UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE ESTATÍSTICA

GEOMARKETING: o uso da regressão logística múltipla no mapeamento de regiões geográficas de alto potencial mercadológico

Wagner Rodeski

Porto Alegre

Dezembro de 2010

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

DEPARTAMENTO DE ESTATÍSTICA

GEOMARKETING: o uso da regressão logística múltipla para o mapeamento de regiões geográficas de alto potencial mercadológico

Wagner Rodeski

Monografia apresentada para a Universidade Federal do Rio grande do Sul para a obtenção do grau de Bacharel em Estatística, sob orientação da Profª. Dra. Jandyra Maria Guimarães Fachel.

Banca Examinadora:

Profª. Dra. Jandyra Maria Guimarães Fachel

Bacharel em Estatística Gustavo Aprile Porto Rossi

Porto Alegre

Dezembro de 2010

DEDICATÓRIA

Dedico este trabalho à minha querida mãe, Terezinha, que lutou incansavelmente para que este momento de vitória chegasse.

Só eu sei e agora orgulhosamente reconheço e exalto todo esforço e dedicação que ela fez por mim.

É com grande emoção e amor incondicional que dedico a ela este trabalho, pois se cheguei até aqui, foi porque tive em todo tempo o seu exemplo de abnegação e implacável determinação.

A grande lição de fé e o compromisso em honrar e esperança em mim depositada certamente foram combustíveis altamente inflamáveis para a concretização deste sonho, agora realizado.

“Por maior que seja a montanha,

nunca tapará o sol.”

Provérbio chinês.

AGRADECIMENTOS

A Deus, meu maior agradecimento, por me dar forças para suportar as infindáveis madrugadas de monografia, por não me deixar perder a lucidez e tampouco o ânimo... por me ajudar a perseverar até o fim.

À minha família, bem mais valioso que possuo, pelo carinho, cuidado, presteza em me ajudar em todos os aspectos e pela fé que depositaram em todo momento em mim.

À minha mãe, Terezinha, pelas orações e pela confiança inabalável na minha capacidade.

Ao meu pai, Nilson, que mesmo com seu jeito nada extrovertido, pode passar a mensagem que estava lá, o tempo todo, pronto para me ajudar em tudo o que estivesse ao seu alcance.

Aos meus irmãos, Gabriel e Pricila, que sempre demonstraram grande orgulho de mim e que, além de acreditarem fielmente neste desfecho feliz, foram grande fonte de alegria e descontração para os momentos mais difíceis e sombrios por que passei.

À minha amável namorada, Anelise, que esteve todo instante ao meu lado, que sacrificou seu tempo livre para estar comigo me incentivando, não me deixando dispersar e, principalmente, tornando esta jornada muito menos árdua através do seu grande amor, carinho e dedicação, além de uma incomparável compreensão ao suportar minha constante ausência.

Aos professores do curso de Estatística, e ao próprio curso, dos quais fui presenteado com um bem que jamais se apagará, o conhecimento, pois segundo Albert Einstein “Uma mente que se abre a uma nova idéia nunca mais volta a seu tamanho original”.

À professora Jandyra, que com paciência me instruiu e me permitiu aprender mais que estatística, me fez enxergar que para se chegar ao sucesso é essencial o envolvimento das pessoas de uma forma tão harmoniosa quanto uma sinfonia.

Ao tutor e colega de empresa, Gustavo Rossi, que além colaborar efetivamente em várias etapas deste trabalho, esteve presente desde a germinação desta idéia, me ajudando a construí-la e torná-la sólida para utilização no contexto em que foi originada.

Por fim minha gratidão aos meus colegas de trabalho, pela compreensão da minha baixa disponibilidade, e pelo apoio e motivação nas horas de maior aflição pelas quais passei.

RESUMO

Em mercados altamente competitivos é intensa a necessidade de ferramentas capazes de proverem soluções com precisão maximizada, e suficientemente robustas para que não percam qualidade sob precárias condições de uso, isto é, quando se precisa de respostas rápidas, precisas e na maioria das vezes alimentadas por informações pouco padronizadas.

Em linha com a tendência de aumento das informações mercadológicas conectadas com informações geográficas, no intuito original de suprir a necessidade das organizações em obter respostas sobre suas dinâmicas geoespaciais, surge recentemente um movimento centrado em tomar dados de mercado disponibilizados sob formas georeferenciadas, e a estes dados sobrecarregá-los de valor agregado, transformando-os em massa crítica para a geração de inteligência de mercado, para que possam viabilizar tomadas de decisão mais assertivas nas ações de expansão comercial. Este novo conceito em análise de mercado tem se propagado como o estudo de Geomarketing.

Com ambição de propor uma solução ainda pouco explorada pelas organizações que já tem o geomarketing na lista de seus ferramentais de análise de mercado, este trabalho apresenta o uso da técnica estatística de Análise de Regressão Logística como instrumento de enriquecimento dos usuais e, comumente, subjetivos estudos de geomarketing.

Com um claro objetivo de trazer à superfície uma forma alternativa de solução para a problemática da expansão comercial, ao fim este trabalho apresentar-se-á um modelo estatístico com a habilidade de predizer a probabilidade de se observar um bom desempenho em um novo ponto de negócio inaugurado em uma localidade qualquer do Brasil, reproduzindo este resultado em mapas temáticos indo de encontro ao enfoque do geomarketing, onde com isso será possível visualizar a distribuição espacial das oportunidades para expansão dos negócios do conglomerado financeiro de crédito cooperativo, cujo nome não foi divulgado neste trabalho pela razão de não expor a empresa a qual teve suas informações tomadas para o desenvolvimento deste estudo.

Palavras-chave: Geomarketing, Regressão Logística, Expansão Comercial.

Sumário

Capítulo 1

Introdução

Este trabalho está centrado em evidenciar que o uso das técnicas estatísticas em conjunto com o uso de métodos tradicionais de outras áreas do conhecimento possui uma alta capacidade na solução de diversos problemas reais, salientando-se o caso em que os conhecimentos de marketing e de modelagem estatística são unidos para a obtenção de uma solução considerada de vanguarda na problemática da expansão comercial.

1.1 Contexto

Vivemos em uma era na qual a informação vem sendo disseminada em alta velocidade, isso graças à globalização promovida pelo desenfreado avanço do mundo digital. Isso naturalmente conduz ao surgimento de incontáveis oportunidades de desenvolvimento e inovação em praticamente todas as áreas do conhecimento. No mundo dos negócios isso não é diferente, ainda mais quando a posse da informação é subsídio principal para a consolidação do conhecimento, e este, é pilar para a geração de inteligência a favor do desenvolvimento dos negócios.

O conglomerado financeiro de crédito cooperativo, doravante denominado "S" (sigilo para preservar o nome da instituição), atua nos dias de hoje em quase 900 municípios, distribuídos em 11 estados brasileiros, através de um contingente de mais de 1000 pontos de atendimento. Em colaboração com a referida instituição é que foi desenvolvido este trabalho, o qual surge a partir de uma demanda crescente não apenas no conglomerado S, mas também em todas as organizações inseridas em cenários de alta competitividade, que é a maximização do sucesso na expansão dos negócios. Embora esse sucesso na abertura de novas unidades seja sem dúvida um elemento de natureza altamente subjetiva, sob uma ótica simplista esse sucesso pode ser tomado como sendo um resultado da ação conjunta de uma expansão executada de forma eficiente e eficaz. Nesse sentido, para a eficiência pode ser atribuída uma otimização do esforço despendido na ação de expansão, e à eficácia, a concretização do objetivo de existência da nova unidade aberta, isto é, a verificação de retorno financeiro à

instituição na proporção projetada, dentre outros aspectos mercadológicos não necessariamente mensuráveis.

Neste contexto, a partir da disponibilização de uma quantidade expressiva de informações de mercado e, da necessidade de uma metodologia com aval científico auxiliando no processo de tomada de decisão em ações de expansão de negócios, tomou-se como referência teórica os conceitos e diretrizes da crescente e inovadora linha de estudos de marketing conhecida como Geomarketing, incorporando a esta o teor científico advindo da técnica estatística chamada Análise de Regressão Logística.

Originado a partir da tradicional análise de marketing, o geomarketing surge como uma ferramenta especial por possuir grande inteligência agregada e, desse modo mais voltado para o uso em decisões gerenciais. A maior parte dessa abordagem reside em adicionar à análise de mercado o fator geográfico de forma relevante, fazendo isso através do georreferenciamento das informações oriundas de diversos tipos e fontes.

Já a análise de regressão logística, como método amplamente difundido na comunidade científica, contribui na modelagem dos fatores que podem influenciar o sucesso na abertura de uma nova unidade de negócio, gerando como resultado um modelo preditivo o qual, considerando elementos econômicos, sociais, infraestruturais e demográficos, dentre outros, tem o poder de discriminar a influência de cada fator na chance de sucesso de uma dada ação de expansão.

Tomando o resultado da análise de regressão logística, que representa a probabilidade de que uma nova unidade de negócio venha a apresentar um sucesso condizente com as expectativas da ação de expansão, esse valor de probabilidade pode ser plotado em um mapa gerando uma visão panorâmica do mercado e suas oportunidades, evocando os conceitos do geomarketing para o uso desse resultado final como ferramenta no planejamento estratégico de uma expansão comercial.

Por fim vale lembrar que a solução apresentada neste trabalho aplica-se estritamente ao particular caso aqui descrito, recomendando-se dessa forma a constante observação ao longo do trabalho dos pressupostos e diretrizes considerados no processo de obtenção dos resultados apresentados.

1.2 Objetivos

Com os resultados deste trabalho espera-se prestar apoio técnico à instituição financeira de crédito cooperativo S, em sua demanda de metodologia com embasamento científico para o planejamento e suporte à tomada de decisão no processo estratégico de expansão comercial.

Disso derivam objetivos pontuais a serem atendidos tais como demonstrar a utilidade da técnica estatística de análise de regressão logística, tendo já sido evidenciado sua extrema robustez e adaptabilidade, na predição das chances de sucesso quando da abertura de novas unidades de negócio, e, ainda entre os objetivos, abordar de uma forma inovadora na literatura acadêmica, o geomarketing com um vínculo direto com a metodologia estatística, adicionando esta metodologia à abordagem subjetiva e conceitual com que geralmente é tratado na literatura corrente. Além disso, figura também entre os objetivos pontuais a apresentação de um manual sucinto, para trazer à tona os procedimentos realizados no software MapInfo, quando da criação dos mapas temáticos requeridos pela visão panorâmica do mercado e suas oportunidades.

O propósito central deste trabalho é apresentar o mapa do Brasil a nível de município, colorido em diferentes tons, onde um tom mais intenso indica uma alta probabilidade de sucesso ao se expandir o negócio para o município portador daquela alta probabilidade, consideradas as características regionais relevantes ao referido sucesso, configurando dessa forma uma ferramenta gráfica com grande apelo visual, e com a atribuição de fornecer insumos suficientes para que a decisão estratégica mais adequada seja tomada.

1.3 Estrutura do Trabalho

Atendendo a proposta apresentada para este trabalho, foram posicionados os assuntos de tal maneira a permitir que a compreensão do tema central ocorresse de forma natural, isto é, segue-se um roteiro para o desenvolvimento do trabalho onde os assuntos abordados são expostos de forma sistematicamente encadeada.

Para tal, é apresentada no capítulo 2 a revisão da literatura a respeito das duas linhas teóricas aqui utilizadas, onde a seção 2.1 trata do *Geomarketing* e o 2.2 da *Análise de Regressão Logística*. A metodologia utilizada, escrita no capítulo 3, está subdivida nas seções 3.1, 3.2 e 3.3, contendo respectivamente os tópicos: *Tratamento dos Dados*, cujo tema central é o procedimento adotado para os casos de dados faltantes, *Obtenção do Modelo*

Estatístico que trata da modelagem estatística propriamente dita e por fim o *Diagnóstico e Validação do Modelo Final*, que faz algumas considerações sobre a ocorrência ou não de características desejáveis à um modelo de regressão logística. Já no capítulo 4 são apresentados os resultados obtidos a partir da metodologia descrita no capítulo 3, e isso em duas etapas de apresentação, onde uma é a *Aplicação do Modelo de Regressão Logística*, seção 4.1, e a outra é o *Mapeamento do Potencial Mercadológico*, seção 4.2. Neste capítulo 4 apresenta-se o que foi citado anteriormente como "propósito central" do trabalho. No encerramento da obra está o capítulo 5 fazendo *Considerações Finais* tais como ponderações sobre particularidades exclusivas deste estudo, assim como recomendações para a continuidade da exploração do tema aqui tratado. As *Referências Bibliográficas* utilizadas na realização deste material estão listadas no capítulo 6 e, por fim, apresenta-se no *Anexo I* o passo a passo do procedimento efetuado no software MapInfo para a confecção dos mapas temáticos.

Capítulo 2

Referencial Teórico

A partir deste ponto será apresentado o embasamento teórico utilizado como guia científico no desenvolvimento da metodologia adotada para este trabalho. Aqui no caso do geomarketing, foi assumida uma postura de exposição mais conceitual, seguindo a linha da literatura disponível para o assunto. Já para o caso da regressão logística, foi adotada uma abordagem mais branda que aquela disponível na maior parte da literatura, isto é, a exposição do assunto deu-se de forma menos aprofundada em termos matemáticos. Esta estratégia realizou-se em virtude do objetivo de favorecer a compreensão do texto por não apenas leitores com sólidos conhecimentos em estatística.

