UNIVERSIDADE ESTADUAL PAULISTA “JULIO DE MESQUITA FILHO”

FACULDADE DE ENGENHARIA DE BAURU

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

RODRIGO FERNANDO MIOLA

USO DE MODELOS ESTATÍSTICOS PARA DADOS DE ESCORE DE CRÉDITO DE UMA INSTITUIÇÃO FINANCEIRA

Bauru/SP

2013

Rodrigo Fernando Miola

USO DE MODELOS ESTATÍSTICOS PARA DADOS DE ESCORE DE CRÉDITO DE UMA INSTITUIÇÃO FINANCEIRA

Dissertação apresentada ao Departamento de Engenharia de Produção da Faculdade de Engenharia da UNESP de Bauru/SP, como requisito para a obtenção do título de Mestre em Engenharia de Produção.

Área de Concentração: Métodos Quantitativos Aplicados

Orientadora: Profª. Drª. Gladys Dorotea Cacsire Barriga

Bauru/SP

2013

Miola, Rodrigo Fernando.
Uso de modelos estatísticos para dados de escore de crédito de uma instituição financeira / Rodrigo Fernando Miola, 2013

90 f.

Orientador: Gladys Dorotea Cacsire Barriga

Dissertação (Mestrado)–Universidade Estadual Paulista. Faculdade de Engenharia, Bauru, 2013

1. Escore de crédito. 2. Análise discriminante. 3. Análise de sobrevivência com fração de cura. I. Universidade Estadual Paulista. Faculdade de Engenharia. II. Título.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE BAURU
FACULDADE DE ENGENHARIA DE BAURU

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE MESTRADO DE RODRIGO FERNANDO MIOLA, DISCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO, DO(A) FACULDADE DE ENGENHARIA DE BAURU.

Aos 11 dias do mês de dezembro do ano de 2013, às 15:00 horas, no(a) Anfiteatro da Seção Técnica de Pós-graduação da Faculdade de Engenharia de Bauru, reuniu-se a Comissão Examinadora da Defesa Pública, composta pelos seguintes membros: Profa. Dra. GLADYS DOROTEA CACSIRE BARRIGA do(a) Departamento de Engenharia de Produção / Faculdade de Engenharia de Bauru, Prof. Dr. MANOEL HENRIQUE SALGADO do(a) Departamento de Engenharia de Produção / Faculdade de Engenharia de Bauru, Profa. Dra. SILVIA INÊS DALLAVALLE DE PÁDUA do(a) Faculdade de Economia, Administração e Contabilidade/USP/Ribeirão Preto, sob a presidência do primeiro, a fim de proceder a arguição pública da DISSERTAÇÃO DE MESTRADO de RODRIGO FERNANDO MIOLA, intitulado "USO DE MODELOS ESTATÍSTICOS PARA DADOS DE ESCORE DE CRÉDITO DE UMA INSTITUIÇÃO FINANCEIRA". Após a exposição, o discente foi arguido oralmente pelos membros da Comissão Examinadora, tendo recebido o conceito final: APROVADO. Nada mais havendo, foi lavrada a presente ata, que, após lida e aprovada, foi assinada pelos membros da Comissão Examinadora.

Profa. Dra. GLADYS DOROTEA CACSIRE BARRIGA

Prof. Dr. MANOEL HENRIQUE SALGADO

Profa. Dra. SILVIA INÊS DALLAVALLE DE PÁDUA

Dedico este trabalho:

A Deus, fonte de minha energia e sabedoria, por me
Possibilitar a realização desse sonho.

Às minhas filhas, Maria Eduarda e Maria Fernanda,
e à minha esposa Flávia, minhas vidas, pelo amor incondicional
e compreensão dos momentos roubados de seu convívio.

Aos meus pais, Marcílio e Janete, minha irmã Patrícia
e a meu irmão Danilo (in memoriam),
pelo amor, valores e ensinamentos transmitidos.

AGRADECIMENTOS

À Prof.ª Dr.ª Gladys Dorotea Cacsire Barriga pela orientação, amizade, ensinamentos e confiança depositada, que contribuíram para o meu crescimento.

Ao Prof.º Dr. Vicente Garibay Cancho pelo auxílio, atenção e colaboração, que foram de suma importância para a conclusão deste trabalho.

Ao Prof.º Dr. Manoel Henrique Salgado e à Prof.ª Dr.ª Silvia Inês Dallavalle de Pádua pelas valiosas e sinceras contribuições em meu exame de Qualificação e Defesa.

A todos os meus amigos e familiares que, de alguma forma, torceram por mim e me ajudaram nesta conquista.

Aos professores do Departamento de Engenharia de Produção da UNESP de Bauru/SP, cujos ensinamentos e estímulos propiciaram condições para a realização deste trabalho.

Aos funcionários da Seção de Pós-graduação da Faculdade de Engenharia da UNESP de Bauru/SP, pelo apoio concedido.

A todos que, de forma direta ou indireta, contribuíram para a realização deste trabalho.

Pouco conhecimento faz com que as pessoas se sintam orgulhosas.
Muito conhecimento, que se sintam humildes...
O conhecimento torna a alma jovem e diminui a amargura da velhice.
Colhe, pois, a sabedoria. Armazena suavidade para o amanhã.

Leonardo da Vinci

RESUMO

Modelos estatísticos de risco têm sido amplamente utilizados nas últimas décadas pelos bancos e outras instituições financeiras devido à forte concorrência por clientes no mercado de crédito e tendo em vista a regulação bancária quanto a risco de crédito, determinado no Acordo de Basiléia II. O objetivo deste trabalho foi usar modelos estatísticos em dados de escore de crédito de uma instituição financeira com o intuito de analisar a ocorrência de inadimplência. Foram utilizadas duas metodologias estatísticas: análise discriminante e análise de sobrevivência com fração de cura. Com a análise discriminante obteve-se um modelo de escore de crédito em (24) para classificar os clientes em grupos (de adimplentes e inadimplentes), resultando que 72,4% dos clientes foram corretamente classificados pelo modelo de escore de crédito obtido, e novos clientes solicitantes de crédito foram classificados em um dos grupos a partir do modelo obtido. Além disso, os resultados apontam que taxas de juros mais elevadas influenciam na classificação do cliente como inadimplente, enquanto que idades mais elevadas influenciam na classificação de clientes como adimplentes. A segunda metodologia, a análise de sobrevivência com fração de cura (clientes adimplentes) foi utilizada para modelar os dados dos tempos de inadimplência, considerando-se que uma proporção substancial de clientes não apresentou inadimplência durante o período do empréstimo. Os resultados obtidos demonstram que o sexo e a renda dos clientes influenciam na proporção de adimplência dos empréstimos e no risco dos clientes tornarem-se inadimplente. Para aplicação das metodologias estatísticas foi utilizada uma amostra de dados reais selecionados da base de dados de empréstimos da modalidade de Crédito Direto ao Consumidor (CDC) de uma instituição financeira com atuação no Brasil. Os modelos e resultados obtidos poderiam ser utilizados pelo banco nas decisões sobre concessão de crédito.

Palavras-chave: Crédito; Inadimplência; Escore de crédito; Análise Discriminante; Análise de sobrevivência com fração de cura; Modelo Weibull Geométrico.

ABSTRACT

Statistical risk models has been widely used in recent decades by banks and other financial institutions owing to high competition for customers in the credit market and in order to bank regulation regarding credit risk, in particular the Basel Accord II. The objective of this paper was to use statistical methods in data credit scoring financial institution in order to analyze the occurrence of default. In this work were used two statistical methods: discriminant analysis and survival analysis with cure fraction. With discriminant analysis obtained a credit score model in (24) to classify customers into groups (non-defaulting and defaulting), resulting that 72.4% of customers were correctly classified by the credit scoring model obtained, and new customers requesting credit were classified into one of the groups from the model obtained. Furthermore, the results indicate that higher interest rates influence the classification of the customers as defaulting, while older ages influence the classification of customers as non-defaulting. The second methodology, the analysis of survival with cure fraction (non-defaulting customers) was used to model the data from the time of default, given that a substantial proportion of customers showed no default during the loan period. The result shows that gender and income influence the proportion of customers' timely payment of loans and the risk of customers becoming default. In the application of statistical methods was used a sample of real data selected from the database of the type of Loans Consumer Credit (LCC) of a financial institution with operations in Brazil. The models and results could be used by the bank in decisions about granting credit.

Keywords: Credit; Default; Credit Scoring; Discriminant Analysis; Survival Analysis with cure fraction; Weibull Geometric Model.

LISTA DE FIGURAS

Figura 1 – Elementos motivacionais para a elaboração da pesquisa.....15
Figura 2 – Estrutura lógica da pesquisa.....22
Figura 3 – Curva de sobrevida de Kaplan-Meier, estratificada por sexo.....42
Figura 4 – Curva de sobrevida de Kaplan-Meier, estratificada por estado civil.....43
Figura 5 – Curva de sobrevida de Kaplan-Meier, estratificada por renda.....43
Figura 6 – Função de sobrevivência dos clientes adimplentes (S_{WG}), à esquerda, e função de sobrevivência populacional (todos os clientes) (S_{pop}), à direita, com $\theta = 0.3$, $\alpha = 2.0$ e $\lambda = 2.0$, sob os diferentes mecanismos de ativação (PA: pontilhada; AA: sólida; UA: tracejada).....50
Figura 7 – Função de densidade do modelo WG, mecanismo de PA, para alguns valores dos parâmetros α, θ e λ.....51
Figura 8 – Função de risco do modelo WG, mecanismo de PA, para alguns valores dos parâmetros α, θ e 52
Figura 9 – Função de densidade do modelo WG, mecanismo de UA, para alguns parâmetros α, θ e λ.....53
Figura 10 – Funções de risco do modelo WG, mecanismo de UA para alguns parâmetros α, θ e λ.....54
Figura 11 – Distribuição da função para o grupo de clientes adimplente (esquerda) e para o grupo de clientes inadimplentes (direita).....66
Figura 12 – Gráfico QQ plot dos resíduos quantis randomizados normalizados com linha de identidade para o mecanismo de primeira ativação.....77
Figura 13 – Curvas de sobrevivência dos clientes adimplentes estratificada por sexo e renda.....78
Figura 14 – Curvas de sobrevivência populacional (todos os clientes) estratificada por sexo e renda.....78

LISTA DE TABELAS

Tabela 1 – Sistematização dos estudos a respeito de modelos de escore de crédito.....29
Tabela 2 – Sistematização dos estudos a respeito de modelos de classificação utilizando análise discriminante em escore de crédito.....34
Tabela 3 – Sistematização dos estudos a respeito de modelos de análise de sobrevivência com fração de cura.....41
Tabela 4 – Faixa de renda dos clientes da instituição bancária.....61
Tabela 5 – Protocolo de estudo.....62
Tabela 6 – Medidas-resumo da amostra por tipo de cliente.....64
Tabela 7 – Correlação entre as variáveis e a função discriminante.....67
Tabela 8 – Resultados da classificação.....67
Tabela 9 – Correlação entre as variáveis e a função discriminante para a amostra ajustada....69
Tabela 10 – Resultados da nova classificação para a amostra ajustada.....70
Tabela 11 – Resultados médios da classificação utilizando o método de validação cruzada para a amostra ajustada.....71
Tabela 12 – Centroides para os tipos de clientes.....73
Tabela 13 – Classificação de novos clientes solicitantes de crédito.....74
Tabela 14 – Estatísticas dos modelos ajustados.....76
Tabela 15 – Estimativas de Máxima Verossimilhança dos parâmetros para o modelo WG com mecanismo de PA.....77

LISTA DE ABREVIATURA E SIGLAS

AA – Mecanismo de Ativação Aleatória

AD – Análise Discriminante

AE – Amostra de estimação

AIC – *Akaike Information Criterion*

AS – Análise de Sobrevivência

AV – Amostra de validação

CDC – Crédito Direto ao Consumidor

FDP – Função Densidade de Probabilidade

FLDF – Função Linear Discriminante de Fisher

MMV – Método de Máxima Verossimilhança

PA – Mecanismo de Primeira Ativação

RL – Análise de Regressão Logística

SBC – *Schwarz Bayesian Criterion*

UA – Mecanismo de Última Ativação

VAR – *Value-at-Risc*

WG – Weibull Geométrico

SUMÁRIO

1. INTRODUÇÃO....12
1.1 Problema de pesquisa....16
1.2 Objetivos....16
1.3 Justificativa....16
1.4 Delimitação do tema....20
1.5 *Estrutura da dissertação....21*

2. REVISÃO DA LITERATURA....23
2.1 Crédito e escore de crédito....23
2.2 Inadimplência....28
2.3 Análise Discriminante – Método de Fisher....30
2.4 Análise de Sobrevivência....34
2.4.1 Distribuição Weibull....38
2.4.2 Modelo de sobrevivência com fração de cura: um modelo de sobrevivência para análise de escore de crédito....40
2.5 Critérios para comparação de modelos....55
2.5.1 AIC – Akaike Information Criterion....56
2.5.2 SBC – Schwarz Bayesian Criterion....57

3. METODOLOGIA DA PESQUISA....58
3.1 Origem e planejamento da pesquisa....58
3.2 *Técnicas estatísticas....63*

4. ANÁLISE DOS DADOS E DISCUSSÃO DOS RESULTADOS....64
4.1 Aplicação da análise discriminante para análise de escore de crédito....64
4.2 Aplicação do modelo de sobrevivência para análise de escore de crédito....75

5. CONSIDERAÇÕES FINAIS....80
5.1 *Conclusões e limitações do estudo....80*
5.2 *Sugestão para trabalhos futuros....82*

REFERÊNCIAS....83

1 INTRODUÇÃO

Desde a década de 1870 a literatura econômica destaca a relevância dos mercados financeiros para o desenvolvimento econômico de um país (COSTA; NAKANE, 2004). No caso brasileiro, o crédito bancário desempenha papel fundamental na intermediação de poupança e na viabilização de projetos de investimentos. Os elevados custos dos empréstimos são causas de restrições nos volumes de crédito e fontes de outros, estruturalmente ligados aos contratos bancários.

As taxas de juros nas concessões de crédito são muito altas e dão origem a problemas de seleção adversa. Costa e Nakane (2004) registram que

> ...dificuldades de colaterização real geram problemas de risco moral, com incentivos errados determinando o comportamento dos agentes. Estas são questões que emergem em ambientes com problemas informacionais e que estão presentes no funcionamento da intermediação financeira.

Para Annibal (2009), os créditos livres referenciais para taxa de juros representam, em média, 82% do total dos créditos com recursos livres e 52% do total dos créditos do Sistema Financeiro Nacional. Essas operações podem ser feitas em diferentes modalidades. No caso de pessoas físicas, as modalidades são: cheque especial; crédito pessoal; financiamento imobiliário; aquisição de bens – veículos automotores; aquisição de bens – outros bens; oriundas de cartão de crédito; e outras. Dessas, o crédito pessoal representa, em média, pouco menos da metade do volume total. A segunda modalidade mais importante é aquisição de bens – veículos automotores. Somadas, essas duas modalidades representam, em média, mais de 80% do volume total de operações referenciais para taxas de juros a pessoas físicas.

No Brasil verificam-se níveis elevados de taxas de empréstimos bancários, que está vinculado ao *spread* praticado pelos bancos no país. De acordo com Costa e Nakane (2004, p. 3), *spread*:

> É definido como sendo a diferença entre o custo de captação dos bancos e o custo cobrado por esse banco quando ele concede um empréstimo. Portanto não se configura aí o lucro do banco, pois há que se deduzirem os custos vinculados à atividade de captação e empréstimo. De forma geral, e particularmente no Brasil, o spread bancário é formado a partir da agregação de vários fatores de custo e de margem. Os fatores de custo se referem a custos administrativos e demais custos operacionais vinculados à atividade bancária; a custos regulatórios da intermediação financeira — e aí entram compulsórios, os subsídios cruzados e custos com as

> contribuições para o sistema de seguro depósito; custos fiscais dados pela incidência de diversos impostos sobre a intermediação financeira e custo de inadimplência, vinculados ao risco de crédito implícito na concessão de empréstimos.

A concessão de crédito começou a ser mais importante para as empresas do setor financeiro, tornando-se uma das principais fontes de receita para os bancos e instituições financeiras em geral (THOMAS, 2010; LOUZADA *et al*., 2012). Devido a este fato, este setor da economia percebeu que foi altamente recomendado aumentar a quantidade de recursos alocados, sem perder a agilidade e a qualidade dos créditos, altura em que a contribuição da modelagem estatística é essencial.

Assim, as instituições financeiras passaram a desenvolver modelos internos para análise de risco de crédito, tendo em vista a regulação bancária quanto a esse tema, determinado no Acordo de Basiléia II (Comitê de Basiléia de Supervisão Bancária, 2006), o que permitiu que os bancos mensurassem com mais qualidade os créditos de sua carteira. Na elaboração dos modelos internos de mensuração do risco de crédito são aplicadas estruturas conceituais utilizadas para análise dos riscos de mercado, utilizando-se de técnicas matemáticas para a apuração de um valor de perda potencial dentro de um intervalo de tempo e um percentual de significância estatística.

Saunders e Schumacher (2000) destacam alguns modelos de análise de risco de crédito utilizados por algumas instituições financeiras nos últimos anos:

- *CreditMetrics* - modelo desenvolvido pelo *JPMorgan Bank Inc.*, baseado na abordagem de migração da qualidade do crédito concedido. Define probabilidades de mudanças da qualidade do crédito, inclusive para falência, dentro de um horizonte de tempo;
- *KMV* - modelo desenvolvido pela *KMV Corporation*, baseado na abordagem estrutural ou avaliação de ativos com base na teoria de opção. O modelo considera o processo de falência endógeno e relacionado à estrutura de capital da empresa. A falência acontece quando o valor dos ativos da empresa cai abaixo de um nível crítico;
- *CreditRisk+* - modelo desenvolvido pelo *Credit Suisse Financial Products* baseado na abordagem atuarial. O modelo procura estabelecer medidas de perda esperada com base no perfil de sua carteira de empréstimos ou títulos e no histórico de inadimplência;

- *CreditPortfolioView* - modelo desenvolvido pela Consultoria *McKinsey* baseado no impacto de variáveis econômicas na inadimplência. O modelo traça cenários multiperíodo onde as chances de falência estão associadas a variáveis como desemprego, patamar de taxas de juros, taxa de crescimento da economia, entre outras.

Os métodos tradicionais de avaliação de crédito em áreas acadêmicas são métodos estatísticos que veem a avaliação como uma classificação de reconhecimento de padrões (MA; TANG, 2007). Segundo esses autores, muitas instituições financeiras têm aplicado previsão de risco de crédito e conseguido algum sucesso. Classificações internas para tomadores de crédito são um procedimento de avaliação agregada de vários fatores financeiros e não financeiros (GRUNERT *et al*., 2005). Na prática bancária, classificações representam a base para a aprovação do empréstimo, preços, monitoramento e provisionamento para perdas com empréstimos, sendo que o papel dos fatores não financeiros permanece ambíguo.

A preocupação com a qualidade na concessão de créditos aos clientes é constante, uma vez que o risco de crédito e risco de mercado são inerentes. Tem sido documentado que as probabilidades de descumprimento dos acordos de crédito e as taxas de recuperação variam através de ciclos de negócios (TANG; YAN, 2010).

Os modelos estatísticos de risco de crédito mais utilizados pelas organizações empresariais, tanto para avaliação de risco de crédito pessoa física como para pessoa jurídica, são modelos de classificação de risco conhecidos como modelo de escore de crédito (*credit scoring).* A consolidação do uso de modelos de escore de crédito ocorreu na década de 1990, quando as mudanças no cenário mundial, como a desregulamentação das taxas de juros e taxas de câmbio, aumento da liquidez e na competição bancária fizeram as instituições financeiras se preocuparem mais com o risco de crédito, ou seja, o risco que estavam correndo ao aceitar alguém como cliente.

Segundo Akkoç (2012), diversos modelos de escore de crédito foram desenvolvidos por bancos e pesquisadores nas últimas décadas a fim de avaliar os pedidos de crédito de clientes utilizando diferentes técnicas e conceitos estatísticos, incluindo a Análise Discriminante (AD).

Em complemento às técnicas estatísticas multivariadas aplicadas em escore de crédito identifica-se a utilização de Análise de Sobrevivência (AS). Com a AS é possível prever não somente se o cliente tornar-se-á inadimplente na amortização de seu empréstimo, mas também

quando serão suscetíveis a essa ocorrência. Assim, os métodos de AS permitem estimar a probabilidade de inadimplência em qualquer horizonte de tempo de escolha (TONG *et al.*, 2012).

Assim, diante da importância da oferta de crédito para a economia de um país, e diante do fato de a inadimplência nesses créditos concedidos impactar diretamente nos resultados, liquidez e solvência das instituições bancárias, mostra-se a de suma importância um estudo sobre o tema. Segundo Alves (2009) diversas pesquisas buscaram analisar questões relacionadas à solvência de instituições bancárias frente a diversos índices econômico-financeiros, porém pouco se tem estudado como se comporta a inadimplência nas carteiras de créditos bancários, fato que pode impactar diretamente na solvência de uma instituição bancária. É de interesse das instituições bancárias a identificação de como se comporta a inadimplência em suas operações de crédito a fim de que sejam capazes de conhecerem a probabilidade de ocorrência da inadimplência e suas características, e tomar ações preventivas para as análises de crédito futuras.

A Figura 1 apresenta um resumo dos elementos motivacionais da presente pesquisa.

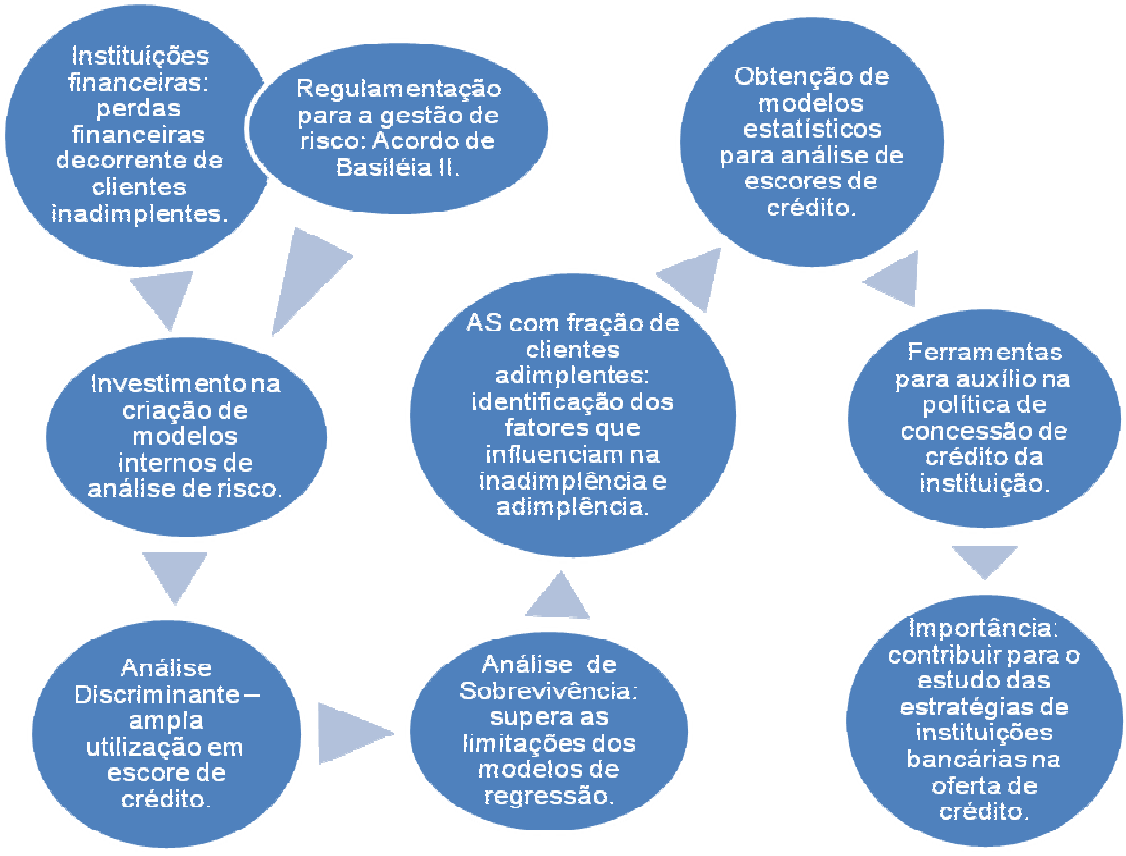

Figura 1: Elementos motivacionais para a elaboração da pesquisa.

Fonte: Elaborado pelo autor.

1.1 Problema de pesquisa

Tendo em vista a utilização de modelos de escores de crédito na resolução de problemas de inadimplência nas concessões de crédito em instituições financeiras, conforme estudado por alguns autores, o presente trabalho de pesquisa busca responder à seguinte questão:

- Como utilizar modelos estatísticos em uma carteira de crédito de uma instituição financeira para analisar e prever a ocorrência de inadimplência considerando os fatores ou variáveis relacionadas ao perfil dos clientes, e/ou características financeiras do crédito tomado?

