



This is an open-access article distributed under the terms of the Creative Commons Attribution License .
Fonte: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1519-70772017000100093&lng=en&nrm=iso. Acesso em: 16 jan. 2018.

REFERÊNCIA

ALBUQUERQUE, Pedro Henrique Melo; MEDINA, Fabio Augusto Scalet; SILVA, Alan Ricardo da. Geographically weighted logistic regression applied to Credit Scoring models. *Revista Contabilidade & Finanças*, São Paulo, v. 28, n. 73, p. 93-112, jan./abr. 2017. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1519-70772017000100093&lng=en&nrm=iso>. Acesso em: 16 jan. 2018. doi: <http://dx.doi.org/10.1590/1808-057x201703760>.

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, Brazil

Fabio Augusto Scalet Medina

Universidade de Brasília, Faculdade de Economia, Administração, Contabilidade e Políticas Públicas, Departamento de Administração, Brasília, DF, Brazil

Alan Ricardo da Silva

Universidade de Brasília, Instituto de Ciências Exatas, Departamento de Estatística, Brasília, DF, Brazil

Received on 05.11.2016 – Desk acceptance on 06.20.2016 – 2nd version approved on 10.11.2016

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.

*Paper presented at the XL ANDAP Congress, Costa do Sauípe, BA, Brazil, September 2016.

1. INTRODUCTION

The main activity of commercial banks is financial intermediation, which consists of raising financial resources and lending them to third parties under pre-established conditions, such as payment period, installment value, and interest rate (Hand & Henley, 1997). As it involves expectation of future receipt, all credit granted is exposed to risks.

The topic “risk management” drew attention in the financial sector after the publishing of the Basel accords, which is a set of documents that serve as a basis for regulation and monitoring of the sector. Advances in technology and computing, together with the development of quantitative methods, have contributed in creating different tools for measuring risk, bringing significant gains in the financial management of institutions.

Credit risk can be defined as the possibility of financial losses occurring, associated with borrowers or counterparties not fulfilling their respective obligations in the agreed terms, with the devaluation of loan contracts because of a deterioration in borrowers’ risk classifications, with reductions in earnings or remunerations, with advantages conceded in renegotiations, and with recovery costs (Brazilian Central Bank [BACEN], 2009). It is one of the main risks that financial institutions are exposed to.

The models used to measure risk when granting credit are called credit scoring models. Due to them involving lower costs and greater agility, objectivity, and predictive power in credit granting decisions, credit scoring models have become popular and are widely used by the financial sector (Hand & Henley, 1997).

Lessmann, Baesens, Seow, and Thomas (2015) carried out a comprehensive study on the classification methodologies used for developing credit scoring models and indicated logistic regression as being the standard methodology in the financial sector.

Logistic regression is a multivariate analysis technique that aims to explain the relationship between a random binary dependent variable and a set of independent predictive variables (Hosmer & Lemeshow, 2000).

Financial institutions use various credit scoring models, which are applied when evaluating different types of clients or credit operations to be contracted. The predictive variables that compose each model can be different, with the aim of improving predictions for the target population.

Geographical location (space) and its relationship with credit risk is the topic of some published studies. Among the most recent, Stine (2011) analyzes the evolution of

defaults on real estate loans in US counties between 1993 and 2010, contemplating pre- and post- subprime crisis periods, and finding evidence of a spatial correlation between default rates in these counties.

Fernandes and Artes (2015) used the Ordinary Kriging methodology to create a variable that reflects spatial risk and applied the Logistic Regression technique to verify the existence of a spatial correlation in defaults on loans taken out by small and medium sized enterprises (SMEs), using data from the *SERASA* credit bureau. The authors developed models with and without the spatial risk variable and confirmed that the inclusion of this variable improves credit scoring model performance.

The Geographically Weighted Regression (GWR) technique, proposed by Brunson, Fotheringham, and Charlton (1996), is used to model spatially heterogeneous (non-stationary) processes; that is, processes that vary (whether in mean, median, variance, etc.) from region to region. The basic idea of GWR is to adjust a regression model to each region in the data set using geographical location of the other observations to weight the parameter estimates. Application of the GWR technique can be observed in different areas of research, such as Geography (See et al., 2015), Health (Gilbert & Chakraborty, 2011), and Economics (Huang & Leung, 2002).

Atkinson, German, Sear, and Clark (2003) used Geographically Weighted Logistic Regression (GWLR) in their study to analyze the dependency of geographical location in the relationship between erosion and geomorphologic controls in a region of Wales. The dummy variable used in this study was the presence or absence of erosion in the areas studied. Applying the GWLR technique resulted in the estimation of models with different parameters (distinct models) for each area studied, revealing the need to adopt different practices to avoid erosion, depending on the region.

This article used data related to transactions involving Consumer Direct Credit (CDC), granted by a Brazilian financial institution to clients residing in the Distrito Federal (DF). The aims were as follows: to verify whether the factors that influence credit risk differ according to borrowers’ geographical locations; 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.

Although the central idea in this article of verifying

whether space influences credit risk is similar to that of Stine (2011) and Fernandes and Artes (2015), the target population and methodology used are different, with no studies being found in the literature that used the GWLR technique for the development of credit scoring models.

One advantage of applying the GWLR technique in relation to the others lies in estimating a model for each region in the study, allowing these models to be distinct in their variables and parameters (Atkinson et al., 2003), whereas a global model, represented by only one formula, may not represent local variations adequately. In relation to credit, different study regions can involve different risks, and if this phenomenon is observed, models that consider local characteristics can better differentiate the credit risk for borrowers residing there and generate

financial gains for the institution.

Another difference from other studies on this topic and an advantage in the GWLR technique involves the use of different samples in developing each local model, giving greater weight to borrowers who are closer geographically, and not using distant information that is outside the radius defined by the weighting function.

Questions regarding endogeneity are not addressed in this study and could be raised by researchers in future papers.

In addition to this introduction, the second section of the article presents the geographically weighted logistic regression methodology and the process for developing the models, the thirds shows the results obtained, and the fourth sets out the conclusion.

2. METHODOLOGY

The flowchart presented in Figure 1 details all of the stages carried out in the process of developing the models in this study.

