ISSN 1808-057X
DOI: 10.1590/1808-057x201703760

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

Geographically Weighted Logistic Regression Applied to Credit Scoring Models

Pedro Henrique Melo Albuquerque
Universidade de Brasília, Faculdade de Economia, Administração, Contabilidade e Políticas Públicas, Departamento de Administração, Brasília, DF, Brasil

Fabio Augusto Scalet Medina
Universidade de Brasília, Faculdade de Economia, Administração, Contabilidade e Políticas Públicas, Departamento de Administração, Brasília, DF, Brasil

Alan Ricardo da Silva
Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Estatística, Brasília, DF, Brasil

Recebido em 11.05.2016 – *Desk* aceite em 20.06.2016 – 2ª versão aprovada em 11.10.2016

RESUMO

Este estudo utilizou dados reais de uma instituição financeira nacional referentes a operações de Crédito Direto ao Consumidor (CDC), concedidas a clientes domiciliados no Distrito Federal (DF), para a construção de modelos de *credit scoring* utilizando as técnicas Regressão Logística e Regressão Logística Geograficamente Ponderada [*Geographically Weighted Logistic Regression*] (GWLR). Os objetivos foram: verificar se os fatores que influenciam o risco de crédito diferem de acordo com a localização geográfica do tomador; comparar o conjunto de modelos estimados via GWLR frente ao modelo global estimado via Regressão Logística, em termos de capacidade de previsão e perdas financeiras para a instituição; e verificar a viabilidade da utilização da técnica GWLR para desenvolver modelos de *credit scoring*. As métricas utilizadas para comparar os modelos desenvolvidos por meio das duas técnicas 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. Os modelos estimados para cada região do DF se mostraram distintos em suas variáveis e coeficientes (parâmetros), concluindo-se que o risco de crédito foi influenciado de maneira distinta em cada região do estudo. As metodologias Regressão Logística e GWLR apresentaram resultados bem próximos, em termos de capacidade de previsão e perdas financeiras para a instituição, e o estudo demonstrou a viabilidade da utilização da técnica GWLR para desenvolver modelos de *credit scoring* para o público-alvo do estudo.

Palavras-chave: risco de crédito, regressão logística geograficamente ponderada, *credit scoring*.

ABSTRACT

This study used real data from a Brazilian financial institution on transactions involving Consumer Direct Credit (CDC), granted to clients residing in the Distrito Federal (DF), to construct credit scoring models via Logistic Regression and Geographically Weighted Logistic Regression (GWLR) techniques. The aims were: to verify whether the factors that influence credit risk differ according to the borrower's geographic location; to compare the set of models estimated via GWLR with the global model estimated via Logistic Regression, in terms of predictive power and financial losses for the institution; and to verify the viability of using the GWLR technique to develop credit scoring models. The metrics used to compare the models developed via the two techniques were the AICc informational criterion, the accuracy of the models, the percentage of false positives, the sum of the value of false positive debt, and the expected monetary value of portfolio default compared with the monetary value of defaults observed. The models estimated for each region in the DF were distinct in their variables and coefficients (parameters), with it being concluded that credit risk was influenced differently in each region in the study. The Logistic Regression and GWLR methodologies presented very close results, in terms of predictive power and financial losses for the institution, and the study demonstrated viability in using the GWLR technique to develop credit scoring models for the target population in the study.

Keywords: *credit risk, geographically weighted logistic regression, credit scoring.*

*Trabalho apresentado no XL Encontro da ANPAD, Costa do Sauípe, BA, Brasil, setembro de 2016.

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 & Henley, 1997). Por envolver expectativa futura de recebimento, todo crédito concedido está exposto a riscos.

O tema "gerenciamento de riscos" ganhou destaque no setor financeiro após a divulgação dos acordos de Basileia, conjunto de documentos que embasam a regulação e fiscalização do setor. Os avanços tecnológicos e computacionais, aliados ao desenvolvimento de métodos quantitativos, contribuíram para a criação de diversas ferramentas para mensuração de riscos, trazendo ganhos significativos para a gestão financeira das instituições.

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 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 (Banco Central do Brasil [BACEN], 2009), e é um dos principais riscos ao qual uma instituição financeira está exposta.

Os modelos utilizados para mensurar o risco no momento da concessão de crédito são denominados modelos de *credit scoring*. Por envolverem menor custo e dar 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 (Hand & Henley, 1997).

Lessmann, Baesens, Seow e Thomas (2015) realizaram uma abrangente pesquisa sobre as metodologias de classificação utilizadas para o desenvolvimento de modelos de *credit scoring* 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 binária dependente e um conjunto de variáveis preditoras independentes (Hosmer & Lemeshow, 2000).

Uma instituição financeira possui diversos modelos de *credit scoring* que são aplicados na avaliação de diferentes tipos de clientes ou operações de crédito a serem contratadas. As variáveis preditoras que compõem cada modelo podem ser distintas, visando a melhorar a predição do seu público-alvo.

A localização geográfica (espaço) e sua relação com o risco de crédito é tema de alguns estudos publicados. Dentre os mais recentes, Stine (2011) analisa a evolução da inadimplência do crédito imobiliário em condados dos Estados Unidos entre 1993 e 2010, contemplando um período pré-crise e um pós-crise do *subprime,* tendo encontrado evidências da existência de correlação espacial entre as taxas de inadimplência daqueles condados.

Fernandes e Artes (2015) usaram a metodologia *Ordinary Kriging* para criar uma variável que reflete o risco espacial e aplicaram 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 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, German, Sear e Clark (2003) utilizam 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) para cada área estudada, revelando a necessidade de adoção de diferentes práticas para se evitar a erosão, a depender da região.

Este artigo utilizou dados referentes à operação de Crédito Direto ao Consumidor (CDC), concedidos por uma instituição financeira nacional a clientes domiciliados

no Distrito Federal (DF), com os seguintes objetivos: verificar se os fatores que influenciam o risco de crédito diferem de acordo com a localização geográfica do tomador; comparar o conjunto de modelos estimados via GWLR frente ao modelo global estimado via Regressão Logística, em termos de capacidade de previsão e perdas financeiras para a instituição; e verificar a viabilidade da utilização da técnica GWLR para desenvolver modelos de *credit scoring*.

Embora a ideia central deste artigo, de verificar se existe influência do espaço no risco de crédito, seja semelhante à de Stine (2011) e Fernandes e Artes (2015), o público-alvo e a metodologia empregada são distintos, não tendo sido encontrado na literatura estudo que utilize a técnica GWLR na construção de modelos de *credit scoring*.