2.1 Geomarketing

O crescimento contínuo e veloz da competitividade entre as organizações tem desencadeado o surgimento de técnicas que embora sejam geralmente simples em seus modos de aplicação, são também em geral extremamente inovadoras em suas concepções. Dentre esses métodos contemporâneos, aquele que lista como um dos seus principais objetivos o aumento da eficácia na expansão dos negócios é sem dúvida o Geomarketing. Esta ferramenta, a qual conta basicamente com dois pilares, onde um é o Sistema de Informação Geográfica (SIG) e o outro é o Marketing propriamente dito, tem auxiliado as organizações a responderem questões mercadológicas de alta relevância (FAGUNDES *et al.*, 2009). No aspecto prático o geomarketing busca proporcionar uma fácil leitura e interpretação dos dados de mercado com um ferramental composto basicamente de relatórios gráficos e mapas temáticos, cujos conteúdos são informações inteligentes distribuídas geograficamente.

De acordo com Aranha (1996), o geomarketing é capaz de produzir informações úteis em diferentes aspectos da análise de marketing dentre os quais a determinação do potencial de mercado, a análise de projeção e resposta de campanhas de marketing, e de forma geral, em estudos onde o fator localização é um componente importante.

2.1.1 Sistema de Informação Geográfica (SIG)

Segundo McGoldrick (1990), o SIG ou GIS (*Geographic Information System*) cujo desenvolvimento iniciou-se por volta de 1960 é definido sucintamente como a ferramenta com a função de relacionar uma dada informação com a sua localização na Terra. Elias (*apud* Erba *et al.*, 2005) refere-se aos Sistemas de Informações Geográficas como ferramentas auxiliares na parametrização de modelos de planejamento, onde nestes modelos o uso dos SIGs permite que as informações sejam visualizadas de forma cartográfica, tornando a compreensão mais fácil e principalmente mais rápida de quando são apresentadas em relatórios ou tabelas, atribuindo dessa forma uma função primária dos SIGs que é a produção de mapas temáticos, onde cada tema é definido de acordo com a informação que alimenta o mapa e a mensagem que se deseja transmitir através desse recurso.

De acordo com Monteiro (2009), um exemplo pioneiro onde a análise espacial foi considerada de forma explícita se deu no século XIX pelo médico britânico John Snow. Segundo os registros históricos, em 1854 ocorria em Londres uma das inúmeras epidemias de cólera trazidas da Índia, das quais pouco se sabia sobre os mecanismos causadores da doença. Correntes científicas distintas buscavam encontrar tais elementos desencadeadores da doença seguindo duas linhas de pesquisa: uma que a relacionava aos miasmas, concentrados nas regiões baixas e pantanosas da cidade, e a outra que indicava a ingestão de água insalubre como sendo o evento causador. No referido contexto o médico John Snow agregou a um mapa de Londres (Figura 1) a localização das residências onde houveram óbitos ocasionados pela doença assim como a localização das bombas de água que abasteciam a cidade. A análise do mapa evidenciou que ao redor da bomba de água localizada em *Broad Street* havia uma concentração de casos de óbitos por cólera, sugerindo que esta região (circulada em vermelho) se tratava do epicentro da epidemia. Embora o agente infeccioso causador da cólera não houvesse sido descoberto até 1905, a simples remoção da bomba de água localizada no epicentro da epidemia foi suficiente para dar fim ao surto de 1854. O caso do médico britânico ilustra a contribuição significativa para a elucidação dos fatos do problema, que a relação entre dados e suas localizações geográficas vistos num panorama cartográfico apresenta, fazendo desta uma das primeiras utilizações da análise espacial.

Figura 1 - Mapa de Londres, onde óbitos por cólera estão representados por pontos, e poços de água estão identificados por cruzes. *Fonte: Monteiro, 2009, p. 21*

Os mapas temáticos desenvolvidos por intermédio do SIG utilizam-se de recursos visuais como jogo de cores e sombras, símbolos e demarcações para identificar e localizar características mercadológicas de regiões sob análise. Segundo Elias *apud* Malhotra (2001), nesses mapas podem ser combinadas as informações geográficas e demográficas, assim como dados de vendas e também outras informações exclusivas de uma empresa.

Em suma, os benefícios de um Sistema de Informação Geográfica incluem um ganho em sofisticação no uso dos dados, um maior detalhamento na demarcação de áreas comerciais e, de uma forma geral, o SIG propicia uma nova abordagem às atuais estratégias de segmentação de mercado que adotam estratos mais amplos, permitindo a convergência para o chamado micromarketing, ou marketing de vizinhança, o que no fim trata-se da sistematização do ferramental necessário para a incorporação minuciosa do aspecto geográfico ao marketing tradicional, onde o produto resultante dessa fusão será o geomarketing.

2.1.2 Geomarketing como Ferramenta de Expansão

O geomarketing pode ser entendido inicialmente como um novo conceito ao marketing tradicional. Isto porque tudo o que se sabe sobre o marketing tradicional está incluso no geomarketing, ou seja, o que ocorre aqui é uma sofisticação do conceito original através da sua projeção ao plano cartográfico. A vantagem do geomarketing está centrada essencialmente no modo com que o resultado da análise de marketing é apresentado e na possibilidade de se agregar a este novo modo de apresentação diferentes conhecimentos não exclusivos da área de marketing, porém tão enriquecedores que se tornam indispensáveis na hora da tomada de decisão.

Embora a evidente relevância prática da utilização do geomarketing nas organizações, na prática o uso desta ferramenta ainda é prematuro, isto porque se trata de uma área de estudo bastante recente se comparada com as teorias clássicas de marketing. Vale lembrar que grande parte dos usuários da análise de mercado ainda não tem pleno domínio do potencial das ferramentas de geomarketing ou até mesmo sequer sabem da existência desse conceito.

Todavia, em organizações privilegiadas com departamentos dedicados à expansão, o geomarketing já predomina as suas diretrizes de trabalho, mesmo que seja em maior ou menor escala, mesmo que seja de modo mais sofisticado metodologicamente ou então de maneira mais informal. A fase ainda não totalmente consolidada desse tema tem permitido que ele seja construído, solidificado e disseminado dia a pós dia, ou seja, embora já haja uma grande exposição do assunto, ainda impera a escassez de manuais técnicos e práticos para a uma fácil implantação nas organizações. Este cenário não deixa de ser positivo, pois é favorável à inovação e ao pioneirismo, até porque, um outro aspecto importante do geomarketing é a sua possibilidade de customização de acordo com a área ou o foco de sua utilização, o que em outras palavras traduz que, pelo menos por enquanto, ainda não existe um jeito absolutamente certo ou errado de se fazer geomarketing, mas sim inúmeras maneiras de fazê-lo tomando como premissa o seu conceito original e remodelando-o para responder, de forma eficaz, diferentes situações de estudo.

De acordo com Sobrinho (2008), através das lentes do geomarketing as empresas têm amplificado, sem perda de precisão, a identificação dos locais com maior potencial de consumo para um determinado produto em um bairro, cidade ou região qualquer. Sem desmerecer as demais formas e capacidades de atuação do geomarketing, a vantagem maior no uso dessa ferramenta é a habilidade de se poder responder à questão "*para onde*

expandir?". De forma geral, todo tipo de informação sobre vendas, perfis de consumo, movimentos pendulares, mercado imobiliário, atividades econômicas, industriais, assim como informações demográficas, dentre muitas outras podem ser georreferenciadas e se transformarem em ferramentas inteligentes indispensáveis para a tomada de decisão no momento da escolha do melhor local para realizar a expansão.

Nesse sentido o geomarketing permite compreender as dinâmicas sociais e estabelecer relações econômicas entre elas, isto é, ele viabiliza o entendimento da interdependência existente entre os mecanismos sociais, demográficos e comerciais, inerentes ao sistema de utilização de um determinado produto ou serviço. No entanto, o modo como isso é feito geralmente tem permanecido incógnito, isso porque a metodologia aqui em questão, além de recente no campo científico, ela é muito mais difundida no meio empresarial que no acadêmico e, por questões muitas das vezes ligadas à vantagem competitiva, o *modus operandi* acaba não sendo suficientemente divulgado, deixando então um campo aberto à inovação e construção de novas metodologias com sustentação no geomarketing. Como existem peças chaves específicas no sistema de utilização de um dado produto ou serviço, é natural que haja também uma abordagem específica para cada tipo de demanda que se deseje estudar. No entanto, a forma genérica e conceitual do geomarketing apresentada ao longo deste capítulo, é capaz de atender qualquer forma de uso desse conceito, todavia, alguns recursos extras como teorias econômicas, projeções demográficas, pesquisas de campo e técnicas estatísticas dentre outros recursos, têm assumido posições de destaque na prática do geomarketing ao trazer um enriquecimento significativo aos resultados obtidos.

2.2 Regressão Logística

A investigação científica tem sido grande aliado no processo de enriquecimento intelectual da humanidade, assim como tem catalisado o progresso tecnológico que há por de traz da globalização cultural e comercial. A utilização em larga escala do poder inerente à investigação científica tem permitido inclusive o estudo de fenômenos naturais, econômicos e antropológicos dentre muitos outros. Neste contexto, surge uma ampla gama de recursos matemáticos e estatísticos que em geral fornecem ao investigador a capacidade de realizar uma leitura da realidade com suficiente rigor científico, bem como modelar essa realidade de forma tão satisfatória que se possa até mesmo predizer o comportamento dela além dos horizontes observados, e, ainda que haja um risco de equívoco nessa extrapolação, no caso dos modelos estatísticos este risco de errar é mensurável e controlado, fazendo com que modelagem estatística seja amplamente utilizada e confiável.

Dentre os muitos recursos estatísticos disponíveis há a Análise de Regressão, ferramenta poderosa na função de agregar informações oriundas desde uma a inúmeras variáveis que visam explicar um dado fenômeno, encontrando nesses dados padrões de relacionamentos entre essas variáveis, apresentando como resultado uma função que expressa matematicamente o modo como tais variáveis podem explicar as alterações no evento em estudo. Em outras palavras, a Análise de Regressão tem o supracitado poder de modelar o comportamento de um fenômeno qualquer.

Segundo Bruni *apud* Galton (1885), o termo regressão teve sua origem na literatura científica quando o estatístico inglês Sir Francis Galton, investigou a relação entre a altura dos pais e a altura dos filhos. Como era de se esperar, suas conclusões indicaram que pais altos tendem a ter filhos altos e pais baixos tendem a ter filhos baixos. No entanto, Francis Galton notou que parte dos pais de estatura alta tinha filhos de baixa estatura e também havia parte dos pais de baixa estatura que tinha filhos de alta estatura. A esse comportamento observado Galton o chamou de *regression toward the mean,* ou seja, de regressão em direção à média.

2.2.1 Modelos de Regressão

Os modelos de regressão são ferramentas estatísticas utilizadas para estudar o comportamento de uma variável dependente, também chamada de variável resposta, por meio da análise de outras variáveis que buscam explicar tal comportamento, ou seja, um

modelo de regressão bem ajustado é capaz de fornecer a relação existente entre variáveis explicativas de tal maneira que estas consigam predizer de forma satisfatória as oscilações inerentes à variável resposta.

Em concordância com Agresti (2000), ao supor uma variável resposta Y e um vetor com p variáveis independentes, $x_i = (x_{i,1},...,x_{i,p})$, pode ser definido um modelo estatístico de regressão da seguinte forma:

$$Y_i = h(x_i) + \varepsilon_i \qquad \text{para } i = 1, ..., n \qquad (2.1)$$

A forma genérica de um modelo de regressão estatística, delineada para n observações, está representada pela a equação (2.1), onde nela $h(x_i)$ é uma função matemática qualquer que expressa a relação existente entre a variável resposta e as variáveis explicativas do modelo e, ε_i é o chamado erro aleatório do modelo, que por ser uma variável aleatória tem uma distribuição de probabilidade associada a ele e, é este termo que adiciona o teor estatístico à estrutura matemática da equação de regressão.

Em geral o formato da função $h(x)$ em conjunto com a natureza da variável resposta é que define com que espécie de regressão se está lidando, pois a partir desta definição uma abordagem específica para cada tipo de regressão deverá então ser adotada. Para ilustrar esta situação suponha um modelo em que Y, variável resposta, seja uma variável de natureza contínua e que haja uma variável x^2 (formato quadrático de $h(x)$) explicando a variação de Y. Para o presente caso estaria configurada uma regressão quadrática, onde seria necessário o uso de artifícios matemáticos aplicados nos valores originais da variável x^2 para então o subseqüente uso de métodos convencionais para a estimação dos coeficientes do modelo. Por outro lado, já não seria necessária tal transformação nos dados originais se fosse o caso de estar uma variável x (forma linear) a modelar Y. Essa abordagem mais simplificada para o modelo linear em relação ao quadrático ou outro qualquer de maior grau, se dá basicamente por três fatores, sendo o primeiro por ele ser de forma geral mais simples, segundo, ele é útil para a resolução da maior parte dos problemas envolvendo modelagem de fenômenos com variável resposta de natureza contínua, o que disso deriva uma ostensiva produção de conhecimento para tais modelos lineares, e, por fim, a possibilidade e prática comumente

realizada de manipulações matemáticas nos dados e nos próprios modelos não lineares de modo a torná-los utilizáveis sob o enfoque linear generalizado.

