Universidade de Brasília
Faculdade de Administração, Contabilidade e Economia
Departamento de Administração

MARCELO FERNANDO FELIX DE OLIVEIRA

ANÁLISE DE MERCADO: Uma ferramenta de mapeamento de oportunidades de negócio baseada em técnicas de Geomarketing e Aprendizado de Máquina

Brasília – DF
2016

Universidade de Brasília

Faculdade de Administração, Contabilidade e Economia

Departamento de Administração

MARCELO FERNANDO FELIX DE OLIVEIRA

ANÁLISE DE MERCADO: Uma ferramenta de mapeamento de oportunidades de negócio baseada em técnicas de Geomarketing e Aprendizado de Máquina

Projeto de monografia apresentado ao Departamento de Administração como requisito parcial à obtenção do título de Bacharel em Administração.

Professor Orientador: Doutor, Pedro Henrique Melo Albuquerque

Brasília – DF

2016

MARCELO FERNANDO FELIX DE OLIVEIRA

ANÁLISE DE MERCADO: Uma ferramenta de mapeamento de oportunidades de negócio baseada em técnicas de Geomarketing e Aprendizado de Máquina

A Comissão Examinadora aprova o Trabalho de Conclusão de Curso de Administração da Universidade de Brasília do aluno:

Marcelo Fernando Felix de Oliveira

Doutor, Pedro Henrique Melo Albuquerque
Professor-Orientador

Pedro Alexandre Moura Barros Henrique
Professor-Examinador

Peng Yao Hao
Professor-Examinador

Brasília, 23 de junho de 2016.

Dedico este trabalho, e tudo que ele representa para mim e para o meu futuro, aos meus pais. Tenho plena consciência de que eu não seria capaz de me formar em uma boa faculdade e me tornar a pessoa que sou hoje se não fosse pelo amor, dedicação e luta que tiveram no decorrer de toda a minha vida. Sei que nem sempre foi fácil e que vocês já passaram por muitos momentos difíceis, mas venceram todos eles e eu espero ter herdado ao menos um pouco do coração e da força de vontade de vocês.

"The principle is competing against yourself. It's about self-improvement, about being better than you were the day before."

(Stevie Young)

RESUMO

O objetivo deste estudo é a elaboração de uma ferramenta quantitativa, baseada em técnicas de *Geomarketing* e Aprendizado de Máquina, capaz de identificar oportunidades de negócio e contribuir para o processo estratégico de escolha locacional de uma rede de franquias, selecionando regiões que possuam uma alta previsão de demanda e uma carência na oferta do produto. Além disso, realizou-se uma análise qualitativa dos pontos comerciais selecionados com base em critérios definidos no decorrer do trabalho. Essa previsão se dá pela construção de um padrão de consumo, definido por um classificador, baseado nas características dos indivíduos que costumam comprar o produto. Inicialmente, para um melhor entendimento a respeito do assunto, foi feito um embasamento teórico abarcando os principais conceitos sobre *Geomarketing* e Aprendizado de Máquina e suas aplicações. Em seguida, para a demonstração dos resultados, optou-se pela aplicação do método para o mercado de chocolates finos (Cacau-Show) no Distrito Federal. As principais bases de dados utilizadas neste trabalho foram provenientes da Pesquisa de Orçamentos Familiares e do Censo Demográfico, ambos desenvolvidos pelo Instituto Brasileiro de Estatística e Geografia (IBGE). Como resultado, obteve-se o Gasto Padronizado (GP), que indica o nível de demanda para cada Setor Censitário, as informações georreferenciadas da concorrência, contendo 44 lojas que possuem como principal produto o chocolate fino, e as malhas digitais do Distrito Federal. O cruzamento dessas informações permitiu a identificação das oportunidades de negócio para o mercado de chocolates finos no Distrito Federal.

Palavras-chave: 1. Geomarketing. 2. Aprendizado de Máquina. 3. Chocolate

ABSTRACT

The aim of this study is to develop a quantitative tool, based on techniques of Geomarketing and Machine Learning, able to identify business opportunities and contribute to the strategic process of locational choice of a franchise network by selecting regions that have a high forecast demand and a shortage in supply of the product. In addition, there was a qualitative analysis of commercial sites selected based on criteria set out in the course of work. This prediction is by building a pattern of consumption, defined by a classifier based on the characteristics of individuals who usually buy the product. Initially, for a better understanding of the subject, it was made a theoretical framework covering the main concepts of *geomarketing* and Machine Learning and its applications. Then, to the income statement, we opted for the application of the method to the market of chocolates (Cacau-Show) in the Distrito Federal. The main databases used in this study were from the Family Expenditure Survey and the Population Census, both developed by the Instituto Brasileiro de Geografia e Estatística (IBGE). As a result, we obtained the Spent Standardized (GP), which indicates the level of demand for each Census Sector, the georeferenced information of competition, containing 44 shops that have the chocolate as their main product, and digital maps of Distrito Federal. The crossing of this information allowed the identification of business opportunities for the market of chocolates in the Distrito Federal.

Keywords: 1. Geomarketing. 2. Machine Leraning. 3. Chocolate.

LISTA DE EQUAÇÕES

Equação 1 - Erro Médio Quadrado.....38
Equação 2 - Gasto Padrão.....40

LISTA DE FIGURAS

Figura 1 - Faturamento em Bilhões (2010 a 2014).....11
Figura 2 - Geração de empregos diretos (2010 a 2014) 11
Figura 3 - Conjunto de exemplos no formato atributo-valor 26
Figura 4 - Processo de classificação por meio do Aprendizado de Máquina 27
Figura 5 - Separação dos exemplos pelo classificador 28
Figura 6 - Maximização entre as fronteiras de cada classe 28
Figura 7 - Mapeamento do consumo de chocolate no Brasil 33
Figura 8 - Esquematização do método de aprendizagem de máquina 40
Figura 9 - Curva de distribuição normal e o Desvio-Padrão.....41

