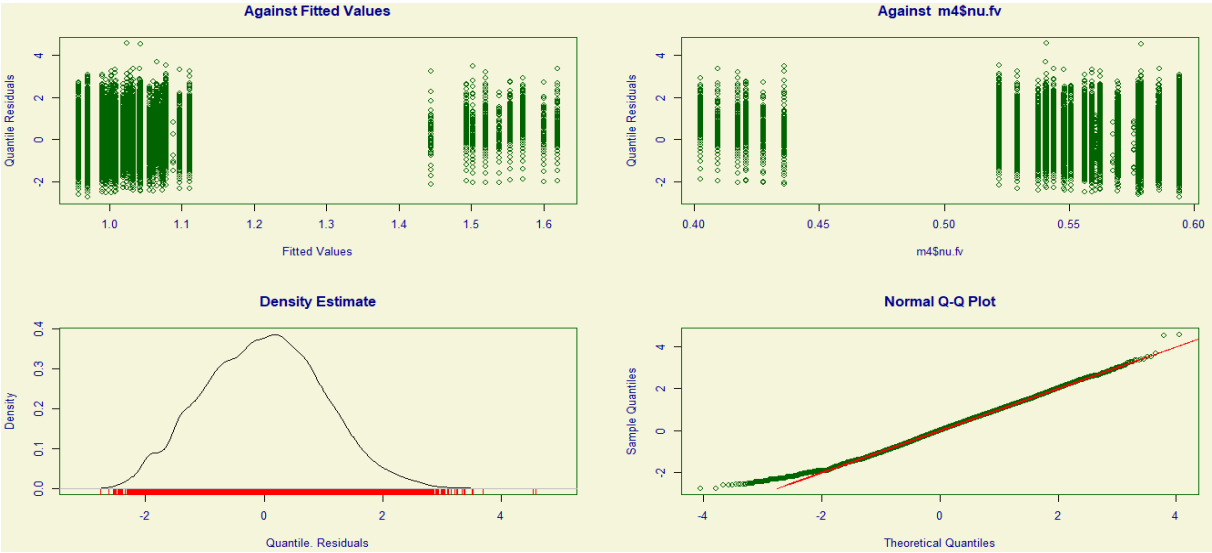
FONTE: Os autores (2018).

FIGURA 21 – Diagnóstico da qualidade de ajuste do modelo com M ~ Binomial Negativa e T~ Weibull



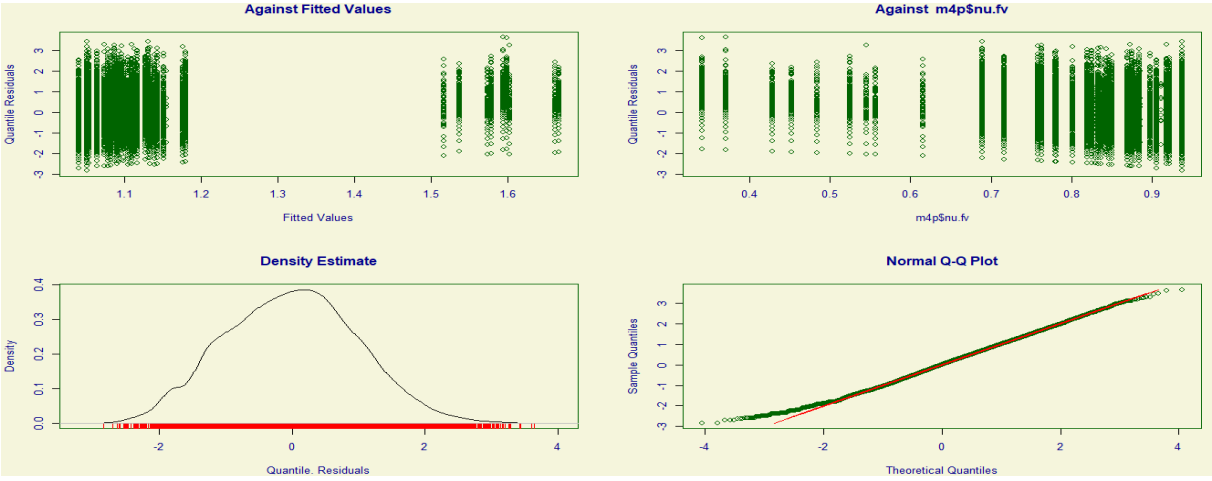
FONTE: Os autores (2018).

FIGURA 22 – Diagnóstico da qualidade de ajuste do modelo com M ~ Bernoulli e T ~ Weibull



FONTE: Os autores (2018).

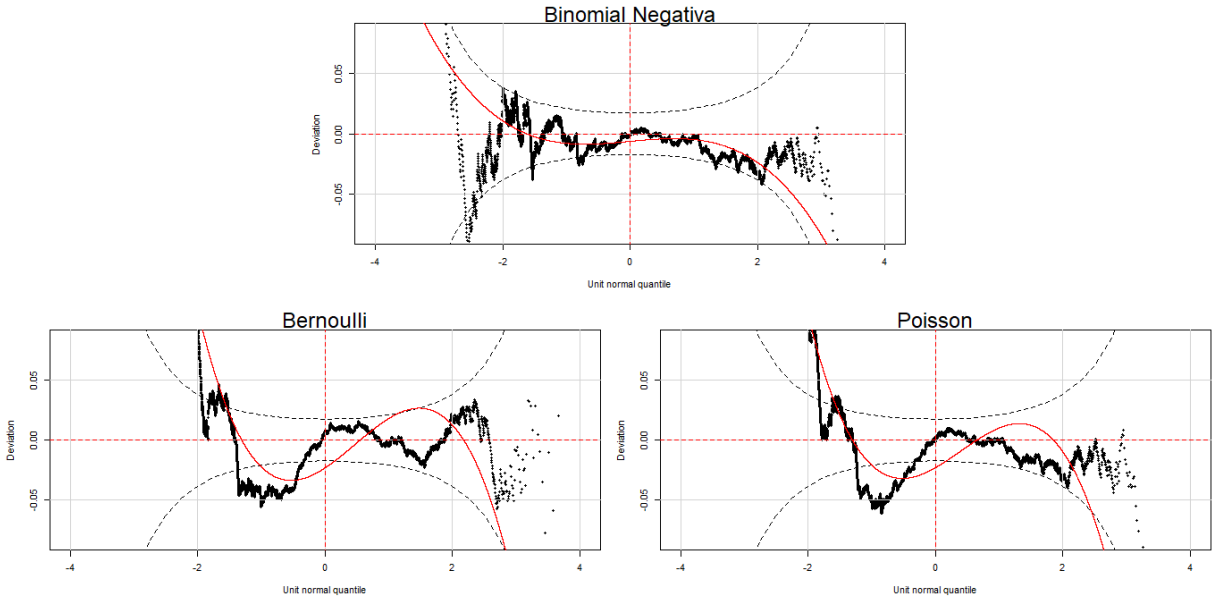
FIGURA 23 – Diagnóstico da qualidade de ajuste do modelo com M ~ Poisson e T ~ Weibull



FONTE: Os autores (2018).

Os gráficos *worm plots* apresentados na Figura 24 auxiliam na tomada de decisão da distribuição de M com melhor ajuste. Em que a linha vermelha representa uma tendência dos resíduos e os tracejados pretos o limite do intervalo de confiança (95%). É possível observar que o modelo com distribuição Binomial Negativa foi o que apresentou o maior número de pontos na região de não rejeição, uma indicação de melhor ajuste.

FIGURA 24 – *Worm plots* dos modelos com T ~ Weibull, p_0 logito e 3 diferentes distribuições para M



FONTE: Os autores (2018).

Para respaldar a tomada de decisão, são mostrados na Tabela 15 os resultados dos cálculos realizados utilizando os critérios AIC e SBC para as três distribuições discutidas. Observa-se que o modelo com $M \sim Binomial\ Negativa$ se destaca como sendo o melhor, de acordo com o AIC e SBC (menores valores).

TABELA 15 – Estatísticas associadas aos três modelos de promoção ajustados aos dados

Distribuição	**Global deviance**	**AIC**	**SBC**
Binomial Negativa-Weibull-Logito	35.695,3	35.731,3	35.873,6
Poisson-Weibull-Logito	36.305,7	36.339,7	36.474,0
Bernoulli Weibull-Logito	36.499,5	36.533,5	36.667,8

FONTE: Os autores (2018).

Dessa forma, com base nos indicadores apresentados, a decisão foi pelo modelo com p_0 logito, $T \sim Weibull$ e com M seguindo a distribuição Binomial Negativa (que representa o número de causas que pode levar a ocorrência do evento). Assim sendo, os indicadores mostrados nas Tabelas 16 e 17 e Figura 25 foram obtidos com

o total da população de desenvolvimento, facilitando os comparativos com os modelos ajustados nas etapas anteriores.

TABELA 16 – Estatísticas associadas ao modelo com M ~ Binomial Negativa considerando a população total de desenvolvimento

Distribuição	Global deviance	AIC	SBC
Binomial Negativa WEI4	338.117,7	338.153,7	338.336,8

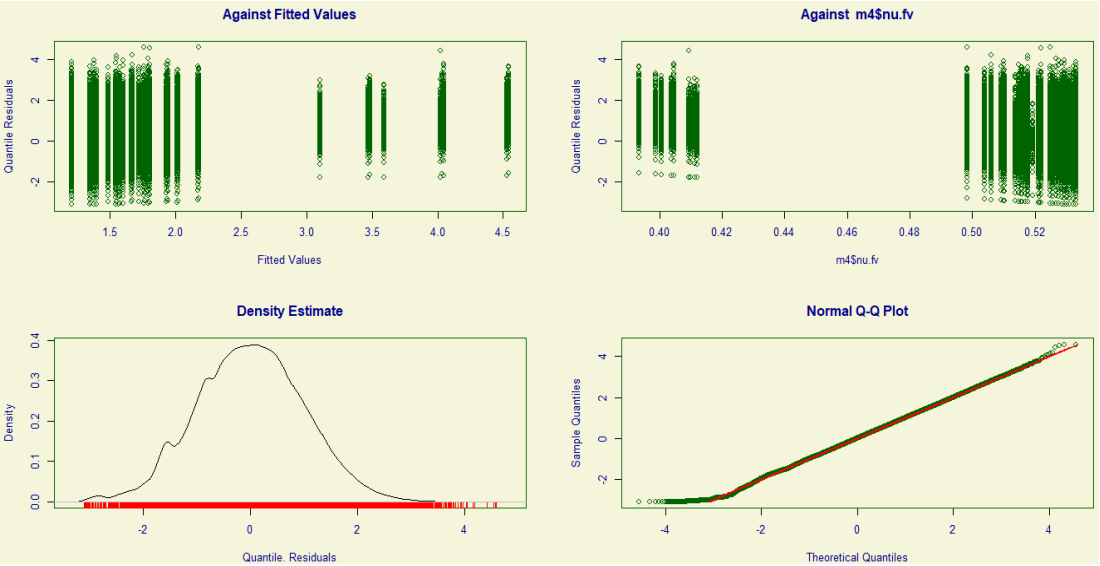
FONTE: Os autores (2018).

TABELA 17 – Estimativas e testes associados ao modelo de tempo de promoção com M ~ Binomial Negativa considerando a população total de desenvolvimento

Parâmetro	Categoria	Log			Logito		
		Estimativa	Erro Padrão	Valor p	Estimativa	Erro Padrão	Valor p
Sigma	-	-13,93350	0,00698	<0,0001	-	-	-
Tau	-	-	-	-	2,800518	0,00090	<0,0001
Intercepto	-	0,185952	0,00205	<0,0001	0,132548	0,00120	<0,0001
Quantidade restritivo LP ativo ou decursado	Igual a 1	0,135510	0,00288	<0,0001	-0,014618	0,00156	<0,0001
	De 2 ou 3	0,210209	0,00378	<0,0001	-0,061982	0,00168	<0,0001
	Mais que 3	0,946003	0,00383	<0,0001	-0,488381	0,00188	<0,0001
Tempo relacionamento (em meses)	De 13 a 23	0,146830	0,00319	<0,0001	-0,008041	0,00153	<0,0001
	Mais que 23	0,265862	0,00300	<0,0001	-0,030199	0,00126	<0,0001
Quantidade total de restritivo ativo	De 4 a 7	0,112547	0,00269	<0,0001	-0,003144	0,00141	0,0252
	Mais que 7	0,115567	0,00301	<0,0001	-0,047035	0,00147	<0,0001

FONTE: Os autores (2018).

FIGURA 25 – Diagnóstico da qualidade de ajuste do modelo com M ~ Binomial Negativa e T ~ Weibull considerando o total da população de desenvolvimento



FONTE: Os autores (2018).

No Anexo 3, podem ser visualizados os gráficos das curvas $S(t|\boldsymbol{x})$ estimadas e observadas para todos os 36 perfis de clientes presentes no estudo. Nota-se que os

perfis que apresentaram uma maior probabilidade de recuperação (bons) são os que apresentaram as maiores distorções entre as curvas estimadas e observadas.

4.5 INTERPRETAÇÃO DOS RESULTADOS

Para interpretar e comparar os resultados dos três modelos distintos que melhor se ajustaram, serão analisados os perfis que apresentaram a melhor propensão de pagamento (se tornar bom cliente) e o pior (com menor probabilidade de pagamento) ao final da janela de performance.

O perfil com melhor propensão de pagamento ao final dos 24 meses de acompanhamento foi o Perfil 02 (Anexo 2), que corresponde aos clientes que apresentaram as categorias: “sem restritivos”, “superior a 23 meses” e “sem restritivos ou até 3 restritivos” para as covariáveis “quantidade de restritivos de operação vencida LP ativo ou decursado”, “tempo de relacionamento em meses no momento do atraso” e “quantidade de restritivos ativo”, respectivamente.

