Universidade de Brasília - UnB

Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação (FACE)

Programa de Pós-graduação em Administração (PPGA)

Curso de Mestrado Acadêmico

FABIO AUGUSTO SCALET MEDINA

Regressão Logística Geograficamente Ponderada Aplicada a Modelos de *Credit Scoring*

Brasília-DF
2016

Universidade de Brasília - UnB

Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação (FACE)

Programa de Pós-graduação em Administração (PPGA)

Curso de Mestrado Acadêmico

FABIO AUGUSTO SCALET MEDINA

Regressão Logística Geograficamente Ponderada Aplicada a Modelos de *Credit Scoring*

Dissertação apresentada ao Programa de Pós-Graduação em Administração (PPGA) da Universidade de Brasília (UnB) como requisito à obtenção do título de Mestre em Administração

Área de Concentração:
Finanças e Métodos Quantitativos

Orientador:
Prof. Dr. Pedro Henrique Melo Albuquerque

Brasília-DF
2016

FICHA CATALOGRÁFICA

MEDINA, Fabio Augusto Scalet.

Regressão Logística Geograficamente Ponderada Aplicada a Modelos de *Credit Scoring*. /. - Brasília, 2016, 92 p.

Dissertação (Mestrado) - Programa de Pós-Graduação em Administração da Universidade de Brasília – UnB. Área de Concentração: Finanças e Métodos Quantitativos.

Orientador: Prof. Dr. Pedro Henrique Melo Albuquerque.

FABIO AUGUSTO SCALET MEDINA

Regressão Logística Geograficamente Ponderada Aplicada a Modelos de *Credit Scoring*

Dissertação apresentada ao Programa de Pós-Graduação em Administração (PPGA) da Universidade de Brasília (UnB) como requisito à obtenção do título de Mestre em Administração

Área de Concentração:
Finanças e Métodos Quantitativos

BANCA EXAMINADORA:

Prof. Dr. Pedro Henrique Melo Albuquerque
Universidade de Brasília- PPGA
Orientador

Prof. Dr. Otávio Ribeiro de Medeiros
Universidade de Brasília - PPGA
Examinador Interno

Prof. Dr. Bernardo Borba de Andrade
Universidade de Brasília- PGEST
Examinador Externo

Brasília-DF, 27 de abril de 2016

Resumo

A presente dissertação de mestrado teve como objetivo principal verificar a aplicabilidade da metodologia Regressão Logística Geograficamente Ponderada (GWLR) para a construção de modelos de credit scoring. As fórmulas do melhor conjunto de modelos locais estimados via GWLR foram comparadas entre si, em termos de valor dos coeficientes e significância das variáveis, e frente ao modelo global estimado via Regressão Logística. Foram utilizados dados reais referentes às operações de Crédito Direto ao Consumidor (CDC) de uma instituição financeira pública nacional concedidas a clientes domiciliados no Distrito Federal (DF). Os resultados encontrados demonstraram a viabilidade da utilização da técnica GWLR para desenvolver modelos de credit scoring. Os modelos estimados para cada região do DF se mostraram distintos em suas variáveis e coeficientes (parâmetros) e três dos cinco indicadores do modelo via GWLR se mostraram superiores aos do modelo via Regressão Logística.

Palavras-chave: Risco de Crédito, Credit Scoring, Regressão Logística Geograficamente Ponderada.

As ideias e opiniões expostas nesse estudo são de responsabilidade do autor, não refletindo a opinião e posição da instituição financeira fornecedora dos dados.

Abstract

This master thesis aimed to verify the applicability of the methodology Geographically Weighted Logistic Regression (GWLR) to develop credit scoring models. The formulas of the best set of local models estimated by GWLR were compared in terms of value of the coefficients and significance of the variables, and against the global model estimated by Logistic Regression. It was used a real granting data of Direct Credit Consumer from a national public financial institution to borrowers domiciled in the Federal District (FD) of Brazil. The results demonstrated the feasibility of using the technique GWLR to develop credit scoring models. The estimated models for each region of FD have showed to be different in their variables and coefficients (parameters) and three out of five indicators calculated for the developed model by GWLR were superiors than indicators of the developed model by Logistic Regression.

Key-words: Credit Risk, Credit Scoring, Geographically Weighted Logistic Regression.

SUMÁRIO

1. INTRODUÇÃO

A principal atividade dos bancos comerciais é a intermediação financeira, que consiste em captar recursos financeiros e emprestá-los a terceiros em condições preestabelecidas tais como prazo de pagamento, valor de prestação e taxa de juros (HAND e HENLEY, 1997). Por envolver expectativa futura de recebimento, todo crédito concedido está exposto a riscos.

O risco de crédito pode ser definido como a possibilidade de ocorrência de perdas financeiras associadas ao não cumprimento pelo tomador ou contraparte de suas respectivas obrigações financeiras nos termos pactuados, à desvalorização de contrato de crédito decorrente da deterioração na classificação de risco do tomador, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação (BACEN, 2009) e é um dos principais riscos ao qual uma instituição financeira está exposta.

O tema gerenciamento de riscos se estabeleceu definitivamente no setor financeiro após a divulgação dos acordos de Basiléia, conjuntos de documentos que embasaram a regulação e fiscalização do setor. Os avanços tecnológicos e computacionais aliados ao desenvolvimento de métodos quantitativos culminaram na criação de diversas ferramentas para mensuração de riscos (THOMAS, 2000).

Nesse contexto, o financista utiliza-se dessas ferramentas e metodologias quantitativas para gerar informações relevantes sobre os riscos aos quais a instituição financeira está exposta, visando minimizar o montante de perdas financeiras, diminuir o valor total provisionado da instituição junto ao órgão regulador e consequentemente melhorar seus resultados financeiros.

As metodologias quantitativas que podem ser aplicadas para a mensuração do risco de crédito variam de acordo com o momento ou a situação do contrato no ciclo de crédito. Os modelos aplicados na concessão de crédito são denominados modelos de *credit scoring* (CS) e possuem o objetivo de quantificar o risco de crédito através da previsão da probabilidade de perda financeira no momento da decisão de concessão (SICSÚ, 2010). Essa quantificação do risco no momento da concessão é de suma importância para o resultado financeiro da empresa pois, diminuindo o número de

tomadores inadimplentes entrantes em sua carteira de crédito, a instituição diminuirá o montante de provisão a ser feito junto ao órgão regulador para fazer frente a esse risco e também diminuirá os gastos com cobrança e recuperação de crédito inadimplente.

Sicsú (2010) destaca as seguintes vantagens da utilização de modelos de *credit scoring*:

1. Consistência nas decisões: Um tomador de crédito possuirá o mesmo escore na instituição financeira independente do analista, da agência bancária ou filial que realizar a avaliação, eliminando assim a subjetividade;
2. Rapidez nas decisões: Recursos computacionais permitem que o escore seja computado logo após o cadastro das informações necessárias para sua estimação, permitindo ao analista uma pronta resposta a um cliente potencial, trazendo vantagem competitiva para a instituição financeira;
3. Gestão do risco de crédito e precificação das operações: O conhecimento da probabilidade de perda de cada cliente permite o cálculo das perdas esperadas das carteiras de operações de crédito, utilizadas para precificar as operações e melhor gestão das carteiras;
4. Definição de políticas de crédito: Através dos escores, os clientes podem ser divididos em classes de risco, permitindo às instituições a adoção de diferentes regras de concessão de crédito para cada classe, como, por exemplo, a redução da taxa de juros a clientes de baixo risco ou a incorporação de garantia (colateral) à operação na concessão à clientes de alto risco;
5. Redução do custo operacional: Os analistas de crédito necessitam de menor experiência em avaliação de concessão de crédito, permitindo às instituições a contratação de mão-de-obra menos qualificada e a redução de gastos com treinamentos.

Segundo Hand e Henley (1997), diante do menor custo e da maior agilidade, objetividade e poder preditivo na decisão da concessão de crédito, os modelos de *credit scoring* se popularizaram e são amplamente utilizados pelo setor financeiro.

Para o seu desenvolvimento, os modelos de CS utilizam-se de informações históricas dos tomadores de crédito, da operação contratada e do comportamento de

pagamento para, através de uma combinação ou ponderação dessas características, produzirem uma pontuação quantitativa, denominada escore (do inglês *score*).

A regressão logística é o método mais utilizado para se obter uma regra de classificação quando a variável preditiva que se deseja analisar é binária. Lessmann et al. (2015) realizaram uma abrangente pesquisa sobre as metodologias de classificação utilizadas para o desenvolvimento de modelos de *credit scoring*, elencando e verificando a acurácia de quarenta e uma (41) metodologias distintas e apontaram a regressão logística como a metodologia padrão do setor financeiro.

A regressão logística é uma técnica de análise multivariada que busca explicar a relação entre uma variável aleatória dicotômica dependente e um conjunto de variáveis explicativas independentes (HOSMER e LEMESHOW, 2000).

Uma instituição financeira possui diversos modelos de *credit scoring* que são aplicados na avaliação de diferentes tipos de clientes (Pessoa Física, Pessoa Jurídica categorizadas por porte de faturamento) ou da operação de crédito a ser contratada. As variáveis explicativas que compõem os modelos podem ser distintas, visando melhorar a predição do risco de crédito do seu público alvo. A localização geográfica do tomador de crédito pode ser uma variável a compor modelos de *credit scoring*, mas qual é a melhor maneira de utilizá-la?

O uso do Código de Endereçamento Postal (CEP) pode ser uma opção para introduzir a informação de localização geográfica nesses modelos, no entanto, por ser uma variável qualitativa com grande número de categorias, pode produzir modelos não parcimoniosos e prejudicar a avaliação de indivíduos de regiões com poucas informações (FERNANDES e ARTES, 2015). Uma solução para incluir o CEP em modelos de CS seria utilizar apenas seus 2 ou 3 dígitos iniciais, uma vez que os números iniciais já contêm a delimitação geográfica.

Outra opção seria subdividir a amostra de tomadores de crédito de acordo com sua região geográfica e desenvolver um modelo para cada subpopulação. As variáveis que irão compor cada modelo serão distintas ou existirão variáveis em comum? As variáveis em comum entre os modelos possuirão coeficientes (parâmetros) das variáveis diferentes? Essas perguntas serão respondidas nessa dissertação.

Stine (2011) analisa a evolução da inadimplência do crédito imobiliário nos condados dos Estados Unidos durante o período de 1993 a 2010, contemplando um período pré e pós crise do *subprime*[1], ocorrida em 2008. Seu estudo apontou evidências de existência de correlação espacial entre as taxas de inadimplência dos condados.

Fernandes e Artes (2015) aplicam a metodologia *Ordinary Kriging* para criar uma variável que reflete o risco espacial e aplicam a técnica de Regressão Logística para verificar a existência de correlação espacial na inadimplência de pequenas e médias empresas (PME) tomadoras de crédito, utilizando dados do *bureau* de crédito SERASA. Os autores desenvolveram modelos com e sem a variável de risco espacial e confirmaram que a inclusão dessa variável melhora o desempenho dos modelos de credit scoring.

A técnica de Regressão Geograficamente Ponderada, em inglês Geographically Weighted Regression (GWR), proposta por Brunsdon, Fotheringham e Charlton (1996), é utilizada para modelar processos heterogêneos (não-estacionários) espacialmente, isto é, processos que variam (seja na média, mediana, variância, etc.) de região para região. A ideia básica da GWR é ajustar um modelo de regressão para cada região do conjunto de dados utilizando a localização geográfica das demais observações para ponderar as estimativas dos parâmetros.

A vantagem de se utilizar a GWR é a possibilidade de variação dos parâmetros do modelo de acordo com a localização geográfica (ATKINSON et al., 2003), enquanto um modelo de regressão global, representado apenas por uma fórmula, pode não representar as variações locais de forma adequada. A aplicação da técnica GWR pode ser observada em diferentes áreas de pesquisa tais como geografia (SEE et al. 2015), saúde (GILBERT; CHAKRABORTY, 2011) e economia (HUANG; LEUNG, 2002).

Atkinson et al. (2003) utiliza em seu estudo a Regressão Logística Geograficamente Ponderada ou *Geographically Weighted Logistic Regression* (GWLR) para analisar a dependência da localização geográfica na relação entre erosão e controles geomorfológicos de uma região do País de Gales. A variável binária utilizada nesse estudo foi a presença ou ausência de erosão nas áreas estudadas. A aplicação da técnica GWLR resultou na estimação de modelos com diferentes parâmetros (modelos distintos)

[1] Detalhes sobre a crise do *subprime* podem ser encontrados em Ackermann (2008).

para cada área estudada, revelando a necessidade de adoção de diferentes práticas para se evitar a erosão a depender da região.

Algumas situações existentes no contexto de risco de crédito sugerem o desenvolvimento de modelos distintos para cada região de estudo, visando a obtenção de melhores resultados frente a um modelo global de formula única e que não considera a localização geográfica em seu desenvolvimento. Pode-se citar como exemplo de situações um bairro de determinado município que apresenta menor quantidade de clientes inadimplentes frente aos demais bairros, ou ainda um município que possui maior renda per capita e menor taxa de desemprego frente aos demais municípios de sua região ou estado. Essas regiões destacadas nos exemplos provavelmente são regiões de menor risco de crédito frente às demais regiões ao seu redor e por isso é razoável a ideia de aplicar a GWLR para desenvolver modelos de regressão que levem em consideração suas particularidades, composto por variáveis preditoras distintas e/ou com pesos diferentes das demais regiões, para melhor discriminar o risco de crédito dos tomadores ali domiciliados.

Travassos et al. (2013) citam em seu artigo o uso da GWLR para incorporar dados de energia elétrica a modelos de *credit scoring* do segmento de microcrédito, no entanto o artigo apresenta somente resultados referentes ao emprego da Regressão Logística tradicional, sob o argumento de menor complexidade e por apresentar resultados semelhantes à GWLR.

Não foram encontrados outros estudos nacionais ou internacionais que utilizaram a GWLR no desenvolvimento de modelos de *credit scoring*. As buscas foram realizadas no portal de periódicos da CAPES e no *Google Scholar* através das expressões RLGP risco de crédito, RLGP *credit scoring*, GWLR *credit scoring*, e GWLR *credit risk.*

O presente estudo utilizou dados referentes à operação de Crédito Direto ao Consumidor (CDC) concedidos por uma instituição financeira pública nacional a clientes domiciliados no Distrito Federal (DF), com o objetivo geral de verificar a viabilidade da aplicação da técnica Regressão Logística Geograficamente Ponderada (GWLR) no desenvolvimento de modelos de *credit scoring.*

Os objetivos específicos dessa dissertação são:

1. Comparar o conjunto de modelos estimados via GWLR frente ao modelo global estimado via Regressão Logística e verificar qual modelo obtém melhores resultados em termos de capacidade de previsão e perdas financeiras para a instituição;
2. Comparar os coeficientes e variáveis significativas do melhor conjunto de modelos locais estimado via GWLR entre si e verificar se existe diferença entre esses modelos.

A presente dissertação está estruturada em cinco capítulos, na qual o primeiro é a presente introdução, o segundo capítulo apresenta a fundamentação teórica, contendo os conceitos de risco, risco de crédito e modelos de *credit scoring.* O terceiro capítulo apresenta a metodologia utilizada nesse estudo, quais sejam regressão logística e regressão logística geograficamente ponderada e o processo de desenvolvimento dos modelos. O quarto capítulo apresenta os resultados obtidos e o quinto capítulo apresenta a conclusão dessa dissertação.

2. REFERENCIAL TEÓRICO

2.1. Riscos

Existem diferentes definições para o termo risco e em finanças um dos primeiros trabalhos publicados sobre o tema foi Markowitz (1952), que apresenta uma solução teórica para a gestão de risco de uma carteira de ativos, determinado pela variância do retorno de cada título e também pela covariância dos retornos de cada par de ativos.

Markowitz (1952) traz também discussões sobre a fronteira eficiente entre risco e retorno e a aversão ao risco de um investidor. Como reconhecimento aos importantes e pioneiros trabalhos desenvolvidos, Harry M. Markowitz, em conjunto com Merton Miller e William Sharpe, foram laureados com o Prêmio Nobel de Economia de 1990.

No contexto das instituições financeiras, risco pode ser definido como possibilidade de ocorrência de prejuízos financeiros (GITMAN, 1997), sendo a gestão de riscos um tema bastante pesquisado e de suma importância para setor financeiro, principalmente após as publicações dos Acordos de Basileia e maior regulação.

Significativas mudanças no mercado financeiro mundial ao longo do tempo acarretaram em uma crescente preocupação com o gerenciamento dos riscos expostos pelas instituições financeiras. Na década de 70, o colapso de *Bretton Woods*[2] gerou um cenário internacional de crescente incerteza, com câmbio e taxas de juros extremamente voláteis. Nesse cenário, alguns bancos adotaram estratégias de negócio que se mostraram erradas ao longo do tempo, culminando em prejuízos financeiros e falências de diversos bancos internacionalmente ativos (DUARTE JÚNIOR; LELIS, 2004).

Os prejuízos acumulados e a queda abrupta no capital dessas instituições impulsionaram os responsáveis pela supervisão bancária dos países do Grupo dos Dez (G-10) a criarem em dezembro de 1974 o Comitê de Regulamentação Bancária e Práticas de Supervisão, também conhecido como Comitê de Basiléia, com o objetivo padronizar a supervisão e aumentar a solidez e estabilidade do sistema bancário internacional.

Em julho de 1988 o Comitê de Basiléia publica o acordo de Convergência Internacional de Mensuração de Capital e Padrões de Capital, também conhecido como Acordo de Basiléia ou Basiléia I. O Acordo padroniza os conceitos de capital e propõe

[2] O sistema *Bretton Woods* foi criado em 1944 para gerenciar a economia global e evitar crises como as registradas após a Primeira Guerra Mundial. Mais detalhes sobre o assunto podem ser encontrados em Bordo (1993) e Eichengreen (1995).

um conjunto mínimo de diretrizes para o cálculo de adequação de capital em bancos, com o objetivo de reduzir os riscos do sistema bancário internacional, fazendo com que as instituições financeiras mantivessem capital suficiente para cobrir as possíveis perdas de valores dos seus ativos e, desse modo, garantir sua solvência e também minimizar as desigualdades competitivas provenientes de diferenças na alocação de capital exigido a bancos de diferentes países (BCBS, 1988; WAGSTER, 1996).

Por ser considerado o principal risco ao qual as instituições financeiras estariam expostas, Basiléia I teve como tema central o risco de crédito e, inicialmente, o requisito mínimo de capital para fazer frente aos riscos foi estipulado em pelo menos 8% dos ativos ponderados pelo risco (RWA – *risk weigthed asset*) (GOODHART, 2005). Algumas classes de ativos e seus respectivos fatores de ponderação estão contidos na tabela 1.

Tabela 2.1 - Fator de ponderação de algumas classes de ativos

Classe de Ativos	**Fator de Ponderação**
Empréstimos Comerciais	100%
Empréstimo com garantias hipotecárias	50%
Títulos de bancos multilaterais de desenvolvimento	20%
Títulos de governos ou bancos centrais de países da OCDE	0%

Fonte: BCBS (1988).

A título de exemplo, um financiamento imobiliário de R$ 20.000,00, que possui um fator de ponderação de 50%, teria um RWA no valor de R$10.000,00. Assim, o capital alocado pelo banco, referente a essa exposição, seria de pelo menos R$ 800,00, equivalente a 8% do RWA.

Segundo Resti e Sironi (2010), originalmente o acordo se aplicava somente a bancos com atuação internacional, no entanto muitas entidades nacionais, dentre elas Estados Unidos e União Europeia, decidiram torná-lo obrigatório para todos os bancos, incluindo aqueles que atuavam somente nos mercados domésticos.

O órgão responsável por regular e supervisionar o Sistema Financeiro Nacional (SFN) é o Banco Central do Brasil (BACEN), que através da Resolução CMN nº 2.099 de 17/08/1994 (BACEN, 1994) regulamentou a implantação do Acordo de Basiléia I no Brasil. Essa Resolução estabeleceu que as instituições autorizadas a operar no mercado brasileiro deveriam constituir o Patrimônio Líquido Exigido (PLE) em um valor igual a, no mínimo, 8% de seus ativos ponderados por fatores de risco, percentual idêntico ao

estabelecido pelo Comitê de Basiléia (BCBS), no entanto esse índice foi alterado posteriormente para 11% por meio da Circular nº 2.784 de 27/11/1997.

Em junho de 2004, o comitê de Basiléia publicou o Novo Acordo de Capitais da Basiléia ou Basiléia II (BCBS, 2004), estruturado em três pilares: o primeiro pilar trata dos requisitos mínimos de capital que os bancos devem possuir para fazer frente aos riscos, com base nos riscos de crédito, mercado e operacional, propondo metodologias que visam melhor estimação e diferenciação entre esses riscos. O segundo pilar concentra-se nas melhores práticas de supervisão, reforçando a responsabilidade dos órgãos supervisores avaliarem a adequação de capital aos riscos expostos pelas instituições e das instituições financeiras adotarem práticas de gerenciamento de riscos com vasta aceitação e utilização pelo mercado. O terceiro pilar discorre sobre disciplina de mercado, exigindo maior transparência na divulgação de informações sobre gestão e riscos, reduzindo a assimetria informacional (BCBS, 2004; BARTH et al., 2004; ANTÃO; LACERDA, 2011).

Segundo Antão e Lacerda (2011), o acordo de Basiléia II foi extremamente inovador em termos de requerimento de capital associado ao Risco de Crédito, sendo uma dessas inovações o uso dos *ratings* de crédito (internos ou externos) para a avaliação dos requerimentos de capital, que se tornaram sensíveis à qualidade creditícia de cada exposição.

Basileia II permite que as instituições decidam entre duas metodologias para cálculo dos ativos ponderados pelo risco: a abordagem padronizada e a abordagem baseada em *ratings* internos (IRB). A abordagem padronizada consiste na adoção de fatores de ponderação de risco preestabelecidos pelo regulador, que variam de acordo com as categorias de exposições, enquanto nas abordagens IRB as instituições são responsáveis pelo cálculo de alguns parâmetros que necessitam de aprovação pelo regulador, o que possibilita maior sensibilidade na mensuração dos riscos (BCBS, 2004).

As abordagens IRB referentes ao risco de crédito utilizam-se dos seguintes parâmetros: Probabilidade de Descumprimento (PD), Exposição no Momento do Descumprimento (EAD), Perda dado o Descumprimento (LGD) e Prazo Efetivo de Vencimento (M) para apuração do requerimento mínimo de capital (BACEN, 2013). Por esse motivo, o desenvolvimento de modelos para estimação dos parâmetros PD, EAD e LGD se tornaram temas de pesquisa populares (LESSMANN et al., 2015). A seguir seguem as definições dos parâmetros de risco extraídas do Artigo 5º da Circular nº 3.648,

de 04/03/2013 (BACEN, 2013), que estabelece os requisitos mínimos para o cálculo da parcela relativa às exposições ao risco de crédito sujeitas ao cálculo do requerimento de capital mediante sistemas internos de classificação do risco de crédito (IRB) (RWA_{IRB}), bem como de alguns estudos relacionados aos temas:

1. PD (*Probability of Default* ou Probabilidade de Descumprimento) – percentual que corresponde à expectativa de longo prazo das taxas de descumprimento para o horizonte temporal de um ano dos tomadores de um determinado nível de risco de credito (*rating*) ou grupo homogêneo de risco (no caso do Varejo). Trabalhos relacionados ao tema: Medema et al. (2009), Volk (2012).

2. EAD (*Exposure at Default* ou Exposição no Momento do Descumprimento) – corresponde ao valor da exposição da instituição, seja ela efetiva ou contingente, perante o tomador ou contraparte no momento da concretização do evento de descumprimento, bruto de provisões e eventuais baixas parciais a prejuízo. Trabalhos relacionados ao tema: Valvonis (2008) e Jacobs (2010);

3. LGD (*Loss Given Default* ou Perda dado o Descumprimento) – corresponde ao percentual, em relação ao parâmetro EAD observado, da perda econômica decorrente do descumprimento, considerados todos os fatores relevantes, inclusive descontos concedidos para a recuperação do crédito e todos os custos diretos e indiretos associados à cobrança da obrigação. Trabalhos relacionados ao tema: Silva et al. (2009), Calabrese (2014) e Yao et al. (2015);

4. M (*Maturity* ou Prazo Efetivo de Vencimento) – corresponde ao prazo remanescente da operação ponderado pelos fluxos de caixa relativos a cada período futuro. Trabalhos relacionados ao tema: Barco (2004), Petrov e Pomazanov (2009).

A implantação de Basiléia II no Brasil é regulamentada por uma série de normas divulgadas pelo BACEN, disponíveis para consulta em seu site[3]. A utilização da

[3] O conjunto de normas que regulamenta Basileia II no Brasil está disponível no seguinte endereço: http://www.bcb.gov.br/nor/basileia/Basileia_Normativos.asp.

abordagem IRB para alocação de capital referente ao Risco de Crédito ainda se encontra em desenvolvimento pelas instituições nacionais e até o presente momento nenhuma das cinco maiores instituições financeiras brasileiras em número de ativos foi autorizada a utilizar tal abordagem[4].

