

Discriminação e Classificação da Exclusão Financeira a partir da Análise de Agrupamentos e Redes Neurais

José Leonardo Ribeiro Macrini

Professor

Universidade Federal Rural do Rio de Janeiro – ITR/UFRRJ

E-mail: lm.macrini@gmail.com.

Roberto Salvador Santolin

Professor

Universidade Federal Rural do Rio de Janeiro – ITR/UFRRJ

E-mail: lm.macrini@gmail.com.

Cinthia Rodrigues de Oliveira

Doutoranda

Universidade Federal Fluminense – UFF

E-mail: cin.roliveira@gmail.com

Teofilo Henrique Pereira de Paula

Professor

Universidade Federal Rural do Rio de Janeiro – ITR/UFRRJ

E-mail: teohpaula@gmail.com.

Resumo: A identificação da exclusão financeira está muitas vezes sujeita a elevado grau de subjetividade. Entretanto, o critério de separação, entre excluídos e não excluídos, é uma questão pouco explorada. Frequentemente, o limite que separa as duas categorias tem caráter difuso e subjetivo. O presente trabalho tem como objetivo propor uma metodologia, baseada em Análise de Agrupamentos e Redes Neurais, para a delimitação em questão. Utiliza-se, para tanto, dados de uma pesquisa empírica realizada no município de Três Rios (Brasil). Espera-se que esta investigação possa contribuir para o delineamento e execução de políticas voltadas para a promoção da inclusão financeira e bancarização.

Palavras-chave: Exclusão Financeira, Classificação, Redes Neurais

Abstract: *Identifying the limit between financially excluded and non-excluded is not an easy task, since the object tends to be diffuse and subjective. This study aims to propose a methodology, based on Cluster Analysis and Neural Networks, for the referred delimitation. It makes use of a field research carried out in the municipality of Três Rios (Brazil). It is expected that this research may contribute to the design and execution of policies aimed at promoting the financial inclusion and bancarization.*

Key words: *Financial Exclusion, Classification, Neural Networks*

JEL Code: C1, G2, R1

INTRODUÇÃO

O acesso a bens e serviços financeiros por parte de uma população é antes de tudo uma questão de cidadania. Não obstante, uma deficiência neste quesito tem diversas implicações que afetam o bem-estar econômico, particularmente no que se refere à economia local. Primeiro, representa uma ineficiência alocativa, dado que a exclusão financeira impede que recursos e capacitações locais sejam utilizados em sua plenitude. Segundo, em economias de crescimento acelerado pode ensejar uma fragilidade financeira que se manifesta no mercado de ativos local, particularmente no mercado imobiliário, uma vez que debilita o processo de financiamento, isto é, a canalização das novas rendas para a geração de novos ativos. Sem esta última, os ativos locais já existentes tendem a uma supervalorização nominal, constituindo ameaça à solidez da estrutura patrimonial de famílias e de empresas, bem como impondo dificuldades ao planejamento e à organização do espaço urbano.

Muitos trabalhos empíricos identificam o problema e descrevem suas principais características em determinados espaços econômicos. Entretanto, a definição de um critério de separação entre excluídos e não excluídos constitui uma questão em aberto, uma vez que o limite que separa as duas categorias tem caráter difuso e subjetivo. Neste sentido, o presente trabalho tem como objetivo propor uma metodologia, baseada em Análise de Agrupamentos e Redes Neurais, para a delimitação em questão. Utiliza-se, para tanto, dados de uma pesquisa empírica realizada no município de Três Rios – RJ (Brasil), a qual teve como foco não somente a identificação do referido fenômeno, mas também as suas implicações para o desenvolvimento local. O tema do desenvolvimento encontra-se, assim, subjacente à discussão metodológica de delimitação da condição de exclusão. Espera-se que esta investigação possa contribuir para o delineamento e execução de políticas públicas voltadas para a mitigação do problema da exclusão financeira.

Após esta breve introdução, o trabalho encontra-se dividido da seguinte forma: a seção 1 faz uma breve discussão sobre o tema da exclusão financeira e sua pertinência quanto ao desenvolvimento local. A seção 2 apresenta a metodologia proposta. Na terceira seção são apresentados os dados utilizados, com destaque para os resultados da pesquisa empírica realizada no município de Três Rios. Segue-se, na quarta seção, a aplicação da metodologia proposta, a descrição dos resultados obtidos e a interpretação dos mesmos. Por fim, são tecidas algumas considerações a título de conclusão.

1. EXCLUSÃO FINANCEIRA E O FINANCIAMENTO DO DESENVOLVIMENTO LOCAL

Nos últimos vinte anos, a pesquisa sobre as relações entre moeda e desenvolvimento regional e urbano experimentou grandes transformações, combinando períodos de euforia e marasmo. No início dos anos 1990, há um ressurgimento do interesse na temática relativa ao papel da moeda e do sistema financeiro na configuração dos territórios, especialmente entre os geógrafos econômicos.

No entanto, dois fatores merecem destaque, todos com implicações territoriais importantes: i) as profundas transformações tecnológicas ocorridas no sistema

financeiro, que possibilitaram não somente o armazenamento de uma quantidade sem precedentes de informações, como também que volumes enormes de recursos financeiros fossem transferidos de um lado para outro do planeta quase que instantaneamente combinada e realimentada pelo processo de liberalização financeira ocorrido em quase todos os países do mundo, reforçado e respaldado pela frenética movimentação de capitais entre países; e ii) o crescente papel da dimensão financeira nas crises econômicas recentes. Este quadro praticamente impôs aos acadêmicos especializados em economia regional e/ou geografia econômica a necessidade de entender melhor o papel da moeda e do sistema financeiro na configuração espacial da economia.

Diversas linhas de investigação foram desenvolvidas ao ponto de Martin (1999) afirmar, ao final dos anos 1990 que uma nova subdisciplina da geografia econômica havia surgido. Dentre estas, no leito da investigação sobre o processo de financeirização e suas consequências espaciais, emerge a discussão sobre exclusão financeira. Segundo Leyshon (2008, p. 447), a exclusão financeira se refere a *“those processes by which individuals and households face difficulties in accessing financial services”*. Posteriormente Dymski (2007a, 2007b) ampliou este conceito ao incluir não somente o acesso, mas também o uso de serviços financeiros que, muitas vezes, se verifica de forma precária e/ou predatória.