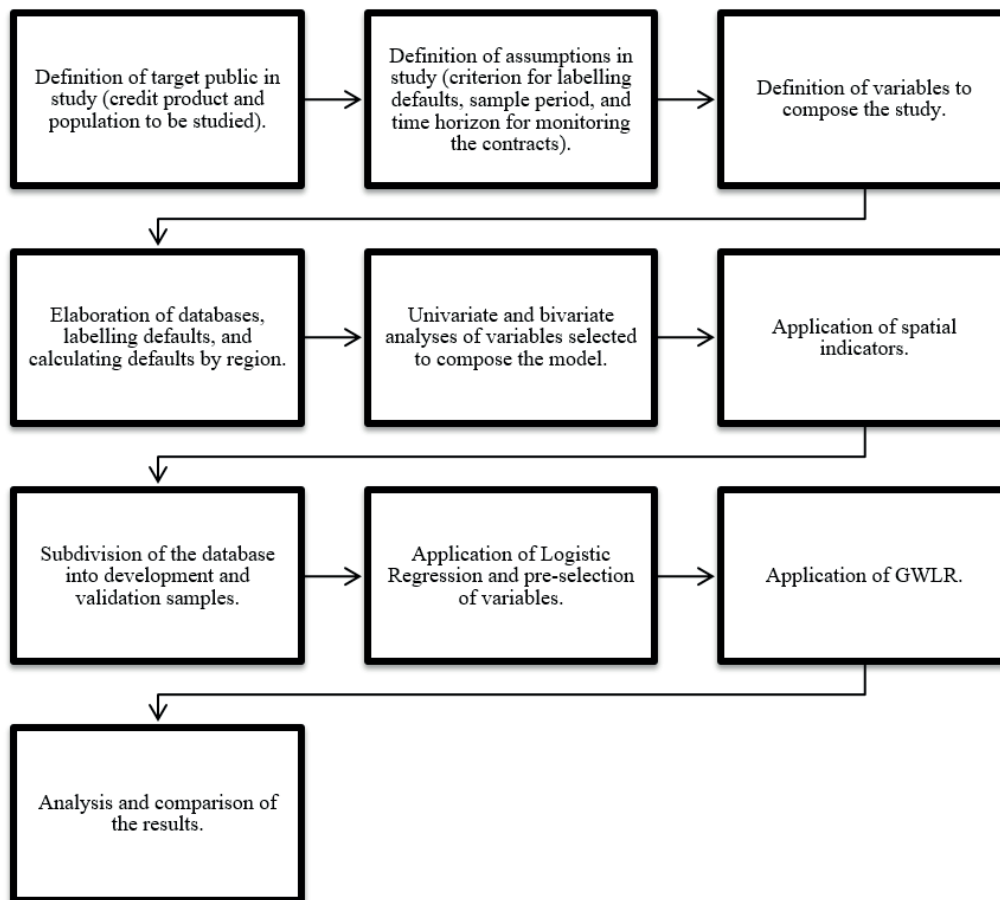


Figure 1. Flowchart of the stages in developing the models.

Source: Prepared by the authors.

2.1. Database

The data related to this study refer to transactions involving Consumer Direct Credit (CDC) granted by a Brazilian financial institution to clients residing in the

Distrito Federal (DF). These transactions are paid in installments over periods between 0 and 36 months and have a maximum contract value of R\$30,000.00.

The territorial division of the DF used in this study was composed of 19 regions, shown in Figure 2.

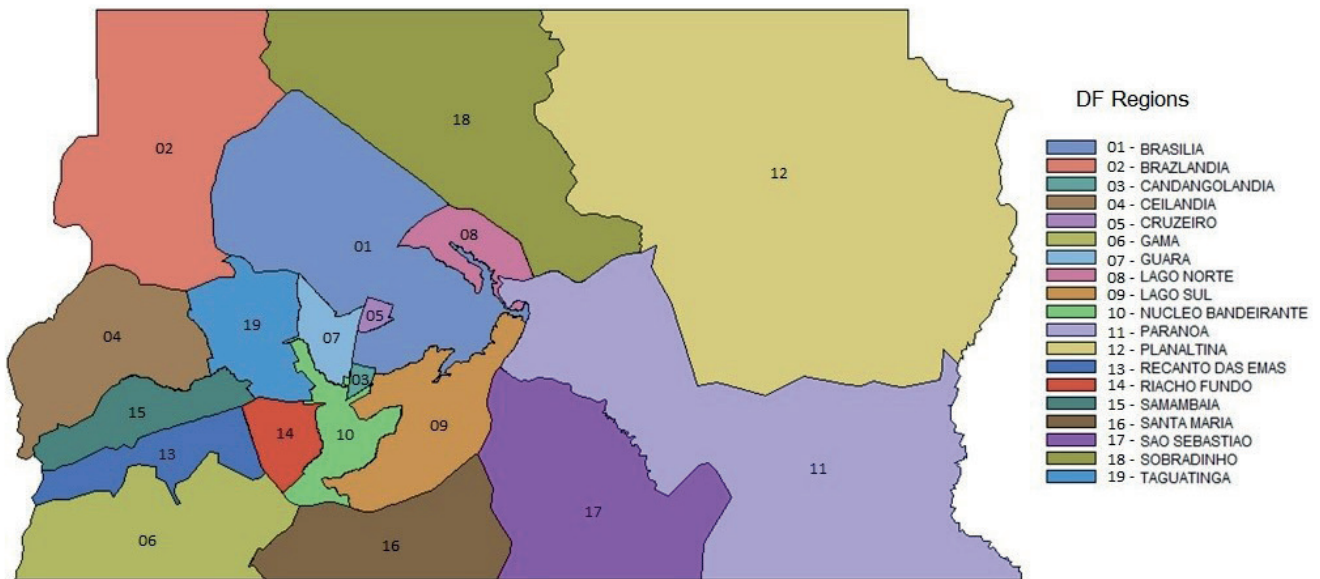


Figure 2. Territorial division of the Distrito Federal used in the study.

Source: Prepared by the authors.

The sample included all loans granted between December 2013 and September 2014, involving 10 rounds of borrowing and a total of 22,132 different loan contracts. Payment performance on these loans was monitored in the twelve months subsequent to the contract agreement date and those that exceeded 90 days in arrears in any of these months were labeled as being in default ($Y=1$). Due to loan arrears performance involving different moments in time, this database is classified as being of the panel data type.

The predictive variables selected to compose the models were: Age, Income, Level of Education, Borrower's Time of Relationship with the Institution, Loan Contract Period, SELIC, Unemployment Rate, and Inflation (*IPCA*). These variables refer to the time credit is taken out (a single point in time), thus involving cross-sectional type data.

The latitude and longitude geographical coordinates for the regions used in the study and needed to apply the GWLR technique were obtained from the *IBGE* website, and refer to the central point in each region and are equal for borrowers residing in the same region.

The database was subdivided into model development

and validation samples according to the date a transaction was contracted, with the development sample composed of the first five rounds (December 2013 to April 2014) and totaling 10,944 records. The validation database is composed of the final five rounds (May to September 2014), totaling 11,188 records.