A crise do *subprime* de 2008 trouxe questionamentos sobre o nível, a qualidade e a pertinência dos mecanismos utilizados pelas políticas de regulação bancária sobre o controle do risco sistêmico (GOODHART, 2008) e contribuiu para a publicação do Acordo de Basiléia III, ocorrido em dezembro de 2010 e revisto em junho de 2011 (BCBS, 2011).

As novas regras apresentadas em BCBS (2011) referem-se à estrutura de capital das instituições financeiras e buscam aperfeiçoar a capacidade das instituições de absorver choques, fortalecendo a estabilidade financeira e a promoção do crescimento econômico sustentável. O aumento da quantidade e qualidade do capital regulamentar mantido por instituições financeiras visa reduzir a probabilidade, a severidade de eventuais crises bancárias e seus consequentes custos para a economia. Também pode-se entender Basiléia III como um esforço global em busca de maior estabilidade dos sistemas bancários via imposição de diversas exigências quanto à manutenção de níveis de liquidez, colchões de capital, reservas, restrições à alavancagem, entre outras, de forma a garantir a maior robustez das instituições bancárias mundiais frente a flutuações econômicas.

Além dos riscos de crédito, de mercado e operacional existem outros tipos de riscos aos quais as instituições financeiras estão expostas, como, por exemplo, o risco de liquidez, o risco legal, o risco reputacional e o risco sistêmico (BCBS, 1997). A seguir são apresentadas suas definições e trabalhos relacionados a cada tema.

1. Risco de Crédito: Por ser objeto dessa dissertação, se encontra detalhado no próximo capitulo;

2. Risco de Mercado: A Resolução CMN nº 3.464, de 26/06/2007 define risco de mercado como a possibilidade de ocorrência de perdas resultantes da flutuação

[4] Foram analisados os balanços financeiros divulgados pelas instituições Banco do Brasil, Itaú-Unibanco, Caixa Econômica Federal, Bradesco e Santander referentes ao primeiro trimestre de 2015, sendo o Bradesco a única instituição autorizada pelo Banco Central para utilizar o modelo IRB para o Risco de Mercado.

nos valores de mercado de posições detidas por uma instituição financeira, incluindo os riscos das operações sujeitas à variação cambial, das taxas de juros, dos preços de ações e dos preços de mercadorias (commodities) (BACEN, 2007). Trabalhos relacionados: Dowd (2007), Jorion (2010) e Chen (2014);

3. Risco Operacional: A Resolução CMN nº 3.380, de 29/06/2006 define risco operacional como a possibilidade de ocorrência de perdas resultantes de falha, deficiência ou inadequação de processos internos, pessoas e sistemas, ou de eventos externos, incluindo o risco legal associado à inadequação ou deficiência em contratos firmados pela instituição, bem como a sanções em razão de descumprimento de dispositivos legais e a indenizações por danos a terceiros decorrentes das atividades desenvolvidas pela instituição. Entre os eventos de risco operacional, incluem-se: fraudes internas, fraudes externas, demandas trabalhistas e segurança deficiente do local de trabalho, práticas inadequadas relativas a clientes, produtos e serviços, danos a ativos físicos próprios ou em uso pela instituição e falhas em sistemas de tecnologia da informação (BACEN, 2006). Trabalhos relacionados: Chavez-Demoulin et al. (2006) e Moscadelli (2004);

4. Risco de Liquidez: Possibilidade de perdas ocorridas devido à insuficiência de recursos para o cumprimento das obrigações da instituição (BCBS, 1997). Trabalho relacionado: Goodhart (2008);

5. Risco Legal: Possibilidade de ocorrência de perdas por falta de suporte das leis ou regulamentações vigentes, incluindo perdas por documentação insuficiente, à execução dos arranjos de liquidação relacionados aos direitos de propriedade e outros interesses que são mantidos pelo sistema de liquidação (DUARTE JÚNIOR, 2001);

6. Risco Reputacional ou de Imagem: Possibilidade de ocorrência de perdas decorrentes da percepção negativa por parte dos clientes, contrapartes, acionistas, investidores, detentores de dívida, analistas de mercado, outros partidos ou reguladores relevantes que podem afetar adversamente a capacidade de um banco para manter ou estabelecer novos relacionamentos de negócio e contínuo acesso

a fontes de financiamento (BCBS, 2009, p. 19). Trabalho relacionado: Haron et al. (2015);

7. Risco Sistêmico: Possibilidade de ocorrência de perdas em virtude de dificuldades financeiras de uma ou mais instituições que provoquem danos substanciais a outras ou ruptura no cenário de normalidade do Sistema Financeiro Nacional - SFN. Trabalho relacionado: Girardi e Ergün (2013) e Rodríguez-Moreno e Peña (2013).

Como o objetivo desse estudo é a obtenção de modelos de previsão de Risco de Crédito, o mesmo será o único tipo de risco apresentado detalhadamente.

2.2. Risco de Crédito

O termo crédito pode ser definido como uma quantidade de dinheiro emprestada por uma instituição financeira a um tomador e que deve ser devolvida com condições preestabelecidas, tais como prazo e taxa de juros (HAND e HENLEY, 1997). Risco de crédito pode ser definido como a possibilidade de ocorrência de perdas financeiras, associadas ao não cumprimento pelo tomador ou contraparte de suas respectivas obrigações financeiras nos termos pactuados, à desvalorização de contrato de crédito decorrente da deterioração na classificação de risco do tomador, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação (BACEN, 2009).

De acordo com Resti e Sironi (2010) e Duarte Júnior (2005), os principais tipos de risco compreendidos pelo risco de crédito são:

1. Risco de inadimplência ou *default*: Possibilidade de ocorrência de perda associada à incapacidade de pagamento da operação de crédito por parte do tomador de crédito;
2. Risco de degradação do crédito ou migração: Possibilidade de ocorrência de perda associada à queda da qualidade creditícia do tomador de crédito, geralmente expressa por rebaixamento de *rating* com base em nova avaliação de risco do cliente ou por agência de classificação externa;

3. Risco de concentração de crédito: Possibilidade de ocorrência de perda associada à alta concentração de operações de crédito em poucos tomadores, poucos setores da economia e poucos ativos;
4. Risco de recuperação da garantia: Possibilidade de ocorrência de perda associada à desvalorização das garantias apresentadas na concessão do crédito, fazendo com que as mesmas não sejam suficientes para cobrir o valor total das obrigações da operação de crédito, ou ainda pela falta de liquidez da garantia no momento da execução da mesma;
5. Risco soberano ou país: Possibilidade de ocorrência de perdas associadas ao não cumprimento de obrigações financeiras nos termos pactuados pelo tomador ou contraparte localizada fora do país, em decorrência de ações realizadas pelo governo do país onde o tomador ou contraparte estão localizados.

Diferentes ferramentas e metodologias são utilizadas pelo setor financeiro para quantificar o risco de crédito de clientes e operações, a distribuição de perdas de carteiras e também para precificar instrumentos financeiros sujeitos ao risco de crédito. Essas ferramentas podem ser classificadas em três categorias: modelos de classificação de risco, modelos estocásticos de risco de crédito e modelos de risco de portfólio (ANDRADE, 2005).

2.2.1. Modelos de Classificação de Risco

Os modelos de classificação de risco avaliam o risco de um tomador ou de uma operação e são utilizados pelas instituições financeiras em seus processos de concessão de crédito. Essas avaliações são expressas através de uma classificação de risco (*rating*) ou pontuação (escore) que representam a expectativa de risco de inadimplência ou *default* desse tomador ou dessa operação de crédito. Dentro dessa categoria de modelos, encontram-se os modelos especialistas, modelos de *credit rating* e modelos de *credit scoring*.

Os modelos especialistas são formados por um conjunto de regras que embasam o analista para a decisão de concessão de crédito. Em sua forma clássica, esses modelos possuíam como principal característica o julgamento subjetivo dessa decisão de

concessão (CAOUETTE et al., 1998). De acordo com Saunders (2000), os sistemas especialistas mais comuns são baseados nos cinco “Cs” do crédito:

1. Caráter: Está associado à índole e à reputação do tomador e sua predisposição em pagar o crédito contraído, podendo ser mensurado através de seu comportamento creditício no mercado e de seu histórico de pagamentos na instituição;
2. Capital: Representa o potencial financeiro do tomador de crédito. A análise da dívida do requerente, os índices de liquidez e as taxas de lucratividade são frequentemente utilizados para avaliar seu capital;
3. Capacidade: Consiste na avaliação da capacidade de o tomador pagar o crédito pleiteado, em que são analisadas as demonstrações financeiras, com ênfase na liquidez e nos fluxos, assim como as projeções de caixa e de endividamento.
4. Colateral: Consiste no somatório de ativos que o tomador oferece em garantia ao empréstimo, aumentando a possibilidade de a instituição financeira reaver os recursos emprestados, caso o tomador do crédito não honre suas obrigações.
5. Condições: Relacionadas ao cenário macroeconômico ou do setor de atuação (no caso de empresas) do solicitante de crédito.

Os modelos especialistas são utilizados atualmente para certos tipos de operações que não possuem massa de dados suficientes ou que essa massa de dados possua uma quantidade irrisória de clientes maus pagadores, o que impossibilita o desenvolvimento de um modelo de *credit scoring.*

Já os modelos de *credit rating* são modelos utilizados para classificar empresas em categorias de risco (*ratings*) e são desenvolvidos internamente pelas instituições financeiras ou por agências externas de *rating*, tais como Moody’s, Standard and Poor’s ou SERASA.

Esses modelos utilizam-se de critérios quantitativos (índices financeiros extraídos das demonstrações contábeis) e qualitativos (qualidade da administração, por exemplo) em sua fórmula para obtenção do *rating*. Em geral, quanto maior o porte da empresa analisada, maior é a influência de critérios qualitativos na atribuição do *rating* (ANDRADE, 2005). Detalhes sobre esses modelos podem ser encontrados em Borges (2001).

Por serem objetos de estudo dessa dissertação, os modelos de *credit scoring* serão detalhados a seguir.

2.2.1.2. Modelos de *Credit Scoring*

De acordo com Thomas (2000), *credit scoring* é, em sua essência, uma ferramenta que permite reconhecer os diferentes grupos que compõem uma população quando não é possível identificar a característica que os separam, mas apenas as variáveis correlatas. O objetivo dos modelos de *credit scoring* é identificar as características do tomador e da operação de crédito que mais determinam a probabilidade de inadimplência e, através de uma combinação ou ponderação dessas características, produzir uma pontuação quantitativa (SAUNDERS, 2000; SICSÚ, 2010).

Dentre as vantagens de se utilizar os modelos de *credit scoring*, Caouette et al (1999, p. 188) destacam a objetividade, a consistência e a rapidez na concessão que, caso sejam desenvolvidos apropriadamente, podem eliminar práticas discriminatórias nos empréstimos e tendem a ser simples e de fácil interpretação e implementação. As metodologias utilizadas para seu desenvolvimento e avaliação são bastante difundidas.

A ideia de distinção entre grupos de uma população foi introduzida por Fisher (1936), que desenvolveu em seu estudo a análise discriminante linear e a utilizou para classificar diferentes espécies de flores do género Íris, com base no comprimento e largura das sépalas e pétalas.

David Durand (1941) foi o primeiro a perceber a aplicabilidade da análise discriminante proposta por Fisher (1936) para diferenciar bons e maus empréstimos. Em seu estudo, realizado para o *National Bureau of Economic Research* dos EUA, Durand (1941) coletou 7.200 observações relativas a empréstimos realizados por 37 instituições, dentre elas bancos comerciais e financeiras de crédito, e utilizou o teste chi-quadrado para identificar as variáveis que melhor discriminavam os bons e os maus empréstimos. Por fim, utilizou a análise discriminante para desenvolver diversos modelos de *credit scoring*, nos quais observou bons resultados de predição para grande parte das empresas.

Myers e Forgy (1963) selecionaram aleatoriamente 600 contratos de financiamentos de uma companhia americana de trailers (*mobile homes*) e aplicaram a regressão logística e a análise discriminante pura e com variações para desenvolver modelos de escoragem. A grande novidade trazida no estudo de Myers e Forgy (1963) foi

a utilização de duas amostras para o desenvolvimento dos modelos: uma denominada amostra inicial, utilizada para desenvolvimento dos modelos, e a amostra *hold-out*, composta por observações que não participaram do desenvolvimento e utilizada para validação. Os autores relatam no artigo que não é possível afirmar sobre a eficácia dos modelos obtidos por Durand (1941), pois não há certeza se os mesmos foram validados em amostras *hold-out*, o que poderia acarretar em uma possível redução na eficácia dos modelos desenvolvidos.

No final dos anos 60, houve grande crescimento no volume de solicitações de cartão de crédito, exigindo dos bancos maior velocidade e automatização nas concessões, culminando na adoção de modelos de *credit scoring*. Esse fato fez com que os bancos percebessem a utilidade e as vantagens do uso de sistemas de escoragem, refletidas pela queda nas taxas de inadimplência do produto e pela possibilidade de contratação de mão de-obra com menor experiência em concessão de créditos. O sucesso observado com os cartões fez com que os bancos passassem a aplicar os modelos de *credit scoring* para a concessão de outros produtos a partir do final dos anos 80 (THOMAS, 2000).

Altman (1968) utilizou a análise discriminante múltipla para desenvolver um modelo de previsão de insolvência de empresas, denominado Z-Score. Sua amostra foi composta por 66 pequenas e médias empresas, das quais 33 se encontravam em insolvência entre os anos de 1946 e 1965. Inicialmente foram selecionados 22 indicadores contábeis para serem testados, sendo que permaneceram no modelo final apenas cinco dessas variáveis. Posteriormente, Altman et al. (1977) desenvolveram outro modelo para previsão de insolvência de empresas, denominado ZETA, obtido através de um refinamento do modelo Z-Score desenvolvido anteriormente.

Ohlson (1980) foi um dos primeiros estudos a utilizar a Regressão Logística para modelos de previsão de insolvência. Ohlson (1980) utilizou uma série coletadas do banco de dados Compustat, que incluía 105 empresas insolventes e 2058 empresas solventes de 1970 a 1976. Sua análise levou em consideração 7 indicadores financeiros e 2 variáveis binárias e o grau de acerto da classificação do seu modelo se mostrou inferior ao relatado em estudos anteriores baseados em análise discriminante múltipla, como, por exemplo, Altman (1968) e Altman et al. (1977).

O avanço computacional das décadas subsequentes contribuiu para o desenvolvimento de outras metodologias quantitativas e consequente aplicação no contexto de *credit scoring*, tais como as redes neurais, análise de sobrevivência e técnicas de aprendizagem de máquinas, como *support vector machine, bagging e boosting*. A

Tabela 2.2 apresenta algumas das principais técnicas quantitativas utilizadas ao longo dos anos para tal finalidade e respectivas referências de estudos:

Tabela 2.2 - Metodologias quantitativas e respectivas aplicações em *credit scoring*.

Metodologia	**Aplicações em *Credit Scoring***
Regressão Logística	Wiginton (1980), Bensic et al. (2005)
Analise Discriminante	Altman (1968), Altman (1994), Kumar e Bhattacharya (2006)
Arvores de Decisão	Bensic et al. (2005), Soltan e Mohammadi (2012)
Redes Neurais	Altman (1994), Desai et al. (1996), West (2000)
Cadeias de Markov	Hurd e Kuznetsov (2007), Frydman e Schuermann (2008)
Análise de Sobrevivência	Stepanova e Thomas (2002), Bellotti e Crook (2009)
Algoritmos Genéticos	Desai et al. (1997), Ong et al. (2005)
Support Vector Machines	Wang et al. (2005), Härdle et al. (2007)
Bagging	Breiman (1996), Optiz e Maclin (1999)
Boosting	Freund e Schapire (1997), Wang et al. (2011)

Fonte: elaborado pelos autores.

Os estudos de Baesens et al. (2003) e Lessmann et al. (2015) apresentam detalhada pesquisa sobre as técnicas aplicadas no desenvolvimento de modelos de *credit scoring* ao longo dos anos.

De acordo com Thomas (2010), os modelos de *credit scoring* utilizados na concessão de crédito podem ser classificados em dois tipos: *Application Scoring e Behavioural Scoring*, no entanto, embora os primeiros estudos relacionados aos modelos de escoragem tenham sidos desenvolvidos para a concessão de crédito e/ou previsão de inadimplência, metodologias com diferentes propósitos foram desenvolvidas ao longo dos anos e podem trazer ganhos significativos na gestão financeira das instituições. A seguir são apresentados alguns tipos de modelos de escoragem que podem ser utilizados em diferentes momentos do ciclo do crédito ou em áreas das instituições:

1. Modelos de *Application Scoring*: São utilizados para estimar a probabilidade de inadimplência de clientes solicitantes de crédito que ainda não possuem relacionamento creditício com a instituição. A variável resposta binária utilizada para desenvolvimento desses modelos é se o cliente foi bom ou mau pagador, geralmente classificado como mau o cliente que atingiu determinado número de dias em atraso na operação. Utilizam-se para tal previsão variáveis cadastrais, financeiras e de comportamento de credito no mercado. Trabalho Relacionado: Makuch (2001);

2. Modelos de *Attrition Scoring*: São utilizados para estimar a probabilidade de um cliente que contratou determinado produto cancelá-lo, podendo auxiliar a instituição na criação de um programa de retenção de clientes. A variável resposta binária utilizada para desenvolvimento desses modelos é se o cliente cancelou ou não determinado produto de crédito ou se o cliente deixou a instituição. Trabalho Relacionado: Xia e Jin (2008);

3. Modelos de *Behavioural Scoring*: Assim como os modelos de *Application Scoring,* são utilizados para estimar a probabilidade de inadimplência de clientes solicitantes de crédito, nesse caso para clientes que já possuam relacionamento creditício com a instituição. Acrescentam-se, dentre as variáveis preditoras, informações sobre o comportamento de crédito desses clientes nas operações já existentes na instituição, tornando esses modelos mais preditivos do que os modelos de *Application.* Esses modelos também são utilizados para reavaliar periodicamente os tomadores de crédito, obtendo informações atualizadas sobre a qualidade das carteiras de crédito. Trabalho relacionado ao tema: Hopper e Lewis (1992), Thomas (2000);

4. Modelos de *Collection Scoring*: São utilizados para estimar a probabilidade de clientes em atraso regularizarem o pagamento desses débitos em determinado período de tempo, com o propósito de ajustar a abordagem e a intensidade do processo de cobrança, maximizar a recuperação, reduzir custos, evitar desgastes desnecessários com o cliente e automatizar os fluxos. A variável resposta binária utilizada para desenvolvimento desses modelos é se o cliente pagou ou não determinado crédito em atraso. Trabalho relacionado ao tema: Souza (2000);

5. Modelos de *Fraud Scoring*: São utilizados para estimar a probabilidade de os clientes fraudarem a instituição no início do relacionamento creditício. Trabalho relacionado ao tema: Moraes (2012);

6. Modelos de *Profit Scoring*: São utilizados para estimar a probabilidade de os clientes serem rentáveis para a instituição financeira. Trabalho relacionado ao tema: Thomas (2000);

7. Modelos de *Propensity Scoring*: São utilizados para estimar a probabilidade de os clientes adquirirem determinados produtos com o objetivo de maximizar o retorno envolvido nas campanhas de *marketing*, em que os participantes selecionados para as campanhas são aqueles com maior probabilidade de contratação do produto. Trabalho relacionado ao tema: Tsai e Yeh (1999).

Os modelos de escoragem são desenvolvidos a partir de base de dados contendo a variável dependente a que se deseja modelar (inadimplência, recuperação do crédito, contratação do produto, fraude, etc.) e informações históricas dos clientes referentes à características do tomador e da operação de crédito contratada (dados cadastrais, demonstrações financeiras, tipo de produto, valor contratado, etc.) (SAUNDERS, 2000).

2.2.2. Modelos Estocásticos de Risco de Crédito

Os modelos estocásticos avaliam o comportamento estocástico do risco de crédito ou das variáveis que o determinam, como, por exemplo, valor de uma empresa, com a finalidade de precificar títulos e derivativos de crédito (ANDRADE, 2005).

De acordo com Duffee e Singleton (1999), esses modelos são divididos em duas categorias: modelos estruturais e modelos de forma reduzida.

Os modelos estruturais surgiram a partir do trabalho de Merton (1974) e relacionam o valor da firma com o processo de *default*. Já modelos de forma reduzida avaliam intensidade de ocorrência de eventos de *default*, independente dos fatores que os provocam. Detalhes sobre esses modelos podem ser encontrados em Bielecki e Rutkowski (2002) e Andrade e Thomas (2007).

2.2.3. Modelos de Risco de Portfólio

Os modelos de risco de portfólio visam a estimar a distribuição estatística das perdas (percentual ou em valor monetário) de uma carteira de crédito. Esses modelos foram desenvolvidos com base em conceitos utilizados para mensuração do risco de mercado e permitem que o risco de crédito seja avaliado de forma agregada, podendo ser utilizados para determinação do *Value at Risk* (VaR) e para cálculo do capital econômico a ser alocado pela instituição.

Segundo Saunders (2000) os principais modelos de risco de portfólio são:

1. *CreditMetrics*: Foi desenvolvido pelo banco J.P. Morgan e é baseado na abordagem de migração da qualidade do credito concedido (GUPTON et al., 1997);

2. *CreditRisk+:* Desenvolvido pela *Credit Suisse Financial Products* (CSFP, 1997), baseado na abordagem atuarial e procura estabelecer medidas de perda esperada com base no perfil de sua carteira e histórico de inadimplência;

3. *CreditPortfolioView*: Desenvolvido pela consultoria McKinsey, baseado no impacto das variáveis macroeconômicas sobre a inadimplência (WILSON, 1997);

4. KMV: Desenvolvido pela consultoria KMV Corporation, baseado na abordagem estrutural e considera o processo de falência endógeno e relacionado à estrutura de capital da firma (KMV, 1993).

Análises comparativas dos modelos de risco de portfólio podem ser encontradas em Gordy (1998) e Crouhy et al. (2000).

3. METODOLOGIA

De acordo com Sicsú (2010), o desenvolvimento de um modelo de *credit scoring* compreende as seguintes etapas:

1. Planejamento e definições;
2. Identificação de variáveis potenciais;
3. Planejamento amostral;
4. Aplicação da metodologia estatística para determinação do escore;
5. Validação e verificação de performance do modelo estatístico;
6. Determinação do ponto de corte ou faixas de escore;
7. Determinação de regra de decisão.

Os capítulos 3 e 4 dessa dissertação discorrem sobre as etapas 1 a 5 supracitadas. As etapas 6 e 7, que se referem às Políticas de Crédito da instituição, não serão abordadas nessa dissertação, onde detalhes podem ser encontrados em Schrickel (1995) e Silva (1998).

O fluxograma contido na Figura 3.1 detalha todas as etapas realizadas no processo de desenvolvimento dos modelos dessa dissertação.

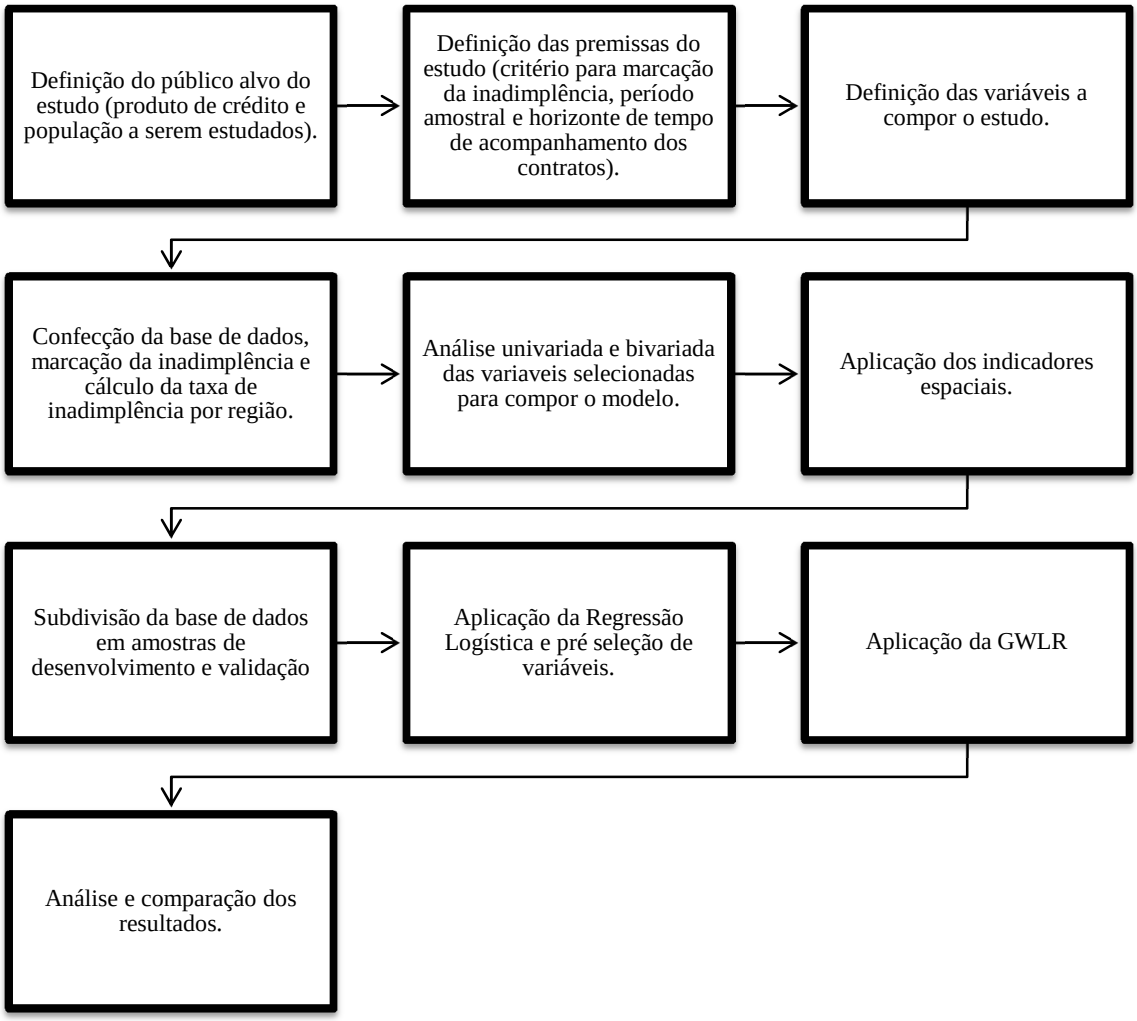

Figura 3.1 – Fluxograma das etapas de desenvolvimento dos modelos.