Tendo em vista a financeirização das atividades em geral e cotidianas, que acabam por imprimir um conteúdo financeiro na grande parte das relações sociais, a exclusão financeira ocasiona consequências no próprio circuito da exclusão social, reforçando-a ou mesmo induzindo-a. Essa exclusão não pode ser caracterizada de forma homogênea e absoluta, mas deve ser vista dentro do contexto de diversas formas de participação social e econômica, o que definirá a natureza da demanda por serviços financeiros sejam estes meios de pagamento, poupança, seguro ou crédito (Crocco, 2014). Reconhece-se, portanto, que as necessidades de inclusão financeira têm natureza heterogênea, de acordo com as especificidades do modo de inclusão socioeconômica em cada espaço regional. Por exemplo, microagricultores no sertão nordestino têm demanda financeira distinta de microempresários populares de regiões metropolitanas, assim como estas diferem da demanda dos jovens, dos idosos e dos profissionais do lar, em cada uma dessas localidades. Em consequência disso, as ações que compõem um projeto de inclusão financeira devem ser capazes de reconhecer essa heterogeneidade.

Outro aspecto importante dos sistemas econômicos contemporâneos diz respeito às instabilidades financeiras que têm acometido países e regiões com frequência crescente. A este respeito, uma das principais contribuições teóricas encontra-se na Hipótese da Instabilidade Financeira (HIF) (Minsky, 1986), segundo a qual a robustez financeira que caracteriza a economia no início do processo de expansão, dará lugar ao longo do tempo a um sistema cada vez mais frágil, resultando invariavelmente em crises financeiras, com impactos diretos sobre a renda e o emprego. Dymski (1998), entretanto, ressalta que o trabalho de Minsky é a-espacial, isto é, supõe implicitamente que o ciclo se desenvolve em uma economia fechada e homogênea. Argumenta que os desequilíbrios comerciais e os fluxos de fatores transfronteiriços devem ser levados em consideração, ou seja, que a fragilidade financeira pode ser intensificada por

desequilíbrios nos fluxos de poupança e fatores entre regiões/países. A possibilidade de emergência de uma crise financeira, bem como a capacidade de reação a essa depende não apenas do ciclo econômico, mas do sucesso da referida economia em mediar as tensões entre os setores financeiro e real inerentes a economias com fronteiras. Tal mediação dependeria, particularmente, da forma como o sistema financeiro se estrutura na referida região de modo a contribuir para a canalização dos influxos de poupanças e reservas para o investimento. O autor argumenta que as diferenças entre países/regiões quanto às “condições de fronteira” e estrutura do setor bancário implicam a necessidade de se espacializar a HIF.

Nesta perspectiva, o ciclo minskyano pode ser caracterizado como um processo de distanciamento entre a valorização nominal e o ritmo de crescimento da oferta de novos ativos reais. A capacidade de reação a este processo depende de quão bem posicionada está a economia para criar ativos reais, isto é, para transformar a entrada de recursos em novos ativos ao invés de meros direitos de propriedade. Uma economia de *boom* é caracterizada por elevados influxos de recursos e quando o sistema financeiro interno falha na transformação deste influxo em ativos reais estabelecem-se as bases para a formação de uma “economia de bolha”.

O arcabouço introduzido por Dymksi e Veitch (1996) destaca o papel do fator terra, uma vez que o crescimento econômico invariavelmente envolve o crescimento urbano. Os autores destacam as particularidades dos ativos geográficos, isto é, os bens físicos de longa vida útil, como as construções residenciais e não residenciais. O investimento nestes ativos tem como característica a irreversibilidade e a presença de externalidades de área (*spillovers*), isto é, da presença ou não de outros ativos na área. Tais aspectos constituem fatores de incerteza e, portanto, combustível para o exercício de atividades especulativas. O financiamento do investimento em ativos geográficos depende não somente das características do tomador como também envolve considerações sobre a área em que o projeto se localizará. Emprestar com sucesso para ativos geográficos implica não apenas comprometer recursos com a obtenção de informações, como também, a coordenação do empréstimo com empréstimos realizados por concorrentes na área de modo a capturar as externalidades positivas e evitar as negativas. Assim, a acumulação de ativos geográficos suficiente para acompanhar os influxos de riqueza depende de quão bem coordenada é a estrutura financeira, de modo a tornar mais atrativo o investimento em ativos reais *vis-à-vis* os ativos financeiros.

Por fim, a relação entre a fragilização financeira e a exclusão financeira é estabelecida em Dymksi (2007b). A partir do princípio de que a exclusão financeira não se refere somente à falta de acesso, mas também a um acesso precário ou predatório aos bens e serviços financeiros, o autor argumenta que nos últimos anos tem ocorrido uma mudança de estratégia no setor bancário, resultando em uma segmentação no mercado de crédito. Mais especificamente, a ideia de um mercado de crédito representativo não mais se aplica. Potenciais tomadores de crédito invariavelmente serão alocados em pelo menos dois grupos distintos. No primeiro se encontrariam aqueles cujos ativos e posições financeiras são considerados seguros, ou seja, aqueles a quem se atribuiria um “bom risco” e que as instituições financeiras gostariam de estabelecer relações sustentáveis de longo prazo. No outro grupo (os financeiramente

excluídos, segundo a conceituação mais ampla) se encontrariam aqueles cujos níveis de renda seriam suficientemente baixos ao ponto de que os contratos seriam firmados com a esperança única de se obter o máximo de retorno no curto prazo de modo a compensar os problemas de insolvência que inevitavelmente surgirão ao longo do tempo. O caso do *subprime* no mercado imobiliário americano, que acabou por desencadear a crise de 2008 constitui um exemplo oportuno.

Considerando-se toda a complexidade e incertezas que envolvem o desenvolvimento urbano, parece claro que mercados de crédito com forte presença de demandantes de crédito do segundo grupo dificultará o processo de criação de novos ativos. Assim, quando este cenário se verifica em economias de rápido crescimento o potencial de geração de bolhas de ativos é elevado, colocando em risco a solidez da estrutura patrimonial de famílias e empresas, bem como precarizando a evolução do espaço urbano. É neste ponto que a devida delimitação da condição de exclusão financeira faz-se pertinente, particularmente por permitir a identificação da população alvo de políticas econômicas que objetivem a prevenção de instabilidades econômicas, o desenvolvimento urbano e o exercício pleno da cidadania.