Uma vantagem da aplicação da técnica GWLR em relação a outras é a estimação de um modelo para cada região do estudo, possibilitando que esses modelos sejam distintos em suas variáveis e parâmetros (Atkinson et al., 2003), enquanto um modelo global, representado apenas por uma fórmula, pode não representar as variações locais de forma adequada. Em relação a crédito, diferentes regiões de estudo podem possuir riscos distintos e, caso esse fenômeno seja observado, modelos que levem em consideração as particularidades locais podem melhor discriminar o risco de crédito dos tomadores ali domiciliados e gerar ganhos financeiros para a instituição.

Mais uma diferença de outros estudos dessa temática e uma vantagem da técnica GWLR é a utilização de amostras distintas no desenvolvimento de cada modelo local, dando um peso maior aos tomadores mais próximos geograficamente, e não utilizando informações distantes que estejam fora do raio delimitado pela função de ponderação.

Questões sobre endogeneidade não são abordadas neste estudo e podem ser levantadas por pesquisadores em trabalhos futuros.

Além desta introdução, a segunda seção do artigo apresenta a metodologia de regressão logística geograficamente ponderada e o processo de desenvolvimento dos modelos, a terceira mostra os resultados obtidos e a quarta expõe a conclusão.

2. METODOLOGIA

O fluxograma apresentado na Figura 1 detalha todas as etapas realizadas no processo de desenvolvimento dos modelos deste estudo.

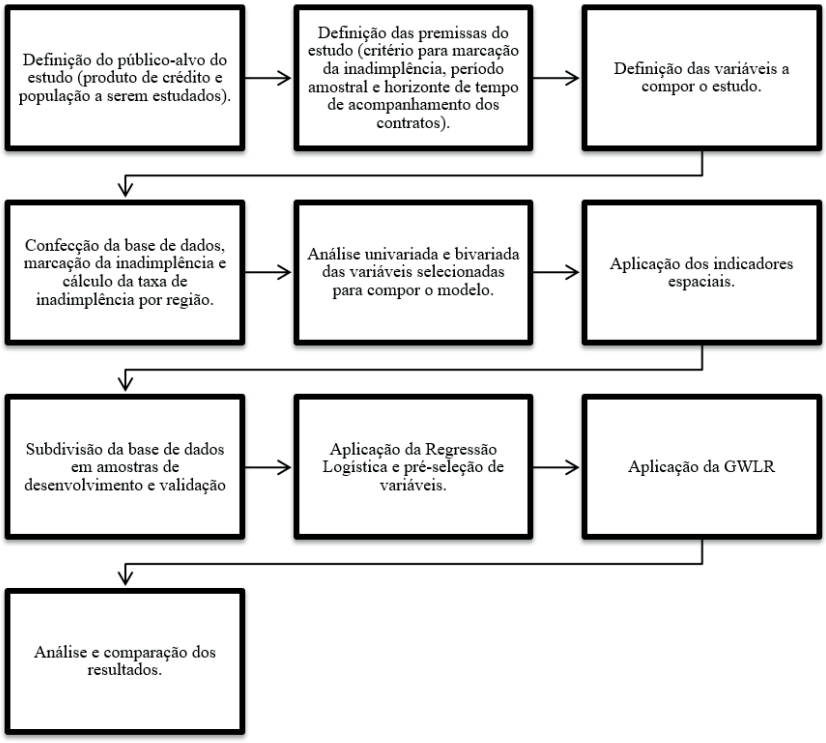

Figura 1. *Fluxograma das etapas de desenvolvimento dos modelos*
Fonte: *Elaborada pelos autores.*

2.1. Base de Dados

Os dados deste estudo referem-se a operações de Crédito Direto ao Consumidor (CDC) concedidas por uma instituição financeira nacional a clientes domiciliados no Distrito Federal (DF). O pagamento dessas operações ocorre de forma parcelada, com prazos de 0 a 36 meses e valor máximo de contratação de R$30.000,00.

A divisão territorial do DF utilizada no estudo foi composta por 19 regiões, expostas na Figura 2.

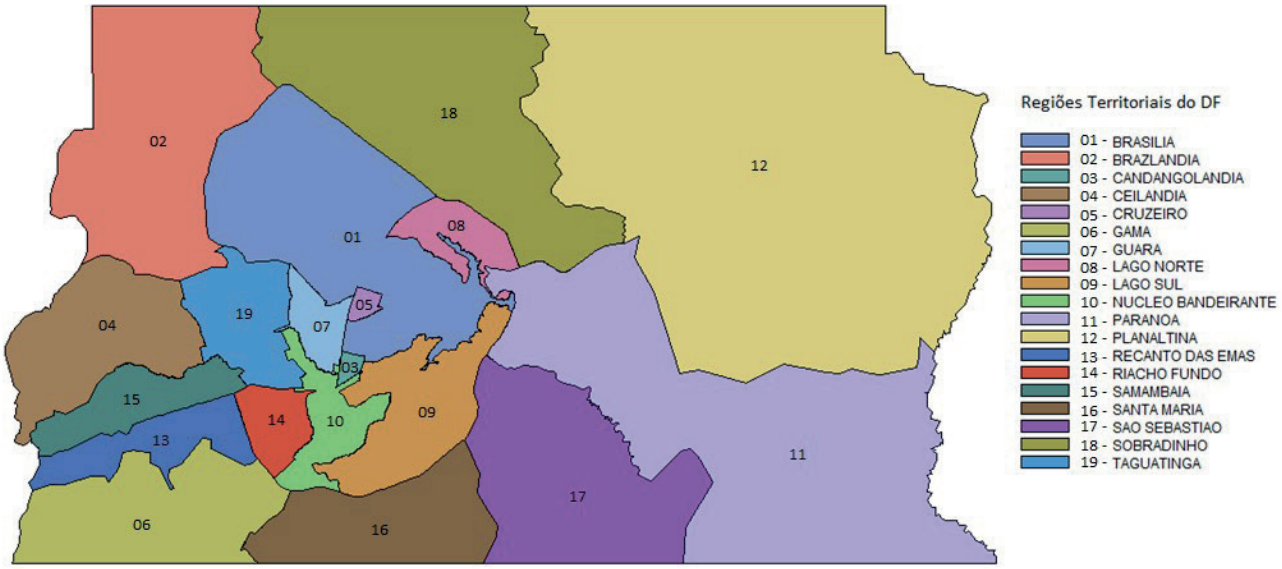

Figura 2. *Divisão territorial do Distrito Federal utilizada no estudo.*
Fonte: *Elaborada pelos autores.*

Foram definidos como amostra todos os contratos concedidos entre os meses de dezembro de 2013 e setembro de 2014, totalizando 10 safras de contratação e um total de 22.132 contratos distintos. O desempenho de pagamento desses contratos foi acompanhado nos doze meses subsequentes à data de contratação e foram marcados como inadimplentes (Y=1) aqueles que ultrapassaram 90 dias em atraso em qualquer um desses meses. 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*).