1.2 Objetivos

O objetivo geral deste trabalho é utilizar modelos estatísticos para dados de escore de crédito de uma instituição financeira com o intuito de analisar e prever a ocorrência de inadimplência.

Os objetivos específicos são:

1. Obter modelo estatístico para classificar os clientes como inadimplentes ou adimplentes e, a partir desse modelo, classificar novos clientes em um dos dois grupos (de adimplentes ou inadimplentes), e analisar a influência das variáveis utilizadas na inadimplência ou adimplência dos clientes;
2. Obter modelo estatístico para modelar os tempos até a ocorrência de inadimplência e a proporção de adimplência, e, a partir desse modelo, determinar os fatores que influenciam na proporção de não inadimplência e nos tempos até a ocorrência de inadimplência.

1.3 Justificativa

O risco de perda financeira decorrente de clientes que não realizam pagamentos conforme contratado, normalmente chamado de risco de crédito ou risco de inadimplência, é

muito importante nas finanças. A gestão do risco de crédito é fundamental para o desempenho das instituições financeiras individualmente e para o funcionamento do mercado financeiro como um todo (DIVINO; ROCHA, 2013).

Na última década os bancos e outras instituições financeiras investiram significativas quantias no desenvolvimento de modelos internos de risco, tendo em vista a regulação bancária quanto a esse tema. A Emenda de Risco de Mercado para o Acordo de Basiléia II (Comitê de Basiléia de Supervisão Bancária, 2006) incorporou formalmente modelos internos de risco de mercado dos bancos no cálculo do capital regulamentar. Com isso, os requisitos de capital para risco de crédito de um banco (o Acordo de Basiléia II incentivou as instituições para gerir o risco de crédito com base em modelos desenvolvidos internamente em uma abordagem de *value-at-risc* (VaR) para determinar a adequação de capital econômico) são uma função do próprio banco (ODEH, 2011). Isso permitir que os bancos mensurem com qualidade os créditos de sua carteira e a qualidade de clientes potenciais e/ou oportunidades de investimentos. Isto torna mais atraente para o banco o desenvolvimento ou compra do seu próprio modelo de avaliação de risco, fornecendo gestão mais rigorosa e eficiente dos ativos do banco, e precificando adequadamente o risco. Com uma previsão mais confiável da probabilidade de perda de crédito, os credores dedicam especial atenção à avaliação de desempenho e procedimentos de aprovação de qualquer pedido de crédito em potencial. Assim, modelos estatísticos são frequentemente utilizados por atingir a maioria dos empréstimos de pequeno volume.

Segundo Eifert (2003), um estudo aplicado à prevenção da inadimplência com as informações disponíveis no momento da análise de crédito deve contribuir para uma avaliação mais criteriosa, possibilitando às instituições bancárias diminuição da inadimplência e, consequentemente, melhores resultados.

A maioria dos modelos para análise de risco de crédito é baseada na ideia de que há um fator, ou um conjunto de fatores de riscos comuns que impulsionam as taxas de inadimplência de todas as posições (ALESSANDRI; DREHMANN, 2010).

A avaliação da probabilidade de inadimplência faz parte da inteligência dos negócios e sistemas de gestão de clientes em relação às instituições financeiras, definindo a probabilidade de que um empréstimo não será reembolsado e caia em inadimplência (PERKO *et al*., 2011). A definição mais consistente da probabilidade de inadimplência e os requisitos de sistema para sua avaliação são encontrados no Acordo de Basiléia II e na Diretiva de Requisitos de Capital do Parlamento Europeu de 2006.

Segundo Odeh (2011) existem diversos estudos em finanças e economia acerca de inadimplência e risco de crédito. Alguns destes estudos desenvolveram modelos que predizem a probabilidade de ocorrência de inadimplência de crédito, outros tratam dos modelos de escore de crédito existentes e propõem métodos para estimar o risco da carteira de crédito.

Além disso, de acordo com Gürtler e Hibbeln (2013), alguns estudos tratam da relação entre a probabilidade de inadimplência e a perda no caso de sua ocorrência. Embora a maior parte da literatura consista em estudos empíricos de empresas, uma fração menor incide sobre empréstimos bancários devido à limitada disponibilidade de dados.

Para Divino e Rocha (2013) as recentes crises financeiras internacionais têm contribuído para a disseminação empírica da aplicação de modelos de risco de crédito. Diferentes modelos surgiram para lidar com a probabilidade de inadimplência de pessoas físicas e empresas.

Diversos modelos de escore de crédito têm sido desenvolvidos utilizando diferentes técnicas e conceitos estatísticos, incluindo a Análise Discriminante (AD), análise de Regressão Logística (RL), de Regressão Multivariada *Splines* Adaptativas, Classificação e Árvore de Regressão, Redes Neurais Artificiais, *Support Vector Machines* e Algoritmo Genético, conforme trabalhos de Abdou *et al.* (2008); Abdou (2009); Angelini *et al.* (2008); Bellotti e Crook (2009); Chen e Huang, (2003); Chuang e Lin (2009); Cinko (2006); Desai *et al.* (1996);. Hsieh e Hung (2010); Hsieh (2004, 2005), Huang *et al.* (2006, 2007); Lee e Chen (2005); Kim e Sohn (2010); Lee *et al.* (2002, 2006); Lee (2007); Li *et al.* (2006), Luo *et al.* (2009). Malhotra e Malhotra (2003); Nanni e Lumini (2009); Ong *et al.* (2005); Paleologo *et al.* (2010); Sustersic *et al.* (2009); Tong *et al.* (2012). Tsai e Wu (2008). Tsai *et al.* (2009); Oeste (2000); Ocidente *et al.* (2005). Yu *et al.* (2008).

Para Thomas (2000), especificamente a AD têm sido amplamente utilizada na pontuação de crédito, com destaque para os trabalhos de Chuang e Lin (2009), que apresenta um modelo de escore de crédito mais preciso, com aplicação em um conjunto de dados em cartão de crédito, e Crook *et al.* (2007), que testam as Máquinas de Vetores de Suporte contra AD de um banco de dados de cartão de crédito de grande porte, e mostram que eles são competitivos e podem ser usados como a base de um método de seleção para descobrir as características que são mais importante na determinação de risco de inadimplência.

No presente trabalho a técnica de AD foi empregada para a obtenção de um modelo de escore de crédito utilizando-se da modalidade de crédito estudada, o Crédito Direto ao

Consumidor (CDC) para clientes pessoa física, senda as variáveis definidas quanto ao perfil do tomador do crédito, como sexo, idade, estado civil e renda, e as variáveis financeiras, como a taxa de juro da operação, o valor do contrato e da prestação, e o prazo para a amortização do crédito.

Na presente pesquisa também foi utilizada a AS, que tem sido bem estabelecida na engenharia (Hosner *et al.*, 2008). De acordo com Tong *et al.* (2012), o método foi introduzido pela primeira vez em classificação de crédito por Narain (1992). Seu uso nesse contexto foi desenvolvido por Banasik *et al.* (1999), Stepanova e Thomas (2002) e Mão e Kelly (2001). Banasik *et al.* (1999) compararam o desempenho de modelos de risco paramétricos e semiparamétricos de regressão. Stepanova e Thomas (2002) demonstraram que a AS pode também ser utilizada para prever o tempo de início de reembolso de empréstimo.

Algumas aplicações de AS podem ser encontradas nos trabalhos de Hoggart e Griffin (2001), Zaider *et al.* (2001), Yin (2005), Perdoná (2006), Chi e Ibrahim (2007), Kim *et al.* (2007), Mizoi *et al.* (2007), Peng *et al.* (2007), Tournoud e Ecochard (2007), Cancho *et al.* (2008) e Sen e Tan (2008).

Diante da característica das análises necessárias dentro da proposta de pesquisa, a aplicação da análise dos dados utilizando AS se justifica tendo em vista que a análise foi realizada em um modelo de escore de crédito buscando a solução do problema de pesquisa proposto. A análise de dados utilizando a técnica de AS é atrativa na medida em que supera as limitações dos modelos de regressão, incluindo tendências longitudinais de desempenho de rastreamento para diferentes grupos examinados (CHAMPLAIN, 2010).

Dentre os trabalhos sobre AS, cabe destaque àqueles que utilizam modelos de sobrevivência com fração de cura. Modelos de sobrevivência com fração de cura (também conhecido como distribuições de sobrevivência de longo prazo) têm merecido um grande interesse por diferentes autores nos últimos anos. Em recentes pesquisas relacionadas a dados de escore de crédito, tem sido demonstrado que o uso de tempo variável em variáveis macroeconômicas com modelos de riscos proporcionais de Cox melhorou a precisão das estimativas de probabilidade inadimplência na pontuação de crédito (BELLOTTI; CROOK, 2009 *apud* TONG *et al.*, 2012). Da mesma forma, Thomas (2009) sugeriu a abordagem de função de risco para estimar o risco de crédito de carteiras de empréstimos ao consumidor em vez de pontuação de crédito em um nível de conta específica. Encontram-se, ainda, estudos recentes que demonstram o desempenho superior dos modelos de cura sobre previsão de falência corporativa (TOPALOGLU; YILDIRIM, 2009), e o estudo sobre previsão de

inadimplência em empréstimos imobiliários corporativos realizado por Yildirim (2008) (TONG *et al.*, 2012).

Banasik e Crook (2010) propõem uma aplicação e demonstram o benefício de pontuação de candidatos a crédito por meio de análise de sobrevivência e pelo modelo de riscos proporcionais de Cox envolvendo modelos mais elaborados, respondendo a perguntas mais específicas, como quando ocorrerá a inadimplência e qual será a sua implicação financeira precisa.

Dentre os trabalhos sobre o tema ressalta-se o trabalho de Tong *et al.* (2012), que considera o modelo de mistura proposto por Boag (1949) para a modelagem do tempo até o cliente tornar-se inadimplente, onde o risco dos clientes não inadimplentes é modelado pelo modelo de riscos proporcionais de Cox (LAWLESS, 2003), e a proporção de não inadimplentes é modelado pela função de ligação logística.

Assim, a proposta desse trabalho consiste, também, em utilizar um modelo de sobrevivência com fração de cura (fração de adimplência ou fração de adimplentes), obtido assumindo que a ocorrência de inadimplência é devida a causas ou riscos latentes não observados, onde o número de causas é modelado por uma distribuição Geométrica e os tempos associados a essas causas são modelados pela distribuição Weibull. A utilização da distribuição Weibull decorre de que ela vem sendo frequentemente utilizada na literatura, e sua popularidade em aplicações práticas se deve ao fato de apresentar uma grande variedade de formas (COLOSIMO; GIOLO, 2006).

1.4 Delimitação do tema

Esta pesquisa considerou um conjunto de dados de uma modalidade de empréstimo (CDC) para pessoa física. Como universo de pesquisa, foram considerados apenas os clientes das agências de determinada cidade da instituição financeira em estudo, tomadores de crédito na citada modalidade. Desta forma, procurou-se delimitar o escopo de análise em relação ao objeto a ser estudado.

O estudo dos dados com a AD foi utilizado nesse trabalho, pois geralmente é aplicado em amostras agrupadas, e restringiu-se a avaliar quais variáveis são importantes para a

discriminação dos grupos conhecidos, e explorar características capazes de serem utilizadas para alocar novos clientes nos diferentes grupos previamente definidos.

A utilização da AS ficou restrita a aplicação de um modelo estatístico para modelar os tempos até a ocorrência de inadimplência dos clientes e a proporção de clientes adimplentes. Para isso foi utilizado um modelo de AS com fração de cura obtido a partir de uma distribuição Weibull e uma distribuição Geométrica, e, a partir desse modelo, a análise restringiu-se a determinar os fatores que influenciam na proporção de não inadimplência e nos tempos até a ocorrência de inadimplência.

1.5 Estrutura da dissertação

O presente trabalho está disposto em seis capítulos, considerando a introdução anteriormente apresentada no primeiro capítulo, que descreve o problema de pesquisa, os objetivos, a justificativa do trabalho e a delimitação do tema abordado. O segundo capítulo contém a fundamentação teórica. Apresenta os aspectos conceituais referentes a crédito e escore de crédito, inadimplência, e das técnicas estatísticas multivariadas de análise discriminante e análise de sobrevivência, além de critérios para comparação de modelos.

Já o terceiro capítulo apresenta a metodologia do trabalho, contendo um detalhamento da população de estudo, com a definição da característica da carteira de crédito estudada, bem como os critérios para seleção da amostra dessa população e definição das variáveis utilizadas nas análises realizadas. No quarto capítulo são realizadas as análises dos dados, interpretação e discussão dos resultados com base nos métodos estatísticos propostos.

No quinto capítulo são discutidos os resultados obtidos e considerações finais do trabalho, e abordadas sugestões para futuras pesquisas sobre o tema. Por fim, são apresentadas as referências utilizadas.

A estrutura lógica da pesquisa é ilustrada na Figura 2:

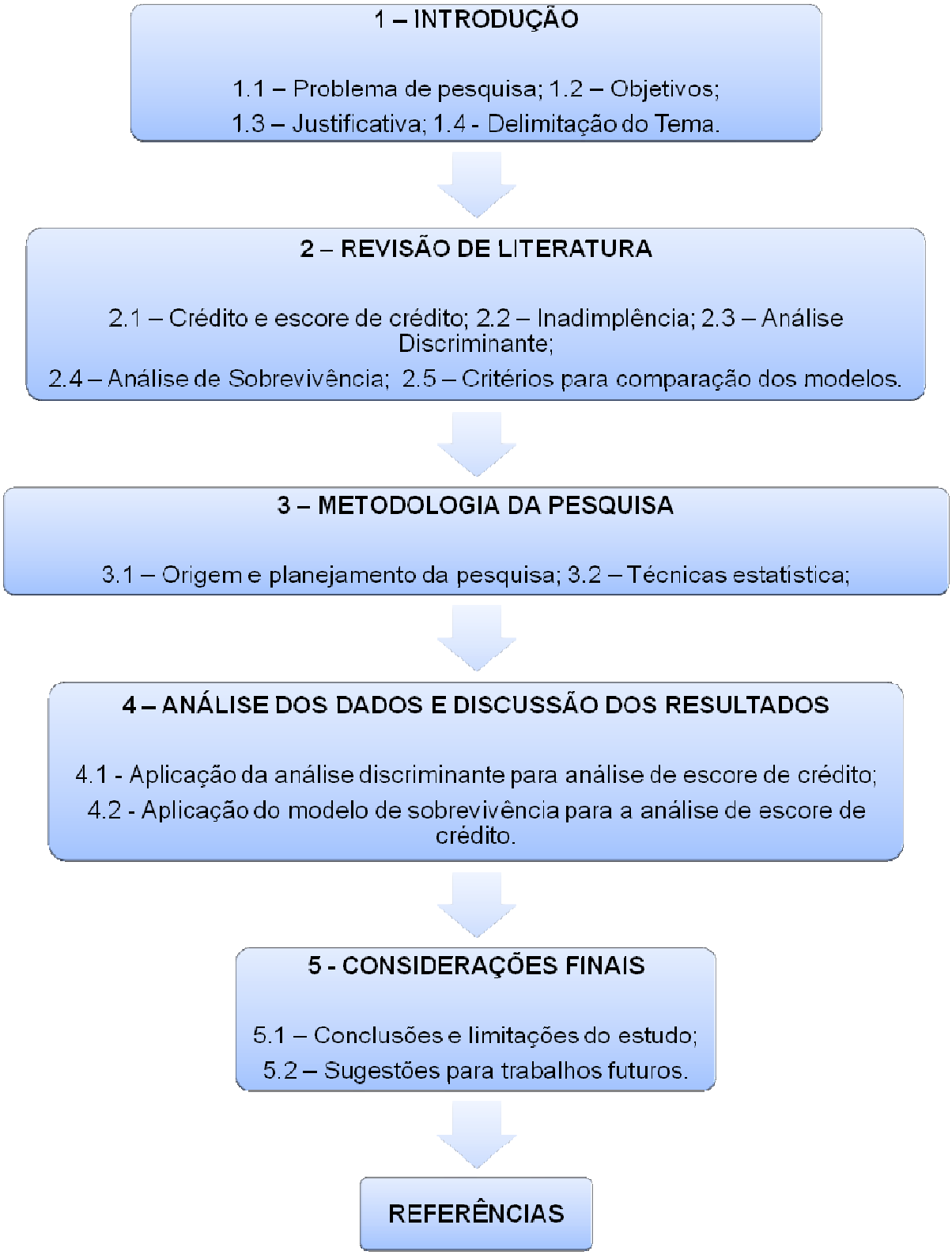

Figura 2: Estrutura lógica da pesquisa

Fonte: Elaborado pelo autor

2 REVISÃO DA LITERATURA

A revisão da literatura, ou referencial teórico, objetiva situar o problema no tempo e no espaço. Esse capítulo apresenta uma discussão teórica dos problemas buscando fundamentá-los nas teorias existentes, disponíveis na literatura mundial. Inicialmente, são abordadas definições, finalidade e aspectos gerais relacionados ao crédito e análise de risco de crédito, conceituando o escore de crédito e levantando um breve histórico de sua aplicação. Em seguida é abordada a definição de inadimplência e como (após quanto tempo) o cliente é considerado inadimplente na presente pesquisa.

Posteriormente, é tratada a técnica estatística de análise discriminante com a definição do método que é utilizado, suas definições e propriedades e objetivos de sua utilização, e são apresentados alguns trabalhos relacionados a escore de crédito onde a técnica AD foi utilizada.

Após, é abordada a técnica estatística de análise de sobrevivência, suas definições, propriedades e objetivos de sua utilização, e obtidos modelos de sobrevivência com fração de cura (clientes adimplentes) para análise de escore de crédito, encerrando com alguns trabalhos relacionados a escore de crédito onde a técnica de AS foi utilizada. Por fim, são apresentados alguns critérios para a comparação de modelos que serão utilizados para a definição do modelo de AS com fração de clientes adimplentes a ser utilizado na presente pesquisa.

2.1 Crédito e escore de crédito

De acordo com Schrickel (2000), crédito “é todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente, parte de seu patrimônio a um terceiro”. O crédito inclui as noções fundamentais de confiança e tempo, este no que se refere ao período determinado entre a aquisição e a liquidação da dívida, e aquela que expressa a promessa de pagamento.

Para Santos (2009), a finalidade do crédito deve estar relacionada com a necessidade do cliente, e para pessoas físicas, o crédito deve atender:

- Créditos emergenciais para atender as necessidades imediatas do cliente, em consequência de eventuais desequilíbrios orçamentários, ou financiamento de compras em operações de crédito de curtíssimo prazo (prazo inferior a um mês);
- Financiamentos de compras para aquisição de produtos e serviços para consumo e bem-estar em operações de crédito de curto prazo (prazo inferior a 12 meses); ou
- Investimento para aquisição de bens de maior valor para integrar seu patrimônio ou desempenhar suas atividades profissionais em operações de crédito de longo prazo (prazo superior a 12 meses).

Para o financiamento das necessidades identificadas, o cliente pode utilizar duas modalidades de crédito: as linhas chamadas rotativas, que consistem em limites de crédito disponibilizados ao cliente, e as linhas pontuais, destinadas para financiamento de forma previamente definida. Dentre as modalidades de créditos constantes na linha pontual, destaca-se o Crédito Direto ao Consumidor, também conhecido como CDC, destina-se a financiar a prestação de serviço ou aquisição e bens duráveis, com amortizações mensais fixas, que contemplam os encargos (como juros, por exemplo) envolvidos. Essa modalidade é objeto de estudo para o presente trabalho.

Na solicitação de crédito em uma instituição bancária, o cliente pessoa física submete-se a uma avaliação, geralmente com base na análise quanto a aspectos pessoais (caráter e capacidade), e aspectos financeiros (capital e condições). O processo de análise de crédito para pessoa física visa a identificar os riscos para a organização que está concedendo o crédito, evidenciar conclusões quanto à capacidade de repagamento do tomador e fazer recomendações sobre o melhor tipo de empréstimo a ser concedido.

Para Schrickel (2000), a análise ocorre conforme as necessidades do solicitante e dentro de um nível de risco aceitável, a partir de documentação apresentada e análise da mesma, objetivando a maximização dos resultados da instituição. O processo de análise de crédito para pessoa física baseia-se na qualidade das informações obtidas e nas decisões decorrentes da análise dessas informações. Essas decisões devem ser práticas e viáveis dentro de um modelo funcional adaptado à realidade da organização, e de acordo com as diretrizes estabelecidas pelo setor de atuação.

Em toda situação de análise de crédito devem ser consideradas três etapas distintas:

- Análise Retrospectiva: avaliação do desempenho histórico do tomador potencial, analisando os riscos inerentes ao mesmo e como foram contornados. Este processo

visa identificar fatores na atual condição do tomador que possam dificultar o pagamento da dívida;

- Análise de Tendências: projeção da condição futura do tomador do crédito, a fim de avaliar o nível de endividamento suportável e quão oneroso será o crédito que se espera obter; e
- Capacidade Creditícia: a partir do grau de risco que o tomador apresenta e a projeção do seu nível de endividamento futuro, avaliar a capacidade creditícia do tomador, ou seja, qual a quantia de capital que ele poderá obter junto ao credor.

Segundo Santos (2009), as fases para o processo de análise de crédito de pessoa física são:

- Análise cadastral;
- Análise de idoneidade;
- Análise financeira;
- Análise de relacionamento;
- Análise patrimonial;
- Análise de sensibilidade;
- Análise do negócio;
- Parâmetros para estabelecer o limite de crédito e o valor de financiamento.

A análise cadastral é a análise de identificação do cliente, contemplando: estado civil, idade, renda, etc. O levantamento e a análise dessas informações cadastrais são de suma importância para a determinação do valor do crédito a ser concedido, o prazo para a amortização do crédito, a taxa de juro da operação, e o valor da prestação a ser paga mensalmente.

Segundo o Manual de Normas e Instruções do Banco Central do Brasil (2006), as instituições financeiras devem conceder crédito a tomadores que possuem adequadas e não restritivas informações cadastrais. Schrickel (2000) reafirma todos estes fatos, dizendo que as instituições de crédito devem munir-se de elementos informativos essenciais e indispensáveis sobre o potencial tomador do crédito, antes de manter qualquer tipo de relacionamento concreto ou formalizar alguma operação de crédito.

Para Santos (2009), os dados que devem ser identificados quando da análise deverão ser:

- Escolaridade;
- Estado Civil;
- Idade;
- Idoneidade;
- Moradia (se própria ou alugada, e o tempo de residência na moradia);
- Número de dependentes;
- Renda (principal e complementar);
- Situação legal dos documentos; e
- Tempo no atual emprego/atividade exercida.

A história do escore de crédito iniciou-se com Durand em 1941, cujo trabalho realizou uma discriminação entre bons e maus empréstimos (THOMAS *et al.*, 2010). Na década de 60, com o aumento do número da utilização de cartões de crédito, o escore de crédito passou a ser uma boa opção para avaliação do risco de crédito, pois automatiza e agiliza o processo de liberação de crédito, e reduz as taxas de inadimplências das empresas. Na década de 1980, devido ao sucesso obtido nos cartões de crédito, os bancos também começam a utilizar o escore de crédito em outros produtos, entre eles empréstimos pessoais, crédito imobiliário e crédito para pequenas empresas (THOMAS *et al.*, 2010). A aplicação de modelos de escore de crédito e outras ferramentas para análises de crédito se iniciaram na década de 1930, em companhias seguradoras, conforme Blatt (1999). Porém, seu desenvolvimento em instituições financeiras ocorreu a partir da década de 1960. Este modelo proporciona uma vantagem competitiva para a organização.

O termo escore de crédito é utilizado para descrever métodos estatísticos adotados para classificar candidatos à obtenção de um crédito em grupos de risco. A partir do histórico de concessões de crédito efetuadas por uma instituição de crédito é possível, através de técnicas estatísticas, identificar as variáveis sócio-econômicas que influenciam na capacidade do cliente em pagar o crédito, ou seja, na qualidade do crédito da pessoa física. O método é baseado na classificação de candidatos a crédito em grupos de acordo com seus prováveis comportamentos de pagamento.

Santos (2009) complementa explicando que o escore de crédito é baseado em informações do passado recente da carteira de crédito e gera notas (escores) para novos candidatos ao crédito que representam a expectativa de que os clientes paguem suas dividas sem se tornarem inadimplentes. Conforme ressalta Saunders (2000), o escore pode ser utilizado para classificação de créditos como adimplentes ou inadimplentes, bons ou maus, desejáveis ou não, de acordo com a pontuação obtida por cada crédito. Esta classificação, por sua vez, pode orientar a decisão do analista em relação à concessão ou não do crédito solicitado.

O objetivo dos modelos de escore de crédito é classificar os solicitantes de crédito de acordo com a sua probabilidade de inadimplência. São atribuídos pesos estatisticamente predeterminados para certos atributos (variáveis) dos solicitantes de crédito, gerando uma pontuação para cada cliente. Caso o cliente tenha um escore maior que um determinado ponto de corte (escore/pontuação mínima para a aprovação do crédito), o crédito é aprovado, caso contrário, é reprovado. A ideia básica destes modelos “é a pré-identificação de certos fatores-chave que determinam a probabilidade de inadimplência e sua combinação ou ponderação para produzir uma pontuação quantitativa” (SAUNDERS, 2000).