2. METODOLOGIA PARA IDENTIFICAÇÃO E CLASSIFICAÇÃO DA EXCLUSÃO FINANCEIRA

A metodologia proposta no presente trabalho consiste em cumprir três etapas, a saber: i) análise de agrupamentos – separar, entre a população de Três Rios, indivíduos excluídos financeiramente ou não no processo de discriminação; ii) seleção de variáveis – separar variáveis pertinentes das redundantes e irrelevantes para o processo de classificação e; iii) redes neurais – classificar indivíduos “fora da amostra” em excluídos financeiramente ou não. Abaixo segue uma discussão mais detalhada de cada uma dessas etapas.

2.1. Análise de agrupamentos

A análise de agrupamentos, também conhecida como análise de conglomerados ou *cluster*, tem como objetivo dividir os elementos da amostra, ou população, em grupos de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (características) que neles forem medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características (HAIR JR. *et al*, 2009).

Dessa forma, é necessário definir um critério a ser utilizado para decidir até que ponto dois elementos de um conjunto de dados podem ser considerados como semelhantes ou não. É necessário considerar medidas que descrevam a similaridade entre elementos amostrais de acordo com as características que neles foram medidas. Ao considerar que para cada elemento amostral têm-se informações de p-variáveis armazenadas em um vetor, a comparação de diferentes elementos amostrais poderá ser feita por meio de medidas matemáticas (métricas), que possibilitem a comparação de

vetores, como as medidas de distância. Assim, é possível calcular as distâncias entre os vetores, de observações dos elementos amostrais e agrupar aqueles de menor distância.

No presente artigo a amostra está separada em duas categorias, quais sejam, excluídos e não excluídos financeiramente. Assim, opta-se pelo método não hierárquico das *k*-Médias (JOHNSON; WICHERN, 2002). Basicamente, cada elemento amostral é alocado ao *cluster* cujo centróide (vetor de médias amostral) é o mais próximo do vetor de valores observados para o respectivo elemento. O método é composto de quatro passos:

- i) Primeiramente escolhem-se *k* centróides, chamados de “sementes” ou “protótipos”, para se inicializar o processo de partição;
- ii) Cada elemento do conjunto de dados é, então, comparado com cada centróide inicial, por meio de uma medida de distância Euclidiana. O elemento é alocado ao grupo cuja distância é a menor;
- iii) Após o passo ii), para cada um dos *n* elementos amostrais, recalcula-se os valores dos centróides para cada novo grupo formado, repete-se o passo ii), considerando os centróides destes novos grupos;
- iv) Os passos ii) e iii) devem ser repetidos até que todos os elementos amostrais estejam “bem alocados” em seus grupos, isto é, até que nenhuma realocação de elementos seja necessária.

2.2. Seleção de variáveis baseada em Informação Mútua sob Distribuição de Informação Uniforme (MIFS-U)

A seleção de variáveis de entrada desempenha um importante papel em sistemas de classificação tais como redes neurais artificiais (RNAs). Tais variáveis podem ser classificadas como pertinentes, irrelevantes ou redundantes e do ponto de vista do gerenciamento de um conjunto de dados, que pode ser extremamente grande, reduzir o número de variáveis, selecionando somente aquelas pertinentes, é extremamente desejável. Desse modo, melhor desempenho com menor esforço computacional é esperado (KWAK; CHOI, 2002). Além disso, Hosmer & Lemeshow (1989) destacam a importância da seleção de variáveis, pois, com um menor número de variáveis, o modelo tende a ser mais generalizável e robusto.

Problemas de seleção de variáveis foram pesquisados por vários autores como Battiti (1994), Joliffe (1986) e Agrawal *et al.* (1993). Um dos métodos mais populares para lidar com este problema é o de análise de componentes principais (*principal component analysis* – PCA) (JOLIFFE, 1986). Entretanto, quando se objetiva preservar os dados originais, este método não é adequado.

O algoritmo abordado neste artigo, a saber, o MIFS-U – Seletor de Variáveis sob Informação Mútua com Distribuição Uniforme de Informação – apresentado por Kwak & Choi (2002), objetivou superar a limitação do seletor de variáveis proposto por Battiti (1994), gerando melhor desempenho no processo de seleção de variáveis. Tal algoritmo, por sua simplicidade, pode ser usado em qualquer sistema de classificação, seja qual for

o algoritmo de aprendizagem. No entanto, seu desempenho pode ser prejudicado como resultado de erros na estimação da informação mútua.

2.2.1. Problema FR_n-k e o Algoritmo de Seleção Ideal

No processo de selecionar variáveis de entrada, é desejável reduzir o número de variáveis, excluindo aquelas que são irrelevantes ou redundantes. Este conceito é formalizado pela seleção das k variáveis mais relevantes de um conjunto de n variáveis, chamado de problema de “redução de variável” (*feature reduction* – FR) (BATTITI, 1994). Tal processo é descrito a seguir:

[$FR_n - k$]: Dado um conjunto inicial de n variáveis, encontre o subconjunto com $k < n$ variáveis que representa “a máxima informação” acerca do desfecho (variável explicada, de resultado ou de saída). O problema de selecionar variáveis de entrada pode ser resolvido calculando a informação mútua (IM) entre variáveis de entrada e o desfecho. Se a informação mútua entre variáveis de entrada e desfecho pudesse ser obtida com exatidão, o problema $FR_n - k$ poderia ser reformulado como segue:

[$FR_n - k$]: Dados um conjunto inicial F com n variáveis e o desfecho D , encontre o subconjunto $S \subset F$ com k variáveis que minimiza $H(D|S)$, isto é, que maximiza a informação mútua $I(D;S)$. O método de seleção adotado aqui é conhecido como “seleção gulosa”. Neste método, a partir do conjunto vazio de variáveis selecionadas, adiciona-se a melhor variável de entrada viável ao conjunto anterior, uma a uma, até que o tamanho do conjunto atinja k . Esse algoritmo de seleção ideal que usa informação mútua é realizado como segue:

- 1) (Inicialização) conjunto $F \leftarrow$ “conjunto inicial com n variáveis”, $S \leftarrow$ “conjunto vazio.”
- 2) (Cálculo da IM com o desfecho), $\forall \phi_i \in F$, compute $I(D; \phi_i)$.
- 3) (Seleção da primeira variável) ache a variável que maximiza $I(D; \phi_i)$, faça $F \leftarrow F \setminus \{\phi_i\}$,
 $S \leftarrow \{\phi_i\}$.
- 4) (Seleção gulosa) repita até que seja alcançado o número desejado de variáveis selecionadas:
 - a) (Cálculo da IM conjunta entre variáveis), $\forall \phi_i \in F$, compute $I(D; \phi_i, S)$.
 - b) (Seleção da próxima variável) escolha a variável $\phi_i \in F$ que maximiza $I(D; \phi_i, S)$ e faça $F \leftarrow F \setminus \{\phi_i\}$, $S \leftarrow \{\phi_i\}$.
- 5) Finalize o conjunto S contendo as variáveis selecionadas.

Na prática, a realização deste algoritmo torna-se inviável, face à alta dimensão do vetor de variáveis no cálculo de $I(D; \phi_i, S)$, visto que, tendo como objetivo a seleção de k ($k < n$) variáveis, o vetor S (composto pelas variáveis já selecionadas), atinge dimensão igual a $(k - 1)$.

2.2.2. O Seletor de Variáveis MIFS-U

O algoritmo ideal tenta maximizar $I(D; \phi_i, \phi_s)$, e isto pode ser reescrito como

$$I(D; \phi_i, \phi_s) = I(D; \phi_s) + I(D; \phi_i | \phi_s). \quad (1)$$

No entanto, calcular $I(D; \phi_i | \phi_s)$ requer tanto esforço quanto calcular $I(D; \phi_i, \phi_s)$. Assim, $I(D; \phi_i | \phi_s)$ é calculada de forma aproximada através de $I(\phi_i; \phi_s)$ e $I(D; \phi_i)$, que são relativamente fáceis de calcular. A informação mútua condicional pode ser representada como:

$$\begin{aligned} I(D; \phi_i | \phi_s) \\ = I(D; \phi_i) - \{I(\phi_i; \phi_s) - I(\phi_i; \phi_s | I)\} \end{aligned} \quad (2)$$

E dessa forma, o passo 4 revisado do algoritmo de seleção ideal toma, assim, a seguinte forma:

4) (Seleção gulosa) repita até que seja alcançado o número desejado de variáveis selecionadas:

- a) (Cálculo da entropia) $\forall \phi_s \in S$, compute $H(\phi_s)$, se ainda não foi avaliada.
- b) (Cálculo da IM entre variáveis), para todos os pares de variáveis (ϕ_i, ϕ_s) com $\phi_i \in F$ e $\phi_s \in S$, compute $I(\phi_i; \phi_s)$, se ainda não foi avaliada.
- c) (Seleção da próxima variável) escolha a variável $\phi_i \in F$ que maximiza $I(D; \phi_i) - \beta \sum_{\phi_s \in S} (I(D; \phi_s) / H(\phi_s)) I(\phi_i; \phi_s)$ e faça $F \leftarrow F \setminus \{\phi_i\}$, $S \leftarrow \{\phi_i\}$.

O parâmetro β oferece flexibilidade ao algoritmo. Caso se adote β igual a zero, a informação mútua entre as variáveis de entrada não é levada em consideração e o algoritmo seleciona tais variáveis na ordem da informação mútua de cada uma delas com o desfecho. A redundância entre as variáveis de entrada não é, portanto, refletida. Quando β aumenta, as variáveis redundantes são excluídas mais eficientemente. Em geral, pode-se tomar $\beta = 1$ (BREIMAN *et al.*, 1984). Caso este em que há um equilíbrio, em termos de peso, entre a redundância da variável candidata e a informação mútua desta com o desfecho.

Kwak & Choi (2002) salientam que o algoritmo MIFS-U pode ser aplicado a problemas complexos sem excessivo esforço computacional.

A estimação da entropia e da informação mútua está baseado na entropia quadrática de Rényi, aliada ao famoso método de estimação de densidade Janela de Parzen, e na definição da Informação Mútua de Cauchy-Schwartz, tornando os cálculos diretos, sem necessidade de um passo de pré-processamento. Nesse artigo utilizamos esse método que é descrito em detalhes em Gonçalves & Macrini (2011).

2.3. Redes Neurais

Rede Neural Artificial (ou simplesmente “Rede Neural”) é um modelo distribuído composto por unidades (chamadas na literatura de “neurônios”) constituídas de funções não-lineares (tipicamente sigmóides e tangentes hiperbólicas). A combinação destas unidades, através de parâmetros estimados a partir dos dados, é o que confere a capacidade deste modelo de inferir relações não-lineares de complexidade arbitrária. Na forma utilizada neste estudo, as unidades são organizadas em camadas, incluindo uma camada oculta, que não está diretamente conectada à saída do modelo. Estas conexões entre as unidades, ou neurônios, são chamadas de pesos (originalmente a terminologia era “pesos sinápticos”). Estes pesos são os parâmetros do modelo que são ajustados por um algoritmo iterativo através dos dados. Uma vez ajustados os pesos, a rede tem a capacidade de representar a relação dos dados de entrada com a variável de saída. A capacidade de aprender através de “exemplos” ou dados (na amostra) e de generalizar (fora da amostra) a informação gerada em ambientes não-lineares complexos, é sem dúvida a grande vantagem das Redes Neurais.

A Rede implementada utilizou o algoritmo de Regularização Bayesiana (MACKAY, 1992). Neste algoritmo, assume-se que os parâmetros da Rede são variáveis aleatórias com distribuições especificadas. Os parâmetros de regularização são variâncias desconhecidas associadas a estas distribuições e podem-se calcular estes parâmetros utilizando então técnicas estatísticas. Portanto, o modelo não é especificado de uma forma arbitrária.