As variáveis preditoras selecionadas para compor os modelos foram: Idade, Renda, Grau de Instrução, Tempo de Relacionamento do Tomador de Crédito com a Instituição, Prazo Contratado, SELIC, Taxa de Desemprego e Inflação (IPCA). Essas variáveis referem-se ao momento da contratação do crédito (um único ponto no tempo), caracterizando-se como dados do tipo *cross-section*.

As coordenadas geográficas latitude e longitude, referentes às regiões utilizadas no estudo e necessárias para aplicação da técnica GWLR, foram obtidas no site do IBGE e referem-se ao ponto central de cada região, sendo iguais para os tomadores residentes na mesma região.

A base de dados foi subdividida em amostras de desenvolvimento e validação do modelo de acordo com a data de contratação da operação, sendo a amostra de desenvolvimento composta pelas cinco safras iniciais (dezembro de 2013 a abril de 2014), totalizando 10.944 registros. A base de validação é composta pelas cinco safras finais (maio a setembro de 2014), que totalizaram 11.188 registros.

A manipulação dos dados, bem como o cálculo das análises univariadas, bivariadas, indicadores espaciais e o desenvolvimento do modelo global via regressão logística foram realizados por meio do software SAS. Os modelos via GWLR foram desenvolvidos através do software GWR4.

2.2. 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 fórmula é dada por:

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

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 indício de heterogeneidade (não estacionariedade) espacial.

A fórmula do Índice Local de Moran é dada por:

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

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:

$$Taxa\ de\ Inadimplência\ da\ Região = \frac{Quantidade\ de\ clientes\ Inadimplentes\ na\ região}{Quantidade\ total\ de\ clientes\ da\ região} \quad (3)$$

Neste 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 esse público-alvo.

2.3. Regressão Geograficamente Ponderada

De acordo com Fotheringham, Brunsdon e Charlton (2002), dado um modelo de regressão linear básico, a expressão equivalente para a GWR é dada por:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (4)$$

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 para estimação dos parâmetros da GWR é dada por:

$$\widehat{\boldsymbol{\beta}}(i) = (\boldsymbol{X}'\boldsymbol{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(u_i, v_i)\boldsymbol{y}, \quad (5)$$

onde

$$\boldsymbol{W}(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{bmatrix} \quad (6)$$

$\boldsymbol{W}(u_i, v_i)$ 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*. A substituição de todos os pesos w_{ij} pelo valor 1 equivale à matriz identidade, que, substituída em (5), a faz retornar ao modelo clássico de regressão linear.

As duas principais funções de ponderação encontradas na literatura são a função Gaussiana (Normal ou, em inglês, *Gaussian*) e a função Biquadrática (em inglês *Bisquare)*. As fórmulas de ambas as funções estão apresentadas na Tabela 1.

Tabela 1 *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, Brunsdon, and Charlton. (2002).*

Nota-se, pela Tabela 1, que existem dois tipos de expressões para cada uma das funções Gaussiana e Biquadrática, que se diferenciam 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; por esse motivo, em situações onde os dados não são igualmente distribuídos entre as regiões, é recomendada a utilização do *bandwidth* variável. A Figura 3 ilustra o *bandwidth* em uma função de ponderação e as Figuras 4 e 5 exemplificam o uso do *bandwidth* fixo ou variável.

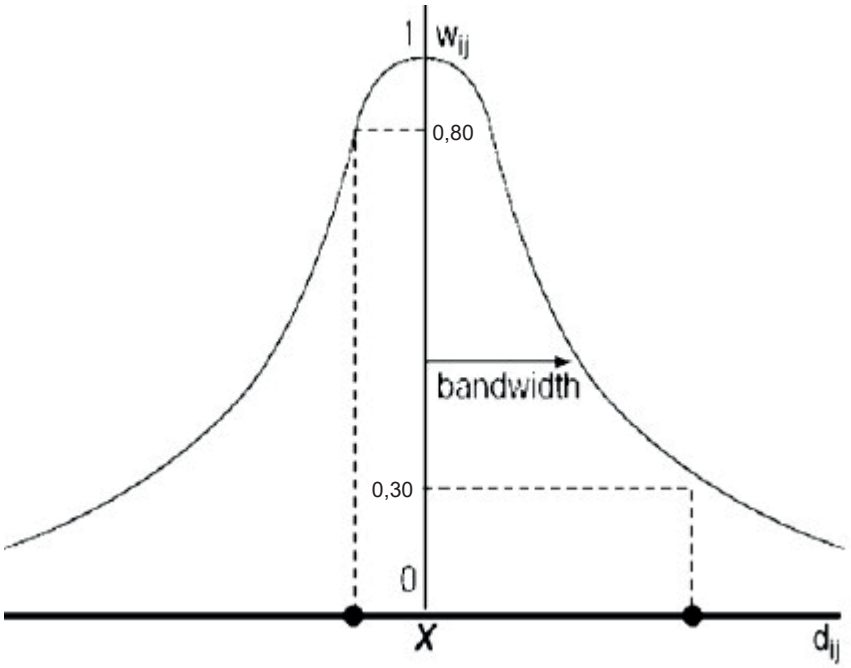

Figura 3. *Bandwidth ou Parâmetro de Suavização.*
Fonte: *Adaptado de Fotheringham et al. (2002).*