3.1. Base de Dados

Os dados utilizados nessa dissertação referem-se às operações de Crédito Direto ao Consumidor (CDC) concedidas por uma instituição financeira pública nacional a clientes domiciliados no Distrito Federal.

Essa operação de crédito possui as seguintes características:

1. Tomadores do crédito: Clientes titulares de conta corrente e/ou poupança;
2. Sem destinação específica;
3. Prazo de Concessão: de 01 a 36 meses;
4. Limites da Operação: Valor mínimo de R$150,00 e máximo de R$30.000,00, definido conforme a capacidade de pagamento do cliente;
5. Encargos: Taxa de juros pré-fixada, IOF e juros de acerto (se for o caso);

6. Contratação: Pode ser realizada em terminais de autoatendimento e *Internet Banking*;
7. Forma de Pagamento: Em prestações mensais que vencem conforme o dia escolhido pelo tomador e são debitadas automaticamente em conta.

A decisão de utilizar dados referentes a essa operação de crédito foi tomada com base em seu grande volume concessões mensais (em torno de 85 mil contratos novos em todo o Brasil durante o ano de 2014), por ser uma operação de crédito parcelada e por não possuir garantia real atrelada à operação (tais como imóveis, automóveis, etc.).

A decisão de utilizar os tomadores domiciliados no Distrito Federal (DF) como público alvo dessa dissertação foi tomada a partir de informações contidas no documento do Instituto de Pesquisa Econômica Aplicada (IPEA, 2011), o qual relata que grande parte dos indicadores sociais do Distrito Federal (DF) está melhor do que a média brasileira, como, por exemplo, a renda domiciliar (a maior no país) e o número de anos de estudo da população residente. No entanto, outros indicadores, especialmente os dados sobre violência entre jovens, desemprego e ritmo de redução da extrema pobreza, destoam e são influenciados pelos níveis de desigualdade de renda: a mais alta (segundo Índice de Gini) entre os estados brasileiros. Outro fator que favoreceu a escolha do Distrito Federal foi o fato de sediar a capital do Brasil e onde está situada a Universidade de Brasília.

A divisão territorial do DF utilizada nessa dissertação foi composta por 19 regiões e está disposta na Figura 3.2.

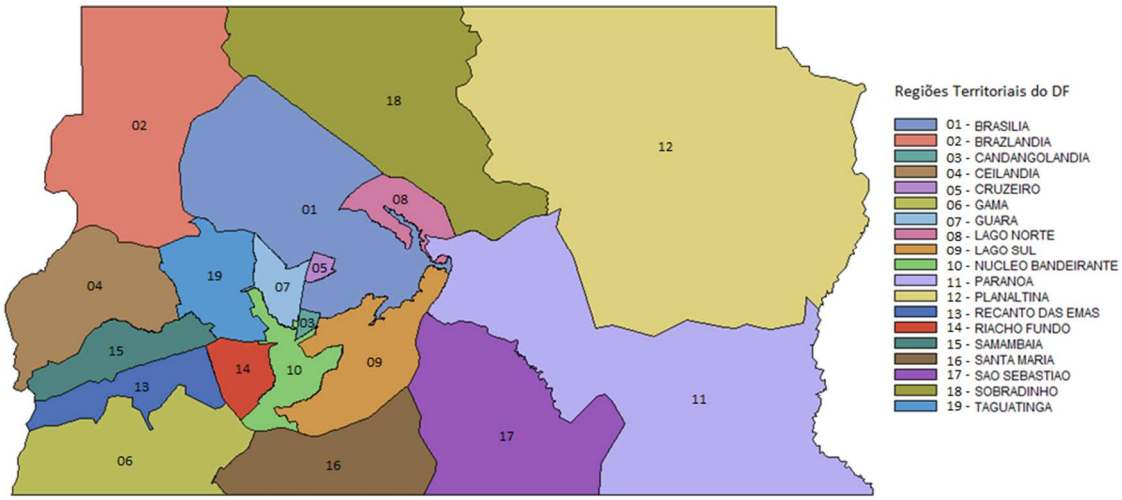

Figura 3.2 – Divisão territorial do Distrito Federal utilizada nesse estudo.
Fonte: elaborado pelo autor.

Após a definição do produto de crédito e do público alvo, foi definido como período amostral todos os contratos concedidos entre os meses de dezembro de 2013 a setembro de 2014, totalizando 10 safras de contratação e um total de 22.132 contratos distintos.

Foi acompanhado o desempenho de pagamento desses contratos nos doze meses subsequentes à data de contratação, os quais foram marcados como inadimplentes (maus), representados pelo número 1 na variável resposta Y, aqueles que ultrapassaram 90 dias em atraso em qualquer período desses doze meses, alinhado com a definição de descumprimento contida no art. 15 da Circular BACEN nº 3.648/13. Por possuir o desempenho de atraso dos contratos em diferentes momentos de tempo, essa base de dados é classificada como do tipo painel (*panel data*).

De acordo com Anderson (2007), as variáveis utilizadas pela literatura no desenvolvimento de modelos de *credit scoring* podem ser subdivididas em três grupos: variáveis socioeconômicas do tomador (idade, renda, escolaridade, endereço residencial, etc.), dados internos da instituição (histórico de empréstimos anteriores, produtos contratados, saldo em aplicação financeira) e dados externos à instituição (o cliente possui relacionamento com outra instituição?, dívida total do tomador no mercado). Dessa forma, buscou-se selecionar variáveis relacionadas a esses três grupos para compor os modelos desenvolvidos nessa dissertação.

Após a seleção inicial de variáveis, algumas foram retiradas do estudo por questão de sigilo, uma vez que fazem parte do atual modelo de *credit scoring* aplicado na instituição financeira. Assim, a seleção final foi composta pelas seguintes variáveis:

1. Idade do Tomador de Crédito: A idade do tomador de crédito é uma das variáveis mais comuns em modelos de *credit scoring* e pode refletir informações não mensuradas diretamente. Essa variável é colhida mediante registro da data de nascimento contida em documento original com foto no ato da solicitação do empréstimo. Espera-se que quanto maior a idade do tomador menor seja seu risco de crédito, pois o tomador mais velho teoricamente possui maior maturidade, responsabilidade, estabilidade e educação financeira, implicando em menor possibilidade de não honrar os compromissos firmados. Espacialmente, o peso dessa variável pode variar, uma vez que podem existir regiões homogêneas quanto à idade dos tomadores (bairros habitados em sua maioria por idosos ou jovens) ou

ainda regiões em que essa variável não discrimine o risco de crédito. Trabalhos que utilizam essa variável no desenvolvimento de modelo de *credit scoring*: Desai et al. (1996) e Van Gool et al. (2012);

2. Renda Formal do Tomador de Crédito: A renda formal influencia diretamente a capacidade de pagamento do tomador de crédito e, consequentemente, é importante para a avaliação da inadimplência. Alguns tomadores possuem somente renda informal, nesses casos essa variável é preenchida com valor zero. Essa variável é colhida mediante comprovante de renda formal no ato da solicitação do empréstimo (holerite ou declaração de imposto de renda). Para diminuir o efeito de queda do valor monetário ao longo do tempo, essa variável fora transformada em salários mínimos (SM) através da divisão pelo valor do SM brasileiro na data de contratação (R$ 678,00 para o mês de dezembro de 2013 e R$ 724,00 para os demais meses). A renda formal reflete a estabilidade financeira do tomador, dado a existência de um contrato de trabalho formal ativo. Espera-se que quanto maior seja a renda do tomador menor seja seu risco de crédito pois, teoricamente, os tomadores com maior renda possuem menor dificuldade ou "aperto" financeiro, implicando em menor possibilidade de não honrar os compromissos firmados. Espacialmente, o peso dessa variável pode variar, uma vez que podem existir regiões com maior desigualdade de renda, onde essa variável pode ser significativa ou regiões com maior concentração e homogeneidade de renda (como é o caso de Brasília), fazendo com que essa variável não discrimine o risco de crédito. Trabalhos que utilizam essa variável no desenvolvimento de modelo de *credit scoring*: Desai et al. (1996) e Harris (2015);
3. Grau de Instrução do Tomador de Crédito: O grau de instrução (escolaridade) mensura o nível educacional do tomador de crédito e, assim como a idade, pode refletir informações não mensuradas. Essa variável é colhida mediante entrevista no ato da solicitação do empréstimo e não necessita de documento comprobatório. Espera-se que, quanto maior a escolaridade do tomador, menor seja seu risco de crédito, pois o tomador com mais anos de estudo teoricamente possui maior clareza, responsabilidade, estabilidade e educação financeira, implicando em menor possibilidade de não honrar os compromissos firmados. Espacialmente, o peso dessa variável também pode variar, uma vez que podem existir regiões com ausência de universidades, implicando em menor e mais homogêneo grau de

instrução da população sendo que em regiões com presença de universidade há uma maior possibilidade de a população ser mais instruída;

4. Tempo de Relacionamento do Tomador de Crédito com a Instituição: Clientes com relacionamento prévio na instituição possuem ou já possuíram produtos financeiros anteriores. Nos casos em que esse produto foi uma operação de crédito, a instituição possui informações sobre o comportamento de pagamento desse tomador e, caso o mesmo não possua um bom histórico creditício, uma nova concessão de crédito geralmente é negada. Essa variável é calculada através da diferença entre a data de contratação do primeiro produto do tomador na instituição e a data de solicitação do novo empréstimo, onde clientes novos possuem valor zero para essa variável. Os clientes mais antigos tendem a prezar por sua reputação perante a instituição e por esse motivo apresentam menor risco de crédito se comparados aos clientes com pouco ou nenhum tempo de relacionamento. O peso dessa variável pode variar de região para região, uma vez que regiões com maior quantidade de agências bancárias tendem a possuir uma população mais heterogênea com relação ao tempo de relacionamento com a instituição frente a uma região rural ou que não possui agências bancárias, onde essa variável pode se mostrar não significativa. Trabalho que utiliza essa variável no desenvolvimento de modelo de *credit scoring*: Khandani et al. (2010);
5. Prazo contratado da operação: As operações contratadas com prazos mais longos estão mais expostas à ocorrência de mudanças inesperadas na vida do tomador, tais como a morte ou perda do emprego. Outro fato comum é os tomadores mais endividados e/ou com mais dificuldades financeiras tomarem o máximo de empréstimo disponível para ele, seja em valores ou prazo, culminando na maior incidência da inadimplência. Dessa forma espera-se um maior risco de crédito para as operações com maiores prazos de vencimento. A variação espacial dessa variável pode ocorrer caso existam regiões com uma população mais endividada, o que acarreta na contratação de operações com prazos mais elevados. Trabalhos que utilizam essa variável no desenvolvimento de modelo de *credit scoring*: Van Gool et al. (2012) e Harris (2015);

6. Taxa SELIC: A Taxa SELIC é uma variável macroeconômica que influencia diretamente a concessão de crédito. Por ser a taxa básica de juros da economia brasileira, seu aumento impacta diretamente nas taxas de juros das operações de crédito, deixando-as mais caras para os tomadores e aumentando o risco de crédito da operação. No momento da contratação do empréstimo o impacto dessa variável já estará embutido na taxa de juros, e, por ser uma operação prefixada, espera-se pouca variação espacial em seus coeficientes. Essa variável está disponível no Sistema Gerenciador de Séries (SGS[5]) do BACEN sob o código 1178;
7. Taxa de Desemprego: A taxa de desemprego também é uma variável macroeconômica muito importante para a inadimplência bancária, pois um aumento dessa taxa significa que mais pessoas estão desempregadas e, consequentemente, sem renda formal, o que também pode acarretar em aumento da inadimplência frente a queda do poder financeiro do tomador. A variação espacial dessa variável dependerá da quantidade de trabalhadores empregados ou setores da economia presentes nas regiões de estudo. Como exemplo, espera-se que essa variável não seja significativa para a região de Brasília, uma vez que sua grande maioria é composta de servidores públicos e que não são afetados pelo desemprego. Essa variável está disponível no Sistema Gerenciador de Séries (SGS) do BACEN sob o código 10777;
8. Inflação (IPCA) acumulado nos últimos 12 meses: O Índice de Preços ao Consumidor Amplo (IPCA) é um índice que tem o objetivo de medir a inflação de um conjunto de produtos e serviços comercializados no varejo, referentes ao consumo pessoal das famílias. Dessa forma, esse indicador reflete o poder de compra da população, sendo que quanto maior o índice menor é o poder de compra. Valores elevados do índice tendem a aumentar os índices de inadimplência, uma vez que o poder de compra dos tomadores de crédito diminui e o pagamento da parcela do empréstimo não seria prioritária frente às demais despesas da família como alimentação, saúde e educação. Por ser uma variável macroeconômica, a variação espacial dessa variável dependerá da renda da população da região, uma vez que a inflação afeta mais populações de menor renda. Novamente citando a região de Brasília como exemplo, espera-se que essa

[5] O SGS possui series históricas de dados referentes a diversos temas de finanças, disponível em < https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries >, acesso em 23/03/2016.

variável possua menor peso nessa região devido à alta renda da população. Essa variável está disponível no Sistema Gerenciador de Séries (SGS) do BACEN sob o formato de variação mensal com o código 433, sendo necessário calcular a taxa acumulada dos últimos 12 meses.

Cabe ressaltar que todas as variáveis selecionadas são referentes ao momento da contratação do crédito (um único ponto no tempo), caracterizando-se como dados do tipo *cross-section.*

Por fim, as coordenadas geográficas latitude e longitude referentes às regiões utilizadas nesse estudo e necessárias para aplicação da técnica GWLR foram obtidas no site do IBGE, sendo importante ressaltar que essas coordenadas são as mesmas para todos os tomadores de crédito residentes na mesma região, onde foram utilizadas as coordenadas referentes ao ponto central de cada região.

Dessa forma, a base de dados final dessa dissertação foi composta pelas seguintes variáveis:

Tabela 3.1 – Composição da base de dados final do estudo.

Variável	Descrição	Tipo	Característica
id_ctr	ID do contrato	Identificadora	Tomador
dt_contratacao	Data de contratação da operação de crédito	Identificadora	Tomador
codigo	Código da região tomador de crédito	Identificadora	Tomador
latitude	Valor da latitude do centro da região do tomador	Identificadora	Tomador
longitude	Valor da longitude do centro da região do tomador	Identificadora	Tomador
Y	Inadimplente (atraso > 90 dias)	Resposta	Tomador
idade	Idade do tomador de crédito	Preditora	Tomador
renda	Renda formal comprovada do tomador (em salários mínimos)	Preditora	Tomador
instrução	Grau de instrução do tomador de crédito	Preditora	Tomador
tempo_rel	Tempo de relacionamento em meses do tomador com a instituição (em meses)	Preditora	Tomador
prazo	Prazo contratado da operação de crédito (em meses)	Preditora	Operação
SELIC	Taxa SELIC anualizada no mês de contratação	Preditora	Macroeconômica
desemprego	Taxa de desemprego no mês de contratação	Preditora	Macroeconômica
inflação	Taxa de inflação (IPCA) acumulada nos últimos 12 meses	Preditora	Macroeconômica

Fonte: elaborado pelo autor.

Para o desenvolvimento dos modelos de regressão, a base de dados foi subdividida em duas amostras: uma para desenvolvimento e outra para validação do modelo. Essa subdivisão foi realizada através da data de contratação da operação, sendo a amostra de desenvolvimento composta pelas 5 safras iniciais de contratação (dezembro de 2013 a abril de 2014), totalizando 10.944 registros e a base de validação composta pelas 5 safras finais (maio a setembro de 2014) que totalizam 11.188 registros. A divisão da população

em amostras de desenvolvimento e validação é muito importante, pois verifica a assertividade do modelo em uma população que não participa do desenvolvimento do mesmo (BARTH, 2004; SICSÚ, 2010). A realização da subdivisão das amostras por meio da data de contratação das operações teve o intuito de simular a aplicação real dos modelos a uma população futura.

3.2. Indicadores Espaciais

O I de Moran (MORAN, 1950) é um dos indicadores globais mais utilizados para verificar a existência de correlação espacial. Os indicadores globais apresentam uma única medida de tendência espacial para toda a região em estudo, permitem testar a hipótese de existência de dependência espacial entre as regiões de acordo com a variável de interesse e são utilizados na análise exploratória dos dados. Sua formula é dada por:

$$I = \frac{n}{\sum_i \sum_{j \neq i} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (3.1)$$

onde n é o número de regiões em estudo, x_i e x_j são os valores da variável de interesse nas regiões i e j e w_{ij} são os elementos da matriz de proximidade espacial, que pode ser calculada de diferentes maneiras, como, por exemplo, através da presença ou ausência de fronteira entre as regiões ou pela distância euclidiana entre elas. O índice de Moran está restrito ao intervalo [-1,1], no qual valores próximos a -1 indicam correlação espacial negativa, valores próximos a 1 indicam correlação espacial positiva e valor igual a 0 indica ausência de correlação espacial ou independência espacial com relação à variável testada.

Enquanto os indicadores globais pressupõem que todas as regiões em estudo podem ser representadas por um único valor, os indicadores locais (do inglês *Local Indicator of Spatial Association* - LISA) desenvolvidos por Anselin (1995) são utilizados para verificar a existência de correlação espacial dentro das unidades geográficas em estudo e buscam as diferenças (peculiaridades) regionais. A presença de áreas com índices locais significativos é um indicio de heterogeneidade (não estacionariedade) espacial.

A fórmula do índice local de Moran é dada por:

$$I_i = \frac{n(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (3.2)$$

A base de dados utilizada na aplicação dos Índices de Moran Global e Local foi a base total de registros (sem subdivisão de amostras), e a variável testada foi a taxa de inadimplência regional, calculada através da seguinte fórmula:

$$Taxa\ de\ Inadimplência\ da\ Região = \frac{Quantidade\ de\ clientes\ Inadimplentes\ na\ região}{Quantidade\ total\ de\ clientes\ da\ região} \quad (3.3)$$

Nesse estudo o índice global de Moran foi utilizado para verificar a existência de correlação espacial da taxa de inadimplência entre as regiões do DF. O índice local de Moran foi utilizado para verificar a existência de regiões distintas quanto à taxa de inadimplência em relação às demais regiões. A existência de regiões significativas (o nível de confiança utilizado para o índice local de Moran foi de 95%) pode indicar que os modelos de regressão desenvolvidos para essas regiões sejam distintos em relação aos modelos das demais regiões do estudo, o que pode justificar a aplicação da GWLR para essa população.

3.3. Regressão Logística

A regressão logística é um caso particular dos Modelos Lineares Generalizados (MLG). Também conhecida como análise *logit*, é uma técnica que estima a probabilidade de ocorrência de determinado evento de variável aleatória binária (variável dependente) a partir de um conjunto de variáveis explicativas (HAIR et al., 2009).

A regressão logística é o método mais utilizado para se obter uma regra de classificação quando a variável preditiva que se deseja analisar é binária. Lessmann et al. (2015) realizaram uma abrangente pesquisa sobre as metodologias de classificação utilizadas para o desenvolvimento de modelos de *credit scoring*, elencando e verificando a acurácia de quarenta e uma (41) metodologias distintas e apontam a regressão logística como a metodologia padrão do setor financeiro.

Suponha que uma variável aleatória binária Y_i segue uma distribuição de Bernoulli e assume os seguintes valores:

$$Y_i = \begin{cases} 1 & \text{se o cliente é inadimplente} \\ 0 & \text{se o cliente é adimplente} \end{cases}$$

Seja $\boldsymbol{x}_i = \left(1, x_1, x_2, \dots, x_{p,}\right)^T$ o vetor de características do cliente *i* e $\pi(\boldsymbol{x}_i)$ a proporção de clientes inadimplentes em função do perfil dos clientes, a distribuição de probabilidades e esperança de Y_i são dadas por:

$$P\left(Y_i = y_i | \, \boldsymbol{x}_i\right) = \pi(\boldsymbol{x}_i)^{y_i} \left(1 - \pi(\boldsymbol{x}_i)\right)^{1-y_i};\ y_i = 0, 1. \tag{3.4}$$

$$\mathbb{E}\left(Y_i \mid \boldsymbol{x}_i\right) = \pi(\boldsymbol{x}_i) \tag{3.5}$$

Dado que a distribuição de Bernoulli pertence à família exponencial temos:

$$g\left(\mathbb{E}(Y_i | \, \boldsymbol{x}_i)\right) = g\left(\pi(\boldsymbol{x}_i)\right) = \ln\left(\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} = \boldsymbol{x}_i{}^T \boldsymbol{\beta} \tag{3.6}$$

Podendo também ser escrito da forma:

$$\pi(\boldsymbol{x}_i) = \frac{e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}}, \tag{3.7}$$

em que $\pi(\boldsymbol{x}_i)$ pode ser interpretado como a probabilidade do *i*-ésimo cliente se tornar inadimplente.

Na expressão (3.7), os valores de $x_{1i}, x_{2i}, \dots, x_{pi}$ são conhecidos e os parâmetros $\beta_0, \beta_1, \dots, \beta_p$ são as únicas quantias desconhecidas que necessitam ser estimadas. Os parâmetros representam a importância de cada variável explicativa para a ocorrência do evento (HAIR et al., 2009) e suas estimativas geralmente são calculadas através do método da máxima verossimilhança (HOSMER; LEMESHOW, 2000).

Sabendo que os dados são oriundos de uma distribuição Bernoulli e uma vez que as observações do conjunto de dados são independentes, a Função de Verossimilhança é dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i} \left(1 - \pi(\boldsymbol{x}_i)\right)^{1-y_i} \tag{3.8}$$

Pelo princípio do método da máxima verossimilhança, os valores estimados de $\boldsymbol{\beta}$ são aqueles que maximizam $L(\boldsymbol{\beta})$. Para obtenção desses valores, calcula-se a derivada dessa função em relação a cada um dos parâmetros e procura-se o ponto crítico no qual a derivada é igual a zero.

Aplicando a transformação monotônica logaritmo natural (ln) à função de verossimilhança, em virtude da propriedade de que o logaritmo de um produto é igual à soma dos logaritmos dos fatores, obtém-se:

$$\ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(\boldsymbol{x}_i)] + (1 - y_i) \ln[1 - \pi(\boldsymbol{x}_i)]\} \tag{3.9}$$

Essa transformação é realizada para simplificar matematicamente o cálculo das derivadas, tendo em vista que os resultados da maximização das funções $L(\boldsymbol{\beta})$ e $\ln[L(\boldsymbol{\beta})]$ são exatamente os mesmos (CASELLA e BERGER, 2010).

Dessa forma, diferenciando $\ln[L(\boldsymbol{\beta})]$ e igualando a zero, obtêm-se as expressões (3.10) e (3.11), conhecidas como equações de verossimilhança:

$$\sum_{i=1}^{n} [y_i - \pi(\boldsymbol{x}_i)] = 0 \tag{3.10}$$

$$\sum_{i=1}^{n} x_i [y_i - \pi(\boldsymbol{x}_i)] = 0 \tag{3.11}$$

Essas equações são não-lineares nos parâmetros e conseguem ser solucionadas via métodos numéricos interativos, como, por exemplo, o método Newton-Raphson.