Sob o modelo logístico, o Perfil 02 apresentou probabilidade de 61,48% de sanar suas dívidas ao final de 24 meses, visto que:

$$\text{logit}\big(\hat{\pi}(\boldsymbol{z_{02}})\big) = -3{,}26743 + 2{,}54935 + 0{,}79195 + 0{,}39358 = -0{,}46745,$$

e, em consequência, a probabilidade estimada para os clientes com Perfil 02 sanar suas dívidas em 24 meses foi: $\hat{\pi}(\boldsymbol{z_{02}}) = \frac{\exp(-0{,}46745)}{\exp(-0{,}46745)+1} = 0{,}6148.$

Já clientes com o Perfil 31 (Anexo 2), foram os que apresentaram a menor probabilidade de pagamento após 24 meses de acompanhamento, apenas 3,67%, pois $\text{logit}\big(\hat{\pi}(\boldsymbol{z_{31}})\big) = -3{,}26743 + 0 + 0 + 0 = -3{,}26743$ e $\hat{\pi}(\boldsymbol{z_{31}}) = \frac{\exp(-3{,}26743)}{\exp(-3{,}26743)+1} = 0{,}0367.$ São os clientes com “mais que 3 restritivos”, “até 12 meses ou sem informação” e “superior a 7 restritivos” para as covariáveis “quantidade de restritivos de operação vencida LP ativo ou decursado”, “tempo de relacionamento em meses no momento do atraso” e “quantidade de restritivos ativo”, respectivamente.

Quando analisados esses mesmos perfis sob o modelo de mistura, foram obtidas as mesmas estimativas para a probabilidade de clientes com esses perfis sanarem suas dívidas após 24 meses. Entretanto, podem também ser obtidas a partir desse modelo, as probabilidades ao longo desses 24 meses, algo que o modelo logístico não consegue mensurar.

Representando coerência nos ajustes, o modelo de mistura logito-Cox apresentou as mesmas estimativas do modelo logístico no que se refere ao componente de incidência e, portanto, as mesmas estimativas para as probabilidades de pagamento após 24 meses. Porém, o componente de latência nos fornece as probabilidades de cada perfil de clientes sanar suas dívidas durante a janela de performance. Conforme as Tabelas 18 e 19 e Figura 26 estima-se, a partir do modelo de mistura ajustado, que o Perfil 02, além de apresentar uma maior probabilidade de recuperação, apresenta também poder de recuperação muito mais rápido nos intervalos dos meses. Por exemplo, enquanto que o Perfil 02 apresenta 72% de clientes 'bons' até o 7º mês, o Perfil 31 atinge esse mesmo percentual somente no 17º mês.

Para o modelo tempo de promoção, assim como realizado para o modelo de mistura, foram estudados os mesmos perfis. Conforme as Tabelas 20 e 21 observamos uma semelhança em relação ao modelo de mistura; o Perfil 02 apresentou 69% de clientes 'bons' até o 7º mês, enquanto o Perfil 31 atingiu esse mesmo percentual somente no 17º mês.

Entretanto, quando se observa a Figura 27 é possível notar que o modelo tempo de promoção estima um percentual de recuperação muito menor para o Perfil 02 do que o modelo de mistura ajustado. Enquanto o modelo de mistura estima não recuperação de 38,5% dos clientes pertencentes ao Perfil 02 (muito próximo ao observado), o modelo tempo de promoção estima não recuperação de 61,9%, mostrando uma menor sensibilidade também para os perfis que apresentaram tendências maiores de recuperação (gráficos dos perfis no Anexo 3).

TABELA 18 – Estimativas de $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$ obtidas sob o modelo de mistura logito-Cox para os clientes com o Perfil 02

Tempo	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
$S(t\|\boldsymbol{x})$	1	0,810	0,650	0,538	0,452	0,384	0,327	0,279	0,234	0,196	0,160	0,128	0,101	0,078	0,060	0,044	0,031	0,020	0,013	0,007	0,003	0,001	0,000	0,000	0,000
$S_p(t\|\boldsymbol{x},\boldsymbol{z})$	1	0,883	0,785	0,716	0,663	0,621	0,586	0,557	0,529	0,506	0,484	0,464	0,447	0,433	0,422	0,412	0,405	0,398	0,393	0,390	0,387	0,386	0,385	0,385	0,385

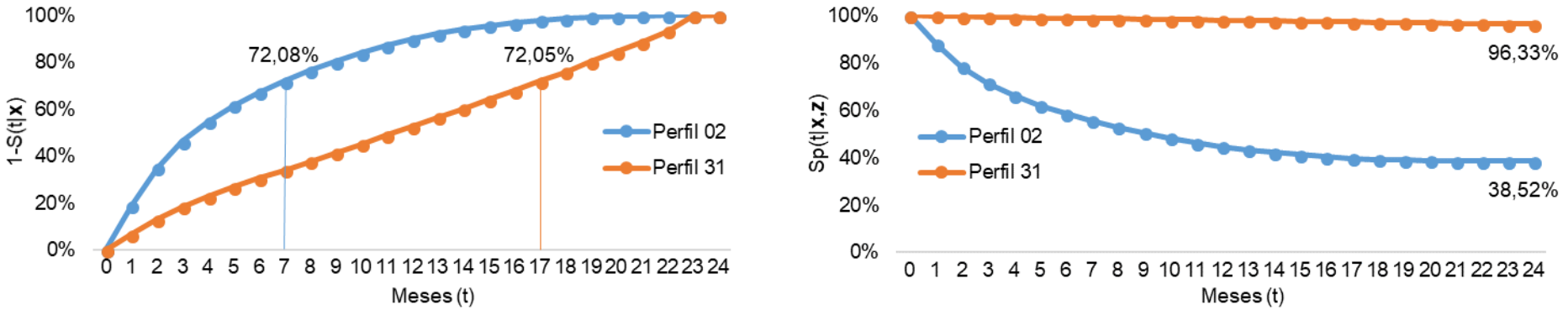
FONTE: Os autores (2018).

TABELA 19 – Estimativas de $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$ obtidas sob o modelo de mistura logito-Cox para os clientes com o Perfil 31

Tempo	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
$S(t\|\boldsymbol{x})$	1	0,933	0,868	0,816	0,771	0,731	0,694	0,659	0,621	0,587	0,549	0,510	0,472	0,434	0,398	0,359	0,322	0,280	0,241	0,198	0,156	0,114	0,066	0,002	0,000
$S_p(t\|\boldsymbol{x},\boldsymbol{z})$	1	0,998	0,995	0,993	0,992	0,990	0,989	0,987	0,986	0,985	0,983	0,982	0,981	0,979	0,978	0,976	0,975	0,974	0,972	0,971	0,969	0,967	0,966	0,963	0,963

FONTE: Os autores (2018).

FIGURA 26 – Acumulado de bons $(1 - S(t|\boldsymbol{x}))$ e estimativa de recuperação $\left(S_p(t|\boldsymbol{x},\boldsymbol{z})\right)$ dos Perfis 02 e 31 em função do tempo t, com t entre 0 e 24 meses, para o modelo de mistura logito-Cox



FONTE: Os autores (2018).

TABELA 20 – Estimativas de $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$ obtidas sob o modelo tempo de promoção p_0 logito, $T{\sim}Weibull$ e $M{\sim}Binomial\ Negativa$ para os clientes com o Perfil 02

Tempo	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
$S(t\|\boldsymbol{x})$	1	0,839	0,662	0,545	0,463	0,400	0,349	0,307	0,271	0,240	0,213	0,189	0,167	0,146	0,128	0,111	0,095	0,081	0,067	0,054	0,042	0,031	0,020	0,010	0,000
$S_p(t\|\boldsymbol{x},\boldsymbol{z})$	1	0,939	0,871	0,827	0,796	0,772	0,752	0,736	0,723	0,711	0,701	0,691	0,683	0,675	0,668	0,662	0,656	0,650	0,645	0,640	0,636	0,631	0,627	0,623	0,619

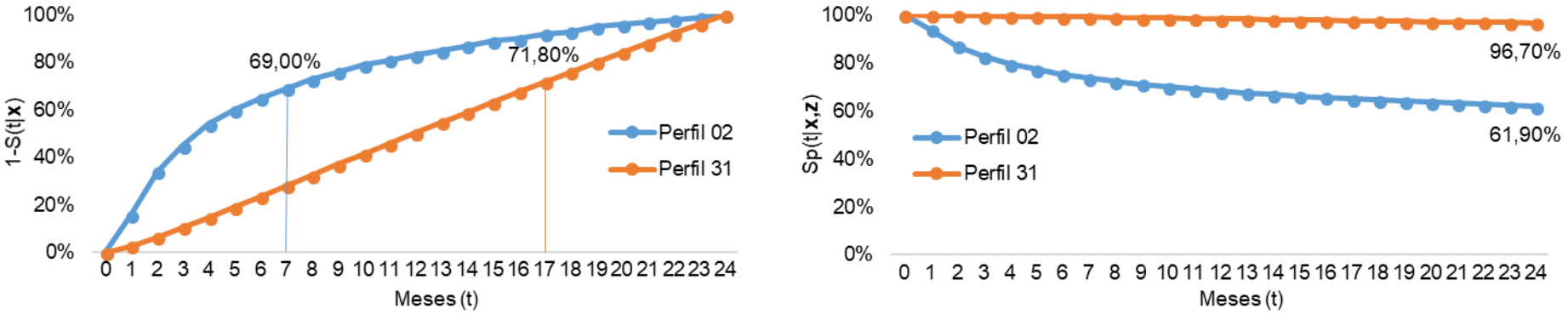
FONTE: Os autores (2018).

TABELA 21 – Estimativas de $S(t|\boldsymbol{x})$ e $S_p(t|\boldsymbol{x},\boldsymbol{z})$ obtidas sob o modelo tempo de promoção p_0 logito, $T{\sim}Weibull$ e $M{\sim}Binomial\ Negativa$ para os clientes com o Perfil 31

Tempo	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
$S(t\|\boldsymbol{x})$	1	0,971	0,934	0,894	0,852	0,808	0,764	0,720	0,675	0,630	0,586	0,541	0,497	0,453	0,410	0,367	0,324	0,282	0,240	0,199	0,158	0,118	0,078	0,039	0,000
$S_p(t\|\boldsymbol{x},\boldsymbol{z})$	1	0,999	0,998	0,996	0,995	0,994	0,992	0,991	0,989	0,988	0,986	0,985	0,983	0,982	0,980	0,979	0,978	0,976	0,975	0,974	0,972	0,971	0,970	0,968	0,967

FONTE: Os autores (2018).

FIGURA 27 – Acumulado de bons $(1 - S(t|\boldsymbol{x}))$ e estimativa de recuperação $\left(S_p(t|\boldsymbol{x},\boldsymbol{z})\right)$ dos Perfis 02 e 31 em função do tempo t, com t entre 0 e 24 meses, para o modelo tempo de promoção com p_0 logito, $T{\sim}Weibull$ e $M{\sim}Binomial\ Negativa$



FONTE: Os autores (2018).

5 CONSIDERAÇÕES FINAIS

Novas técnicas de exploração de dados, como, por exemplo *big data*, *machine learning* e análise de sobrevivência, que auxiliam na identificação dos mais variados tipos de perfis de clientes e seus comportamentos começam a ser melhor recebidas e utilizadas nas instituições financeiras. O trabalho aqui presente buscou, além de revisar o conhecimento estatístico obtido no decorrer do curso aplicados a dados reais, agregar informações com a técnica de análise de sobrevivência para a recuperação de crédito da instituição financeira que, gentilmente, forneceu o banco de dados.

A técnica de regressão atualmente utilizada pela instituição financeira foi utilizada como forma de revisão e comparação com as novas possibilidades propostas. Ela se mostrou coesa e estável ao banco de dados, com um bom ajuste. As covariáveis mais importantes e que mostraram efeito significativo foram: "q*uantidade de restritivos de operação vencida LP (lucros e perdas) ativo ou decursado no ponto de observação", "tempo de relacionamento em meses no momento do atraso",* e "q*uantidade de restritivos ativo no ponto de observação"*.

Como forma alternativa de análise dos dados, foram utilizados dois modelos no contexto de análise de sobrevivência com fração de curados: (a) o modelo de mistura, ajustado com o auxílio da macro SAS proposta por Corbiére e Joly (2007), e (b) o modelo tempo de promoção, ajustado com o auxílio do pacote *gamlss* do R proposto por Castro et al. (2010). Em ambos os casos, as covariáveis que permaneceram nos modelos foram as mesmas do modelo logístico.

Os modelos no contexto de análise de sobrevivência se mostraram muito competitivos na discriminação entre clientes bons e maus, quando comparado ao modelo logístico. A informação adicional obtida a partir dos modelos de sobrevivência é que a dimensão tempo está embutida na variável resposta. Assim, enquanto o modelo logístico, tradicionalmente utilizado nas instituições financeiras, fornece a probabilidade de pagamento (ou não) ao final da janela de performance, os modelos de sobrevivência fornecem essa mesma informação para cada intervalo de tempo observado. Ou seja, é possível tomar decisões antecipadas dependendo do tempo estimado de recuperação de um determinado perfil de clientes. Por exemplo, para um perfil de clientes com poucos dias de atraso e com longo tempo estimado para recuperação, pode-se efetuar estratégia mais agressiva, trazendo o valor presente

como desconto para uma negociação antecipada. Já para clientes com atrasos mais elevados e com longo tempo estimado para recuperação, é possível antecipar a venda desse perfil de clientes, agregando assim valor de mercado.

Neste contexto, e com os resultados obtidos a partir dos modelos ajustados, pode-se concluir que o modelo de mistura logito-Cox se mostrou eficiente para a modelagem dos dados, agregando ganho da informação “tempo”, se comparado ao modelo de regressão logística atualmente utilizado pela instituição, bem como um melhor ajuste aos perfis em relação ao modelo tempo de promoção. Tendo como população clientes inadimplentes, a principal vantagem está em poder estimar o tempo que cada perfil de clientes necessita até sanar suas dívidas, o que possibilita a tomada de decisões (cobranças) diferenciadas.

Como sugestão de trabalhos futuros, estudos de segmentações e tratamentos de modelos com variáveis abertas (sem categorizar), bem como também sazonalidade, poderão ser realizados. Além disso, por se tratar de uma população que pode navegar, com idas e vindas no mundo das adimplências e inadimplências, estudos com eventos recorrentes também podem ser úteis para modelar os perfis dos clientes em cobrança.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, Boston, v. 19, n. 6, p. 716-723, 1974.