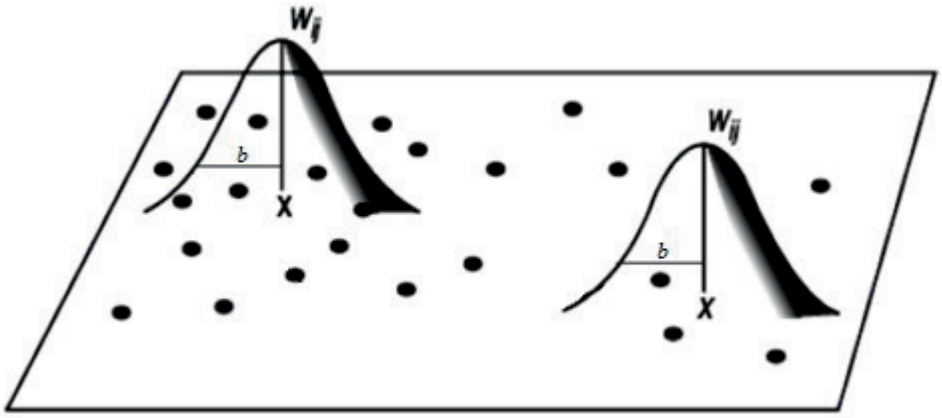

Figura 4. *Funções de ponderação espacial com Bandwidth fixo.*
Fonte: *Adaptado de Fotheringham et al. (2002).*

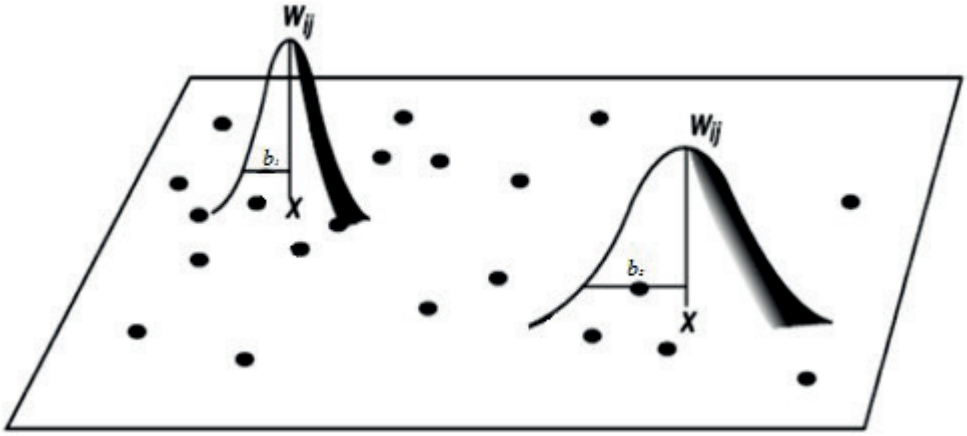

Figura 5. *Funções de ponderação espacial com Bandwidth variável.*
Fonte: *Adaptado de Fotheringham et al. (2002).*

No desenvolvimento de um modelo via GWR utilizando o *bandwidth* fixo, ele 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 ser utilizado nos modelos e, com base nessa quantidade *k*, o valor do *bandwidth* varia entre as regiões do estudo.

2.4. 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 por:

$$\ln\left(\frac{\pi(\boldsymbol{x}_j)}{1-\pi(\boldsymbol{x}_j)}\right) = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{jk} + \varepsilon_i \qquad (7)$$

ou, ainda, na forma:

$$\pi(\boldsymbol{x}_j) = \frac{e^{\beta_0(u_i,v_i)+\sum_{k=1}^{p}\beta_k(u_i,v_i)x_{jk}}}{1+e^{\beta_0(u_i,v_i)+\sum_{k=1}^{p}\beta_k(u_i,v_i)x_{jk}}}, \qquad (8)$$

onde $\pi(\boldsymbol{x}_j)$ é a probabilidade do *j*-ésimo cliente se tornar inadimplente e 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).

A estimação dos parâmetros da GWLR é realizada via método da máxima verossimilhança, sendo a função de verossimilhança da GWLR representada pela seguinte expressão:

$$L(\boldsymbol{\beta}(u_i, v_i)) = \left\{\prod_{j=1}^{n}\left[1+\exp\left(\sum_{k=0}^{p}\beta_k(u_i, v_i)x_{jk}\right)\right]^{-1}\right\}\exp\left[\sum_{k=0}^{p}\left(\sum_{j=1}^{n}y_j x_{jk}\right)\beta_k(u_i, v_i)\right] \qquad (9)$$

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

$$\ln[L(\boldsymbol{\beta}(u_i, v_i))] = \sum_{k=0}^{p}\left(\sum_{j=1}^{n}y_j x_{jk}\right)\beta_k(u_i, v_i) - \sum_{i=1}^{n}\ln\left\{1+\exp\left(\sum_{k=0}^{p}\beta_k(u_i, v_i)x_{jk}\right)\right\} \qquad (10)$$

A matriz $\boldsymbol{W}(u_i, v_i)$ descrita em (6) possui em seus elementos os pesos w_{ij} (calculados através das funções de ponderação expostas na Tabela 1) 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 atribuir um peso maior para as observações mais próximas geograficamente da região *i* na estimação dos seus parâmetros e atribuir 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:

$$\ln[L^*(\boldsymbol{\beta}(u_i, v_i))] = \sum_{k=0}^{p}\left(\sum_{j=1}^{n} w_j(u_i, v_i)\, y_j x_{jk}\right)\beta_k(u_i, v_i) - \sum_{j=1}^{n} w_j(u_i, v_i)\ln\left\{1 + \exp\left(\sum_{k=0}^{p}\beta_k(u_i, v_i)x_{jk}\right)\right\} \quad (11)$$

Similar ao modelo de regressão logística, após diferenciar (11) 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.

Inicialmente foram desenvolvidos quatro modelos distintos utilizando cada uma das funções de ponderação apresentadas na Tabela 1. O melhor modelo com base no AICc foi selecionado para comparação com o modelo global e 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 estimativas dos coeficientes das variáveis.

2.5. Comparação Entre os Modelos

As métricas utilizadas para comparação entre os modelos desenvolvidos via GWLR e Regressão Logística foram: o critério informacional AICc (Hurvich, Simonoff, & Tsai, 1998), 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.

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

Tabela 2 *Matriz de Confusão*

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

Nota. *VP: Verdadeiro Positivo - quantidade de clientes bons classificados como bons; VN: Verdadeiro Negativo - quantidade de clientes maus classificados como maus; FP: Falso Positivo - quantidade de clientes maus classificados como bons; FN: Falso Negativo - quantidade de clientes bons classificados como maus.*
Fonte: *Adaptado de Crook, Edelman e Thomas (2007).*

De acordo com a Tabela 2, 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 este 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. Dessa forma, o percentual de FP foi uma das métricas utilizadas para comparação entre os modelos.

A somatória do saldo devedor de todos os tomadores classificados como FP foi mensurada para verificar o valor monetário que entraria em inadimplência devido ao erro de classificação do modelo.