O aprendizado ou treinamento de uma rede neural tem tipicamente por objetivo reduzir a soma dos quadrados dos erros (FORESEE; HAGAN, 1997), conforme a seguinte equação:

$$\hat{\Psi} = \arg \min_{\Psi} Q_1(\Psi) = \arg \min_{\Psi} \sum_{t=1}^N (y_t - G(x, \Psi))^2 \quad (3)$$

Em que $(x, \psi) \in X \times \Psi$, sendo $x = [x_1, x_2, \dots, x_l]$ vetores de variáveis independentes e ψ o vetor de parâmetros $\psi = [\alpha, \gamma]$, composto pelos vetores de pesos da camada de saída e da camada oculta respectivamente; y_t é a saída alvo da Rede e $G(x, \psi)$ é a saída estimada pela Rede. Assim como outros modelos flexíveis não-lineares, as Redes Neurais podem sofrer de *overfitting*. Este problema ocorre quando é utilizado um número excessivo de neurônios na camada oculta, que levarão a uma perda da capacidade de generalização (fora da amostra). Em contrapartida, se o número de neurônios em excesso for reduzido, ocorre a perda da capacidade de aproximar o processo gerador dos dados (MEDEIROS; PEDREIRA, 2001).

Atualmente, diversas metodologias são utilizadas para solucionar o problema de *overfitting* (HAYKIN, 2001). Neste estudo, será utilizado o procedimento desenvolvido por Mackay (1992), chamado de Regularização Bayesiana, que consiste em adicionar um termo de penalização (regularização) à função objetivo, de forma que o algoritmo de estimação faça com que os parâmetros irrelevantes convirjam para zero, reduzindo assim o número de parâmetros efetivos utilizados no processo.

Seguindo a notação utilizada por Medeiros e Pedreira (2001), o problema de estimação passa a ser definido como:

$$\hat{\Psi} = \arg \min_{\Psi} Q_T(\Psi) = \arg \min_{\Psi} \sum_{t=1}^N (\eta Q_1(\Psi) - Q_2(\Psi))^2 \quad (4)$$

Em que ϕ e η são parâmetros de regularização e $Q_2(\Psi)$ é a função de penalização, que é dada pela soma do quadrado dos parâmetros (α, γ) , vetores de pesos da camada de saída e da camada oculta, respectivamente, conforme a seguinte equação:

$$Q_2(\Psi) = \sum_{h=0}^H \alpha_h^2 + \sum_{h=0}^H \sum_{i=0}^I \gamma_{hi}^2 \quad (5)$$

O problema de regularização é otimizar a função objetivo de forma a encontrar valores para os parâmetros de regularização ϕ e η . Este problema de otimização requer o cálculo da matriz Hessiana como pode ser visto em (MACKAY, 1992). O algoritmo desenvolvido por Foresee e Hagan (1997) propõe a aproximação da matriz Hessiana pelo algoritmo de Levenberg-Marquardt (LEVENBERG, 1944; MARQUARDT, 1963), reduzindo o custo computacional.

Todos os modelos utilizados neste estudo tiveram como arquitetura da rede neural uma camada de entrada, uma camada escondida com dez neurônios e uma camada de saída com um neurônio. A função de ativação tangente hiperbólica foi utilizada em todos os neurônios da camada oculta e a função sigmóide na camada de saída. Os pesos e os *bias* foram inicializados através do algoritmo de Nguyen-Widrow (1989).

3. EXCLUSÃO FINANCEIRA NO MUNICÍPIO DE TRÊS RIOS (BRASIL)

Para o desenvolvimento do trabalho as seguintes fontes de informações foram utilizadas: i) fontes secundárias para a obtenção de informações sobre renda, população, indicadores de desenvolvimento, estrutura etária, dentre outras. Particularmente, utilizou-se o sítio do Instituto Brasileiro de Geografia e Estatística – IBGE; ii) fonte primária, por meio de pesquisa de campo (DE PAULA, 2015), com aplicação de questionário¹. Este último tem como base aquele utilizado em pesquisa do Banco Mundial (World Bank, 2005), acrescido de questionamentos que incluem características próprias da economia local.

Os dados da pesquisa foram extraídos de uma população de mais de 70.000 pessoas dentro do perímetro urbano Município de Três Rios. A população-alvo são todas as pessoas com faixa etária de 15 anos ou mais. A amostra é dada por 386 questionários individuais, considerando um erro de 5%, com coeficiente de confiança de 95%. O levantamento inicial foi realizado ao longo do mês de maio de 2015. A seleção da amostra foi baseada na probabilidade proporcional ao tamanho dos bairros e regiões da cidade. Levou-se em conta também a distribuição da população em relação ao gênero, idade, escolaridade e classe de rendimentos. Para estratificação de classes sociais em A, B, C, D

¹ O questionário da pesquisa pode ser acessado na página do Grupo de Pesquisa em Moeda e Desenvolvimento, da Universidade Federal Rural do Rio de Janeiro (ITR/UFRRJ), disponível em: <https://gmdesenvolvimento.wixsite.com/grupo/pesquisas-em-andamento>.

e E aplicou-se o “Critério Brasil”, de acordo com as definições estabelecidas em ABEP (2015).

Os questionários foram aplicados a partir das zonas de classificação, obedecendo à distribuição do Censo 2010 (Instituto Brasileiro de Geografia e Estatística – IBGE). Para a definição do plano amostral foram reunidas informações a partir do cadastro básico agregado no Setor Censitário que permite o menor recorte territorial de áreas urbanas e rurais, de modo a captar as características específicas das respectivas zonas e de aplicá-las à totalidade da população.

A seleção das variáveis visa a captar as várias formas de exclusão financeira, as quais são sintetizadas por Anderloni *et al.* (2007), tal como se segue:

- i) **Acesso geográfico:** refere-se à existência ou não de agências e/ou postos de atendimento em uma determinada área geográfica;
- ii) **Exclusão de acesso:** restrição de acesso a serviços bancários resultantes de processos de avaliação de risco por parte dos bancos (custo de avaliação pode ser elevado - *redlining*);
- iii) **Exclusão por condicionantes:** situação onde os condicionantes atrelados a produtos ofertados fazem com que estes não atendam às necessidades de alguns grupos de clientes;
- iv) **Exclusão por preço:** incapacidade de determinados grupos de indivíduos terem acesso a serviços financeiros, devido ao valor de comissões e/ou tarifas associadas a produtos e serviços;
- v) **Exclusão por marketing:** situação onde alguns segmentos do mercado são excluídos pela forma como vendas e publicidades são direcionadas; e
- vi) **Auto Exclusão:** situação onde parcela da população se recusa a procurar bancos porque acreditam que suas demandas por produtos e serviços serão recusadas.