LISTA DE MAPAS

Mapa 1 - Legenda e disposição geral 59
Mapa 2 - Subdistrito Brasília: Indicação urbana 60
Mapa 3 - Subdistrito Brasília: Análise de mercado 60
Mapa 4 - Subdistrito Brazlândia: Indicação urbana 61
Mapa 5 - Subdistrito Brazlândia: Análise de demanda 61
Mapa 6 - Subdistrito Candangolândia: Indicação urbana 62
Mapa 7 - Subdistrito Candangolândia: Análise de demanda 62
Mapa 8 - Subdistrito Ceilândia: Indicação urbana 63
Mapa 9 - Subdistrito Ceilândia: Análise de demanda 63
Mapa 10 - Subdistrito Cruzeiro: Indicação urbana 64
Mapa 11 - Subdistrito Cruzeiro: Análise de demanda 64
Mapa 12 - Subdistrito Gama: Indicação urbana 65
Mapa 13 - Subdistrito Gama: Análise de demanda 65
Mapa 14 - Subdistrito Guará: Indicação urbana 66
Mapa 15 - Subdistrito Guará: Análise de demanda 66
Mapa 16 - Subdistrito Lago Norte: Indicação urbana 67
Mapa 17 - Subdistrito Lago Norte: Análise de demanda 67
Mapa 18 - Subdistrito Lago Sul: Indicação urbano 68
Mapa 19 - Subdistrito Lago Sul: Análise de demanda 68
Mapa 20 - Subdistrito Núcleo Bandeirante: Indicação urbana 69
Mapa 21 - Subdistrito Núcleo Bandeirante: Análise de demanda 69
Mapa 22 - Subdistrito Paranoá: Indicativo urbano 70
Mapa 23 - Subdistrito Paranoá: Análise de demanda 70
Mapa 24 - Subdistrito Planlatina: Indicativo urbano 71
Mapa 25 - Subdistrito Planaltina: Análise de demanda 71
Mapa 26 - Subdistrito Recanto das Emas: Indicativo urbano 72
Mapa 27 - Subdistrito Recanto das Emas: Análise de demanda 72
Mapa 28 - Subdistrito Riacho Fundo: Indicação urbana 73
Mapa 29 - Subdistrito Riacho Fundo: Indicação urbana 73
Mapa 30 - Subdistrito Samambaia: Indicação urbana 74
Mapa 31 - Subdistrito Samambaia: Análise de demanda 74

Mapa 32 - Subdistrito Santa Maria: Indicação urbana.....75
Mapa 33 - Subdistrito Santa Maria: Análise de demanda 75
Mapa 34 - Subdistrito São Sebastião: Indicação urbana.....76
Mapa 35 - Subdistrito São Sebastião: Análise de demanda.....76
Mapa 36 - Subdistrito Sobradinho: Indicação urbana.....77
Mapa 37 - Subdistrito Sobradinho: Análise de demanda 77
Mapa 38 - Subdistrito Taguatinga: Indicação urbana 78
Mapa 39 - Subdistrito Taguatinga: Análise de demanda.....78

LISTA DE TABELAS

Tabela 1 - Cesta de Produtos relacionados ao chocolate fino 30
Tabela 2 - Descrição das variáveis sociodemográficas.....33
Tabela 3 - Variáveis selecionadas na base T_MORADOR_S.txt.....35
Tabela 4 - Variáveis selecionadas na base T_CADERNETA_DESPESA_S 36
Tabela 5 - Classificação dos setores censitários por demanda 41
Tabela 6 - Identificação da concorrência para uma franquia da Cacau-Show 44

TABELA DE SIGLAS

ABF – Associação Brasileira de Franchising

GP – Gasto Padrão

IBGE – Instituto Brasileiro de Geografia e Estatística

IBOPE – Instituto Brasileiro de Opinião Pública e Estatística

POF – Pesquisa de Orçamentos Familiares

SE – Sistemas Especialistas

SIG – Sistemas de Informação Geográfica

DF – Distrito Federal

SUMÁRIO

1 INTRODUÇÃO 11
1.1 Formulação do problema 12
1.2 Objetivo Geral 13
1.3 Objetivos Específicos 13
1.4 Justificativa 14
2 REFERENCIAL TEÓRICO 15
2.1 Aplicações do SIG 17
2.2 Sistemas de Informação Geográfica (SIG) e Marketing. 21
2.3 Sistemas Especialistas como subsídio tomada de decisão locacional. 22
2.4 A localização Geográfica como recurso Organizacional. 24
2.4.1 Sistemas de Informação Geográfica 24
2.4.2 Sistemas Especialistas (Aprendizado de Máquina) 25
3 MÉTODOS E TÉCNICAS DE PESQUISA 29
3.1 Descrição geral da pesquisa 29
3.2 Elaboração da Cesta de Produtos 30
3.3 Seleção das Variáveis 32
3.4 Tratamento de dados e aplicação do método 35
3.4.1 Tratamento dos dados 35
3.4.2 Treinamento da Máquina 37
3.4.3 Aplicação do método para o Distrito Federal 39
4 RESULTADOS 43
4.1 Oportunidades de negócios 43
5 CONCLUSÃO E RECOMENDAÇÕES 53
5.1 Conclusão 53
5.2 Limitações e recomendações 54
REFERÊNCIAS 56
Apêndice A – Mapas desenvolvidos no QGIS 59
Apêndice B – Programação no RStudio 80

1 INTRODUÇÃO

Segundo a Associação Brasileira de *Franchising* - ABF, o sistema de *Franchising*, em sua definição atual, teve sua origem nos Estados Unidos após a guerra civil, no século XIX, onde a empresa I. M Singer &Co., atuante no mercado de máquinas de costura, estabeleceu uma rede de revendedores. No Brasil, as primeiras franquias surgiram nos anos 60 com a introdução de cursos de inglês, como o CCAA e a Wizard, pautados por essa modalidade de negócio. A partir de então, a expansão do setor foi notória no País.

Nos últimos anos, o sistema de franquias vem assumindo um papel de destaque na economia Brasileira. Segundo dados divulgados pela ABF – Associação Brasileira de Franchising, em 2014 o setor alcançou, no Brasil, 2.942 redes franqueadoras. Esse fato ocasionou em uma participação do setor no PIB (2014) de aproximadamente 0,25%, atingindo o valor de 127,331 bilhões de reais. Isso representa um crescimento de 7,7% em relação ao ano de 2013. O setor também é responsável pela geração de mais de um milhão de empregos diretos.

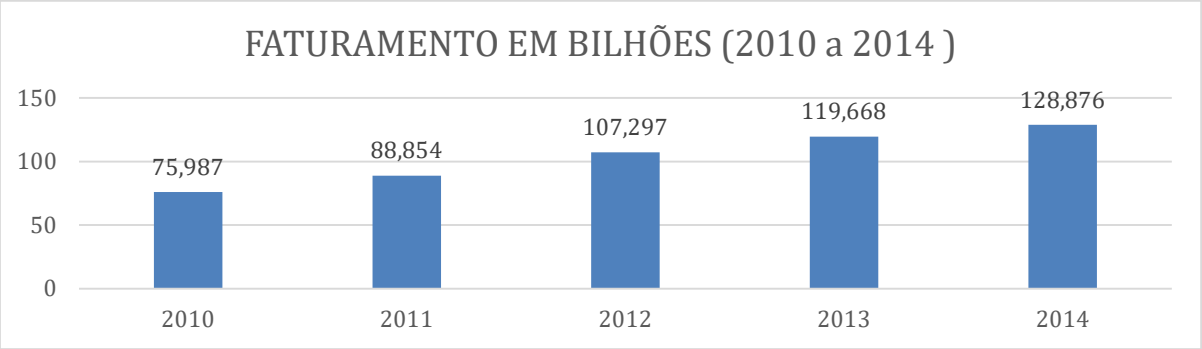


Figura 1 – Faturamento em Bilhões (2010 a 2014)
Fonte: Associação Brasileira de Franchising – ABF