A acurácia do modelo é calculada pela proporção de VP e VN em relação ao total, conforme a seguinte fórmula:

$$\frac{VP + VN}{VP + VN + FP + FN} \quad (12)$$

O valor monetário esperado de inadimplência da carteira foi calculado por meio da fórmula da esperança das distribuições discretas, dada por:

$$E(X) = \sum_{i=1}^{n} x_i * P(Y_i = 1), \quad (13)$$

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 de o 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.

3. RESULTADOS

3.1. Análises Univariada e Bivariada

Os resultados das taxas de inadimplência geral e por região estão dispostos nas Tabelas 3 e 4 e a distribuição espacial das taxas de inadimplência se encontra na Figura 6.

Tabela 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: *Elaborada pelos autores.*

Tabela 4 *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 SEBASTIÃO	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: *Elaborada pelos autores.*

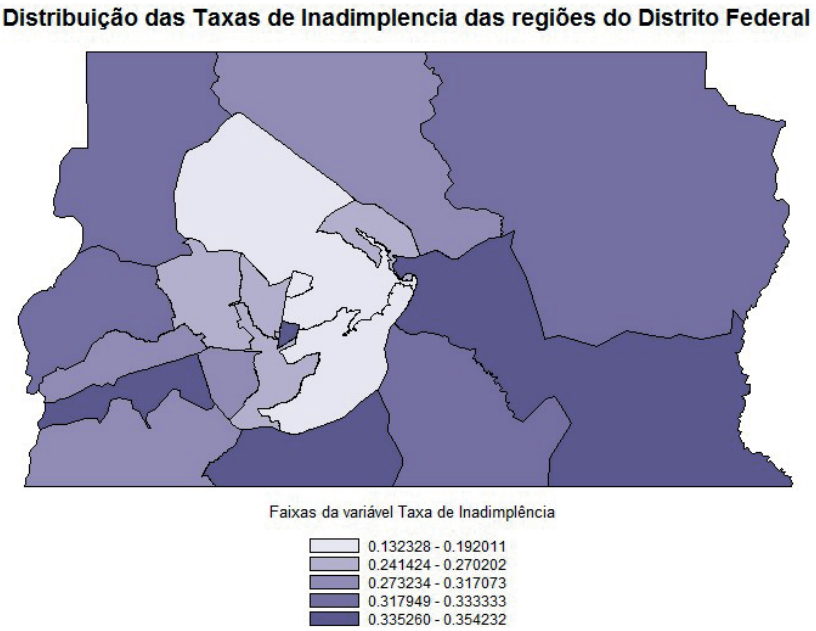

Figura 6. *Distribuição espacial das taxas de inadimplência do Distrito Federal.*
Fonte: *Elaborada pelos autores.*

Conforme exposto na Tabela 3, a taxa de inadimplência geral do DF foi de 27,66%; assim, pode-se observar na Tabela 4 que apenas sete regiões (Lago Sul, Cruzeiro, Brasília, Guará, Lago Norte, Taguatinga e Núcleo Bandeirante) possuem taxas de inadimplência abaixo da média geral. Nota-se tambem que a região do Lago Sul foi a que apresentou a menor taxa de inadimplência entre as regiões estudadas, seguida das regiões Cruzeiro e Brasília. Como pode ser observado na Figura 6, as três regiões estão localizadas no centro do Distrito Federal.

Ainda analisando a Figura 6, nota-se que à medida que se afasta do ponto central do DF, as taxas de inadimplência aumentam (representadas pelas áreas mais escuras do mapa). Destaque negativo para as regiões de Santa Maria, Recanto das Emas e Paranoá, que apresentam as piores taxas de inadimplência.

Foram calculadas as frequências e estatísticas média, mediana, máximo, mínimo e quartis das variáveis candidatas a compor os modelos e, por não terem apresentado inconsistências, valores *missing* ou *outliers*, nenhuma variável foi retirada nessa etapa do estudo.

A análise bivariada consistiu no cálculo da frequência cruzada entre as variáveis preditoras e a variável resposta, com o objetivo de identificar as variáveis que discriminam o risco de crédito do público-alvo do estudo. As variáveis foram categorizadas com base no Risco relativo (14) e, a partir dessa categorização, foram criadas variáveis *dummies* para compor os modelos.

$$Risco\ Relativo\ da\ categoria = \frac{\frac{Total\ de\ clientes\ bons\ na\ categoria}{Total\ de\ clientes\ bons}}{\frac{Total\ de\ clientes\ maus\ na\ categoria}{Total\ de\ clientes\ maus}} \quad (14)$$

As variáveis taxa de desemprego e inflação apresentaram todos os atributos com níveis semelhantes de risco de crédito e, por esse motivo, foram excluídas do estudo. As categorias das demais variáveis encontram-se na Tabela 5.

Tabela 5 *Categorização e Risco Relativo das variáveis.*

Variável	Classe	Categorização	Risco Relativo	Quantidade de Bons	Quantidade de Maus	Total
Renda Formal (salários mínimos)	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
Grau de Instrução	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
Idade (anos)	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
Prazo Contratado (meses)	1	< = 12	1,9630	724	141	865
	2	] 12 ; 24]	1,4197	3.747	1.009	4.756
	3	< = 24	0,8875	11.540	4.971	16.511
Tempo de Relacionamento	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
Taxa SELIC	1	>= 10	1,0115	14.515	5.486	20.001
	2	< 10	0,9007	1.496	635	2.131

Fonte: *Elaborada pelos autores.*

Observa-se na Tabela 5 que tomadores com maior Renda Formal apresentaram menor risco de crédito. Observa-se também que quanto maior é o Grau de Instrução do tomador, menor é seu risco, com os doutores apresentando um risco relativo bem superior aos demais. Os resultados também apontaram que, quanto maior a idade do tomador de crédito e quanto menor o prazo contratado da operação, menores são os riscos de crédito. Com relação ao tempo de relacionamento do tomador com a instituição, aqueles que possuem o menor tempo apresentaram maior risco de crédito.

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 seria a inadimplência e o risco de crédito. No entanto, conforme observado na Tabela 5, 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%. No entanto, 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.

A partir dessa categorização foram criadas variáveis *dummies* para serem utilizadas na composição dos modelos de regressão.

3.2. Indicadores Espaciais

A etapa seguinte do estudo consistiu em aplicar os Índices de Moran Global e Local com o objetivo de verificar a existência de correlação espacial da variável taxa de inadimplência e regiões singulares no universo de estudo.