A partir da classificação acima, o Quadro 1 apresenta: i) na primeira coluna, as variáveis que foram inicialmente consideradas pertinentes (perguntas do questionário) para a Análise de Agrupamentos (*clusters*); e ii) na segunda coluna, o resultado do teste de significância estatística de Mann-Whitney para duas amostras independentes, ao nível de 1% de probabilidade, referente às respectivas variáveis.

Quadro 1: Variáveis e respectiva significância estatística

Variáveis (Perguntas do Questionário)	Apresenta significância estatística?
1. Está empregado no momento?	Sim
2. Seu trabalho é formal?	Sim
3. Qual a sua condição de moradia?	Não (p -valor<0.140)
4. Possui conta corrente?	Sim
5. Qual instituição financeira, ou canal de relacionamento, você considera mais importante?	Sim
6. Realizou algum tipo de empréstimo ou obtenção de crédito nos últimos 12?	Não (p -valor<0.260)
7. Qual a sua principal fonte de renda atualmente?	Sim
8. Como você recebe a sua renda?	Sim
9. Como você paga as suas despesas?	Sim
10. A água utilizada na sua casa é proveniente de:	Não (p -valor<0.999)
11. Qual o tipo de calçamento da rua onde você reside?	Não (p -valor<0.811)

Fonte: Elaboração própria com o uso do software SPSS.

4. RESULTADOS OBTIDOS A PARTIR DA APLICAÇÃO DA METODOLOGIA PROPOSTA

A partir dos elementos discutidos na seção anterior a Análise de Agrupamentos apresentou o seguinte resultado:

Tabela 1: Resultado da Análise de Agrupamentos

	Frequência	Percentual
Excluídos	222	72,8
Não Excluídos	83	27,2
Total	305	100,0

Fonte: Elaboração própria com o uso do software SPSS.

Segundo esta divisão, podemos observar isoladamente algumas variáveis que se mostram bastante significativas na discriminação de indivíduos que podem ser classificados como excluídos e não excluídos. A Tabela A1 (ver Anexo 1) apresenta as estatísticas referentes a estas variáveis.

Encontrou-se diferença significativa em relação às questões estar empregado, trabalho formal, possui conta correte, canal de relacionamento importante, fonte de renda e como paga a suas contas (p <0.001). Não houve diferença significativa nas questões sobre condições de moradia (p <0.140), realização de empréstimos nos últimos 12 meses (p <0.260), água utilizada na residência (p <0.999) e o calçamento da tua (p <0.811), o que não prejudica a análise pois o questionário não especifica a qualidade desses serviços.

Após identificar cada indivíduo do banco de dados como sendo excluído financeiramente ou não, o banco de dados ficou formado por 18 variáveis de entrada. Como variável de saída tem-se o *Cluster* de identificação de um indivíduo como sendo excluído (0) ou não excluído (1) financeiramente. Ao utilizarmos o algoritmo MIFS-U para ordenar essas variáveis por ordem de importância, com o Desfecho obteve-se o resultado do Quadro 2 abaixo.

Cabe ressaltar que este resultado deve ser analisado como um vetor multivariado e sua leitura devem ser feita, conjuntamente, pela ordem de importância. As variáveis Idade e Escolaridade são duas variáveis que não são mencionadas na literatura e que são identificadas pelo método não linear aplicado com poder discriminante com a região onde o indivíduo mora.

Quadro 2: Variáveis Seleccionadas por Ordem de Importância com o Desfecho

Ordem	Variável
1	Idade
2	Região
3	Escolaridade
4	Estar Empregado
5	Fonte Renda
6	Principal canal de Relacionamento
7	Sexo
8	Realização de empréstimo 12 Meses
9	Tipo de calçamento rua onde mora
10	Meio de pagamento das despesas
11	Condição de trabalho (formal ou informal)
12	Moradia (alugada ou própria)
13	Critério Brasil
14	Meio pelo qual recebe a renda
15	Possui Conta-Corrente
16	Raça
17	Meio locomoção
18	Bairro

Fonte: Elaboração própria.

Nota: O algoritmo separa as variáveis pertinentes das redundantes. Como uma região engloba vários bairros, note que ele foi capaz de separar essas variáveis.

O banco de dados apresenta um total de 304 casos. Este conjunto de dados foi dividido em um conjunto para treinamento da Rede Neural. Do número total de casos 70% sendo 156 casos de excluídos e 58 casos de não excluídos, num total de 214 casos foram utilizados para o treinamento da rede. Os 30% restante dos dados foram utilizados na generalização com 66 casos de excluídos e 24 casos de não excluídos num total de 90 casos.

Como trabalhamos com 18 variáveis foram realizados 18 experimentos, sendo que no primeiro experimento trabalhamos com uma variável de entrada, no caso “Idade”, no segundo experimento as variáveis de entrada foram “Idade” e “Região”, e assim por diante até o último experimento contendo todas as 18 variáveis de entrada. O melhor resultado *out sample* encontrado foi quando se utilizou as 13 primeiras variáveis seleccionadas pelo algoritmo MIFS-U, isto é: Idade, Região, Escolaridade, Esta Empregado,

Fonte de Renda, Canal de Relacionamento, Sexo, Empréstimo nos últimos 12 Meses, Calçamento da Rua, Forma de Pagamento das Despesas, Trabalho Formal/Informal, Condição de Moradia e Critério Brasil.