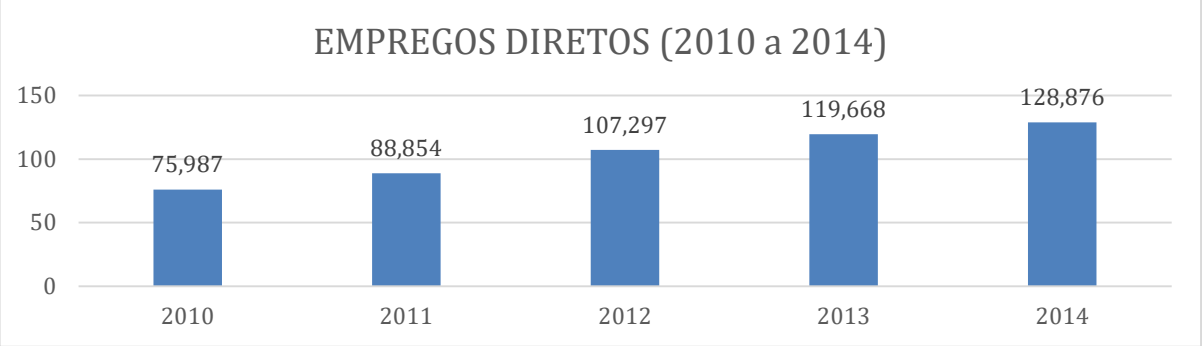


Figura 2 – Geração de Empregos Diretos (2010 a 2014)
Fonte: Associação Brasileira de Franchising – ABF

De acordo com a Lei 8.955/94, “franquia empresarial é o sistema pelo qual um franqueador cede ao franqueado o direito de uso da marca ou patente, associado ao direito de distribuição exclusiva ou semi-exclusiva de produtos ou serviços”.

Dessa forma, o sistema de franquias é, do ponto de vista do Franqueador, uma estratégia segura para expandir o seu negócio. Isso se explica pelo fato de se tratar de um canal direto e exclusivo que terá o seu crescimento subsidiado por recursos financeiros e humanos de terceiros. A descentralização da administração permite a entrada do seu conceito de negócio em regiões desconhecidas pelo franqueador, proporcionando uma maior cobertura de mercado.

Percebe-se, portanto, que uma unidade franqueada, ao se instalar em um novo ponto comercial, possui algumas vantagens em relação à sua concorrência. Isso ocorre visto que já possui uma marca estabelecida e um *know-how* consolidado.

Entretanto, é preciso conhecer profundamente as características do mercado que está sendo alvo da expansão. Fatores como a demanda, hábitos de consumo dos clientes e a concorrência variam de região para região e podem influenciar diretamente os resultados de uma franquia.

Nesse contexto, o desenvolvimento de um método quantitativo, como o *Geomarketing*, para a escolha locacional de uma nova unidade integrante de uma rede de *Franchising* pode representar uma vantagem competitiva para o franqueador, podendo minimizar riscos, reduzir custos e otimizar os resultados organizacionais.

1.1 Formulação do problema

A localização é, sem sombra de dúvidas, um item crucial para o sucesso ou fracasso de um empreendimento. De acordo com pesquisa realizada pelo SEBRAE-SP (2014), 46% dos empreendedores que fecharam suas empresas não conheciam o número de clientes que teriam e quais os seus hábitos de consumo; 38% não sabiam o número de concorrentes presentes na região; e 37% não sabiam qual a melhor localização para a instalação do seu negócio. Estes dados expressam a importância da escolha do ponto para o futuro de uma organização.

Segundo Cliquet (2006), a cobertura territorial, tanto no nível regional quanto no internacional, é pelo menos tão importante quanto o volume de vendas para

determinar a força de uma rede de lojas. Nesse contexto, e considerando a necessidade que uma rede de franquias tem de possuir um processo criterioso e estratégico de decisão na determinação de um ponto comercial, este trabalho possui como problema de pesquisa a seguinte questão: “Poderia o *Geomarketing*, combinado com um Sistema de Aprendizado de Máquina, representar uma vantagem competitiva sustentável para uma rede de Franchising, auxiliando o seu processo de expansão estratégica? ”

1.2 Objetivo Geral

A elaboração de um método quantitativo que auxilie o processo de decisão locacional de uma rede de franquias a partir da utilização de técnicas de *Geomarketing* e Aprendizado de Máquina.

1.3 Objetivos Específicos

I. a elaboração de uma cesta de produtos condizentes com o que é oferecido pela rede de franquias escolhida e que estejam presentes na Pesquisa de Orçamentos Familiares 2008-2009;
II. a identificação das variáveis, constantes na Pesquisa de Orçamentos Familiares 2008-2009, que influenciam, direta ou indiretamente, os resultados de uma unidade da franquia;
III. construção de um Sistema de Aprendizado de Máquina que seja capaz de prever o comportamento dos consumidores, de acordo com as informações coletadas na Pesquisa de Orçamentos Familiares 2008-2009;
IV. aplicação do padrão estabelecido pelo classificador, na base de dados do Censo demográfico do IBGE (2010), para obter previsões sobre o consumo da população do Distrito Federal;
V. mapear, com o auxílio do software QGis e técnicas de Geomarketing, a concorrência e a distribuição da demanda pelo o produto escolhido ao longo do território do Distrito Federal;
VI. apontar, por meio de critérios eliminatórios e classificatórios, os locais que representam, no Distrito Federal, oportunidades de negócio para a rede de franquias escolhida.

1.4 Justificativa

É essencial para qualquer empresa conhecer detalhadamente o seu mercado de atuação. Entretanto, isso só é possível através da combinação de uma base de dados que forneça informações comportamentais dos seus clientes em potencial – tais como seus hábitos de consumo, renda, escolaridade, entre outras – e da combinação desses dados com as informações espaciais desses mesmos clientes. Essa junção permite uma análise muito mais abrangente e se faz possível através das técnicas de *Geomarketing.*

Esse conhecimento sobre os consumidores, acrescido, pela ferramenta de *Geomarketing*, das informações sobre onde e como estes clientes estão situados graficamente no mercado, é fundamental para a implementação de uma abordagem de *marketing*. Segundo Cliquet (2006), existe uma necessidade cada vez maior por uma compreensão mais precisa do mercado, o qual se manifesta através de segmentações crescentes e específicas. Essa segmentação acontece devido à gradual fragmentação da população e pela necessidade de se definir uma estratégia diferenciada para cada segmento.

Trata-se, portanto, de uma ferramenta de extrema utilidade para os gestores, auxiliando na tomada de decisão e fazendo com que a identificação de oportunidades e ameaças ao seu negócio sejam apontadas de forma mais eficiente. Além disso, é possível afirmar que as técnicas de *Geomarketing* contribuem para uma melhor alocação dos recursos disponíveis; para a redução dos riscos envolvidos no processo de abertura de um novo ponto comercial, prevenindo questões como a escassez de demanda, má aceitação do público ou excesso de concorrência; e para a otimização dos resultados organizacionais, visto que é capaz de identificar o ambiente ideal para o desenvolvimento do negócio e, no caso das franquias, apontar regiões com deficiência de cobertura, o que acaba representando um custo para uma rede de lojas.