The data manipulation, as well the univariate, bivariate, and spatial indicator calculations, along with those for developing the global model via logistic regression analysis, were carried out using the SAS software. The GWLR models were developed using the GWR4 software.

2.2. Spatial indicators

Moran's *I* (Moran, 1950) is one of the most widely used global indicators for verifying the existence of spatial correlation. Global indicators present a single measure of spatial tendency for the whole region being studied, they allow the hypothesis of the existence of spatial dependency between regions to be tested in accordance with the variable of interest, and are used in exploratory analysis of data. The formula is given by:

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

in which n is the number of regions being studied, x_i and x_j are the values of the variable of interest in regions i and j , and w_{ij} are the spatial proximity matrix elements, which can be calculated in different ways, such as via the presence or absence of a frontier between the regions or by the Euclidian distance between them. The Moran index is restricted to the interval $[-1,1]$, in which values close to -1 indicate negative spatial correlation, values close to 1 indicate positive spatial correlation, and a value equal to 0 indicates the absence of spatial correlation or spatial

independence with relation to the variable tested. Whereas the global indicators assume that all of the regions studied can be represented by a single value, the local indicators of spatial association (LISA), developed by Anselin (1995), are used to verify the existence of spatial correlation within the geographical units studied and look for regional differences (peculiarities). The presence of areas with significant local indices is an indication of spatial (non-stationary) homogeneity.

The Moran Local Index formula is given by:

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

The database used in applying the Moran Global and Local Indices was the total database of records (without

subdivision of samples) and the variable tested was the regional default rate, calculated via the following formula:

$$\text{Default Rate in the Region} = \frac{\text{Number of defaulting clients in the region}}{\text{Total number of clients in the region}} \tag{3}$$

In this study the Moran Global Index was used to verify the existence of spatial correlation in the default rate between the regions in the DF. The Moran Local Index was used to verify the existence of regions with different default rates in relation to the others. The existence of significant regions (the confidence level used for the Moran Local Index was 95%) may indicate that the regression models developed for these regions are different in relation to

the models for the other regions in the study, which may warrant applying the GWLR to this target population.

2.3. Geographically Weighted Regression

According to Fotheringham, Brunson, and Charlton (2002), given a basic linear regression model, the equivalent expression for the GWR is given by:

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

It is noted from the expression above that the model parameters represented by the function $\beta_k(u_i, v_i)$ vary according to the values (u_i, v_i) , which represent the latitude and longitude geographical coordinates for observation (region) i , resulting in a different model for each region in

the study. The assumptions of the classical linear regression model remain in place for GWR.

The matrix form for estimating the GWR parameters is given by:

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

in which

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

$W(u_i, v_i)$ is a diagonal matrix and different for each point i of coordinates (u_i, v_i) , containing the weights w_{ij} in its main diagonal, obtained via the weighting functions, or kernel. The substitution of all the weights w_{ij} for the value 1 equates to the identity matrix, which, substituted in (5),

turns it back into the classical linear regression model.

The two main weighting functions found in the literature are the Normal or Gaussian function and the Bisquare function. The formulas for both functions are presented in Table 1.

Table 1 Weighting functions or kernels.

Weighting Functions	Weighting Function Formulas
Fixed Gaussian	$w_{ij} = \exp\left\{-\frac{1}{2}(d_{ij}/b)^2\right\}$
Fixed Bisquare	$w_{ij} = [1 - (d_{ij}/b)^2]^2$ if $d_{ij} < b$, and $w_{ij} = 0$ otherwise
Adaptive Gaussian	$w_{ij} = \exp\left\{-\frac{1}{2}(d_{ij}/b_{i(k)})^2\right\}$
Adaptive Bisquare	$w_{ij} = [1 - (d_{ij}/b_{i(k)})^2]^2$ if $d_{ij} < b_{i(k)}$, and $w_{ij} = 0$ otherwise

Source: Fotheringham, Brunson, and Charlton. (2002).

It is noted from Table 1 that there are two types of expressions for each one of the Gaussian and Bisquare functions, which differ in the method of choosing the b (bandwidth) parameter to be used (whether fixed or variable). The d_{ij} parameter contained in the weighting functions represents the distance from point i to point j , the b parameter is the fixed bandwidth (smoothing parameter), and the $b_{i(k)}$ parameter represents the adaptive bandwidth, with the letter k representing the number of

neighbors closest to point i .

The bandwidth parameter controls the variance in the weighting function; for this reason, in situations in which the data are not equally distributed between regions, use of the bandwidth adaptive is recommended. Figure 3 illustrates the bandwidth in the weighting function and Figures 4 and 5 exemplify the use of fixed or adaptive bandwidths.

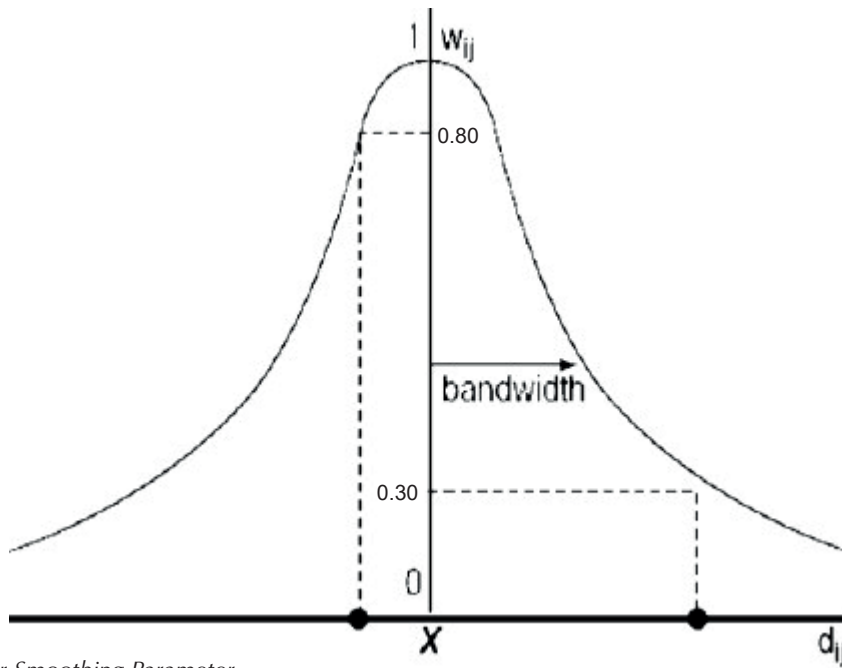


Figure 3. Bandwidth or Smoothing Parameter.
Source: Adapted from Fotheringham et al. (2002).