Para Silva (2006), é importante considerar que o uso de métodos quantitativos como o escore de crédito não elimina a necessidade de que as organizações empresariais tenham claras definições políticas e estratégicas e de que seus profissionais sejam devidamente treinados em crédito.

Destacam-se como vantagens dos modelos de escore de crédito a possibilidade de revisões constantes de crédito. Além disso, tendem a eliminar práticas discriminatórias de concessão de crédito, permitem uma melhor organização das informações, demonstram objetividade e consistência, são simples, de fácil interpretação e instalação, proporcionam maior eficiência no tratamento de dados e nos processos de concessão. Como desvantagens, podem-se destacar a degradação com o passar do tempo, caso a população a ser aplicado o modelo seja divergente da população original quando do seu desenvolvimento, o excesso de confiança dos usuários, e a falta de dados e informações causam problemas na sua utilização (ALTMAN; SAUNDERS, 1998; PARKINSON; OCHS, 1998).

Importante evidenciar que mesmo com os investimentos e os esforços direcionados à concessão de créditos de qualidade, muitos clientes não conseguem, por diferentes motivos, honrar com o compromisso de pagamento firmado dentro do prazo estabelecido, tornamdo-se, assim, inadimplente.

2.2 Inadimplência

Inadimplência, ou *default* (conceito que será considerado com o mesmo sentido de inadimplência, conforme terminologia utilizada na literatura internacional), "é o descumprimento de um contrato ou de qualquer de suas condições" (Dicionário MICHAELIS). Diante da concessão de crédito de um banco para um cliente, significa o não pagamento da quantia acordada dentro do prazo acordado.

Annibal (2009) considera que é difícil obter um consenso para uma definição operacional de inadimplência, pois os objetivos das análises podem ser conflitantes. Todavia, o autor explicita algumas definições existentes na literatura, como a definida por Westgaard e Wijst (2001), onde afirmam que entrar em inadimplência "é fracassar em pagar uma quantia devida a um banco", e a descrita por Bessis (1998), conforme abaixo:

> Considera-se ter ocorrido *default* em relação a um devedor específico quando um ou ambos os eventos seguintes tenham acontecido:
>
> • O banco considera improvável que o devedor pague na totalidade suas obrigações ao conglomerado financeiro sem que este tenha que recorrer a ações tais como a realização de garantias (se possuir);
>
> • O devedor está atrasado em mais de 90 dias em alguma obrigação material com o conglomerado financeiro. Saques a descoberto são considerados como operações em atraso quando o cliente infringir um limite recomendado ou tenha lhe sido recomendado um limite menor que a dívida atual.

Um título é considerado inadimplente quando: não há pagamento do principal ou de juros; o pagamento é efetivado após o vencimento; se a empresa devedora pedir concordata ou quando sua falência é decretada.

Ainda segundo Annibal (2009), a prática mais comum de mercado é a utilização do prazo de 90 dias de atraso para a caracterização da inadimplência. Para o presente trabalho o cliente é considerado inadimplente conforme essa prática, ou seja, depois de constatado atraso de 90 dias ou mais em seu contrato.

Para melhor entendimento, os clientes não inadimplentes, ou seja, aqueles que não descumpriram o pagamento de suas obrigações por mais de 90 dias serão identificados como adimplentes. Assim, adimplente é aquele "que cumpre suas obrigações contratuais no prazo certo" (Dicionário MICHAELIS)

Uma definição importante para o desenvolvimento de uma medida de inadimplência é a unidade de referência de crédito, que determina se a inadimplência é medida através da quantidade de empréstimos (contratos) de crédito ou de clientes que se tornaram inadimplentes, ou ainda pelos valores das dívidas que eles representam ou possuem. Assim, mensura-se a inadimplência de acordo com a quantidade de títulos, ou seja, se apenas um título de crédito de um conjunto de cem tornar-se inadimplente, a incidência de inadimplência é de apenas 1%. Por outro lado, se for adotada os valores dos títulos como referência, a participação deste título no valor agregado do conjunto de títulos pode ser muito diferente de 1% (ANNIBAL, 2009).

A Tabela 1 apresenta alguns modelos de escore de crédito encontrados na revisão da literatura referente a análise de risco de crédito e inadimplência.

Tabela 1 – Sistematização dos estudos a respeito de modelos de escore de crédito

Modelos de escore de crédito	**Pesquisa**
Desenvolvimento de modelos que predizem a probabilidade de ocorrência de inadimplência de crédito.	Barry, Escalante, e Ellinger (2002); Roszbach e Jacobson (2003); Katchova e Barry (2005); Lopez (1999) e Stein (2003).
Abordam modelos de escore de crédito existentes e propõem um método para estimar o risco de carteira de crédito.	Jacobson e Roszbach (2003).
Abordam a distância-padrão para determinar o *Value at Risk* (VaR) para uma amostra de carteiras de crédito de agricultores.	Katchova e Barry (2005).
Tratam da relação entre a probabilidade de inadimplência e a perda no caso de sua ocorrência.	Frye (2000); Altman *et al. (*2005); Acharya *et al.* (2007); Bade *et al.* (2011).
Diferentes modelos para lidar com a probabilidade de inadimplência de pessoas físicas e empresas.	Lane, Looney e Wansley (1986); Banasik, Crook, e Thomas (1999); Stepanova e Thomas (2001); Stepanova e Thomas (2002); Balzarotti , Falkenheim e Powell (2002); Andreeva (2006); Bellotti e Crook (2009); Jimenez *et al.* (2008); e Ioannidou, Ongena e Peydro (2008).

Fonte: Elaborado pelo autor.

2.3 Análise Discriminante – Método de Fisher

Segundo Huang *et al.* (2007), de acordo com a associação de grupos estabelecida nas amostras, a função discriminante é calculada para distinguir os agregados, minimizando a variabilidade dentro do agrupamento, enquanto maximiza a variabilidade entre os aglomerados, isto é, minimiza o erro de classificação. Existem Muitos métodos de discriminação disponíveis, entre eles o de Fisher, que é utilizado amplamente devido à sua não restrição à variação da distribuição (HUANG *et al*., 2007).

Para Regazzi (2000) a realização da discriminação entre dois ou mais grupos, visando classificação ulterior, foi abordado inicialmente por Fisher (1936), e consiste em obter funções matemáticas capazes de classificar um individuo qualquer em um dos diferentes grupos, com base em medidas de um número de características, buscando minimizar a probabilidade de um indivíduo ser classificado erroneamente em determinado grupo, quando realmente deveria pertencer a outro grupo. Seu objetivo é classificar observações desconhecidas e verificar quais variáveis são as mais importantes para a discriminação entre os grupos (ANDERSON; FARRAR; THOMAS, 2009). Análise discriminante é um método típico para a extração de recursos e redução de dimensionalidade, onde se encontra uma transformação linear que maximiza a dispersão entre classes e minimiza a dispersão dentro de cada classe para conseguir a separabilidade máxima entre as classes (ZENG *et al.*, 2010).

Para Hair *et al.* (2005), a AD é a técnica multivariada adequada para estudar problemas em que a variável estatística (combinação linear de variáveis com pesos determinados empiricamente) é dicotômica e, portanto, não métrica. É indicada para construir modelos de previsão de inadimplência, cujo objetivo principal é a classificação de um cliente solicitante de crédito em um determinado grupo, nesse caso, de provável adimplente ou inadimplente.

Uma vez que a associação de agrupamento é estabelecida, a AD, como um método de reconhecimento de padrões supervisionado, é aplicada para prever a adesão a determinado grupo para novos processos cuja associação é indeterminada (HUANG *et al.*, 2007).

Para Jonhnson e Wichern (2007), os objetivos imediatos de discriminação e classificação na AD são dois, conforme abaixo:

- Descrever, de forma gráfica ou algebricamente, as características diferenciais de observações de várias populações conhecidas. Tenta-se encontrar

discriminante cujos valores numéricos são tais que os grupos são separados tanto quanto possível;

- Para classificar observações em dois ou mais grupos rotulados. A ênfase está na derivação de uma regra que pode ser usada para otimizar novos objetos (indivíduos) nos grupos marcados.

Considerando duas classes (populações) π_1 e π_2, os objetos ou observações são ordinariamente separados ou classificados com base nas medidas de associação à variável $\underline{X}$, vetor aleatório de característica das populações. Os valores observados de $\underline{X}$ diferem de uma classe para outra.

Se os valores de $\underline{X}$ não forem muito diferentes dos objetos em π_1 e π_2, as classes serão indistinguíveis e novos objetos poderiam ser designados aleatoriamente a qualquer uma das classes. Estas duas populações podem ser descritas pelas respectivas FDP $f_1(\underline{x})$ e $f_2(\underline{x})$, e consequentemente, pode-se falar na designação de observações às populações.

Fisher propôs transformar as observações multivariadas $\underline{X}$ em observações univariadas Y tal que Y são obtidas a partir das populações π_1 e π_2, e são o mais distante possível. Sugeriu tomar combinações lineares dos componentes de $\underline{X}$ para criar as variáveis Y (JOHNSON; WICHERN, 2007).

Assim, supõe-se π_i uma população por um vetor aleatório $\underline{X}$ tal que :

$$\mathrm{E}[\underline{X} \mid \pi_i] = \underline{\mu}_i \text{ e } \mathrm{Var}(\underline{X} \mid \pi_i) = \Sigma_i,\ i = 1,2.$$

Seja $\underline{a}$ um vetor em R^{p} de constantes fixadas e definida por $Y = \underline{a}^T \underline{X}$.

Então, sob a i-ésima população temos:

$$\mathrm{E}[\mathrm{Y} \mid \pi_i] = \mu_{iY} = \underline{a}^T \underline{\mu}_i \text{ e } \mathrm{Var}(\mathrm{Y} \mid \pi_i) = \underline{a}^T \Sigma_i \underline{a}^T,\ i = 1,2.$$

Finalmente, assume-se que as populações π_1 e π_2 têm a mesma covariância Σ, sendo as matrizes de covariâncias Σ_1 e Σ_2 iguais nas duas populações: $\Sigma_1 = \Sigma_2 = \Sigma$.

A matriz de covariância comum Σ é estimada pela matriz combinada S_c definida por:

$$S_c = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] \cdot S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] \cdot S_2,$$

em que:

- S_c é o estimador da matriz comum de covariâncias Σ;
- n_1 é o número de observações da população π_1;
- n_2 é o número de observações da população π_2;
- S_1 é o estimador da matriz de covariâncias da população π_1;
- S_2 é o estimador da matriz de covariâncias da população π_2;

Ainda segundo Jonhnson e Wichern (2007), posteriormente, seleciona-se $\underline{a}$ tal que a distância quadrada entre μ_{1Y} e μ_{2Y} relativa à variabilidade dos Y seja a maior possível. Ou seja, seleciona-se $\underline{a}$ tal que a razão $\frac{(\mu_{1Y} - \mu_{2Y})^2}{Var(Y)} = \frac{[\underline{a}^T(\underline{\mu}_1 - \underline{\mu}_2)]^2}{\underline{a}^T \sum \underline{a}}$ seja máxima.

Fazendo $\underline{\omega} = \underline{\mu}_1 - \underline{\mu}_2$, tem-se que maximizar a razão $\frac{(\underline{a}^T \underline{\omega})^2}{\underline{a}^T \Sigma \underline{a}}$, cuja solução é dada, utilizando a desigualdade de Cauchy-Schwarz, por $\underline{a} \propto \Sigma^{-1} \underline{\omega}$, com valor máximo dado por $\underline{\omega}^T \Sigma^{-1} \underline{\omega}$.

Tomando $\underline{a} = \Sigma^{-1}$ e $\underline{\omega} = \Sigma^{-1}\left(\underline{\mu}_1 - \underline{\mu}_2\right)$, tem-se a função discriminante linear de Fisher dada por $Y = \left[\underline{\mu}_1 - \underline{\mu}_2\right]^T \cdot \Sigma^{-1} \cdot \underline{X}$.

O ponto médio entre duas populações univariadas $\underline{\mu}_1$ e $\underline{\mu}_2$ é:

$$m = \frac{1}{2}[\mu_1 - \mu_2]' \cdot \Sigma^{-1} \cdot [\mu_1 + \mu_2] \rightarrow m = \frac{1}{2}[Y_{\mu 1} + Y_{\mu 2}],$$

em que $Y_{\mu 1}$ e $Y_{\mu 2}$ são os centroides das duas populações consideradas.

Na prática, os parâmetros da população não são conhecidos, e a função discriminante linear amostral de Fisher é obtida substituindo-se os parâmetros μ_1, μ_2 e Σ pelas quantidades amostrais respectivas, $\bar{x}_1$, $\bar{x}_2$ e S_c (JONHNSON ; WICHERN, 2007). Assim, tem-se:

$Y = [\bar{x}_1 - \bar{x}_2]^T \cdot S_c^{-1} \cdot \underline{X}$, e

$$\overline{m} = \frac{1}{2}[\bar{x}_1 - \bar{x}_2]^T \cdot S_c^{-1} \cdot [\bar{x}_1 + \bar{x}_2] \rightarrow \overline{m} = \frac{1}{2}[Y_{\bar{x}_1} + Y_{\bar{x}_2}],$$

em que $Y_{\bar{x}1}$ e $Y_{\bar{x}_2}$ são os centroides das duas amostras consideradas.

A regra de classificação baseada na função discriminante linear amostral de Fisher é:

$$\begin{cases} \text{Alocar } \underline{X} \text{ em } \pi_1 \text{ se } Y = \left[\underline{\bar{x}}_1 - \underline{\bar{x}}_2\right]^T \cdot S_c^{-1} \cdot \underline{X} \geq \overline{m}\,; \\ \\ \text{Alocar } \underline{X} \text{ em } \pi_2 \text{ se } Y = \left[\underline{\bar{x}}_1 - \underline{\bar{x}}_2\right]^T \cdot S_c^{-1} \cdot \underline{X} < \overline{m}\,. \end{cases} \quad (1)$$

A análise discriminante de Fisher tem atraído muita atenção e tem sido aplicada em muitos problemas porque possui elegância conceitual e performance do estado da arte (XU *et al.*, 2004). Segundo Huang *et al.* (2007), a ideia básica do método de Fisher é reduzir um grande conjunto de medições múltiplas a compostos lineares das variáveis originais. Ao fazer isso, um conjunto de dados multivariados é transformado em um conjunto de dados univariados. O composto linear é chamado de função discriminante linear de Fisher, e os dados univariados são chamados de pontuação discriminante (*score discriminant*), que é uma projeção de cada ponto no eixo discriminante.

De acordo com Pasiouras e Tanna (2010), ao longo dos últimos 35 anos foram realizadas pesquisas significativas para desenvolver modelos de classificação, notadamente com AD, para a previsão de alvos de aquisição em diferentes países.

Já Doumpos *et al.* (2002) descrevem a complexidade do processo de avaliação de risco de crédito, que exigiu a construção de modelos de avaliação com base na abordagem de classificação, que pode ser usado por analistas financeiros e de crédito tanto como sistemas de avaliação de novos clientes que buscam empréstimos e/ou financiamentos. Uma ampla revisão da avaliação de risco de crédito ao longo das últimas duas décadas é apresentada por Altman e Saunders (1998).

Segundo Akkoç (2012), Durand (1941) foi o primeiro a utilizar AD em escore de crédito, procurando diferenças entre bons e maus grupos de credores. Desde então, essa técnica estatística tem sido utilizada em diversos estudos com aplicações em escore de crédito, como em Altman (1968), Martin (1977), Meyer e Pifer (1970), Sinkey (1975) e Oeste (1985).

A Tabela 2 apresenta alguns modelos de classificação de crédito encontrados na revisão da literatura referente à análise de risco de crédito e inadimplência utilizando-se de Análise Discriminante.

Tabela 2 – Sistematização dos estudos a respeito de modelos de classificação utilizando análise discriminante em escore de crédito

Modelos de classificação utilizando Análise Discriminante para escore de crédito	**Pesquisa**
Desenvolvimento de modelos de classificação para a previsão de alvos de aquisição nos EUA.	Espahbodi e Espahbodi (2003).
Desenvolvimento de modelos de classificação para a previsão de alvos de aquisição no Reino Unido.	Powell (2001).
Desenvolvimento de modelos de classificação para a previsão de alvos de aquisição no Canadá.	Belkaoui (1978).
Desenvolvimento de modelos de classificação para a previsão de alvos de aquisição na Grécia.	Slowinskiet *et al.* (1997).
Construção de ferramentas de triagem de clientes incluídos na carteira de crédito de um banco ou uma instituição de crédito.	Lane (1972); Altman *et al.* (1981); Grablowsky e Talley (1981); Srinivasan e Kim (1987); Srinivasan e Ruparel (1990).
Testam as Máquinas de Vetores de Suporte contra AD de um banco de dados de cartão de crédito de grande porte, e mostram que eles são competitivos e podem ser usados como a base de um método de seleção para descobrir as características que são mais importante na determinação de risco de inadimplência,	Crook *et al.* (2007).
Apresentam um modelo de escore de crédito mais preciso utilizando AD, com aplicação em um conjunto de dados em cartão de crédito.	Chuang e Lin (2009).

Fonte: Elaborado pelo autor.

2.4 Análise de Sobrevivência

A teoria de AS está relacionada com a análise de dados de tempo de ocorrência do evento de interesse, comumente censurados ou incompletos, sendo aplicada em diferentes áreas, como Medicina, Biologia e Engenharias. Para Colosimo e Giolo (2006) a técnica AS é uma das áreas da estatística que mais cresceu nas últimas décadas, evidenciando o número de aplicações em medicina.

Para Alves (2009), análise de sobrevivência consiste:

> ... na análise dos tempos de duração de um equipamento, indivíduo ou empresa no atual estado em que se encontram a fim de estimar as variáveis que possam explicar o comportamento destes tempos. Para tanto, os modelos estatísticos envolvidos em análise de sobrevivência são capazes de estimar a probabilidade de que estes elementos continuem em seus determinados estados. O resultado desta estimação consiste na chamada função de sobrevivência que corresponde à função das probabilidades, em tempos diferentes, de um elemento permanecer no atual estado em que se encontra. Sua utilização é de suma importância à medida que os resultados encontrados, correspondente à variável resposta e às variáveis explicativas, são capazes de auxiliar nas tomadas de decisão visando o aumento da probabilidade de sobrevivência.

O trabalho de Kiefer (1988) apresenta uma pesquisa introdutória e elucidativa sobre AS. Esse autor apresenta a função probabilidade condicional de falha, que representa o conceito central de AS, e consiste na estimação das probabilidades condicionais de determinado evento ocorrer em instantes distintos. Na análise de um evento a AS considera, além da probabilidade de ocorrência do evento em si, a probabilidade de que o mesmo evento ocorra supondo uma condição anterior.

Em AS a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse, denominado momento de falha (COLOSIMO; GIOLO, 2006). Contudo, frente às características dessa pesquisa, o evento de interesse é o tempo em que o cliente tomador do crédito se tornar inadimplente. A AS difere-se de outras análises na construção de modelos de previsão, pois utiliza os tempos de sobrevivência dos elementos em questão no estudo, onde a variável resposta dos modelos elaborados por esta técnica estatística corresponde a uma função dos tempos de sobrevivência, ou seja, a função de sobrevivência.

A AS também permite a investigação dos efeitos de mitigação onde a intensidade de falha varia com o tempo desde o início do empréstimo. O método também pode ser útil para a construção de modelos de risco de crédito, em conformidade com o Acordo de Basiléia II, considerando a modelagem do momento em que o cliente torna-se inadimplente, ao invés de considerar apenas uma classificação estruturada como mau ou bom pagador.

A delimitação do tempo em que o cliente se torna inadimplente permite a elaboração da variável tempo até a ocorrência desse evento, que por sua vez denomina-se tempo de sobrevivência. Este tempo tem fundamental importância em AS já que a variável resposta dos modelos existentes nesta análise corresponde a uma função do tempo de sobrevivência em questão. Cox e Oakes (1984), ao apresentar os conceitos da AS, definiram que qualquer

observação incompleta sobre o tempo até a ocorrência do evento de interesse é denominado censura.

Para Colosimo e Giolo (2006), a principal característica de dados de sobrevivência é a presença de censura, definida como a observação parcial da resposta. Nesse sentido, estudos que utilizam como ferramenta estatística a AS devem mencionar qual é a censura existente em relação aos tempos de duração, censura à direita, censura à esquerda, ou ambas.

Para Strapasson (2007), mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser analisados, pois mesmo incompletas, as observações censuradas fornecem informações sobre o tempo de vida do objeto de estudo, e sua omissão no cálculo de interesse pode acarretar conclusões viciadas**.**

Colosimo e Giolo (2006) definem três tipos de censuras:

- Censura do tipo I: ocorre em estudos que ao serem finalizados após um período pré-estabelecido de tempo registram, em seu término, indivíduos que ainda não apresentaram o evento de interesse;
- Censura do tipo II: ocorre em estudos os quais são finalizados após a ocorrência do evento de interesse em um número pré-estabelecido de indivíduos;

  Para as censuras acima, todos os indivíduos entram no estudo ao mesmo tempo.
- Censura do tipo aleatória: ocorre quando um indivíduo é retirado no decorrer do estudo sem ter ocorrido o evento de interesse.

Esses três mecanismos de censura apresentados são conhecidos por censura à direita, pois o tempo de ocorrência do evento de interesse está à direita do tempo registrado. É o tipo mais frequentemente encontrado, porém pode ocorrer a censura à esquerda e a censura intervalar. A censura à esquerda ocorre se o evento de interesse já aconteceu quando o indivíduo foi observado.

Tempos exatos bem como tempos censurados à direita e à esquerda são casos especiais de dados de sobrevivência intervalar. Na censura intervalar não se sabe o tempo exato de ocorrência do evento de interesse, sabe-se apenas que ele ocorreu dentro de um intervalo.

A função de sobrevivência é dada em termos probabilísticos como exposto a seguir (COLOSIMO; GIOLO, 2006):

$$S(t) = P(T > t),$$

em que $S(t)$ corresponde à função de sobrevivência que é definida como a probabilidade de uma observação não falhar antes do tempo *t*, ou seja, a probabilidade de uma observação durar um período de tempo *T* maior que *t*.

Isto é, a função de sobrevivência é a probabilidade de que evento de interesse (ocorrência da inadimplência, por exemplo) ocorra após o tempo especificado (t). Normalmente assume-se $S(0) = 1$, embora possa ser inferior a 1 se houver a possibilidade de morte imediata ou fracasso.

Ainda segundo Colosimo e Giolo (2006), a função de sobrevivência não deve ser maior que $S(u) \leq S(t)$ se $u \geq t$. Esta propriedade segue diretamente porque $T > u$ implica $T > t$. Isso reflete a noção de que a sobrevivência até uma idade mais "avançada" só é possível se todas as idades mais jovens são atingidas. Dada essa propriedade, a função de distribuição e a função densidade são bem definidas. A função de sobrevivência é geralmente assumida que se aproximam de zero à medida que aumenta a idade, sem limite, ou seja, $S(t) \rightarrow 0$ quando $t \rightarrow \infty$, embora o limite pudesse ser maior do que zero, se a vida eterna é possível.

Com base nesse conceito, a função de distribuição acumulada é definida como a probabilidade de uma observação não durar até o tempo *t*, ou seja:

$$F(t) = 1 - S(t) .$$

Se F é diferenciável, então a derivada, o qual é a função densidade, é convencionalmente denominada por *f*,

$$f(t) = F'(t) = \frac{d}{dt} F(t) .$$

A função de sobrevivência pode ser expressa em termos de distribuição e função densidade

$$S(t) = P(T \geq t) = \int_t^{\infty} f(u)du = 1 - F(t) .$$

Do mesmo modo, a função de densidade do evento de sobrevivência pode ser definida como

$$s(t) = S'(t) = \frac{d}{dt} S(t) = \frac{d}{dt} \int_t^{\infty} f(u)du = \frac{d}{dt}[1 - F(t)] = -f(t) .$$

Para a análise proposta é fundamental que seja definida a função de probabilidade condicional de falha, chamada ainda de função de risco ou *hazard*, denotada por $\lambda(t)$, também apresentada como a variável resposta de modelos elaborados com base na AS, que consiste na probabilidade de certo evento ocorrer e na probabilidade de sua ocorrência dado que o mesmo evento não ocorreu até o instante *t*. Essa função é assim definida:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}.$$

Essa função de risco também pode ser descrita como a razão entre a função densidade de probabilidade (FDP) e a própria função de sobrevivência, além de ser o resultado da derivação do logaritmo neperiano da função de sobrevivência, conforme abaixo:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\ln(S(t)).$$

Agora, é necessário determinar a função acumulada da função de risco para obter a função de sobrevivência. Assim, a relação entre a função de risco acumulado, denotada por $\Lambda(t)$ e a função sobrevivência é dada por:

$$\Lambda(t) = \int_0^t \lambda(u)du = -\ln(S(t)).$$

E, da mesma forma, a função de sobrevivência é estimada por:

$$S(t) = \exp\{-\Lambda(t)\} = \exp\{-\int_0^t \lambda(u)du\}.$$

2.4.1 Distribuição Weibull

Para uma variável aleatória *T* com distribuição de Weibull, tem-se a função de densidade de probabilidade dada por:

$$f(t;\gamma,\alpha) = \alpha \cdot \gamma \cdot t^{\alpha-1} \exp\{-\gamma \cdot t^{\alpha}\},$$

em que γ, o parâmetro de forma, e α, o de escala, são ambos positivos. O parâmetro α tem a mesma unidade de medida de *t* e γ não tem unidade.