A aplicação de Sistemas Especialistas (SE) na literatura de *Marketing* é algo de extrema escassez. Apesar da grande variedade de disciplinas que utilizam essa ferramenta como objeto de estudo, como a agricultura, saúde e engenharia, Cui e Curry (2005) afirmaram não haver qualquer utilização deste tipo de sistema dentro da literatura de *Marketing*.

A única aplicação encontrada de Sistemas Especialistas, integrados com um Sistema de Informação Geográfica (SIG), na literatura Brasileira foi o artigo "*A localização geográfica como recurso organizacional: utilização de sistemas especialistas para subsidiar a tomada de decisão locacional do setor bancário",* publicado na Revista de Administração Contemporânea*.* Neste sentido, a principal motivação teórica para a resolução deste trabalho é a falta de exploração deste campo do conhecimento e a contribuição que este estudo poderá proporcionar para o desenvolvimento acadêmico nessa área.

2 REFERENCIAL TEÓRICO

Um sistema de informações geográficas, ou qualquer outro sistema de informação, não representa um fim em si próprio. O valor de qualquer sistema de informações advém da utilidade do seu produto informacional resultante. Este produto deve ser a resposta para uma pergunta específica, que nasce no contexto de surgimento de algum problema que precisa ser resolvido, estando este conectado à alguma função administrativa (DE MAN, 1988).

Essa função administrativa, citada anteriormente, pode incluir os campos de planejamento, decisão, pesquisa e monitoramento. Essas funções se relacionam com os fins a serem atingidos e os meios a serem utilizados, de acordo com o contexto em que ocorre cada situação.

Percebe-se, pela relação entre meios e fins, que as funções administrativas sempre têm um contexto espacial distinto. Meios existem para serem utilizados em algum lugar, o contexto em que ocorrem precisa ter uma dimensão espacial e até os objetivos geralmente possuem uma localização específica.

Informações e bases de dados, em geral, podem ser considerados como tendo três dimensões diferentes (STEINER et al*.*, 1972; SALOMONSSON, 1980): conteúdo; tempo; e espaço (ou localização). Os Sistemas de Informação Geográfica (SIG) representam uma forma especializada e complexa de coleta, armazenamento, recuperação e preservação desses dados, possuindo, portanto, características indispensáveis para a utilização dados georreferenciados.

A informação só é necessária quando existe risco presente na operação (EISGRUBER; FOSTER, 1978). Sendo assim, a busca por informações só tem

sentido quando a questão a ser estudada é, simultaneamente, de relativa importância e envolve incerteza.

A noção de incerteza entre os meios e os fins é importante para entender a estruturação do problema. A especificação de um problema se encontra entre dois extremos, onde o primeiro se dá quando as três funções administrativas (meio, contexto e finalidade) são totalmente conhecidos e estão sobre o controle do analista. Esse cenário é caracterizado como tendo “ausência incertezas”, enquanto que, no outro extremo, as funções são totalmente desconhecidas, configurando um cenário de total incerteza. Obviamente, de forma prática, esses pontos extremos não existem, sendo os problemas definidos entre esses dois polos, tendo níveis diferentes de incerteza.

As decisões de rotina se encaixam nos problemas bem estruturados (com baixo nível de incertezas entre as funções). Já as situações em que existe um alto nível de incertezas, o que configura um problema mal estruturado, estão as decisões inovadoras e de difícil diagnóstico, onde apenas “possíveis soluções” podem ser indicadas (BOSMAN, 1973; DAENZER, 1978; DE MAN, 1985; SIMON, 1960).

Dessa forma, pode-se dizer que o valor da informação e o valor potencial de um sistema de informação dependem do potencial de redução das incertezas entre os meios e suas finalidades, dentro do contexto das funções administrativas.

Segundo Erik De Man (1988), além do valor intrínseco da informação, existem ao menos três motivos para que o valor da informação e dos sistemas de informação seja considerado. Primeiramente, a coleta de informações é um procedimento custoso. Portanto, os possíveis benefícios trazidos por essas informações devem estar explícitos para que possam ser comparados com seus custos. Segundo, existem várias formas de conseguir informação. Cada uma delas exigem técnicas diferenciadas que resultam em diferentes produtos. É necessária, então, uma medida para que essas diferentes técnicas possam ser comparadas entre si. Terceiro, os sistemas de informação possuem impactos sobre as organizações, oferecendo um quadro integrado de dados que podem ser usados por diversas áreas de atividade. Isso contribui para o desenvolvimento organizacional, o que significa dizer que agrega valor para a instituição.

A utilização da informação implica que se tenha a capacidade de lidar e interpretar esses dados. Pensar que quanto mais informação melhor é muito

comum, porém de nada vale uma grande quantidade de dados se não existe a capacidade de organiza-los e saber interpretar os seus resultados (SIMON, 1976).

Um sistema de informações lida com base de dados espaciais e não espaciais. Portanto um Sistema de Informações Geográficas (SIG) deve ter como requisito a capacidade de interagir esses dados e com o fluxo de informação. Um SIG que lida somente com dados espaciais, por exemplo, opera somente com uma parte muito pequena da informação.

2.1 Aplicações do SIG

O avanço das técnicas e ferramentas na área de cartografia tem se mostrado cada vez mais útil e eficaz nas mais diversas áreas do conhecimento. É crescente a quantidade de dados que podem ser inseridos nos mapas, o que fez com que o sistema analógico (manual) perdesse espaço, visto que a tendência é a inserção de inúmeros dados dentro de um mesmo mapa. O aumento da demanda por uma representação cada vez maior de dados dentro de um mapa, paralelo ao avanço dos computadores, fez com que surgisse uma poderosa ferramenta, os Sistemas de Informações Geográficas (SIG).

Segundo Furlan A. A. (2011), “Sistemas de informações geográficas (SIG) são conjuntos de aplicativos computacionais desenvolvidos para tratar informações relacionadas à superfície da Terra”. Eles têm a capacidade de integrar dados de fontes diversas com um banco de dados georreferenciado, a fim de realizar, a partir do resultado dessa integração, análises diversas e complexas. A capacidade de integrar e combinar informações faz com que o Sistema de informações geográficas seja capaz de gerar novas representações que auxiliam no processo de tomada de decisão. As três principais formas de utilização dos SIG são: ferramenta para produção de mapas; suporte para análise espacial de fenômenos; e como banco de dados geográficos, com funções de armazenamento e recuperação de informação espacial. (CÂMARA; QUEIROZ, 2004)

Com o desenvolvimento dos Sistemas de Informações Geográficas e a ascensão das máquinas ao estágio de avanço tecnológico moderno, o Geoprocessamento – “conjunto de conhecimentos destinado ao tratamento das informações referentes aos objetos, ocorrências ou fenômenos que são associados a posições relativas da superfície da Terra” (FURLAN, 2011) – tornou-se um

elemento dinâmico no processo de conhecimento e representação da superfície terrestre.