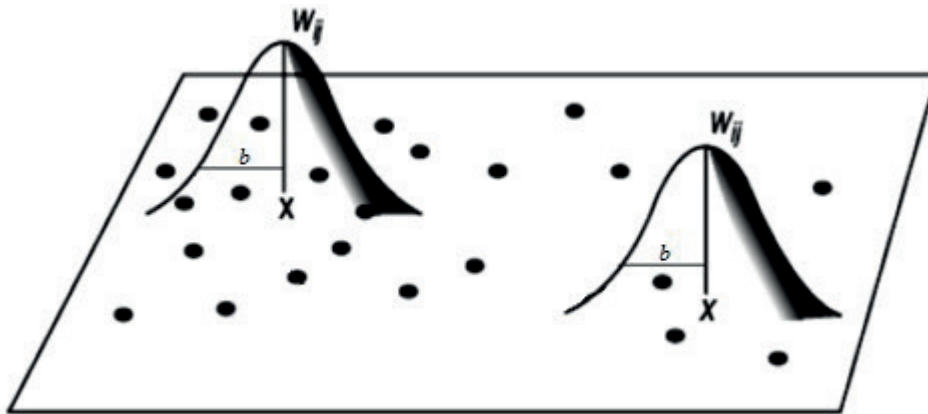


Figure 4. Spatial weighting functions with fixed Bandwidth.
Source: Adapted from Fotheringham et al. (2002).

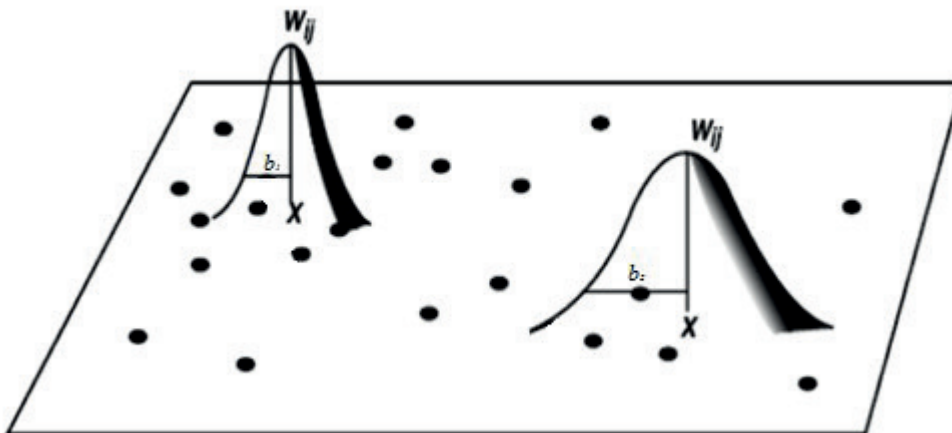


Figure 5. Spatial weighting functions with adaptive Bandwidth.
Source: Adapted from Fotheringham et al. (2002).

When developing a model via GWR using the fixed bandwidth, it should be specified by its value in unit of distance; however, in using the adaptive bandwidth, a k (fixed) number of closest neighbors to be used in the models should be defined, and based on this quantity k , the value of the bandwidth varies between the regions being studied.

2.4. Geographically Weighted Logistic Regression

When the response variable is binary, GWR should be applied via Geographically Weighted Logistic Regression (GWLR), in which the formula for obtaining the probability of the event of interest occurring is given by:

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

or still, in the form:

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

in which $\pi(\mathbf{x}_j)$ is the probability of the j^{th} client defaulting and the function $\beta_k(u_i, v_i)$ represents the parameters (coefficients) of the k variables in the model, which vary according to the region i of latitude and longitude

coordinates (u_i, v_i) .

The GWLR parameters are estimated via the maximum vraisemblance method and the GWLR vraisemblance function is represented by the following expression:

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

By applying the natural logarithm transformation (ln) and developing the formula, we obtain:

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

The $W(u_i, v_i)$ matrix described in (6) features weights w_{ij} (calculated via the weighting functions shown in Table 1) and is used to geographically weight the observations in the estimation of each set of parameters $\beta_k(u_i, v_i)$. That is, this matrix is responsible for assigning a higher weight to the geographically closest observations to region i in

the estimation of its parameters, and assigning a lower or zero weight (depending on the weighting function chosen) for the most distant observations from region i in question in the estimation of its parameters $\beta_k(u_i, v_i)$. The $W(u_i, v_i)$ matrix also varies according to the location of each borrower and composes the likelihood function in the following way:

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

11

Similar to the logistic regression model, after differentiating (11) in function of $\beta(u_i, v_i)$ and equating to zero, the model parameters are estimated using interactive numerical methods, such as the interactively reweighted least squares (IRLS) method. It should be noted that this maximization procedure is carried out for each one of the functions related to each region i in the study.

Initially, four different models were developed using each one of the weighting functions presented in Table 1. The best model based on AICc was selected for comparison with the global model and to compare between the local models (the models generated for each region in the DF) in terms of significance of the variables that composed

the final formula and estimations of the coefficients of the variables.

2.5. Comparison Between the Models

The metrics used to compare the models developed via GWLR and Logistic Regression were: the AICc informational criteria (Hurvich, Simonoff, & Tsai, 1998), 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 defaults compared with the monetary value of defaults observed.

The accuracy of the models and percentage of false positives were obtained via the confusion matrix, given by:

Table 2 Confusion Matrix

		Value Observed	
		0	1
Value Predicted	0	TP	FP
	1	FN	TN

Note. TP: True Positive – number of good clients classified as good; TN: True Negative – number of bad clients classified as bad; FP: False Positive – number of bad clients classified as good; FN: False Negative – number of good clients classified as bad.

Source: Adapted from Crook, Edelman, and Thomas (2007).

According to Table 2, there are two types of error that a classifying model can commit: rejecting good clients (False Negative – FN), or approving bad clients (False Positive – FP). The latter, also known as a Type II Error, is considered to be the worst of the two errors, since these clients would be approved and could generate financial losses for the institution. Thus, the FP percentage was one of the metrics used to compare the models.

The sum of the outstanding balance of all borrowers classified as FP was measured to verify the monetary value that would enter into default due to classification error in the model.

The accuracy of the model is calculated using the proportion of TP and TN in relation to the total, as in the following formula:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

12

The expected monetary value of portfolio defaults was calculated using the expected discrete distributions formula, given by:

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

in which n is the total number of borrowers in a portfolio, x_i is the outstanding balance on the credit transaction for borrower i , and $P(Y_i = 1)$ is the probability of borrower i defaulting, resulting in the credit scoring models. This

value was compared with the value of the sum of defaulting client debts, with the aim of verifying which model comes closest to the real default value.