Para esta distribuição, a função de sobrevivência é

$$S_W(t) = \exp\{-\gamma \cdot t^{\alpha}\}. \qquad (2)$$

para $t \geq 0$, α e $\gamma > 0$. Observe que, quando $\gamma = 1$, tem-se a distribuição exponencial, um caso particular da distribuição de Weibull, sendo a função de risco dada por $\lambda(t) = \gamma \cdot t^{\alpha-1}$.

A distribuição Weibull é utilizada nesse trabalho visto que ela vem sendo frequentemente citada na literatura e por apresentar uma grande variedade de formas. Segundo Ortega *et al.* (2011) nas últimas décadas foram desenvolvidas diferentes novas distribuições para fração de cura com base em extensões da distribuição Weibull, entre elas: Mudholkar *et al.* (1995) introduziram uma distribuição Weibull Exponenciada, Xie e Lai (1995) apresentaram uma distribuição aditiva Weibull, Lai *et al.* (2003) propuseram a distribuição de Weibull modificada, Famoye *et al.* (2005) e Wahed (2006) propuseram um método geral de construção de famílias extensas de distribuições de uma distribuição contínua a partir da linha de base, Lee *et al.* (2007) estudaram distribuição Weibull Beta, Carrasco *et al.* (2008) definiram a distribuição Weibull Modificada Generalizada, e Wahed *et al.* (2009) apresentou uma generalização da distribuição Weibull com a aplicação de um conjunto de dados de câncer de mama.

Em modelos tradicionais de AS, assume-se que em dado momento o evento de interesse ocorre para todos os indivíduos observados após determinado tempo, ou seja, todos os indivíduos são suscetíveis ao evento durante a realização do estudo. No entanto, em determinadas análises, alguns indivíduos podem nunca apresentar o evento de interesse, pois estão curados ou são considerados imunes ao evento. No caso deste trabalho, similarmente, muitos clientes não se tornam inadimplentes durante a vida útil do empréstimo.

Modelar estes dados ignorando a existência dessa parcela de “curados” (os clientes que não se tornam inadimplentes) na população de estudo pode conduzir o trabalho a conclusões distorcidas. Quando esta característica é incorporada ao modelo, pode-se saber, por exemplo, quais variáveis influenciam na proporção de adimplentes. Para esse trabalho, onde existe uma fração de clientes que não se tornam inadimplentes dentro do prazo do empréstimo tomado, é pertinente que se utilize da AS considerando a fração de clientes adimplentes (ou curados).

Em AS os modelos paramétricos e não paramétricos tradicionais assumem que a fração de cura é zero ao longo do tempo. Porém, incluir a possibilidade de cura tem sido reconhecido em diversas aplicações (YAKOVLEV; TSODIKOV, 1996).

2.4.2 Modelo de sobrevivência com fração de cura: um modelo de sobrevivência para análise de escore de crédito.

Os modelos com fração de cura são uma extensão para o modelo de sobrevivência padrão. Foram recentemente utilizados em medicina para modelos de sobrevivência em ensaios clínicos de câncer, em termos de duas subpopulações distintas (Sy e Taylor, 2000 *apud* Tong *et al.*,2012). Em uma subpopulação, os pacientes não são susceptíveis (curados) e livres de câncer após um tratamento, enquanto o outro contém uma subpopulação de pacientes que são susceptíveis (não curados) e, com o tempo, irão apresentar reincidência do câncer.

Modelos que acomodam fração de cura têm sido amplamente desenvolvidos e aplicados na literatura. A teoria inicial decorre do modelo de mistura introduzido por Boag (1949) e Berkson e Gage (1952), onde se presume que certa percentagem da população de interesse é curada, no sentido de que não apresentam o evento de interesse durante um longo período de tempo, e podem ser vistos como “imunes” ou curados (MALLER; ZHOU, 1996). Da teoria inicial decorre o modelo de mistura introduzida no campo da estatística médica por Farewell (1982), onde o modelo geral incorpora dois componentes, um para prever a incidência dos indivíduos suscetíveis à falha, e um modelo de latência para prever o tempo de sobrevivência de indivíduos condicionais a se tornarem suscetíveis à falha. O componente de incidência é essencialmente um modelo de classificação binária (regressão logística, por exemplo).

Para a parte de latência, o modelo de mistura proposto por Farewell (1982) utiliza um modelo de sobrevivência paramétrica baseada na distribuição de Weibull para calcular tempos de sobrevivência. No entanto, durante a última década, modelos semiparamétricos também têm sido desenvolvidos, e, proporcionam maior flexibilidade, uma vez que não requerem uma distribuição de sobrevivência específica para o componente de incidência (Tong *et al.*, 2012).

A maior referência acerca de modelos com fração de cura é Maller e Zhou (1996). Perdoná e Louzada-Neto (2011) introduziram um modelo de fração de cura que generaliza várias distribuições habituais de fração de cura. Recentemente, Cancho *et al.* (2012) introduziram o modelo com fração de cura *Birnbaum-Saunders* geométrica para analisar dados de sobrevivência na presença de uma fração de cura. Seus interesses práticos estão bem estabelecidos em ciências biomédicas, criminologia e de engenharia como um método de dados de modelagem de tempo para o evento de interesse.

Tabela 3 – Sistematização dos estudos a respeito de modelos de análise de sobrevivência com fração de cura.

Modelos de AS com fração de cura - abordagem teórica e aplicações em escore de crédito.	**Pesquisa**
Estudo de modelos de longa duração adotando como embasamento conceitos de processos estocásticos.	Zaider *et al.* (2001)
Modificam o modelo Weibull-exponenciado possibilitando acomodar uma fração de cura.	Cancho e Bolfarine (2001)
Discutem a estimação por máxima verossimilhança em um modelo semiparamétrico.	Chen e Ibrahim (2001)
Desenvolve uma fórmula de generalização de modelos de longa duração baseados na distribuição de Weibull.	Perdoná (2006)
Verificam a flexibilidade na fração de cura utilizando um modelo bayesiano.	Chi e Ibrahim (2007)
Propõem um modelo semiparamétrico dinâmico bayesiano.	Kim *et al.* (2007)
Dedicam-se a um modelo com fração de cura com erros de medição nas covariáveis.	Mizoi *et al.* (2007)
Abordam um modelo com estrutura de riscos proporcionais para dados correlacionados.	Peng *et al.* (2007)
Tratam de um modelo com fração de cura em situações em que a exposição a um fator de risco ocorre em diversas ocasiões.	Tournoud e Ecochard (2007)
Aplicam técnicas de avaliação de influência local a um modelo Weibull com fração de cura.	Cancho *et al.* (2008)
Discutem estimação não paramétrica da fração de cura na presença de censura intervalar.	Sen e Tan (2008)
Desenvolvem um modelo para o tempo até que um indivíduo deixa de ser cliente de um banco.	Hoggart e Griffin (2001)
Estudo sobre previsão de inadimplência em empréstimos imobiliários corporativos.	Yildirim (2008)
Sugere uma abordagem de função de risco para estimar o risco de crédito em carteiras de empréstimos ao consumidor em vez de pontuação de crédito em um nível de conta específica	Thomas (2009)
Demonstram o desempenho superior de modelos de cura sobre previsão de falência corporativa.	Topaloglu e Yildirim (2009)
Propõem uma aplicação e demonstram o benefício de pontuação de candidatos a crédito por meio de AS e pelo modelo de riscos proporcionais de Cox.	Banasik e Crook (2010)
Discutem a aplicação da modelagem de fração de cura para prever o tempo para a inadimplência em uma carteira de crédito pessoal no Reino Unido.	Tong *et al.* (2012)

Fonte: Elaborado pelo autor.

Na Tabela 3 são apresentados alguns modelos de sobrevivência com fração de cura encontrados na revisão da literatura contemplando uma abordagem teórica e aplicações em carteiras de crédito de instituições financeiras.

Nesta pesquisa, com foco em modelagem de escore de crédito, distribuições com fração de cura são particularmente muito úteis, pois uma proporção substancial de observações (clientes) não é observada, simplesmente porque não experimentam o evento de interesse (inadimplência) durante o período de vida da linha de crédito (empréstimo). A fim de alinhar a terminologia para a modelagem de escore de crédito, cura é denotada por adimplência, e o termo fração de cura denominado de fração de adimplência, ou fração de adimplentes.

Considerando a amostra retirada da instituição financeira objeto de estudo a partir de uma carteira de crédito pessoal, observa-se que uma quantidade substancial de clientes do banco não apresenta inadimplência durante o período de empréstimo.

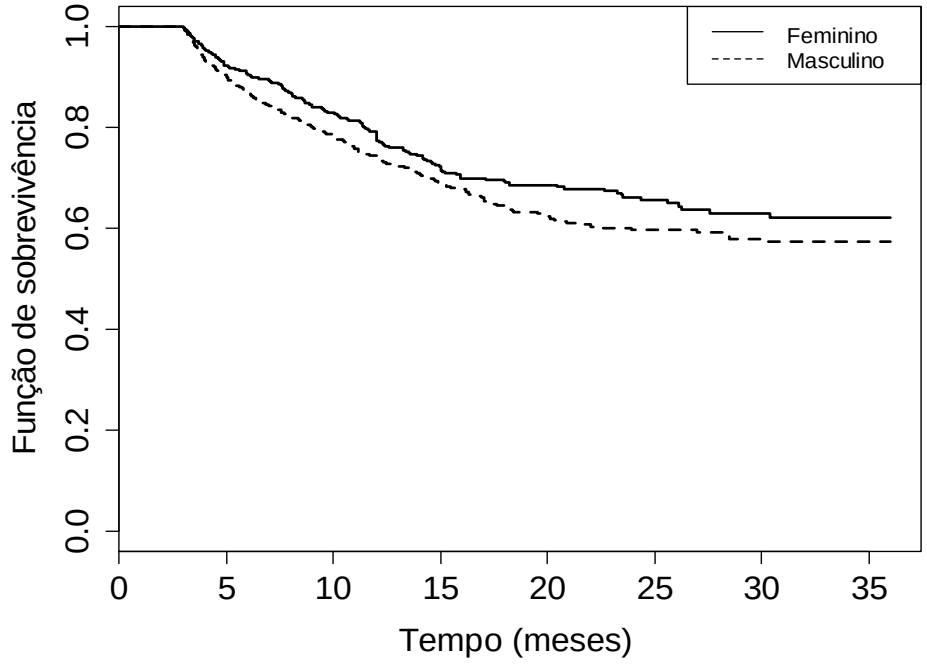

Figura 3 – Curva de sobrevida de Kaplan-Meier, estratificada por sexo.

Fonte: Elaborado pelo autor.

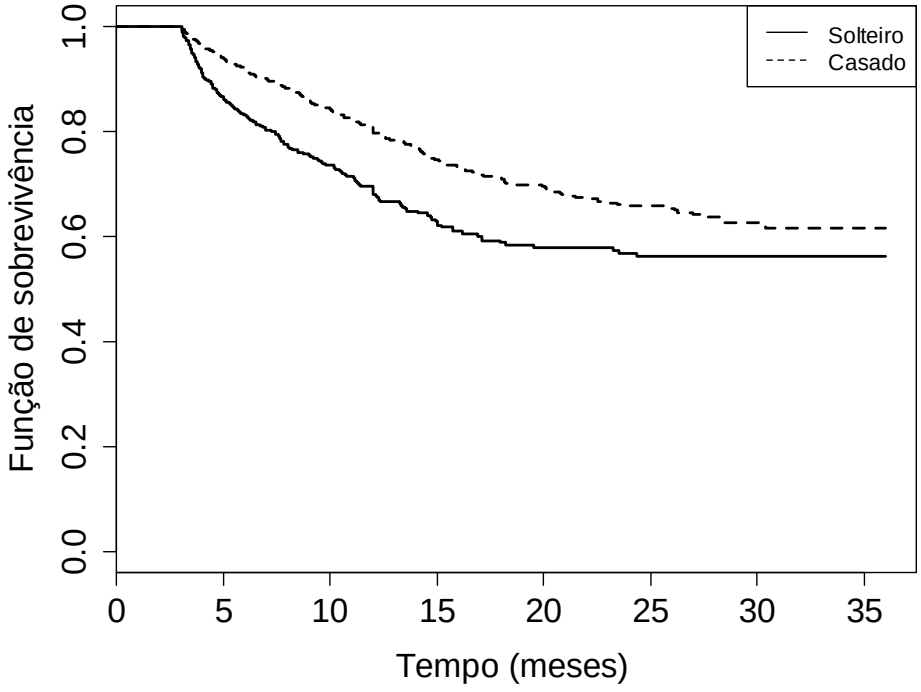

Figura 4 – Curva de sobrevida de Kaplan-Meier, estratificada por estado civil.

Fonte: Elaborado pelo autor.

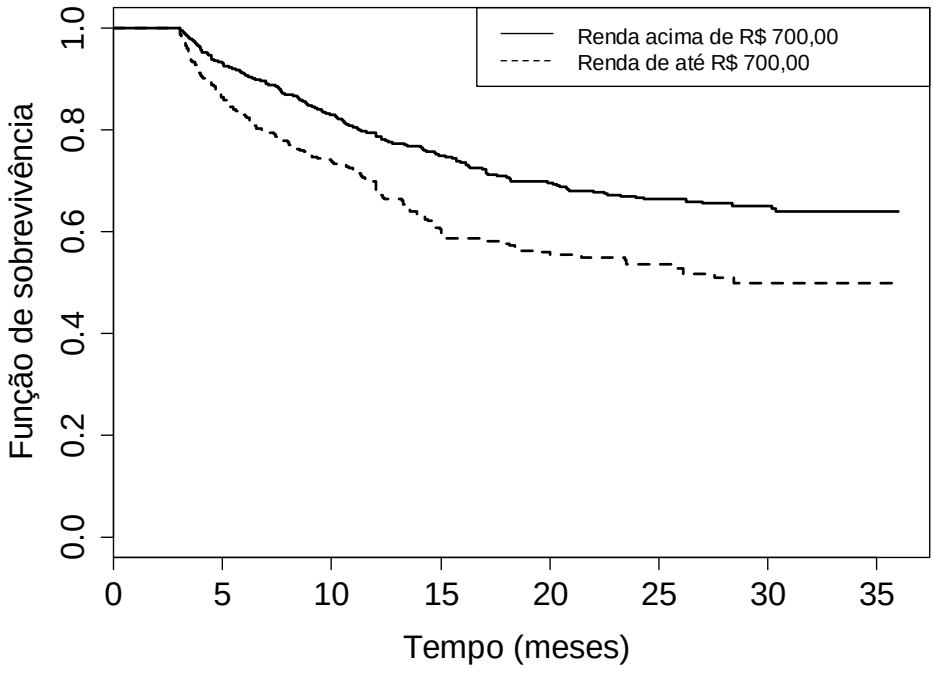

Figura 5 – Curva de sobrevida de Kaplan-Meier, estratificada por renda.

Fonte: Elaborado pelo autor.

A estimativa de Kaplan-Meier da função de sobrevivência dada na Figura 3, estratificada por sexo, demonstra a presença de um planalto em níveis acima de 0,5 para ambos os sexos a partir do 30° mês, e indicam a presença de fração de clientes adimplentes na amostra. Além disso, nota-se que clientes do sexo feminino apresentam menor probabilidade de tornarem-se inadimplentes no decorrer do empréstimo em relação aos clientes do sexo masculino. Na Figura 4, estratificada por estado civil, observa-se também a presença de um planalto em níveis acima de 0,5 tanto para clientes solteiros, a partir do 25° mês, como para clientes casados, a partir do 30° mês, indicando a presença de fração de clientes adimplentes na amostra. Nota-se, também, que clientes casados apresentam menor probabilidade de tornarem-se inadimplentes no decorrer do empréstimo em relação aos clientes solteiros. Já na Figura 5, estratificada por renda, observa-se a presença de um planalto em níveis acima de 0,5 para clientes com renda de até R$ 700,00, a partir do 30° mês, e a presença de um planalto em níveis acima de 0,6 para clientes com renda acima de R$ 700,00, a partir do 30° mês, indicando a presença de fração de clientes adimplentes na amostra. Nota-se, também, que clientes com maior renda apresentam menor probabilidade de tornarem-se inadimplentes no decorrer do empréstimo. Vale ressaltar que o planalto no nível 1,0, entre o mês 0 e 3 nas três Figuras (3, 4 e 5), decorre da delimitação de o cliente ser considerado inadimplente somente depois de decorridos 90 dias (3 meses) da não realização do pagamento da prestação mensal do empréstimo.

Com base no modelo de mistura padrão (BOAG, 1949; BERKSON; GAGE, 1952), que considera uma mistura de distribuições, onde uma representa o tempo de sobrevivência da população não curada (de inadimplentes) e outra dada por uma distribuição degenerada com tempos infinitos para os imunes, ou seja, a fração dos clientes que não se tornam inadimplentes durante o período de empréstimo (adimplentes), assume-se que uma fração p_0 $(0 < p_0 < 1)$ da população está curada, e o restante *1 - p_0*, não está curada. E seja *S(t)* a função de sobrevivência para os clientes inadimplentes, de tal modo que, $\lim_{t\to\infty} S(t) = 0$. Utilizando uma partição em imunes *A* (no caso os adimplentes) e não imunes *I* (no caso os inadimplentes), com $P(A) = p_0$ e $P(I) = 1 - p_0$ a função de sobrevivência para a população, denotada por $S_{pop}(A)$, é dada por:

$$S_{pop}(A) = P(T > t)$$

$$S_{pop}(A) = P(T > t \mid A)P(A) + P(T > t \mid I)P(I)$$

$$S_{pop}(A) = p_0 + (1 - p_0)S(t)\,.$$

em que $P(T > t \mid I) = S(t)$ denota a função de sobrevivência para a população de inadimplentes e $P(T > t \mid A) = 1$ denota a função de sobrevivência para a população de adimplentes, para todo $t > 0$.

Uma desvantagem com essa modelagem é que é baseada na suposição de que apenas uma causa é responsável pela ocorrência do evento de interesse (inadimplência). No entanto, na medida em que o evento de interesse acontece, pode ser causado por diferentes motivos, como o esquecimento de realização do pagamento, a perda do emprego, o pagamento mensal comprometer excessiva percentagem da renda, a falência pessoal, entre outras causas. Além disso, estas causas são, em geral, latentes no sentido em que não há informação sobre o que foi responsável pela ocorrência do evento de interesse.

Nesse contexto, a literatura sobre distribuições que acomoda diferentes causas latentes é rica, onde se pode considerar o livro de Ibrahim *et al.* (2001), e os artigos de Tsodikov *et al*. (2003), Cancho e Bolfarine (2001), Ortega *et al*. (2008), Rodrigues *et al.* (2009), Ortega *et al.* (2009) e Cancho *et al.* (2009), apesar da aplicação em áreas diferentes.

Yin e Ibrahim (2005) publicam uma proposta que consiste em uma classe de modelos que naturalmente reúne uma família de funções de sobrevivência próprias e impróprias. Rodrigues *et al.* (2009) contribuem com um modelo unificado que inclui o modelo de mistura padrão e o modelo de tempo de promoção (YAKOVLEV, 1994) como dois casos especiais. Uma abordagem semiparamétrica que permite utilizar dados correlacionados no modelo de mistura padrão é discutida em Peng *et al.* (2007). Um teste para avaliar a suficiência do tempo de acompanhamento em uma ampla classe de modelos de fração de cura foi proposto por Klebanov e Yakovlev (2007).

Assim, é considerado o modelo Weibull Geométrico (WG) com fração de cura, obtido a partir da composição de uma distribuição de Weibull e uma Geométrica, e concebida dentro de um cenário latente de causas concorrentes, em que o evento de interesse (inadimplência) pode ser causado por diferentes mecanismos de ativação. De acordo com Cooner *et al.* (2007), as causas latentes são assumidas para formar uma sequência estocástica disposta, os quais induzem a ocorrência do evento de interesse por meio de um mecanismo de ativação subjacente. Em suma, mecanismo de ativação é o momento em que o cliente torna-se inadimplente. Se o tempo até a ocorrência da inadimplência é definido como uma variável aleatória dada pelo valor mínimo dos tempos de vida associado às causas latentes, o

mecanismo de primeira ativação (PA) é observado. Como mecanismo de contrapartida, se o tempo observado até a ocorrência da inadimplência é definido como uma variável aleatória dada pelo valor máximo dos tempos de vida associado às causas latentes, o mecanismo de última ativação (UA) é observado.

Propõe-se um tipo de mecanismo de ativação começando com o mecanismo de PA, passando por um mecanismo de ativação aleatório (AA), e terminando com o mecanismo de UA, mas assumindo que não há informações sobre qual motivo foi o responsável pelo tempo para a ocorrência de inadimplência individual. Uma vantagem para essa abordagem é que o modelo considerado pode explorar todos os mecanismos de ativação subjacentes da primeira para a última. No contexto de classificação de crédito, a dificuldade é visualizar se foi a primeira ou a última causa, ou até mesmo uma aleatória, ordenados por seu tempo de ocorrência, a responsável pela ocorrência de inadimplência. Entretanto, considerando-se essa modelagem pode-se ajustar o modelo WG com fração de clientes adimplentes considerando os três mecanismos de ativação e decidir pelo melhor frente a análise dos dados.

O modelo WG com fração de cura considerando é obtido como segue. Para um indivíduo (cliente) na população, seja M o número de causas não observáveis do evento de interesse (inadimplência) para este indivíduo. Suponha que M segue uma distribuição geométrica com parâmetro θ, $0<\theta<1$, a FDP é dada por

$$P(M=m)=\theta(1-\theta)^m,\ \ m=0,1,... \quad . \qquad (3)$$

O tempo para a *j-ésima* causa produzir o evento de interesse (tempos de evento latente) é denotada por Z_j, $j = 1,2,...,M$. Assume-se que, condicionalmente em M, os Z_j são diferentes com distribuição Weibull, e $Z_1,Z_2,...$ são independentes de M. O tempo de inadimplência observável é definido pela variável aleatória $Y = Z_{(R)}$, onde R depende de M, $Z_{(1)} \le Z_{(2)} \le ... \le Z_{(R)} \le ... \le Z_{(M)}$ são estatísticas de ordem e $Y=\infty$ se $M = 0$. Na área financeira, R pode ser interpretado como um fator de resistência relacionado com o número de possíveis causas latentes de um cliente resistir até a inadimplência. Se o evento de interesse ocorre (o cliente tornar-se inadimplente), então a variável aleatória Y, tempo de inadimplência, converte o valor da *R-ésima* estatística de ordem $Z_{(R)}$. Em outras palavras, como em Cooner *et al.* (2006) e Cooner *et al.* (2007), R causas de M são necessárias para produzir a inadimplência. O fator de resistência pode ser uma constante, uma função de M ou uma variável aleatória especificada por meio de uma distribuição condicional em M.

Adaptando a terminologia empregada em Cooner *et al.* (2006) e Cooner *et al.* (2007), o presente trabalho vai tratar três especificações para *R*. Primeiro assume-se *R* aleatoriamente, direcionando a um mecanismo de ativação aleatória (AA). Assim, pode-se analisar toda a gama de possíveis mecanismos de ativação. Em seguida, são considerados os dois casos extremos, fixando $R = 1$ para dirigir o mecanismo de PA e $R = M$ para direcionar ao mecanismo de UA. A questão prática é que causa é ativada por meio de um mecanismo de AA no sentido *R* das *M* causas são necessários para ativar a inadimplência, ou ele é ativado por meio de um mecanismo de PA no sentido de ocorrer a primeira causa e ativar o momento de inadimplência, ou ele é ativado por meio de um mecanismo de UA, no sentido de que é necessário o último motivo acontecer em ordem para ativar o momento de inadimplência.

Das condições do modelo tem-se que a função de sobrevivência populacional é dada por

$$S_{pop}(w) = P(Y \geq y) = P(M = 0) + P(Z_{(R)} > y,\ 1 \leq R \leq M)$$

$$= P(M = 0) + \sum_{k=1}^{\infty} \sum_{R=1}^{k} P(Z_{(R)} > y \mid R, M = k) P(R \mid M = k) P(M = k)\ , \qquad (4)$$

em que

$$P(Z_{(R)} > y \mid R, M = k) = \sum_{i=o}^{R-1} \binom{k}{i} (F_W(y))^i (S_W(y))^{k-i}\ . \qquad (5)$$

A função (5) mostra a FDP de uma distribuição binomial, com ensaios *k* e probabilidade de sucesso $F_W(y) = 1 - S_W(y)$. Se $R = 1$, então, $P(Z_{(1)} > y \mid R, M = k) = S_W(y)^k$, que inclui os modelos propostos por Tsodikov *et al.*, (2003).