BREIMAN, L.; FRIEDMAN, J. H; OLSHEN, R. A; STONE, C. J. **Classification and Regression Trees**. Belmont, California, Wadsworth, 1984.

CHEN, M. H.; IBRAHIM, J. G.; SINHA, D. A new Bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, v. 94, p. 909–919, 1999.

COLOSIMO, E. A; GIOLO, S. R. **Análise de sobrevivência aplicada**. São Paulo: Blucher, 2006.

CORBIÈRE, F.; JOLY, P. A SAS macro for parametric and semiparametric mixture cure models. **Computer Methods and Programs in Biomedicine**, v. 83, n. 2, p. 173-180, 2007.

COX, D. R. Regression models and life table. **Journal of the Royal Statistical Society**. Series B, v. 34, p. 187-220, 1972.

COX, D. R.; HINKLEY, D. V. **Theoretical Statistics**. Chapman & Hall, London, 1974.

DE CASTRO, M.; CANCHO, V. G.; RODRIGUES, J. A hands-on approach for fitting long-term survival models under the GALMSS framework. **Comp. Meth. and Prog in Biom.**, v. 97, n. 2, p.168-177, 2010.

DURAND, D. **Risk Elements in Consumer Installment Financing**. New York: NBER, 1941.

FILLIBEN, J. J. **The probability plot correlation coefficient test for normality**. Technometrics, v. 17, n. 1, p. 111-117, 1975.

GONÇALVES, E. B.; GOUVÊA, M. A.; MANTOVANI, D. M. N. Análise de risco de crédito com o uso de regressão logística. **Revista Contemporânea de Contabilidade,** v. 10, n. 20, p. 139-160, 2013.

HAND, D. J.; HENLEY, W. E. Statistical Classification Methods in Consumer Credit Scoring: a Review. **Journal of the Royal Statistical Society**: Série A, v.160, n. 3, p. 523–541, 1997.

HANREJSZKOW, A; STROMBERG, E. **Aplicação de regressão logística e modelos de mistura em um estudo sobre clientes inadimplentes de uma empresa de telecomunicações**. Monografia (Graduação em Estatística) - Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2013.

KAPLAN, E. L; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, v. 53, p. 457-81, 1958.

KLEIN, J. P.; MOESCHBERGER, M. L. **Survival analysis**: techniques for censored and truncated data. 2. ed. New York: Springer, 1997.

LAWLESS, J. **Statistical Models and Methods for Lifetime Data**. Jonh Wiley & Sons, New York, 1982.

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

LEE, E. T.; WANG, J. W. **Statistical Methods for Survival Data Analysis.** John Wiley & Sons, New York, 2003.

MALLER, R.; ZHOU. X. **Survival Analysis with Long-Term Survivors**. Wiley, New York, 1996.

MIOLA, R. F. **Uso de modelos estatísticos para de escore de crédito de uma instituição financeira**. 2013. Dissertação (Mestrado em Engenharia de Produção), Faculdade de Engenharia da UNESP, São Paulo, 2013.

QUIDIM, I. L. **Análise de sobrevivência com fração de fidelizados**: uma aplicação na área de marketing. Dissertação (Mestrado em Estatística), IME - Instituto de Matemática e Estatística, Universidade de São Paulo, 2005.

R CORE TEAM.: A language and environment for statistical computing. Viena, Austria, 2015. Disponível em: <http://www.R-project.org/>.

SAS/STAT© Software: Enterprise Guide, 7.1 Copyright, SAS Institute Inc. Cary, NC, USA, 2016.

SCHWARZ, G. **Estimating the dimensional of a model.** Annals of Statistics, Hayward, v.6, n.2, p.461-464, Mar. 1978.

SOUZA, R. B. **O modelo de Collection scoring como ferramenta para a gestão estratégica do risco de crédito**. Dissertação (Mestrado em Administração), Fundação Getúlio Vargas – FGV, São Paulo, 2000.

SPC Brasil. **Indicador de inadimplência de Pessoas Jurídicas SPC Brasil e CNDL.** Janeiro, 2018. Disponível em:< https://webcache.googleusercontent.com/search?q=cache:JTYc6Ci1NxEJ:https://www.spcbrasil.org.br/wpimprensa/wp-content/uploads/2018/02/An%25C3%25A1lise-PJ_janeiro_2018.pdf+&cd=2&hl=ptBR&ct=clnk&gl=br>. Acesso em: 15/03/2018.

THOMAS, L.C.; EDELMAN, D. B.; CROOK, J. N. **Credit Scoring and Its Applications**. Siam: Philadelphia, 2002.

TONEGI, L. **Modelo com fração de inadimplentes: uma aplicação a dados financeiros**. Monografia (Graduação em Estatística) - Setor de Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2017.

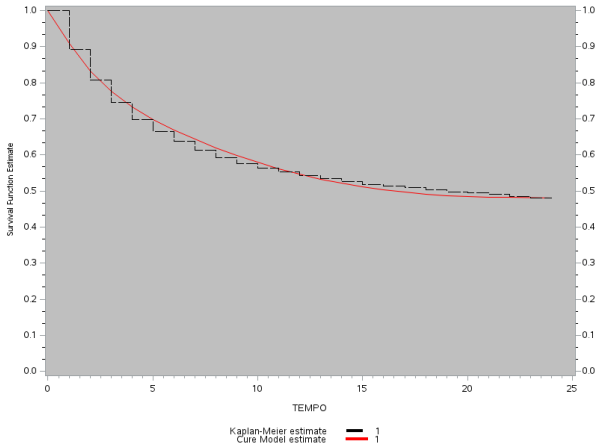
WALD, A. Tests of Statistical Hypotheses concerning Several Parameters when the number of Observations is Large, **Trans. Amer. Math. Soc**., v. 54, p. 426-482, 1943.

YAKOVLEV, A.; TSODIKOV, A. D. **Stochastic Models of Tumor Latency and their Biostatistical Application**. 1ST Edition, World Scientific, Singapore, 1996.

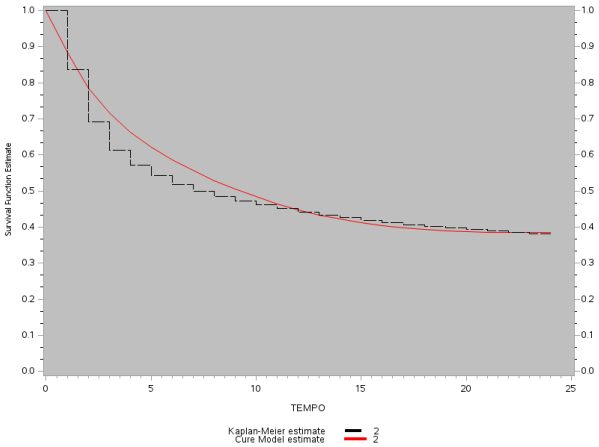
YAKOVLEV, A.; ASSELAIN, B.; BARDOU, V.; FOURQUET, A.; HOANG, T.; ROCHEFEDIERE, A.; TSODIKOV, A. A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. **Biometric et Analyse de Donnes Spatio-Temporelles**, v. 12, p. 67-82, 1993.

ANEXO 1 – CURVAS OBSERVADAS E ESTIMADAS A PARTIR DO MODELO DE MISTURA PARA TODOS OS PERFIS DE CLIENTES

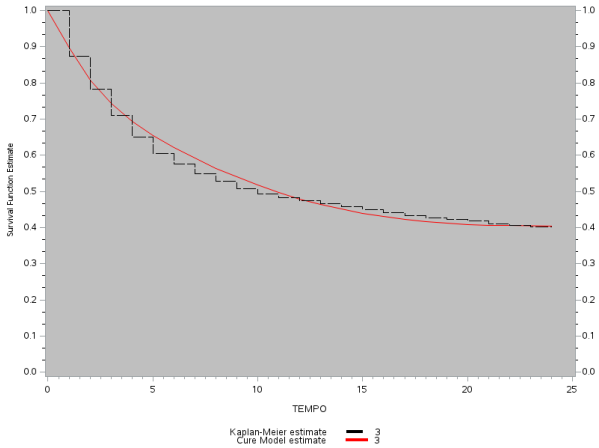
Perfil 01 (Anexo 2)



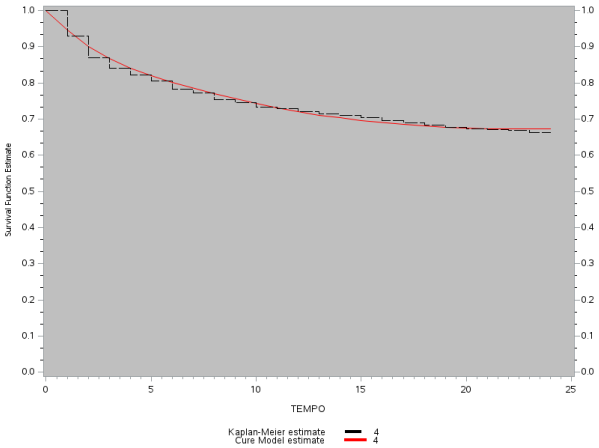
Perfil 02 (Anexo 2)



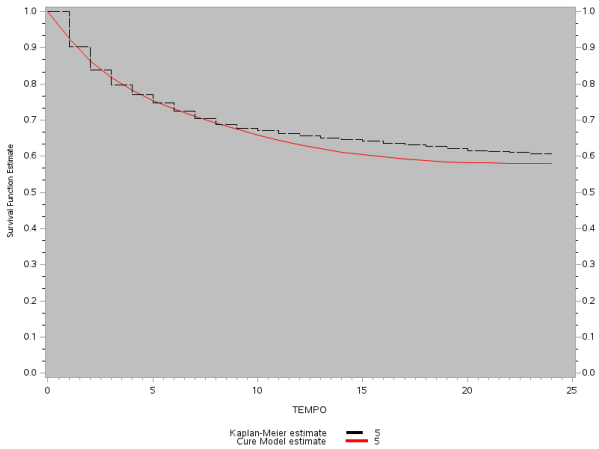
Perfil 03 (Anexo 2)



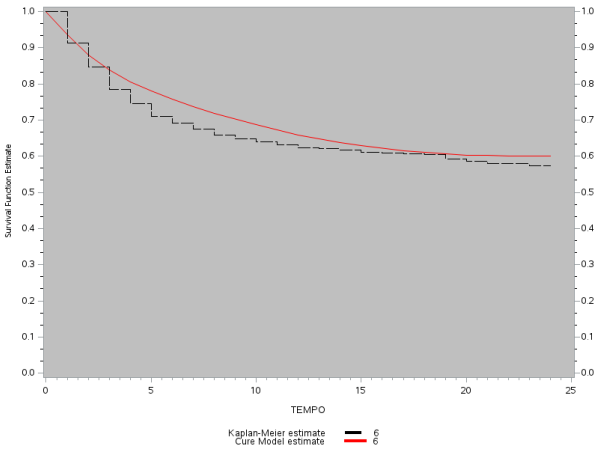
Perfil 04 (Anexo 2)



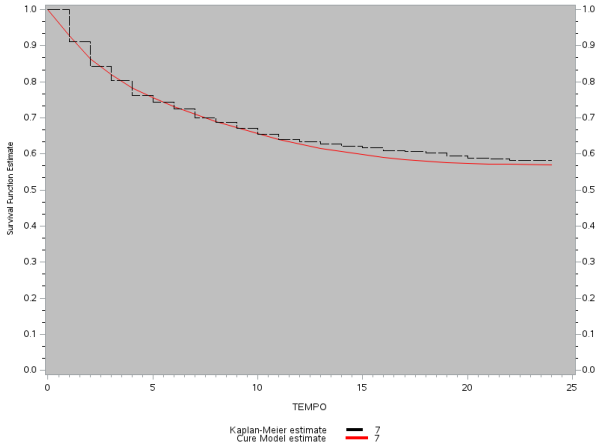
Perfil 05 (Anexo 2)



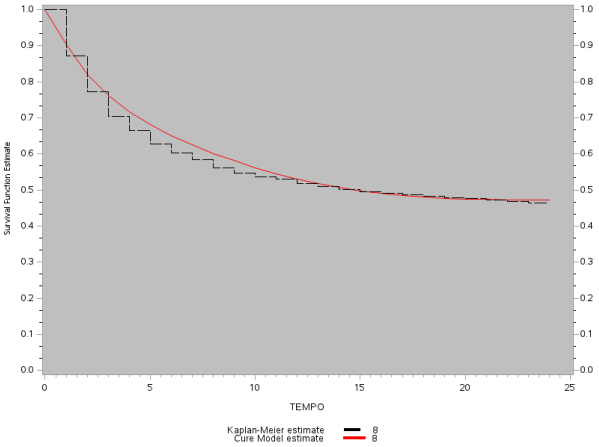
Perfil 06 (Anexo 2)



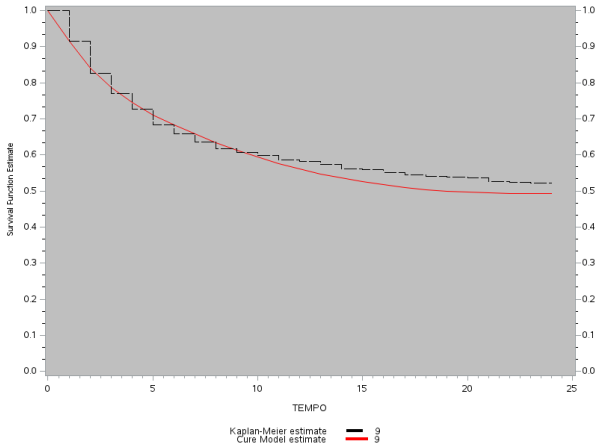
Perfil 07 (Anexo 2)