Os estimadores possuem diversas características, as quais destacam-se:

1. Eficiência: O estimador mais eficiente é aquele de menor variância;
2. Consistência: Um estimador é dito consistente quando o mesmo converge, em probabilidade, para o seu valor populacional quando o tamanho da amostra *n* tende para infinito;
3. Viés: Um estimador não enviesado é aquele em que a esperança do estimador é o seu valor populacional, ou seja, $\mathbb{E}(\hat{\beta}) = \beta$.

A significância dos estimadores pode ser testada através do Teste da Razão de Verossimilhança, que tem o intuito de comparar os valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável que se deseja testar.

A comparação dos observados com os valores preditos pode ser realizada através da estatística *Deviance* (D) que se baseia na função de verossimilhança e é dada pela seguinte expressão:

$$D = -2ln\left[\frac{verossimilhança\ do\ modelo\ testado}{verossimilhança\ do\ modelo\ saturado}\right] \tag{3.12}$$

O teste utilizado nesse estudo para verificar a significância dos coeficientes (parâmetros) da regressão foi o Teste de Wald, que se baseia na distribuição Normal Padrão e possui as seguintes hipóteses a serem testadas:

$$\begin{cases} \mathrm{H}_0: \beta_j = 0 \\ \mathrm{H}_1: \beta_j \neq 0 \end{cases} \text{ para } j = 1,2,\dots,p$$

A estatística do teste é dada por:

$$Z_j = \frac{\widehat{\beta_j}}{\widehat{DP}\left(\widehat{\beta_j}\right)} \tag{3.13}$$

onde $\widehat{\beta_j}$ é o EMV de β_j e $\widehat{DP}\left(\widehat{\beta_j}\right)$ é o Desvio Padrão estimado de $\widehat{\beta_j}$.

Sob a hipótese nula (H_0), Z_j tem aproximadamente uma distribuição normal com média zero e variância um (normal padrão).

Os modelos de regressão logística podem ser aplicados por meio de diferentes tipos de método de seleção de variáveis, os mais difundidos são os métodos *Forward, Backward* e *Stepwise*. A presente dissertação utilizou o método *Stepwise* para pré-selecionar as variáveis a compor os modelos de regressão desenvolvidos via GWLR, utilizando como critério de permanência no modelo as variáveis com p-valores abaixo de 0,10. O método *stepwise* possui a vantagem de retirar variáveis já presentes no modelo que se tornam não significativas (de acordo com o ponto de corte definido) após inclusão de novas variáveis no modelo.

A aplicação da Regressão Logística para desenvolvimento de modelos de *credit scoring* pode ser encontrada nos estudos de Wiginton (1980) e Bensic et al. (2005).

3.4. Regressão Geograficamente Ponderada

A técnica de Regressão Geograficamente Ponderada, em inglês *Geographically Weighted Regression* (GWR) foi proposta por Brunsdon, Fotheringham e Charlton (1996) e é utilizada para modelar processos heterogêneos (não estacionários) espacialmente. Sua ideia básica é ajustar um modelo de regressão para cada ponto no conjunto de dados com base nas observações mais próximas geograficamente.

Dado um modelo de regressão linear básico, a expressão equivalente para a GWR é dada por:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{3.14}$$

Nota-se pela expressão acima que os parâmetros do modelo, representados pela função $\beta_k(u_i, v_i)$ variam de acordo com os valores de u_i, v_i, que representam as coordenadas geográficas latitude e longitude da observação (região) i, resultando em um modelo distinto para cada região do estudo. Os pressupostos do modelo clássico de regressão linear permanecem para a GWR.

A forma matricial da estimação dos parâmetros de um modelo de regressão geograficamente ponderada (GWR) é dada por:

$$\widehat{\boldsymbol{\beta}}(i) = (\boldsymbol{X'W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X'W}(u_i, v_i)\boldsymbol{y}, \tag{3.15}$$

onde

$$\boldsymbol{W}(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{bmatrix} \tag{3.16}$$

Note que $\boldsymbol{W}(u_i, v_i)$ é uma matriz diagonal e distinta para cada ponto i de coordenadas (u_i, v_i), contendo em sua diagonal principal os pesos w_{ij} obtidos por meio das funções de ponderação ou em inglês *kernel*. Note que a substituição de todos os pesos

w_{ij} pelo valor 1 equivale à matriz identidade, que substituída em (3.15) a faz retornar ao modelo clássico de regressão linear.

As duas principais funções de ponderação encontradas na literatura são as funções Gaussiana (Normal ou em inglês *Gaussian*) e a função Biquadrática *(*em inglês *Bisquare).* As fórmulas de ambas as funções estão contidas na tabela Tabela 3.3.

Tabela 3.3 – Funções de Ponderação ou *kernels*.

Funções de Ponderação	**Fórmula das Funções de Ponderação**
Gaussiana Fixa	$w_{ij} = \exp\left\{-\frac{1}{2}(d_{ij}/b)^2\right\}$
Biquadrática Fixa	$w_{ij} = [1-(d_{ij}/b)^2]^2$ se $d_{ij} < b$, e $w_{ij} = 0$ caso contrário
Gaussiana Variável	$w_{ij} = \exp\left\{-\frac{1}{2}(d_{ij}/b_{i(k)})^2\right\}$
Biquadrática Variável	$w_{ij} = [1-(d_{ij}/b_{i(k)})^2]^2$ se $d_{ij} < b_{i(k)}$, e $w_{ij} = 0$ caso contrário

Fonte: Fotheringham *et al.* (2002).

Nota-se pela Tabela 3.3 que existem dois tipos de expressões para cada uma das funções Gaussiana e Biquadrática, que se diferenciam entre si por meio da escolha do parâmetro *b (bandwidth)* a ser utilizado (se fixo ou variável). O parâmetro d_{ij} contido nas funções de ponderação representa a distância do ponto *i* ao ponto *j*, o parâmetro *b* é o *bandwidth* (parâmetro de suavização) fixo e o parâmetro $b_{i(k)}$ representa o *bandwidth* variável, sendo que a letra *k* representa o número de vizinhos mais próximos do ponto *i*.

O parâmetro *bandwidth* controla a variância da função de ponderação. Quando os dados são esparsos (espacialmente dispersos ou quando as áreas têm tamanhos diferentes), um mesmo *bandwidth* pode ser adequado para algumas localidades e inadequados para outras, pois, nesse último caso, os parâmetros estimados poderiam ter grandes erros padrões devido a poucos registros utilizados na estimação dos modelos (SILVA, 2009). Por esse motivo, em situações onde os dados não são igualmente distribuídos dentre as regiões (algumas regiões um número grande de registros enquanto outras possuem poucos registros) é recomendado a utilização do *bandwidth* variável. A Figura 3.2 ilustra o *bandwidth* em uma função de ponderação e as Figuras 3.3 e 3.4 exemplificam o uso do *bandwidth* fixo ou variável.

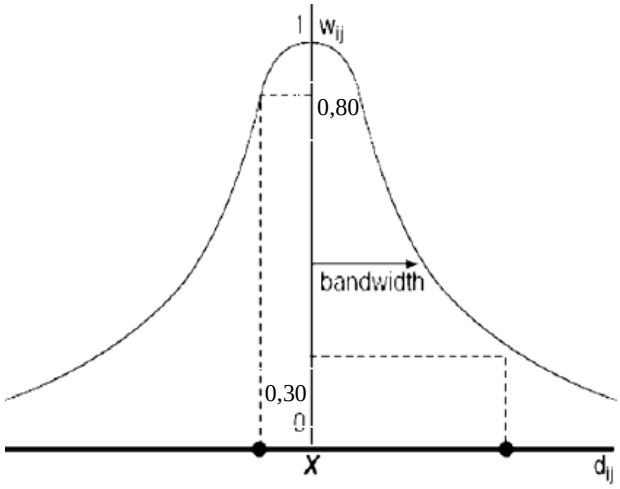

Figura 3.2: *Bandwidth* ou Parâmetro de Suavização.
Fonte: Fotheringham *et al.* (2006), com adaptações.

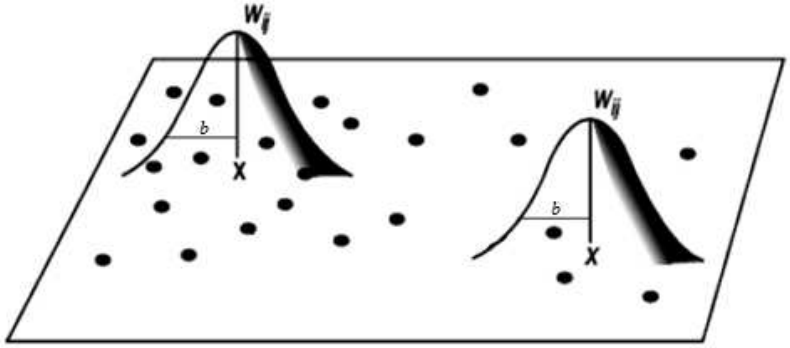

Figura 3.3: Funções de ponderação espacial com *Bandwidth* fixo.
Fonte: Fotheringham *et al.* (2006), com adaptações.

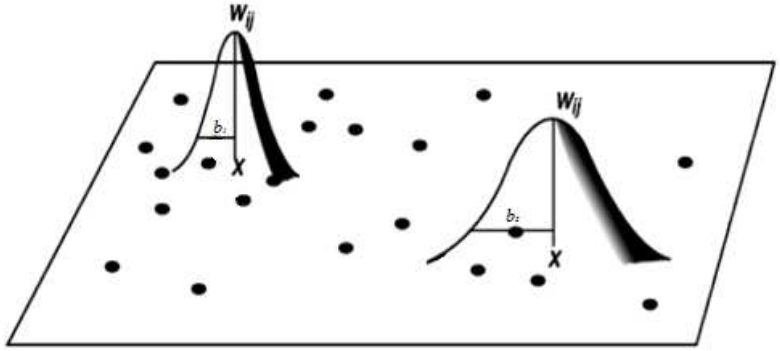

Figura 3.4: Funções de ponderação espacial com Bandwidth variável.
Fonte: Fotheringham et al. (2006), com adaptações.

No desenvolvimento de um modelo via GWR utilizando o *bandwidth* fixo, o mesmo deve ser especificado por seu valor em unidade de distância, no entanto, na utilização do *bandwidth* variável, deve-se definir um número *k* (fixo) de vizinhos mais próximos a serem utilizados nos modelos e, com base nessa quantidade *k*, o valor do *bandwidth* (que continua sendo expresso por um valor de distância) varia entre as regiões do estudo.

A tabela 3.4 simula a diferença entre os valores dos pesos w_{ij} calculados para diferentes *bandwidths*, d_{ij} e funções de ponderação.

Tabela 3.4 – Simulação de valores dos pesos w_{ij}.

Bandwidth (em km)	d_{ij} (em km)	Gaussiana	Biquadrática
b = 10	1	0,9950	0,9801
	5	0,8825	0,5625
	25	0,0439	-
	50	0,0000	-
	100	0,0000	-
b = 25	1	0,9992	0,9968
	5	0,9802	0,9216
	25	0,6065	-
	50	0,1353	-
	100	0,0003	-
b = 50	1	0,9998	0,9992
	5	0,9950	0,9216
	25	0,8825	0,5625
	50	0,6065	-
	100	0,1353	-
b = 100	1	0,9999	0,9998
	5	0,9988	0,9950
	25	0,9692	0,8789
	50	0,8825	0,5625
	100	0,6065	-

Fonte: elaborado pelo autor.

Note que, quanto maior é a diferença entre o *bandwidth* e a distância entre *i* e *j* (d_{ij}), menor é a diferença dos pesos w_{ij} calculados através das duas funções de ponderação. A medida que essa diferença diminui, os pesos w_{ij} atribuídos através da função Gaussiana são maiores do que os pesos atribuídos pela função Biquadrática. Note também que a função Gaussiana continua a atribuir pesos aos pontos com distância superior ao *bandwidth*, isso ocorre devido à característica assintótica da curva normal em relação ao eixo das abscissas e fazem com que os pesos nunca cheguem ao valor zero, enquanto a função Biquadrática atribui peso zero a pontos com distância igual ou superior ao *bandwidth*, isto é, são utilizados no desenvolvimento do modelo somente os pontos com distâncias d_{ij} inferiores ao *bandwidth.*

3.5. Regressão Logística Geograficamente Ponderada

Quando a variável resposta de interesse é binária, a aplicação da GWR deve ser realizada por meio da Regressão Logística Geograficamente Ponderada ou *Geographically Weighted Logistic Regression* (GWLR), cuja fórmula para obtenção da probabilidade de ocorrência do evento de interesse é dada a partir da substituição dos parâmetros de (3.6) pelos parâmetros dispostos em (3.14), representada pela seguinte fórmula:

$$\ln\left(\frac{\pi(\boldsymbol{x}_j)}{1-\pi(\boldsymbol{x}_j)}\right) = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{jk} + \varepsilon_i \tag{3.17}$$

Ou ainda, na forma demonstrada em (3.7):

$$\pi(\boldsymbol{x}_j) = \frac{e^{\beta_0(u_i,v_i)+\sum_{k=1}^{p}\beta_k(u_i,v_i)x_{jk}}}{1+e^{\beta_0(u_i,v_i)+\sum_{k=1}^{p}\beta_k(u_i,v_i)x_{jk}}}, \tag{3.18}$$

onde $\pi(\boldsymbol{x}_j)$ é a probabilidade do *j*-ésimo cliente se tornar inadimplente, a função $\beta_k(u_i, v_i)$ representa os parâmetros (coeficientes) das *k* variáveis do modelo, que variam de acordo com a região i de coordenadas latitude e longitude (u_i, v_i).

Assim como a regressão logística, a estimação dos parâmetros da GWLR também é realizada via método da máxima verossimilhança, sendo a função de verossimilhança da GWLR representada pela seguinte expressão:

$$L(\boldsymbol{\beta}(u_i, v_i)) = \left\{\prod_{j=1}^{n}\left[1+\exp\left(\sum_{k=0}^{p}\beta_k(u_i, v_i)x_{jk}\right)\right]^{-1}\right\}\exp\left[\sum_{k=0}^{p}\left(\sum_{j=1}^{n}y_j x_{jk}\right)\beta_k(u_i, v_i)\right] \tag{3.19}$$

Aplicando a transformação logaritmo natural (ln) e desenvolvendo a fórmula obtém-se:

$$\ln[L(\boldsymbol{\beta}(u_i, v_i))] = \sum_{k=0}^{p}\left(\sum_{j=1}^{n}y_j x_{jk}\right)\beta_k(u_i, v_i) - \sum_{i=1}^{n}\ln\left\{1+\exp\left(\sum_{k=0}^{p}\beta_k(u_i, v_i)x_{jk}\right)\right\} \tag{3.20}$$

A matriz $\boldsymbol{W}(u_i, v_i)$ descrita em (3.16) possui em seus elementos os pesos w_{ij} (calculados através das funções de ponderação expostas na Tabela 3.3) e é utilizada para ponderar geograficamente as observações na estimação de cada conjunto de

parâmetros $\beta_k(u_i, v_i)$, ou seja, essa matriz é responsável por dar um peso maior para as observações mais próximas geograficamente da região *i* na estimação dos seus parâmetros e dar um peso menor ou zero (a depender da função de ponderação escolhida) para as observações mais distantes da região *i* em questão na estimação dos seus parâmetros $\beta_k(u_i, v_i)$. A matriz $\boldsymbol{W}(u_i, v_i)$ também varia de acordo com a localidade de cada tomador de crédito e compõe a função de verossimilhança da seguinte maneira:

$$\begin{aligned} \ln[L^*(\boldsymbol{\beta}(u_i, v_i))] = &\sum_{k=0}^{p}\left(\sum_{j=1}^{n} w_j(u_i, v_i)\, y_j x_{jk}\right)\beta_k(u_i, v_i) \\ &- \sum_{j=1}^{n} w_j(u_i, v_i) \ln\left\{1 + \exp\left(\sum_{k=0}^{p} \beta_k(u_i, v_i) x_{jk}\right)\right\} \end{aligned} \tag{3.21}$$

Similar ao modelo de regressão logística, após diferenciar (3.21) em função de $\boldsymbol{\beta}(u_i, v_i)$ e igualar a zero, os parâmetros do modelo são estimados utilizando-se métodos numéricos iterativos, como, por exemplo, o método dos mínimos quadrados reponderados iterativos (MQRI). Cabe ressaltar que esse procedimento de maximização é realizado para cada uma das funções referentes a cada região *i* do estudo.

Conforme já dito na introdução dessa dissertação, o único trabalho que faz referência ao uso da técnica GWLR para desenvolvimento de modelo de *credit scoring* foi o de Travassos et al. (2013), no entanto o artigo não apresenta os resultados encontrados.

Não foram encontrados outros estudos nacionais ou internacionais que utilizaram a GWLR no desenvolvimento de modelos de *credit scoring*, com buscas realizadas no portal de periódicos da CAPES e no *Google Scholar* através das expressões RLGP risco de crédito, RLGP *credit scoring*, GWLR *credit scoring*, e GWLR *credit risk.*

Como o objetivo principal desse estudo foi verificar a aplicabilidade da GWLR no desenvolvimento de modelos de *credit scoring*, inicialmente foram desenvolvidos quatro modelos, sendo um para cada combinação das duas funções de ponderação (Gaussiana e Biquadrática) com os dois tipos de *bandwidth* (fixo ou variável). O melhor modelo estimado via GWLR, bem como o valor do *bandwidth* ótimo para compor esses modelos, foram definidos por meio do critério informacional AIC corrigido (AICc).

3.6. Comparação Entre os Modelos

As métricas utilizadas para comparação entre os modelos desenvolvidos por meio das metodologias GWLR e Regressão Logística foram: o critério informacional AICc, a acurácia dos modelos, o percentual de falsos positivos, a soma do valor da dívida dos falsos positivos e o valor monetário esperado de inadimplência da carteira frente ao valor monetário de inadimplência observado.

O critério informacional AIC corrigido (AICc) foi desenvolvido para a GWR por Hurvich et al. (1998) e foi o critério utilizado para comparação entre todos os modelos desenvolvidos nessa dissertação e também para definir o melhor *bandwidth* e os *k* vizinhos mais próximos a serem utilizados. Sua fórmula é dada por:

$$AIC_c = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + \frac{n(n+tr(\mathbf{R}))}{n-2-tr(\mathbf{R})}\ , \qquad (3.22)$$

onde a $\hat{\sigma}$ é a estimativa de máxima verossimilhança de σ e a matriz **R** é obtida através das matrizes de $\hat{\mu}$ e *y*. Detalhes sobre o cálculo dessa matriz pode ser encontrado em Fotheringhan et al. (2002) e Hurvich et al. (1998).

A acurácia dos modelos e o percentual de falsos positivos foram obtidos através da matriz de confusão, dada por:

Tabela 3.5 – Matriz de Confusão.

		Valor Observado	
		0	1
Valor Predito	0	VP	FP
	1	FN	VN

Fonte: Crook et al. (2007), com adaptações.

onde:

i. VP: Verdadeiro Positivo - quantidade de clientes bons classificados como bons;
ii. VN: Verdadeiro Negativo - quantidade de clientes maus classificados como maus;
iii. FP: Falso Positivo - quantidade de clientes maus classificados como bons;
iv. FN: Falso Negativo - quantidade de clientes bons classificados como maus.

De acordo com a Tabela 3.5, existem dois tipos de erro que um modelo classificador pode cometer: reprovar clientes bons (Falso Negativo - FN) ou aprovar clientes maus (Falso Positivo - FP), sendo que esse último, também conhecido como Erro do tipo II, é considerado o pior dos dois erros pois esse cliente seria aprovado e poderia gerar prejuízos financeiros para a instituição.

Também foi mensurada, para ambos os modelos, a somatória do saldo devedor de todos os tomadores classificados como FP, com o intuito de calcular o valor monetário que entraria em inadimplência devido ao erro de classificação do modelo e que poderia acarretar em prejuízo financeiro para a instituição.

A partir da matriz de confusão também é possível calcular a Acurácia ou a proporção de acertos do modelo, obtida pela proporção de Verdadeiro Positivo e Verdadeiro Negativo em relação ao total, conforme a seguinte fórmula:

$$\frac{VP + VN}{VP + VN + FP + FN} \tag{3.23}$$

Outras medidas calculadas a partir da matriz de confusão são as medidas Sensibilidade (S) e Especificidade (E), utilizadas nessa dissertação para definição do ponto de corte e dadas por:

$$S = \frac{VP}{VP + FN} \tag{3.24}$$

$$E = \frac{VN}{VN + FP} \tag{3.25}$$

O valor do ponto de corte foi definido pelo escore que minimiza a distância entre a Sensibilidade e Especificidade.

O valor monetário esperado de inadimplência da carteira, calculado por meio da esperança das distribuições discretas, também foi utilizado para comparação entre os dois modelos, cuja fórmula é dada por:

$$E(X) = \sum_{i=1}^{n} x_i * P(Y_i = 1), \quad (3.24)$$

onde *n* é a quantidade total de tomadores da carteira, x_i é o saldo devedor da operação de crédito do tomador *i* e $P(Y_i = 1)$ é a probabilidade do tomador *i* se tornar inadimplente, resultante dos modelos de *credit scoring*. Esse valor foi confrontado com o valor da somatória das dívidas dos clientes inadimplentes, com o intuito de verificar qual modelo mais se aproxima do valor real de inadimplência.

O melhor modelo estimado via GWLR também foi utilizado para comparar os modelos locais (os modelos gerados para cada região do DF) entre si em termos de significância das variáveis que compuseram a fórmula final e faixas de estimativas dos coeficientes das variáveis, onde os resultados se encontram no capítulo 4.

4. RESULTADOS

4.1. Análise Univariada

A base de dados utilizada nessa etapa foi a base completa se subdivisão de amostras, que totaliza 22.132 registros distribuídos conforme as Tabelas 4.1 e 4.2 em termos de região e safras de contratação.

Tabela 4.1 – Distribuição de frequências das regiões do DF.

Região	Frequência	Percentual	Frequência Acumulada	Percentual Acumulado
BRASÍLIA	2203	9,95%	2203	9,95%
BRAZLÂNDIA	390	1,76%	2593	11,72%
CANDANGOLÂNDIA	173	0,78%	2766	12,50%
CEILÂNDIA	2671	12,07%	5437	24,57%
CRUZEIRO	772	3,49%	6209	28,05%
GAMA	1136	5,13%	7345	33,19%
GUARÁ	1545	6,98%	8890	40,17%
LAGO NORTE	331	1,50%	9221	41,66%
LAGO SUL	597	2,70%	9818	44,36%
NÚCLEO BANDEIRANTE	396	1,79%	10214	46,15%
PARANOÁ	638	2,88%	10852	49,03%
PLANALTINA	1323	5,98%	12175	55,01%
RECANTO DAS EMAS	778	3,52%	12953	58,53%
RIACHO FUNDO	697	3,15%	13650	61,68%
SAMAMBAIA	1488	6,72%	15138	68,40%
SANTA MARIA	1031	4,66%	16169	73,06%
SÃO SEBASTIAO	667	3,01%	16836	76,07%
SOBRADINHO	1614	7,29%	18450	83,36%
TAGUATINGA	3682	16,64%	22132	100,00%

Fonte: elaborado pelo autor.

Tabela 4.2 – Distribuição de frequências das safras de contratação.

Safra de Contratação	Frequência	Percentual	Frequência Acumulada	Percentual Acumulado
201312	2.131	9,63%	2.131	9,63%
201401	2.300	10,39%	4.431	20,02%
201402	2.310	10,44%	6.741	30,46%
201403	1.923	8,69%	8.664	39,15%
201404	2.280	10,30%	10.944	49,45%
201405	2.382	10,76%	13.326	60,21%
201406	2.366	10,69%	15.692	70,90%
201407	2.047	9,25%	17.739	80,15%
201408	2.248	10,16%	19.987	90,31%
201409	2.145	9,69%	22.132	100,00%

Fonte: elaborado pelo autor.

Pode-se notar que as regiões de Taguatinga, Ceilândia e Brasília foram as três regiões que mais possuem contratos de crédito no estudo e que as regiões Candangolândia, Lago Norte e Brazlândia são as que possuem o menor número de contratos. Com relação às safras, nota-se pela Tabela 4.2 que as quantidades estão bem distribuídas, não havendo nenhuma que apresente um número elevado de observações frente às demais.

Em seguida foram calculadas as estatísticas descritivas das variáveis candidatas a compor os modelos. Essa análise é muito importante pois possibilita ao analista identificar a presença de inconsistências, valores faltantes (*missings*), valores discrepantes (*outliers)*, variáveis com valor único, com poucos valores distintos ou com percentual elevado da população em determinado valor. Variáveis com inconsistências ou número elevado de valores faltantes ou discrepantes podem gerar erros nas estimativas dos modelos e as variáveis com valores únicos, poucos valores distintos ou valores concentrados em determinado atributo geralmente não discriminam o risco de crédito, pois toda ou a maior parte da população está contida em um único valor. Diante desses resultados, cabe ao analista a decisão de excluir, permanecer ou tratar as variáveis de forma que elas estejam aptas a participar do desenvolvimento do modelo.