O Índice de Moran Global apresentou o valor de 0,05, indicando uma dependência espacial quase nula.

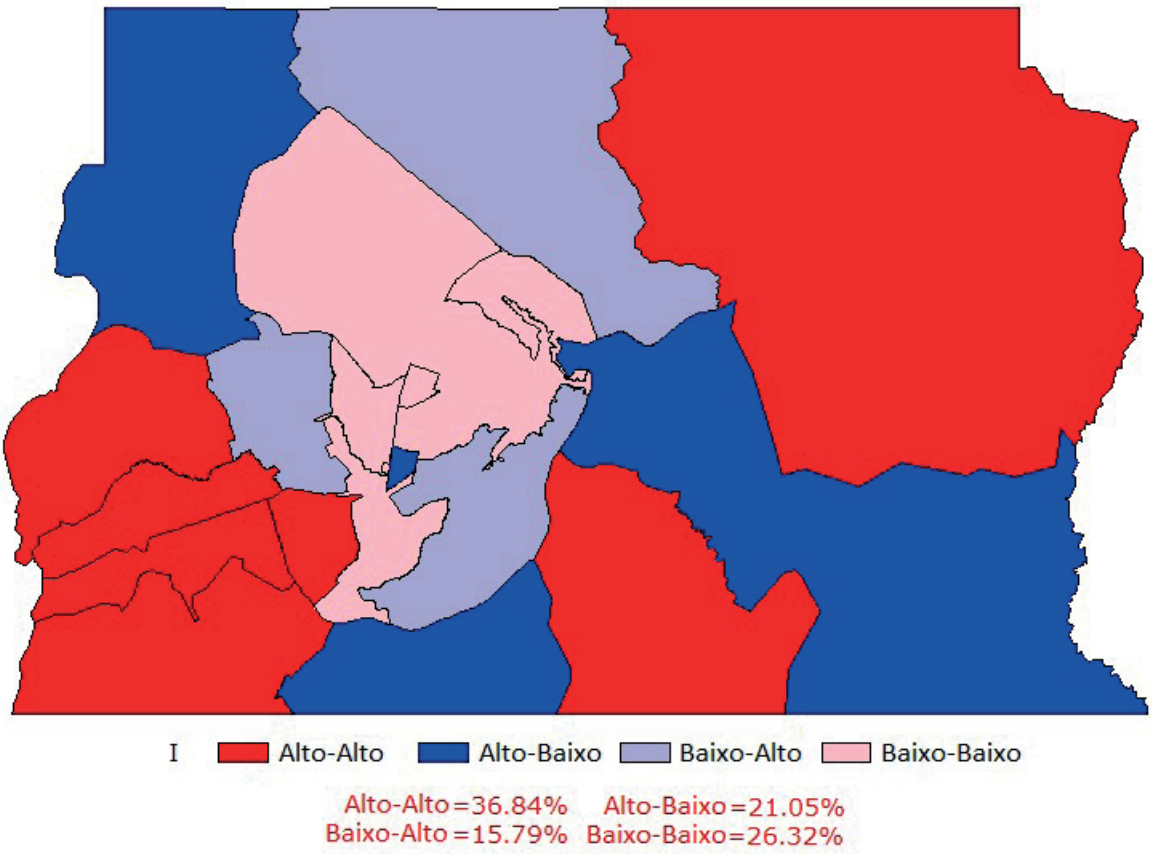

Figura 7. *Mapa de espalhamento de Moran.*
Fonte: *Moran (1950).*

A Figura 7 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 "Baixo-Baixo" são as que apresentaram as menores taxas de inadimplência, seguidas das regiões "Baixo-Alto", "Alto-Baixo" e "Alto-Alto", 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 do público-alvo de campanhas de recuperação de crédito, em que a cobrança dos clientes residentes nas regiões "Alto-Alto" deve ser o foco inicial das ações, visando a melhorar o resultado financeiro da empresa.

Os resultados encontrados para o Índice de Moran Local, utilizando um nível de significância de 95%, são apresentados no Mapa de Moran, na Figura 8.