Ainda no que tange à definição do tipo de regressão a ser realizada, há a diferenciação pela natureza da variável resposta, onde esta é geralmente classificada como contínua ou categórica. Quando contínua, é comum o uso de modelos lineares para a análise de regressão, já para o caso de ela ser categórica, ou seja, assumir valores discretos ou nominais (em geral codificados numericamente), algumas propriedades da regressão linear perdem eficácia ou até mesmo validade, gerando a necessidade de uma nova abordagem e de uma reformulação nas técnicas de regressão linear. Como solução à lacuna metodológica criada pela modelagem de um fenômeno com resposta categórica é que surge a análise de regressão logística, cujo entendimento desta é o foco desde capítulo e consequentemente será discutida com maior profundidade a seguir.

2.2.2 Modelos de Regressão Linear

Antes de tratar sobre a regressão logística é relevante discutir, ainda que brevemente, a forma mais simples da análise de regressão, que é a regressão linear simples. Este tipo de análise relaciona uma variável resposta de natureza contínua a uma única variável explicativa por intermédio de uma relação linear. Sendo assim, a fórmula matemática genérica para a regressão linear simples é do tipo:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad (2.2)$$

Está representado em (2.2) a equação de regressão linear simples, onde β_0 e β_1 são os coeficientes e ε o erro da regressão. Como entendimento preliminar dos elementos da referida equação, o coeficiente β_0 representa o valor da variável resposta quando a variável explicativa x for igual a zero e, o coeficiente β_1 representa a magnitude da variação ocorrida na variável resposta quando se faz a variável x oscilar em uma unidade, de modo que o sinal deste coeficiente indica se o acréscimo de uma unidade na variável x ocasiona também um acréscimo em Y, onde neste caso o sinal será positivo, ou se este aumento em x gera um decréscimo em Y, indicando que β_1 tem sinal negativo. Por se tratar de um modelo estatístico, ou seja, não determinístico, ao utilizar-se do modelo ajustado para estimar qual

valor da variável resposta Y está associado a um dado valor da variável x, é natural que as estimativas apresentem uma variação em relação ao verdadeiro valor de Y, e a leitura que se faz desse desvio entre o valor estimado e o valor observado é o que outrora foi chamado de erro da regressão, que no caso da estimação do modelo ele é denotado por e.

Vale lembrar que o domínio das estimativas de $\hat{Y}$ abrange toda a reta real $\Re$, ou seja, o valor estimado para a variável resposta pode assumir qualquer valor desde $-\infty$ a $+\infty$. Por fim, uma generalização da regressão linear simples é a inclusão de mais variáveis explicativas ao modelo, o que naturalmente incorre numa adaptação matemática visto que o poder de predição do modelo deve levar em consideração, quando da obtenção dos coeficientes, todas as variáveis explicativas (também chamadas de preditivas) simultaneamente, o que de maneira intuitiva já sugere uma maior complexidade no cálculo da solução.

2.2.3 Modelos de Regressão Logística

A regressão logística vem assumindo um papel protagonista servindo ao propósito da investigação científica em diversas áreas do conhecimento, com um notório crescimento nos últimos 30 anos (SOUZA *apud* CRAMER, 2002). Descoberta no século XIX para descrever o crescimento das populações e as reações químicas em cursos de autocatálise, hoje em dia ela agrega alto valor em estudos epidemiológicos apoiando na determinação de fatores de risco para as doenças estudadas, bem como tem apresentado uma extrema eficácia como ferramenta estatística para a obtenção de escores de risco de inadimplência na área de concessão de crédito, ou mais popularmente falando, nos estudos de *credit scoring*, além de inúmeras aplicações desde a descrição das chances de ocorrência de um determinado evento em função de fatores específicos até a obtenção de modelos preditivos em economia, ciências sociais, geoanálise, marketing, etc., dentre outras aplicações. Essa popularização da regressão logística deve-se a basicamente a dois fatores em que, um é a capacidade que a técnica tem em fornecer uma probabilidade de ocorrência de um dado fenômeno em estudo, o que também pode ser visto alternativamente como o comportamento das chances de ocorrência desse fenômeno a partir de uma alteração ocorrida em um determinado fator explicativo da regressão, e, outro elemento a favor dessa alta disseminação é a flexibilidade que o modelo de regressão logística tem em aceitar todo tipo de variável explicativa, ou seja, isso postula que tanto variáveis contínuas quanto categóricas são permitidas, amplificando extremamente as possibilidades de uso de todo tipo de informação para modelar um fenômeno.

Segundo Souza (2006), na análise de regressão logística a variável resposta é de natureza dicotômica, ou seja, assume o valor 1 quando ocorre o evento de interesse (na linguagem estatística é comum o uso do termo *sucesso*) e o valor 0 quando ocorre o evento complementar (*fracasso*), onde os eventos correspondentes aos valores 1 e 0 devem necessariamente ser mutuamente exclusivos, isto é, a ocorrência de um evento implica na não ocorrência do outro e vice-versa, e, a soma desses dois eventos representa todas as possibilidades de resultados da variável resposta em questão.

A ocorrência dos eventos "sucesso" e "fracasso" se dá com probabilidades $\pi(x) = P(Y = 1 \mid X = x)$ e $1 - \pi(x) = P(Y = 0 \mid X = x)$, respectivamente. Como Y só pode assumir os valores 0 e 1, a probabilidade $\pi(x)$ será igual a $E(Y \mid X = x)$, que é a média (na estatística é comum o uso dos termos *esperança* ou, *valor esperado*, para referir-se à média) condicional de Y dado x. Conforme o trabalho de Hosmer e Lemeshow (2000), a distribuição acumulada de probabilidade tem sido utilizada para a obtenção de modelos para a média condicional de Y dado x quando as variáveis resposta são dicotômicas. Embora muitas funções de distribuição de probabilidade têm sido propostas para darem suporte a modelos estatísticos de resposta binária, a distribuição logística de probabilidade continua imperando nas escolhas para modelagem devido a sua extrema flexibilidade e fácil utilização, isso do ponto de vista matemático, assim como sua capacidade de proporcionar interpretações ricas em significados práticos. Dessa forma, este trabalho apresenta a distribuição logística de probabilidade como referencial teórico ao modelo de regressão com variável resposta dicotômica. Sendo assim a probabilidade de ocorrência de um evento de interesse na regressão logística é expressa da seguinte forma:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \qquad (2.3)$$

A equação (2.3) representa em síntese a entrega final da análise de regressão logística. Neste ponto surgem duas importantes diferenças entre a regressão logística e a regressão linear. Primeiro, embora a mesma quantidade a ser modelada em qualquer problema de regressão seja o valor médio da variável resposta dado os valores das variáveis explicativas, genericamente denotado por $E(Y \mid X = x)$, na regressão logística por ser este valor médio uma probabilidade ele assume apenas valores entre 0 e 1, enquanto que na regressão linear

esta média assume valores de $-\infty$ a $+\infty$. A outra grande diferença reside na distribuição de probabilidade dos erros da regressão, que enquanto na regressão linear estes seguem uma distribuição Normal com média 0 e variância constante qualquer, na regressão logística esses erros têm média 0 e variância $\pi(x)[1-\pi(x)]$, em que neste caso é a distribuição Binomial a mais adequada para descrever a distribuição de probabilidade dessas quantidades aleatórias.

O cerne da técnica de regressão logística jaz sobre a transformação *logit* realizada em $\pi(x)$, definida da seguinte maneira:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \qquad (2.4)$$

Aqui, a transformação apresentada em (2.4) tem suma importância porque dela derivam muitas propriedades desejáveis de uma regressão linear. De acordo com Hosmer e Lemeshow (2000), o *logit*, $g(x)$, é linear em seus parâmetros, e pode compreender valores de $-\infty$ a $+\infty$, dependo apenas da amplitude dos valores de x.

O processo de obtenção do modelo de regressão logística consiste na determinação dos coeficientes desconhecidos β_0 e β_1, procedimento este realizado por intermédio do método da máxima verossimilhança, o qual sumariamente tem a função de estimar os coeficientes desconhecidos de modo que a partir desta estimativa, a probabilidade de obtenção de dados tais como aqueles observados na amostra seja maximizada.

Desse modo, de acordo com Bruni (2007), para uma amostra de tamanho n, a contribuição de um dado par de observações (x_i, y_i) para a função de verossimilhança se dá conforme a expressão (2.5):

$$\zeta(x_i) = \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{(1-y_i)} \qquad (2.5)$$

Onde $\pi(x_i) = f(\beta_0, \beta_1)$

Como do delineamento amostral parte que as observações Y_i sejam todas independentes, a função de verossimilhança para amostra observada é obtida pelo produtório da equação apresentada em (2.5), o que gera o seguinte resultado:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{(1-y_i)} \qquad (2.6)$$

Onde $\boldsymbol{\beta} = (\beta_0, \beta_1)$

Conforme postula o método da máxima verossimilhança, as estimativas de β_0 e β_1 são obtidas de modo que $l(\boldsymbol{\beta})$ seja maximizada. Sob o ponto de vista de manipulação matemática é conveniente que a expressão (2.6) seja tratada na sua forma logarítmica, definida em (2.7) como:

$$L(\boldsymbol{\beta}) = \ln\left[l(\boldsymbol{\beta})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\} \qquad (2.7)$$

O procedimento matemático necessário para a determinação dos valores de $\boldsymbol{\beta}$ que maximizam a expressão (2.7) será a derivação desta em relação à β_0 e β_1, e a seguir tornar a expressão resultante igual zero. O resultado dessa operação será a geração de duas equações conhecidas como expressões de verossimilhança as quais apresentam a seguinte forma:

$$\sum_{i=1}^{n} \left[y_i - \pi(x_i) \right] = 0 \qquad (2.8)$$

E

$$\sum_{i=1}^{n} x_i \left[y_i - \pi(x_i) \right] = 0 \qquad (2.9)$$

Segundo Hosmer e Lemeshow (2000), nos modelos de regressão logística a solução das equações (2.8) e (2.9) requer o uso de métodos especiais. No particular caso da regressão

logística, Hosmer e Lemeshow citam o trabalho de McCullagh e Nelder (1989) como referência em métodos para a solução das referidas equações, onde o método iterativo de mínimos quadrados ponderados é apresentado como o mais adequado para o caso aqui em questão.

Como o foco deste capítulo é a compreensão da técnica de regressão logística em sua forma mais conceitual, até então foi considerado o modelo de regressão logística com apenas uma variável explicativa. No entanto é útil a visualização da forma assumida pelo modelo quando mais de uma variável independente é utilizada na modelagem.

Basicamente, em termos conceituais a diferença reside apenas no acréscimo de variáveis preditoras ao modelo, o que conduz a algumas sutis adaptações na formulação matemática do modelo de regressão simples, e que podem ser vistas como segue:

$$g(\mathbf{x}) = \ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1} \qquad (2.10)$$

Onde $\mathbf{x} = (x_1, x_2, ..., x_{p-1})$

Do mesmo modo, as demais expressões antes apresentadas para o caso do modelo de regressão logística simples seguem valendo para o modelo de regressão logística múltipla, onde a notação $\boldsymbol{\beta}'\mathbf{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1}$ passa a ser inserida. Além do mais, a probabilidade se *sucesso* no caso do modelo múltiplo poderá ser obtida através da expressão:

$$\pi(\mathbf{x}) = \frac{e^{\boldsymbol{\beta}'\mathbf{x}}}{1+e^{\boldsymbol{\beta}'\mathbf{x}}} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_{p-1}x_{p-1}}} \qquad (2.11)$$

2.2.4 Testes de Significância dos Coeficientes

De posse dos princípios básicos que sustentam a análise de regressão logística, parte-se para a etapa de verificação da qualidade do modelo ajustado. Etapa esta crucial para a interpretação da análise, pois nela o pesquisador pode tanto compreender o que a estatística

tem a evidenciar sobre o fenômeno em estudo, quanto intervir no modelo resultante de acordo com seu conhecimento sobre o problema em questão. É neste momento que se faz a leitura daquilo que os dados "têm a dizer", e, que se enxergam as interdependências entre todas as variáveis envolvidas na modelagem, ocasionando a tomada de decisão no tocante à escolha do modelo final.

Anteriormente foi citado que para a configuração de um modelo de regressão logística múltipla, essencialmente a mudança em relação ao modelo simples era o acréscimo de variáveis preditoras, no entanto, este acréscimo não pode ocorrer de forma descontrolada, mas sim resguardando os princípios científicos envolvidos na análise de regressão. Seguindo um procedimento puramente estatístico a inclusão ou exclusão de variáveis independentes ao modelo se dá de forma simples, porém metódica, seguindo entre outros critérios os de significância estatística dos coeficientes das variáveis independentes.

De acordo com Hosmer e Lemeshow (2000), o ponto de partida para a avaliação da entrada ou saída de uma determinada variável no modelo é a comparação de modelos com e sem esta determinada variável sob teste no que diz respeito a qual dos dois modelos diz mais sobre a variável resposta, ou em outras palavras, qual modelo explica mais o comportamento do fenômeno em estudo, aquele com ou o sem a variável explicativa em avaliação.

A supracitada comparação é realizada através da razão de verossimilhanças, expressão (2.12), onde esta razão é a verossimilhança do modelo ajustado dividida pela verossimilhança do modelo saturado (definido como o modelo cujos *outputs* correspondem exatamente aos valores observados). A quantidade "menos duas vezes" o logaritmo natural da razão de verossimilhanças segue uma distribuição qui-quadrado com υ graus de liberdade, onde υ é obtido pela diferença do número de graus de liberdade entre os dois modelos.

$$D = -2\ln\left[\frac{\text{verossimilhança do modelo atual}}{\text{verossimilhança do modelo saturado}}\right] \qquad (2.12)$$

Usando as expressões expostas em (2.7) e (2.12) e com alguma manipulação matemática tem-se o seguinte resultado:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\pi(x_i)}{y_i}\right) + (1 - y_i)\ln \right] \qquad (2.13)$$

A expressão (2.13) representa a estatística *Deviance*, que corresponde à chamada soma de quadrados residuais na regressão linear, onde ambas estatísticas visam a mensuração do ajuste do modelo de regressão em relação aos dados.