3. RESULTS

3.1. Univariate and Bivariate Analyses

The results on general default rates and those by region are shown in Tables 3 and 4 and the spatial distribution of default rates is found in Figure 6.

Table 3 Distribution of frequencies of response variable Y

Y	Frequency	Percentage	Accumulated Frequency	Accumulated Percentage
0	16,011	72.34%	16,011	72.34%
1	6,121	27.66%	22,132	100.00%

Source: Prepared by the authors.

Table 4. Default rates by DF region.

Region	Number of Defaults	Total Number	Default Rate
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%

Source: Prepared by the authors.

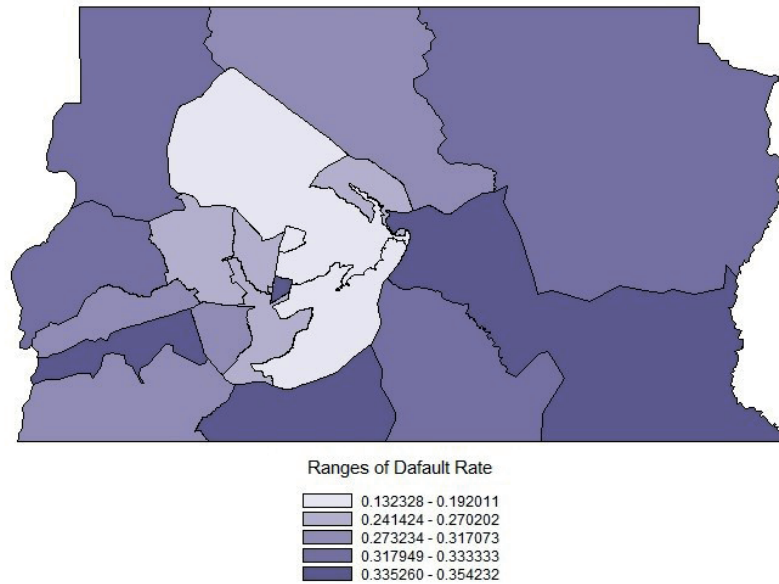


Figure 6. Spatial distribution of default rates in the Distrito Federal.
Source: Prepared by the authors.

As shown in Table 3, the general default rate in the DF was 27.66%; thus, it can be observed in Table 4 that only seven regions (Lago Sul, Cruzeiro, Brasília, Guar, Lago Norte, Taguatinga, and Ncleo Bandeirante) have lower default rates than the general average. It is also noted that the Lago Sul region presented the lowest default rate of the regions studied, followed by the Cruzeiro and Brasilia regions. As can be observed in Figure 6, the three regions are located in the center of the Distrito Federal.

Also by analyzing Figure 6, it is noted that the greater the distance from the central point in the DF, the more default rates increase (represented by the darkest areas on the map). The Santa Maria, Recanto das Emas, and Parano regions stand out in negative terms by presenting

the worst default rates.

The frequencies were calculated, along with the mean, median, maximum, minimum, and quartiles statistics for the candidate variables for composing the models, and as there were no inconsistencies, missing values, or outliers, no variable was removed in this stage of the study.

The bivariate analysis consisted of calculating the cross frequency between the predictive variables and the response variable, with the aim of identifying the variables that differentiate credit risk among the target population in the study. The variables were categorized based on relative Risk (14), and using this categorization, dummy variables were created to compose the models.

$$Relative\ Risk\ of\ the\ category = \frac{\frac{Total\ good\ clients\ in\ the\ category}{Total\ good\ clients}}{\frac{Total\ bad\ clients\ in\ the\ category}{Total\ bad\ clients}} \quad 14$$

All attributes of the rate of unemployment and inflation variables presented similar levels of credit risk, and for this

reason, they were excluded from the study. The categories for the other variables are found in Table 5.

Table 5 Categorization and Relative Risk of the variables.

Variable	Class	Categorization	Relative Risk	Number of Good Clients	Number of Bad Clients	Total
Formal Income (minimum wages)	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
Level of Education	1	PhD	6.1168	48	3	51
	2	Masters	2.1941	132	23	155
	3	Specialization or Completed College Degree	1.5530	4,570	1,125	5,695
	4	Incomplete College Degree or lower Level of Education	0.8662	11,261	4,970	16,231
Age (years)	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
Period Contracted (months)	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
Time of Relationship	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
SELIC Rate	1	>= 10	1.0115	14,515	5,486	20,001
	2	< 10	0.9007	1,496	635	2,131

Source: Prepared by the authors.

It is observed in Table 5 that borrowers with higher Formal Incomes presented a lower credit risk. It is also observed that the higher the borrower's Level of Education, the lower the risk, with PhDs presenting a much higher relative risk than the rest. The results also indicated that the older the borrower and the shorter the loan period, the lower the credit risks. With relation to the borrower's time of relationship with the institution, those with shorter times presented a greater credit risk.

The SELIC rate is the basic interest rate in the Brazilian economy. An increase in the SELIC makes it more expensive for financial institutions to raise funds, which consequently makes credit transactions more expensive. Higher interest rates in credit transactions reduce the purchasing power of borrowers, and because of this, it is expected that the higher the SELIC rate, the greater defaults and credit risk will be. However, as observed in Table 5, the results obtained were the opposite from expected, with lower relative risk (greater credit risk)

for SELIC values below 10.00%, and lower credit risk for values above 10.00%. However, even in light of the results presented, the decision was made to maintain the SELIC rate variable in the study due to it being the only remaining macroeconomic variable. Subsequent studies using a more comprehensive target population should be conducted to better assess this variable.

Based on this categorization, dummy variables were created to be used for composing the regression models.

3.2. Spatial Indicators

The next stage in the study involved applying the Moran Global and Local Indices with the aim of verifying the existence of a spatial correlation between the default rate variable and the individual regions in the study population.

The Moral Global Index presented a value of 0.05, indicating an almost null spatial dependency.