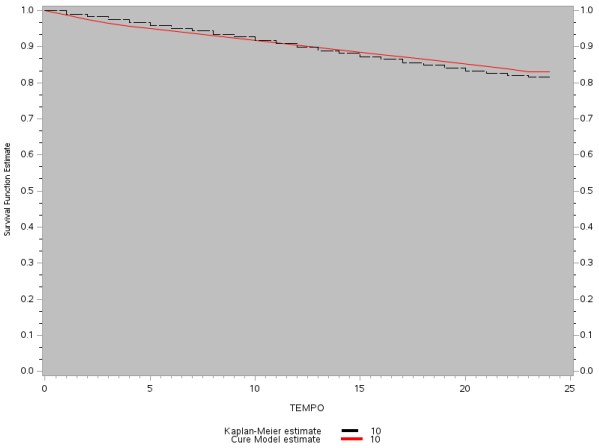
Perfil 08 (Anexo 2)



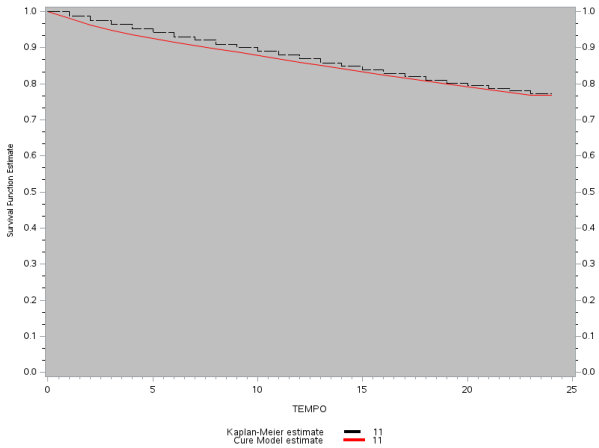
Perfil 09 (Anexo 2)



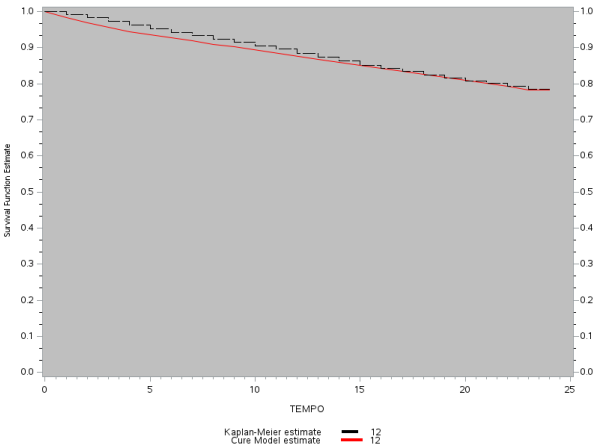
Perfil 10 (Anexo 2)



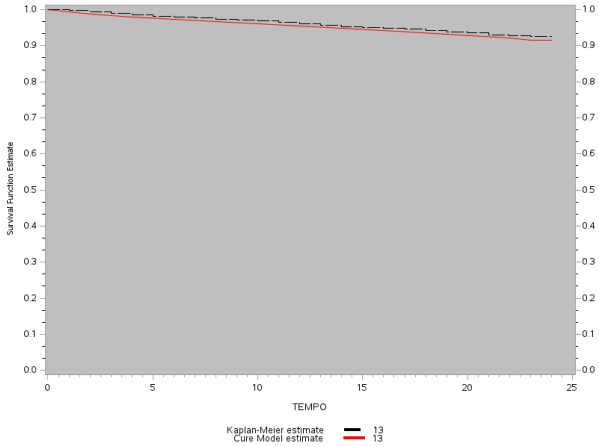
Perfil 11 (Anexo 2)



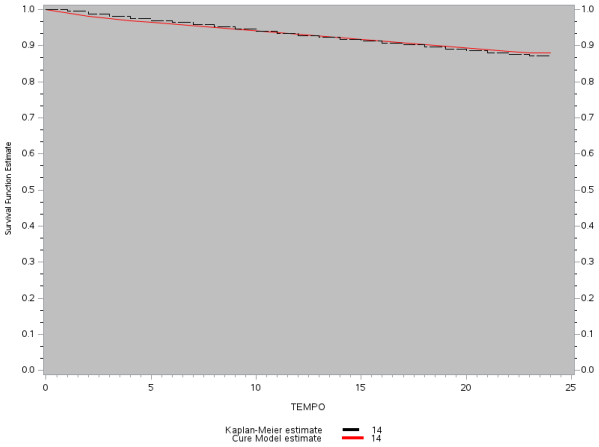
Perfil 12 (Anexo 2)



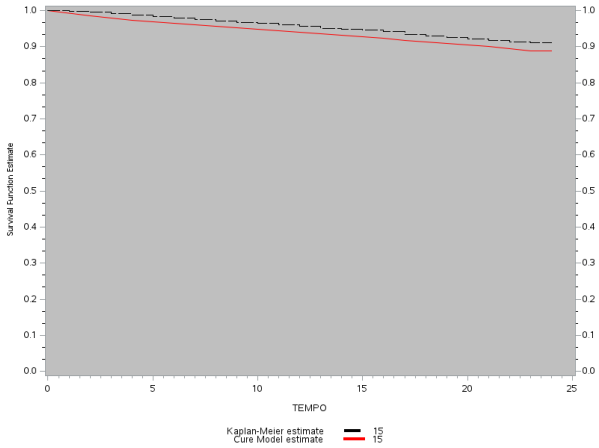
Perfil 13 (Anexo 2)



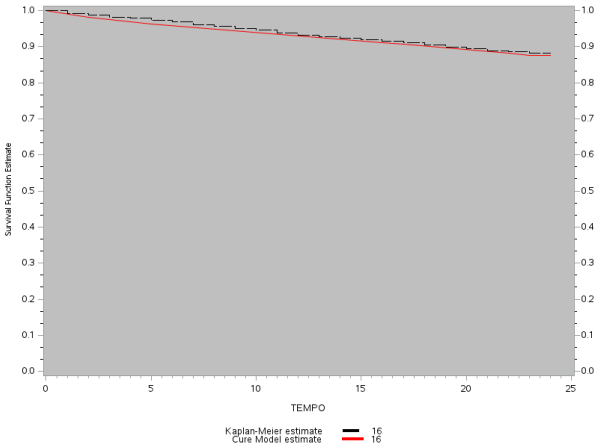
Perfil 14 (Anexo 2)



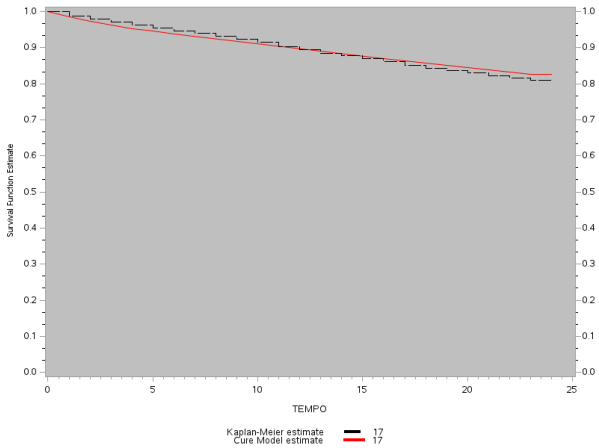
Perfil 15 (Anexo 2)



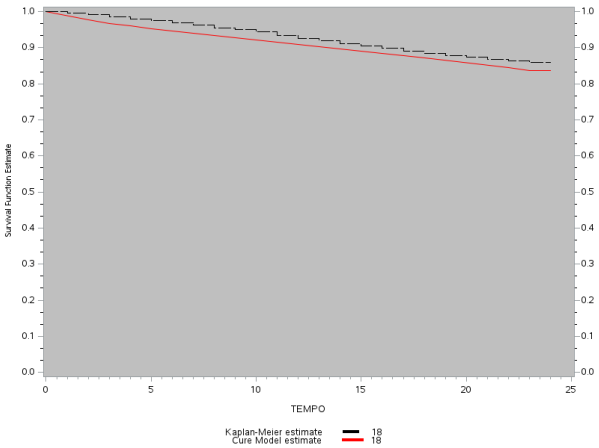
Perfil 16 (Anexo 2)



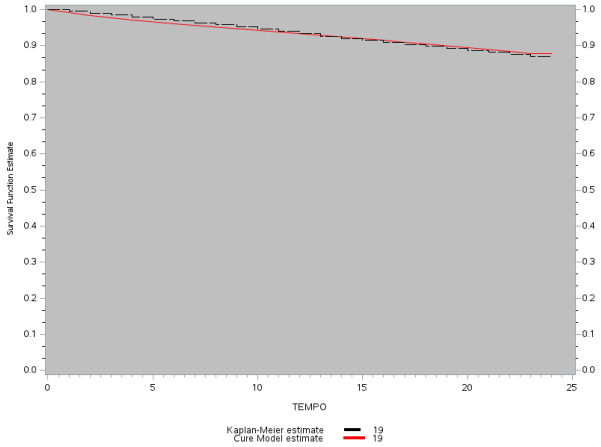
Perfil 17 (Anexo 2)



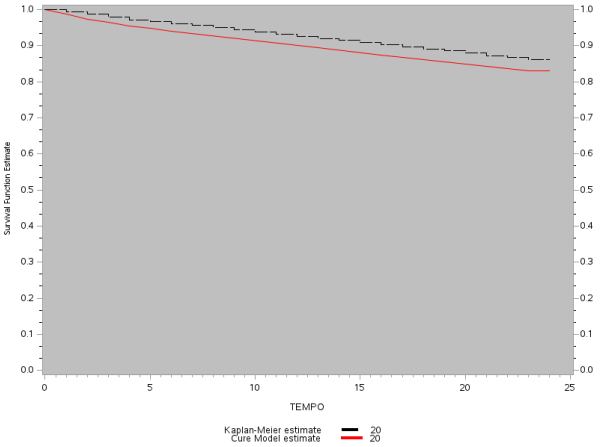
Perfil 18 (Anexo 2)



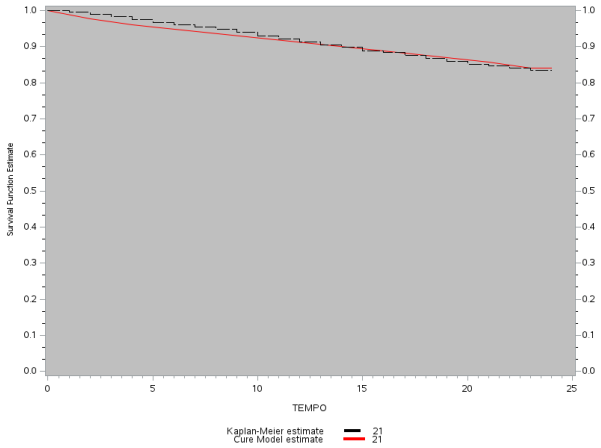
Perfil 19 (Anexo 2)



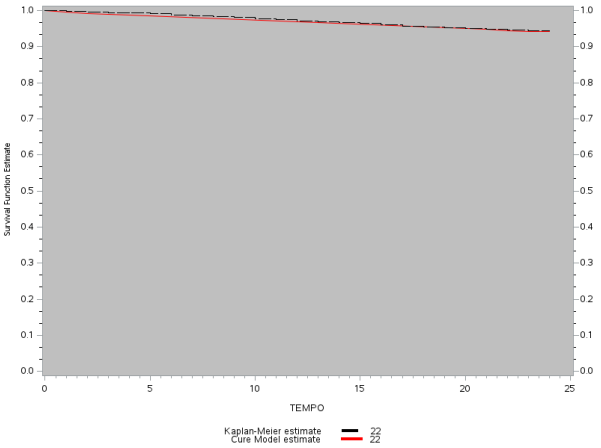
Perfil 20 (Anexo 2)



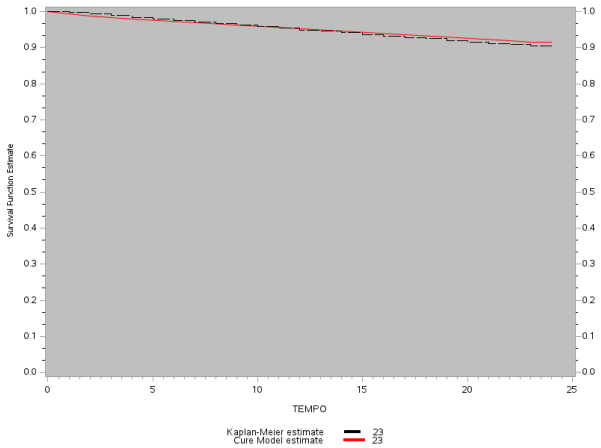
Perfil 21 (Anexo 2)



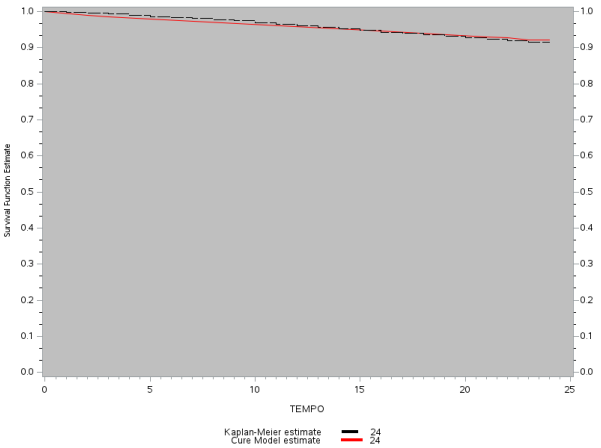
Perfil 22 (Anexo 2)



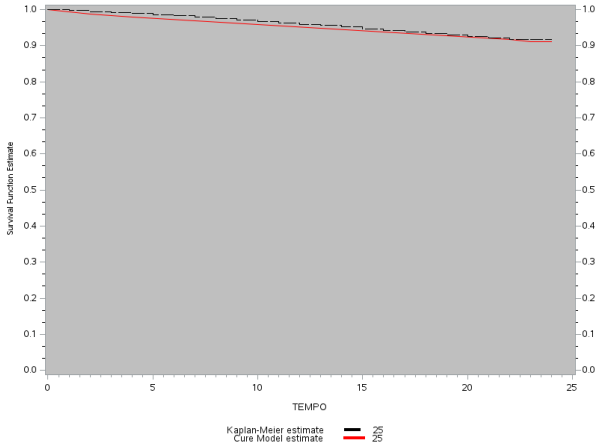
Perfil 23 (Anexo 2)