Quando da avaliação da significância de uma determinada variável explicativa, é comparado o valor da estatística *D* referente ao modelo na presença dessa variável, com o valor de *D* no caso do modelo sem a referida variável explicativa.

Sendo assim, pode-se elaborar um conceito de mensuração da diferença entre os dois modelos simplesmente computando a alteração observada em *D* devida a inclusão da variável explicativa ao modelo, conforme expressão abaixo:

$$G = D(\text{para o modelo sem a variável}) - D(\text{para o modelo com a variável}) \qquad (2.14)$$

A equação (2.14) diz respeito à estatística usada para se verificar a significância de uma dada variável independente recém inclusa no modelo de regressão. Ainda sobre a estatística *G*, relembrando (2.12) verifica-se que a verossimilhança do modelo saturado é comum aos dois valores de *D*, o que após algum desenvolvimento algébrico conduz a uma nova expressão para a estatística *G*:

$$G = -2\ln\left[\frac{\text{verossimilhança do modelo sem a variável}}{\text{verossimilhança do modelo com a variável}}\right] \qquad (2.15)$$

Tomando como exemplo o caso da modelagem com apenas uma variável independente, tem-se que sob a hipótese de β_1 ser igual à zero, a estatística *G* seguirá uma distribuição qui-quadrado (χ^2) com 1 grau de liberdade.

Ainda como recurso adicional para a verificação de significância estatística de uma variável independente em um modelo de regressão existe o teste de Wald, o qual é

computado a partir da comparação do coeficiente estimado, $\hat{\beta}$, da variável independente a qual se quer testar a significância, com a estimativa do seu erro padrão, $\hat{SE}$ (*Standard Error*).

A premissa do teste é de que sob a hipótese de $\hat{\beta} = 0$, a estatística *W* expressa em (2.16) seguirá uma distribuição normal padrão, $N(0,1)$.

$$W = \frac{\hat{\beta}}{\hat{SE}\left(\hat{\beta}\right)} \qquad (2.16)$$

Segundo Hosmer e Lemeshow *apud* Hauck e Donner (1977), quando examinada a eficácia do teste de Wald, este tem frequentemente se comportado de maneira atípica, falhando em rejeitar a significância de um coeficiente quando este de fato era significativo, onde a partir dessas constatações os autores recomendam que o teste da razão de verossimilhança seja utilizado.

Capítulo 3

Metodologia

A partir deste ponto inicia-se o processo de modelagem das características regionais que favoreceriam a ocorrência de um bom desempenho em uma nova unidade de atendimento a ser aberta em um dado município. Este modelo preditivo é obtido através da análise de regressão logística (Figura 2) e, após, o seu resultado é incorporado a uma ferramenta de geomarketing, propiciando uma visão espacial e gerencial da análise do mercado e suas oportunidades para a subseqüente tomada de decisão no tocante aos movimentos de expansão a serem realizados.

Para tanto, é necessário enfatizar o cerne deste estudo, que é detectar qual o perfil dos municípios, onde em média é observado um desempenho positivo para as unidades de atendimento já presentes no referido município. Uma vez encontrado o perfil da cidade que tem um alto potencial de viabilizar a ocorrência de uma expansão bem sucedida, parte-se então para a busca de cidades que melhor se enquadrem nesse perfil, para que a estas se possa atribuir uma maior probabilidade de sucesso ao se inaugurar novos negócios, já que foram previamente sinalizadas pelo modelo como regiões de alto potencial mercadológico.

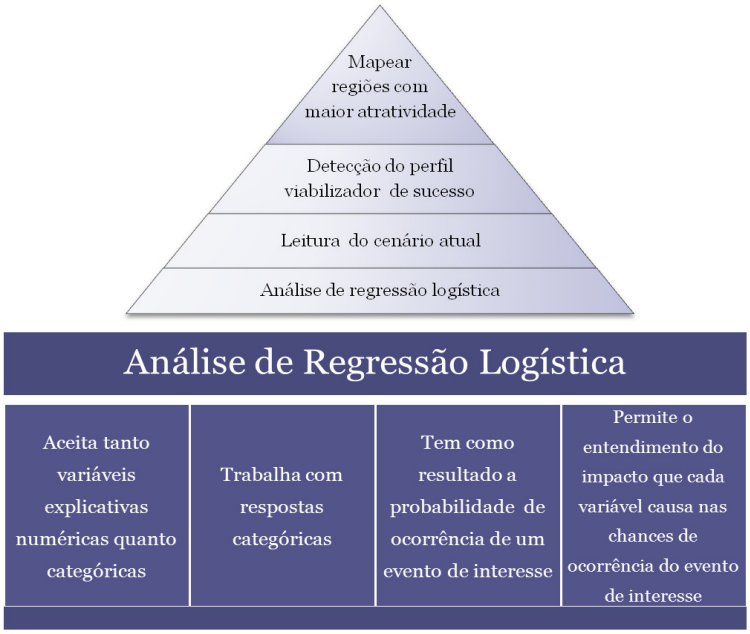

Figura 2 – Estrutura do estudo evidenciando a regressão logística como base.
Fonte: Autor

3.1 Variável Resposta

A apuração do desempenho das unidades de atendimento da instituição financeira S, definido como variável resposta, se deu através da soma dos valores de volume de negócios e rentabilidade verificados em 2010, ou seja, a cidade cujo perfil econômico-estrutural propiciou que em uma unidade de atendimento fosse verificado um montante expressivo ou em termos de volume de negócios ou em rentabilidade ou em ambos, indica que o perfil dessa cidade é um viabilizador da ocorrência de um bom desempenho.

A leitura do desempenho medido em uma unidade de atendimento pode ser tomada como sendo a composição de dois elementos, onde um expressa o aquecimento do mercado naquela região, mensurado através do volume de negócios, e o outro reflete a propensão da população local em utilizar os serviços disponibilizados pela empresa S, medida através da rentabilidade observada naquela unidade de atendimento.

Como o objeto desse estudo concentra-se em obter uma probabilidade para o desempenho de uma futura unidade de atendimento, resguardando uma ampla flexibilidade de modelagem e um maior grau de riqueza em interpretações, optou-se por tomar a regressão logística binária como ferramenta estatística de modelagem, a qual exige que a variável resposta seja expressa de forma binária. No intuito de obter-se a referida resposta binária, foi estabelecido um ponto de corte cujo critério de escolha deste ponto foi de ser uma medida de tendência central, onde se atribuiria valor *zero* se a variável resposta fosse menor que o ponto de corte, e valor *um* para o caso contrário.

Para tanto, após a apuração dos valores brutos da variável resposta para cada unidade de atendimento, como medida de contenção da intensa variabilidade observada nesses resultados procedeu-se com o ordenamento destes valores passando-os para uma base 100, isto é, atribuindo um escore 100 para o valor mais alto e após dividindo cada um dos demais valores por este máximo e então multiplicando este quociente por 100, obtendo dessa forma valores indo de zero a cem dentro das devidas proporções. Terminada esta etapa de ordenamento, ao observar um comportamento na distribuição desses escores de desempenho ligeiramente concentrado na média, Figura 3, adotou-se esta medida de tendência central como critério na determinação do ponto de corte.

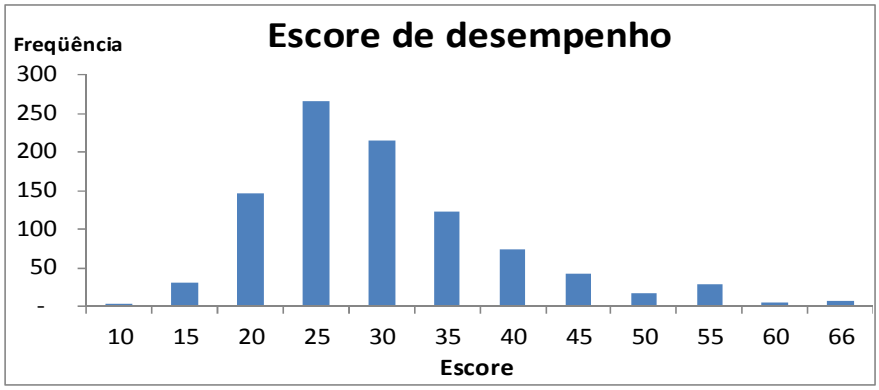

Figura 3 – Escores de desempenho concentrados em torno da média.

Fonte: Autor

3.2 Variáveis Explicativas

O processo de escolha das variáveis, cujo papel é explicar o comportamento do desempenho observado nas unidades de atendimento, foi desenvolvido em cima de quatro pilares os quais se julga caracterizarem bem o município que contém um determinado ponto de negócio. Estes pilares são as características econômicas, sociais, de infraestrutura e demográficas do município, onde dentro de cada pilar foram introduzidas algumas variáveis de um total de 40, Quadro 1, com maior relevância e com algum particular interesse de investigação, para posteriormente terem suas significâncias no modelo avaliadas.

Quadro 1 – Estrutura das variáveis explicativas

Economia	Quantidade de Empresas	Total empresas
		Indústrias em geral
		Indústrias de construção
		Comércios em geral
		Comércio atacadista
		Comércio varejista
		Serviços em geral
		Saúde
		Educação
		Financeiros
		Alojamento e alimentação
		Agronegócio
	Indicadores	Índice de potencial de consumo (IPC)
		População economicamente ativa (PEA)
		Produto interno bruto (PIB) *per capita*
	Informações Financeiras	Agências bancárias
		Depósitos à vista
		Depósitos a prazo
		Poupança
		Crédito
Demografia		Taxa de crescimento demográfico
		Densidade demográfica
		% população entre 20-29 anos
		% população entre 30-49 anos
		% população com 50 anos ou mais
		% população urbana
Infra-estrutura		Frota
		Distância da capital
		Transportes
		Total de domicílios
		% domicílios urbanos
		% domicílios urbanos-classe B
		% domicílios urbanos-classe C
		% domicílios urbanos-classe D e E
		Área plantada
Social		% população alfabetizada
		Médicos / mil habitantes
		IFDM * emprego
		IFDM * educação
		IFDM * saúde

* *Índice FIRJAN (Federação das Indústrias do Rio de Janeiro) de Desenvolvimento Municipal*

Fonte: Autor

Dada a intensa variação existente, quando observados os dados referentes aos quatro pilares adotados pelo estudo, nos cerca de 900 municípios onde a instituição financeira S está hoje presente, aliada com um dos propósitos deste trabalho que é lançar uma proposta vanguardista de estudo na referida instituição e juntando a isso algumas limitações de recursos ocorridas, foi decidido categorizar todas as variáveis explicativas em duas categorias, onde determinada variável assumiria valor *zero* caso seu valor original estivesse abaixo do ponto de corte obtido para esta variável, e assumiria valor *um* para o caso contrário. Além do mais, tal como foi para a variável resposta, aqui também foi utilizada uma medida de tendência central, para a determinação do ponto de corte de cada variável.

Vale citar que um fator de cunho matemático considerado na decisão de se categorizar as variáveis explicativas foi a grande quantidade de variáveis a serem modeladas e com todas apresentando intensa variabilidade, o que em termos de modelagem estatística causaria a rejeição de variáveis importantes que seriam excluídas pelo algoritmo de modelagem em função de que seus coeficientes, por terem um alto desvio-padrão, seriam considerados não significativos para o modelo.

Tomando o mesmo racional utilizado na escolha da medida de tendência central balizadora do ponto de corte da variável resposta, aqui no caso das variáveis explicativas, embora a variabilidade observada fosse ainda maior que aquela observada nos dados brutos da variável resposta, aqui não houve interesse em criar um ordenamento para contornar essa intensa variabilidade, isso devido ao atendimento do aspecto prático da utilização do modelo final, onde seria mais complexa e mais demorada a obtenção da ordenação de um novo valor, ou seja, fugiria do escopo do trabalho, que é prover uma ferramenta com alto nível de operacionalidade e rapidez de resultados. Todavia, para não ignorar a grande variabilidade mencionada e também considerar eventuais assimetrias nos dados, elegeu-se a mediana como medida determinadora do ponto de corte, onde é válido citar como propriedade da mediana a sua baixa contaminação com valores extremos, o que não acontece com a média, que é fortemente afetada por valores extremos.

Uma consideração relevante a se fazer neste ponto é que, no presente trabalho, os referidos valores dos pontos de corte tanto para cada variável explicativa quanto para a variável resposta, não foram divulgados em ordem de preservar o conglomerado financeiro S de uma possível competitividade desleal, pois essa informação é crucial para a visualização e para o dimensionamento numérico dos fatores chaves, para o sucesso observado na referida instituição.

3.3 Tratamento dos Dados

Como ponto de partida para a execução da etapa de modelagem, procedeu-se com o processo de exploração de dados nas mais de mil observações de desempenho computadas para todas as unidades de atendimento da empresa S. A partir desta análise exploratória foram identificadas algumas observações faltantes para a variável resposta, onde o resgate dessas informações não foi possível de ocorrer em tempo hábil para a etapa de modelagem, conduzindo à exclusão dessas observações no banco de dados, resultando dessa forma em um conjunto de dados contendo 956 observações, onde cada observação refere a uma unidade de atendimento da empresa S. No entanto, como algumas destas unidades de atendimento estão localizadas dentro de um mesmo município, para se obter um único valor de *performance* por município, calculou-se a média dos desempenhos observados nas unidades de uma mesma cidade, o que fez com que o número total de observações fosse reduzido para 827.