Segundo Maguire et al. (1993), o geoprocessamento é classificado em três grandes áreas: socioeconômica; ambiental; e gerencial. A primeira, socioeconômica, trata do uso da terra, contendo dados da agroindústria e irrigação, trata da ocupação antrópica, que inclui cadastros urbanos, regionais e sistemas para serviços de utilidade pública, e também inclui atividades econômicas, que abarcam o *geomarketing* e atividades industriais.

O *geomarketing,* segundo Junior (2007, p. 3), pode ser classificado como uma atividade inovadora mesmo sendo algo que é praticado há tempos, visto que era feita de maneira analógica. Basta olhar para as empresas de varejo, por exemplo, que possuem mapas com alfinetes de cores distintas para diferenciar as lojas próprias das concorrentes, determinando a localização de cada ponto.

Levando em conta os conhecimentos que podem ser gerados – como, por exemplo, a identificação de segmentos de mercado, o conhecimento aprofundado de cada grupo de consumidores, suas demandas e especificidades, bem como sua localização no espaço – a importância do *geomarketing* tende a se consolidar cada vez mais, visto que é vital para as empresas, especialmente em momentos de crise, como é o cenário Brasileiro atual, a utilização de técnicas e ferramentas que tornem possível a captação de clientes e o aprimoramento de estratégias de vendas e propaganda direcionada, visando a maximização dos resultados.

Para entender melhor a relação entre o espaço geográfico e o *marketing,* é preciso conhecer a literatura de alguns autores muito influentes nessa área, sejam eles: Von Thünen, Alfred Weber, August Lösch, William Reilly e Walter Christaller. Vejamos então, de forma sucinta, algumas dessas teorias.

Von Thünen foi o primeiro a desenvolver um modelo analítico entre mercado, produção e distância. (RODRIGUE, 2004). Ele afirmou, observando as paisagens agrícolas, que o custo relativo de transporte de *commodities* para os centros comerciais é que determinavam qual a forma de uso da terra para fins agrícolas nos arredores da cidade. Ou seja, aquelas atividades com maior produtividade irão se localizar mais próximas dos centros urbanos, enquanto que aquelas com menor nível de produção se instalarão em regiões mais distantes. Desta forma, seguindo a teoria, seria possível determinar a atividade econômica em qualquer lugar através do

estudo dos processos decisórios e dos fatores geográficos de cada área (JUNIOR, 2007, p. 7).

Deste argumento resultaram-se os *Aneis de Thünen,* que, por sua vez, é alvo de críticas por vários outros autores. Vieira, Albert e Bagolin (2007, p. 35) fazem uma crítica dizendo que “Von Thünen considerou que o único fator a diferenciar o custo de produção é o custo de transporte do produto até a cidade”, indicando que existem outros fatores que influenciam os custos de produção e, portanto, este modelo dificilmente poderia ser reproduzido inteiramente na realidade.

Já segundo o modelo de Alfred Weber (1909), o que determina a localização industrial é a busca por custos mínimos de matéria prima e transporte. Ele afirma que a melhor localização seria aquela em que o custo de produção é mínimo e, ao mesmo tempo, proporcione acesso rápido e fácil à matéria prima, o que resultaria em uma redução nos custos de produção (VIEIRA; ALBERT; BAGOLIN, 2007, p. 37).

Lösch (1940) criticou o modelo de Weber (1919) introduzindo um elemento de extrema importância para a determinação da localização da indústria: a demanda. Dito isso, admite-se que a elasticidade da curva de demanda varia entre as regiões devido a fatores como a preferência dos indivíduos, que são diferentes entre si, e o seu poder aquisitivo (JUNIOR, apud MANGAZOL, 1985).

Entre todos esses autores, o que mais se aproximou da explicação da realidade quanto à explicação dos fatores locacionais das atividades econômicas foi Walter Christaller. Em sua obra “*os lugares centrais no Sul da Alemanha”,* de 1933, ele diz que existem leis que determinam o tamanho e a distribuição das cidades, estabelecendo uma hierarquia entre elas. A partir dessa hierarquização, Chirstaller diz que existe uma distribuição uniforme de pontos centrais e áreas de influência a partir deles (JUNIOR, 2007, p. 9). Segundo Busciolli (apud CAVALCANTE, 2005), a produção de bens e serviços dentro das cidades resulta de uma escala de produção que alcança o seu ponto ótimo e que existe uma tendência à formação de arranjos hexagonais específicos para a distribuição das cidades.

O geoprocessamento, segundo Rodrigues (1988), “é um sistema de coleta e tratamento de informações espaciais e cartográficas interpoladas com dados estatísticos, junto a um *software* que processa esses dados de forma gerencial”. O sistema de Informação Geográfica (SIG) permite a coleta e o cruzamento de informações georreferenciadas que compõem o senário de políticas públicas ou

características de determinada região ou município, atuando como uma ferramenta de auxílio à decisão e ao planejamento urbano e rural (ARONOFF, 1989).

As novas geotecnologias estão presentes na vida dos indivíduos em todos os momentos. A cada passo que se dá novos dados georreferenciados são criados e essas informações são utilizadas pelo governo em ações de saúde, segurança, educação, e pelas empresas, que as utilizam para identificar padrões e características do mercado, direcionar propagandas e alocar recursos específicos para cada segmento.

Segundo a *revista Cidade do Brasil* (2005), estima-se que cerca de 80% das atividades efetuadas numa prefeitura, por exemplo, sejam dependentes do fator localização. A implantação do geoprocessamento e a atualização da sua base cadastral resultam em um aumento da arrecadação, trata-se de um investimento com alta taxa de retorno.

O resultado que se obtém pela utilização do SIG (Sistemas de Informação Geográfica) é uma melhora no processo de tomada de decisão de carácter tático de planejamento, otimizando a alocação de recursos, devido, principalmente, a dois fatores: a exatidão de dados com parâmetros científicos e a “gestão corporativa”, com a integração entre secretarias no compartilhamento de informações, ações e procedimentos.

O geoprocessamento também pode trazer grandes benefícios para outras questões, como, por exemplo, a prevenção de catástrofes, por meio da seleção de imóveis que se encontram em áreas de probabilidade de deslizamento e da classificação das áreas de risco; adoção de medidas preventivas e corretivas; fornecimento de relatórios históricos de chuvas; atendimento de forma rápida e dinâmica às situações de risco; identificação dos pontos de socorro mais próximos, entre outras medidas.

Outra situação em que o uso do SIG se faz muito útil é na área da saúde pública. No artigo “*uso do SIG na determinação da acessibilidade a serviços de saúde em áreas urbanas”,* é demonstrada a utilização desse recurso no monitoramento da oferta dos serviços de saúde, através da identificação e distribuição da demanda no espaço e a localização dos pontos de oferta desses serviços. Dessa forma foi possível a identificação das áreas com maior carência de cobertura mediante o cálculo de índices de acessibilidade.