Em uma primeira configuração, considera-se o mecanismo de AA com $M \geq 1$ e assume-se que a distribuição condicional de *R* dada $M=k$ com $\{1,...,k\}$, tem-se a função de sobrevivência de *Y* dada por

$$S_{pop}(y) = P(M = 0) + \sum_{k=1}^{\infty} \left\{ \sum_{R=o}^{k} (k - R) B(R; k, F_W(y)) \right\} \frac{1}{k} \theta (1-\theta)^k$$

$$= P(M = 0) + S_W(y) \sum_{k=1}^{\infty} \theta (1-\theta)^k$$

$$= \theta + (1-\theta) S_W(y), \qquad (6)$$

em que $B(x;k,F_W(y)) = P(X = x)$ e $Z \sim Binomial(k, F(y))$. Observa-se o $S_{pop}(y)$ em (6) podendo ser visto como um modelo determinado na mistura de (2) com fração de clientes adimplentes $p_o = P(M = 0) = \lim_{y\to\infty} S_{pop}(y) = \theta$. De (6) a função densidade é dada por:

$$f_{pop}(y) = -S'_{pop}(y) = (1-\theta) f_W(y), \qquad (7)$$

onde $f_W(y)$ denota a FDP Weibull dada em (2). Além disso, a função de risco correspondente é

$$h_{pop}(y) = \frac{(1-\theta) f_W(y)}{\theta + (1-\theta) S_W(y)}.$$

Como uma segunda configuração, considera-se o mecanismo de PA supondo que a inadimplência acontece devido à primeira dessas causas latentes (eventos). Portanto, neste caso, *R = 1*, o tempo para o evento é $Y = Z_{(1)} = \min\{Z_{(1)}...Z_{(M)}\}$. De (4) e (5) a função sobrevivência de *Y* é dada por:

$$\begin{aligned} S_{pop} &= P(M = 0) + \sum_{k=1}^{\infty} S_W(t)^k P(M = k) \\ &= \theta + \theta \sum_{k=1}^{\infty} [S_W(y)(1-\theta)]^k \\ &= \frac{\theta}{1-(1-\theta)S_W(y)}. \end{aligned} \qquad (8)$$

A fração de clientes adimplentes é dada por $p_0 = \theta$. A função densidade associada a (8) é dada por

$$f_{pop}(y) = \theta(1-\theta) f_W(y)[(1-(1-\theta)S_W(y)]^{-2}, \qquad (9)$$

com função de risco

$$h_{pop}(y) = (1-\theta) f_W(y)[(1-(1-\theta)S_W(y)]^{-1}.$$

Como uma terceira configuração, considera-se o mecanismo de UA, assumindo que a inadimplência ocorre após todas as causas *M* terem sido atingidas. Assim, *R = M* e o tempo observado até a ocorrência da inadimplência é $Y = Z_{(M)} = \max\{Z_1,...,Z_M\}$. De (4) e (5) tem-se que a função de sobrevivência de *Y* é dada por:

$$S_{pop}(y) = P(M=0) + \sum_{k=1}^{\infty}[1-F_W(t)^k]P(M=k)$$

$$= 1-\theta\sum_{k=1}^{\infty}[F_W(y)(1-\theta)]^k$$

$$= 1+\theta-\frac{\theta}{1-(1-\theta)F_W(y)}. \qquad (10)$$

de modo que a fração de adimplência é $p_o = \theta$. A função de sobrevivência em (10) conduz para a função densidade

$$f_{WG}(y) = \theta(1-\theta)\, f_W(y)[1-(1-\theta)F_W(y)]^{-2}, \qquad (11)$$

com função de risco

$$h_{WG}(y) = \frac{\theta(1-\theta)\, f_W(y)}{[1-(1-\theta)F_W(y)]^2\left[1+\theta-\frac{\theta}{1-(1-\theta)F_W(y)}\right]}.$$

Verifica-se que $f_{WG}(y)$ e $h_{WG}(y)$ são funções impróprias decorrentes das funções de sobrevivência $S_{pop}(y)$ impróprias. Além disso, os modelos diferem por suas funções de densidade e de risco, porém a fração de adimplentes é a mesma, e também a relação de ordem entre as funções de sobrevivência.

A Figura 6 mostra os comportamentos distintos das funções de sobrevivência em função do tempo *y* (em anos) e ilustram a flexibilidade concedida pelo modelo. A curva assintótica com patamar acima de zero (painel esquerdo) indica a existência de clientes adimplentes. Além disso, observa-se que o mecanismo de PA apresenta melhores resultados em relação aos demais mecanismos de ativação.

A função de sobrevivência (própria) para a população de adimplentes, isto é, para o tempo observado até a ocorrência da inadimplência, denotada por $S_{WG}(y)$, é calculada por $S_{WG}(y) = P(Y \geq y \mid M \geq 1)$.

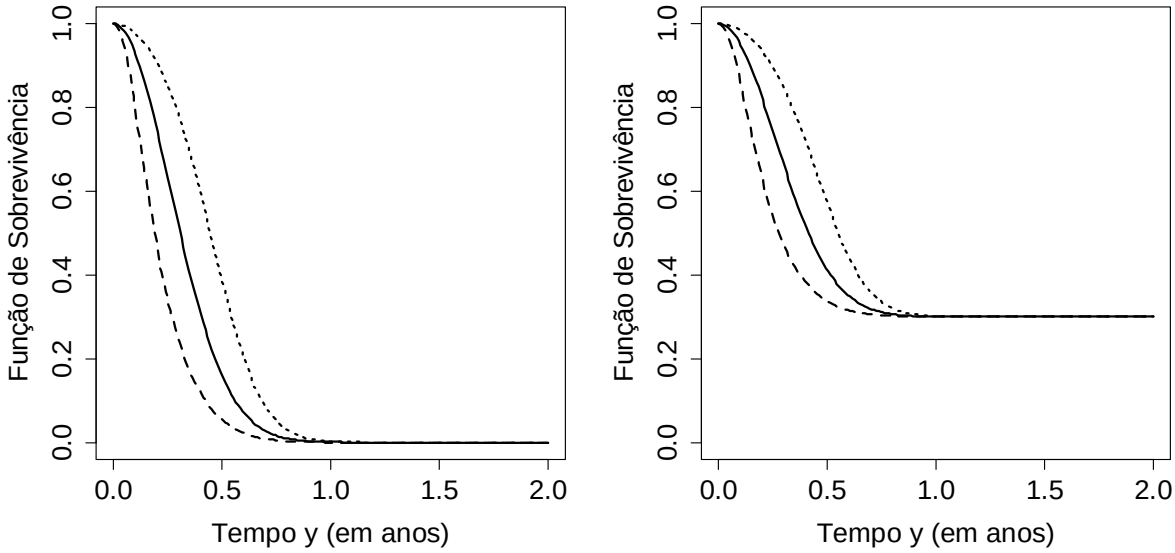

Figura 6: Função de sobrevivência dos clientes adimplentes (S_{WG}), à esquerda, e função de sobrevivência populacional (todos os clientes) (S_{pop}), à direita, com $\theta = 0.3$, $\alpha = 2.0$ e $\lambda = 2.0$, sob os diferentes mecanismos de ativação (PA: pontilhada; AA: sólida; UA: tracejada).

Fonte: Elaborado pelo autor.

Para o mecanismo de AA, S_{WG} é a mesma distribuição de Z_j, isto é, $S_{WG}(y)=S_W(\gamma)$, com $S_W(\gamma)$ dado em (2).

Para o mecanismo de PA, a função de sobrevivência para a população de adimplentes é dada por

$$S_{WG}(y) = \frac{\theta \exp\{-\lambda y^{\alpha}\}}{1-(1-\theta)\exp\{-\lambda y^{\alpha}\}}, \quad y > 0. \tag{12}$$

Nota-se que $S_{WG}(0)=1$ e $S_{WG}(\infty)=0$, então (12) é uma função de sobrevivência própria.

A FDP para o mecanismo de PA é dado por

$$f_{WG}(y) = \frac{\theta\lambda\alpha y^{\alpha-1}\exp\{-\lambda y^{\alpha}]}{[1-(1-\theta)\exp\{-\lambda y^{\alpha}\}]^2}, \quad y > 0. \tag{13}$$

Na Figura 7 podem-se observar a FDP do modelo WG com mecanismo de PA para alguns valores fixados de α, θ e λ, conforme descrita em (13), indicando que a distribuição

WG-PA é muito flexível e os valores de α tem um efeito substancial em sua curtose e θ em sua assimetria.

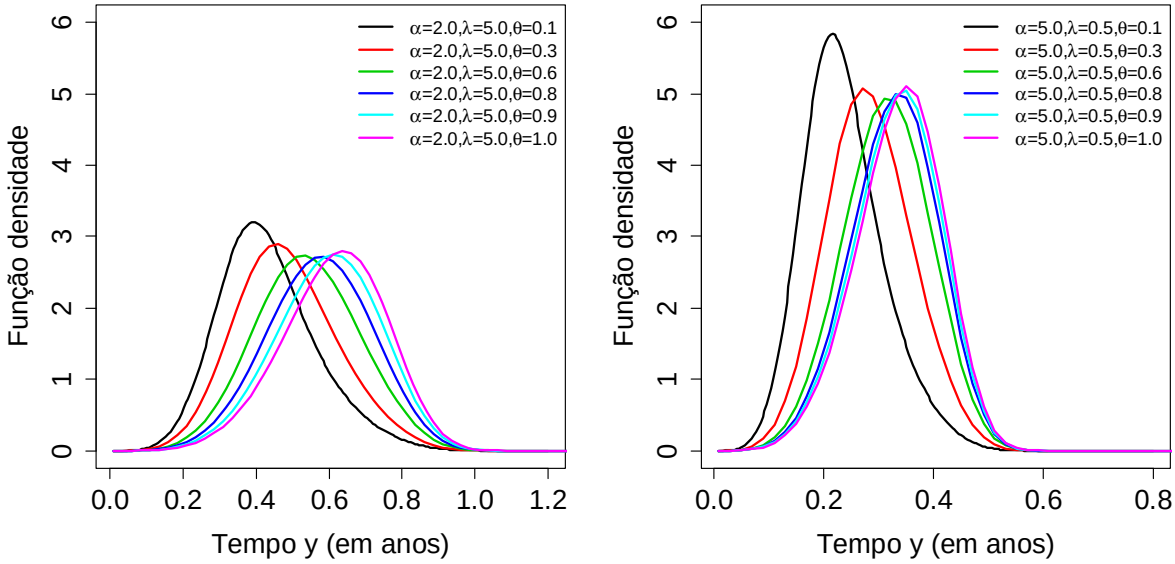

Figura 7: Função de densidade do modelo WG, mecanismo de PA, para alguns valores dos parâmetros α, θ e λ.

Fonte: Elaborado pelo autor.

De (12) e (13) obtém-se a função de risco do modelo WG da população de inadimplentes, no qual é dada por

$$h_{WG}(y) = \frac{h_W(y)}{1-(1-\theta)S_W(y)}, \quad y > 0. \qquad (14)$$

em que $h_w(y)$ e $S_w(y)$ são, respectivamente, a função de risco e função de sobrevivência de uma distribuição Weibull.

A Figura 8 mostra algumas formas da função de risco do modelo WG considerando o mecanismo de PA para alguns valores de parâmetros fixos.

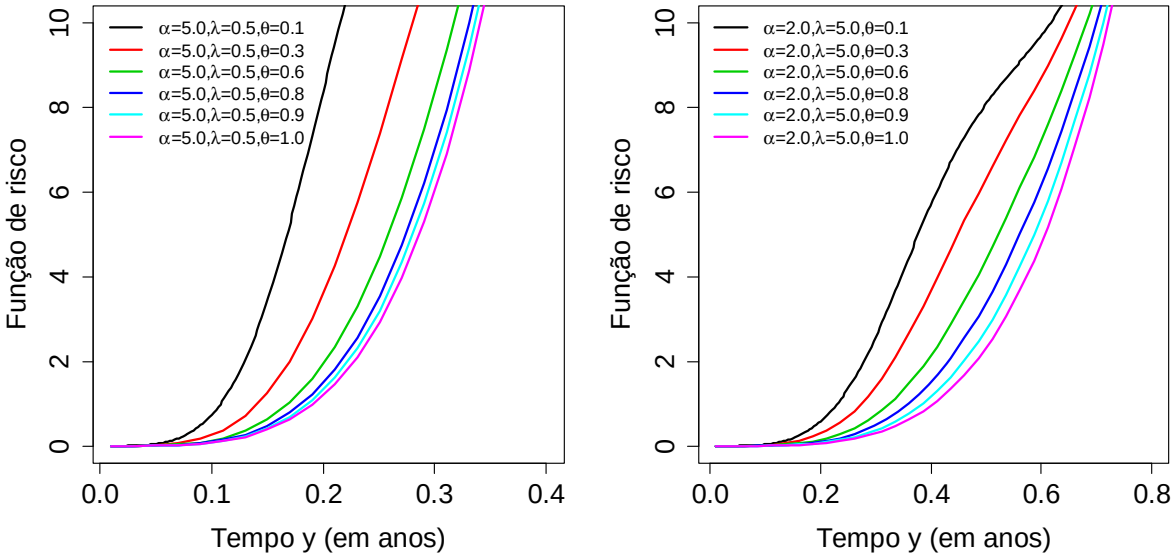

Figura 8: Função de risco do modelo WG, mecanismo de PA, para alguns valores dos parâmetros α, θ e λ.

Fonte: Elaborado pelo autor.

Para o mecanismo de UA, a função de sobrevivência para a população de adimplentes é dada por:

$$S_{WG}(y) = \frac{\exp\{-\lambda y^{\alpha}\}}{1-(1-\theta)[1-\exp\{-\lambda y^{\alpha}\}]}. \qquad (15)$$

Observa-se que $S_{WG}(0)=1$ *e* $S_{WG}(\infty)=0$, então (15) é uma função de sobrevivência própria.

A FDP correspondente é dada por

$$f_{WG}(y) = \frac{\theta\lambda\alpha y^{\alpha-1}\exp\{-\lambda y^{\alpha}]}{\{1-(1-\theta)[1-\exp\{-\lambda y^{\alpha}\}]\}^2}, \quad y>0. \qquad (16)$$

Na Figura 9, verificam-se as funções densidade do modelo WG para o mecanismo de UA, conforme descrita em (16), para alguns valores fixos de α, θ e λ. Estes gráficos mostram que o modelo WG-UA é também muito flexível e que os valores de θ e λ tem um efeito substancial em sua assimetria e curtose.

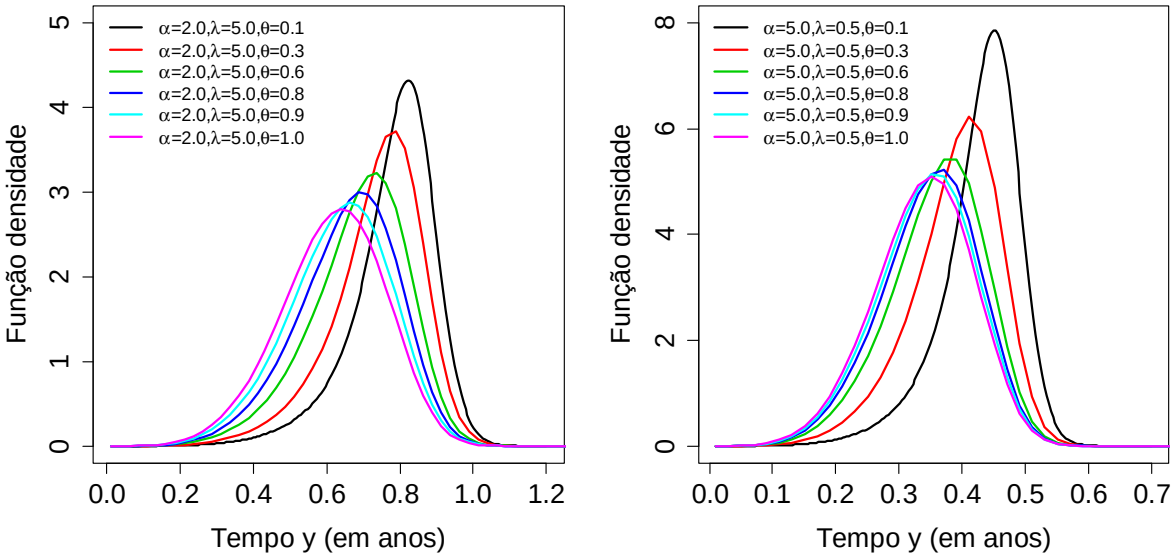

Figura 9: Função de densidade do modelo WG, mecanismo de UA, para alguns parâmetros α, θ e λ.

Fonte: Elaborado pelo autor.

De (15) e (16) obtém-se a função de risco do modelo WG da população de inadimplentes, no qual é dada por

$$h_{WG}(y) = \frac{h_W(y)}{1-(1-\theta)S_W(y)}, \quad y > 0. \qquad (17)$$

em que $h_w(y)$ e $S_w(y)$ são, respectivamente, a função e risco e função de sobrevivência de uma distribuição Weibull.

A Figura 10 mostra algumas formas da função de risco do modelo WG considerando o mecanismo de UA para alguns valores de parâmetros fixos.

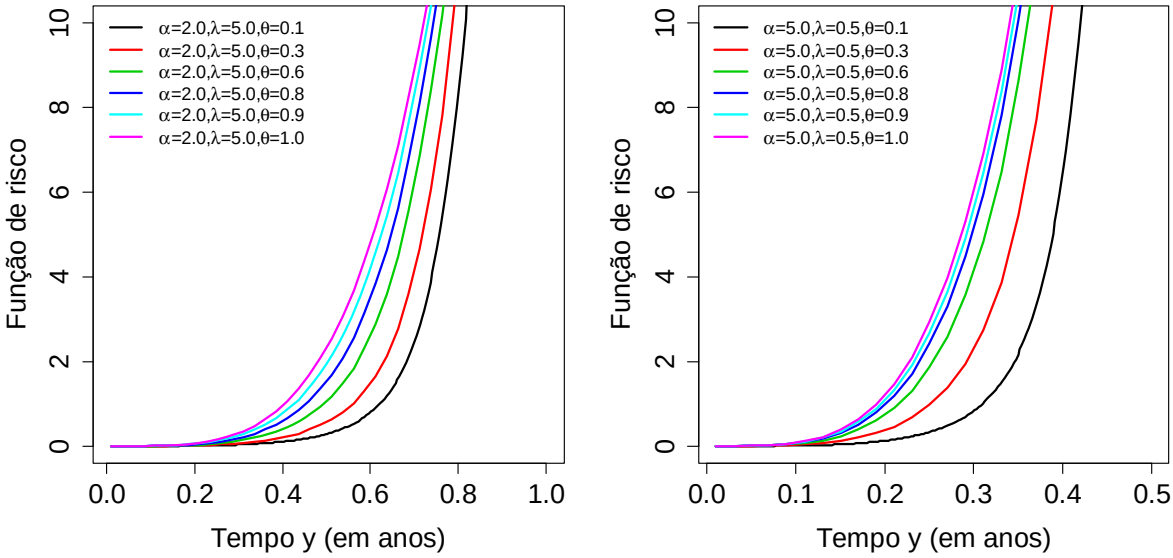

Figura 10: Funções de risco do modelo WG, mecanismo de UA para alguns parâmetros α, θ e λ.

Fonte: Elaborado pelo autor.

Há uma relação matemática entre os modelos (8) e (10) e o modelo de mistura padrão (2) (BOAG, 1949; BERKSON; GAGE, 1952), podendo-se escrever

$$S_{pop}(y) = \theta + (1-\theta)S_{WG}(y),$$

em que S_{WG} é dado por (12) ou (15).

Assim, $S_{pop}(y)$ é um modelo com fração de clientes do sistema financeiro, com fração $p_0 = \theta$, e função de sobrevivência $S_{WG}(y)$ para a população de adimplentes. O resultado implica que cada modelo de fração de adimplentes corresponde a algum modelo da forma (8), com mecanismo de PA, ou (10), com mecanismo de UA, para qualquer θ e $S_W(\cdot)$ (este resultado é válido para qualquer função de distribuição).

Inferência estatística (para aplicação na amostra utilizada)

Considera-se a situação em que o tempo até a ocorrência de inadimplência Y descrito não é completamente observado e está sujeito a censura à direita. C_i denota a censura do tempo até a ocorrência de inadimplência. Em uma amostra de tamanho n, observa-se

$T_i = \min\{Y_i, C_i\}$ e $\delta = I(Y_i \leq C_i\}$, onde $\delta = 1$ se T_i é o tempo até a inadimplência e $\delta = 0$ se for censurada à direita, para $i = 1,...,n$. Para completar o modelo, assume-se a existência de um vetor de covariáveis x_{ji}, com $j = 1,2$, e $i = 1,2,...,n$, em que x_{1i} está relacionado com a proporção de adimplentes por meio de uma função de ligação logística e x_{2i} está relacionado com o parâmetro de escala da distribuição de Weibull por meio de uma função de ligação logarítmica, o qual implica que a função de riscos proporcionais de Cox (LAWLESS, 2003; COLOSIMO; GIOLO, 2006)

$$\log\left(\frac{p_{0i}}{1-p_{0i}}\right) = x_{1i}\beta_1 \text{ e } \log(\lambda_i) = x_{2i}\beta_2 \Rightarrow h(z) = h_w(z) = \exp\{\beta x_i\}, \quad (21)$$

em que β_1 e β_2 denotam os coeficientes dos vetores de modo que em cada grupo de indivíduos representados por x_{1i} e x_{2i}, tem-se uma possível fração de adimplentes diferente e parâmetros de escala. Nota-se a função de verossimilhança de $\vartheta = (\beta_1, \beta_2, \alpha)^{\mathrm{T}}$ sob uma censura não informada é dada por

$$L(\vartheta, D) = \prod_{i=1}^{n} f_{WGcr}(t_i; \vartheta)^{\delta_i} S_{WGcr}(t_i; \vartheta)^{1-\delta_i}, \quad (22)$$

onde $D = (t, \delta, x_1, x_2)$, $t = (t_1,...,t_n)^{\mathrm{T}}$ e $\delta = (\delta_1,...,\delta_n)^{\mathrm{T}}$ denotam os vetores *n-dimensionais* do tempo até a ocorrência de inadimplência e censura, respectivamente, $x_j = (x_{j1},...,x_{jn})^{\mathrm{T}}$, $j = 1,2$, foram definidos acima, e $f_{pop}(\cdot;\vartheta)$ e $S_{pop}(\cdot;\vartheta)$ são FDP e funções de sobrevivência impróprias dadas em (6) - (11).

A partir da função de verossimilhança em (22), a estimativa de máxima verossimilhança do parâmetro ϑ é obtida maximizando a função log-verossimilhança $l(\vartheta, D) = \log\{L(\vartheta, D)\}$, realizada utilizando o software R (*R Development Core Team*, 2010).

Inferência para os parâmetros podem ser baseados nas estimativas de probabilidade máxima e os seus erros padrão estimados. Além disso, modelos diferentes podem ser comparados em relação ao seu ajuste usando o bem conhecido *Akaike information criterion* (AIC) e o *Schwartz-Bayesian criterion* (SBC), descritos nas seções seguintes.

2.5 Critérios para comparação de modelos

Comparar diferentes modelos é uma questão muito importante em modelagem estatística. A modelagem estatística é um processo iterativo, onde os modelos são construídos

gradualmente e comparados entre si para a escolha do mais adequado entre os de melhores ajustes. Além disso, a variância dos dados em torno da média nem sempre pode ser admitida como uniforme ao longo do intervalo da variável explanatória (PAIVA; FREIRE; CECATTI, 2008).

Existem diversas ferramentas estatísticas para analisar e selecionar o modelo mais adequado, com destaque para (VAN BUUREN; FREDRIKS, 2001):

- Inspeção visual da forma das curvas de referência;
- Gráfico quantil-quantil dos z-escores ("*qq-plot*" e "*detrended qq-plot*");
- Critério de informação de Akaike (AIC);
- Critério bayesiano de Schwarz (SBC);

Todavia, não existe uma regra para a escolha das ferramentas, nem em que sequência essas deve ser utilizada. A análise visual dos gráficos produzidos depende da subjetividade do pesquisador e nem sempre é suficiente. Por isso, critérios objetivos devem ser utilizados na seleção de modelos. Os critérios numéricos são bastante usados para discriminar modelos porque ponderam menos parâmetros com melhor adequação, ou seja, com menores desvios. A regra é selecionar entre os modelos candidatos que produzam o menor valor do AIC ou SBC. Menores valores para AIC e SBC indicam melhor ajuste, de modo que ambos os critérios permitem comparação entre modelos, penalizando aqueles com maior número de parâmetros.

2.5.1 AIC – *Akaike Information Criterion*

O AIC fundamenta-se no conceito de entropia por oferecer uma medida relativa da informação perdida quando um determinado modelo é usado para descrever a realidade e pode ser dito para descrever o equilíbrio entre polarização e variância na construção do modelo, ou, então, da precisão e complexidade do modelo (BURNHAM; ANDERSON, 1998). O cálculo do AIC pode ser descrito como

$$AIC = 2k - 2ln(L)$$

em que k é o número de parâmetros estimados no modelo, e L o valor maximizado da função de verossimilhança para o modelo estimado.