Ainda no processo de exploração de dados, nas variáveis explicativas foram detectados dados faltantes em *População economicamente ativa (PEA)* e *médicos / mil habitantes*, sendo que para este problema a solução adotada foi a imputação dos dados faltantes.

Atualmente existem diversos métodos disponíveis na literatura quando o assunto é imputação de dados faltantes, dentre os quais vale destacar o mais simples que é a média, onde neste método, aos valores ausentes ocorridos em uma variável é atribuído o valor da média dessa variável. Ainda neste sentido, um outro método bastante disseminado merece destaque, que é o da interpolação, a qual pode ser linear, polinomial ou ainda trigonométrica. Todavia, aqui se buscou realizar uma imputação das informações faltantes com uma maior coerência com os dados presentes, descartando assim os métodos de imputação mais mecanizados e com pouco sentido prático. Para tanto, realizou-se um procedimento mais personalizado lançando mão da técnica estatística de Análise de *Clusters*.

A análise de *clusters* tem a peculiaridade de fazer o reconhecimento de padrões em uma matriz de dados, entregando como resultado a discriminação e o agrupamento das observações com comportamentos semelhantes, isto é, o resultado da análise de *clusters* é a identificação de grupos de dados em que cada elemento de um grupo é mais similar a outro de dentro do mesmo grupo que de outro elemento fora daquele grupo.

Em suma, o procedimento adotado para a imputação de dados neste trabalho foi o de se detectar variáveis com algum comportamento similar à variável que se desejava imputar

informações, tendo como critério de similaridade, o conhecimento empírico sobre o relacionamento entre as variáveis em geral e a medida de correlação estatística entre elas, onde o pré-julgamento sobre a existência de alguma relação, aliado à significância da medida estatística de correlação entre as variáveis, determinou qual ou quais variáveis deveriam ser utilizadas para a análise de *clusters*. Após a obtenção dos agrupamentos formados a partir da referida análise, foi computada a média da variável a ter seus valores faltantes imputados dentro de cada grupo, e a estes valores faltantes foram atribuídos os valores da média observada em seus respectivos grupos. Para ajudar a elucidar este procedimento, a seguir é apresentado o roteiro utilizado neste processo.

a) imputação dos dados faltantes para a variável *População economicamente ativa (PEA)*:

1. Foram identificadas as variáveis *% população urbana* e *% população alfabetizada*, que além de terem uma expressiva relação com a variável *PEA*, apresentaram as correlações estatísticas mais significativas (p-valor < 0,001) com esta variável.
2. Realizou-se então a análise de *clusters* com as variáveis *% população urbana* e *% população alfabetizada*.
3. A partir do passo anterior foram obtidos 11 grupos com alta homogeneidade interna e alta distinção entre grupos.
4. Para cada grupo referido anteriormente computou-se a média, isto é, foram calculadas 11 médias, uma para cada um dos 11 grupos respectivamente.
5. Fez-se a imputação registrando em qual grupo uma dada observação com dado faltante estava inclusa e a ela foi atribuído o valor da média do grupo em que ela se enquadrou.
6. Repetiu-se o processo descrito acima sucessivamente até que todas as observações faltantes tivessem seus valores imputados.

b) imputação dos dados faltantes para a variável *Médicos / mil habitantes*:

Executou-se o mesmo procedimento descrito no item (a), apenas com alguns diferenciais, a citar, no passo 1 a variável escolhida por afinidade e alta significância da medida de correlação (p-valor < 0,001) com a variável aqui

tratada foi a variável *População total*; no passo 3 foram obtidos 6 grupos com as mesmas características citadas no item (a), e no restante, tudo ocorreu nos mesmos moldes do processo apresentado acima, resguardando as devidas modificações em função de que aqui foi utilizada a variável *População total* na análise de *clusters*.

3.4 Obtenção do Modelo Estatístico

Efetuada a mineração e formatação dos dados a serem utilizados no processo de modelagem, deu-se início a uma etapa crucial neste trabalho, que é aquela onde os dados são examinados extensivamente por técnicas estatísticas, as quais buscam diagnosticar um comportamento padrão nas características municipais (variáveis explicativas) que possuem a propriedade de predizer o desempenho a ser observado em uma nova unidade de negócio recém aberta em um município qualquer.

Conforme o exposto no capítulo 2.2, a análise de regressão logística possui o ferramental teórico completo para o atendimento dos objetivos deste estudo, sendo esta a técnica aqui utilizada através do software SPSS, chegando-se à obtenção de um modelo estatístico, composto de variáveis preditoras criteriosamente selecionadas, habilitado a predizer a probabilidade de sucesso em movimentos de expansão de negócios da instituição financeira S.

Em modelagem estatística, o princípio da parcimônia, isto é, o menor número de variáveis possíveis, é sempre considerado um primeiro requisito a ser atendido, isto porque dessa forma o modelo mais parcimonioso adquire características desejáveis como a baixa dependência dos dados observados, um menor erro padrão e, portanto, uma estabilidade numérica suficiente para tornar o modelo altamente generalizável. Todavia a vasta literatura sobre modelagem também é flexível em aceitar que variáveis com alta relevância conceitual sejam incluídas no modelo, independentemente de serem ou não estatisticamente significativas.

3.4.1 Seleção de Variáveis Preditoras

Segundo Klück (2004), ao se fazer uso de softwares estatísticos para a realização da análise de regressão logística, algumas alternativas são apontadas para a seleção das variáveis preditoras, sendo que dentre essas possíveis abordagens uma delas é o método manual de inclusão ou exclusão de variáveis no modelo. Este método, cujo uso é bastante recomendado

em pesquisas médicas, consiste basicamente em avaliar passo a passo a significância e relevância de uma nova variável a ser incluída no modelo, assim como avaliar pelos mesmos critérios supracitados a exclusão de uma dada variável. Se no método de seleção manual a interação do pesquisador é total, em uma segunda abordagem feita através do método *stepwise* ocorre um processo mais automatizado, isto é, aqui neste método de seleção as variáveis do modelo são obtidas de forma automática, fazendo uso apenas de critérios estatísticos calculados automaticamente em cada etapa interna do procedimento executado pelo software estatístico. Ainda no método *stepwise* são possíveis dois modos de seleção de variáveis, o modo *forward* e modo *backward*. Na modalidade *forward*, a cada etapa uma variável é adicionada ao modelo, onde a primeira variável incluída por este processo é aquela mais significativa e, por este mesmo critério, as demais variáveis são submetidas a esta avaliação para entrar ou não no modelo. Já na modalidade *backward*, todas as variáveis disponíveis e escolhidas pelo pesquisador para o estudo são selecionadas para o modelo inicial, e a variável menos significativa é retirada, repetindo-se esta exclusão sequencialmente até que, assim como no modo *forward*, só restem variáveis estatisticamente significativas no modelo final.

Seguindo as recomendações da literatura de regressão logística, para este trabalho foi inicialmente realizada uma análise de correlação entre cada um das variáveis propostas para o estudo e a variável resposta, onde foi observado o nível de significância estatístico para cada valor de correlação, adotando o critério de levar para a etapa de modelagem, variáveis com p-valor $< 0,25$. A partir deste critério verificou-se que todas as 40 variáveis propostas para o estudo apresentaram um p-valor abaixo deste nível, devendo então todas elas serem disponibilizadas para a etapa de modelagem.

Para a seleção das variáveis preditoras deste estudo foi realizado inicialmente um procedimento que visou integrar requisitos básicos com os mais variados pontos relevantes dos métodos tradicionais de seleção de variáveis explicativas. Neste sentido, através do software estatístico SPSS executou-se o método *stepwise* na sua modalidade *backward*, que sugeriu algumas variáveis as quais apresentaram certas incongruências conceituais em seus coeficientes, além de variáveis com presença de colinearidade. Para contornar a falha observada no método automatizado do SPSS, migrou-se para o método de seleção manual, onde cada variável foi pré-avaliada estatisticamente e analisada a sua contribuição, em termos conceituais, para a construção de um modelo preditivo com alto valor estatístico.

3.4.2 Modelagem

Em continuidade ao processo de obtenção do modelo preditivo mais adequado, embora três modelos iniciais tenham sido obtidos através do método de seleção de variáveis *stepwise backward*, todos foram reprovados em suas avaliações, o que abriu porta para a obtenção de um modelo gerado através do método manual de seleção de variáveis explicativas, resultando em um modelo com características expostas no Quadro 2. Além do mais, o trabalho de modelagem aqui descrito foi baseado em 827 observações, das quais 484 (59%) apresentaram valor *zero* na variável resposta e consequentemente 343 (41%) apresentaram valor *um*.

Quadro 2 – Modelo estatístico final

Pilar	Variável	Coeficiente	Significância	Odds Ratio
Economia	Volume de crédito	1,153	0,000	3,168
	Agronegócio	0,377	0,105	1,458
	Indústrias	0,587	0,021	1,799
	% pop. economicamente ativa	0,279	0,190	1,321
	Índice de potencial de consumo	0,223	0,465	1,249
	Concorrentes / mil habitantes	-1,786	0,055	0,168
Demografia	% pop. urbana	0,017	0,003	1,017
	% Domicílios urbanos classe B	0,372	0,227	1,451
	% Domicílios urbanos classes D e E	0,440	0,139	1,553
	% pop. entre 20 e 29 anos	0,253	0,265	1,288
Infraestrutura	Distancia da capital	-0,204	0,249	0,815
	População / área plantada	-0,417	0,044	0,659
Social	IFDM saúde	-0,306	0,110	0,736
	IFDM emprego	0,100	0,597	1,106
	% pop. alfabetizada	0,195	0,371	1,215
	Constante	-2,891	0,000	0,056

Fonte: Autor

Algumas propriedades do modelo composto pelas variáveis apresentadas no Quadro 2 podem ser vistas na tabela abaixo.

Tabela 1 – Propriedades estatísticas modelo final

-2 Log Verossimilhança	Teste Hosmer & Lemeshow
863,663	0,793

Fonte: Autor

Conforme a Tabela 1, a estatística -2 log verossimilhança tem sua utilidade principal para a comparação de modelos distintos, o que de fato se mostrou superior à estatística dos outros modelos descartados, isto é, apresentou um valor absoluto menor que o observado nos

outros modelos. Também consta na Tabela 1 o resultado do teste de Hosmer e Lemeshow, o qual mede quão próximo estão os valores preditos pelo modelo em relação aos valores observados na amostra, onde a hipótese nula deste teste é de que não há diferença entre os valores preditos e os observados. Como para o presente modelo o teste de Hosmer e Lemeshow apresentou um p-valor de 0,793 indicando que não há diferença significativa, a um nível de confiança de 95%, entre os valores preditos e os observados, torna-se consistente a conclusão de se ter aqui um modelo suficientemente apropriado para representar os dados deste estudo, uma vez que os significados conceituais das variáveis apresentadas no Quadro 2 confirmam que suas permanências no modelo são apropriadas.

3.4.3 Diagnóstico do Modelo Final

Concluindo a etapa de modelagem, além da verificação do percentual de acertos que o modelo final obteve ao ter seus valores preditos comparados com os valores observados nos dados do estudo, realizou-se a análise de sensibilidade e especificidade do modelo através da análise da curva ROC.

Segundo Klück (2004), o uso da curva ROC (*Receiver Operating Characteristics*) está relacionado à determinação da acurácia do modelo, isto é, ao poder de discriminação que o modelo tem e que, em outras palavras, pode ser definido como a habilidade em distinguir adequadamente os casos de sucesso daqueles de não sucesso.

A origem da curva ROC está na teoria de detecção de sinal, onde a curva tem a função de indicar o modo de operação de um receptor na existência de um sinal com presença de ruído, plotando a probabilidade de detectar um sinal verdadeiro (sensibilidade) e um sinal falso (1-especificidade) para um conjunto de possíveis pontos de corte. Dessa forma, a área encontrada abaixo da curva ROC, a qual varia entre 0 e 1, irá determinar a habilidade que o modelo tem em discriminar quais municípios apresentam os fatores viabilizadores do sucesso na abertura de uma nova unidade de negócio, daqueles que não apresentam tais fatores.

Tomando como referência os critérios encontrados na literatura, de forma pragmática, pode-se estabelecer (Tabela 2) a seguinte relação para os possíveis valores da área sob a curva ROC:

Tabela 2 – Regra geral para avaliação da acurácia do modelo

Área sob curva ROC	Avaliação
ROC = 0,5	Não há discriminação, resultado obtido ao acaso
0,5 ≤ ROC < 0,7	Resultado não aceitável, pequena relevância
0,7 ≤ ROC < 0,8	Discriminação aceitável
0,8 ≤ ROC < 0,9	Discriminação excelente
ROC ≥ 0,9	Discriminação destacada, quase perfeita

Fonte: Klück, 2004, p. 85.

Sendo assim, foi realizado no software SPSS a análise da curva ROC, Figura 4, a qual evidenciou, através de uma área sob a curva de 0,811 - um poder de discriminação excelente para modelo estatístico final.