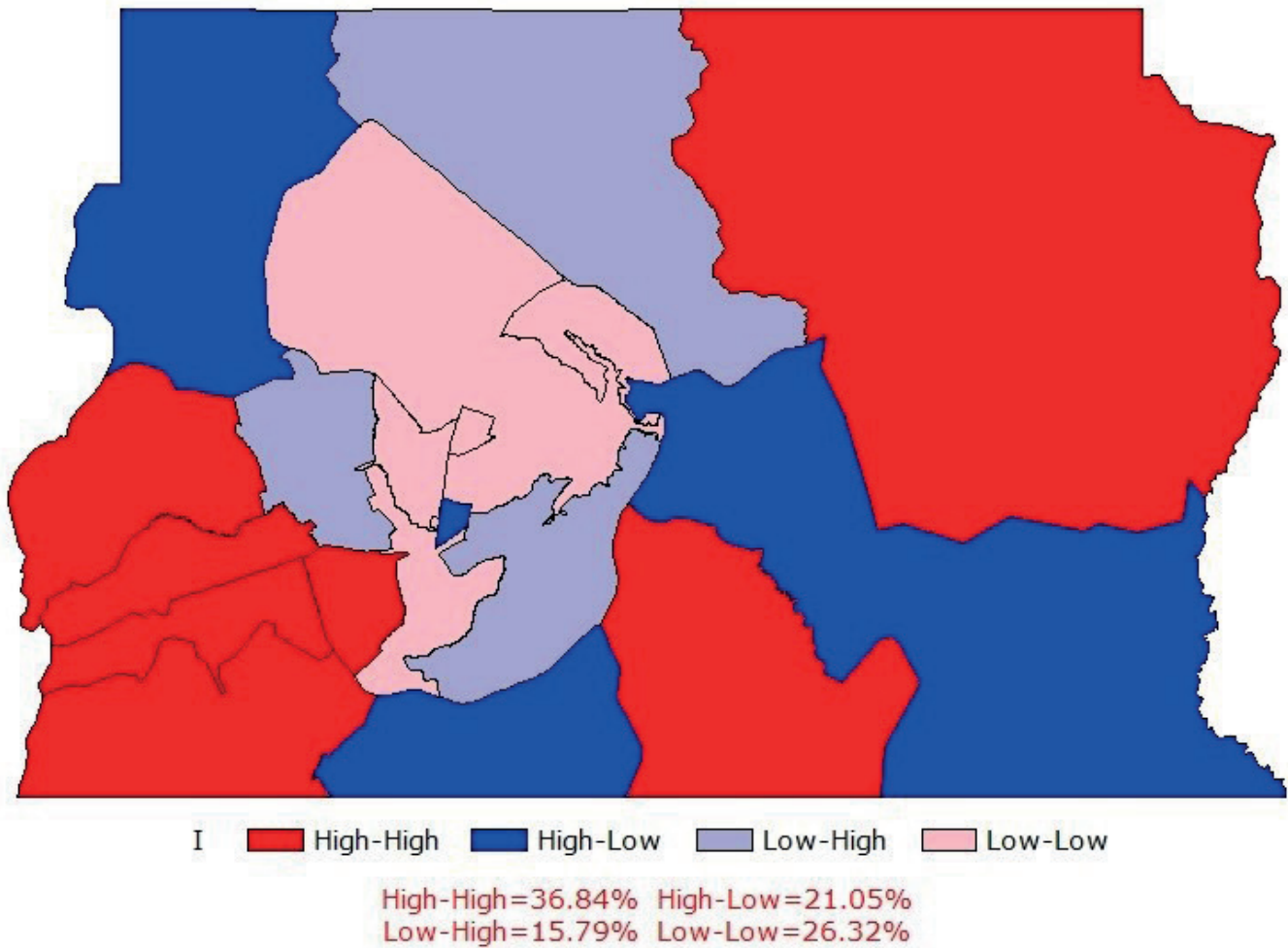


Figure 7. Moran dispersion map.
Source: Moran (1950)

Figure 7 presents the Moran dispersion map in which the regions colored in red tones present positive spatial dependency, whereas the regions colored in blue tones present negative spatial dependency. The “Low-Low” type regions presented the lowest default rates, followed by the “Low-High”, “High-Low”, and “High-High” regions. These results can be considered as spatial clusters of the default rate variable. This information could be used by

the financial institution to define the target population in loan recovery campaigns, in which obtaining payment from clients residing in the “High-High” regions should be the initial focus of activities, with the aim of improving the company’s financial results.

The results found for the Moran Local Index, using a 95% level of significance, are presented in the Moran Map in Figure 8.

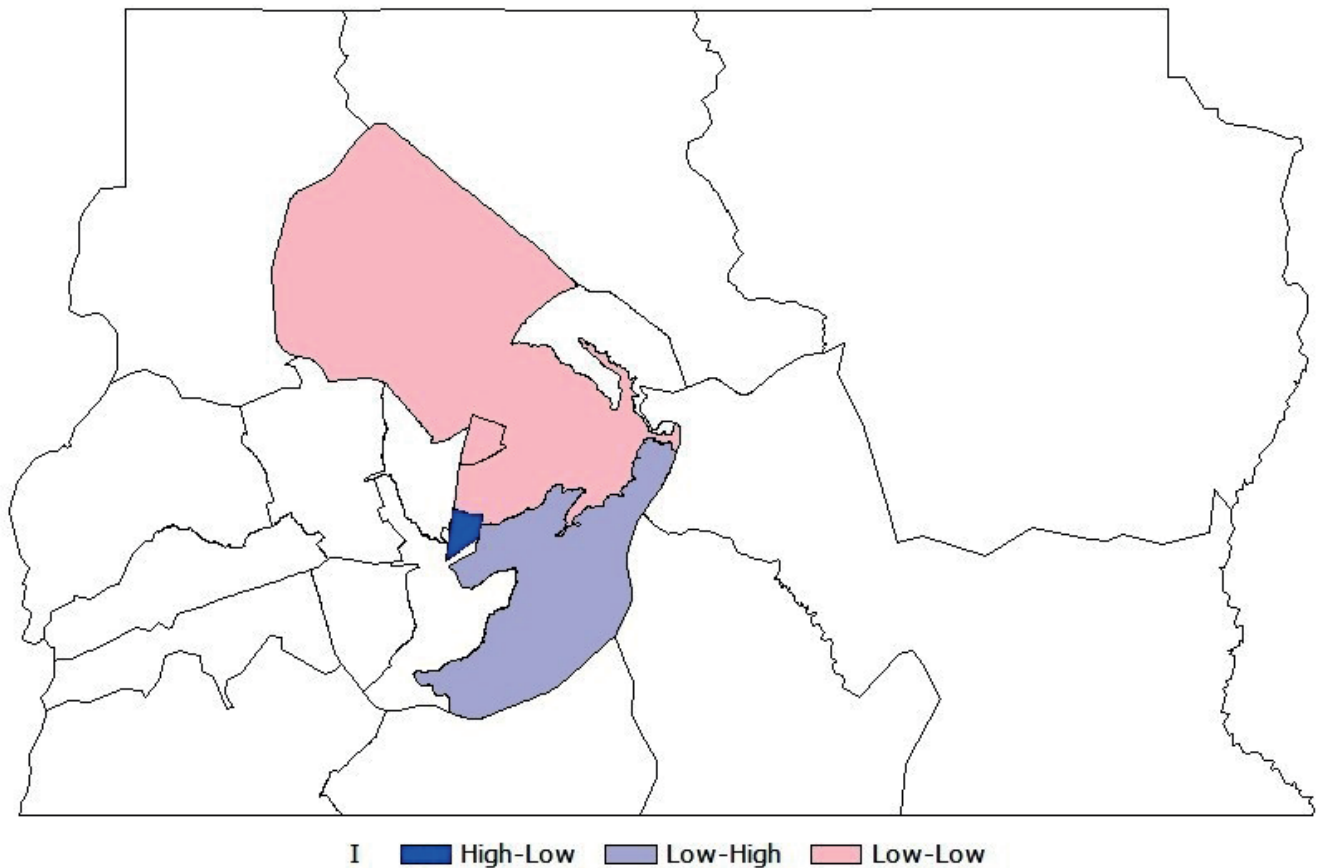


Figure 8. Moran Map with 95% confidence.
Source: Prepared by the authors.