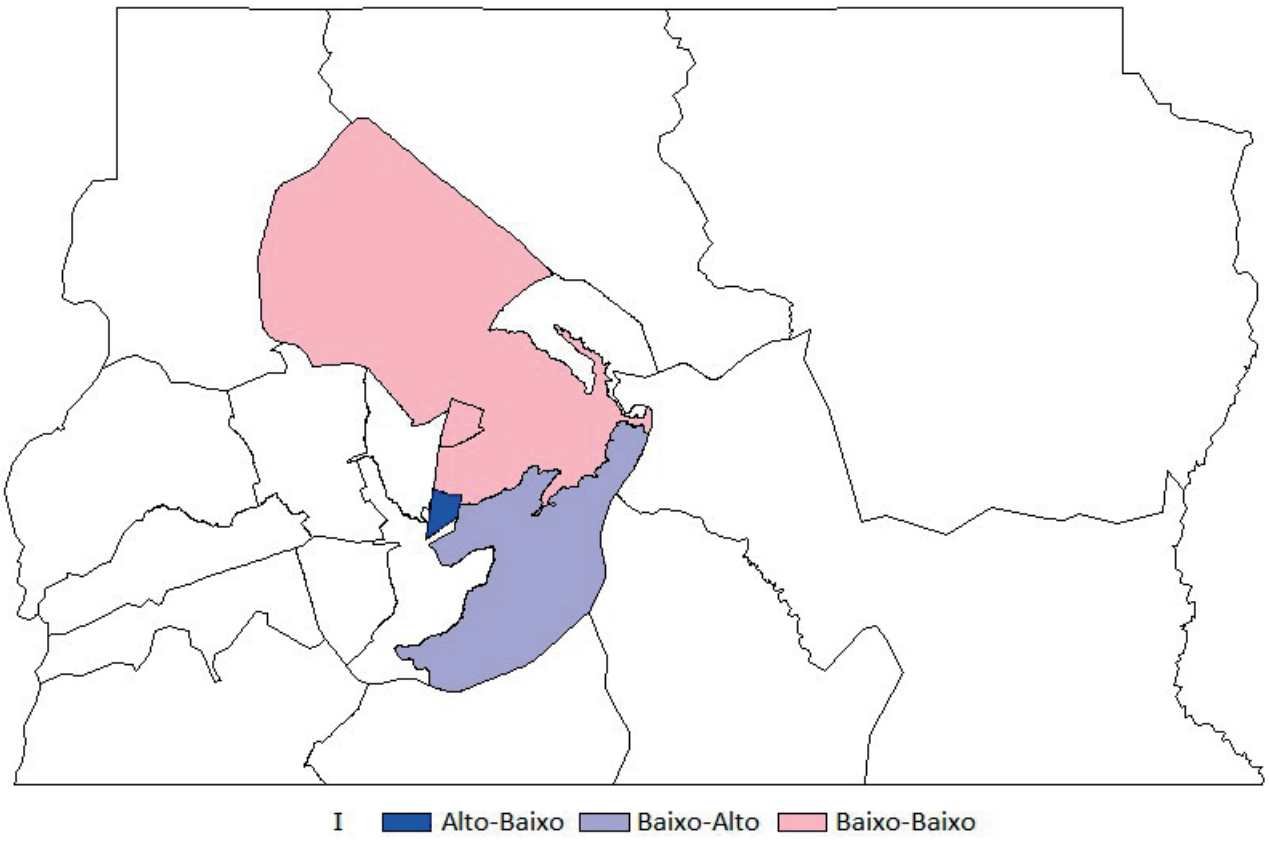

Figura 8. *Mapa de Moran a 95% de confiança.*
Fonte: *Elaborada pelos autores.*

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 8 são Brasília e Cruzeiro (Baixo-Baixo), Lago Sul (Baixo-Alto) e Candangolândia (Alto-Baixo). De acordo com Fotheringham et al. (2002), a existência de valores significativos para o Índice de Moran Local justifica a aplicação da técnica GWLR.

3.3. Modelo Global via Regressão Logística

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

As variáveis utilizadas no desenvolvimento do modelo foram todas as *dummies* criadas a partir das categorizações apresentadas na Tabela 5. Utilizando o método de seleção de variáveis *stepwise*, as variáveis com p-valor abaixo de 0,10 (nível de significância de 10%) e que foram selecionadas para compor o modelo final de regressão logística (modelo global) são apresentadas na Tabela 6.

Tabela 6 *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*	0,46
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: *Elaborada pelos autores.*

A variável SELIC não se mostrou significativa e 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.

Além disso, 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 inversões de risco relativo em suas faixas de valores quando categorizadas de forma granular. Outra possível explicação é que a categorização foi realizada com a base total de registros e o modelo foi desenvolvido com a base de desenvolvimento, que compreende um número menor de registros.

A nomenclatura das variáveis *dummies* respeita a nomenclatura das categorias expostas na Tabela 5. 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 esta 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, de o cliente se tornar inadimplente. Desta maneira, pode-se notar na Tabela 6 que todos os coeficientes da regressão global, exceto os da variável Renda Formal, 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 de o cliente se tornar inadimplente. Esta análise é denominada análise de congruência; é importante para verificar se existem inversões nos coeficientes e se a categorização das variáveis foi realizada de maneira correta.

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

3.4. Modelos Locais via Regressão Logística Geograficamente Ponderada (GWLR)

Conforme descrito na metodologia, foram desenvolvidos quatro modelos utilizando a técnica GWLR, sendo um para cada função de ponderação exposta na Tabela 1. As variáveis preditoras utilizadas foram aquelas selecionadas pelo modelo de regressão logística, expostas na Tabela 6.

O melhor modelo via GWLR, segundo o critério AICc, foi o modelo Gaussiano Variável com valor de 2.022 vizinhos mais próximos para estimar os *bandwidths* variáveis.

A Tabela 7 contém as estatísticas descritivas dos coeficientes estimados pelo modelo GWLR, onde se nota que as médias dos coeficientes ficaram bem próximas dos coeficientes do modelo global apresentados na Tabela 6.

Tabela 7 *Estatísticas dos coeficientes estimados do modelo GWLR 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: *Elaborada pelos autores.*

A Tabela 8 contém a fórmula final dos modelos estimados via GWLR Gaussiano Variável para as 19 regiões do DF.

Tabela 8 *Fórmulas de Regressão Locais estimadas pelo modelo GWLR 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 NS	0,438	0,231	-0,581	-0,321	0,520	0,291	0,100 NS	-0,468	-0,371
BRAZLÂNDIA	-1,320	-0,571	-0,363	0,097 NS	0,522	0,310	-0,691	-0,330	0,496	0,367	0,109 NS	-0,685	-0,431
CANDANGOLÂNDIA	-1,256	-0,740	-0,340	0,011 NS	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 NS	0,548	0,351	-0,763	-0,421	0,537	0,477	0,147 NS	-0,712	-0,448
CRUZEIRO	-1,326	-1,015	-0,351	-0,099 NS	0,313	0,234	-0,485	-0,231 NS	0,557	0,268	0,107 NS	-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 NS	-0,046 NS	0,372	0,323	-0,527	-0,265	0,430	0,217	0,101 NS	-0,685	-0,418
LAGO NORTE	-1,378	-0,755	-0,156	0,148 NS	0,522	0,212	-0,555	-0,289	0,554	0,327	0,043 NS	-0,377	-0,370
LAGO SUL	-1,257	-0,716	-0,308	0,057 NS	0,489	0,268	-0,745	-0,407	0,423	0,292	0,122 NS	-0,528	-0,396
NÚCLEO BANDEIRANTE	-1,258	-0,678	-0,344	0,060 NS	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 NS	0,176	0,585	0,283	-0,808	-0,409	0,350	0,374	0,069 NS	-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 NS	-0,556	-0,434
RECANTO DAS EMAS	-1,253	-0,628	-0,372	0,079 NS	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 NS	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 NS	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 SEBASTIÃO	-1,273	-0,624	-0,247	0,141	0,567	0,289	-0,822	-0,408	0,373	0,354	0,108 NS	-0,507	-0,418
SOBRADINHO	-1,392	-0,568	-0,151 NS	0,216	0,578	0,285	-0,625	-0,273	0,456	0,364	0,025 NS	-0,470	-0,412
TAGUATINGA	-1,271	-0,666	-0,312	0,027 NS	0,485	0,335	-0,582	-0,322	0,439	0,259	0,173	-0,756	-0,452

Nota. *NS: Coeficiente não significativo com 90% de confiança (p-valor acima de 0,10).*
Fonte: *Elaborada pelos autores.*

Nota-se na Tabela 8 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.