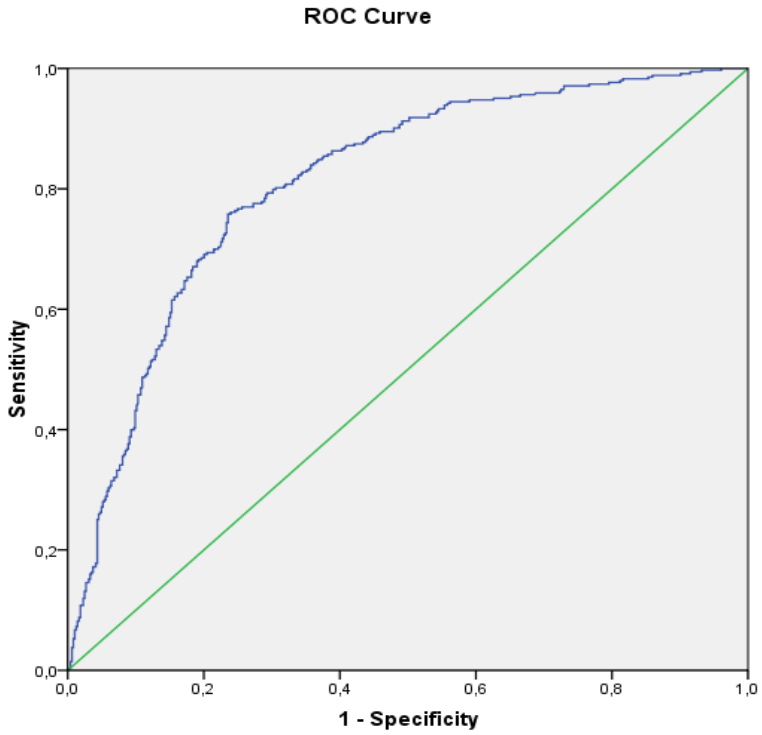

Figura 4 – Plotagem da curva ROC com área abaixo igual a 0,811.

Fonte: Autor

Conforme citado anteriormente, foi computada a tabela de classificação do modelo de regressão logística, num esforço de se gerar um outro tipo de avaliação complementar da adequação do modelo, a qual é apresentada a seguir na Tabela 3.

Tabela 3 – Tabela de classificação do modelo de regressão logística

			Valores preditos		
			Variável resposta		%
			0	**1**	**Correto**
Valores observados	Variável resposta	**0**	376	108	77,7
		1	102	241	70,3
% Correto geral					**74,6**

Fonte: Autor

A partir da análise da Tabela 3, que evidencia uma boa acurácia geral (74,6 %), uma questão importante atendida é a congruência das medidas de sensibilidade (77,7 %) e especificidade (70,3 %), isto é, o modelo obtido neste trabalho desempenha bem tanto a função de detectar a região geográfica viabilizadora de sucesso quanto aquela ausente em atributos relacionados ao bom desempenho dos negócios.

Ainda vale citar que a leitura dos valores da Tabela 3 pode ser feita da seguinte forma: dos 484 valores *zero* observados nos dados, em 376 deles o modelo os predisse corretamente, assim como nos 343 valores *um* observados, em 241 deles o modelo também os assinalou como sendo valor *um*.

Finalmente, pode-se anunciar que o modelo selecionado para este trabalho é composto da seguinte forma:

Y = -2,89 + 1,15*(volume de crédito) + 0,38*(agronegócio) + 0,59*(indústrias) + 0,28*(% população economicamente ativa) + 0,22*(índice de potencial de consumo) - 1,79*(concorrentes por mil habitantes) + 0,02*(% população urbana) + 0,37*(% domicílios urbanos de classe B) + 0,44*(% domicílios urbanos de classes D e E) + 0,25*(% população entre 20 e 29 anos) – 0,20*(distância da capital) – 0,42*(população por área plantada) – 0,31*(IFDM saúde) + 0,10*(IFDM emprego) + 0,20*(% população alfabetizada)

Onde para as variáveis *"% população urbana"* e *"concorrentes por mil habitantes"* entram no modelo os seus respectivos valores brutos, e para as demais variáveis atribui-se valor *zero* ou *um*, dependendo de onde seu valor nominal se encontre em relação ao ponto de corte de cada variável.

Capítulo 4

Resultados

De posse de um modelo estatístico habilitado a predizer a probabilidade de sucesso ao se inaugurar uma nova unidade de atendimento em um dado município, efetuou-se o uso deste modelo computando os dados da cada município do Brasil de acordo com as variáveis estabelecidas na etapa de modelagem, obtendo dessa forma um valor de probabilidade de sucesso em uma ação de expansão para cada cidade brasileira. Após a geração desta listagem de probabilidades, plotou-se este resultado em mapas temáticos por intermédio do software MapInfo obtendo como produto final visualizações gráficas das melhores oportunidades de expansão para a instituição financeira S.

4.1 Interpretação do Modelo

Recorrendo à ressalva feita no início do Capítulo 3, a compreensão das variáveis selecionadas para o modelo final assim como a interpretação de seus respectivos coeficientes, restringem-se a fornecer a leitura do mercado em que a empresa S está inserida, conjugando o seu modo de atuação, isto é, seu modelo cooperativista de negócio, com o perfil dos usuários de serviços financeiros aderentes a esse modo de atuação.

Dessa forma, no intuito de colaborar com o melhor entendimento do modelo obtido, cabe aqui postular três sentenças sobre o conglomerado financeiro S:

- A instituição tem suas raízes históricas originadas no meio rural;
- Atualmente, é um fato a sua baixa presença nos grandes centros urbanos;
- Existe ainda uma baixa difusão da cultura cooperativista nos meios urbanos;

Neste sentido, vale dizer que as variáveis que mais pesam no resultado da predição são as variáveis ”*volume de crédito*”, “*concorrentes por mil habitantes*” e “*indústrias*”. Explorando um pouco mais estas variáveis temos que para a primeira citada, os municípios que apresentam um volume de crédito acima do ponto de corte estabelecido, têm cerca de 3 vezes mais chances de sucesso quando comparados àqueles com um volume de crédito demandado abaixo do referido ponto de corte. Já no caso da segunda variável mencionada,

ocorre um movimento inverso, isto é, à medida que se observa o acréscimo de um concorrente para cada mil habitantes, as chances de sucesso em uma nova abertura de unidade são reduzidas em cerca de 83,2%, e, para o caso da terceira variável citada, quantidade total de indústrias, a existência de um número de indústrias acima do ponto de corte determinado para esta variável, propicia uma melhora nas chances de sucesso em um município na ordem de 1,8 vezes, se comparado com as chances de sucesso em um outro município onde o número de indústrias está aquém do ponto de corte para esta variável.

Outro entendimento relevante sobre o modelo final obtido refere às variáveis "*distância da capital*" e "*população por área plantada*", onde ambas figuram no modelo com sinal negativo em seus coeficientes, indicando que a ultrapassagem de seus respectivos pontos de corte acarreta em uma redução nas chances de sucesso quando comparadas com as chances associadas à municípios em que os valores destas variáveis encontram-se abaixo dos referidos pontos de corte. Em outras palavras pode-se dizer que quanto mais perto da capital do estado estiver o município, ou, quanto maior for a quantidade de área plantada por habitante, melhores serão as chances de sucesso quando comparadas com o caso oposto.

Corroborando com o perfil da instituição financeira S, a variável "*agronegócio*", que representa a quantidade de empresas no município que atuam no ramo do agronegócio, desempenha no modelo um papel de potencializar as chances de sucesso na abertura de novas unidades, em municípios onde se observa uma quantidade destas empresas acima do ponto de corte, aumentando estas chances em cerca de 1,5 vezes em relação aos casos em que o volume de empresas de agronegócios é menor que o ponto de corte supracitado.

4.2 Mapeamento do Potencial Mercadológico

Nesta etapa é apresentado um exemplo real da aplicação dos resultados obtidos a partir deste trabalho. A realização deste caso aplicado ocorreu congregando-se a base de dados dos municípios brasileiros e as informações referentes à atual área de abrangência da instituição financeira S.

Dessa forma, foram computados 709 municípios com alto potencial mercadológico, onde este alto potencial foi atribuído aos municípios cuja probabilidade de sucesso predita pelo modelo foi de 0,7 ou mais (máximo 1). Fazendo-se a intersecção destes 709 municípios com aqueles contidos na área de abrangência da empresa S, foram relacionados 346 (49%) municípios presentes nesta intersecção. Elucidando melhor a relação entre a quantidade de

municípios e suas respectivas probabilidades de proporcionar a observação de um bom desempenho em uma nova unidade de atendimento, é apresentada a seguir a Figura 5.

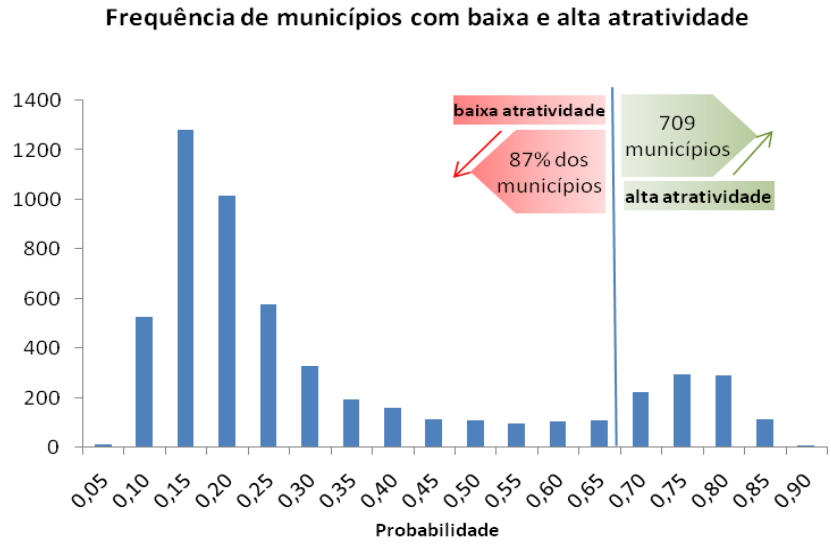

Figura 5 – Distribuição da ocorrência de municípios com baixa e alta atratividade.

Fonte: Autor

Agora, materializando uma parcela dos resultados obtidos com o estudo, apresenta-se no Quadro 3 a lista com o *rank* dos 30 municípios mais atrativos, nos termos deste trabalho.

Quadro 3 – Lista com os 30 municípios mais atrativos

Rank	UF	Município	Está na área de abrangência	Probabilidade
1	RS	CRUZ ALTA	✓	88,8%
2	GO	RIO VERDE	✓	88,3%
3	MG	CORONEL FABRICIANO		87,5%
4	MG	RIBEIRAO DAS NEVES		86,6%
5	MS	SAO GABRIEL DO OESTE		86,1%
6	MS	DOURADOS	✓	86,1%
7	GO	CATALAO		85,5%
8	RS	CACHOEIRA DO SUL	✓	85,1%
9	RS	ESTEIO	✓	85,0%
10	RS	CARAZINHO	✓	84,9%
11	SC	CAMPOS NOVOS	✓	84,9%
12	RJ	SEROPEDICA		84,4%
13	RS	VACARIA	✓	84,3%
14	RO	VILHENA	✓	83,9%
15	RJ	ARARUAMA		83,6%
16	SE	NOSSA SENHORA DO SOCORRO		83,5%
17	PE	PAULISTA		83,3%
18	PR	IRATI	✓	83,2%
19	MG	BRUMADINHO		83,2%
20	PA	ANANINDEUA		82,9%
21	RJ	SAO GONCALO		82,8%
22	RJ	SAO JOAO DE MERITI		82,7%
23	MT	TANGARA DA SERRA	✓	82,7%
24	BA	LAURO DE FREITAS		82,7%
25	RJ	NOVA IGUACU		82,7%
26	RS	SAO LOURENCO DO SUL	✓	82,6%
27	PR	SARANDI	✓	82,5%
28	PE	OLINDA		82,4%
29	GO	CERES		82,3%
30	PR	PALMEIRA	✓	82,3%

Sendo assim, a seguir na Figura 6, apresenta-se o mapa que representa a distribuição das oportunidades de negócios ao longo dos 5.564 municípios Brasil, isto é, como estão alocadas as probabilidades de sucesso ao se abrir novas unidades de atendimento. Neste e nos demais mapas apresentados, a legenda "*Baixa probabilidade*" refere-se à faixa de probabilidade que vai de 0 à 0,4 (exclusive), já a legenda "*Probabilidade indiferente*" trata-se do intervalo de probabilidade que vai de 0,4 à 0,55 (exclusive), "*Média probabilidade*" compreende as probabilidades de 0,55 à 0,7 (exclusive) e por fim a legenda "*Alta probabilidade*" refere-se à faixa de probabilidade que vai de 0,7 à 1.