No AIC não há hipótese sendo testada. O critério permite que se determine qual modelo é o mais correto e quanto é mais correto. O método pode ser utilizado para comparar

qualquer tipo de modelo: linear, não linear, etc. A base teórica matemática do método combina a teoria da máxima verossimilhança, a teoria da informação e o conceito de entropia da informação (FLORIANO *et al.*, 2008). Para escolha do modelo pelo AIC deve ser respeitada a sequência de procedimentos a seguir:

1. Ajustar as equações;
2. Anotar a soma de quadrados do erro de cada modelo (usando a soma ponderada de quadrados do erro no caso de fatores ponderados);
3. Determinar o número de observações *n*; em sendo utilizada uma variável de frequência, esta deve ser considerada como multiplicador, e o valor total de *n* deve ser aquele determinado pela soma das frequências de cada classe de valor;
4. Determinar o valor de *k* para cada modelo;
5. Calcular o AIC;
6. O modelo com menor AIC é o mais próximo de ser o correto.

Somente devem ser comparados modelos que se ajustam bem aos dados, eliminando-se, anteriormente, todos os modelos que não apresentam bons resultados. O AIC é mais apropriado para dados provenientes de pequenas amostras.

2.5.2 SBC – *Schwarz Bayesian Criterion*

Também conhecido como critério bayesiano de Schwarz ou critério de informação bayesiano (BIC), o SBC foi desenvolvido por Schwarz em 1978, como uma estatística semelhante ao critério de Akaike, porém com a característica de impor uma penalidade maior pela inclusão de coeficientes adicionais a serem estimados.

O cálculo geral do SBC pode ser descrito por

$$SBC = k\,ln(n) - 2ln(L)$$

onde *k* é o número de parâmetros estimados no modelo, *n* o tamanho da amostra e *L* o valor maximizado da função de verossimilhança para o modelo estimado.

Segundo Schwarz (1978), para um grande número de observações os procedimentos AIC e SBC diferem marcadamente entre si. Desta forma, o critério de *Akaike* não pode ser assintoticamente ótimo. Portanto, o critério SBC é recomendado para dados provenientes de grandes amostras.

3 METODOLOGIA DA PESQUISA

Para Gil (1996), pesquisa é um “procedimento racional e sistemático que tem como objetivo proporcionar respostas aos problemas que são propostos”. A pesquisa é realizada quando não se dispõe de informações suficientes para responder ao problema, e quando a informação disponível encontra-se não organizada.

No presente trabalho foi realizado um levantamento das informações e dados históricos de uma instituição financeira com atuação no Brasil. O trabalho caracteriza-se como uma pesquisa aplicada, classificada como e*x-post facto* (a partir do fato passado), tendo em vista não haver controle sobre as variáveis independentes, dado que o evento já ocorreu (SAMPIERI; COLLADO; LUCIO, 2006). Além disso, é assim classificada por apresentar a determinação das relações existentes entre a variável resposta, probabilidade de um cliente se tornar inadimplente, e as variáveis explicativas. Do ponto de vista dos procedimentos técnicos, a pesquisa caracteriza-se como qualitativa e quantitativa, em que, com o uso de modelos matemáticos e estatísticos, busca-se a resolução de um problema real da instituição financeira estudada. É classificada como qualitativa, pois considera a instituição financeira como fonte direta de dados e o pesquisador como instrumento fundamental, frente a seu caráter descritivo e porque considera um corte temporal-espacial. Além disso, a contribuição ao trabalho de pesquisa é uma mistura de procedimentos de cunho racional e intuitivo capazes de contribuir para a melhor compreensão dos fenômenos (POPE; MAYS, 1995). É classificada como quantitativa, pois permite a descoberta de uma característica ou grupo de características de um grupo, gera medidas precisas e confiáveis que permitem análise estatística e permite a criação de modelos de previsão com base em característica observáveis. Frente ao caráter quantitativo desta pesquisa, mostra-se importante a escolha de um método estatístico que seja capaz de mensurar com qualidade e confiabilidade os dados que serão analisados.

3.1 Origem e planejamento da pesquisa

Em princípio, foi realizada análise e revisão para identificação, na literatura nacional e internacional, de pesquisas sobre técnicas de análise multivariada de dados com aplicações em instituições financeiras e modelos de escore de crédito. Verificou-se que, embora exista vasta

produção sobre o tema, pouco existe sobre a utilização de modelos de escore de crédito para descrever a inadimplência em carteiras de clientes pessoa física, e particularmente quanto aos fatores que influenciam para sua ocorrência.

Despertou interesse pela pesquisa sobre o tema, também, a informação de que o comportamento futuro desconhecido dos clientes ser muito importante para a gestão do relacionamento com o cliente, inclusive em instituições bancárias, e, assim, de crucial importância prever o comportamento futuro para que se possam tomar ações precisas no ato da solicitação de crédito.

Definido o tema a ser tratado, inicialmente foi realizada uma busca na literatura dos métodos estatísticos que permitissem e justificassem a análise dos dados e possibilitassem o alcance dos objetivos propostos. Posteriormente, definiu-se uma estrutura conceitual teórica que sustentasse os conhecimentos necessários para a efetiva realização da proposta. Após o levantamento do referencial teórico, procurou-se realizar a análise considerando os dados da instituição bancária objeto de estudo.

O passo seguinte foi definir a abordagem metodológica que determinou os métodos e técnicas, tanto para a coleta quanto para a análise dos dados. Com relação à coleta de dados, foi realizado um levantamento nas carteiras de créditos da instituição financeira estudada. Assim, a população inicialmente considerada contemplou todos os credores que adquiriram crédito na instituição estudada. Dessa população foi retirada uma amostra não probabilística para aplicação dos métodos estatísticos, correspondente a uma base de dados financeiros de tomadores de crédito na modalidade de Crédito Direto ao Consumidor (CDC) para clientes pessoa física. Essa modalidade de crédito foi escolhida, pois é a modalidade com maior número de contratos na instituição em estudo.

Todavia, frente à imensa quantidade de contratos, a amostra foi limitada aos tomadores de crédito na citada modalidade, considerando o período de 01/01/2010 a 31/12/2012 (36 meses), solicitados e administrados em todas as agências do banco instaladas em determinada cidade. Foram considerados, ainda, somente os contratos com prazo de 12, 24 e 36 meses. A data de observação foi 28/03/2013, e considerados inadimplentes os clientes com pagamento pendente (em atraso) por mais de 90 dias (ou três meses). Esse prazo foi considerado tendo em vista ser essa a definição de prazo para caracterização de inadimplência mais utilizada na literatura para os clientes com atraso no pagamento.

Dos clientes tomadores do crédito nas condições e limitações supracitadas, que totalizaram 2.302 clientes, foram consideradas, para cada cliente, as informações cadastrais prestadas e analisadas quando da solicitação do crédito em questão, que foram retiradas do banco de dados da instituição financeira na data de observação do estudo. Essas informações compõem as variáveis que são consideradas nas análises realizadas no presente estudo. Os dados foram tratados e tabulados utilizando-se planilhas eletrônicas, e o tratamento estatístico foi realizado utilizando-se da ferramenta computacional (software) R (*R Development Core Team*, 2010) e o software estatístico SPSS (PASW *Statistics* 18).

A base de dados da instituição objeto de estudo permitiu a obtenção das seguintes variáveis:

- Sexo: para efeitos das análises, os clientes do gênero feminino são considerados numericamente como 0, e os clientes do gênero masculino como 1;
- Idade: foram calculadas a partir da diferença entre a data de observação do estudo e a data de nascimento do cliente;
- Estado civil: foram consideradas as informações declaradas pelos clientes quando da solicitação do crédito, enquadrados dentre os grupos abaixo estabelecidos:
    - Solteiro(a), considerados numericamente como 0;
    - Casado(a), onde foram enquadrados todos aqueles que assim se declararam, indiferentemente do regime de comunhão de bens, e da forma de união com o cônjuge (casamento, união estável, etc.), e aqueles que já não mantenham mais essa condição (como os separados e divorciados). Considerados numericamente como 1;
- Valor da prestação, em reais (R$), conforme valor solicitado pelo cliente, considerando o limite máximo de 30% da renda mensal total declarada e/ou comprovada pelo cliente, conforme estabelecido pelo banco;
- Valor do contrato, em reais (R$), conforme valor solicitado pelo cliente, considerando o valor máximo decorrente do valor da prestação acordada e o prazo para pagamento do empréstimo;
- Prazo, determinando a quantidade máxima de meses em que o cliente deve amortizar a dívida contraída. Para a análise foram considerados os prazos de 12, 24 e 36 meses somente;

- Taxa de juro, referente ao índice de remuneração mensal da operação. A taxa de juro varia para cada solicitação de crédito, pois depende do relacionamento de cada cliente com a instituição bancária, e com a política de crédito do banco no momento da solicitação do empréstimo.
- Carteira (renda): a instituição financeira objeto do estudo segmenta seus clientes em distintas carteiras, definidas a partir de sua renda declarada e/ou comprovada e do total de investimentos com o banco. Para efeito desse estudo, é considerada somente a faixa de renda ao qual o cliente está inserido, e de acordo com a Tabela 4, com evidência de sua identificação numérica e respectiva faixa de renda declarada e/ou comprovada.

Tabela 4 - Faixa de renda dos clientes da instituição bancária

Pessoa Física	**Renda**
0	Até R$ 700,00
1	De R$ 700,01 a R$ 3.000,00
2	De R$ 3.000,01 a R$ 7.000,00
3	Acima de R$ 7.000,01

Fonte: Elaborado pelo autor.

É importante considerar que a renda do cliente influencia diretamente no valor da prestação a ser assumida, pois está não pode superar a margem de 30% da renda mensal total declarada e/ou comprovada do cliente, e, consequentemente, no valor máximo que pode ser contratado frente ao prazo almejado e possível.

O protocolo de estudo de caso para esta pesquisa é apresentado na Tabela 5.

Tabela 5 – Protocolo de estudo.

Questão principal de pesquisa	Como utilizar métodos estatísticos em uma carteira de crédito de uma instituição financeira para analisar e prever a ocorrência de inadimplência considerando os fatores ou variáveis relacionadas ao perfil dos clientes, e/ou características financeiras do crédito tomado?
Questões secundárias	Como obter um modelo estatístico (utilizando análise discriminante) para classificar os clientes como adimplentes e inadimplentes com altos índices de classificações corretas? Com esse modelo, como classificar novos clientes? Quais fatores (variáveis) influenciam na ocorrência de adimplência e da inadimplência dos clientes e como ocorre essa influência?
	Como utilizar um modelo estatístico (utilizando análise de sobrevivência) para modelar os tempos até a ocorrência da inadimplência e calcular a proporção de adimplência? Quais outros fatores (variáveis) influenciam na ocorrência de adimplência e da inadimplência dos clientes?
Unidade de análise	Inadimplência de clientes em uma carteira de Crédito Direto ao Consumidor (CDC).
Limites de Tempo	Créditos tomados entre 01/01/2010 a 31/12/2012 (36 meses), com prazos para pagamento de 12, 24 e 36 meses.
Local	Todas as agências do banco instaladas em determinada cidade.
Validade de Construtos	Baseado nas pesquisas de revisão bibliográfica sobre o tema.
Validade Interna	Utilização de amostra retirada do banco de dados da instituição financeira.
	Validação dos modelos utilizados.

Fonte: Elaborado pelo autor.

3.2 Técnicas estatísticas

Com referência às técnicas de análide multivariada de dados existentes, neste trabalho foi utilizada a AD, pois é aplicada em grupos predefinidos, característica deste caso, onde é possível identificar e agrupar os clientes como inadimplentes e adimplentes. Diante dessa característica, foi possível avaliar quais variáveis são importantes para a discriminação dos clientes na amostra existente, sendo possível a obtenção de uma função discriminante para essa classificação. Além disso, a função discriminante permitiu explorar as características capazes de serem utilizadas para alocar novos clientes nos grupos previamente identificados (de clientes adimplentes e clientes inadimplentes).

Maior confiabilidade foi agregada ao modelo estatístico obtido com a realização da AD utilizando-se da validação estratificando-se a amostra inicial (no caso a amostra após ajustada), dividindo-a aleatoriamente em duas subamostras com cerca de metade dos clientes cada uma, sendo uma chamada amostra de estimação (AE), ou amostra de análise, que foi utilizada para estimar a função discriminante, e a outra metade, chamada amostra de validação (AV), ou amostra retida, utilizada para fins de validação da função discriminante (HAIR *et al.*, 2005). Com esse procedimento sendo realizado diversas vezes, foi possível realizar o que se denomina validação cruzada, método amplamente utilizado na literatura.

Também foi realizada a análise dos dados utilizando um modelo de AS com fração de clientes adimplentes formulado com base em uma distribuição Weibull e uma Geométrica. O modelo permitiu determinar os fatores que influenciam na inadimplência e adimplência, conhecendo o tempo até que um cliente se torne inadimplente, e considerando que a inadimplência pode ser causada por diferentes mecanismos de ativação dados em (12) e em (15). Cada mecanismo de ativação foi considerado como um submodelo para aplicação nos dados da amostra. A análise foi realizada começando com a aplicação do submodelo com o mecanismo de PA, passando por um mecanismo de ativação aleatória (AA), e terminando com o mecanismo de UA, considerando que não há informações sobre qual motivo foi o responsável pelo tempo para a ocorrência de inadimplência. Os submodelos foram comparados utilizando-se dos critérios AIC (*Akaike Information Criterion*,) e do critério bayesiano SBC (*Schwartz Bayesian Criterion*), para identificar qual mecanismo de ativação melhor se enquadra ao modelo proposto e dados analisados.

4 ANÁLISE DOS DADOS E DISCUSSÃO DOS RESULTADOS

4.1 Aplicação da análise discriminante para análise de escore de crédito

Primeiramente foi identificado na amostra os tipos de clientes, que consistem nos clientes adimplentes, identificados numericamente na análise como 0, e dos clientes inadimplentes, identificados numericamente na análise como 1, denominando-se a variável dependente qualitativa. As variáveis independentes foram consideradas de acordo com as informações disponíveis na amostra coletada, limitando-se somente às variáveis quantitativas, quais sejam: Idade, Valor do Contrato (identificada como ValorContrato), Prazo, Taxa e Valor da Prestação (identificada como ValorPrestacao). Para auxílio na realização da AD foram utilizadas planilhas eletrônicas e o software estatístico SPSS 18 (PASW *Statistics* 18).

Previamente à realização da análise dos resultados com a aplicação de AD, foi realizada uma análise estatística descritiva da amostra utilizada, conforme Tabela 6.

Tabela 6 - Medidas-resumo da amostra por tipo de cliente.

Tipo de cliente (grupos)		**Média**	**Desvio-padrão**	**Coeficiente de variação (%)**
0 - Adimplentes n = 1784	Idade (em anos)	46,23	13,73	29,69
	Valor do Contrato (em milhares de R$)	2,9594	3,8298	129,41
	Prazo (meses)	30,25	8,82	29,16
	Taxa (% ao mês)	3,7688	1,0514	27,90
	Valor da Prestacão (em milhares de R$)	0,1711	0,22	126,66
1 - Inadimplentes n = 518	Idade (em anos)	40,47	14,09	34,81
	Valor do Contrato (em milhares de R$)	2,123	0,2248	105,96
	Prazo (meses)	31,88	7,40	23,23
	Taxa (% ao mês)	4,6142	0,7318	15,86
	Valor da Prestacão (em milhares de R$)	0,1365	0,1479	108,35

Fonte: Elaborado pelo autor.

Conforme se observa, da amostra total com 2.302 cliente, cerca de 22,5% são inadimplentes, contra 77,5% de adimplentes. No grupo de clientes adimplentes, contemplando 1.784 clientes, a idade média dos clientes da amostra é de 46 anos, com variabilidade de

29,69%. Já no grupo de clientes inadimplentes, com 518 clientes, observa-se uma idade média de 40 anos, com variabilidade de 34,81%. Assim, pode-se observar que os clientes inadimplentes possuem, em média, idade inferior aos clientes adimplentes.

O valor médio dos empréstimos contratados pelos clientes adimplentes girou em torno de R$ 2.959,42. Para clientes inadimplentes verifica-se um valor médio de R$ 2.121.27 para essa variável, indicando que os empréstimos dos clientes adimplentes possui maior variação do que os clientes inadimplentes. O prazo médio da amostra dos grupos representaram valores aproximados, em que para os clientes adimplentes o prazo médio foi de 30 meses para a quitação do empréstimo, e no grupo de inadimplentes, de quase 32 meses.

A taxa de juros dos créditos tomados pelos clientes adimplentes ficou, em média, em 3,77% ao mês, apresentando maior variação em relação aos clientes inadimplentes. A taxa média de juros no grupo de inadimplentes foi consideravelmente superior, de 4,61%. Já o valor médio das prestações mensais dos empréstimos contradados girou em torno de R$ 171,14 para os clientes adimplentes e de R$ 136,50 para os clientes inadimplentes.

Com os dois grupos de clientes (adimplentes e inadimplentes) conhecidos do banco, a AD foi aplicada buscando descrever os aspectos que diferenciam os grupos e avaliar quais variáveis são importantes para discriminar a variável dependente (tipo de cliente). Segundo Hair *et al.* (2005), a AD é muito sensível à proporção entre o tamanho da amostra e o número de variáreis preditoras, e estudos sugerem uma proporção de 20 observações para cada variável preditora, sendo o mínimo recomendado de 5 observações para cada variável independente. Como a amostra considerada para a análise conta com 2.302 clientes e serão consideradas (inicialmente) cinco variáveis independentes, a proporção para a realização da AD é satisfatória.

Para uma aplicação apropriada da AD é recomendável que se atenda a certas condições. As suposições iniciais para determinar a função discriminante são a normalidade multivariada das variáveis independentes. Assim, foi realizada uma análise gráfica descritiva das variáveis independentes utilizadas, separadas conforme a variável dependente (Tipo de Cliente, ou seja, clientes adimplentes e clientes inadimplentes). Como se observa nos gráficos da Figura 11, ambos os grupos sugerem distribuição normal multivariada dos dados.

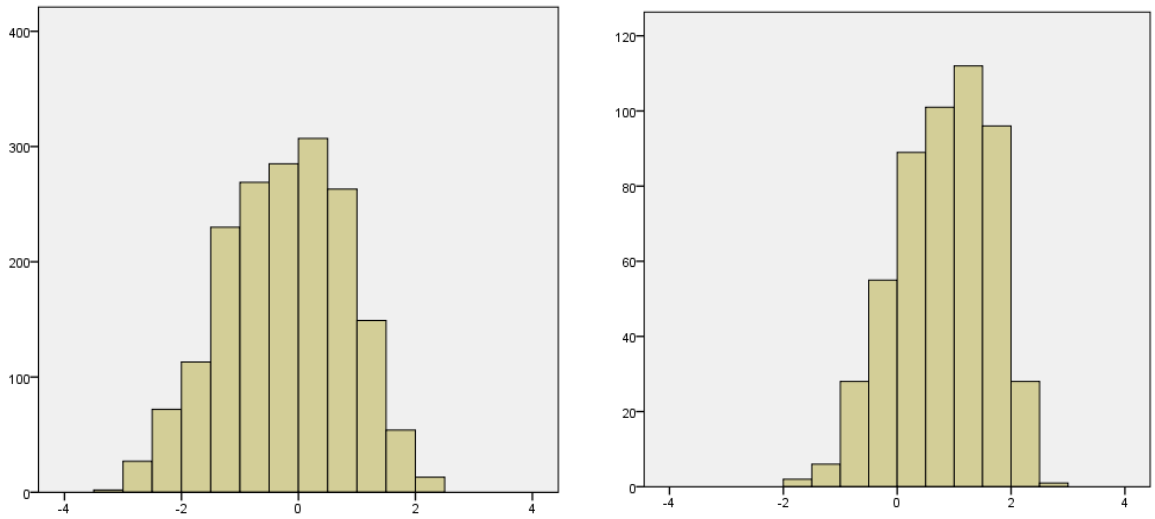

Figura 11 – Distribuição da função para o grupo de clientes adimplente (esquerda) e para o grupo de clientes inadimplentes (direita).

Fonte: Elaborado pelo autor.

Outra importante hipótese a ser testada refere-se à igualdade das matrizes de covariância de cada grupo (de adimplentes e de inadimplentes) para verificar a homogeneidade das matrizes, ou seja, $\sum_1 = \sum_2$ (HAIR *et al.*, 2005). Assim, supõe-se:

$$\begin{cases} H_0 = As\ \ matrizes\ \ de\ \ \text{cov}\,ariância\ \ dos\ \ dois\ \ grupos\ \ são\ \ iguais; \\ H_1 = As\ \ matrizes\ \ de\ \ \text{cov}\,ariância\ \ dos\ \ dois\ \ grupos\ \ não\ \ são\ \ iguais. \end{cases}$$

Para essa verificação foi utilizado o teste M de Box, que resultou em um *p-value* de 0.000, não demonstrando, assim, a homogeneidade das matrizes de covariância. De acordo com Tabachnick e Fidell (2012), o teste M de Box é altamente sensível e, assim, se *p-value*<0,001 e o tamanho das amostras são desiguais, o teste deve ser ignorado.

Realizadas as análises e considerações descritas acima foi realizada a AD, em que todas as observações (100%) foram consideradas na análise dos 2302 clientes da amostra, e foi possível calcular os coeficientes da função discriminante linear de Fisher. Esses coeficientes, que também são chamados de pesos discriminantes, podem ser utilizados para avaliar a importância relativa de cada variável explicativa para a função discriminante. De acordo com os resultados obtidos, o modelo estatístico para classificar os clientes é dado pela função discriminante linear de Fisher em (23).

$$I = -3 - 0{,}036 \cdot Idade - 0{,}035 \cdot ValorContrato + 0{,}039 \cdot \text{Pr}azo + 0{,}893 \cdot Taxa + 0{,}094 \cdot Valor\text{Pr}estacao \quad (23)$$

Com os coeficiente das variáveis independentes da função discriminante é possível verificar quais variáveis possuem maior peso na discriminação dos clientes adimplentes e inadimplentes. Assim, é possível verificar que a variável Taxa possui maior peso positivo na discriminação dos clientes, enquanto a variável Idade possui maior discriminação negativa.

Todavia, por meio do cálculo e análise das correlações combinadas dentro de grupos entre variáveis discriminantes e a função discriminante linear de Fisher (variáveis ordenadas pelo tamanho da correlação dentro da função) é possível identificar de forma clara e mais precisa essa correlação, cujos resultados são apresentados na Tabela 7.

Tabela 7 – Correlação entre as variáveis e a função discriminante

Variáveis	**Função**
Taxa	0,812
Prazo	0,181
ValorPrestacao	-0,162
ValorContrato	-0,225
Idade	-0,396

Fonte: Elaborado pelo autor.

Como observado, a variável Taxa possui maior correlação positiva com a função discriminante. Dessa forma, há indícios de que altas taxas de juro dos empréstimos tomados pelos clientes do banco estudado exerça maior influência que as demais variáveis na ocorrência de inadimplência dos contratos. Por outro lado, a variável idade possui maior correlação negativa com a função discriminante, e, ao contrário da taxa de juro da operação, é provável que clientes mais velhos influenciem (mais que as demais variáveis) na manutenção da adimplência dos contratos.

Tabela 8 – Resultados da classificação

	Grupos	**0**	**1**	**Total**
Quantidade	0	1709	75	1784
	1	385	133	518
%	0	95,8	4,2	100
	1	74,3	25,7	100

Fonte: Elaborado pelo autor.

Após a determinação da função discriminante linear de Fisher em (23), foi possível calcular os índices de classificações corretas. Conforme resultados mostrados na Tabela 8, 80% ((1709+133)/2302) dos clientes foram classificados corretamente na análise realizada.

Identifica-se, também, que o grupo de clientes adimplentes apresenta maior porcentagem de classificações corretas (95,8%) do que o grupo de inadimplentes (25,7%).

Não obstante tenha-se verificado um alto percentual (80%) de classificações corretas para análise da carteira de clientes do banco em estudo é de suma importância que o modelo de escore de crédito resulte em elevados índices de classificação correta dos clientes inadimplentes, que é o objetivo do banco. Conforme o baixo resultado de classificações corretas para os clientes inadimplentes de 25,7% encontrado, se comparado com a expectativa do banco para as corretas classificações desses clientes, que é de um percentual acima de 62,5% (percentual que será justificado adiante), a amostra inicial foi ajustada para a realização de uma nova AD, buscando um modelo estatístico que classifique com melhor índice de classificação correta os clientes inadimplentes.

Ajuste do tamanho da amostra

Para a realização da AD, além do tamanho da amostra geral, deve-se considerar o tamanho da amostra de cada grupo. Se os grupos variam muito em tamanho, como na amostra objeto deste estudo, pode causar impacto na estimação da função discriminante e na classificação dos clientes. Para esses casos, pode-se retirar uma amostra aleatoriamente a partir do grupo maior, de clientes adimplentes, reduzindo seu tamanho a um nível comparável ao do grupo menor, de cientes inadimplentes (HAIR *et al.*, 2005).