Este raciocínio de uso dos sistemas de informação geográfica pode se estender para outras áreas, como saneamento, epidemiologia, educação preventiva, alimentação, segurança, turismo, agricultura e outras. Esses levantamentos de dados georreferenciados dão condições para que os agentes públicos possam tomar decisões estratégicas e efetuar um planejamento mais efetivo de políticas públicas.

2.2 Sistemas de Informação Geográfica (SIG) e Marketing.

A mudança na capacidade e na forma de se armazenar dados e a maior acessibilidade a esses recursos têm causado uma revolução na forma como as empresas conduzem suas atividades de *marketing.*

Pela junção dos conceitos de geografia e *marketing,* tendo o primeiro como sendo a distribuição territorial dos fenômenos e o segundo como “o ato de conhecer o mercado de atuação de uma organização, para posteriormente oferecer, de forma inovadora e criativa, produtos e serviços que esse mercado deseja” (JUNIOR, apud ZELA, 2004), define-se *geomarketing* como “a disciplina que estuda as relações existentes entre as estratégias e políticas de *Marketing* e o território ou espaço, onde a instituição, seus clientes, fornecedores e pontos de distribuição se localizam” (JUNIOR, apud DAVIES, 1976).

A base do sucesso de um *marketing* de relacionamento se dá pela identificação, em detalhes, de grupos de clientes específicos que possuem características homogêneas. Para identificar esses segmentos é necessário um número muito grande de informação para, assim, conhecer as particularidades de cada grupo e poder satisfazer as suas necessidades. O processo de segmentação é longo e complexo, pois exige a confirmação de que os segmentos existem, a determinação das suas características e localização para que, a partir dessas informações, se possa elaborar formas de alocar cada cliente no segmento correto (SHEPARD, 1993).

Os Sistemas de Informação Geográfica (SIG) podem ser de grande utilidade para esse tipo de tarefa. O SIG foi desenvolvido, segundo PAREDES (1994), na década de 60, inicialmente para as companhias mineradoras, petrolíferas e para uso do governo. Essa ferramenta foi descoberta recentemente (1996) pelas empresas

industriais e comerciais devido aos avanços da tecnologia e da diminuição dos custos da sua utilização, tornando-a algo mais acessível.

O diferencial dos Sistemas de Informação Geográfica é que ele permite que os dados e informações de um banco de dados possam ser associados a um mapa. Isso torna mais dinâmica a identificação de padrões dentro de uma região, entender tendências, quantificar variáveis, analisar mercados e alocar recursos. Os SIG permitem que o gestor, ou tomador de decisão, tenha compreensões, a respeito dos dados, que não estariam disponíveis de outra forma, visto que representações gráficas dificilmente são capazes de fazer relações espaciais de maneira satisfatória.

Os SIG integram três tipos de arquivos: banco de dados; arquivos geográficos; e arquivos de pontos. O banco de dados são as informações puramente externas à empresa, contendo, por exemplo, dados econômicos, demográficos e sociais do mercado. Os arquivos geográficos contêm as entidades geográficas definidas por suas coordenadas e servem para a produção dos mapas. O terceiro tipo de arquivo é a união dos dois primeiros, onde os dados coletados ficam associados à sua localização geográfica. A junção desses três arquivos torna possível a criação de mapas e a aplicação de cores, padrões e símbolos, representando simultaneamente diversos tipos de dados. O resultado final é a análise de potencial de mercado, segmentação, localização de clientes ou a identificação do melhor ponto comercial para determinada organização (ARANHA, 1996).

2.3 Sistemas Especialistas como subsídio tomada de decisão locacional.

A vantagem competitiva, e todos os fatores que a determinam, é o tema central dos estudos sobre estratégia (PORTER, 1991) e daqueles estudos que associam a visão baseada em recursos (WERNERFELT, 1984; BARNEY, 1991). Os recursos são de extrema importância para uma organização, pois, quando estes são inovadores, raros e difíceis de imitar, aumentam a competitividade da empresa. Segundo Barney (1996) e Barney &Hesterly (1996), os recursos podem ser divididos da seguinte forma: financeiros; físicos (localização geográfica, instalações); organizacionais (planejamento, controle) e humanos (experiência, treinamento, inteligência). Para Grante (1996), são adicionadas à divisão acima duas categorias:

reputação (visão dos clientes sobre a empresa); e recursos tecnológicos (*software, hardware,* equipamentos e máquinas).

A decisão sobre a localização geográfica de uma empresa é determinante para o seu sucesso ou fracasso. Sendo assim, decidir a localização de um empreendimento exige decisões estratégicas embasadas em estudos criteriosos, e não apenas em questões de bom senso e sugestões de especialistas. Para tal, existem ferramentas e técnicas computacionais disponíveis para auxiliar este processo decisório, como os Sistemas de informações Geográficas (SIG) e os Sistemas Especialistas (SE).

A visão baseada em recursos (WERNERFELT, 1984) busca explicar o melhor desempenho das organizações pautando-se em uma visão baseada em recurso, que podem ser definidos como sendo “os ativos, tangíveis e intangíveis, que são vinculados de forma semipermanente à empresa, que não podem ser transferidos a outra sem custos e, que a capacita a conceber e executar estratégias de modo a obter e/ou manter certa posição competitiva” (BARNEY, 1992,1994,1996; HESTERLY, 1996; WERNERFELT, 1984).

Segundo Barney (1991), recursos são “todos os ativos, capacidades, processos organizacionais, atributos da empresa, informações, conhecimentos, controlados pela empresa e que lhe permitem compreender e colocar em ação as estratégias suscetíveis de aumentar sua eficácia e sua eficiência na concepção e implementação de estratégias”. Essa definição não garante que todos os recursos representem vantagens competitivas sustentáveis (BARNEY, 1991; PETERAF, 1993), visto que, segundo Barney (1991), a vantagem competitiva deve agregar valor para uma organização durante a implementação de uma estratégia que não está sendo simultaneamente executada por nenhuma outra empresa. Ou seja, quando os concorrentes não são capazes de duplicar o benefício gerado pela estratégia implementada. Dessa forma tem-se a vantagem competitiva sustentável.

Nesse sentido, Barney (1991,1996) atribui uma lista de características para os recursos que, quando eficientemente executadas, resultam em uma vantagem competitiva sustentável:

I. valiosos: os recursos devem explorar ou neutralizar as ameaças do ambiente, gerando uma redução de custos ou aumento das receitas da empresa;

II. raros: os recursos devem ser raros entre os concorrentes atuais e potenciais da empresa;
III. difíceis de serem imitados: ocorre quando as demais organizações, que não possuem o recurso, enfrentariam desvantagens de custo para obtê-lo;
IV. difíceis de serem substituídos: quando não existe uma estratégia equivalente e que seja, também, rara e difícil de ser imitada entre os concorrentes;

2.4 A localização Geográfica como recurso Organizacional.

O desenvolvimento de sistemas computacionais facilitou a análise espacial, devido à integração de dados de diversas fontes e à criação de bancos de dados georreferenciados. Estes sistemas são denominados Sistemas de Informações Geográficas (SIG).