A Tabela 2 e as Figuras 1 e 2 abaixo mostram o número de acertos em valor percentual encontrado na generalização:

Tabela 2: Nível de acerto na generalização da Rede

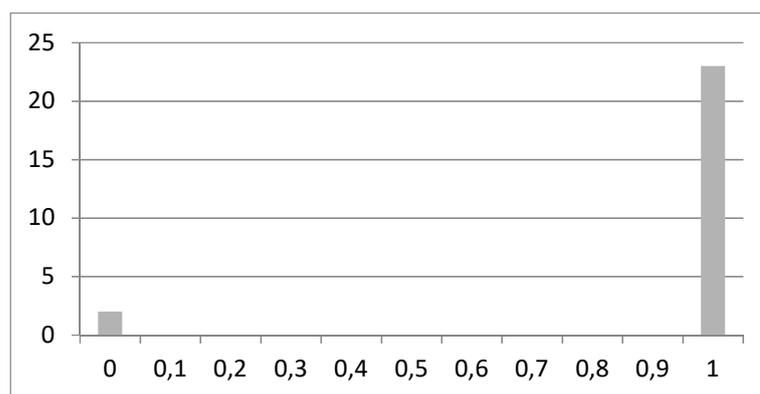
	Previsão		
Realidade	Excluídos	Não Excluídos	Total
Excluídos	84,2	15,8	100,00
Não Excluídos	0,0	100,0	100,00

Fonte: Elaboração própria.

Ou seja, um percentual de acertos de 84,2% para os excluídos e de 100,00% para os não excluídos, com um acerto médio de aproximadamente 93,3%.

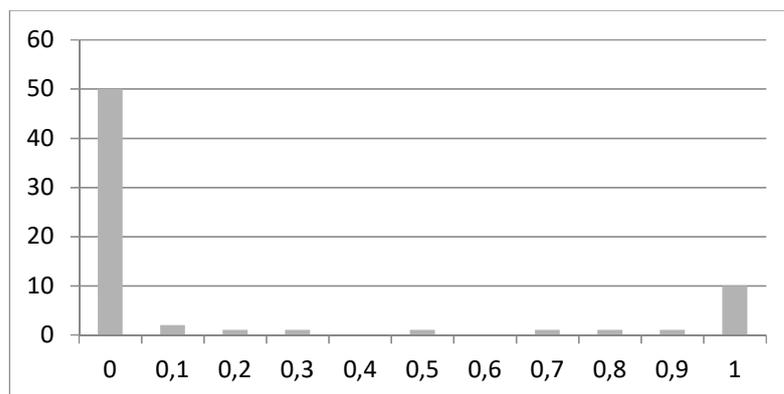
É importante ressaltar também no que diz respeito ao poder discriminante da Rede. O desfecho do modelo é uma variável *dummy* com valor 0 (excluídos financeiramente) ou 1 (não excluídos financeiramente). A saída da Rede é o resultado de uma função sigmóide com valores entre 0 e 1. Adotando-se como ponto de corte 0.5 assume-se que valores menores que 0.5 serão considerados excluídos financeiramente valores maiores que 0.5 serão considerados não excluídos financeiramente. A saída da Rede será mais discriminatória à medida que os valores se aproximem o mais perto possível de 0 ou de 1. Para mostrar este resultado, nos quadros 1 e 2 estão os histogramas da saída efetiva da Rede na generalização. Desta forma, vê-se que a Rede consegue uma boa separabilidade ao classificar a exclusão financeira.

Figura 1: Discriminar para os dados dos não excluídos financeiramente:



Fonte: Elaboração própria

Figura 2: Discriminar para os dados dos excluídos financeiramente:



Fonte: Elaboração própria

CONSIDERAÇÕES FINAIS

O presente trabalho apresenta uma metodologia para a identificação da condição de exclusão financeira contribuindo, dessa forma, para a superação de uma dificuldade que se refere ao caráter difuso da separação entre as condições de excluído e não excluído. Essa ferramenta pode conferir objetividade ao planejamento público e privado na definição de populações alvo de políticas de inclusão financeira. O método é constituído por três etapas: análise de agrupamentos, seleção de variáveis e Redes Neurais, sendo que a última apresentou um elevado grau de acerto. A importância da aplicação de Redes Neurais se verifica ainda no fato de que um novo indivíduo pode imediatamente ser classificado como excluído ou não.

O método revelou ainda, a partir do processo de seleção de variáveis, quais variáveis são de fato importantes para a caracterização da condição financeira dos indivíduos de uma população. Tal fato proporciona uma redução dos esforços de pesquisa de campo ao permitir trabalhar com um questionário significativamente mais simples, baseado apenas nas variáveis selecionadas como estatisticamente relevantes. Não obstante, é importante ressaltar a necessidade de se verificar se estes resultados se repetem em outros espaços econômicos.

Por fim, como uma proposta para desenvolvimentos futuros, o método pode ser aprimorado no sentido de incluir uma terceira categoria, qual seja a dos “incluídos de forma precária”, a qual, segundo alguns autores, constitui uma forma igualmente importante de exclusão financeira.

REFERÊNCIAS BIBLIOGRÁFICAS

ABEP (Associação Brasileira de Empresas de Pesquisa) (2015). “Alterações na aplicação do critério Brasil”. ABEP, São Paulo, Brasil. Disponível em <<http://www.abep.org/criterio-brasil>>. Acesso em maio de 2015.

AGRAWAL, Rakesh; IMIELINSKI, Tomasz; SWAMI, Arun. Database mining: A performance perspective. *IEEE Trans on Neural Networks*, Washington, p. 914-925, dez. /1993.

BATTITI, Roberto, Using mutual information for selecting features in supervised Neural net learning. *IEEE Trans on Neural Networks*, Washington, v. 5, p. 537-550, jul. /1994.

BREIMAN, Leo; FRIEDMAN, Jerome; OLSHEN, Richard; STONE, Charles. *Classification and Regression Trees*. Belmont: Wadsworth, 1984.

CROCCO, Marco. *Globalização, Financeirização, Sistemas Financeiros e seus Rebatimentos no Território*. Laboratório de estudos em moeda e território – LEMTe, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, 2014.

DE PAULA, Teófilo. Exclusão financeira e bolha de ativos em economias de crescimento acelerado: o caso do município de Três Rios (RJ). *Projeto de Pesquisa Faperj APQ1* (Edital 2014). Departamento de Ciências Econômicas – DCEEX, Universidade Federal Rural do Rio de Janeiro – UFRRJ, Três Rios/Brasil, 2015.