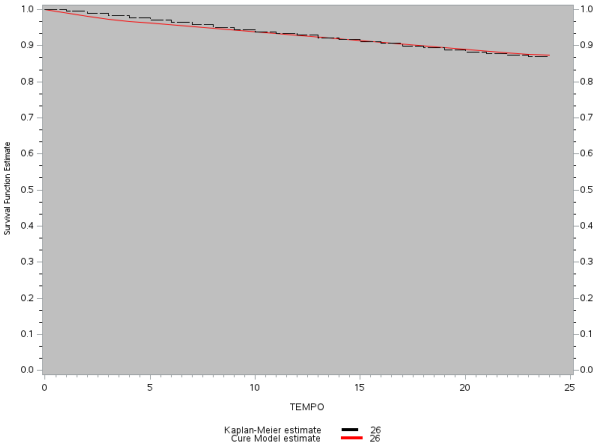
Perfil 24 (Anexo 2)



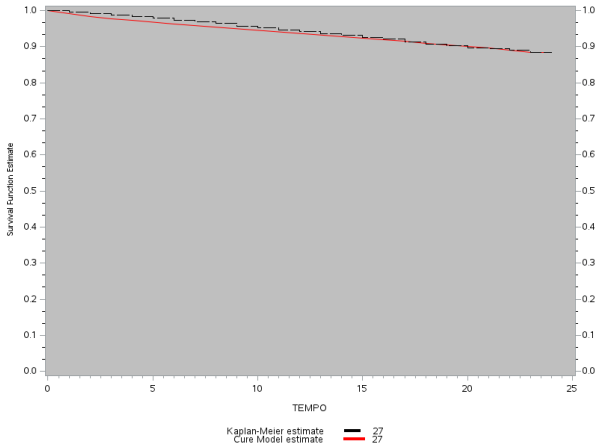
Perfil 25 (Anexo 2)



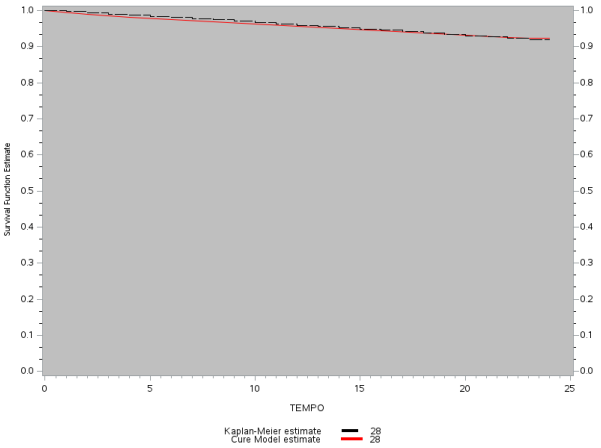
Perfil 26 (Anexo 2)



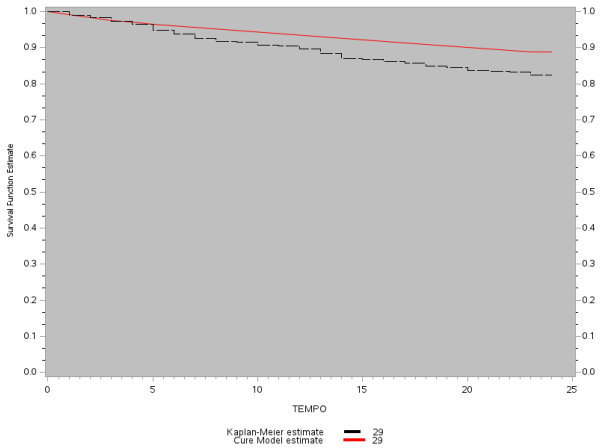
Perfil 27 (Anexo 2)



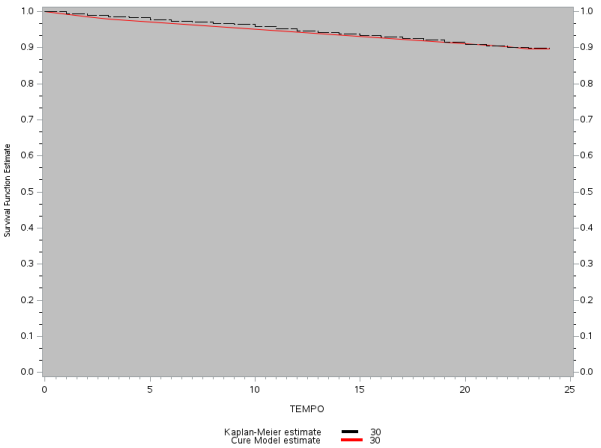
Perfil 28 (Anexo 2)



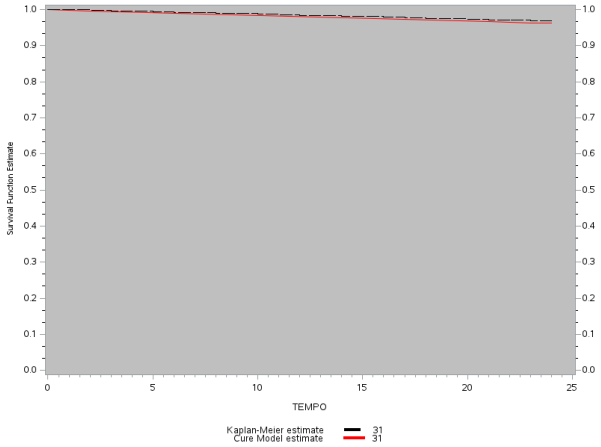
Perfil 29 (Anexo 2)



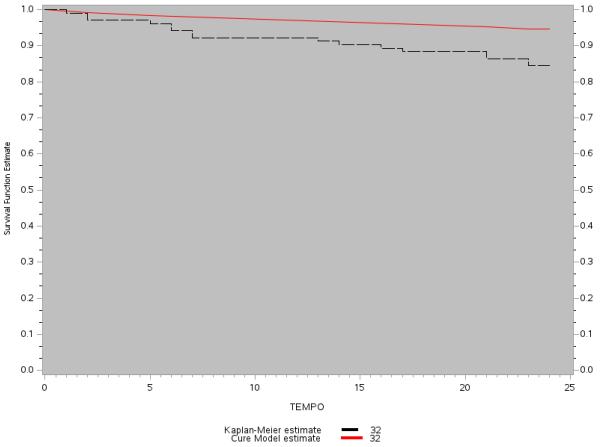
Perfil 30 (Anexo 2)



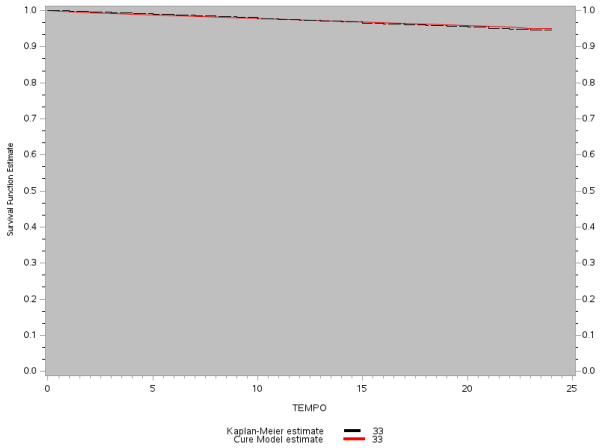
Perfil 31 (Anexo 2)



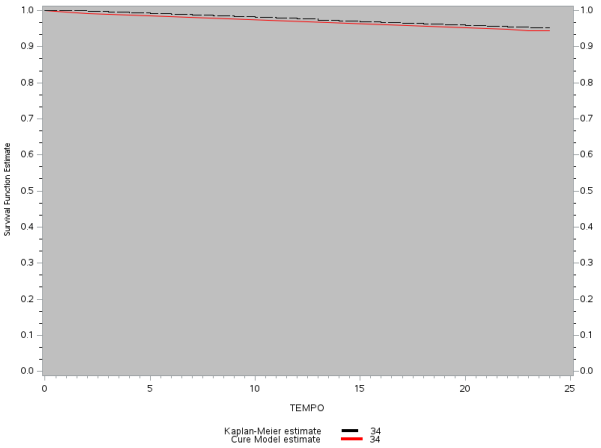
Perfil 32 (Anexo 2)



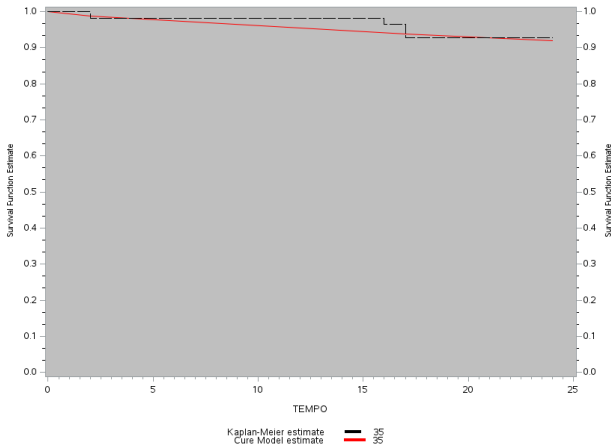
Perfil 33 (Anexo 2)



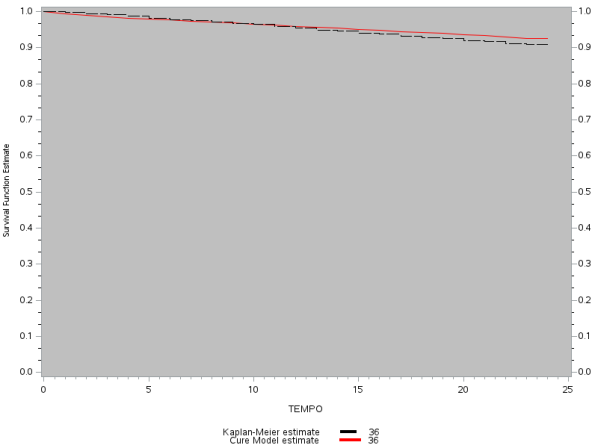
Perfil 34 (Anexo 2)



Perfil 35 (Anexo 2)



Perfil 36 (Anexo 2)