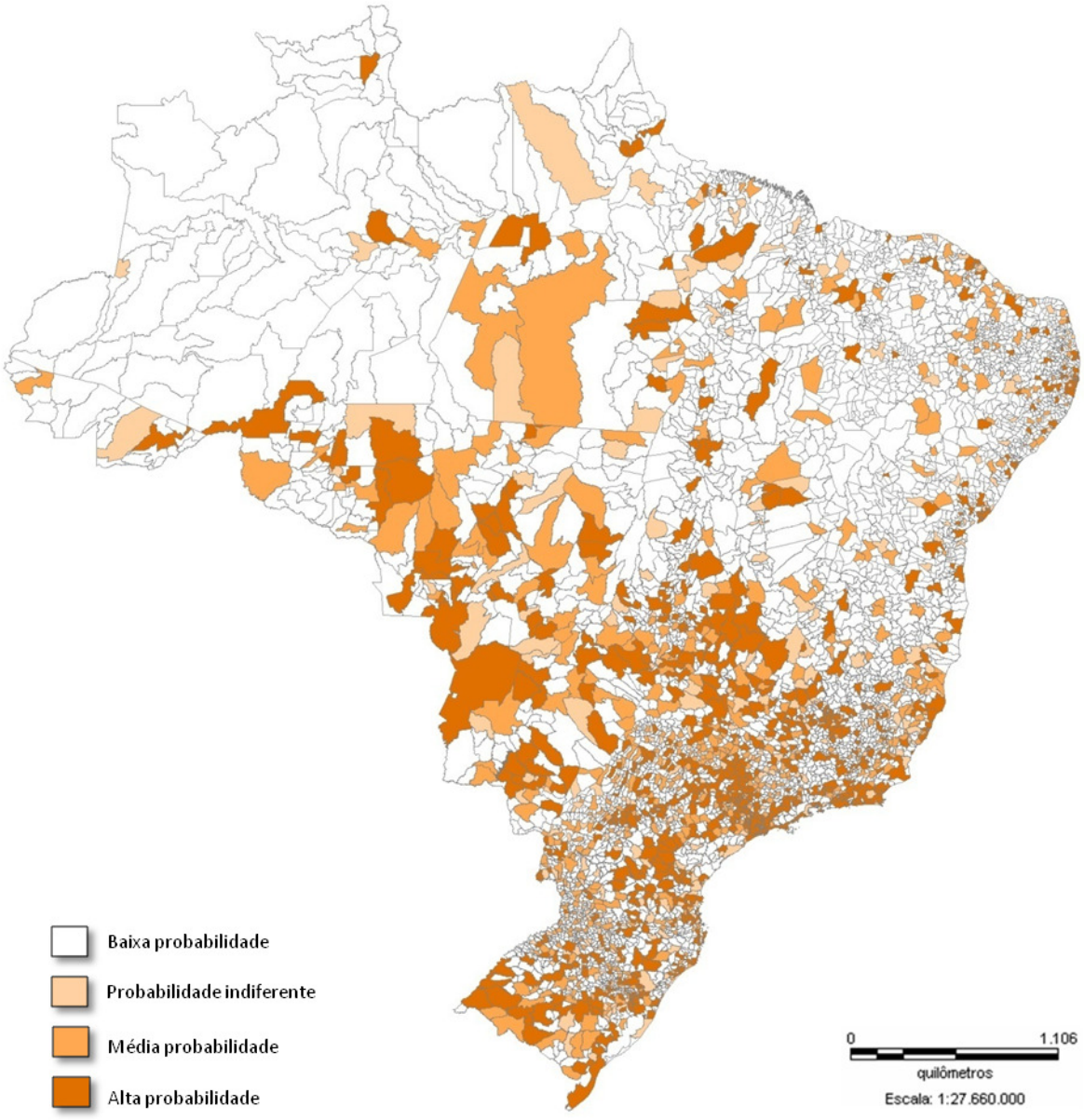

Figura 6 – Mapeamento das regiões com maior potencial mercadológico.

Fonte: Autor

A seguir, para melhor elucidar a configuração do cenário atual, a Figura 7 representa o mapa com a distribuição da área de abrangência do conglomerado financeiro S, a qual compreende 1.694 municípios.

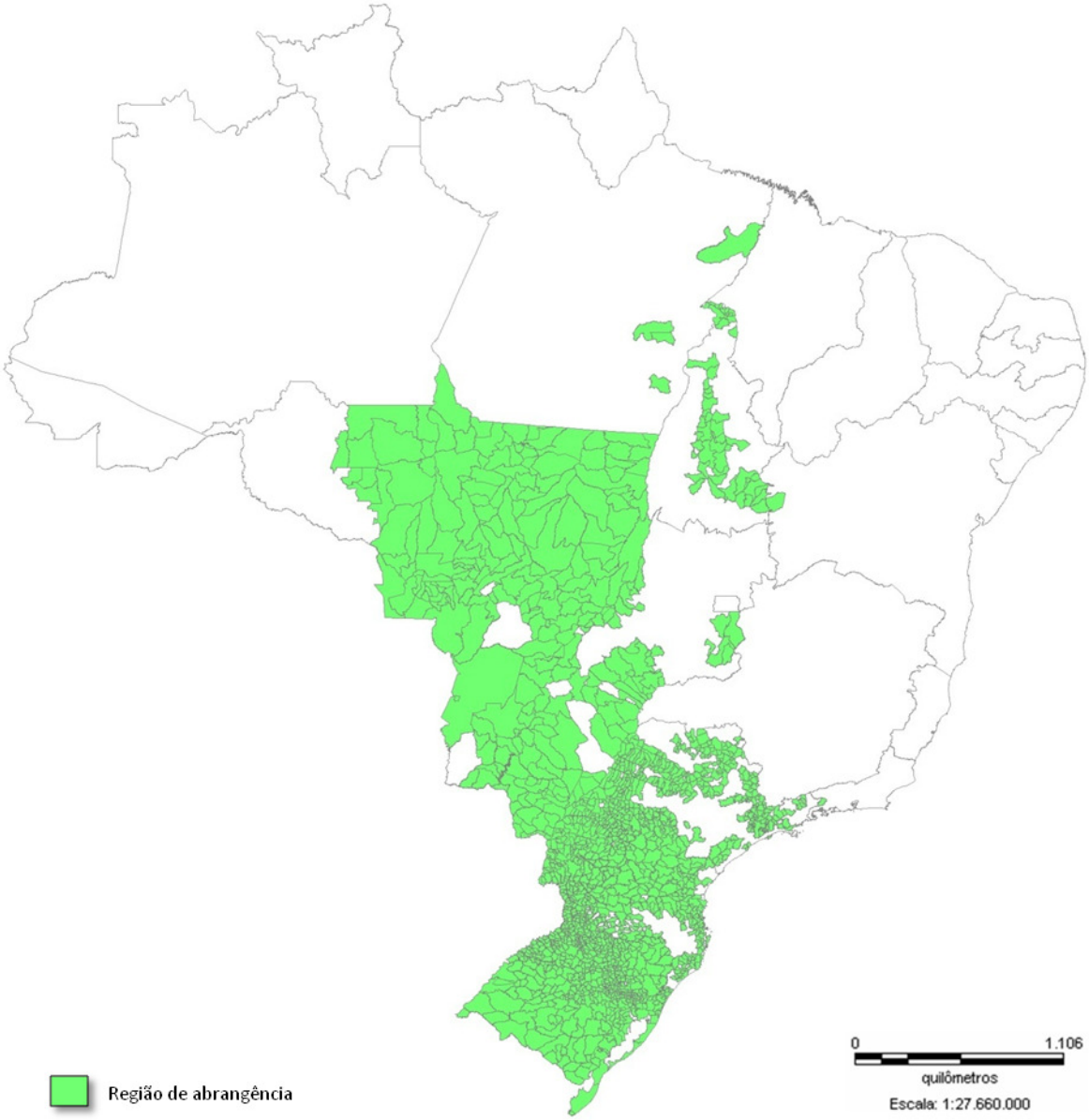

Figura 7 – Visualização da área de abrangência da instituição financeira S.
Fonte: Autor

Para fins de comparação e esclarecimento do cenário atual sob o prisma do potencial mercadológico, apresenta-se na Figura 8 uma progressão do mapa anterior, isto é, a área de abrangência da instituição financeira S acrescida da sinalização dos 346 municípios com alto potencial de mercado, contidas nesta área de abrangência.

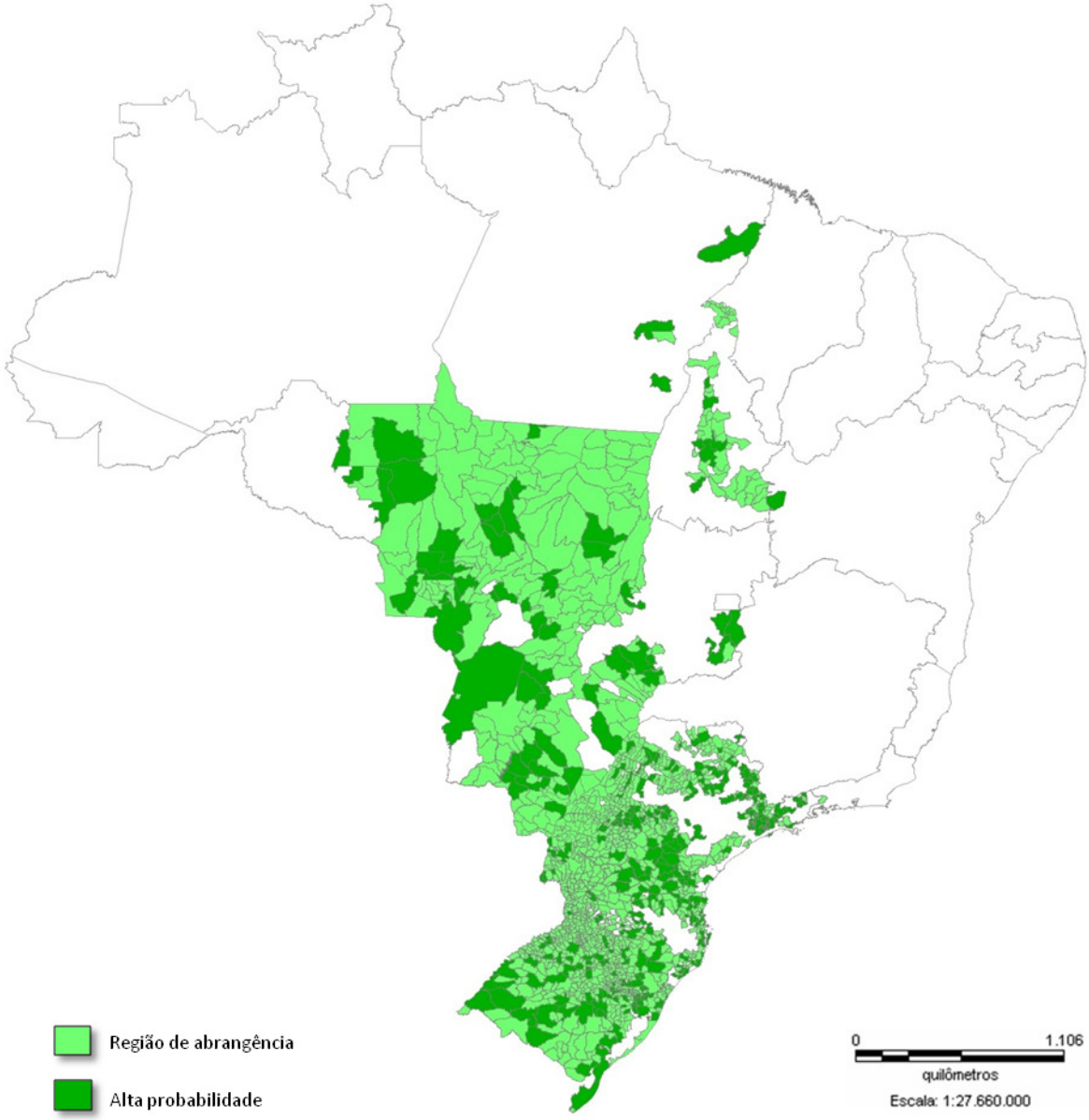

Figura 8 – Ocorrência de regiões com alto potencial mercadológico dentro da área de abrangência.

Fonte: Autor

Por fim, na Figura 9 tem-se uma visão panorâmica de como o conglomerado financeiro S está inserido geograficamente no Brasil e de como a sua área de abrangência tangencia as regiões (709 municípios) apontadas pelo modelo como sendo de alto potencial mercadológico, além da indicação de regiões fora das fronteiras da referida instituição financeira onde este alto potencial está presente.