Para as variáveis qualitativas ou quantitativas com pequena quantidade de valores distintos, a análise univariada foi realizada através das frequências. Para as variáveis quantitativas com grande quantidade de valores distintos foram calculadas as estatísticas média, mediana, máximo, mínimo e quartis da distribuição, onde os resultados encontrados estão dispostos nas Tabelas 4.3 a 4.7.

Tabela 4.3 – Distribuição de frequências da variável resposta Y.

Y	Frequência	Percentual	Frequência Acumulada	Percentual Acumulado
0	16.011	72,34%	16.011	72,34%
1	6.121	27,66%	22.132	100,00%

Fonte: elaborado pelo autor.

A Tabela 4.3 demonstra que a taxa de inadimplência geral (Y = 1) foi de 27,66%.

Tabela 4.4 – Distribuição de frequências da variável taxa de desemprego.

Taxa de Desemprego	Frequência	Percentual	Frequência Acumulada	Percentual Acumulado
4,6	4.646	20,99%	4.646	20,99%
4,7	4.429	20,01%	9.075	41,00%
4,8	4.068	18,38%	13.143	59,38%
4,9	2.300	10,39%	15.443	69,78%
5	4.558	20,59%	20.001	90,37%
5,1	2.131	9,63%	22.132	100,00%

Fonte: elaborado pelo autor.

Tabela 4.5 – Distribuição de frequências da variável taxa SELIC.

Taxa SELIC	Frequência	Percentual	Frequência Acumulada	Percentual Acumulado
9,9	2.131	9,63%	2.131	9,63%
10,4	2.300	10,39%	4.431	20,02%
10,65	4.233	19,13%	8.664	39,15%
10,9	13.468	60,85%	22.132	100,00%

Fonte: elaborado pelo autor.

Tabela 4.6 – Distribuição de frequências da variável inflação (IPCA).

Inflação (IPCA)	Frequência	Percentual	Frequência Acumulada	Percentual Acumulado
5,585	2.300	10,39	2.300	10,39
5,680	2.310	10,44	4.610	20,83
5,911	2.131	9,63	6.741	30,46
6,153	1.923	8,69	8.664	39,15
6,280	2.280	10,3	10.944	49,45
6,375	2.382	10,76	13.326	60,21
6,502	2.047	9,25	15.373	69,46
6,513	2.248	10,16	17.621	79,62
6,524	2.366	10,69	19.987	90,31
6,746	2.145	9,69	22.132	100

Fonte: elaborado pelo autor.

As Tabelas 4.4, 4.5 e 4.6 demostram a distribuição de frequências das variáveis macroeconômicas selecionadas para o estudo. Essas variáveis referem-se ao momento da

contratação do crédito e apresentaram poucos valores distintos pois o estudo utilizou somente 10 meses de contratação das operações de crédito.

Nota-se através da Tabela 4.4 que a variável taxa de desemprego apresentou somente 6 valores distintos, no entanto não existe concentração excessiva em um desses valores.

Nota-se através da Tabela 4.5 que a variável taxa Selic apresentou somente 4 valores distintos, sendo que 60% da população está contida no atributo de valor “10,9”.

Nota-se através da Tabela 4.6 que a variável inflação foi a que apresentou mais valores distintos dentre as variáveis macroeconômicas, sendo um valor distinto para cada mês de contratação e, por esse motivo, a distribuição de frequências dessa variável é semelhante à distribuição de frequências das safras, expostas na Tabela 4.2.

Tabela 4.7– Distribuição de frequências da variável Grau de Instrução.

Grau de Instrução	**Frequência**	**Percentual**	**Frequência Acumulada**	**Percentual Acumulado**
Analfabeto	123	0,56%	123	0,56%
Ensino fundamental incompleto	1.814	8,20%	1.937	8,76%
Ensino fundamental completo	1.288	5,82%	3.225	14,58%
Ensino Médio Incompleto	1.788	8,08%	5.013	22,66%
Ensino Médio Completo	8.898	40,20%	13.911	62,86%
Superior Incompleto	2.320	10,48%	16.231	73,34%
Superior Completo	4.782	21,61%	21.013	94,95%
Especialização	913	4,13%	21.926	99,08%
Mestrado	155	0,70%	22.081	99,78%
Doutorado	51	0,23%	22.132	100,00%

Fonte: elaborado pelo autor.

A Tabela 4.7 refere-se à distribuição de frequências da variável grau de instrução (escolaridade) dos tomadores de crédito. Apesar de ser uma variável informada pelo tomador durante a solicitação de crédito e não ser obrigatória apresentação de nenhum documento comprobatório, os resultados encontrados se mostraram coerentes, com a maior frequência de tomadores com Ensino Médio Completo, seguido de Superior Completo e Superior Incompleto. Conforme esperado Tomadores com Doutorado e Mestrado apresentaram as menores frequências.

Tabela 4.8 – Estatísticas descritivas das variáveis quantitativas.

Variável	Média	Mediana (Q2)	Máximo	Mínimo	Q1	Q3
Idade	41	40	99	16	31	50
Tempo de Relacionamento	33,8	5	750	0	0	35
Renda Formal	5,1817	2,8587	455,4394	0	1,2707	6,5502
Prazo Contratado	28	31	36	12	23	35

Fonte: elaborado pelo autor.

A Tabela 4.8 demonstra as estatísticas calculadas para as variáveis quantitativas com grande número de valores distintos.

Observa-se que a base de dados contém tomadores de 16 a 99 anos de idade. Cabe ressaltar que a concessão de crédito para cliente menores de idade é permitida caso o menor seja emancipado.

A variável Tempo de Relacionamento indica que mais da metade dos tomadores de crédito possui relacionamento recente com a instituição, uma vez que a mediana da distribuição foi de 5 meses e que pelo menos um quarto dos tomadores não possuíam relacionamento prévio com a instituição. Nota-se também um cliente que possui relacionamento há mais de 62 anos (750 meses).

A variável Renda Formal indica que pelo menos um quarto dos tomadores não possui renda formal e metade recebe até 2,85 salários mínimos. A população total do estudo possui média de 5,18 salários mínimos de renda formal mensal.

Nota-se também que o prazo médio contratado para esse produto é de 28 meses. Uma vez que o prazo máximo é de 36 meses, pode-se considerar essa média elevada.

Diante dos resultados obtidos, decidiu-se manter todas as variáveis para a realização da próxima etapa do estudo, que consistiu na análise bivariada.

4.2. Análise Bivariada

A análise bivariada consistiu em realizar uma frequência cruzada entre as variáveis preditoras candidatas a compor o modelo com a variável resposta, com o objetivo de verificar se essas variáveis discriminam o risco de crédito.

Através dessa análise é possível categorizar as variáveis preditoras de acordo com seu comportamento de risco e, a partir dessa categorização, criar variáveis *dummies* para compor os modelos finais.

A métrica utilizada para quantificar o risco de crédito de cada categoria das variáveis preditoras foi o Risco Relativo, dado pela seguinte fórmula:

$$Risco\ Relativo\ da\ categoria = \frac{\frac{Totalde\ clientes\ bons\ na\ categoria}{Total\ de\ clientes\ bons}}{\frac{Total\ de\ clientes\ maus\ na\ categoria}{Total\ de\ clientes\ maus}} \quad (4.1)$$

Nota-se por (4.1) que o risco relativo é determinado pelo percentual de bons da categoria com relação ao total de bons, dividido pelo percentual de maus da categoria com relação ao total de maus. Assim, quanto maior o valor dessa métrica, maior é quantidade de bons na categoria em relação aos maus e, consequentemente, menor é o risco de crédito da categoria. Valores próximos de 1 indicam que essa categoria é neutra em relação ao risco de crédito e valores abaixo de 1 indicam que a categoria possui maior risco de crédito.

A categorização das variáveis foi realizada da seguinte maneira: para variáveis que possuem poucos atributos (grau de instrução, taxa SELIC e taxa de desemprego) foi calculado o risco relativo para todos os possíveis valores e, em seguida, agrupados os atributos com valores próximos de risco relativo. Para as variáveis quantitativas que possuem um número grande de possíveis valores foram criadas 20 categorias iniciais, baseadas nos percentis da distribuição e, em seguida, essas categorias foram agrupadas de acordo com o risco relativo, respeitando as seguintes premissas:

1. A categorização deve ser monotônica crescente ou decrescente com relação aos valores dos atributos e do risco relativo das categorias; e
2. As categorias devem possuir intervalos superiores a 0,10 unidades de risco relativo entre si.

Após realizado esse procedimento, notou-se que os possíveis valores das variáveis taxa de desemprego e inflação, expostos nas Tabelas 4.9 e 4.10, apresentaram valores muito próximos de risco relativo, indicando que todos os atributos dessas variáveis correspondem a níveis semelhantes de risco de crédito e, consequentemente, não discriminam o risco de crédito do público alvo dessa dissertação. Nota-se também que ocorre inversões no risco relativo conforme os valores das variáveis aumentam, enquanto os valores esperados de uma variável que discrimine risco de crédito sejam monotônicos no valor dos atributos e no valor do risco relativo.

Tabela 4.9 – Risco Relativo da variável taxa de desemprego.

Taxa de Desemprego	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
4,60	1,0207	3.380	1.266	4.646
4,70	1,0194	3.221	1.208	4.429
4,80	1,0100	2.951	1.117	4.068
4,90	0,9439	1.637	663	2.300
5,00	1,0321	3.326	1.232	4.558
5,10	0,9007	1.496	635	2.131

Fonte: elaborado pelo autor.

Tabela 4.10 – Risco Relativo da variável inflação.

Inflação (IPCA)	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
5,5852	0,9439	1.637	663	2.300
5,6797	1,0398	1.689	621	2.310
5,9108	0,9007	1.496	635	2.131
6,1530	0,9617	1.376	547	1.923
6,2797	1,0420	1.668	612	2.280
6,3750	1,0701	1.755	627	2.382
6,5023	0,9646	1.466	581	2.047
6,5129	1,0243	1.637	611	2.248
6,5236	1,0008	1.712	654	2.366
6,7464	1,0564	1.575	570	2.145

Fonte: elaborado pelo autor.

Uma possível explicação para o resultado da variável taxa de desemprego é a presença expressiva de pessoas concursadas residentes no DF, sejam funcionários públicos federais ou estaduais, onde variações nos índices de desemprego podem não afetar grande parte dessa população. Outro fato a ser considerado é a utilização de um curto período de tempo no estudo, fazendo com que ambas as variáveis apresentassem poucos valores distintos. Diante do exposto, as variáveis taxa de desemprego e inflação foram excluídas do estudo.

A variável Renda Formal apresentou um comportamento inesperado durante sua categorização, pois esperava-se que quanto maior a renda do tomador menor seria seu risco de crédito, no entanto, conforme categorização inicial exposta na Tabela 4.11, a classe 3 (clientes com renda formal de 3,5 a 4 salários mínimos) possui risco relativo melhor do que os clientes com faixas de renda superiores, sendo que essa inversão também ocorre com outras faixas de renda tais como as classes 7 e a 10. Essa inversão pode influenciar os resultados das estimativas dessas categorias nos modelos de

regressão, gerando incoerências ou até fazendo com que essas classes não sejam significativas estatísticamente.

Tabela 4.11 – Categorização e Risco Relativo iniciais da variável Renda Formal.

Classe	Categorização	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
1	< 0,5	0,6923	2.010	1.110	3.120
2	[0,5 ; 3,5[	0,9041	6.558	2.773	9.331
3	[3,5 ; 4,0[	1,4540	677	178	855
4	[4,0 ; 5,5[	1,0918	1.525	534	2.059
5	[5,5 ; 6,0[	1,3202	518	150	668
6	[6,0 ; 7,5[	1,0556	1.121	406	1.527
7	[7,5 ; 8,0[	1,9509	347	68	415
8	[8,0 ; 8,5[	1,3691	265	74	339
9	[8,5 ; 9,0[	1,0407	245	90	335
10	[9,0 ; 9,5[	2,3455	227	37	264
11	[9,5 ; 10 [	1,0384	201	74	275
12	[10 ; 15 [	1,6485	1.285	298	1.583
13	> = 15	1,1992	1.032	329	1.361

Fonte: elaborado pelo autor.

Após junção das classes, a categorização final dessa variável está exposta na Tabela 4.12.

Tabela 4.12 – Categorização e Risco Relativo finais da variável Renda Formal.

Classe	Categorização	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
1	> = 7,5	1,4196	3.602	970	4.572
2	[3,5 ; 7,5[	1,1580	3.841	1.268	5.109
3	< 3,5	0,8435	8.568	3.883	12.451

Fonte: elaborado pelo autor.

A categorização final das demais variáveis selecionadas encontra-se nas tabelas 4.13 a 4.17.

Tabela 4.13 - Categorização e Risco Relativo da variável Grau de Instrução.

Classe	Categorização	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
1	Doutorado	6,1168	48	3	51
2	Mestrado	2,1941	132	23	155
3	Especialização ou Superior Completo	1,5530	4.570	1.125	5.695
4	Superior Incompleto ou menor Grau de Instrução	0,8662	11.261	4.970	16.231

Fonte: elaborado pelo autor.

Nota-se pela Tabela 4.13 que quanto maior o Grau de Instrução do tomador de crédito menor é seu risco, com os doutores apresentando um risco relativo bem superior aos demais.

Tabela 4.14 – Categorização e Risco Relativo da variável Idade.

Classe	Categorização	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
1	> 55	2,2855	3.019	505	3.524
2	] 49 ; 55]	1,5760	1.954	474	2.428
3	] 40 ; 49]	1,1970	3.610	1.153	4.763
4	] 30 ; 40]	0,8634	4.275	1.893	6.168
5	< = 30	0,5751	3.153	2.096	5.249

Fonte: elaborado pelo autor.

Nota-se pela Tabela 4.14 que quanto maior a idade do tomador de crédito menor é seu risco. Destaque negativo para os tomadores de crédito menores do que 30 anos, que apresentaram um risco relativo de 0,57.

Tabela 4.15 – Categorização e Risco Relativo da variável Prazo Contratado.

Classe	Categorização	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
1	< = 12	1,9630	724	141	865
2	] 12 ; 24]	1,4197	3747	1.009	4.756
3	< = 24	0,8875	11.540	4.971	16.511

Fonte: elaborado pelo autor.

Nota-se pela Tabela 4.15 que quanto menor o prazo contratado, menor é o risco de crédito da operação. Esse fato pode ser explicado pela menor possibilidade decorrência de imprevistos no curto prazo, favorecendo o cumprimento das obrigações do tomador de crédito.

Tabela 4.16 – Categorização e Risco Relativo da variável Tempo de Relacionamento.

Classe	Categorização	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
1	> 50	2,9392	3.798	494	4.292
2	] 20 ; 50]	1,6576	2.337	539	2.876
3	] 4 ; 20]	1,0095	3.343	1.266	4.609
4	< = 4	0,6535	6.533	3.822	10.355

Fonte: elaborado pelo autor.

Nota-se pela tabela 4.16 que os clientes com menor tempo de relacionamento com a instituição são os que possuem maior risco de crédito. Os clientes que possuem mais de 50 meses de relacionamento possuem menor risco de crédito.

Tabela 4.17 – Categorização e Risco Relativo da variável Taxa SELIC.

Classe	Categorização	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
1	>= 10	1,0115	14.515	5.486	20.001
2	< 10	0,9007	1.496	635	2.131

Fonte: elaborado pelo autor.

A taxa SELIC é a taxa básica de juros da economia brasileira. O aumento da SELIC faz com que a captação de recursos por parte das instituições financeiras fique mais cara o que, consequentemente, encarece as operações de crédito. Juros maiores nas operações de crédito diminuem o poder de compra do tomador de crédito, e, por esse motivo, esperava-se que quanto maior a taxa SELIC maior seja a inadimplência e o risco de crédito. No entanto, conforme pode ser observado na Tabela 4.17, os resultados obtidos foram o inverso do esperado, com risco relativo menor (maior risco de crédito) para valores de SELIC abaixo de 10,00% e menor risco de crédito para valores acima de 10,00%.

Mesmo diante dos resultados apresentados, decidiu-se manter a variável taxa SELIC no estudo por ser a única variável macroeconômica remanescente. Estudos posteriores utilizando um público alvo mais abrangente devem ser realizados para um melhor diagnóstico dessa variável.

Variáveis *dummies* são variáveis binárias (assumem valor 0 ou 1) criadas a partir da categorização das variáveis originais e serão utilizadas na composição dos modelos finais de regressão.

4.3. Indicadores Espaciais

Após as etapas de análise univariada e bivariada, o próximo passo do estudo consistiu em calcular a taxa de inadimplência das regiões do Distrito Federal para, em seguida, aplicar os Índices de Moran Global e Local com o objetivo de verificar a existência de correlação espacial ou regiões singulares no universo de estudo.

Os resultados das taxas de inadimplência por região estão dispostos na Tabela 4.18 e a distribuição espacial se encontra na Figura 4.1.

Tabela 4.18 – Taxas de Inadimplência por região do DF.

Região	Quantidade de Inadimplentes	Quantidade Total	Taxa de Inadimplência
LAGO SUL	79	597	13,233%
CRUZEIRO	136	772	17,617%
BRASÍLIA	423	2.203	19,201%
GUARÁ	373	1.545	24,142%
LAGO NORTE	82	331	24,773%
TAGUATINGA	921	3.682	25,014%
NÚCLEO BANDEIRANTE	107	396	27,020%
SOBRADINHO	441	1.614	27,323%
GAMA	330	1.136	29,049%
SAMAMBAIA	441	1.488	29,637%
RIACHO FUNDO	221	697	31,707%
BRAZLÂNDIA	124	390	31,795%
CEILÂNDIA	882	2.671	33,021%
SÃO SEBASTIAO	222	667	33,283%
PLANALTINA	441	1.323	33,333%
CANDANGOLÂNDIA	58	173	33,526%
SANTA MARIA	347	1.031	33,657%
RECANTO DAS EMAS	267	778	34,319%
PARANOÁ	226	638	35,423%

Fonte: elaborado pelo autor.

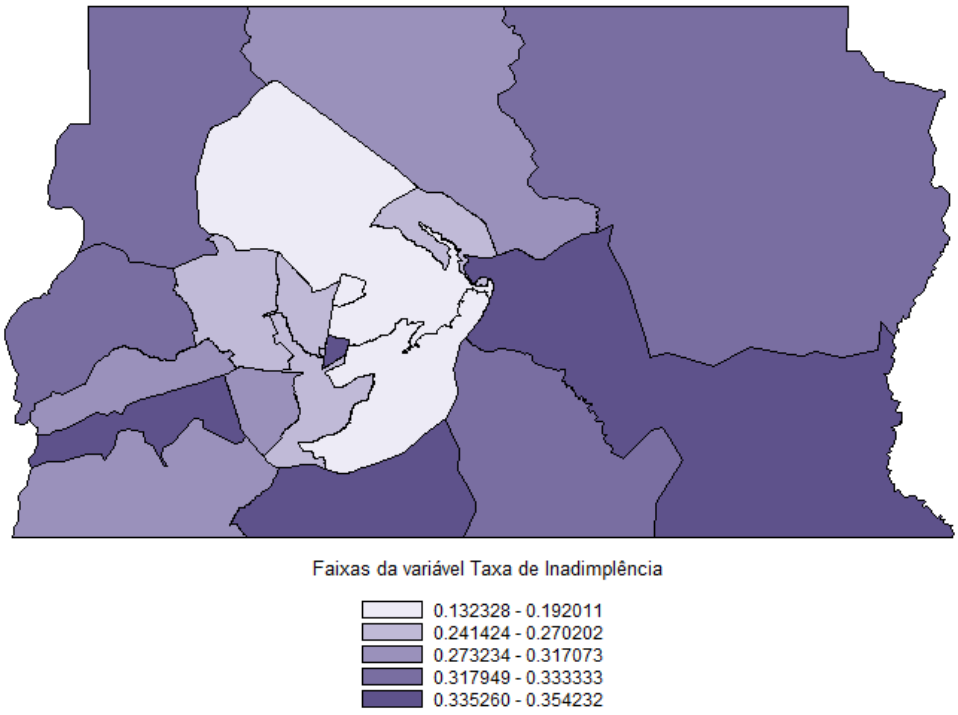

Figura 4.1– Distribuição espacial das taxas de inadimplência do Distrito Federal.
Fonte: elaborado pelo autor.

Nota-se através da Tabela 4.18 que a região do Lago Sul foi a que apresentou a menor taxa de inadimplência dentre as regioes estudadas, seguidas das regiões Cruzeiro e Brasília, com todas apresentando taxas inferiores a 20%. Nota-se também, através da Figura 4.1, que as três regiões estão localizadas no centro do Distrito Federal.

Ainda analisando a Figura 4.1 nota-se que a medida que se afasta do ponto central do DF, as taxas de inadimplência aumentam (representadas pelas áreas mais escuras do mapa), ou seja, pode-se concluir que as regiões mais afastadas do ponto central do DF possuem maior risco de crédito no produto CDC para a instituição financeira estudada. Destaque negativo para as regiões de Santa Maria, Recanto das Emas e Paranoá, que apresentam as piores taxas de inadimplência (33,657%, 34,319% e 35,423% respectivamente).

Conforme exposto na Tabela 4.3, a taxa de inadimplência geral do DF foi de 27,66%, assim, pode-se observar pela Tabela 4.18 que apenas 7 regiões (Lago Sul, Cruzeiro, Brasília, Guará, Lago Norte, Taguatinga e Núcleo Bandeirante) possuem taxas de inadimplência abaixo da média geral.

Conforme já dito anteriormente, o teste de correlação espacial da variável taxa de inadimplência da operação de crédito CDC do Distrito Federal foi realizado através do Índice de Moran global, que apresentou o valor de 0,05, indicando uma dependência espacial muito baixa.

Esse resultado traz à luz a seguinte discussão: suponhamos que o valor obtido para essa correlação fosse próximo de 1, indicando uma forte correlação positiva; nesse caso, quando houvesse um aumento na inadimplência de uma das regiões, a taxa de inadimplência das demais regiões também aumentaria, pois a correlação entre elas seria muito alta. Nesse caso de correlações positivas altas faria sentido construir modelos distintos de *credit scoring* para cada região?

Suponhamos agora que essa correlação seja próxima de -1, indicando uma forte correlação negativa: quando houvesse um aumento na taxa de inadimplência de determinada região, as demais diminuiriam em decorrência da correlação negativa. Nesse caso de correlações negativas, faria sentido possuir modelos distintos de *credit scoring* (por exemplo, os parâmetros das variáveis poderiam possuir sinais opostos), pois o comportamento das mesmas em relação à inadimplência é oposto. Já uma correlação inexistente não implica que as regiões sejam independentes entre si, mas não descarta a possibilidade de estimar modelos distintos para as mesmas.

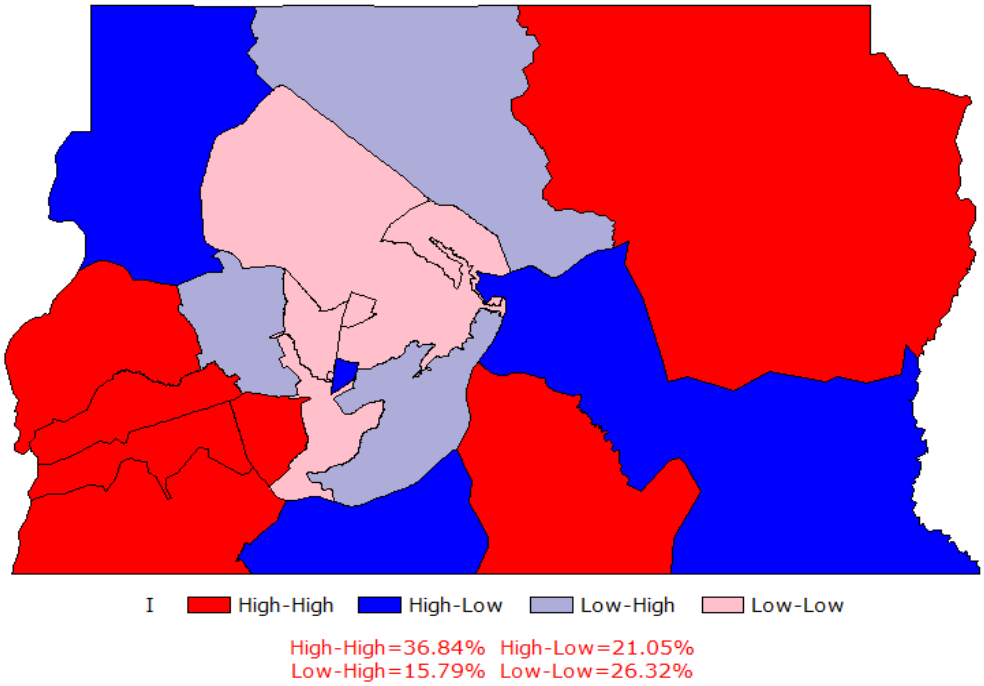

Figura 4.2– Mapa de espalhamento de Moran.
Fonte: elaborado pelo autor.