The Moran Map indicates the existence of local correlations in some regions that are significantly different from the others, revealing indications of spatial heterogeneity. The significant regions in the local index and which are labeled in Figure 8 are Brasília and Cruzeiro (Low-Low), Lago Sul (Low-High), and Candangolândia (High-Low). According to Fotheringham et al. (2002), the existence of significant values for the Moran Local Index warrants applying the GWLR technique.

3.3. Global Model via Logistic Regression

The global model was developed using the development sample, containing 10,944 records.

The variables used in developing the model were all of the dummies created based on the categorizations presented in Table 5. Using the stepwise variable selection method, the variables with p-values under 0.10 (10% level of significance) and which were selected to compose the final logistic regression model (global model) are presented in Table 6.

Table 6 Final variables in the global model and respective coefficients.

Variables	Coefficients	Standard Deviation	Wald Statistic	Ratio of Chances
Intercept	-1.3068	0.0893	-14.6338*	-
d_age1	-0.5665	0.084	-6.7440*	0.567
d_age2	-0.2891	0.0907	-3.1874*	0.749
d_age4	0.1481	0.0635	2.3323*	1.160
d_age5	0.5684	0.0653	8.7044*	1.765
d_education4	0.3019	0.0614	4.9169*	1.352
d_time_rel1	-0.7764	0.0862	-9.0070*	0.46
d_time_rel2	-0.3529	0.0844	-4.1813*	0.703
d_time_rel4	0.4206	0.0566	7.4311*	1.523
d_income1	0.3742	0.0705	5.3078*	1.454
d_income2	0.1135	0.06	1.8917**	1.120
d_pd_contract1	-0.6099	0.1398	-4.3627*	0.543
d_pd_contract2	-0.4165	0.0541	-7.6987*	0.659

* *p*-value below 0.05.

** *p*-value below 0.10.

Source: Prepared by the authors.

The SELIC variable was not significant and was not selected to compose the final global regression model. One possible explanation for this fact is the use of a short loan contract period, leading to few distinct values for this variable.

Moreover, the coefficients for the Formal Income variable were inverted, in which the best income bands (d_income1 and d_income2) obtained worse coefficients with relation to the worst band (d_income4, the coefficient for which is zero). This result can be explained by the variable's behavior, with inversions of relative risk in its value ranges when categorized granularly. Another possible explanation is that the categorization was carried out based on total records and the model was developed using the development database, which covers a smaller number of records.

The nomenclature for the dummy variables respects the nomenclature for the categories shown in Table 5. For example, the dummy d_age1 represents the age category "> 55 years old" and is the best category of this variable with relation to credit risk, and the dummy d_education4 represents clients in the category "Incomplete College Degree or lower level of education", with this being the worst category for the Level of Education variable with relation to credit risk.

Response variable *Y* involves the occurrence of defaults ($Y=1$) as the event of interest, with the probability resulting from the logistic regression models and via GLWR referring to the probability of this event occurring; that

is, of the client defaulting. Thus, it can be noted in Table 6 that all of the global regression coefficients, except for the Formal Income variable, are coherent, since the best categories for each variable with relation to credit risk presented lower coefficients in relation to the higher risk categories for the same variable; that is, the presence of the best categories for each variable reduces the probability of a client defaulting. This analysis is called congruence analysis; it is important for verifying whether there are inversions in the coefficients and whether categorization of the variables was carried out correctly.

The value found for the AICc informational criterion of the global model was 12,098.29, with this value being used for comparison with the models estimated via GWLR, the results from which are presented below.

3.4. Local Models via Geographically Weighted Logistic Regression (GWLR)

As described in the methodology, four models using the GWLR were developed, one for each weighting function shown in Table 1. The predictive variables used were those selected by the logistic regression model, shown in Table 6.

The best model using GWLR, following the AICc criterion, was the Adaptive Gaussian model, with a value of 2,022 closest neighbors to estimate the adaptive bandwidths.

Table 7 contains the descriptive statistics of the coefficients estimated by the GWLR model, in which it is noted that the averages for the coefficients were very close to the coefficients for the global model presented in Table 6.

Table 7 Statistics of the coefficients estimated in the Gaussian Adaptive GWLR model.

Variable	Mean	Standard Deviation	Minimum	Maximum	Range	Q1	Median (Q2)	Q3
Intercept	-1.2950	0.0432	-1.3923	-1.2006	0.1917	-1.3201	-1.2847	-1.2689
d_age1	-0.6557	0.1193	-1.0145	-0.4850	0.5295	-0.7164	-0.6283	-0.5676
d_age2	-0.3230	0.0950	-0.4969	-0.1507	0.3462	-0.3586	-0.3319	-0.2660
d_age4	0.0749	0.0760	-0.0987	0.2164	0.3151	0.0272	0.0616	0.1320
d_age5	0.5054	0.0696	0.3130	0.5910	0.2780	0.4852	0.5275	0.5605
d_education4	0.3004	0.0376	0.2124	0.3518	0.1394	0.2851	0.2979	0.3347
d_time_rel1	-0.6720	0.1019	-0.8264	-0.4858	0.3406	-0.7626	-0.6894	-0.5817
d_time_rel2	-0.3436	0.0513	-0.4208	-0.2314	0.1894	-0.3716	-0.3465	-0.3213
d_time_rel4	0.4614	0.0543	0.3498	0.5573	0.2075	0.4393	0.4430	0.5201
d_income1	0.3272	0.0732	0.2173	0.4769	0.2596	0.2680	0.3222	0.3638
d_income2	0.1255	0.0443	0.0247	0.1791	0.1544	0.0996	0.1469	0.1669
d_pd_contract1	-0.6241	0.1160	-0.7555	-0.3766	0.3789	-0.7183	-0.6849	-0.5065
d_pd_contract2	-0.4134	0.0332	-0.4516	-0.3327	0.1189	-0.4479	-0.4177	-0.3904

Source: Prepared by the authors.