Com relação à idade do tomador, as variáveis d_idade1 e d_idade5 se mostraram significativas para todas as regiões do Distrito Federal, enquanto as variáveis d_idade2 e d_idade4 não foram significativas para algumas regiões, indicando que a idade do tomador de crédito influencia o risco de maneira distinta, a depender da região em estudo.

A variável d_instrução4 também se mostrou significativa para todas as regiões do Distrito Federal, apresentando pequena variação dos coeficientes entre as regiões.

Com relação ao Tempo de Relacionamento do tomador de crédito com a instituição, as variáveis d_tempo_rel1 e d_tempo_rel4 se mostraram significativas para todas as regiões do Distrito Federal, enquanto a variável d_tempo_rel2 não se mostrou significativa para a região de Cruzeiro.

Com relação à Renda do tomador de crédito, a variável d_renda1 se mostrou significativa para todas as regiões do Distrito Federal enquanto 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, indicando que a Renda do Tomador também influencia o risco de crédito de maneira distinta entre as regiões.

As variáveis d_pz_contratação1 e d_pz_contratação2, que representam o Prazo de Contratação, se mostraram significativas para todas as regiões do Distrito Federal.

3.5. Comparação Entre os Modelos

A comparação entre o modelo de Regressão Logística 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 9 mostra as estatísticas descritivas dos escores obtidos por ambos os modelos selecionados na amostra de validação.

Tabela 9 *Análise 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: *Elaborada pelos autores.*

As médias dos escores dos modelos ficaram bem próximas, com diferença apenas na terceira casa decimal; no entanto, o modelo via GWLR apresentou uma 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 Sensitividade e Especificidade e seu valor foi de 0,30.

Tabela 10 *Matriz de Confusão dos modelos via RL.*

		Valor Observado RL		Valor Observado GWLR	
		0	1	0	1
Valor Predito	0	48,7%	11,3%	49,0%	11,2%
	1	24,0%	16,0%	23,8%	16,0%

Fonte: *Elaborada pelos autores.*

Pode-se notar na Tabela 10 que os modelos apresentaram resultados bem próximos quanto à classificação dos clientes.

A Tabela 11 contém todas as métricas utilizadas para comparação entre os modelos, onde se nota pequena diferença entre os valores dos indicadores dos dois modelos.

Tabela 11 *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: *Elaborada pelos autores.*

Na Tabela 11, todos os valores obtidos para as métricas dos dois modelos também ficaram muito próximos, 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; já o modelo via RL 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.

4. CONCLUSÃO

Neste artigo foram utilizados dados reais de uma instituição financeira nacional referentes à operação de Crédito Direto ao Consumidor, concedidas a clientes domiciliados em 19 regiões do Distrito Federal para o desenvolvimento de modelos de *credit scoring* através de duas metodologias distintas: Regressão Logística e Regressão Logística Geograficamente Ponderada.

A metodologia Regressão Logística é bastante difundida no setor financeiro, sendo utilizada neste estudo para desenvolver um modelo global de *credit scoring* para todo o Distrito Federal.

A metodologia Regressão Logística Geograficamente Ponderada é pouco difundida e utiliza a localização geográfica do tomador de crédito para ponderar as observações no desenvolvimento de modelos distintos para cada região de estudo.

Os indicadores utilizados para comparação entre os modelos desenvolvidos através das duas metodologias se mostraram bem próximos e, baseados nos resultados obtidos, pode-se considerar que as metodologias são semelhantes em termos de capacidade de previsão de perdas financeiras para a instituição.

O 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 fatores, 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, mostrando que os pesos (importância) das variáveis variaram de região para região.

Os resultados demonstraram a viabilidade da aplicação da metodologia GWLR para desenvolvimento de modelos de *credit scoring* para o público-alvo deste estudo. As fórmulas obtidas são aplicáveis somente a esse público, no entanto, acredita-se que essa metodologia pode ser expandida para outras operações de crédito e níveis espaciais (e. g. bairros, municípios, UFs).

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.

Com relação às limitações do estudo, o uso de poucas variáveis preditoras 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; entretanto, os valores dos seus coeficientes se mostraram invertidos. Estudos considerando outra categorização ou outro público-alvo devem ser realizados para verificar a relevância dessa variável para o risco de crédito.

Como tópicos de pesquisas futuras, sugere-se aplicar a metodologia GWLR para desenvolver modelos de *credit scoring* para outros públicos-alvo (por exemplo, diferentes operações de crédito ou regiões geográficas),

realizar comparações com outras metodologias (tais como *Support Vector Machines* ou *Boosting*), utilizar outras variáveis preditoras, aplicar a metodologia GWLR para o desenvolvimento de modelos em outras áreas de uma instituição financeira, como nas áreas de estratégia e marketing, ou utilizar outras funções, como a Log Binomial, para desenvolver modelos geograficamente ponderados.

REFERÊNCIAS

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, *27*(2), 93-115.

Atkinson, P. M., German, S. E., Sear, D. A., & Clark, M. J. (2003). Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geographical Analysis*, *35*(1), 58-82.

Banco Central do Brasil (2009). Resolução CMN nº 3.721, de 30/04/2009. Recuperado de http://www.bcb.gov.br.

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, *28*(4), 281-298.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*(3), 1447-1465.

Fernandes, G. B., & Artes, R. (2016). Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research*, *249*(2), 517-524.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester: John Wiley & Sons.

Gilbert, A., & Chakraborty, J. (2011). Using geographically weighted regression for environmental justice analysis: Cumulative cancer risks from air toxics in Florida. *Social Science Research*, *40*(1), 273-286.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523-541.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Hoboken, NJ: John Wiley & Sons.

Huang, Y., & Leung, Y. (2002). Analysing regional industrialisation in Jiangsu province using geographically weighted regression. *Journal of Geographical Systems*, *4*(2), 233-249.

Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(2), 271-293.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124-136.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17-23.

See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., Comber, A., ..., & Obersteiner, M. (2015). Building a hybrid land cover map with crowdsourcing and geographically weighted regression. *ISPRS Journal of Photogrammetry and Remote Sensing*, *103*, 48-56.

Stine, R. (2011). Spatial temporal models for retail credit. In *Proceedings of the Credit Scoring and Credit Control Conference,* Edinburgh, UK.

Endereço para correspondência:
Pedro Henrique Melo Albuquerque

Universidade de Brasília, Departamento de Administração
Campus Universitário Darcy Ribeiro, Bloco A-2, 1º andar, Sala A1-54/7 – CEP: 70910-900
Asa Norte – Brasília – DF – Brasil
E-mail: pedroa@unb.br