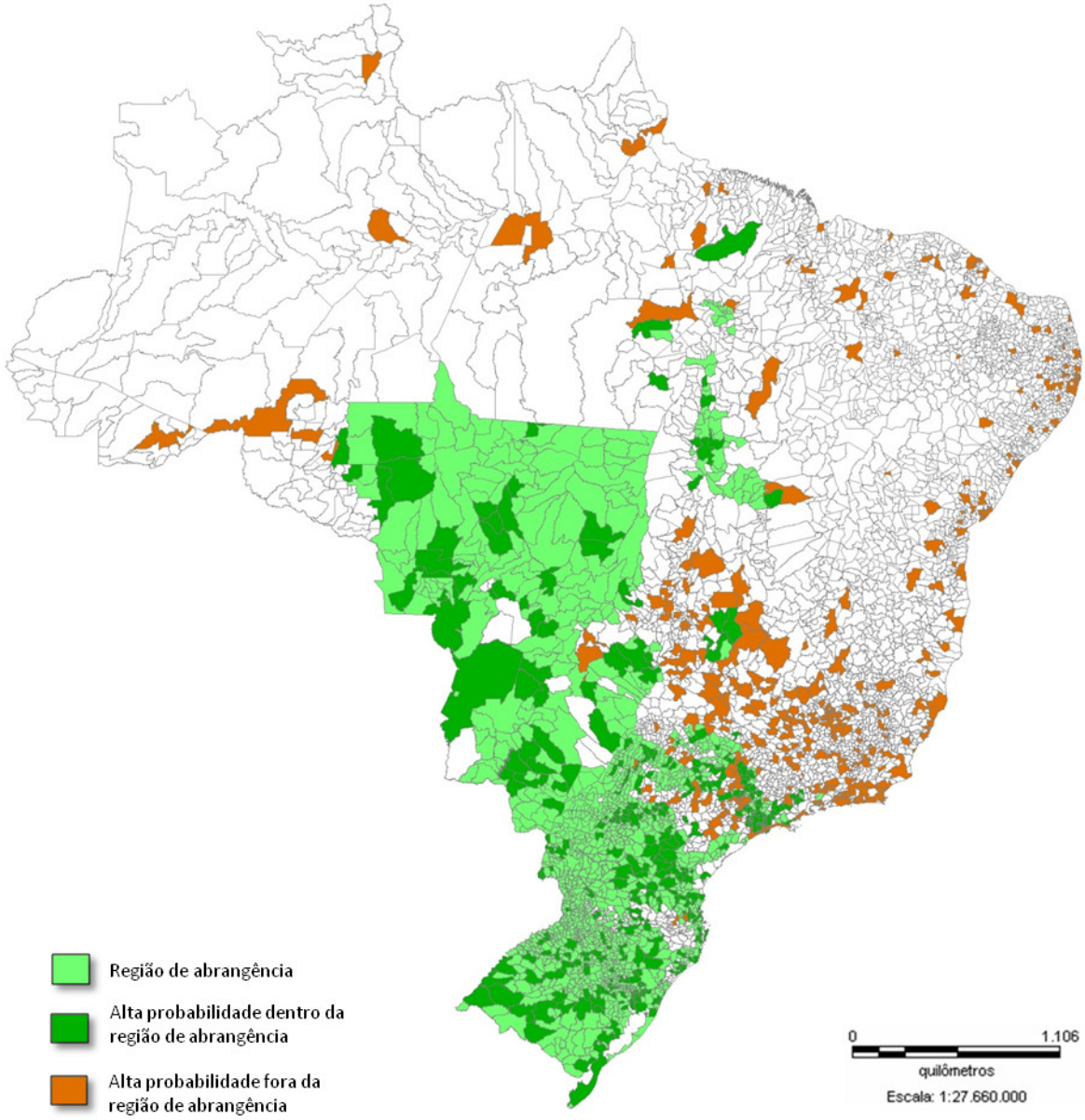

Figura 9 – mapeamento do potencial mercadológico dentro e fora da área de abrangência.

Fonte: Autor

Como se pode observar no mapa acima, existe uma ampla gama de oportunidades para movimentos de expansão além das fronteiras de abrangência, sendo que o leque de opções para abertura de novas unidades de atendimento em locais com alta probabilidade de sucesso chega próximo de 400 municípios.

Capítulo 5

Considerações Finais

Diante de um volume expressivo de conclusões e conhecimentos novos consolidados, gerados até aqui através deste estudo, e, dado o caráter estratégico que estas novas informações representam, alguns pontos devem ser enfatizados de modo a coibir o mau uso e a má interpretação dos resultados obtidos com este trabalho.

Nesta linha, é de extrema importância que durante todo tempo o entendimento da variável resposta adotada neste estudo esteja claro, isto pelo fato de ela ter sido definida como a soma do volume de negócios observado na unidade de atendimento com a rentabilidade nela observada. Implicitamente aqui se está definindo que ambas as parcelas desta soma apresentam pesos iguais, o que provavelmente impacta diretamente em todos os números envolvidos na análise de potencial mercadológico. Por outro lado, isto também significa que a composição da variável resposta pode ser, arbitrariamente ou com embasamento técnico, recalibrada para mapear os municípios que favoreçam mais o volume de negócios ou então mais a rentabilidade.

De forma geral, este trabalho não possui a pretensão de resolver de forma definitiva a extensa problemática da expansão comercial, porém busca sim introduzir uma nova forma de se tratar este assunto. Decorrente disso, acontece a partir desta obra a abertura de inúmeras portas para o aprimoramento daquilo que foi realizado através deste trabalho, a citar a possível e altamente relevante aplicação da metodologia aqui desenvolvida para o caso de setores censitários, onde se teria uma visão mais aproximada do potencial de mercado dentro de um único município.

Encerrando, vale recapitular a sistemática da análise utilizada neste estudo, a qual está centrada em olhar para o cenário atual da instituição financeira S, analisando este cenário, buscar a detecção dos padrões envolvidos nos casos onde se observou um bom desempenho, e, por fim, assumir que estes padrões detectados (ligação com o modelo final obtido), quando presentes em um município, proporcionarão uma chance maior de que seja observado também um bom desempenho tal como aquele observado na análise do cenário atual.

Referências

- AGRESTI, Alan. **Categorical Data Analysis**. 2ed. New Jersey: John Wiley & Sons, 2002. 721 p.

- ARAÚJO, Elaine Aparecida; CARMONA, Charles U. de Moutreuil. Construção de Modelos *Credit Scoring* com Análise Discriminante e Regressão Logística para a Gestão do Risco de Inadimplência de uma Instituição de Microcrédito. **Revista Eletrônica de Administração**, Porto Alegre, v. 15, n. 1, jan/abril, 2009.

- BATISTELA, Gislaine Cristina; RODRIGUES, Sergio Augusto; BONONI, Júlia T. Carrer Martinelli. Estudo sobre a evasão escolar usando regressão logística: análise dos alunos do curso de administração da Fundação Educacional de Ituverava. **Tékhne e Lógos**, Botucatu, v. 1, n.1, out. 2009.

- BRUNI, Eduardo Serur. **Uso de Regressão Logística para precificação de *credit default swaps***. 2007. 104 f. Monografia (Graduação em Engenharia de Produção) - Universidade de São Paulo, São Paulo, 2007.

- COSTER, Rodrigo. **Um alerta sobre o uso de amostras pequenas na regressão logística**. 2009. 25f. Monografia (Graduação em Estatística) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

- FIGUEIRA, Cleonis Viater. **Modelos de Regressão Logística**. 2006. 149f. Dissertação (Mestrado em Matemática) - Programa de Pós-Graduação em Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

- GIMENO, Suely G. Agostinho; SOUZA, José Maria Pacheco de. Utilização de Estratificação e Modelo de Regressão Logística na análise de dados de estudo caso-controle. **Revista de Saúde Pública**, São Paulo, v. 29, n. 4, ago. 1995.

- HARTMAN, Julia. **An Interactive Tutorial for SPSS 10.0 for Windows©**: Binomial Logistic Regression. Alabama / EUA, 2000.

- HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. 2 ed. New York: John Wiley & Sons, 2000. 375 p.

- KLUCK, Mariza Machado. **Metodologia para ajuste de indicadores de desfechos hospitalares por risco prévio do paciente**. 2004. 128 f. Tese (Doutorado em Epidemiologia) - Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

- SILVA, Eduardo A. R. **Regressão Logística**. 2010. 09 f. Dissertação (Graduação em Engenharia de Produção) - Universidade Federal de São Carlos, São Carlos, 2010.

- TSUCHIYA, Ítalo. **Regressão Logística aplicada na análise espacial de dados arqueológicos**. 2002. 103 f. Dissertação (Mestrado em Ciências Cartográficas) - Universidade Estadual Paulista, Presidente Prudente, 2002.

- WIKIPÉDIA. Disponível em: <http://pt.wikipedia.org/wiki/Regress%C3%A3o_log%C3%ADstica>. Acessado em: 4 out. 2010.

- WIKIPÉDIA. Disponível em: <http://pt.wikipedia.org/wiki/Logit>. Acessado em: 4 out. 2010.

- ZEILHOFER, Peter et al. SIG e Regressão Logística para mapeamento de risco de contaminação por pesticidas nos mananciais superficiais da bacia do Alto Rio das Mortes - MT. In: XIII Simpósio Brasileiro de Sensoriamento Remoto (SBSR), Florianópolis. **Anais**. São José dos Campos: INPE, 2007. P. 3623-3630.

- ALONSO, Joaquim Mamede; CASTRO, Pedro M. Ribeiro. SIG – Sistemas de Informação Geográfica: o Geobusiness como área de aplicação dos sistemas de informação geográfica e espaço de oportunidade. **Nexus revista empresarial**, Viana do Castelo, n. 2, p. 10-11, set. 2006.

- CAVION, Renata; PHILIPS, Jürgen. Os Fundamentos do Geomarketing: Cartografia, Geografia e Marketing. In: Congresso Brasileiro de Cadastro Técnico Multifinalitário, Florianópolis. **Anais**. Florianópolis: COBRAC, 2006.

- COSTA, Átila Mendes; NEVES, João Adamor Dias. Geomarketing e pequenas empresas: análise espacial dos postos de combustível de Fortaleza. In: II Encontro de Marketing da ANPAD, Rio de Janeiro. **Anais**. Rio de Janeiro: ANPAD, 2006.

- ELIAS, Wanda Luquine. **Segmentação geodemográfica:** modelos mentais dos profissionais do ramo imobiliário de Presidente Prudente/SP e seus influenciadores *versus*

modelo com dados oficiais gerados a partir do geomarketing. 2009. 142 f. Dissertação (Mestrado em Administração de Organizações) - Programa de Pós-Graduação em Administração, Universidade de São Paulo, Ribeirão Preto, 2009.

- FAGUNDES, André Francisco Alcântara et al. Geomarketing: Um Estudo de Caso de Uma Empresa de Telecomunicações. In: III Encontro de Marketing da ANPAD, Curitiba. **Anais**. Rio de Janeiro: ANPAD, 2008.

- FIGUEIREDO, Willian Augusto de. Geomarketing aplicado à instituições educativas. **Estudos & Pesquisas**, Lins, v. 9, n. 1, p. 157-169, jun. 2006.

- FREITAS, Eduardo; CORRET, Fabíola. Novas tendências do geomarketing: Saiba como acertar na mosca usando análise geográfica. **Revista InfoGEO**, Curitiba, ed. 59, jan/abril, 2010.

- GREGORI, Reinaldo. Geomaketing ganha espaço nas empresas. [23 de agosto, 2010]. São Paulo: **Jornal DCI - Comércio, Indústria e Serviços**. Entrevista concedida a Roberto Müller.

- GREGORI, Reinaldo G. Insight' 09: novidades em tecnologia e metodologia no mundo do geomarketing. **Revista InfoGEO**, Curitiba, ed. 57, jul/set, 2009.

- IBRAHIM, Rafael. **Seminário mostra importância do geomarketing nos negócios**. Disponível em: <http://geoeasy.com.br/blog/?p=505>. Acesso em: 4 out. 2010.

- LOURENÇO, Fátima. **Geomarketing ganha espaço no franchising brasileiro**. Disponível em: <http://www.dcomercio.com.br/materia.aspx?id=49240&canal=70>. Acesso em: 28 set. 2010.

- MARQUEZ, Rafael; BRITO, Rodrigo. **Implementing Geomarketing as a Strategic Planning Tool**. 2006. Trabalho apresentado na ESRI International User Conference, San Diego, ago. 2006.

- MONTEIRO, Jaimar de Barros. **Indicador de Criminalidade Geral Baseado em Métodos Multivariados e Estatística Espacial para Controle do Estado**. 2009. 85 f. Monografia (Graduação em Estatística) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

- PAESE, Cíntia. Impactos da Inteligência Estratégica na Orientação para o Mercado e Desempenho das Organizações. In: III Encontro de Marketing da ANPAD, Curitiba. **Anais**. Rio de Janeiro: ANPAD, 2008.

- REIS, Solange Garcia; NEVES JÚNIOR, Idalberto José das; MORGAN, Beatriz Fátima. Definição de Metas para Avaliação de Desempenho de Agências Bancárias. **Revista de Administração Mackenzie**, São Paulo, v. 8, n. 4, out/dez, 2007.

- RISSOLI, V. C.; MENDES, A. A. Geomarketing: potencialidades de mercado e viabilidade para a implantação de uma franquia do ramo *fast-food* no município de Rio Claro/SP. In: IX Semageo – Semana de Geografia, Rio Claro. **Anais**. Rio Claro: Unesp, 2008.

- SÁ, Sylvia de. **O que é e como se faz Geomarketing**. Disponível em: <http://www.mundodomarketing.com.br/8,12357,o-que-e-e-como-se-fazgeomarketing.htm>. Acesso em: 4 out. 2010.

- SILVA, Sandra Maria; SILVA, Wesley Vieira; CORSO, Jansen Maia Del; DUCLÓS, Luiz Carlos. Segmentação de mercado: análise do perfil sócio-econômico dos municípios do Paraná. **Revista GEPEC**, Toledo, v.10, n.2, jul/dez, 2006. P. 9-28.

- SOBRINHO, Alfredo O. de Macedo. **A geotecnologia como ferramenta de marketing para otimizar o processo de segmentação geográfica de mercados e a localização de empreendimentos comerciais**. 2008. 35 f. Monografia (Graduação em Administração com habilitação em Marketing) - Faculdade do Pará, Belém, 2008.

- TERRA, Thiago. **Diageo, Coca-Cola e Dufry ampliam pontos de contato:** Empresas apresentam cases durante o Seminário Marketing 360º no Rio. Disponível em: <http://www.mundodomarketing.com.br/10,9619,diageo-coca-cola-e-dufry-ampliampo ntos-de-contato.htm>. Acesso em: 28 set. 2010.

- TISCOSKI, Rogério. **Geografia + Marketing = Geomarketing**. Disponível em: <http://www.administradores.com.br/informe-se/artigos/geografia-marketinggeomarketing/47799/>. Acesso em: 28 set. 2010.

- WALTER, Silvana Anita et al. Lealdade de Estudantes de uma Instituição de Ensino Superior: um Modelo de Regressão Logística para curso de Administração. In: III Encontro de Marketing da ANPAD, Curitiba. **Anais**. Rio de Janeiro: ANPAD, 2008.