A Figura 4.2 apresenta o mapa de espalhamento de Moran, onde as regiões coloridas em tons de vermelho apresentam dependência espacial positiva, enquanto as regiões coloridas em tons de azul apresentam dependência espacial negativa. As regiões do tipo "*Low-Low*" são as que apresentaram as menores taxas de inadimplência, seguidas das regiões *"Low-High"*, *"High-Low"* e *"High-High"*, sendo que esses resultados podem ser considerados *clusters* espaciais da variável taxa de inadimplência. Essa informação poderia ser utilizada pela instituição financeira para a definição de público alvo de campanhas de recuperação de crédito, em que a cobrança dos clientes residentes nas regiões *"High-High"* devem ser o foco inicial das ações, visando melhorar o resultado financeiro da empresa. A Tabela 4.19 apresenta os grupos de regiões apresentados na Figura 4.2.

Tabela 4.19 – I de Moran das regiões do DF.

Região	I de Moran
CEILÂNDIA	*High-High*
GAMA	*High-High*
PLANALTINA	*High-High*
RECANTO DAS EMAS	*High-High*
RIACHO FUNDO	*High-High*
SAMAMBAIA	*High-High*
SÃO SEBASTIAO	*High-High*
BRAZLÂNDIA	*High-Low*
CANDANGOLÂNDIA	*High-Low*
PARANOÁ	*High-Low*
SANTA MARIA	*High-Low*
LAGO SUL	*Low-High*
SOBRADINHO	*Low-High*
TAGUATINGA	*Low-High*
BRASÍLIA	*Low-Low*
CRUZEIRO	*Low-Low*
GUARÁ	*Low-Low*
LAGO NORTE	*Low-Low*
NÚCLEO BANDEIRANTE	*Low-Low*

Fonte: elaborado pelo autor.

Os resultados encontrados para o Índice Local de Moran utilizando um nível de significância de 95% são apresentados no Mapa de Moran, contido na Figura 4.3.

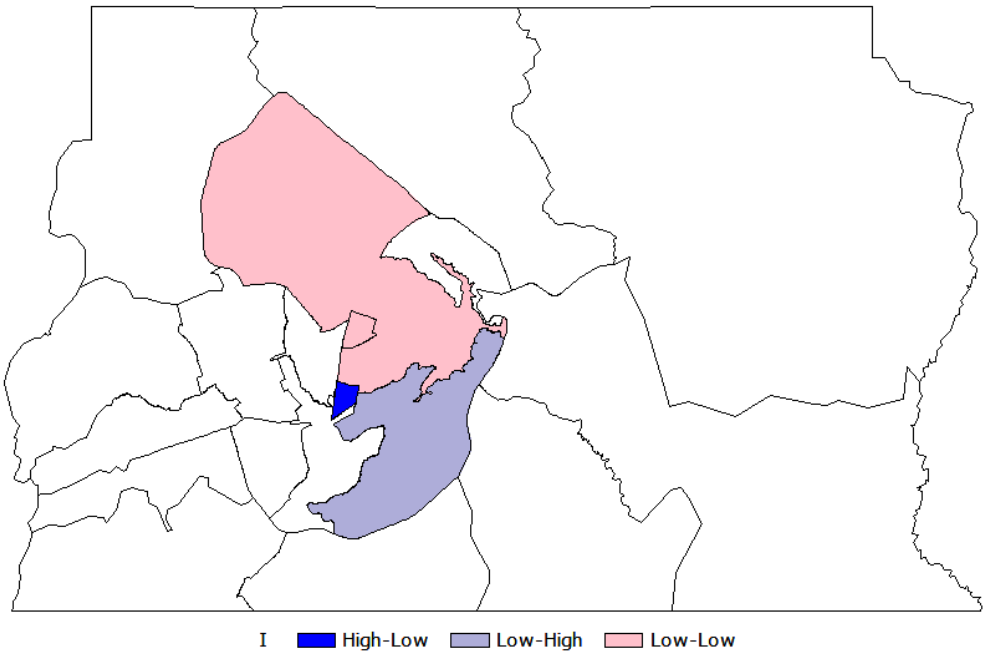

Figura 4.3 – Mapa de Moran a 95% de confiança.
Fonte: elaborado pelo autor.

O mapa de Moran indica a existência de correlações locais em algumas regiões que são significativamente diferentes das demais, revelando indícios de heterogeneidade espacial. As regiões significativas no índice local e que estão demarcadas na Figura 4.3 são: Brasília e Cruzeiro (*Low-Low*), Lago Sul (*Low-High*) e Candangolândia (*High-Low*). A existência de valores significativos para o índice de Moran local justifica a aplicação da técnica GWLR para verificar se as fórmulas dos modelos de regressão a serem obtidas para essas regiões são distintas das demais regiões.

4.4. Modelo Global via Regressão Logística

O modelo global foi desenvolvido utilizando a amostra total de desenvolvimento contendo 10.944.

As variáveis utilizadas no desenvolvimento do modelo foram todas as *dummies* criadas a partir das categorizações apresentadas nas Tabelas 4.12 a 4.17. Utilizando o método de seleção de variáveis *stepwise*, as variáveis com p-valor abaixo de 0,10 (nível de significância $\alpha = 10\%$) e que foram selecionadas para compor o modelo final de regressão logística (modelo global) são apresentadas na Tabela 4.20.

Tabela 4.20 – Variáveis finais do modelo global e respectivos coeficientes.

Variáveis	Coeficientes	Desvio Padrão	Estatística de Wald	Razão de Chances
Intercepto	-1,3068	0,0893	-14,6338*	-
d_idade1	-0,5665	0,084	-6,7440*	0,567
d_idade2	-0,2891	0,0907	-3,1874*	0,749
d_idade4	0,1481	0,0635	2,3323*	1.160
d_idade5	0,5684	0,0653	8,7044*	1.765
d_instrucao4	0,3019	0,0614	4,9169*	1.352
d_tempo_rel1	-0,7764	0,0862	-9,0070*	0,460
d_tempo_rel2	-0,3529	0,0844	-4,1813*	0,703
d_tempo_rel4	0,4206	0,0566	7,4311*	1.523
d_renda1	0,3742	0,0705	5,3078*	1.454
d_renda2	0,1135	0,06	1,8917**	1.120
d_pz_contratado1	-0,6099	0,1398	-4,3627*	0,543
d_pz_contratado2	-0,4165	0,0541	-7,6987*	0,659

* p-valor abaixo de 0,05.
** p-valor abaixo de 0,10.

Fonte: elaborado pelo autor.

Note que a variável SELIC não se mostrou significativa e portanto não foi selecionada para compor o modelo final de regressão global. Uma possível explicação

para esse fato é a utilização de um período curto de contratação, culminando em poucos valores distintos para essa variável.

Note também que os coeficientes para a variável Renda Formal se mostraram invertidos, onde as melhores faixas de renda (d_renda1 e d_renda2) obtiveram piores coeficientes com relação à pior faixa (d_renda3, cujo coeficiente é zero). Esse resultado pode ser explicado pelo comportamento da variável, que possui diversas inversões em suas faixas de valores. A categorização foi realizada com a base total de registros e o modelo foi desenvolvido com a base de desenvolvimento, e isso pode ter influenciado tal resultado inesperado.

A nomenclatura das variáveis *dummies* respeita a nomenclatura das categorias expostas nas Tabelas 4.12 a 4.17. Por exemplo, a *dummy* d_idade1 representa a categoria de idade "> 55 anos" e é a melhor categoria dessa variável com relação ao risco de crédito e a *dummy* d_instrucao4 representa os clientes que possuem a categoria "Superior Incompleto ou menor grau de instrução", sendo essa a pior categoria da variável Grau de Instrução com relação ao risco de crédito.

A variável resposta Y possui como evento de interesse a ocorrência da inadimplência (Y=1), sendo que a probabilidade resultante dos modelos de regressão logística e via GWLR referem-se à probabilidade de ocorrência desse evento, ou seja, do cliente se tornar inadimplente. Isso posto, pode-se notar através da Tabela 4.20 que todos os coeficientes da regressão global se mostraram coerentes, uma vez que as melhores categorias de cada variável com relação ao risco de crédito apresentaram menores coeficientes em relação às categorias de maior risco da mesma variável, isto é, a presença das melhores categorias de cada variável diminui a probabilidade do cliente se tornar inadimplente. Essa análise é denominada análise de congruência e é importante para verificar a existência de inversões nos coeficientes e se a categorização das variáveis foi realizada de maneira correta.

A Razão de Chances (em inglês *Odds Ratio*) contida na Tabela 4.20 é calculada por meio da probabilidade de ocorrência do evento inadimplência (Y=1) dividido pela probabilidade de ocorrência do evento adimplência (Y=0) na variável em questão. A variável d_tempo_rel1 foi a que apresentou o menor valor para a Razão de Chances (0,46) e sua interpretação é que a chance de ocorrência do evento inadimplência (Y=1) nessa variável é de 0,46 para 1, ou seja, a chance de ocorrência de adimplência nessa categoria

é maior. Valores iguais a 1 significam que a chance de ocorrência de ambos os eventos é igual (1 para 1) e valores acima de 1 indicam que a chance de ocorrência de inadimplência é maior para essa variável/categoria. Note que as melhores categorias possuem Razão de Chances menores frente às demais categorias da mesma variável (por exemplo, Razão de Chances de d_idade1 = 0,586, enquanto a Razão de Chances de d_idade2 = 0,769), indicando que a chance de ocorrência de clientes inadimplentes em d_idade2 é maior do que a chance de ocorrência em d_idade1 e demonstrando resultados coerentes. O valor encontrado da Razão de Chances para todas as variáveis também se mostrou coerente frente aos valores dos coeficientes das variáveis, onde os piores coeficientes (maiores em termos de valor) apresentaram os maiores valores de Razão de Chances (d_idade5, d_tempo_rel4 e d_selic2).

O valor encontrado para o critério informacional AICc do modelo global foi 12.098,29, sendo esse o valor que foi utilizado para a comparação com os modelos estimados via GWLR, cujos resultados são apresentados a seguir.

4.5. Modelos Locais via GWLR

Conforme relatado no capitulo 3 dessa dissertação, foram desenvolvidos inicialmente 4 modelos utilizando a técnica Regressão Logística Geograficamente Ponderada (GWLR), sendo que o modelo 1 utilizou a função de ponderação Gaussiana Fixa, o modelo 2 utilizou a função de ponderação Gaussiana Variável, a função de ponderação Biquadrática Fixa e o modelo 4 utilizou a Biquadrática Variável.

As variáveis utilizadas para compor todos os modelos foram as mesmas selecionadas pelo modelo de regressão logística e expostas na Tabela 4.20.

O valor do *bandwidth* fixo ótimo, o número *k* ótimo de vizinhos mais próximos para estimar os *bandwidths* variáveis e os resultados do critério informacional AICc de cada modelo encontram-se na Tabela 4.21.

Tabela 4.21 – Resultados do Modelos via GWLR para diferentes Funções de Ponderação.

Modelo GWLR	**Valor do *Bandwidth* (em km)**	**Valor de *k* vizinhos mais próximos**	**AICc**
Gaussiano Fixo	57	-	12.097,83
Gaussiano Variável	-	2.022	12.091,19

Biquadrático Fixa	57	-	12.098,13
Biquadrático Variável	-	10.944	12.095,21

Fonte: elaborado pelo autor.

Nota-se através da Tabela 4.21 que o valor encontrado para o bandwidth ótimo dos modelos Gaussiano Fixo e Biquadrático Fixo foi de 57 km, sendo esse valor o valor máximo possível a ser utilizado e que abrange todas as regiões do estudo. O modelo Biquadrático Variável também selecionou o número total de observações para desenvolvimento do modelo (10.944) e todas as regressões estimadas via GWLR obtiveram melhor resultado frente ao modelo global, cujo valor do critério informacional AICc foi de 12.098,29 (quanto menor o valor do AICc melhor é o modelo), indicando ganhos no desenvolvimento de modelos regionais. O melhor modelo estimado via GWLR foi o modelo 2, estimado via função de ponderação Gaussiana Variável, cuja performance foi comparada com o modelo global desenvolvido via regressão logística e os resultados são apresentados a seguir.

A Tabela 4.22 contêm as estatísticas descritivas dos coeficientes estimados do modelos Gaussiano Variável, eleito o melhor entre os modelos via GWLR, onde nota-se que a média dos coeficientes ficaram bem próximas com relação aos coeficientes do modelo global expostos na Tabela 4.20.

Tabela 4.22 – Estatísticas dos coeficientes estimados do modelo Gaussiano Variável.

Variável	Média	Desvio Padrão	Mínimo	Máximo	Amplitude	Q1	Mediana (Q2)	Q3
Intercepto	-1,2950	0,0432	-1,3923	-1,2006	0,1917	-1,3201	-1,2847	-1,2689
d_idade1	-0,6557	0,1193	-1,0145	-0,4850	0,5295	-0,7164	-0,6283	-0,5676
d_idade2	-0,3230	0,0950	-0,4969	-0,1507	0,3462	-0,3586	-0,3319	-0,2660
d_idade4	0,0749	0,0760	-0,0987	0,2164	0,3151	0,0272	0,0616	0,1320
d_idade5	0,5054	0,0696	0,3130	0,5910	0,2780	0,4852	0,5275	0,5605
d_instrucao4	0,3004	0,0376	0,2124	0,3518	0,1394	0,2851	0,2979	0,3347
d_tempo_rel1	-0,6720	0,1019	-0,8264	-0,4858	0,3406	-0,7626	-0,6894	-0,5817
d_tempo_rel2	-0,3436	0,0513	-0,4208	-0,2314	0,1894	-0,3716	-0,3465	-0,3213
d_tempo_rel4	0,4614	0,0543	0,3498	0,5573	0,2075	0,4393	0,4430	0,5201
d_renda1	0,3272	0,0732	0,2173	0,4769	0,2596	0,2680	0,3222	0,3638
d_renda2	0,1255	0,0443	0,0247	0,1791	0,1544	0,0996	0,1469	0,1669
d_pz_contratado1	-0,6241	0,1160	-0,7555	-0,3766	0,3789	-0,7183	-0,6849	-0,5065
d_pz_contratado2	-0,4134	0,0332	-0,4516	-0,3327	0,1189	-0,4479	-0,4177	-0,3904

Fonte: elaborado pelo autor.

A Tabela 4.23 contém a formula dos modelos estimados via GWLR Gaussiano Variável para as 19 regiões do Distrito Federal presentes nessa dissertação.

Tabela 4.23 – Fórmulas de Regressão Locais estimadas pelo modelo Gaussiano Variável.

Região	Intercepto	d_idade1	d_idade2	d_idade4	d_idade5	d_instrucao4	d_tempo_rel1	d_tempo_rel2	d_tempo_rel4	d_renda1	d_renda2	d_pz_contratado1	d_pz_contratado2
BRASÍLIA	-1,304	-0,839	-0,266	0,028*	0,438	0,231	-0,581	-0,321	0,520	0,291	0,100*	-0,468	-0,371
BRAZLÂNDIA	-1,320	-0,571	-0,363	0,097*	0,522	0,310	-0,691	-0,330	0,496	0,367	0,109*	-0,685	-0,431
CANDANGOLÂNDIA	-1,256	-0,740	-0,340	0,011*	0,438	0,282	-0,636	-0,346	0,455	0,261	0,128	-0,618	-0,396
CEILÂNDIA	-1,342	-0,485	-0,497	0,090*	0,548	0,351	-0,763	-0,421	0,537	0,477	0,147*	-0,712	-0,448
CRUZEIRO	-1,326	-1,015	-0,351	-0,099*	0,313	0,234	-0,485	-0,231*	0,557	0,268	0,107*	-0,426	-0,333
GAMA	-1,285	-0,619	-0,334	0,132	0,572	0,296	-0,826	-0,366	0,443	0,323	0,179	-0,647	-0,363
GUARÁ	-1,248	-0,758	-0,359*	-0,046*	0,372	0,323	-0,527	-0,265	0,430	0,217	0,101*	-0,685	-0,418
LAGO NORTE	-1,378	-0,755	-0,156	0,148*	0,522	0,212	-0,555	-0,289	0,554	0,327	0,043*	-0,377	-0,370
LAGO SUL	-1,257	-0,716	-0,308	0,057*	0,489	0,268	-0,745	-0,407	0,423	0,292	0,122*	-0,528	-0,396
NÚCLEO BANDEIRANTE	-1,258	-0,678	-0,344	0,060*	0,492	0,290	-0,709	-0,363	0,442	0,281	0,145	-0,642	-0,390
PARANOÁ	-1,289	-0,609	-0,172*	0,176	0,585	0,283	-0,808	-0,409	0,350	0,374	0,069*	-0,455	-0,428
PLANALTINA	-1,315	-0,542	-0,205	0,193	0,591	0,298	-0,771	-0,346	0,363	0,394	0,072*	-0,556	-0,434
RECANTO DAS EMAS	-1,253	-0,628	-0,372	0,079*	0,530	0,300	-0,741	-0,378	0,459	0,321	0,155	-0,692	-0,398
RIACHO FUNDO	-1,201	-0,664	-0,357	0,043*	0,484	0,278	-0,682	-0,372	0,434	0,271	0,159	-0,739	-0,404
SAMAMBAIA	-1,269	-0,623	-0,408	0,062*	0,527	0,317	-0,689	-0,364	0,482	0,346	0,147	-0,718	-0,429
SANTA MARIA	-1,286	-0,628	-0,332	0,124	0,561	0,297	-0,807	-0,367	0,439	0,322	0,167	-0,627	-0,373
SÃO SEBASTIAO	-1,273	-0,624	-0,247	0,141	0,567	0,289	-0,822	-0,408	0,373	0,354	0,108*	-0,507	-0,418
SOBRADINHO	-1,392	-0,568	-0,151*	0,216	0,578	0,285	-0,625	-0,273	0,456	0,364	0,025*	-0,470	-0,412
TAGUATINGA	-1,271	-0,666	-0,312	0,027*	0,485	0,335	-0,582	-0,322	0,439	0,259	0,173	-0,756	-0,452

* p-valor acima de 0,10 (coeficiente não significativo com 90% de confiança).

Fonte: elaborado pelo autor.

A significância dos coeficientes estimados bem como a distribuição espacial desses coeficientes, estão expostos nas Figuras 4.4 a 4.17.

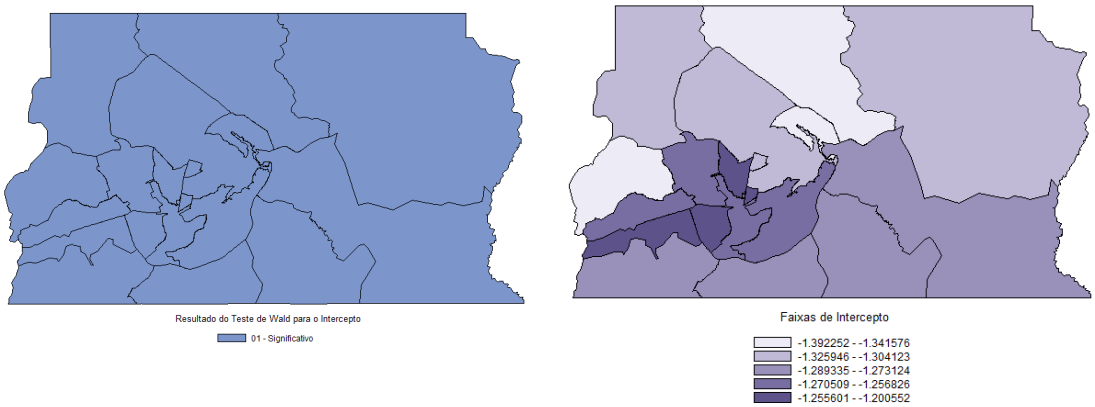

Figura 4.4 – Distribuição espacial da significância e das estimativas do Intercepto.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.4 que o intercepto foi significativo para todas as regiões do Distrito Federal e variou de -1,3922 a -1,2005, indicando diferença regional entre os valores estimados.

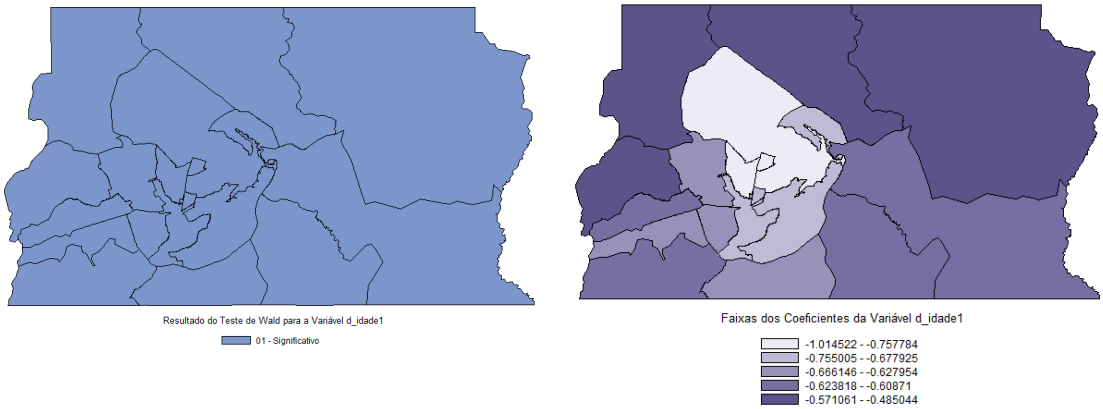

Figura 4.5– Distribuição espacial da significância e das estimativas da variável d_idade1.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.5 que a variável d_idade1 também se mostrou significativa para todas as regiões do Distrito Federal. Observa-se que os valores do coeficiente variaram de -1,0145 a -0,4850, sendo a região que apresentou o menor valor (-1,0145) foi o Cruzeiro e o maior (-0,4850) foi a região de Ceilândia. Essa variação entre as regiões comprova que a variável d_idade1 influencia o risco de crédito de maneira distinta de região para região, fazendo com que o desenvolvimento de modelos regionais seja justificável. Esse comportamento foi observado em todas as variáveis, demonstradas a seguir.

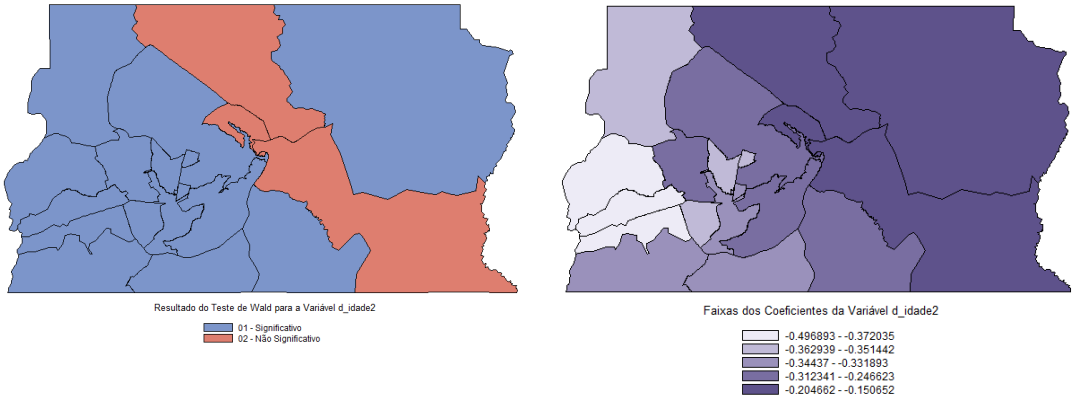

Figura 4.6– Distribuição espacial da significância e das estimativas da variável d_idade2.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.6 que a variável d_idade2 não se mostrou significativa para as regiões Sobradinho, Lago Norte e Paranoá. Nota-se que as três regiões são limítrofes e um dos motivos para que uma variável apresente o mesmo resultado (significante ou não) para regiões limítrofes é o fato da GWLR dar maior peso para as informações mais próximas geograficamente. A região na qual a presença desse atributo mais influencia a diminuição da probabilidade de inadimplência é Ceilândia (-0,497), enquanto a região em que essa variável menos influencia (dentre as regiões cujo coeficiente foi significante) é Planaltina (-0,205).