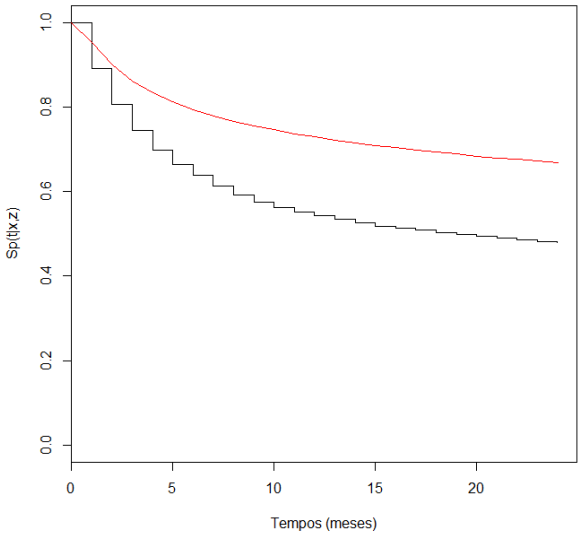


ANEXO 2 – ESTIMATIVAS OBTIDAS VIA O MODELO DE MISTURA LOGITO-COX PARA OS PERFIS DE CLIENTES ESTUDADOS

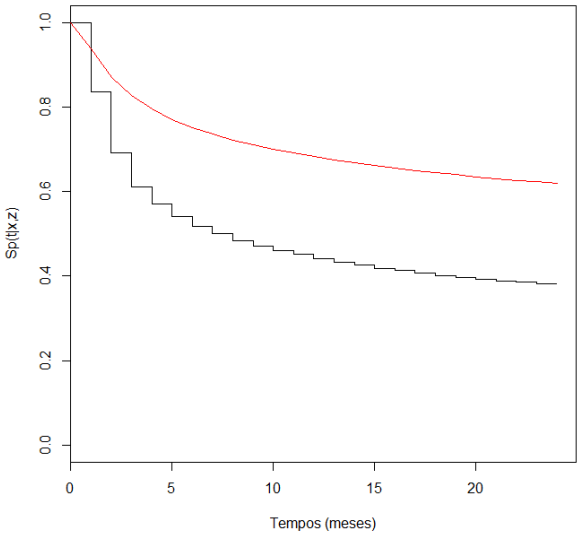
Perfil	Quantidade de Restritivos de Operação Vencida LP Ativo ou Decursado	Tempo de relacionamento em meses no momento do atraso	Quantidade de restritivos Ativo	Volume	Acúmulos de futuros bons em relação ao tempo final (24 meses)				Taxa de inadimplentes ao longo do tempo				%Bons ao final da Janela de Performance	Tempo até obter % do total de Bons		
					t = 6	t = 12	t = 18	t = 24	t = 6	t = 12	t = 18	t = 24		25%	50%	75%
01	Sem restritivo	Superior a 23 meses	Superior a 7 restritivos	3.083	36%	12%	2%	0%	67%	55%	49%	48%	52%	2	4	9
02	Sem restritivo	Superior a 23 meses	Sem restritivos ou até 3	7.941	33%	10%	1%	0%	59%	45%	39%	39%	61%	2	4	8
03	Sem restritivo	Superior a 23 meses	De 4 a 7 restritivos	2.921	36%	12%	2%	0%	62%	48%	42%	40%	60%	2	4	9
04	Sem restritivo	Até 12 meses ou sem informação	Superior a 7 restritivos	642	39%	15%	3%	0%	80%	72%	68%	67%	33%	2	5	10
05	Sem restritivo	Até 12 meses ou sem informação	Sem restritivos ou até 3	2.261	36%	12%	2%	0%	73%	63%	59%	58%	42%	2	4	9
06	Sem restritivo	Até 12 meses ou sem informação	De 4 a 7 restritivos	1.025	39%	15%	3%	0%	76%	66%	61%	60%	40%	2	5	10
07	Sem restritivo	De 13 até 23 meses	Superior a 7 restritivos	714	37%	13%	2%	0%	73%	63%	58%	57%	43%	2	4	9
08	Sem restritivo	De 13 até 23 meses	Sem restritivos ou até 3	2.033	34%	11%	1%	0%	65%	53%	48%	47%	53%	2	4	8
09	Sem restritivo	De 13 até 23 meses	De 4 a 7 restritivos	883	37%	13%	2%	0%	68%	56%	50%	49%	51%	2	4	9
10	Apenas 1	Superior a 23 meses	Superior a 7 restritivos	8.412	66%	43%	20%	0%	94%	90%	86%	83%	17%	4	11	17
11	Apenas 1	Superior a 23 meses	Sem restritivos ou até 3	17.229	64%	40%	17%	0%	92%	86%	81%	77%	23%	4	10	16
12	Apenas 1	Superior a 23 meses	De 4 a 7 restritivos	7.448	66%	43%	20%	0%	93%	88%	83%	78%	22%	4	11	17
13	Apenas 1	Até 12 meses ou sem informação	Superior a 7 restritivos	2.948	68%	46%	23%	0%	97%	95%	93%	92%	8%	5	11	18
14	Apenas 1	Até 12 meses ou sem informação	Sem restritivos ou até 3	9.421	66%	43%	20%	0%	96%	93%	90%	88%	12%	4	11	17
15	Apenas 1	Até 12 meses ou sem informação	De 4 a 7 restritivos	4.296	69%	46%	23%	0%	96%	94%	91%	89%	11%	5	11	18
16	Apenas 1	De 13 até 23 meses	Superior a 7 restritivos	2.797	67%	44%	21%	0%	96%	93%	90%	88%	12%	5	11	17
17	Apenas 1	De 13 até 23 meses	Sem restritivos ou até 3	7.265	64%	41%	18%	0%	94%	90%	86%	83%	17%	4	10	17
18	Apenas 1	De 13 até 23 meses	De 4 a 7 restritivos	3.118	67%	44%	21%	0%	95%	91%	87%	84%	16%	5	11	17
19	Com 2 ou 3	Superior a 23 meses	Superior a 7 restritivos	8.648	67%	45%	22%	0%	96%	93%	90%	88%	12%	5	11	18
20	Com 2 ou 3	Superior a 23 meses	Sem restritivos ou até 3	7.788	65%	41%	19%	0%	94%	90%	86%	83%	17%	4	10	17
21	Com 2 ou 3	Superior a 23 meses	De 4 a 7 restritivos	7.612	68%	45%	22%	0%	95%	91%	88%	84%	16%	5	11	18
22	Com 2 ou 3	Até 12 meses ou sem informação	Superior a 7 restritivos	6.938	70%	48%	24%	0%	98%	97%	96%	94%	6%	5	12	18
23	Com 2 ou 3	Até 12 meses ou sem informação	Sem restritivos ou até 3	8.018	67%	44%	21%	0%	97%	95%	93%	92%	8%	5	11	18
24	Com 2 ou 3	Até 12 meses ou sem informação	De 4 a 7 restritivos	9.389	70%	48%	25%	0%	98%	96%	94%	92%	8%	5	12	18
25	Com 2 ou 3	De 13 até 23 meses	Superiores a 7 restritivos	4.004	68%	46%	23%	0%	97%	95%	93%	91%	9%	5	11	18
26	Com 2 ou 3	De 13 até 23 meses	Sem restritivos ou até 3	3.950	66%	42%	20%	0%	96%	93%	90%	87%	13%	4	11	17
27	Com 2 ou 3	De 13 até 23 meses	De 4 a 7 restritivos	4.451	68%	46%	23%	0%	96%	94%	91%	88%	12%	5	11	18
28	Mais do que 3	Superior a 23 meses	Superiores a 7 restritivos	11.793	67%	44%	21%	0%	97%	96%	94%	92%	8%	5	11	18
29	Mais do que 3	Superior a 23 meses	Sem restritivos ou até 3	455	65%	41%	18%	0%	96%	93%	91%	89%	11%	4	10	17
30	Mais do que 3	Superior a 23 meses	De 4 a 7 restritivos	4.222	67%	44%	21%	0%	97%	94%	92%	90%	10%	5	11	18
31	Mais do que 3	Até 12 meses ou sem informação	Superiores a 7 restritivos	16.258	69%	47%	24%	0%	99%	98%	97%	96%	4%	5	12	18
32	Mais do que 3	Até 12 meses ou sem informação	Sem restritivos ou até 3	103	67%	63%	25%	0%	98%	98%	96%	95%	5%	5	13	17
33	Mais do que 3	Até 12 meses ou sem informação	De 4 a 7 restritivos	6.458	69%	47%	24%	0%	98%	97%	96%	95%	5%	5	12	18
34	Mais do que 3	De 13 até 23 meses	Superiores a 7 restritivos	6.393	68%	45%	22%	0%	98%	97%	96%	94%	6%	5	11	18
35	Mais do que 3	De 13 até 23 meses	Sem restritivos ou até 3	56	85%	85%	23%	0%	99%	99%	94%	92%	8%	16	16	17
36	Mais do que 3	De 13 até 23 meses	De 4 a 7 restritivos	2.375	68%	45%	22%	0%	98%	96%	94%	93%	7%	5	11	18

ANEXO 3 – CURVAS OBSERVADAS E ESTIMADAS A PARTIR DO MODELO TEMPO DE PROMOÇÃO PARA TODOS OS PERFIS DE CLIENTES

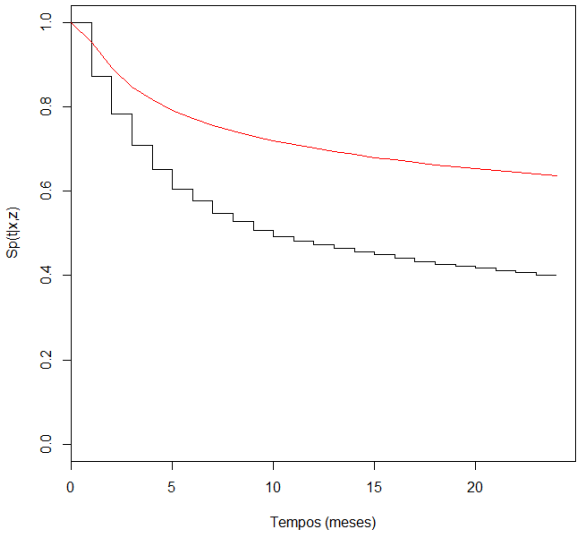
Perfil 01 (Anexo 2)



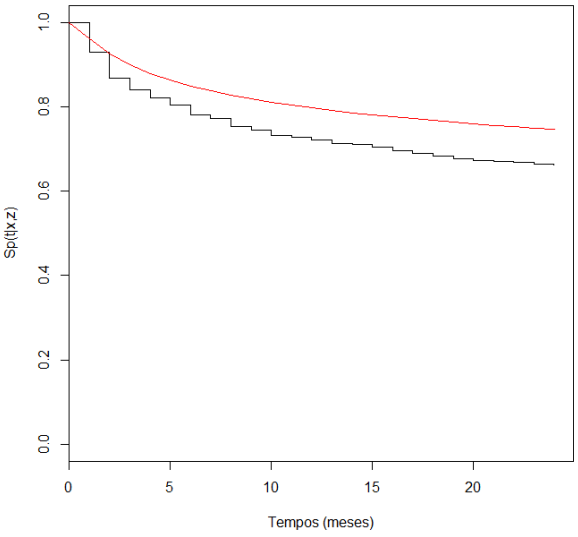
Perfil 02 (Anexo 2)



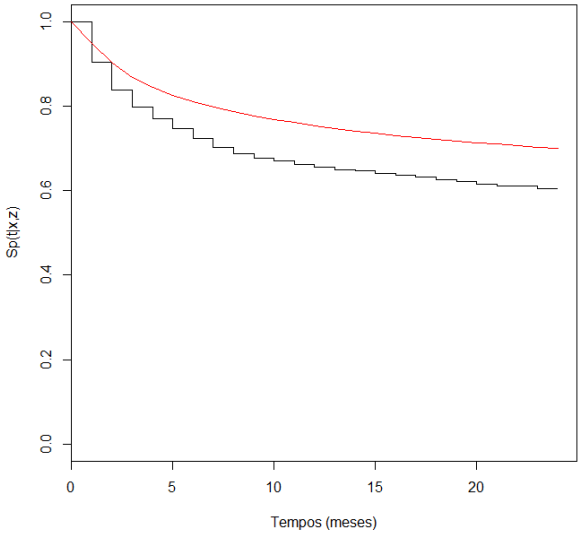
Perfil 03 (Anexo 2)



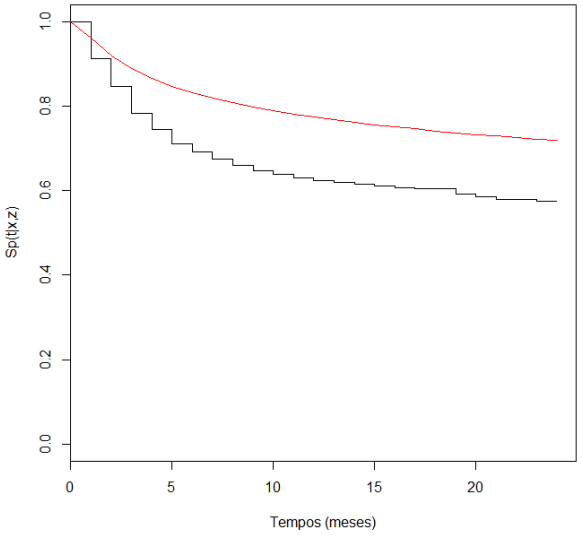
Perfil 04 (Anexo 2)



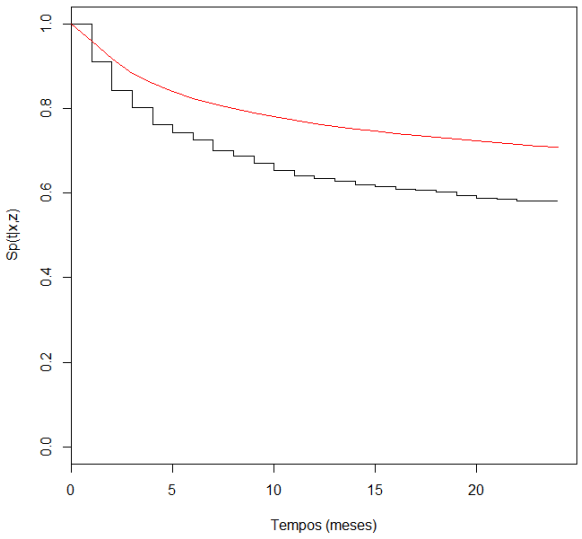
Perfil 05 (Anexo 2)



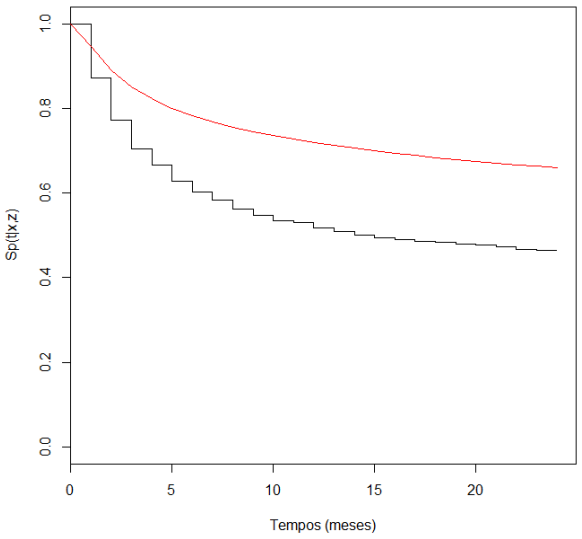
Perfil 06 (Anexo 2)



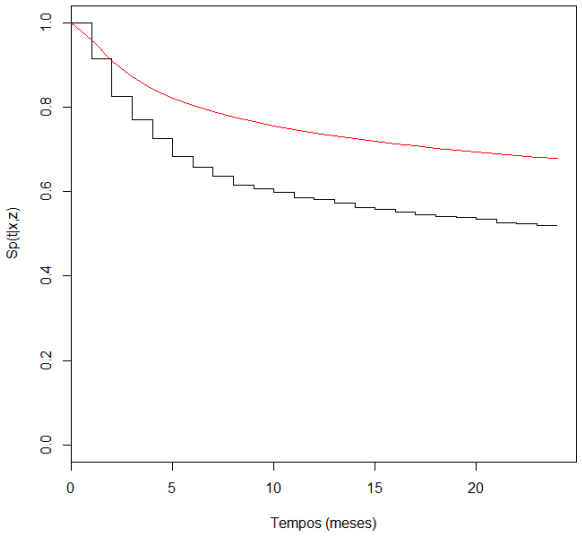
Perfil 07 (Anexo 2)



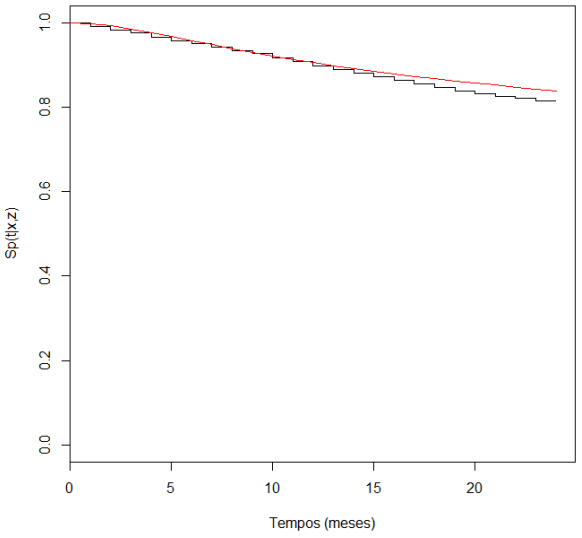
Perfil 08 (Anexo 2)