2.4.1 Sistemas de Informação Geográfica

Segundo Câmara e Queiroz (2006), SIG são sistemas que fazem o tratamento de dados geográficos e conseguem informações com base em características alfanuméricas e de localização espacial.

A partir das várias definições de SIG apresentadas neste artigo e do conhecimento construído acerca do assunto, destacam-se para a realização deste trabalho as definições de Burrough e McDonnell (1998, p.11), onde é dito que SIG é um “conjunto de ferramentas para coletar, armazenar, recuperar, transformar e representar visualmente dados espaciais”; e de Cowen (1988, p. 1554), que diz que SIG “é um sistema de apoio à decisão que envolve a integração de dados espacialmente referenciados, em um ambiente para resolução de problemas”.

Com base nessas definições, fica claro que o SIG é uma ferramenta bastante útil de auxílio à tomada de decisão e análises espaciais. A possibilidade de se combinar dados referentes à localização das agências bancárias e postos de atendimento, tanto da própria rede quanto dos concorrentes, com dados socioeconômicos, como renda, escolaridade e densidade demográfica, representa um avanço na análise de dados, que antes, de maneira isolada, não geravam grandes benefícios para as organizações.

2.4.2 Sistemas Especialistas (Aprendizado de Máquina)

Feigenbaum, um dos principais pesquisadores de Sistemas Especialistas (SE), segundo Waterman (1983) e Harmon e Kink (1988), define um Sistema Especialista como um “programa inteligente de computador que usa conhecimento e procedimentos inferenciais para resolver problemas que requerem perícia humana para a sua solução”. Segundo Waterman (1986), SE são programas de computadores que manipulam conhecimento para resolver problemas eficientemente em uma área específica.

Um sistema especialista é, então, composto por uma base extensa de conhecimentos e regras sobre determinado assunto e por um processador de inferência, que utiliza a base para tomar conclusões e produzir julgamentos sobre aquele assunto. A máquina interpreta e decide como as regras devem ser utilizadas e em que ordem, deduzindo, assim, novos conhecimentos. (GENARO, 1986).

Estes sistemas estão sendo utilizados para auxiliar os especialistas nas etapas do planejamento, diagnósticos de doenças, localização de depósitos minerais e em várias outras áreas. De acordo com Eldrandaly, Eldin e Sui (2003), vários sistemas especialistas buscam resolver problemas de localização geográfica, auxiliando o especialista no processo decisório em relação à escolha de um local.

Neste estudo será utilizado o Aprendizado de Máquina como mecanismo de Sistema Especialista. Segundo Monard e Baranauskas (2003), um Sistema de Aprendizado de Máquina é um programa de computador que consegue tomar decisões tomando como base experiências acumuladas.

Este programa utiliza a **inferência indutiva** para derivar conhecimento novo e predizer eventos futuros. A indução é uma forma de inferência lógica que permite a **generalização** de um modelo, validado para uma amostra específica. Por essa razão, é preciso ser cauteloso na escolha da quantidade e da qualidade dos exemplos que serão apresentados, visto que isso pode fazer com que as hipóteses geradas sejam de pouco valor e não preservem a verdade (MONARD; BARANAUSKAS, 2003).

O aprendizado indutivo se divide entre **supervisionado** e **não-supervisionado**. No primeiro caso são fornecidos ao indutor (algoritmo de aprendizagem) dados de treinamento que contenham, individualmente, uma série de características, bem como a classe associada a elas. Ou seja, no aprendizado

supervisionado é fornecido para o algoritmo um conjunto de exemplos, onde cada um deles está associado à um grupo de características que definem uma determinada classe, pertencente a um conjunto discreto (nominal) de classes {C1, C2, ..., Ck}. Dessa forma, o algoritmo de indução será capaz de determinar corretamente a classe de um novo exemplo que apresente apenas o seu grupo de características como informação (MONARD; BARANAUSKAS, 2003). A **Figura 3** descreve o formato padrão de um conjunto de exemplos T, com t exemplos e m atributos. A coluna Y é o que o indutor tentará predizer partindo dos atributos associados a cada exemplo.

	X_1	X_2	...	X_m	Y
T_1	x_{11}	x_{12}	...	x_{1m}	y_1
T_2	x_{21}	x_{22}	...	x_{2m}	y_2
⋮	⋮	⋮	⋯	⋮	⋮
T_n	x_{n1}	x_{n2}	...	x_{nm}	y_n

Figura 3: Conjunto de exemplos no formato atributo-valor
Fonte: Adaptada de Monard; Baranauskas, 2003, p. 44

Já no aprendizado não-supervisionado o indutor analisa os exemplos fornecidos e procura determinar se existe alguma forma de agrupá-los, formando os chamados *clusters.* Passada essa etapa, geralmente é necessária uma análise para identificar o que cada agrupamento significa no contexto do problema que está sendo estudado (CHEESEMAN; STUTZ, 1990).

Segundo Michalski (1983) e Kubat, Bratko, Michalski (1988), os sistemas de aprendizado podem ser classificados em duas grandes categorias:

I. sistemas tipo caixa-preta: são os sistemas que não apresentam resultados internos claros sobre o conceito criado. Ou seja, sua representação interna e o processo de reconhecimento não podem ser facilmente interpretados por humanos;
II. sistemas orientados ao conhecimento: Objetivam a criação outputs no formato de estruturas simbólicas que sejam compreensíveis por humanos;

A **Figura 4** abaixo é capaz de resumir o processo de classificação por meio da ferramenta de Aprendizado de Máquina. De maneira geral, ocorre a especificação do problema e a seleção do conjunto de exemplos que servirão de entrada para indutor. Após induzido, é gerado um classificador capaz de tomar decisões futuras com base nas informações fornecidas a ele na primeira etapa. Posteriormente é feita uma validação do classificador, onde, considerando sua precisão, são feitas mudanças na especificação do problema e na seleção dos dados, com o intuito de melhorar o sistema como um todo.