DYMSKI, Gary. Exclusão e eficiência: a transformação global do core banking, um estudo para o Brasil. In: DE PAULA, Luis Fernando; ORERO, José Luís. (Org.). *Sistema Financeiro: uma análise do setor bancário brasileiro*. Rio de Janeiro: Elsevier, p. 255-284, 2007a.

DYMSKI, Gary. From financial exploitation to global banking instability: two overlooked roots of the subprime crisis. *Working Paper*, University of California, Sacramento/USA, p. 1-30, dez. / 2007b.

DYMSKI, Gary. Economia de bolha e crise financeira no Leste Asiático e na Califórnia: uma perspectiva espacializada de Minsky. *Economia e Sociedade*, Campinas, p. 73-136, dez. /1998.

DYMSKI, Gary; VEITCH, John. Financial transformation and the metropolis: booms, busts, and banking in Los Angeles. *Environment and Planning*, Oxford, v. 28, n. 7, p. 1233-1260, jul. /1996.

FORESEE, Dan; HAGAN, Martin. Gauss-Newton approximation to Bayesian regularization. *Proceedings of the 1997 International Joint Conference on Neural Networks*. June 9-12. IEEE Ed. Houston, Texas, 1997.

GONÇALVES, Leonardo; MACRINI, Leonardo. Rényi Entropy and Cauchy-Schwartz Mutual Information Applied to MIFS-U Variable Selection Algorithm: A Comparative Study. *Pesquisa Operacional* v.1(3). Ed. Sobrapo, Rio de Janeiro, Brasil, p. 499-519, 2011.

HAIR JR, Joseph. *Análise Multivariada de Dados*. Porto Alegre: Bookman, 2009.

HAYKIN, Simon. *Redes Neurais: Princípios e Prática*. Porto Alegre: Bookman, 2001.

HOSMER, David; LEMESHOW, Stanley. *Applied Logistic Regression*. John Wiley & Sons, Inc: New York, 1989.

IBGE (Instituto Brasileiro de Geografia e Estatística). “Censo demográfico”. Ed. IBGE, Rio de Janeiro, Brasil, 2010.

JOHNSON, Richard; Wichern, DEAN. *Applied Multivariate Statistical Analysis*. 4th Ed., Prentice-Hall: New Jersey, 2002.

KWAK, Nojun; CHOI, Chong-Ho. Input Feature Selection for Classification Problems. *IEEE Trans on Neural Networks*, New York, v. 13, n.1, p. 143-159, ago. /2002.

LEVENBERG, Kenneth. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, USA, n. 2, p. 164-168, mar. /1944.

LEYSHON, Andrew; FRENCH, Shaun; SIGNORETTA, Paola. Financial exclusion and the geography of bank and building society closure in Britain. *Transactions of the Institute of British Geographers*, New Jersey, nº 33, p. 447–465, jun. /2008.

MACKAY, David. Bayesian interpolation. *The MIT Press Journals – Neural Computation*, Cambridge, v. 4, n. 3, p. 415-447, mai. /1992.

MARQUARDT, Donald. An Algorithm for Least Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, Philadelphia/USA, vol. 11, n. 2, p. 431-441, jun. /1963.

MARTIN, Ron. The new economic geography of money. In: Ron Martin (Ed.). *Money and the Space*. *Economy New York*, USA: John Willey & Sons, p. 3 – 27, 1999.

MEDEIROS, Marcelo; PEDREIRA, Carlos. What are the effects of forecasting linear time series with neural networks? *Engineering Intelligent Systems*, United Kingdom, vol. 9, p. 237-242, dez. /2011.

MINSKY, Hyman. *Stabilizing an Unstable Economy*. New Haven, USA: Yale University Press, 1986.

NGUYEN, Derrick; WIDROW, Bernard. The truck backer-upper: An example of self-learning in neural networks. *Proceedings of the 1989 International Joint Conference on Neural Networks*, Washington/USA, v.2, p. jun. /1989.

Universidade Federal Rural do Rio de Janeiro (2012). “IBM SPSS Statistics for Windows” [IBM SPSS], Version 21.0. IBM Corp. Armonk, USA.

World Bank. *Brazil: Access to Financial Services*. Anjali Kumar (coord.). Washington/USA: World Bank, 2005.

Anexo 1

Tabela A1: Estatísticas referentes à Análise de Agrupamentos

Pergunta	Resposta	Grupo 0 - Excluídos			Grupo 1 – Não excluídos		
		Frequência	Percentual	Percentual Acumulado	Frequência	Percentual	Percentual Acumulado
Pergunta 8 - Está Empregado?	Sim	136	61.3	61.3	77	92.8	92.8
	Não	86	38.7	100	6	7.2	100
	Total	222	100		83	100	
Pergunta 15 - Possui Conta Corrente	Não	108	48.6	48.6	18	21.7	21.7
	Sim	106	47.7	96.4	54	65.1	86.7
	Mais de Uma	8	3.6	100	11	13.3	100
	Total	222	100		83	100	
Pergunta 21 - Canal de relacionamento Importante	Agência bancária	57	25.7	25.7	48	57.8	57.8
	Financeira	1	0.5	26.1			
	Correios	3	1.4	27.5			
	Lotéricas	94	42.3	69.8	16	19.3	77.1
	Lojas/farmácias/ Supermercados	66	29.7	100	19	22.9	100
Total	222	100		83	100		
Pergunta 25 - Qual sua Principal Fonte de Renda Atualmente	Salários	115	51.8	51.8	60	72.3	72.3
	Diárias	12	5.4	57.2	2	2.4	74.7
	Rendimentos de atividades empresariais	7	3.2	60.4	9	10.8	85.5
	Rendimento de atividade informal	21	9.5	69.8	6	7.2	92.8
	Renda de aplicações financeiras ou aluguéis	3	1.4	71.2			
	Seguro Social	25	11.3	82.4	2	2.4	95.2
	Aposentadoria	24	10.8	93.2	3	3.6	98.8
	Outros	13	5.9	100			
Total	222	100		83	100		

Fonte: Elaboração própria a partir dos dados da pesquisa.
Nota: Foi utilizado o Software SPSS na elaboração desta tabela.