Assim, do grupo de clientes adimplentes da amostra inicial, com 1.784 clientes, foi retirada, aleatoriamente, uma subamostra com tamanho igual ao grupo de clientes inadimplentes (com 518 clientes). A nova amostra, aqui denominada amostra ajustada, conta, então, com 1036 clientes, sendo metade pertencente ao grupo de clientes adimplentes e a outra metade ao grupo de clientes inadimplentes.

Realizando nova AD com a amostra ajustada a normalidade multivariada das variáveis foi analisada descritivamente, cujos resultados se mostraram muito próximos aos apresentados na Figura 11, e, assim, foi dada continuidade na AD. Todas as observações (100%) foram consideradas na análise dos 1036 clientes da amostra ajustada, e obteve-se o modelo estatístico para discriminar os clientes, ou seja, o modelo de escore de crédito dado em (24) para classificar os clientes da instituição financeira:

$$I = -3{,}876 - 0{,}034 \cdot Idade - 0{,}079 \cdot ValorContrato + 0{,}045 \cdot \Pr azo + 0{,}968 \cdot Taxa + 0{,}824 \cdot Valor\,\Pr estacao \quad (24)$$

Com os coeficiente das variáveis independentes dessa função discriminante é possível verificar que a variável Taxa possui maior peso positivo na discriminação dos clientes, enquanto a variável Valor do Contrato possui maior discriminação negativa.

Consultando a Tabela 9 podem-se identificar as novas correlações combinadas dentro de grupos entre variáveis discriminantes e a função discriminante. A variável Taxa possui maior correlação positiva com a função discriminante, demonstrando mais uma vez que a taxa de juro dos empréstimos tomados pelos clientes do banco estudado exerce maior influência que as demais variáveis na ocorrência de inadimplência dos contratos. A variável Idade possui maior correlação negativa com a função discriminante, indicando que a idade do cliente exerce maior influência que as demais variáveis na manutenção da adimplência dos contratos.

Tabela 9 – Correlação entre as variáveis e a função discriminante para a amostra ajustada

Variáveis	**Função**
Taxa	0,831
Prazo	0,180
ValorPrestacao	-0,132
ValorContrato	-0,209
Idade	-0,373

Fonte: Elaborado pelo autor.

Utiliza-se a função discriminante linear de Fisher em equação (24) como meio de classificação porque ela fornece uma representação concisa, simplificando o processo de interpretação e a avaliação da contribuição de variáveis independentes, e deve-se garantir que ela forneça diferenças entre os grupos de adimplentes e inadimplentes. Segundo Hair *et al.* (2005), para essa verificação pode ser realizado um teste de significância Wilks' Lambda para verificar diferenças entre os grupos. Para comprovar a significância o valor do Wilks' Lambda deve estar entre 0 e 1, com *p-value* inferior a 0.01. O teste realizado resultou em Wilks' Lambda igual a 0.749, com *p-value* 0.000. Assim, os grupos mostraram diferenças significativas, demonstrando que a função discriminante em (24) representa diferenças entre os grupos.

Após a determinação do modelo de escore de crédito em (24), foram calculados os novos índices de classificações corretas. Conforme resultados mostrados na Tabela 10, 72,4% ((349+401)/1036) dos clientes foram classificados corretamente nessa nova análise para a

amostra ajustada. O grupo de clientes adimplentes apresenta menor porcentagem de classificações corretas (67,4%) em relação às análises ao do grupo de clientes inadimplentes (77,4%).

Tabela 10 – Resultados da nova classificação para a amostra ajustada.

	Grupos	**0**	**1**	**Total**
Quantidade	0	349	169	518
	1	117	401	518
%	0	67,4	32,6	100
	1	22,6	77,4	100

Fonte: Elaborado pelo autor.

Apesar da redução do percentual médio de classificações corretas para 72,4% após o ajuste da amostra, notadamente o percentual de classificações corretas dos clientes inadimplentes subiu expressivamente para 77,4%. Assim, após o ajuste da amostra, o cálculo de escores que resultaram em altos índices de classificações corretas dos clientes inadimplentes, que é de suma importância para a instituição bancária, permitindo a ela a utilização do modelo de escore de crédito como ferramenta auxiliar e complementar ao modelo de análise de risco de crédito existente na política de concessão de crédito aos clientes, considerando a modalidade de crédito estudada (CDC).

Todavia, após a adequação do tamanho da amostra, e considerando os altos percentuais de classificações corretas apresentados na Tabela 10, é imprescindível que o modelo seja validado. Dessa forma, a análise utilizando-se da amostra ajustada será validada por meio de validação cruzada (HAIR *et al.,* 2005).

Validação cruzada

Realizando o método de validação cruzada com base nos dados da amostra ajustada, foi possível calcular o percentual médio de classificações corretas, conforme resultados apresentados na Tabela 11.

Como se pode observar em média 71,8% dos clientes foram corretamente classificados com a validação cruzada. No grupo de clientes adimplentes, em média 67,8% dos clientes foram corretamente classificados, e em média 76,3% foram corretamente classificados no

grupo dos clientes inadimplentes, valores esses muito próximos aos encontrados na análise com a amostra ajustada, corroborando, assim, para a conclusão de que a função discriminante linear de Fisher obtida em (24) discrimina e classifica corretamente os clientes.

Tabela 11 – Resultados médios da classificação utilizando o método de validação cruzada para a amostra ajustada

	Grupos	**0**	**1**	**Total**
Quantidade	0	349	169	518
	1	123	395	518
%	0	67,4	32,6	100
	1	23,7	76,3	100

Fonte: Elaborado pelo autor.

Todavia, apesar de as proporções de sucesso serem altas, elas devem ser comparadas com os critérios de chance máxima e de chance proporcional para avaliar sua efetividade (HAIR *et al.*, 2005). O critério de chance máxima é a proporção de sucesso obtida se se designa todas as observações para o grupo com a maior probabilidade de ocorrência. Como as amostras em ambos os casos possuem cerca de 50% de clientes adimplentes e 50% de clientes inadimplentes, a probabilidade mais alta para o critério de chance máxima seria de 50%.

Já o critério de chance proporcional é calculado por:

$$C = p^2 + (1-p)^2,$$

onde p é a proporção de clientes no grupo de adimplentes e $1 - p$ é a proporção de clientes no grupo de inadimplente. Assim, o valor de chance proporcional calculado é também de 50%. Todavia, como referência para verificação de ajuste dos modelos foi considerada uma referência 25% maior que o valor critério (que é de 50%). Assim, a proporção de sucesso deve exceder a 62,5% (50% + 25% de 50%). A proporção de sucesso de 72,4% na amostra ajustada, considerando ambos os grupos supera esse critério (62,5%) que é assumido pelo banco.

Igualmente, deve-se avaliar a taxa de classificação para os grupos individualmente. Para o grupo de clientes adimplentes o percentual de proporção de sucesso é de 67,4% na amostra ajustada, superando o percentual de referência considerado (62,5%). Para o grupo de clientes inadimplentes a proporção de sucesso é de 77,4% na amostra ajustada, também superando o critério definido (62,5%).

Diante dos resultados apresentados, pode-se concluir que se verificou o ajuste do modelo em (24) para a amostra considerada. As evidências para essa conclusão são verificadas, principalmente, no teste de significância Wilks' Lambda para diferenças entre os grupos para a amostra ajustada, que demonstra as diferenças significativas entre os grupos para a função discriminante, e a proporção de sucesso de classificação do modelo, tanto em relação à classificação considerando a média dos grupos, quanto os considerando individualmente, que resultou em percentuais elevados de classificações corretas em relação aos critérios estabelecidos. Assim, pode-se concluir que os dados foram bem classificados após o ajuste da amostra, que foi ratificado com a realização da validação cruzada (HAIR *et al.*, 2005).

Dessa forma, a função discriminante linear de Fisher dada em equação (24) pode ser considerada como o modelo de escore de crédito e classificação dos clientes da instituição bancária para clientes solicitantes de crédito na modalidade CDC. Essa classificação poderia permitir ao banco analisar se o cliente será inadimplente com as variáveis independentes consideradas. Uma aplicação do modelo de escore de crédito é realizada a seguir.

Classificação de clientes solicitantes de crédito

Novos clientes solicitantes de crédito na modalidade CDC puderam ser classificados realizando-se o cálculo do escore de crédito correspondente. Para isso foram inseridos na função discriminante linear de Fisher em (24) os respectivos valores das variáveis independentes de cada novo cliente. O escore resultante para cada cliente foi comparado com o valor de referência $\overline{m}$, encontrado de acordo com a regra de classificação baseada na função discriminante linear amostral de Fisher dada em (1), p. 33.

Para isso foram selecionados, aleatoriamente, 30 novos clientes do banco de dados de solicitantes de crédito na modalidade CDC nas agências consideradas na amostra inicial da instituição bancária, créditos esses solicitados no mês de setembro de 2013. Além disso, foram considerados somente os clientes solicitantes do referido crédito com prazos de 12, 24 ou 36, conforme definido na amostra inicialmente utilizada.

Para cada um dos 30 clientes da amostra foi realizado o cálculo do escore de crédito correspondente, cujo escore resultante foi comparado com o valor de referência $\overline{m}$. De acordo com (1), $\overline{m}$ foi calculado utilizando os valores dos centroides mostrados na Tabela 12.

Tabela 12: Centroides para os tipos de clientes

Tipo de Cliente	Centroide
0 - Adimplente	-0,579
1 - Inadimplente	0,579

Fonte: Elaborado pelo autor.

Diante dos valores dos centroides o valor de referência $\overline{m}$ resultou em zero. Assim, os clientes foram classificados da seguinte forma (JOHNSON; WICHERN, 2007):

- O cliente $\underline{X}$ foi alocado no grupo de clientes inadimplentes para escore resultante maior ou igual a zero;
- O cliente $\underline{X}$ foi alocado no grupo de clientes adimplentes para escore resultante menor que zero.

Os valores de cada variável independente, o escore resultante para cada cliente e sua classificação pode ser verificado na Tabela 13. Como se pode obsevar com a classificação dos 30 clientes solicitantes de crédito, 26,67% dos clientes foram classificados no grupo de clientes inadimplentes. Isto posto, no caso de as ferramentas de análise de risco de crédito do banco permitirem a liberação do crédito a esses solicitantes, os resultados apresentados podem subsidiar a decisão de conceder ou não o crédito ao cliente, ou, também, ser considerado para o ajuste de alguns fatores financeiros do empréstimo para diminuição da probabilidade de inadimplência, alterando o valor a ser emprestado, o prazo para pagamento, a taxa de juro mensal e/ou o valor da prestação, por exemplo.

Além disso, os escores determinados podem ser analisados como subsídio para a tomada decisão quanto à liberação de crédito ao cliente. Escores negativos indicam que o cliente será classificado no grupo de adimplentes, e quanto menor o valor do escore, maior a probabilidade de que o cliente seja um bom pagador. Por outro lado, escores positivos indicam classificação no grupo de clientes inadimplentes, e quanto mais alto o valor do escore do cliente, maior a probabilidade de que ele não honre com os pagamentos do empréstimo.

Outro fator que é possível observar é o peso de algumas variáveis na determinação do escore de cada cliente. Taxas de juros mais elevadas influenciam na classificação do cliente como inadimplente, enquanto que idades mais elevadas influenciam na classificação de clientes como adimplentes.

Tabela 13 – Classificação de novos clientes solicitantes de crédito.

Cliente	Idade (em anos)	Prazo (em meses)	Taxa (% ao mês)	Valor do Contrato (em milhares de R$)	Valor da Prestação (em milhares de R$)	Escore	Classificado no grupo
1	39,98	12	3,88	2,5000	0,2747	-0,9105	*Adimplente*
2	86,39	12	3,89	2,0000	0,2161	-2,4877	*Adimplente*
3	30,28	12	2,39	1,0000	0,0980	-2,0504	*Adimplente*
4	25,60	12	2,70	3,0000	0,3014	-1,5816	*Adimplente*
5	38,69	12	3,88	0,5000	0,0545	-0,8902	*Adimplente*
6	25,25	12	4,95	1,5000	0,1749	0,6228	*Inadimplente*
7	59,86	12	4,13	0,5000	0,0548	-1,3676	*Adimplente*
8	57,95	12	4,78	0,7000	0,0791	-0,6693	*Adimplente*
9	39,27	24	3,59	4,0000	0,3215	-0,7073	*Adimplente*
10	75,16	24	4,65	0,7340	0,0533	-0,8642	*Adimplente*
11	56,19	24	5,40	2,5000	0,2023	0,4900	*Inadimplente*
12	39,53	24	5,40	0,5000	0,0407	1,0812	*Inadimplente*
13	30,78	24	4,35	1,0000	0,0773	0,3529	*Inadimplente*
14	75,25	24	2,39	2,0000	0,1154	-3,1038	*Adimplente*
15	44,41	24	3,99	3,0000	0,2035	-0,5131	*Adimplente*
16	78,54	24	5,45	15,0000	1,1869	-0,3976	*Adimplente*
17	25,66	24	3,88	29,9000	2,0284	-0,6034	*Adimplente*
18	56,30	36	2,39	8,0000	0,3397	-2,2086	*Adimplente*
19	49,25	36	3,88	2,2000	0,1171	-0,2519	*Adimplente*
20	51,67	36	3,40	0,5000	0,0247	-0,7407	*Adimplente*
21	58,22	36	4,13	4,0000	0,2192	-0,3731	*Adimplente*
22	48,44	36	4,78	2,0000	0,1196	0,6645	*Inadimplente*
23	25,10	36	4,69	0,5500	0,0333	1,4144	*Inadimplente*
24	48,59	36	5,40	0,5000	0,0329	1,3066	*Inadimplente*
25	33,78	36	3,40	3,3000	0,1639	-0,2389	*Adimplente*
26	49,81	36	3,50	0,5000	0,0255	-0,5799	*Adimplente*
27	38,42	36	4,65	1,3000	0,0780	0,9005	*Inadimplente*
28	57,73	36	4,65	13,0000	0,7798	-0,1020	*Adimplente*
29	41,04	36	3,40	0,7500	0,0497	-0,3786	*Adimplente*
30	75,16	36	2,98	1,1206	0,0768	-1,9519	*Adimplente*

Fonte: Elaborado pelo autor.

Os resultados descritos acima, se utilizados na realização da análise de risco de crédito da instituição, poderiam permitir ao banco maior segurança na liberação dos créditos solicitados na modalidade estudada, numa perspectiva de redução da ocorrência da inadimplência, e, consequentemente, diminuição do risco de liquidez e insolvência da instituição financeira. Todavia, outros fatores (variáveis) podem ser considerados no

momento da análise para concessão do crédito ao cliente como subsídio na decisão de conceder ou não o crédito. Dentre esses fatores pode ser analisada a renda do cliente, onde rendas mais elevadas podem pesar positivamente para a decisão de concessão de crédito ao cliente.

4.2 Aplicação do modelo de sobrevivência para análise de escore de crédito

Para ilustrar e aplicar a modelagem de AS com fração de clientes adimplentes obtida e discutida na seção 2.4.2 (p. 40 a 55), considera-se a amostra dos 2.302 clientes do banco, tomadores de empréstimo na modalidade CDC, que compreendeu o prazo de 36 meses de duração do empréstimo, conforme descrito no capítulo 3 (p. 58 a 63). O tempo observado *y* (em anos) refere-se ao tempo até a ocorrência da inadimplência do cliente, ou o tempo de censura para os clientes que não apresentaram inadimplência no decorrer do período de empréstimo. Clientes que se tornaram inadimplentes por outras causas, como o esquecimento de realização do pagamento, a perda do emprego, o pagamento mensal comprometer excessiva percentagem da renda, a falência pessoal, entre outras causas, e não constam como inadimplentes na amostra, bem como clientes que não se tornaram inadimplentes até o fim do estudo (clientes censurados), compreendem 77,5% do total de clientes.

O principal objetivo da análise é entender como se comporta o tempo até a ocorrência da inadimplência de acordo com o sexo e a renda do cliente e como ele pode ser representado em termos da função de distribuição de probabilidade. Além disso, os analistas de dados do banco estudado afirmam que o sexo e a renda dos clientes influenciam na inadimplência, em que clientes do sexo feminino são menos suscetíveis à inadimplência em relação aos clientes do sexo masculino, e clientes com baixa renda são mais suscetíveis à inadimplência em relação àqueles que possuem maior renda, ocorrências que também foram analisadas.

A renda dos clientes, descrita na Tabela 4, p. 61, foi reagrupada para utilização na AS. Assim, foram mantidos no grupo 0 os clientes com renda de até R$ 700,00, cuja proporção é de 29,54% dos clientes da amostra, e no grupo 1 foram considerados todos os clientes com renda superior a R$ 700,00, abrangendo 70,46% dos clientes. O agrupamento foi realizado dessa forma, pois o objetivo foi agrupá-los em apenas dois grupos buscando a menor diferença entre as proporções de clientes em cada grupo.

Para a realização da análise foram consideradas como covariáveis o sexo (gênero) *(x_1)*, sendo que a amostra contém 58,34% de clientes do gênero masculino e 41,66% do feminino; a renda *(x_2);* e o estado civil *(x_3)*, sendo 38,66% dos clientes declarados como solteiros e 61,34% como casados.

Inicialmente foram ajustados os modelos propostos em (6), (8) e (10), p. 47 a 49, com a função de ligação dada em (21), p. 55, com todas as covariáveis. Após um teste da razão de verossimilhança ficou-se com o seguinte modelo: $\theta_i = p_0 = \exp(\beta_1 x_i)/(1+\exp(\beta_1 x_i))$ e $\lambda_i = \exp(\beta_2 x_i)$, com $x_i = (1, x_1, x_2, x_3)^{\mathrm{T}}$ denotando o vetor de covariáveis.

Na Tabela 14 são mostrados os resultados numéricos dos cálculos realizados utilizando os critérios AIC e SBC, e com esses resultados é possível selecionar um submodelo dentre os submodelos candidatos com relação aos mecanismos de ativação. Observa-se que o submodelo de fração de clientes adimplentes com o mecanismo de primeira ativação (PA) se destaca como um dos melhores, de acordo com os critérios AIC e SBC (menores valores).

Tabela 14 – Estatísticas dos modelos ajustados.

	Estatísticas	
Mecanismos de ativação	**AIC**	**SBC**
Aleatório	4.849.313	4.883.762
Primeira Ativação	**4.824.247**	**4.858.696**
Última Ativação	4.884.738	4.919.187

Fonte: Elaborado pelo autor.

A Figura 12 mostra o gráfico QQ plot dos resíduos quantis randomizados normalizados (DUNN; SMYTH, 1996; RIGBY; STASINOPOULOS, 2005), referente ao mecanismo de PA. Cada ponto na figura corresponde à média de cinco conjuntos de resíduos ordenados, levando-se em conta os critérios na Tabela 14. Os pontos distribuídos próximos à linha de identidade sugerem que o modelo WG com fração de clientes adimplentes com mecanismo de PA produz um *t* aceitável. Assim, levando-se em conta os critérios e resultados obtidos com a Tabela 14 e o gráfico QQ plot apresentado na Figura 12, o modelo WG com fração de clientes adimplentes com mecanismo de PA foi selecionado como o modelo de trabalho frente sua melhor adequação e aplicação aos dados da amostra.

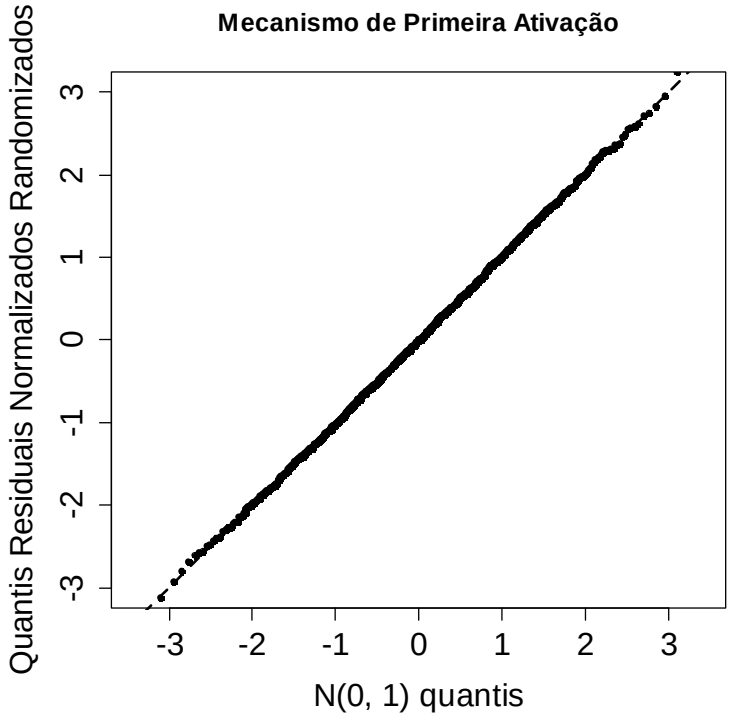

Figura 12: Gráficos QQ plot dos resíduos quantis randomizados normalizados com linha de identidade para o mecanismo de primeira ativação.

Fonte: Elaborado pelo autor.

As estimativas de máxima verossimilhança e os erros padrão dos parâmetros do modelo WG com mecanismo de PA, com $\alpha = 5\%$, são dados na Tabela 15.

Tabela 15 – Estimativas de Máxima Verossimilhança dos parâmetros para o modelo WG com mecanismo de PA.

Parâmetro	**Estimativa (Est)**	**Erro Padrão (EP)**	**|Est| / EP**
alpha	2.1159350	0.07233015	29.253846
$\beta_{10\text{ - Intercepto}}$	0.8054218	0.10674085	7.545.582
$\beta_{11\text{ – Sexo}}$	-0.4227374	0.10803861	3.912836
$\beta_{12\text{ – Carteira}}$	-0.6542153	0.10861169	6.023433
$\beta_{20\text{ - Intercepto}}$	-5.1878446	0.18534694	27.989912
$\beta_{21\text{ – Estado Civil}}$	-0.8506819	0.11964500	7.110049

Fonte: Elaborado pelo autor.

Na Figura 13 tem-se o gráfico da função de sobrevivência da fração de clientes adimplentes estratificada por sexo e renda dos clientes. Como se observa, as curvas assumem valores muito próximos no decorrer do tempo, com a curva se tornando plana por volta de 2,5 anos.

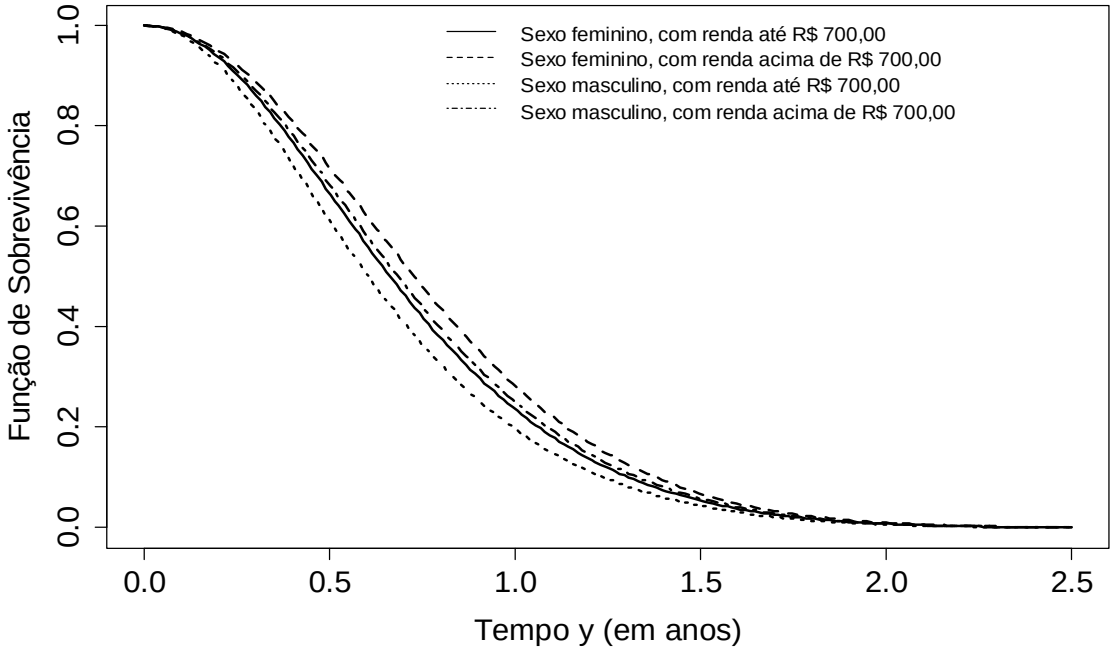

Figura 13: Curvas de sobrevivência dos clientes adimplentes estratificada por sexo e renda.

Fonte: Elaborado pelo autor.