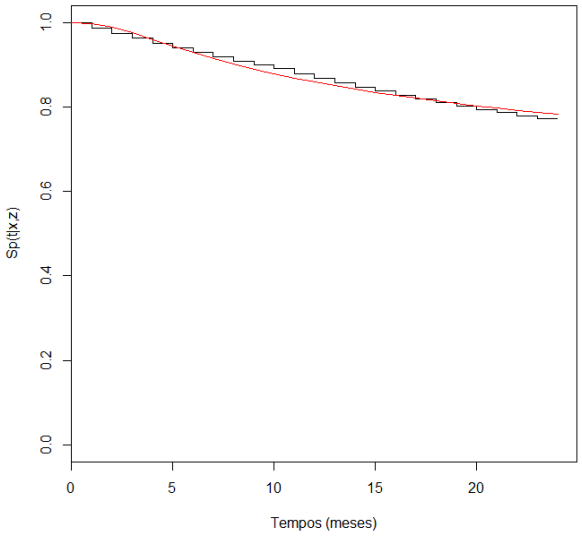
Perfil 09 (Anexo 2)



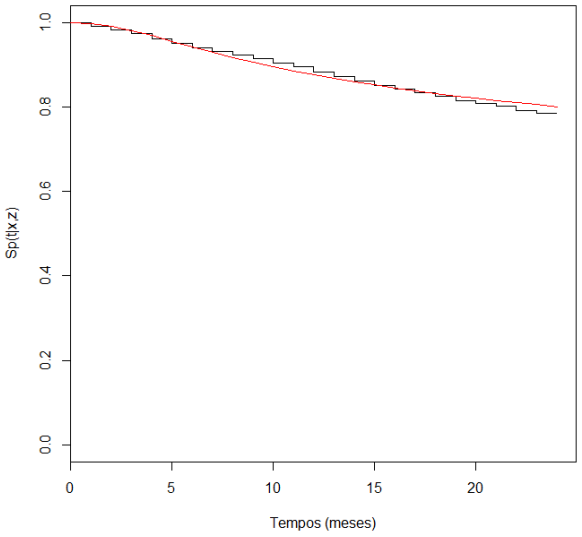
Perfil 10 (Anexo 2)



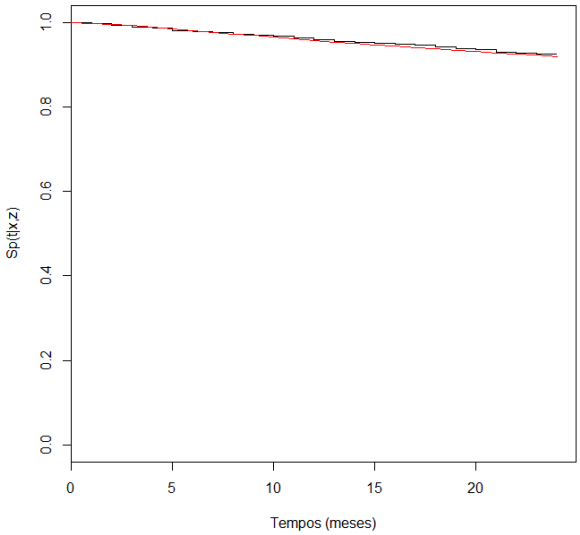
Perfil 11 (Anexo 2)



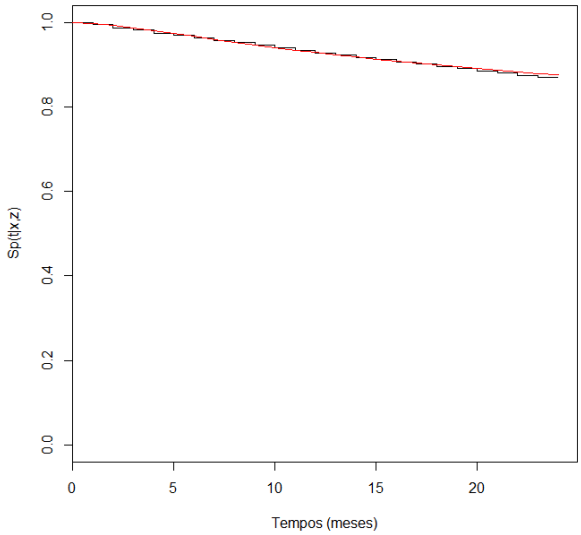
Perfil 12 (Anexo 2)



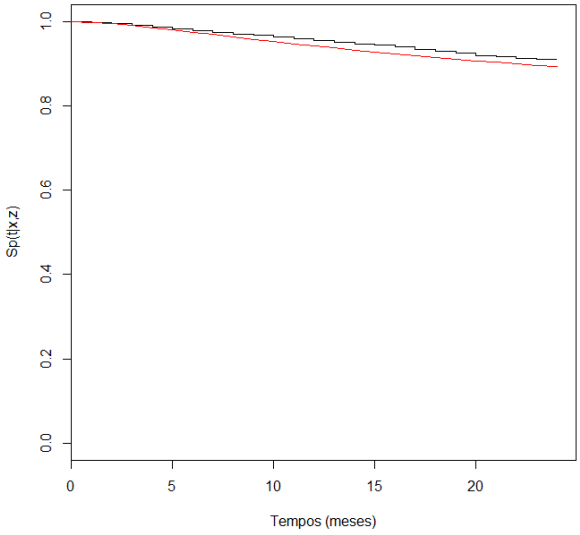
Perfil 13 (Anexo 2)



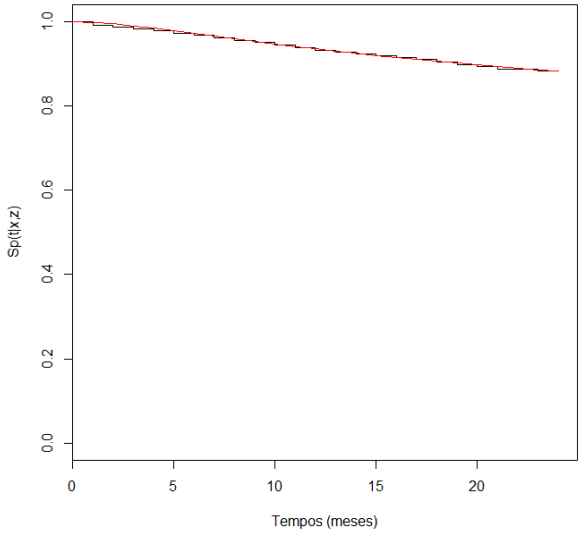
Perfil 14 (Anexo 2)



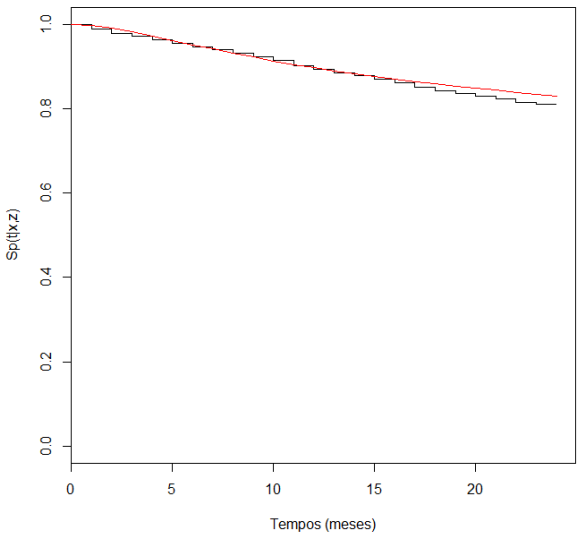
Perfil 15 (Anexo 2)



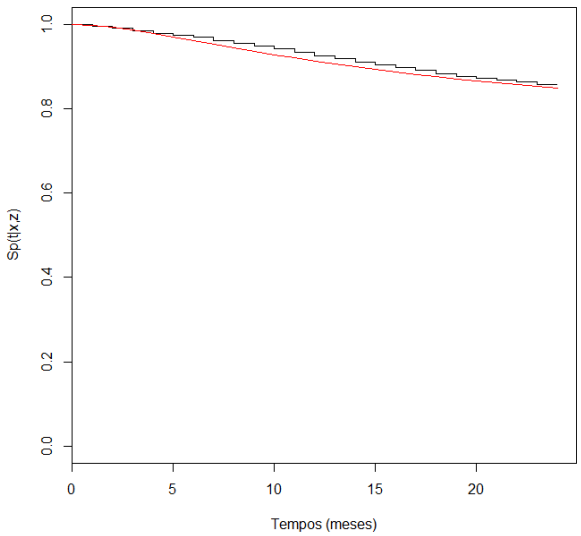
Perfil 16 (Anexo 2)

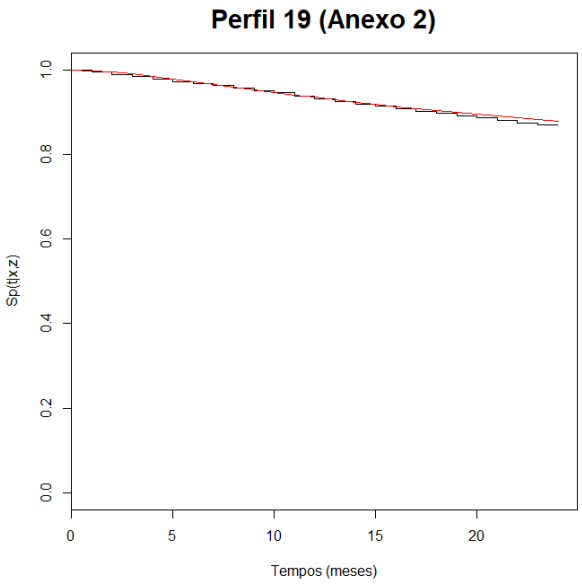


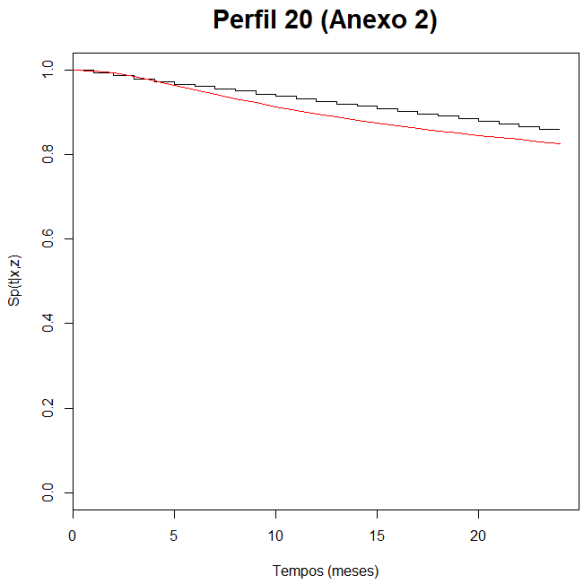
Perfil 17 (Anexo 2)

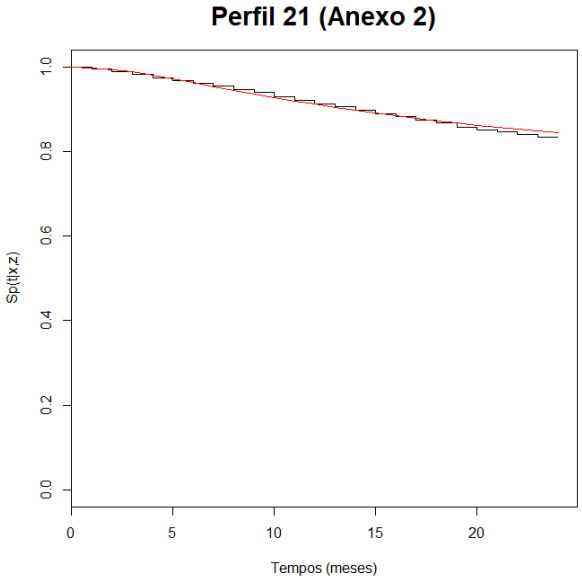


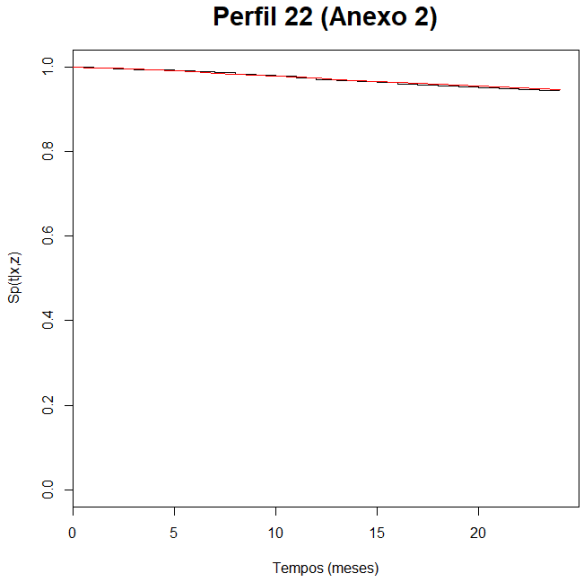
Perfil 18 (Anexo 2)