Table 8 contains the final formula of the models estimated via Adaptive Gaussian GWLR for the 19 regions in the DF.

Table 8 Formulas of Local Regressions estimated via the Adaptive Gaussian GCLR model.

Region	Intercept	d_age1	d_age2	d_age4	d_age5	d_education4	d_time_rel1	d_time_rel2	d_time_rel4	d_income1	d_income2	d_pd_contract1	d_pd_contract2
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
BRAZILÂNDIA	-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

Note. NS: Coefficient not significant with 90% confidence (p-value above 0.10).

Source: Prepared by the authors.

It is noted in Table 8 that the Intercept was significant for all the regions in the Distrito Federal and varied from -1.3922 to -1.2005, indicating a regional difference between the values estimated.

With relation to the borrower's age, the variables *d_age1* and *d_age5* were significant for all of the regions in the Distrito Federal, whereas the variables *d_age2* and *d_age4* were not significant for some regions, indicating that the borrower's age influences risk differently, depending on the region studied.

The *d_education4* variable was also significant for all of the regions in the Distrito Federal, presenting a small variation in coefficients between the regions.

With relation to the borrower's Time of Relationship with the institution, the variables *d_time_rel1* and *d_time_rel4* were significant for all of the regions in the Distrito Federal, whereas the *d_time_rel2* variable was not significant for the Cruzeiro region.

With relation to the borrower's Income, the *d_income1* variable was significant for all of the regions in the Distrito Federal, whereas the *d_income2* variable was significant only for the regions of Candangolândia, Gama, Núcleo Bandeirante, Recanto das Emas, Riacho Fundo,

Samambaia, Santa Maria, and Taguatinga, indicating that the borrower's Income also influences credit risk differently between the regions.

The variables *d_pd_contract1* and *d_pd_contract2*, which represent the Loan Contract Period, were significant for all of the regions in the Distrito Federal.

3.5. Comparison Between the Models

The comparison between the Logistic Regression model and the GWLR Adaptive Gaussian model was made using the following metrics: International AICc Criterion, Accuracy, Percentage of False Positives, Sum of Value of False Positive Debt, and Expected Monetary Value of Defaults in the portfolio compared with the monetary value of defaults observed.

Except for the AICc informational criterion, calculated when developing the model, the other metrics were calculated based on the validation database, composed of 11,188 records.

Table 9 shows the descriptive statistics for the scores obtained by both the models selected in the validation sample.

Table 9 Descriptive Analysis of Model Scores.

Model	Mean	Minimum	Q1	Median (Q2)	Q3	Maximum	Range
LR	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

Source: Prepared by the authors.

The means for the model scores were very close, with a difference only in the third decimal place; however, the model using GWLR presented a greater range of scores. The use of few predictive variables meant that the scores produced by the models did not present values greater than 0.585 and 0.639.

To calculate the confusion matrix, a cut-off point had to be defined in terms of score, so that borrowers could be classified as good or bad (0 or 1). This cut-off point was defined based on the shortest distance between Sensitivity and Specificity and its value was 0.30.

Table 10 Confusion Matrix of the models using LR.

		Value Observed LR		Value Observed GWLR	
		0	1	0	1
Value Predicted	0	48.7%	11.3%	49.0%	11.2%
	1	24.0%	16.0%	23.8%	16.0%

Source: Prepared by the authors.

It can be noted in Table 10 that the models presented very close results with regards to client classification.

Table 11 contains all of the metrics used for comparison between the models, in which a small difference is noted

between the values of the indicators in the two models.

Table 11 Comparison between the LR and GWLR models

Model	AICc	Accuracy	% FP	Sum of Debt Value FP	Expected Default Value
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

Source: Prepared by the authors.

In Table 11, all of the values obtained for the metrics of the two models were also very close, with the model using GWLR being the one with the best (lowest) AICc informational criterion and best (highest) Accuracy, which indicates a better percentage of hits and lower percentage of False Positives. The model using LR was slightly higher in the metrics Sum of the Value of False Positives - this metric can be considered as an estimate

of the monetary value that would be granted and enter into default, resulting in financial loss for the institution - and Expected Value of Defaults, since the sum of the value of debt of all of the contracts in default ($Y=1$) in the validation database of the model was R\$ 12,026,290.09, and the value that comes closest is the value from the model using LR.

4. CONCLUSION

In this article, real data were used from a Brazilian financial institution on transactions involving Consumer Direct Credit, granted to clients residing in 19 regions in the Distrito Federal, to develop credit scoring models using two different methodologies: Logistic Regression and Geographically Pondered Logistic Regression.

The Logistic Regression methodology is quite widespread in the financial sector, and is used in this study to develop a global credit scoring model for the whole Distrito Federal.

The Geographically Weighted Logistic Regression methodology is quite rare and uses the borrower's geographical location to weight observations when developing different models for each region studied.

The indicators used for comparison between the models developed via the two methodologies were very close, and based on the results obtained, the methodologies can be considered as similar in terms of their power to predict financial losses for the institution.

The study demonstrated that some variables were significant for all of the regions, whereas others were significant only for particular regions, concluding that credit risk is influenced by different factors, depending on the region studied.

It was also observed that all of the regression models developed using GWLR (regional models) presented different values for the coefficients (parameters) of the variables, showing that the weights (importance) of the

variables varied from region to region.

The results demonstrated the viability of applying the GWLR methodology for developing credit scoring models for the target population in this study. The formulas obtained are applicable only to this population, however, it is believed that this methodology could be extended to other credit transactions and spatial levels (e.g. neighborhoods, municipalities, federal units).

Due to great advances in computing and technology occurring in recent decades, institutions granting credit have robust credit risk evaluation systems, which makes the implantation and use of a set of models estimated via GWLR viable.

With relation to the limitations of the study, the use of few predictive variables meant that the models presented low ranges of scores.

Categorization of the Formal Income variable was carried out so that the classes were monotonic with relation to relative risk; however, the values of their coefficients were inverted. Studies considering another categorization or target population should be carried out to verify the relevance of this variable for credit risk.

For future study topics, it is suggested that: the GWLR methodology is applied to develop credit scoring models for other target populations (for example, different credit transactions or geographical regions); comparisons are carried out with other methodologies (such as Support Vector Machines or Boosting); other predictive variables

are used; the GWLR methodology is applied to develop models in other areas of a financial institution, such as strategy and marketing; or other functions are used, such

as the Log Binomial, to develop geographically weighted models.

REFERENCES

- 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. Retrieved from <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.

Address for Correspondence: 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 – Brazil
Email: pedroa@unb.br