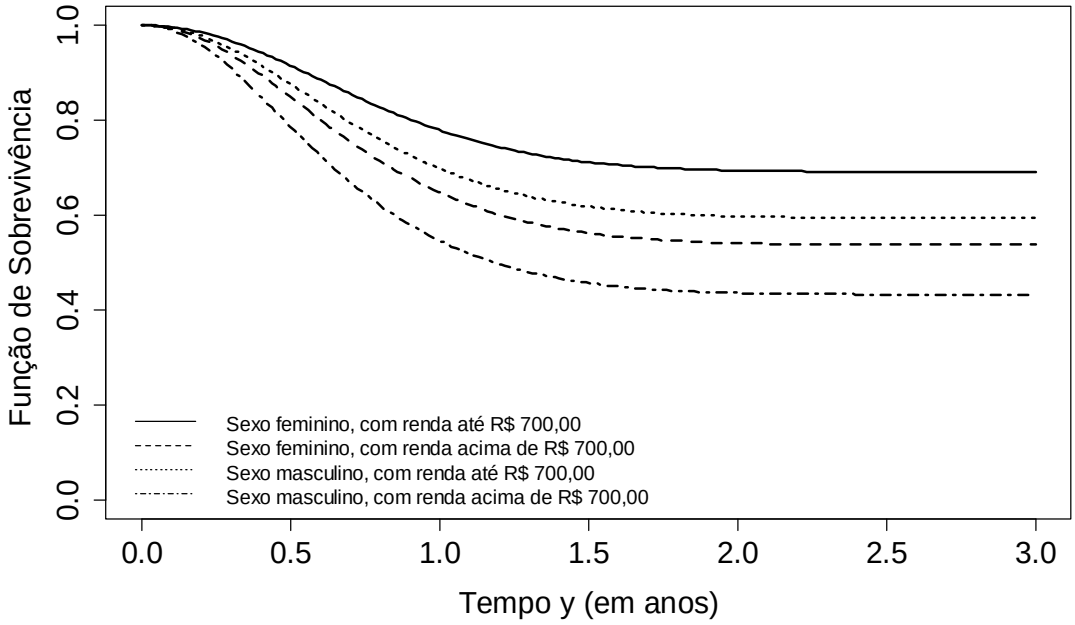

Figura 14: Curvas de sobrevivência populacional (todos os clientes) estratificada por sexo e renda.

Fonte: Elaborado pelo autor.

Na Figura 14 observa-se o gráfico da função de sobrevivência populacional com diferenciação quanto ao sexo e renda dos clientes, onde se observa que clientes do sexo masculino com renda acima de R$ 700,00 apresentam o menor nível na função de sobrevivência, ou seja, menor probabilidade de adimplência, e observa-se que a curva apresenta-se plana a um nível pouco acima de 0,4, a partir do segundo ano após a obtenção do empréstimo, apresentando, assim, a maior probabilidade de ocorrência de inadimplência no decorrer do tempo. Similarmente, pode-se verificar que clientes do sexo feminino com renda acima de R$ 700,00 apresentam a segunda maior probabilidade de ocorrência de inadimplência no decorrer do tempo, verificando-se que a curva apresenta-se plana a um nível pouco acima de 0,5, a partir do segundo ano. A curva da função de sobrevivência de clientes do sexo masculino com renda de até R$ 700,00 apresenta-se plana a um nível próximo de 0,6 também a partir do segundo ano após a obtenção do empréstimo. Por fim, clientes do sexo feminino com renda de até R$ 700,00 apresentam o maior nível na função de sobrevivência, verificando-se que a curva apresenta-se plana a um nível próximo a 0,7 após dois anos da obtenção do empréstimo, apresentando, assim, a menor probabilidade de ocorrência de inadimplência no decorrer do tempo.

Os resultados do estudo reforçam a necessidade de manter o atendimento ao cliente e acompanhamento quanto aos pagamentos das prestações mensais em todo o período de empréstimo, com evidência que os clientes de gêneros diferentes e com rendas diferentes se comportam diferentemente para o pagamento do empréstimo estudado, com destaque para os clientes do gênero masculino com renda acima de R$ 700,00 que apresentam maior probabilidade de se tornarem inadimplentes, e no outro extremo os clientes do gênero feminino com renda até R$ 700,00 que apresentam a menor probabilidade de ocorrência de inadimplência no decorrer do pagamento do empréstimo. Esses resultados podem contribuir para a análise de concessão de créditos da instituição financeira na modalidade de crédito estudada (CDC), complementando as ferramentas de escore de crédito e subsidiando a tomada de decisão.

5 CONSIDERAÇÕES FINAIS

5.1 Conclusões e limitações do estudo

Neste trabalho foram propostas duas metodologias estatísticas para análise de dados de escore de crédito de clientes de um sistema bancário, considerando-se como primeira metodologia a análise discriminante multivariada e a segunda metodologia a análise de sobrevivência. Com a metodologia de analise discriminante foi desenvolvida uma função de discriminação linear (24), p. 69, denominado modelo de escore de crédito, o qual foi utilizado para verificar se os clientes foram corretamente classificados nos grupos a priori existentes (de clientes adimplentes e de clientes inadimplentes). O modelo foi validado utilizando-se de validação cruzada após o ajuste da amostra inicial, e apresentou proporção de sucesso de 72,4% na classificação dos clientes (adimplentes e inadimplentes). Para o grupo de clientes adimplentes a proporção de classificações corretas foi de 67,4%, e para os clientes inadimplentes de 77,4%. Esse modelo (24) de escore de crédito possibilitou a classificação de novos clientes requerentes de crédito em um dos grupos conhecidos, de acordo com fatores relacionados ao tomador do crédito (como sua idade) e outros fatores financeiros relacionados ao empréstimo solicitado (como o prazo e a taxa de juro, valor da prestação e do contrato). Essa classificação permite à instituição bancária outra forma de análise dos empréstimos a serem concedidos, além dos métodos e ferramentas já utilizadas, podendo o modelo agregar informações a serem consideradas nas estratégias relacionadas à oferta do crédito na modalidade estudada (CDC).

A taxa de juros dos empréstimos impacta na decisão do cliente quanto à escolha da instituição de crédito onde tomará o crédito, pois geralmente procura-se pela instituição com as menores taxas praticadas pelo mercado. O estudo de Miola e Barriga (2011) apresenta uma análise das taxas de juros praticadas por algumas instituições financeiras com atuação no Brasil, quanto a algumas modalidades de créditos, dentre elas o CDC, com evidência do grupo de instituições que praticavam as menores taxas de juros nas modalidades estudadas. Além disso, foi possível identificar que taxas de juros mais elevadas nos empréstimos exercem maior influência que as demais variáveis na ocorrência de inadimplência dos contratos, e, opostamente, a idade (mais elevada) do cliente exerce maior influência que as demais variáveis na manutenção da adimplência dos contratos.

Diante dos resultados obtidos pode-se concluir que os objetivos da pesquisa quanto à aplicação da técnica estatística de AD foram atingidos. Todavia, o modelo de escore de crédito encontrado poderia incluir a renda dos clientes como variável independente para analisar qual a influência desse fator na inadimplência ou adimplência dos clientes.

Com a metodologia de análise de sobrevivência foi obtido outro modelo (8), p. 48, para a modelagem dos tempos até o cliente tornar-se inadimplente. Em particular foi considerado um modelo com fração de cura (fração de adimplentes). A partir desse modelo proposto foram identificados os diferentes fatores que influenciam na proporção de clientes adimplentes e no risco de um cliente tornar-se inadimplente, e sua aplicabilidade foi discutida na área de escore de crédito para clientes pessoa física do banco em estudo na modalidade de CDC em relação ao sexo e renda dos clientes, onde se verificou que esses fatores influenciam na ocorrência de inadimplência dos empréstimos tomados.

Do ponto de vista prático, o mecanismo de PA está de acordo com o fato de que se supõe que um cliente pode ser considerado inadimplente quando parou de pagar o empréstimo como primeiro recurso que poderia utilizar. Por outro lado, no entanto, o pressuposto de PA pode ser questionado no escore de crédito do cliente, uma vez que pode ser considerado que um conjunto de causas pode afetar o não pagamento de cada um dos clientes, e que se observou apenas a causa quanto ao tempo de vida mínimo ou máximo (tempo até a ocorrência da inadimplência). Embora estes pressupostos realmente dependam do tipo de carteira em análise, vale a pena considerá-los a partir de uma perspectiva de modelagem de crédito. Além disso, o estudo aponta para uma possível maior aceitação do empréstimo na carteira considerada para clientes do sexo feminino, enquanto mostra que cuidados especiais devem ser tomados pelo banco ao fornecer os empréstimos para o público masculino, e que a renda também deve ser considerada na análise do crédito.

Finalmente, é importante ressaltar que os modelos de escore de crédito obtidos poderiam ser utilizados em conjunto e complementarmente às ferramentas existentes na instituição financeira na análise de solicitações de crédito na modalidade CDC. Esses modelos podem ser considerados preliminarmente ao desenvolvimento de um modelo preditivo que apresente uma probabilidade de não ocorrência de inadimplência com base em um conjunto de covariáveis, ou serem incorporados às ferramentas de escore de crédito da instituição.

Diante dos resultados obtidos pode-se concluir que os objetivos da pesquisa quanto à aplicação da técnica estatística de AS com fração de clientes adimplentes foram atingidos.

A importância da pesquisa está no fato de que pode contribuir para o estudo das estratégias de instituições bancárias na oferta de crédito, que poderão utilizar os resultados da análise como conhecimento do atual comportamento da inadimplência na carteira de crédito estudada, possibilitando a utilização de medidas preventivas em relação à ocorrência da inadimplência. Para a comunidade científica, a pesquisa contribui para o estado da arte pois refere-se à aplicação de técnicas estatísticas comumente utilizadas em diferentes áreas, em dados da carteira de crédito de uma instituição financeira.

5.2 Sugestões para trabalhos futuros

Para futuras pesquisas pode-se considerar a aplicação deste estudo em outras empresas do setor financeiro para créditos concedidos na modalidade CDC. Pode-se estudar e discutir sua aplicabilidade, também, em outras modalidades de crédito, como Cheque Especial, Crédito Imobiliário ou Crédito para obtenção de Veículos. Também podem ser discutidos os modelos a partir de outras (ou mais) variáveis, além das utilizadas no presente trabalho. Em relação ao modelo de escore de crédito obtido com a AD, pode-se obter um novo modelo após a exclusão, considerando a amostra utilizada, daqueles clientes com resultados discrepantes (*outliers*) na tentativa de obtenção de um modelo que classifique com maior precisão os clientes inadimplentes.

Em relação ao modelo de sobrevivência com fração de cura obtido pode-se realizar simulações com diferentes tamanhos de amostra para verificação e discussão dos valores resultantes em relação ao mecanismo de ativação a ser considerado, objetivando verificar e utilizar aquele que melhor se ajuste ao modelo e aos dados utilizados.

REFERÊNCIAS

AKKOÇ, S. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. **European Journal of Operational Research**, v. 222, p. 168-178, 2012.

ALESSANDRI, P.; DREHMANN, M. An economic capital model integrating credit and interest rate risk in the banking book. **Journal of Banking & Finance**, v. 34, p. 730–742, 2010.

ALTMAN, E. I. The success of business failure prediction models. **Journal of Banking & Finance**, v. 8, p. 171-198, 1984.

ALTMAN, E.I., Saunders, A., Credit risk measurement: Developments over the last 20 years. **Journal of banking and Finance**, v. 21, p. 1721–1742, 1998.

ALVES, Karina Lumena de Freitas. **Análise de sobrevivência de bancos privados no Brasil.** Dissertação (Mestrado em Engenharia de Produção) – Universidade de São Paulo, São Paulo, 2009.

ANDERSON, R. H.; FARRAR, D. B.; THOMS, S. R. Aplication of discriminant analysis with clustered data to determine anthropogenic metals contamination. **Science of the total environment**, v. 408, p. 50-56, 2009.

ANNIBAL, C. A. **Inadimplência do Setor Bancário Brasileiro: uma avaliação de suas medidas**. Trabalhos para Discussão n. 192 – Banco Central do Brasil, p. 1-36, 2009.

BANASIK, J.; CROOK, J. Reject inference in survival analysis by augmentation. **Journal of the operational research society**, v. 61, ed. 3, p. 473-485, 2010.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, v. 47, p. 501-515, 1952.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. **Journal of the Royal Statistical Society**, v. 11, p. 15-53, 1949.

BOCK, H. H. Probabilistic models in cluster analysis. **Computational Statistics & Data Analysis**, v. 23, p. 5-28, 1996.

BURNHAM, K. P.; ANDERSON, D. R. **Model Selection and Inference - A practical information – theoretic approach**. Ed. Springer, 488p, 1998.

CANCHO, V. G.; BOLFARINE, H. Modeling the presence of immunes by using the exponentiated-Weibull model. **Journal of Applied Statistics**, v. 28, p. 659-671, 2001.

CANCHO, V.; LOUZADA-NETO, F.; BARRIGA, G. D. C. The geometric birnbaum-saunders regression model with cure rate. **Journal of Statistical Planning and Inference**, v. 142, p. 993-1000, 2012.

CHAMBLESS, L. E.; CUMMISKEY, C. P.; CUI, G. Several methods to assess improvement in risk prediction models: Extension to survival analysis. **Statistics Medicine**, v. 30, p. 22-38, 2011.

CHAMPLAIN, A. F. Limitations of Common Regression Models to Analyze Performance Trends. **National Board of Medical Examiners**, Philadelphia, PA, USA, 2010.

CHEN, M. H. IBRAHIM, J. G. Maximum likelihood methods for cure rate models with missing covariates. **Biometrics**, v. 57, p. 43-52, 2001.

CHI, Y. Y. IBRAHIM, J. G. Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions. **Statistica Sinica**, v. 17, p. 445-462, 2007.

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

BANCO CENTRAL DO BRASIL. **Comitê da Basiléia de Supervisão Bancária.** Princípios gerais para continuidade de atividades. Fórum conjunto. Bank for International Settlements, 2006. Disponível em <http://www.bcb.gov.br/htms/spb/Principios_gerais_continuidade_atividades.pdf>. Acesso em Set. de 2013.

COONER, F.; BANERJEE, S.; MC BEAN, A. M. Modeling geographically referenced survival data with a cure fraction. **Statistical Methods in Medical Research**, v.15, p. 307-324, 2006.

COONER, F.; BANERJEE, S.; CARLIN, B. P. Flexible cure rate modeling under latent activation schemes. **Journal of the American Statistical Association**, v. 102, 560-572, 2007.

CORRAR, L. J., PAULO, E., DIAS FILHO, J. M. (Coord). **Análise Multivariada para os Cursos de Administração, Ciências Contábeis e Economia.** São Paulo: Atlas, 2007.

COSTA, A. C. A. e NAKANE, M. I. Revisitando a metodologia de decomposição do spread bancário no Brasil. **Anais do XXVI Encontro da Sociedade Brasileira de Econometria**. 2004.

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

COX, D. R.; OAKES, D. **Analysis of survival data**. New York: Chapman and Hill, 1984.

COX, D. R.; SNELL, E. **Analysis of binary data.** London: Chapman and Hall, 1989.

DIVINO, J. A.; ROCHA, L. C. S. Probability of default in collateralized credit operations. **The North American Journal of Economics and Finance**, v. 25, p. 276-292, 2013.

DOUMPOS *et al.* Credit risk assessment using a multicriteria hierarchical discrimination approach: A comparative analysis. **European Journal of Operational Research**, v. 138, p. 392–412, 2002.

DONG, G.; LAI, K. K.; YEN, J. Credit scorecard based on logistic regression with random coefficients. **Procedia Computer Science – International Conference on Computational Science**, v. 1, p. 2463-2468, 2010.

DUNN, K. P.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, v. 5, p. 1-10, 1996.

DURBIN, J. Maximum likelihood estimation of the parameters of a system of simultaneous regression equations. **Econometric Theory**. USA. v. 4, p. 159-170, 1988.

EIFERT, D. S. **Análise quantitativa na concessão de crédito versus inadimplência: um estudo empírico**. Dissertação (Mestrado em Administração) – Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 2003.

EISENHARDT, K. M. Building theories from case study research. **Academy of Management Review**, v. 14, n. 4, p. 532-550, 1989.

EVRENSEL, A. Y. Banking crisis and financial structure: A survival-time analysis. **International Review of Economics and Finance**, v. 17, p. 589-602, 2008.

FLORIANO, E. P. *et al*. Ajuste e seleção de modelos tradicionais para série temporal de dados de altura de árvores. **Revista Ciência Florestal**. Santa Maria, v. 16, n. 2, p. 177-199, 2008.

GIBBERT, M.; RUIGROK, W. The ‘‘What’’ and ‘‘How’’ of case study rigor: Three strategies based on published work. **Organizational Research Methods,** v. 13, n. 4, p. 710-737, 2010.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. São Paulo: Atlas, 1996.

GRUNERT, J.; NORDEN, L.; WEBER, M. The role of non-financial factors in internal credit ratings. **Journal of Banking & Finance**, v. 29, p. 509-531, 2005.

GÜRTLER, M.; HIBBELN, M. Improvements in loss given default forecasts for bank loans. **Journal of Banking & Finance**, v. 37, p. 2354-2366, 2013.

HAIR, J.; F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. **Análise multivariada de dados.** 5ª ed. Porto Alegre: Bookman, 2005.

HEIJDEN, H. V. D. Decision support for selecting optimal logistic regression model. **Expert System with Applications**, v. 39, p. 8573-8583, 2012.

HOSMER, D.; LEMESHOW, S.; MAY, S. **Applied survival analysis: regression modeling of time to event data**. 2ª Ed. New York, New York: John Wiley & Sons, 2008.

HUANG, J. Y.; GUO, X. P.; QIU, Y. B.; CHEN, Z. Y. Cluster and discriminant analysis of electrochemical noise data. **Electrochemical Acta**, v. 53, p. 680-687, 2007.

IBRAIM, J. G., CHEN, M.-H. & SINHA, D. **Bayesian Survival Analysis**. Springer, New York, 2001.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. Sixth edition. New Jersey: Prentice Hall, 2007.

KIM, S., CHEN, M. H., DEY, D. K. GAMERMAN, D. Bayesian dynamic models for survival data with a cure fraction. **Lifetime Data Analysis**, v. 13, p. 17-35, 2007.

LAM, K. F.; MOY, J. W. Combining discriminant methods in solving classification problems in two-group discriminant analysis. **European Journal of Operational Research**, v. 138, p. 294-301, 2002.

LOUZADA-NETO, F.; FERREIRA-SILVA, P.; DINIZ, C. A. R. On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data. **Expert Systems with Applications**, v. 39, p. 8071-8078, 2012.

LOUZADA-NETO, F. Poly-hazard regression models for lifetime data. **Biometrics**, v. 55, p. 1121-1125, 1999.

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

MA, Ruo-wei; TANG, Chun-yang. Building up default predicting model based on logistic model and misclassification loss. **System Engineering – Theory & Practice**, v. 27, 8ª edição, p. 33-38, 2007.

MIGUEL, P. A. C. (Coord.). **Metodologia de pesquisa em engenharia de produção e gestão de operações**. Rio de Janeiro: Elsevier, 2010.

MIOLA, R. F.; BARRIGA, G. D. C. Um estudo comparativo das taxas de juros praticadas pelas instituições bancárias com atuação no Brasil na oferta de crédito. **Anais do XVIII Simpósio de Engenharia de Produção**. 2011.

MIZOI, M. F., BOLFARINE, H. LIMA, A. C. P. Cure rate model with measurement error. **Communications in Statistics** - *Theory and Methods*, v. 36, p. 185-196, 2007.

NAGHETTINI, M.; PINTO, E. J. A. **Hidrologia estatística**. Belo Horizonte: CPRM, 2007.

NIKOLIC *et. al.* The application of brute force logistic regression to corporate credit scoring models: Evidence from Serbian financial statement. **Expert Systems with Applications,** v. 40, p. 5932-5944, 2013.

NIE, G.; ROWE, W.; ZHANG, L.; TIAN, Y.; SHI, Y. Credit card churn forecasting by logistic regression and decision tree. **Expert Systems with** Applications, v. 38, p. 15273-15285, 2011.

ORTEGA, E.; CANCHO, V.; LACHOS, V. Assessing influence in survival data with a cure fraction and covariates. **Statistics and Operations Research Transactions (SORT)**, v. 32, p. 115-140, 2008.

ORTEGA, E.; CANCHO, V. G.; PAULA, G. A. Generalized log-gamma regression models with cure fraction. **Lifetime Data Analysis**, v. 15, p. 79-106, 2009.

ORTEGA, E.; CORDEIRO, G M.; HASHIMOTO, E. M. A Log-Linear Regression Model for the Beta-Weibull Distribution. **Communications in Statistics: Simulation & Computation,** v. 40, p. 1206-1235, 2011.

PAIVA, C. S. M.; FREIRE, D. M. C.; CECATTI, J. G. Modelos Aditivos Generalizados para Posição, Escala e Forma (GAMLSS) na Modelagem de Curvas de Referência. **Revista Brasileira de Ciências da Saúde**. João Pessoa, v.12, n.3, p.289-310, 2008.

PASIOURAS, F.; TANNA, S. The prediction of bank acquisition targets with discriminant and logit analyses: Methodological issues and empirical evidence. **Research in International Business and Finance,** v. 24, p. 36-61, 2010.

PERDONÁ, G. S. C. **Modelos de riscos aplicados à análise de sobrevivência**. Tese de doutorado, Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo, São Carlos/SP, 2006.

POPE, C.; MAYS, N., Reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health service research. **British Medical Journal**, v. 311, p. 42-45, 1995.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 54, p. 507-554, 2005.

RODRIGUES, J.; CANCHO, V.; DE CASTRO, M.; LOUZADA-NETO, F. On the unification of long-term survival models. **Statistics and Probability Letters**, v. 79, 753-759, 2009.

SAMOLADAS, I.; ANGELIS, L.; STAMELOS, I. Survival analysis on the duration of open source projects. **Information and Software Technology**, v. 52, p. 902–922, 2010.

SANTOS, J. O. **Análise de Crédito: Empresas, Pessoas Físicas, Agronegócios e Pecuária.** 3. ed. São Paulo: Atlas, 2009.

SCHRICKEL, W. K. **Análise de Crédito: Concessão e Gerência de Empréstimos.** 5. ed. São Paulo: Atlas, 2000.

SOHN, S. Y.; SHIN, H. W. Reject inference in credit operations based on survival analysis. **Expert Systems with Applications,** v. 31, p. 26–29, 2006.

STRAPASSON, E. **Comparação de modelos com censura intervalar em análise de sobrevivência**. Tese (Doutorado em Agronomia) – Universidade de São Paulo. Piracicaba - São Paulo, 2007.

SCHWARZ, G. Estimating the Dimension of a Model. **Annals of Statistics**, v. 6, n. 2, p. 461-464, 1978.

TABACHNICK, B. G.; FIDELL, L.S. **Using Multivariate Statistics.** New York. Sixth edition: Pearson, 2012.

TANG, D. Y.; YAN, H. Market conditions, default risk and credit spreads. **Journal of Banking & Finance**, v. 34, p. 743–753, 2010.

THOMAS, L. C. Modeling the credit risk for portfolios of consumer loans: Analogies with corporate loan models. **Mathematics and Computer in Simulations,** v. 79, p. 2525-2534, 2009.

THOMAS, L. C. Consumer finance: Challenges for operational research. **Journal of the Operational Research Society**, v. 61, p. 42-52, 2010.

TONG, E. N. C.; MUES, C.; THOMAS, L. C. Mixture cure models in credit scoring: If and when borrowers default. **European Journal of Operations Research**. v. 218, p. 132-139, 2012.

TSODIKOV, A. D.; IBRAHIM, J. G.; YAKOVLEV, A. Y. Estimating cure rates from survival data: an alternative to two-component mixture models. **Journal of the American Statistical Association**, v. 98, p. 1063-1078, 2003.

VAN BUUREN, S; FREDRIKS, A. M. Worm plot: A simple diagnostic device for modeling growth reference curves. **Statistics in Medicine**, v. 20, p. 1259-1277, 2001.

WEIBULL, W. A statistical theory of the strength of materials. **Ingenious Vetenskaps Akademien - Handlingar**, v. 151-3, p. 45–55, 1939.

WEIBULL, W. A statistical distribution function of wide applicability. **Journal Applied Mechanical**, v.18, p. 293–297, 1951

XU, Young.; YANG, Jing-yu; YANG, Jian. A reformative kernel Fisher discriminant analysis. **Pattern Recognition**, v. 37, p. 1299-1302, 2004.

YAKOVLEV, A. Y. ; TSODIKOV, A. D. **Stochastic Models of Tumor Latency and Their Biostatistical Applications**. World Scientific, Singapore, 1996.

YIN, G. Bayesian cure rate frailty models with application to a root canal therapy study. **Biometrics**, v. 61, p. 552-55, 2005.

ZAIDER, M. et. al. A survival model for fractionated radiotherapy with an application to prostate cancer. **Physics in Medicine and Biology**, v. 46, p. 2745-2758, 2001.

ZENG, W.; LI, X.; ZHANG, X.; CHENG, E. Kernel-based nonlinear discriminant analysis using minimum squared errors criterion for multiclass and undersampled problems. **Signal Processing,** v. 90, p. 2333-2343, 2010.