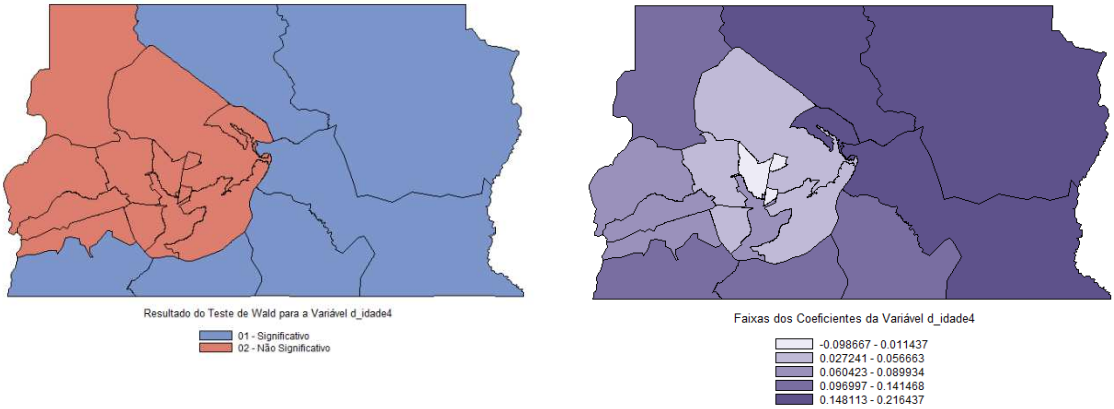

Figura 4.7– Distribuição espacial da significância e das estimativas da variável d_idade4.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.7 que a variável d_idade4 se mostrou significativa para as regiões Gama, Santa Maria, São Sebastião, Paranoá, Sobradinho e Planaltina, onde novamente observamos a influência da localização geográfica para a significância das variáveis. Apesar de ter apresentado valores negativos e positivos, os valores negativos não se mostraram com 90% de confiança.

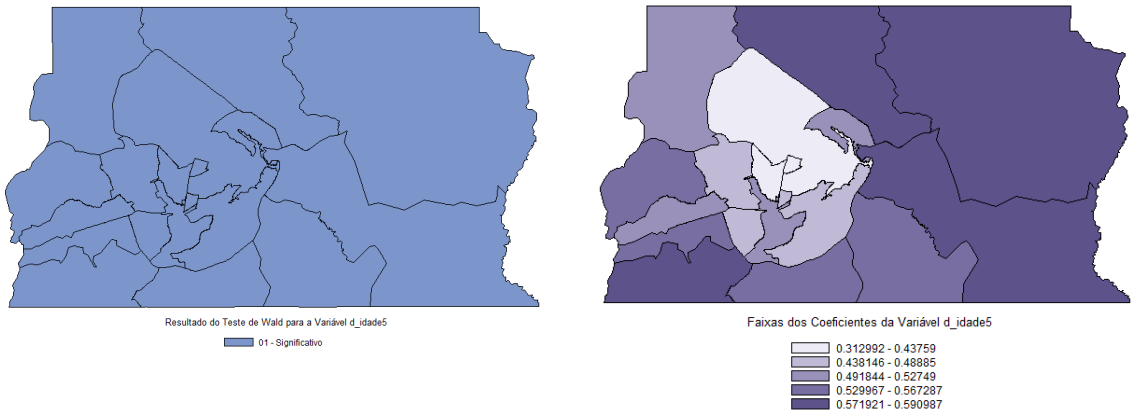

Figura 4.8 – Distribuição espacial da significância e das estimativas da variável d_idade5.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.8 que a variável d_idade5 se mostrou significativa para todas as regiões do Distrito Federal. Observa-se que os valores do coeficiente variaram de 0,313 a 0,590, o que significa que essa variável aumenta a probabilidade de inadimplência do tomador enquadrado nessa faixa de idade em todas as regiões.

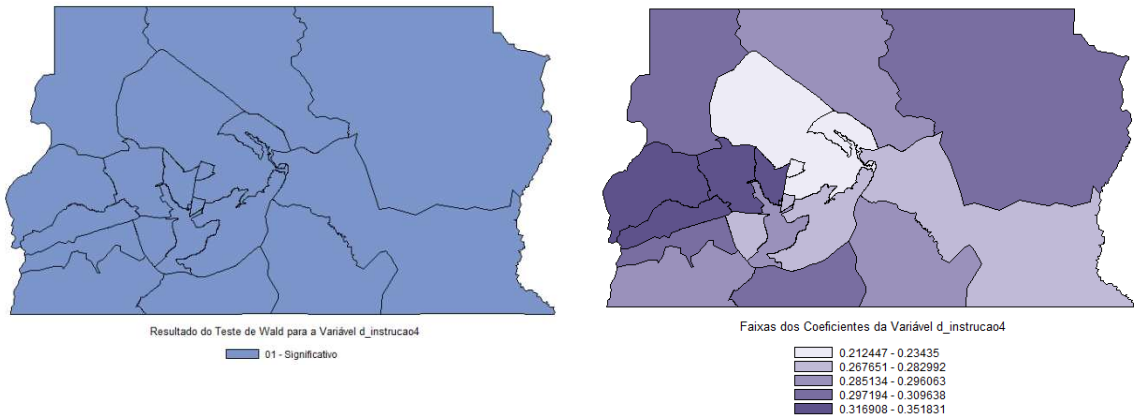

Figura 4.9– Distribuição espacial da significância e das estimativas da variável d_instrução4.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.9 que a variável d_instrução4 também se mostrou significativa para todas as regiões do Distrito Federal, cuja região que mais influencia o aumento do risco de crédito é o Ceilândia (0,351) e a que possui a menor influência no aumento do risco de crédito é a região do Lago Norte (0,212). Nota-se a pouca variação dos coeficientes dessa variável dentre as regiões.

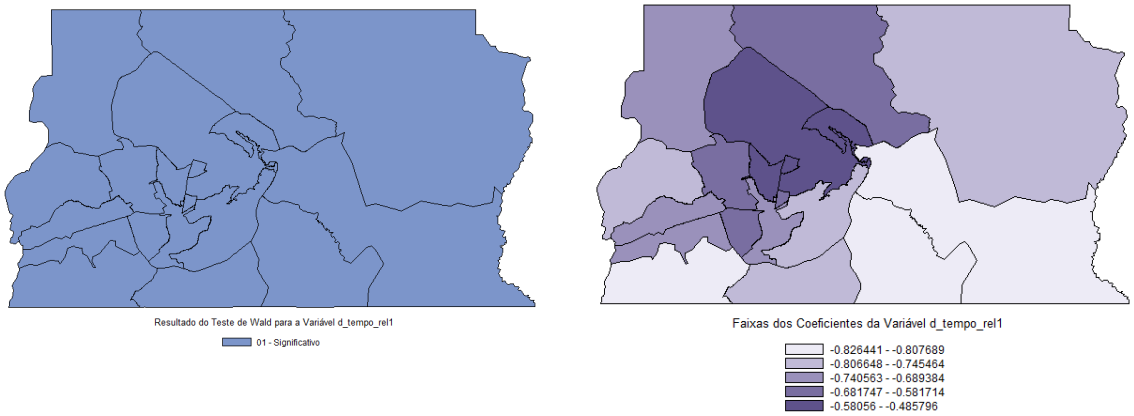

Figura 4.10 - Distribuição espacial da significância e das estimativas da variável d_tempo_rel1.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.10 que a variável d_tempo_rel1 se mostrou significativa para todas as regiões do Distrito Federal. Observa-se que os valores dos coeficientes variaram de -0,826 a -0,485, sendo a região que apresentou o menor valor foi a região do Gama e o maior foi a região do Cruzeiro.

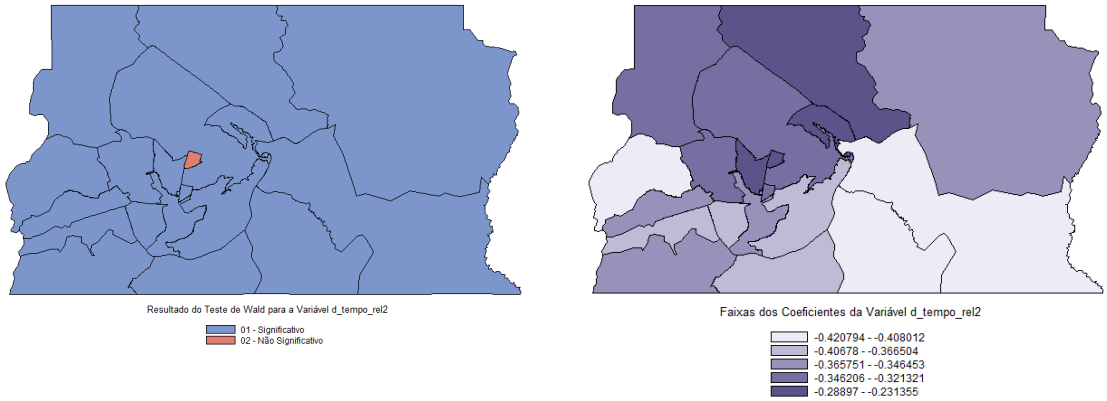

Figura 4.11 – Distribuição espacial da significância e das estimativas da variável d_tempo_rel2.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.11 que a variável d_tempo_rel2 não se mostrou significativa para a região do Cruzeiro. A região de Ceilândia foi a que apresentou o menor valor de coeficiente para essa variável (-0,421), sendo que para a região do Guará essa variável apresentou o maior coeficiente dentre os significativos (-0,265). Note também que houve pequena variação dentre os coeficientes dessa variavel.

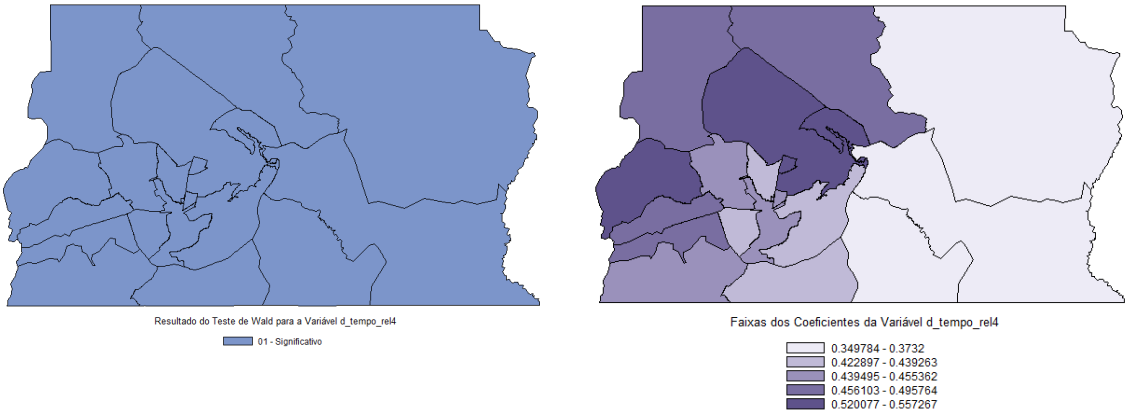

Figura 4.12 – Distribuição espacial da significância e das estimativas da variável d_tempo_rel4.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.12 que a variável d_tempo_rel4 se mostrou significativa para todas as regiões do Distrito Federal. Observa-se que os valores dos coeficientes variaram de 0,350 a 0,557, demonstrando que a presença dessa variável implica em um aumento do risco de crédito do tomador (coeficientes positivos) em todas as regiões do DF.

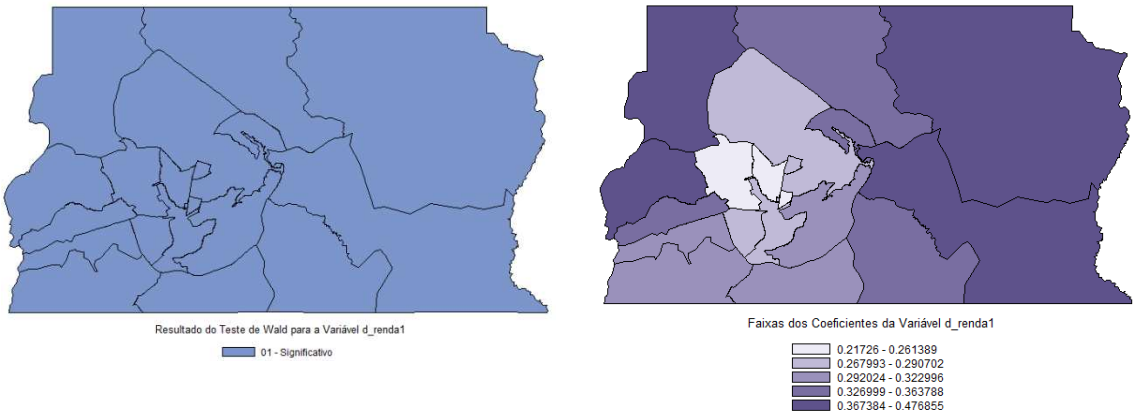

Figura 4.13– Distribuição espacial da significância e das estimativas da variável d_renda1.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.13 que a variável d_renda1 se mostrou significativa para todas as regiões do Distrito Federal. Observa-se que os valores dos coeficientes variaram de 0,217 a 0,477, sendo positivos em todas as regiões do DF e indicando que a presença desse atributo nas características do tomador aumenta seu risco de crédito. Assim como na Regressão global, esses valores de coeficientes foram inesperados, indicando que nova categorização deve ser realizada para essa variável.

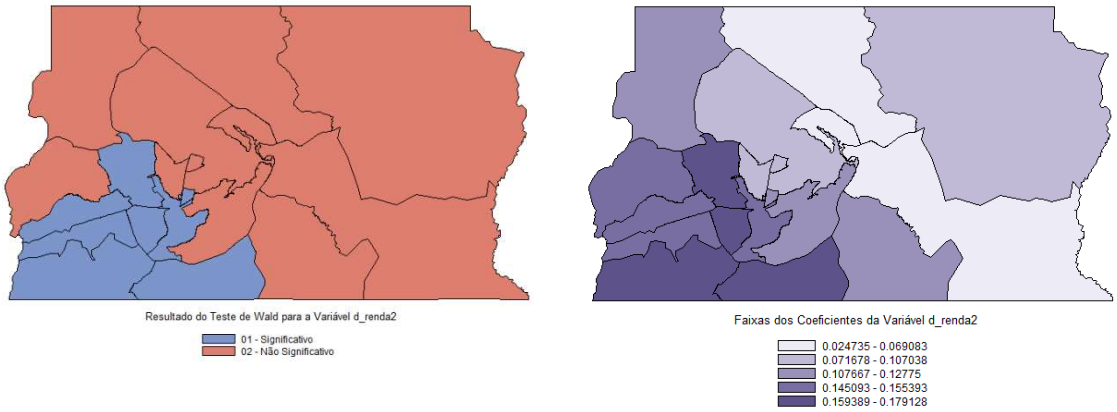

Figura 4.14– Distribuição espacial da significância e das estimativas da variável d_renda2.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.14 que a variável d_renda2 se mostrou significativa somente para as regiões Candangolândia, Gama, Núcleo Bandeirante, Recanto das Emas, Riacho Fundo, Samambaia, Santa Maria e Taguatinga, onde novamente observou-se valores positivos para todos os coeficientes.

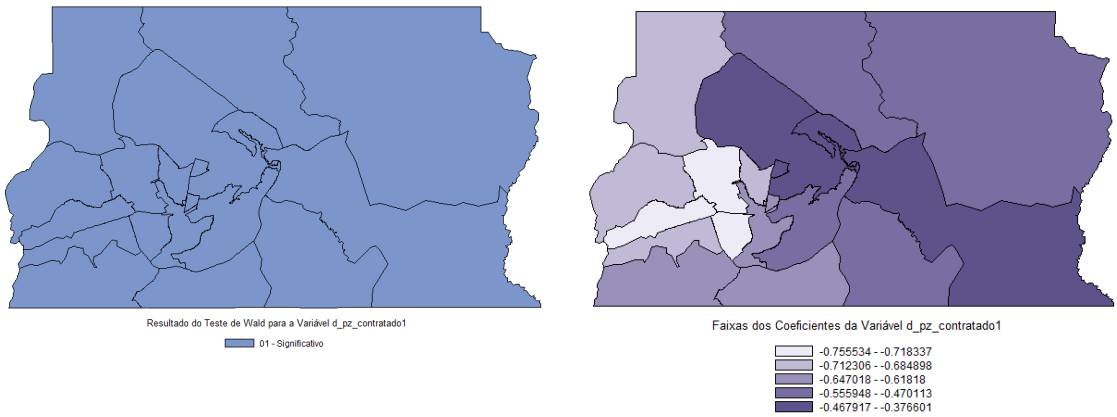

Figura 4.15– Distribuição espacial da significância e das estimativas da variável d_pz_contratação1.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.15 que a variável d_pz_contratação1 se mostrou significativa para todas as regiões do Distrito Federal. Observa-se que os valores dos coeficientes variaram de -0,755 a -0,376, sendo negativos em todas as regiões do DF e indicando que a presença desse atributo nas características do tomador diminui seu risco de crédito.

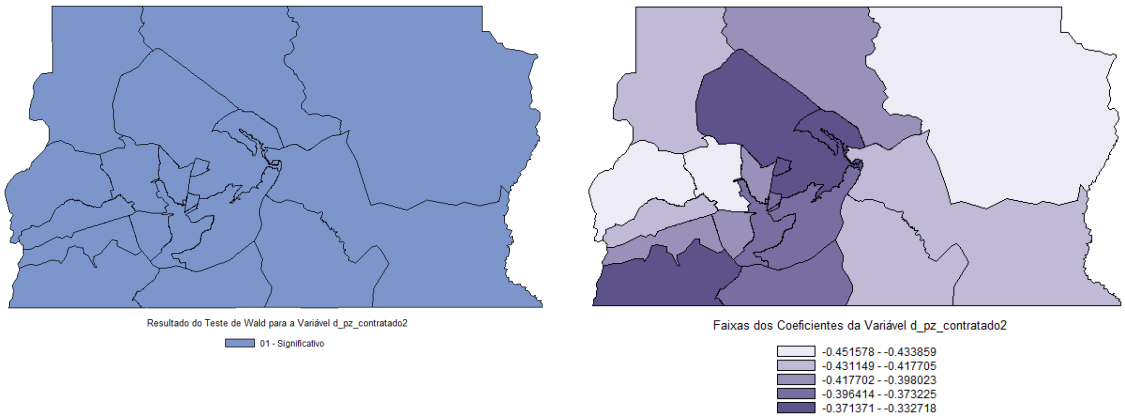

Figura 4.16– Distribuição espacial da significância e das estimativas da variável d_pz_contratação2.
Fonte: elaborado pelo autor.

Nota-se através da Figura 4.16 que a variável d_pz_contratação2 se mostrou significativa para todas as regiões do Distrito Federal. Observa-se que os valores dos coeficientes variaram de -0,451 a -0,332, sendo negativos em todas as regiões do DF e indicando que a presença desse atributo nas características do tomador diminui seu risco de crédito. Nota-se também que a amplitude das estimativas dessa variável foi o menor dentre toas, comum valor de 0,119.

4.6. Comparação Entre os Modelos

A comparação entre o modelo de Regressão Logística Global (LR) e o modelo de GWLR Gaussiano Variável se deu através de cinco métricas: Critério Informacional AICc, Acurácia, Percentual de Falsos Positivos, Somatória do Valor da Dívida dos Falsos Positivos e Valor Monetário Esperado de Inadimplência da carteira frente ao valor monetário de inadimplência observado.

Exceto o critério informacional AICc, calculado no desenvolvimento do modelo, as demais métricas foram calculadas a partir da base de validação, composta por 11.188 registros.

A Tabela 4.24 demonstra as estatísticas descritivas dos escores obtidos por ambos os modelos selecionados aplicados à amostra de validação.

Tabela 4.24 – Analise Descritiva dos Escores dos Modelos.

Modelo	Média	Mínimo	Q1	Mediana (Q2)	Q3	Máximo	Amplitude
RL	0,277	0,036	0,172	0,268	0,392	0,585	0,551
GWLR	0,272	0,035	0,166	0,270	0,378	0,639	0,603

Fonte: elaborado pelo autor.

Note que a média dos escores dos modelos ficaram bem próximas, com diferença apenas na terceira casa decimal, no entanto o modelo via GWLR apresentou amplitude maior de escores. O uso de poucas variáveis preditoras fez com que os escores produzidos pelos modelos não apresentassem valores superiores a 0,585 e 0,639.

Para o cálculo da matriz de confusão, foi necessário definir um ponto de corte, em termos de nota do escore, para então classificar os tomadores em bons ou maus (0 ou 1). Esse ponto de corte foi definido com base na menor distância entre a Sensibilidade e Especificidade exposta na Tabela 4.25, gerada a partir da base de desenvolvimento do modelo global via regressão logística.

Tabela 4.25 – Tabela de Classificação do Modelo Global.

Ponto de Corte (Escore)	Classificação Correta		Classificação Incorreta		Acurácia	Sensibilidade (S)	Especificidade (E)	Diferença (E - S)
	Maus	Bons	Maus	Bons				
0,04	3078	0	7866	0	28,1	100	0	100
0,06	3069	113	7753	9	29,1	99,7	1,4	98,3
0,08	3048	355	7511	30	31,1	99	4,5	94,5
0,10	2990	880	6986	88	35,4	97,1	11,2	85,9
0,12	2951	1203	6663	127	38	95,9	15,3	80,6
0,14	2881	1695	6171	197	41,8	93,6	21,5	72,1
0,16	2828	2101	5765	250	45	91,9	26,7	65,2
0,18	2708	2594	5272	370	48,4	88	33	55
0,2	2618	2912	4954	460	50,5	85,1	37	48,1
0,22	2517	3266	4600	561	52,8	81,8	41,5	40,3
0,24	2307	3700	4166	771	54,9	75	47	28
0,26	2240	4056	3810	838	57,5	72,8	51,6	21,2
0,28	2129	4444	3422	949	60,1	69,2	56,5	12,7
0,30	**1841**	**5152**	**2714**	**1237**	**63,9**	**59,8**	**65,5**	**5,7**
0,32	1737	5400	2466	1341	65,2	56,4	68,6	12,2
0,34	1664	5590	2276	1414	66,3	54,1	71,1	17
0,36	1449	6049	1817	1629	68,5	47,1	76,9	29,8
0,38	1408	6095	1771	1670	68,6	45,7	77,5	31,8
0,4	837	7060	806	2241	72,2	27,2	89,8	62,6
0,42	705	7102	764	2373	71,3	22,9	90,3	67,4
0,44	682	7258	608	2396	72,6	22,2	92,3	70,1
0,46	648	7298	568	2430	72,6	21,1	92,8	71,7
0,48	648	7298	568	2430	72,6	21,1	92,8	71,7
0,50	109	7801	65	2969	72,3	3,5	99,2	95,7
0,52	100	7808	58	2978	72,3	3,2	99,3	96,1
0,54	41	7853	13	3037	72,1	1,3	99,8	98,5
0,56	41	7853	13	3037	72,1	1,3	99,8	98,5
0,58	41	7853	13	3037	72,1	1,3	99,8	98,5
0,60	0	7866	0	3078	71,9	0	100	100

Fonte: elaborado pelo autor.

Diante dos resultados foi definido o valor de 0,30 como ponto de corte para construção das Matrizes de Confusão expostas a seguir.

Tabela 4.26 – Matriz de Confusão do modelo via RL.

		Valor Observado	
		0	1
Valor Predito	0	48,7%	11,3%
	1	24,0%	16,0%

Fonte: elaborado pelo autor.

Tabela 4.27 – Matriz de Confusão do modelo via GWLR.

		Valor Observado	
		0	1
Valor Predito	0	49,0%	11,2%
	1	23,8%	16,0%

Fonte: elaborado pelo autor.

Nota-se pelas Tabelas 4.26 e 4.27 que os modelos apresentaram resultados bem próximos quanto à classificação dos clientes.

A Tabela 4.28 contém todas as métricas utilizadas para comparação entre os modelos.

Tabela 4.28 – Comparação entre os modelos RL e GWRL.

Modelo	**AICc**	**Acurácia**	**% FP**	**Soma do Valor Dívida FP**	**Valor Esperado Inadimplência**
RL	12.098,29	64,7%	11,3%	R$ 5.271.027,78	R$ 11.909.313,79
GWLR	12.091,19	65%	11,2%	R$ 5.484.464,08	R$ 11.611.161,58

Fonte: elaborado pelo autor.

Nota-se através da Tabela 4.28 que todos os valores obtidos para as métricas dos dois modelos também ficaram muito próximas, sendo que o modelo via GWLR foi o modelo que apresentou o melhor (menor) critério informacional AICc, melhor (maior) acurácia, que indica um melhor percentual de acertos e menor percentual de falsos positivos, enquanto o modelo via LR foi levemente superior nas métricas Soma do valor dos Falsos Positivos, sendo que essa métrica pode ser considerada uma estimativa do valor monetário que seria concedido e entraria em inadimplência, resultando em perda financeira para a instituição e Valor Esperado de Inadimplência, uma vez que a somatória

do valor da dívida de todos os contratos inadimplentes (Y=1) da base de validação do modelo foi de R$ 12.026.290,09 e o valor que mais se aproxima é o valor do modelo via RL.

5. CONCLUSÃO

Nessa dissertação foram utilizados dados reais, referentes à operação de Crédito Direto ao Consumidor de uma instituição financeira pública nacional concedidas a clientes domiciliados no Distrito Federal, para o desenvolvimento de modelos de *credit scoring* através de duas metodologias distintas: Regressão Logística (RL) e Regressão Logística Geograficamente Ponderada (GWLR).

A metodologia Regressão Logística (RL) é bastante difundida no setor financeiro, sendo utilizada nessa dissertação para desenvolvimento de um modelo global de *credit scoring* para todo o Distrito Federal.