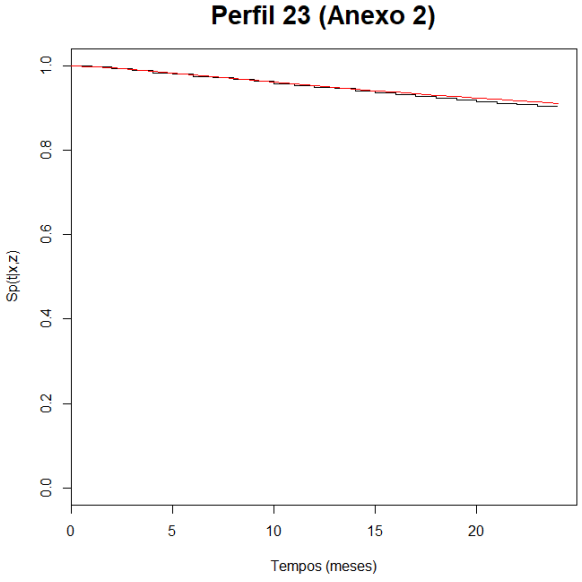


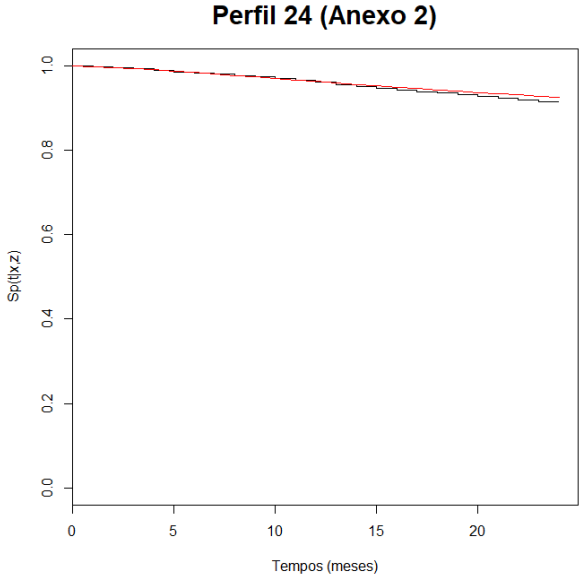

Perfil 19 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

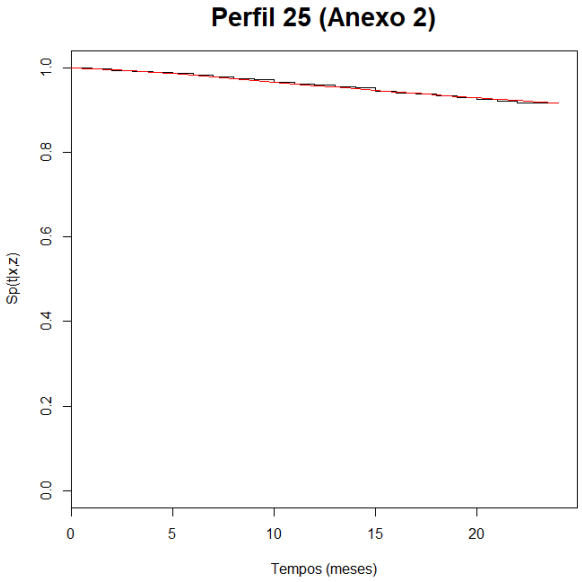

Perfil 20 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

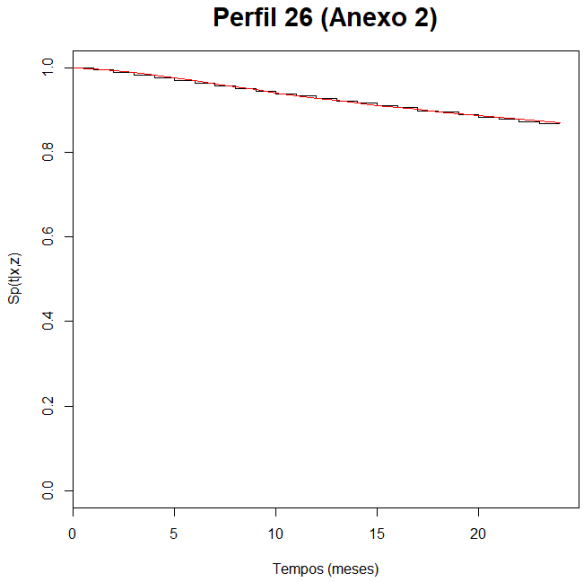

Perfil 21 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

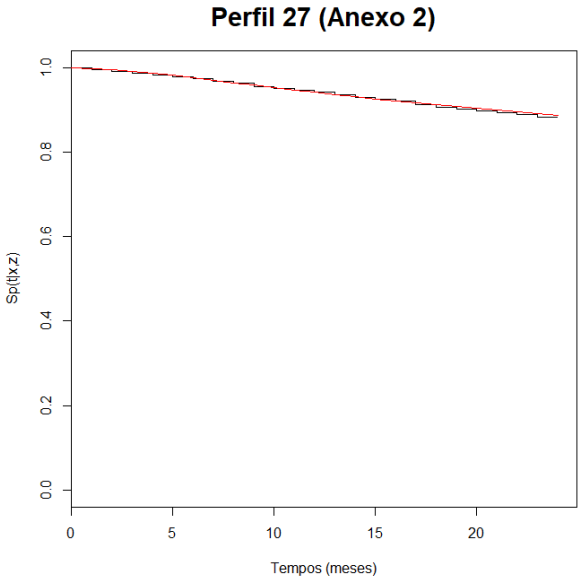

Perfil 22 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

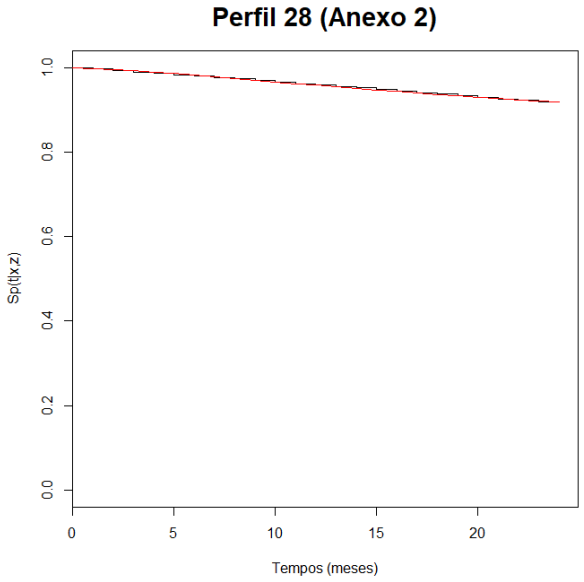

Perfil 23 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

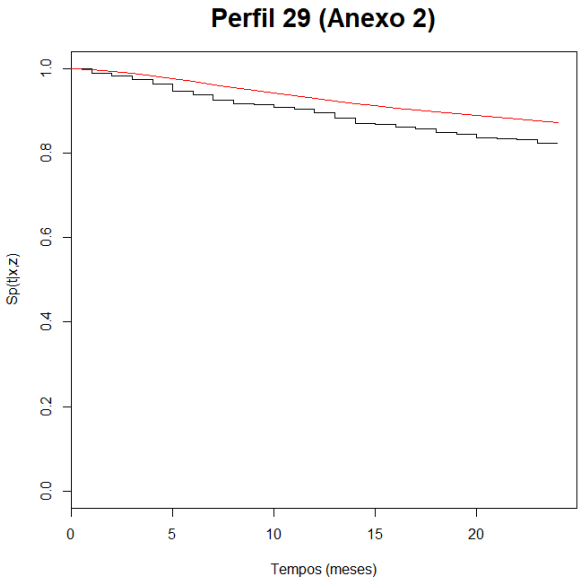

Perfil 24 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

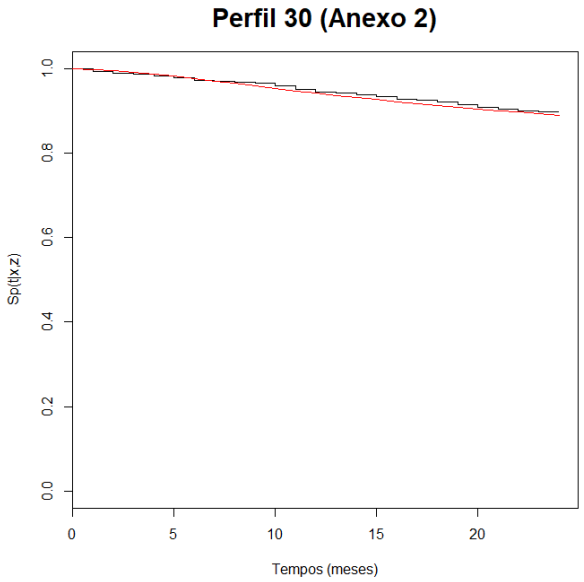

Perfil 25 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

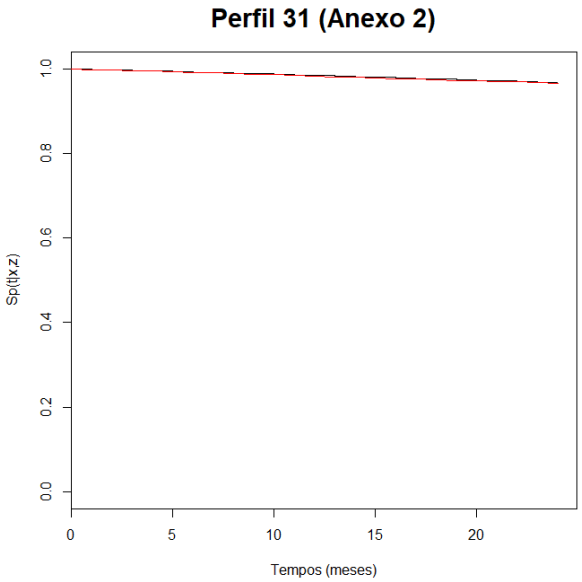

Perfil 26 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

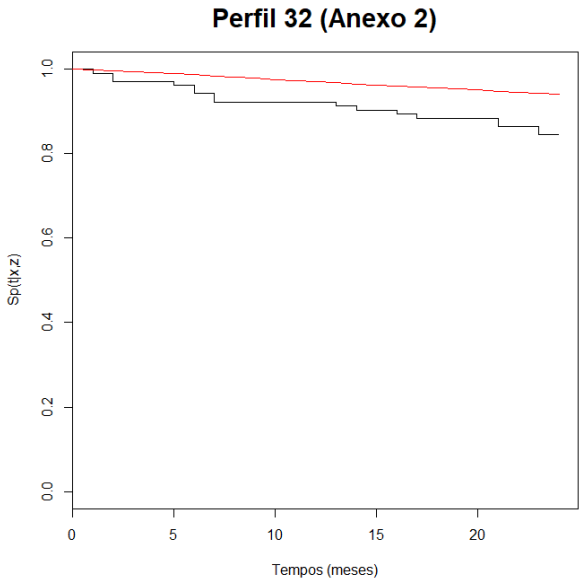

Perfil 27 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

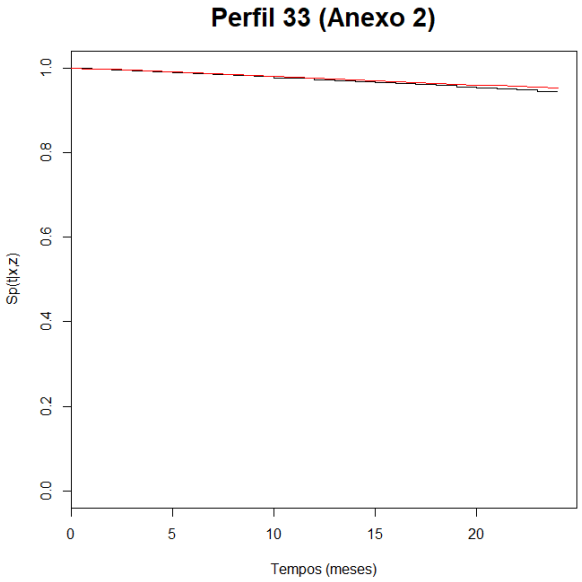

Perfil 28 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

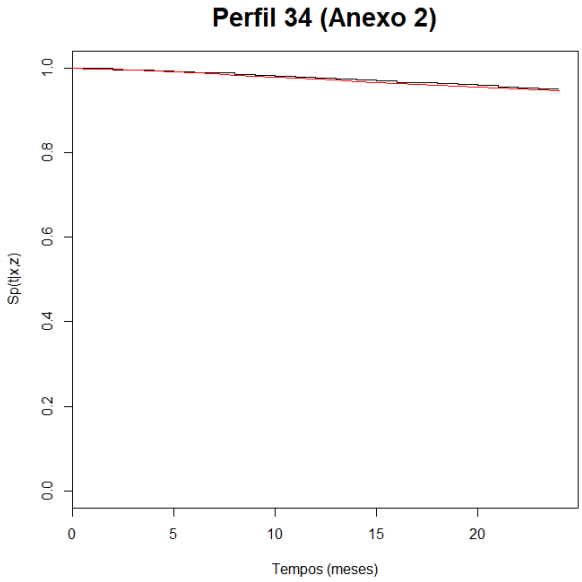

Perfil 29 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

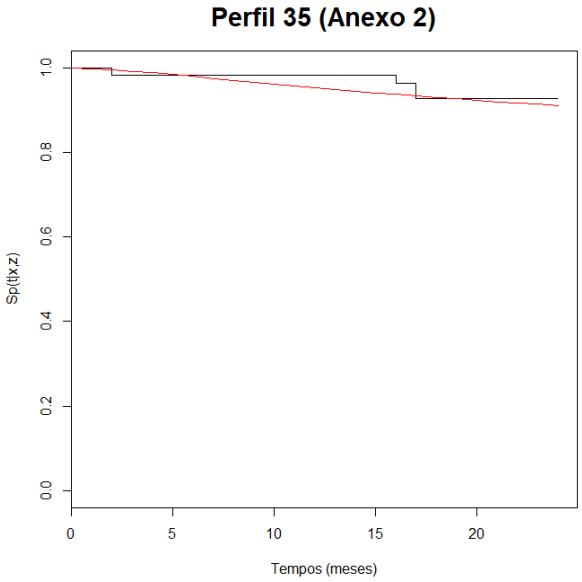

Perfil 30 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)

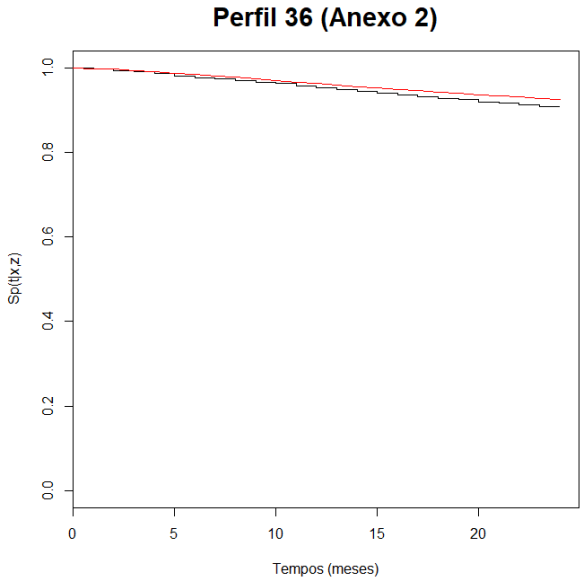

Perfil 31 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)


Perfil 32 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)


Perfil 33 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)


Perfil 34 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)


Perfil 35 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)


Perfil 36 (Anexo 2)
Sp(t|x,z)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Tempos (meses)