A metodologia Regressão Logística Geograficamente Ponderada (GWLR) é pouco difundida no setor financeiro e utiliza a localização geográfica do tomador de crédito para ponderar as observações no desenvolvimento de modelos distintos par cada região de estudo. Nesse estudo essa metodologia foi utilizada para desenvolver um modelo de *credit scoring* distinto para cada uma das 19 regiões do Distrito Federal.

Através dos resultados observados pode-se concluir que a técnica GWLR é viável para ser aplicada no desenvolvimento de modelos de *credit scoring*.

Os indicadores utilizados para comparação entre os modelos desenvolvidos através das duas metodologias se mostraram bem próximos entre si e, apesar do modelo desenvolvido via GWLR ter superado o modelo via RL em 3 das 5 métricas, pode-se considerar que as metodologias obtiveram um empate técnico em termos de capacidade de previsão e perdas financeiras para a instituição.

Esse estudo demonstrou que algumas variáveis foram significativas para todas as regiões, enquanto outras se mostraram significativas somente para determinadas regiões, concluindo que o risco de crédito é influenciado por diferentes variáveis a depender da região em estudo.

Observou-se também que todos os modelos de regressão desenvolvidos pela GWLR (modelos regionais) apresentaram valores distintos para os coeficientes (parâmetros) das variáveis, demonstrando que o peso (importância) das variáveis também varia de região para região.

As variáveis macroeconômicas utilizadas nesse estudo não se mostraram significativas para o público alvo estudado.

Devido ao grande avanço computacional e tecnológico ocorrido nas últimas décadas, as instituições concessoras de crédito possuem sistemas robustos de avaliação de risco de crédito, o que viabiliza a implementação e utilização de um conjunto de modelos estimados via GWLR.

5.1. Limitações

Os modelos desenvolvidos nessa dissertação são aplicáveis a somente aos clientes solicitantes da operação de crédito CDC e domiciliados no Distrito Federal da instituição financeira em questão, uma vez que esse foi o público utilizado no desenvolvimento dos modelos.

Para expandir o uso desses modelos a outras operações de crédito ou ainda serem utilizado por outras instituições financeiras é necessário realizar testes de aderência para os públicos alvo desejados.

A ausência de variáveis macroeconômicas segregadas para as regiões do DF foi um limitador para a utilização de mais variáveis dessa natureza, como, por exemplo, a renda per capita ou o PIB municipal.

O uso de poucas variáveis preditoras no estudo fez com que os modelos apresentassem baixas amplitudes de escores.

A categorização da variável Renda Formal foi realizada para que as classes ficassem monotônicas com relação ao risco relativo, no entanto, os valores dos seus coeficientes se mostraram invertidos. Estudos considerando outro público alvo devem ser realizados para demostrar a relevância dessa variável.

5.2. Trabalhos Futuros

Conforme já relatado, o público alvo dessa dissertação foi composto por tomadores de crédito da operação CDC e domiciliados no Distrito Federal, foi utilizado um conjunto pequeno de variáveis preditoras e foram utilizadas as metodologias GWLR e RL para desenvolvimento de modelos de *credit scoring*. Diante do exposto, seguem algumas sugestões para trabalhos futuros:

1. Aplicar a metodologia GWLR para desenvolver modelos de *credit scoring* para outros públicos alvo (diferentes operações de crédito ou regiões geográficas) e compara-los com a Regressão Logística;
2. Aplicar a metodologia GWLR para desenvolver modelos de *credit scoring* e compara-la frente a outras metodologias (por exemplo *Support Vector Machines* ou *Boosting*);
3. Utilizar outras variáveis preditoras (por exemplo os 2 ou 3 dígitos iniciais do CEP ou o PIB Municipal) para desenvolver modelos de *credit scoring* através da metodologia GWLR e verificar se seu incremento melhora a predição dos modelos;
4. Aplicar a metodologia GWLR para o desenvolvimento de modelos em outras áreas de uma instituição financeira, como por exemplo em áreas de estratégia e marketing.
5. Utilizar outras funções, como por exemplo a função Log Binomial, para desenvolver modelos geograficamente ponderados.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. **Categorical data analysis**. New York: John Wiley, 1990.

ALTMAN, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. **The Journal of Finance**, v. 23, n. 4, p. 589-609, 1968.

ALTMAN, E. I.; HADELMANN, R.; NARAYANAN, P. ZETA analysis: a new model to identify bankruptcy risk of corporations. **Journal of Banking and Finance**. p. 470-492, 1977.

ALTMAN, E. I. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian Experience). **Journal of Banking & Finance**, 18(3), 505–529, 1994.

ANDERSON, R. **The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation: Theory and Practice for Retail Credit Risk Management and Decision Automation.** OUP Oxford, 2007.

ANDRADE, F. W. M. Desenvolvimento de Modelo de Risco de Portfolio para Carteiras de Crédito a Pessoa Física. Tese de Doutorado apresentada ao Curso de Doutorado em Administração de Empresas da EAESP/FGV, Área de Concentração: Controle, Finanças e Contabilidade. São Paulo: EAESP/FGV, 2004. 196 p. Disponível em < http://bibliotecadigital.fgv.br/dspace/handle/10438/2513>. Acesso em: 27/10/2015.

ANDRADE, F. W. M.; THOMAS, L. C. Structural models in consumer credit. **European Journal of Operational Research**, v. 183, n. 3, p. 1569-1581, 2007.

ANSELIN, L. Local Indicators of Spatial Association – LISA. **Geographical Analysis**, 27(2):93-115, 1995.

ANTÃO, P.; LACERDA, A. Capital requirements under the credit risk-based framework. **Journal of Banking & Finance**, v. 35, n. 6, p. 1380-1390, 2011.

ATKINSON, P. M.; GERMAN, S. E.; SEAR, D. A.; CLARK, M. J. Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. **Geographical Analysis**, v. 35, n. 1, p. 58-82, 2003.

BAESENS, B.; VAN GESTEL, T.; VIAENE, S.; STEPANOVA, M.; SUYKENS, J.; VANTHIENEN, J. Benchmarking state-of-the-art classification algorithms for credit scoring. **Journal of the Operational Research Society**, v. 54, n. 6, p. 627-635, 2003.

BACEN - BANCO CENTRAL DO BRASIL. Resolução CMN nº 2.099 de 17/08/1994.

BACEN - BANCO CENTRAL DO BRASIL. Circular nº 2.784, de 27/11/1997.

BACEN - BANCO CENTRAL DO BRASIL. Resolução CMN nº 3.380, de 29/06/2006.

BACEN - BANCO CENTRAL DO BRASIL. Resolução CMN nº 3.464, de 26/06/2007.

BACEN - BANCO CENTRAL DO BRASIL. Resolução CMN nº 3.721, de 30/04/2009.

BACEN - BANCO CENTRAL DO BRASIL. Circular nº 3.648, de 04/03/2013.

BARCO, M. Credit portfolio risk Bringing credit portfolio modelling to maturity. **Risk-London-Risk Magazine Limited**, v. 17, n. 1, p. 86-90, 2004.

BARTH, N. L. **Inadimplência: construção de modelos de previsão**. São Paulo: Nobel, 2004. 98p.

BCBS - BASEL COMMITTEE ON BANKING SUPERVISION. International Convergence of Capital Measurement and Capital Standards. **Bank for International Settlements**, 1988. Disponível em < http://www.bis.org/publ/bcbs04a.htm>. Acesso em: 03/10/2015.

BCBS - BASLE COMMITTEE ON BANKING SUPERVISION. Core Principles for Effective Banking Supervision. **Bank for International Settlements**, 1997. Disponível em: <http://www.bis.org/publ/bcbs30a.htm >. Acesso em: 03/10/2015.

BCBS - BASEL COMMITTEE ON BANKING SUPERVISION. International Convergence of Capital Measurement and Capital Standards: A Revised Framework. **Bank for International Settlements**, 2004. Disponível em <http://www.bis.org/publ/bcbs107.htm>. Acesso em: 03/10/2015.

BCBS - BASEL COMMITTEE ON BANKING SUPERVISION. Proposed Enhancements to the Basel II Framework. **Bank for International Settlements**, 2009. Disponível em <http://www.bis.org/publ/bcbs150.htm>. Acesso em: 03/10/2015.

BCBS - BASEL COMMITTEE ON BANKING SUPERVISION. Basel III: A global regulatory framework for more resilient banks and banking systems - revised version. **Bank for International Settlements**, 2011. Disponível em <http://www.bis.org/publ/bcbs189.htm>. Acesso em: 03/10/2015.

BELLOTTI, T.; CROOK, J. Credit scoring with macroeconomic variables using survival analysis. **Journal of the Operational Research Society**, v. 60, n. 12, p. 1699-1707, 2009.

BENSIC, M.; SARLIJA, N.; ZEKIC-SUSAC, M. Modelling small-business credit

scoring by using logistic regression, neural networks and decision trees**. Intelligent Systems in Accounting, Finance and Management**, v. 13, n. 3, p. 133-150, 2005.

BIELECKI, T. R.; RUTKOWSKI, M. **Credit risk: modeling, valuation and hedging.** Springer Science & Business Media, 2002.

BORDO, M. D. The Bretton Woods international monetary system: a historical overview. In: **A retrospective on the Bretton Woods system: Lessons for international monetary reform**. University of Chicago Press, p. 3-108, 1993.

BORGES, O. Rating de Crédito: Considerações sobre os modelos. **Revista Tecnologia de Crédito – Serasa**, volume 24, 14-27, 2001.

BREIMAN, L. Bagging Predictors**. Machine Learning**, v. 26, 123-140, 1996.

BROOKS, C. **Introductory econometrics for finance**. 2nd ed. New York: Cambridge University Press, 2008. 648 p.

BRUSDON, C.; FOTHERINGHAM, A. S.; CHARLTON, M. Geographically weighted regression: a method for exploring spatial nonstationarity. **Geographical Analysis**, 28(4): 281-298, 1996.

CALABRESE, R. Downturn loss given default: Mixture distribution estimation. **European Journal of Operational Research**, 237 (2014), pp. 271–277, 2014.

CAOUETTE, J. B.; ALTMAN, E. I.; NARAYANAN, P.; NIMMO, R. **Managing Credit Risk: The Great Challenge for the Global Financial Markets**. 2nd ed. New Jersey: John Wiley & Sons Inc. 2008. 627p.

CASELLA, G.; BERGER, R. L. **Inferência Estatística**. 2ª Edição. Cengage Learning, 2010. 588p.

CHAVEZ-DEMOULIN, V.; EMBRECHTS, P.; NEŠLEHOVÁ, J. Quantitative models for operational risk: extremes, dependence and aggregation. **Journal of Banking & Finance**, v. 30, n. 10, p. 2635-2658, 2006.

CHEN, J. M., Measuring Market Risk Under the Basel Accords: VaR, Stressed VaR, and Expected Shortfall A estimation. **The IEB International Journal of Finance**, v. 8, pp.

184-201, 2014.

CROUHY, M., GALAI, D., MARK, R. A comparative Analysis of Current Credit Risk Models. **Journal of Banking and Finance**, v. 24, p. 59-117, 2000.

CSFP - CREDIT SUISSE FINANCIAL PRODUCTS. **CreditRisk+: A Credit Risk Management Framework** – Credit Suisse Financial Products, 1997.

DESAI, V. S.; CROOK, J. N.; OVERSTREET JR, G. A. A comparison of neural networks and linear scoring models in the credit union environment. **European Journal of Operational Research**, 95(1), 24–37, 1996.

DESAI, V. S.; CONWAY, D. G.; CROOK, J. N.; OVERSTREET JR, G. A. Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. **IMA Journal of Management Mathematics**, 8(4), 323-346, 1997.

DOWD, K. **Measuring market risk**. John Wiley & Sons, 2007.

DUARTE JÚNIOR, A. M. **Riscos: definições, tipos, medição e recomendações para seu gerenciamento em gestão de riscos e derivativos**. São Paulo: Atlas, 2001.

DUARTE JÚNIOR, A. M.; LELIS, R. J. F. Unificando a alocação de capital em bancos e seguradoras no Brasil. RAE - **Revista de Administração de Empresas**, v. 44, n. 2, 2004.

DUARTE JÚNIOR, A. M. **Gestão de Riscos para Fundos de Investimentos**. São Paulo: Prentice Hall, 2005.

DUFFIE, D.; SINGLETON, K. J. Modeling Term Structures of Defaultable Bonds. **Review of Financial Studies**, 12, pp.687-720, 1999.

DURAND, D. Risk elements in consumer instalment financing. (Technical edition) By David Durand. **National bureau of economic research**. New York, 1941.

EICHENGREEN, B. História e Reforma do Sistema Monetário Internacional. **Economia e Sociedade**. Campinas (SP), n. 4, p.53-78, 1995.

FERNANDES, G. B.; ARTES, R. Spatial dependence in credit risk and its improvement in credit scoring. **European Journal of Operational Research**, 2015.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of eugenics**, v. 7, n. 2, p. 179-188, 1936.

FOTHERINGHAM, A. S.; BRUSDON, C.; CHARLTON, M. Geographically Weighted Regression, John Wiley & Sons Ltd, England,, 2002.

FOTHERINGHAM, A. S.; BRUNSDON, C.; CHARLTON, M. Geographically Weighted Regression – the analysis of spatially varying relationships. John Wiley & Sons Ltd, England, 2006.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of computer and system sciences**, v. 55, n. 1, p. 119-139, 1997.

FRYDMAN, H.; SCHUERMANN, T. Credit rating dynamics and Markov mixture models. **Journal of Banking & Finance**, v. 32, n. 6, p. 1062-1075, 2008.

GILBERT, A.; CHAKRABORTY, J. Using geographically weighted regression for environmental justice analysis: Cumulative cancer risks from air toxics in Florida. **Social Science Research**, v. 40, n. 1, p. 273-286, 2011.

GIRARDI, G.; ERGÜN, A. T. Systemic risk measurement: Multivariate GARCH estimation of CoVaR. **Journal of Banking & Finance**, v. 37, n. 8, p. 3169-3180, 2013.

GITMAN, L. **Princípios da administração financeira**. 7ª ed. São Paulo: Harbra, 1997.

GOODHART, C., Financial regulation, credit risk and financial stability. **National Institute Economic Review** 192 - 1, 118 – 127, 2005.

GOODHART, C., Liquidity risk management. **Financial Stability Review**, issue 11, pages 39-44, 2008.

GORDY, M. B. A **Comparative Anatomy of Credit Risk Models**. Working Paper. Board of Governors of the Federal Reserve System, 1998.

GUPTON, G. M.; FINGER, C. C.; BHATIA, M. **CreditMetrics: technical document.** JP Morgan & Co., 1997.

HAIR JR., J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. Análise

Multivariada de Dados. 6ª ed. Porto Alegre: Bookman, 2009.

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. **Journal of the Royal Statistical Society**. Series A (Statistics in Society), p. 523-541, 1997.

HÄRDLE, W. K.; MOROB, A. R.; SCHÄFER, D. **Estimating Probabilities of Default with Support Vector Machines**. Discussion Paper, Series 2: Banking and Financial Studies Nº 18/2007. Deutsche Bundesbank, 2007, 32p.

HARON, M. S.; RAMLI, R.; INJAS, M. M. Y.; INJAS, R. A. Reputation Risk and Its Impact on the Islamic Banks: Case of the Murabaha**. International Journal of Economics and Financial Issues**, v. 5, n. 4, p. 854-859, 2015.

HARRIS, Terry. Credit scoring using the clustered support vector machine. **Expert Systems with Applications**, v. 42, n. 2, p. 741-750, 2015.

HOPPER, M. A.; LEWIS, E. M. **Behaviour scoring and adaptive control systems**. In: Thomas, L. C., Crook, J. N., & Edelman, D. B. (Eds.), Credit scoring and credit control, Oxford University Press, Oxford, pp. 257–276, 1992.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. 2nd ed. New York: John Wiley & Sons, 2000.

HUANG, Y.; LEUNG, Y. Analysing regional industrialisation in Jiangsu province using geographically weighted regression. **Journal of Geographical Systems**, v.4, n. 2, p. 233-249, 2002.

HURD, T.; KUZNETSOV, A. Affine Markov chain models of multifirm credit migration. **Journal of Credit Risk**, v. 3, n. 1, p. 3-29, 2007.

HURVICH, C. M.; SIMONOFF, J. S.; TSAI, C. L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. **Journal of the Royal Statistical Society**: Series B (Statistical Methodology), v. 60, n. 2, p. 271-293, 1998.

IPEA. Instituto de Pesquisa Econômica Aplicada. Situação Social nos Estados: o caso do Distrito Federal. Brasília: Ipea, 2011, 67p.

JACOBS, M. An empirical study of exposure at default. **Journal of Advanced Studies in Finance (JASF)**, n. 1, p. 31-59, 2010.

JORION, P. **Value at Risk: a nova fonte de referência para a gestão do risco financeiro**. São Paulo: BM&FBOVESPA: Bolsa de Valores, Mercadorias e Futuros, 2010. 487p.

KMV Corporation. **Modeling Default Risk**. KMV Corporation, San Francisco, 1993.

KHANDANI, Amir E.; KIM, Adlar J.; LO, Andrew W. Consumer credit-risk models via machine-learning algorithms. **Journal of Banking & Finance**, v. 34, n. 11, p. 2767-2787, 2010.

KUMAR, K.; BHATTACHARYA, S. Artificial neural network vs linear discriminant analysis in credit ratings forecast: A comparative study of prediction performances. **Review of Accounting and Finance**, v. 5, n. 3, p. 216-227, 2006.

LESSMANN, S.; BAESENS, B.; SEOW, H. V.; THOMAS, L. C. Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. **European Journal of Operational Research**, 247, p.124–136, 2015.

MAKUCH, W. M. The basics of a better application score. **Handbook of Credit Scoring**, p. 127-48, 2001.

MARKOWITZ, H. Portfolio selection. **The Journal of Finance**, v. 7, n. 1, p. 77-91, 1952.

MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. 2nd ed. London: Chapman and Hall, 1989. 511p.

MEDEMA, L.; KONING, R. H.; LENSINK, R. A practical approach to validating a PD model. **Journal of Banking & Finance**, v. 33, n. 4, p. 701-708, 2009.

MERTON, R. C. On the Pricing of Corporate Debt: The Risl Structure of Interest Rates. **Journal of Finance**, 29, pp. 449-470, 1974.

MORAES, D. Modelagem de fraude em cartão de crédito. Dissertação de Mestrado, Universidade Federal de São Carlos - Departamento de Estatística, São Carlos-SP, 2008,

120 f.

MORAN, P. A. P. Notes on Continuous Stochastic Phenomena. **Biometrika** 37 (1): 17–23, 1950.

MOSCADELLI, M. The modelling of operational risk: experience with the analysis of the data collected by the Basel Committee. **Available at SSRN 557214**, 2004.

MYERS, J. H.; FORGY, E. W. The development of numerical credit evaluation systems. **Journal of the American Statistical Association**, v. 58, n. 303, p. 799-806, 1963.

OHLSON, J. A. Financial ratios and the probabilistic predictions of bankruptcy. **Journal of Accounting Research**, v. 18, n. 1, p. 109-131, Spring 1980.

ONG, C.; HUANG, J.; TZENG, G. Building credit scoring models using genetic programming. **Expert Systems with Applications**, v. 29, n. 1, p. 41-47, 2005.

OPTIZ, D.; MACLIN, R. Popular Ensemble Methods: An Empirical Study**. Journal of Artificial Intelligence Research**, 11, 169-198, 1999.

PETROV, D.; POMAZANOV, M. Validation method of maturity adjustment formula for Basel II capital requirement. **The Journal of Risk Model Validation**, v. 3, n. 3, p. 81-97, 2009.

RESTI, A.; SIRONI, A. **Gestão do risco na atividade bancária e geração de valor para o acionista: modelos de medição de risco a políticas de alocação de capital**. 1ª ed. Rio de Janeiro: Qualitymark, 2010.992 p.

RODRÍGUEZ-MORENO, M.; PEÑA, J. I. Systemic risk measures: The simpler the better? **Journal of Banking & Finance**, v. 37, n. 6, p. 1817-1831, 2013.

SAUNDERS, A. Medindo o Risco de Crédito – novas abordagens para value at risk e outros paradigmas. Rio de Janeiro: Qualitymark, 2000. 200p.

SCHRICKEL, W. K. **Analise de Crédito: Concessão e Gerência de Empréstimos**, São Paulo: Atlas, 1995.

SEE, L. et al. Building a hybrid land cover map with crowdsourcing and geographically weighted regression. ISPRS **Journal of Photogrammetry and Remote Sensing**, v. 103,

p. 48-56, 2015.

SICSÚ, A. L. **Credit Scoring: desenvolvimento, implantação, acompanhamento**. São Paulo: Blucher, 2010. 180p.

SILVA, A. R. **Metodologia para Avaliação e Distribuição de Recursos para o Transporte Escolar Rural**. Tese de Doutorado, Publicação T.TD-001A/2009, Departamento de Engenharia Civil e Ambiental, Faculdade de Tecnologia, Universidade de Brasília, DF, 161p., 2009

SILVA J. P. **Gestão e análise de risco de crédito**. 5ª ed. São Paulo: Atlas, 1998.

SILVA A.; MARINS J.; NEVES M. **Loss Given Default: um estudo sobre perdas em operações prefixadas no mercado brasileiro.** Working Paper Series, Banco Central do Brasil, 2009.

SOLTAN, A.; MOHAMMADI, M. A hybrid model using decision tree and neural network for credit scoring problem. **Management Science Letters**, v. 2(5), p. 1683-1688, 2012.

SOUZA, R. B. O modelo de collection scoring como ferramenta para a gestão estratégica do risco de crédito. Dissertação de Mestrado, FGV, São Paulo-SP, 2000.

STEPANOVA, M.; THOMAS, L. C. Survival analysis methods for personal loan data. **Operations Research**, v. 50, n. 2, p. 277-289, 2002.

STINE, R. Spatial temporal models for retail credit. **Proceedings of credit scoring and credit control conference 2011**. Edinburgh, UK, 2011.

THOMAS, L. C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. **International journal of forecasting**, v. 16, n. 2, p. 149-172, 2000.

THOMAS, L.C. Consumer Credit Models: Pricing, Profit, and Portfolios, Oxford University Press, New York, 2009.

THOMAS, L. C. Consumer finance: Challenges for operational research. **Journal of the Operational Research Society**, v. 61, n. 1, p. 41-52, 2010.

TRAVASSOS, A. P. et al. Indicadores de microcrédito baseados em energia elétrica: inovação e sustentabilidade na concessão de crédito e no risco de inadimplência. **Cad. CPqD Tecnologia**, v. 9, n. 2, p. 121-130, 2013.

TSAI, H. T. YEH. H. C A two-stage screening procedure for mailing credit assessment. IMA **Journal of Mathematics Applied in Business and Industry**, 10, 317-329, 1999.

VALVONIS, V. Estimating EAD for retail exposures for Basel II purposes. **Journal of Credit Risk**, v. 4, n. 1, p. 79-101, 2008.

VAN GOOL, J., VERBEKE, W., SERCU, P., BAESENS, B. Credit scoring for microfinance: is it worth it? **International Journal of Finance & Economics**, v. 17, n. 2, p. 103-123, 2012.

VOLK, M. Estimating probability of default and comparing it to credit rating classification by banks. **Economic and Business Review**, v. 14, n. 4, 2012.

WAGSTER, J. D. Impact of the 1988 Basle Accord on International Banks**. Journal of Finance**, vol. 51, no. 4, pp. 1321-1346, 1996.

WANG, Y.; WANG, S.; LAIET, K. K. A New Fuzzy Support Vector Machine to Evaluate Credit Risk. **IEEE Transactions on Fuzzy Systems**, Vol. 13, Nº. 6, 2005.

WANG, G.; HAO, J.; MA, J.; JIANG, H. A comparative assessment of ensemble learning for credit scoring. **Expert Systems with Applications**, Elsevier BV, v. 38, n. 1, p. 223–230, Jan 2011.

WEST, D. Neural network credit scoring models. **Computers & Operations Research**, 27(11–12), 1131–1152, 2000.

WIGINTON, J. C. A note on the comparison of logit and discriminant models of consumer credit behavior. **Journal of Financial and Quantitative Analysis**, v. 15, n. 03, p. 757-770, 1980.

WILSON, T. C. Measuring and Managing Credit Risk Portfolio: Part I: Modelling Systemic Default Risk. Risk, 10 (9), September 1997.

WILSON, T. C. Measuring and Managing Credit Risk Portfolio: Part II: Portfolio Loss Distributions. Risk, 10 (10), October 1997.

XIA, G; JIN, W. Model of consumer churn prediction on support vector machine. **Systems Engineering – Theory and Practice**, v. 28, n.1, p. 71-77, 2008.

YAO, X.; CROOK J.; ANDREEVA, G. Support vector regression for loss given default modelling. **European Journal of Operational Research**, 240, pp. 528–538, 2015.