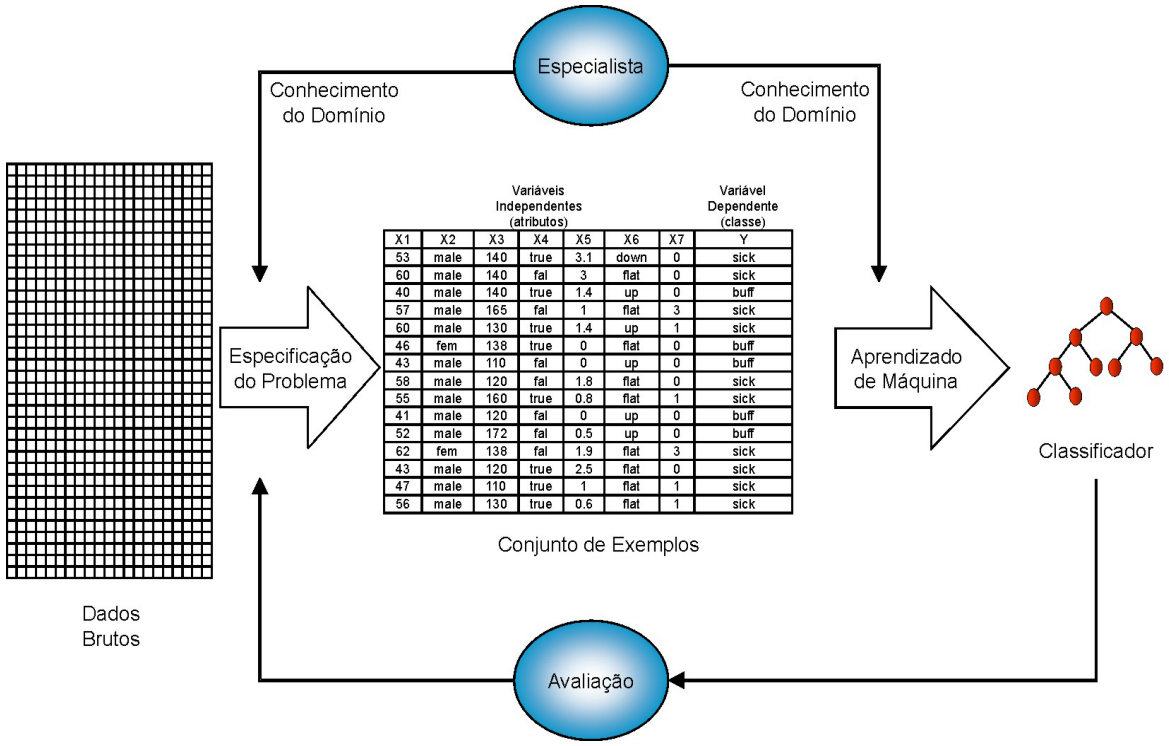


Figura 4: Processo de classificação por meio do Aprendizado de Máquina
Fonte: Monard; Baranauskas, 2003, p. 42.

Outra forma interessante de ilustrar o processo de Aprendizado de Máquina é imaginando um gráfico contendo todos os exemplos oferecidos ao indutor. Neste cenário, o objetivo do classificador é identificar a qual classe pertence cada exemplo e separá-los de uma forma linear, como mostra a **Figura 5**:

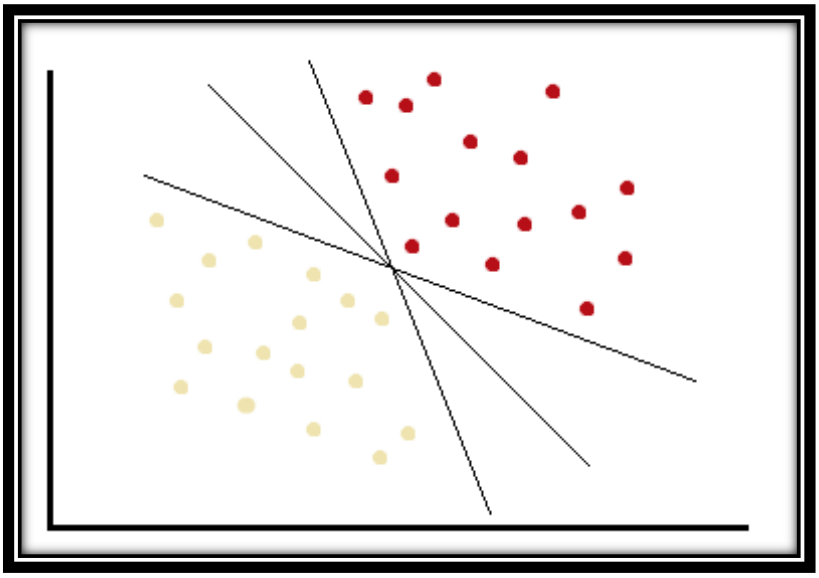


Figura 5 - Separação dos exemplos pelo classificador
Fonte: Elaborado pelo autor.

Como se pode perceber, existem infinitas formas de separar os classificadores. Portanto, o classificador irá buscar aquela que forneça a maior margem entre as classes, tendo como objetivo aumentar a distância (n) entre os limites de cada classe. Quanto maior o valor de n menor será a probabilidade de erro de classificação, como ilustra a **Figura 6:**

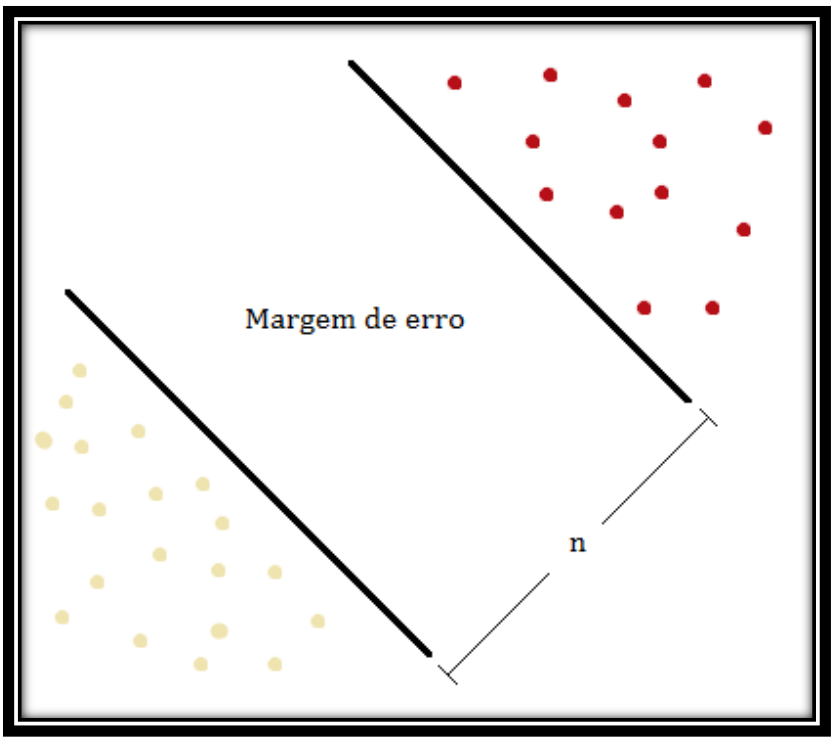


Figura 6 – Maximização entre as fronteiras de cada classe.
Fonte: Elaborado pelo autor.

3 MÉTODOS E TÉCNICAS DE PESQUISA

3.1 Descrição geral da pesquisa

As bases de dados utilizadas neste trabalho foram a Pesquisa de Orçamentos Familiares (POF) 2008-2009 e o Censo Demográfico de 2010, disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Segundo o IBGE, a POF 2008-2009 teve por objetivo a composição dos orçamentos domésticos por meio da coleta de dados sobre os hábitos de consumo, alocação de gastos e da distribuição dos rendimentos, considerando, também, as características dos domicílios e das pessoas entrevistadas. Também de acordo com o IBGE, o censo Demográfico é uma pesquisa, realizada a cada dez anos, onde pesquisadores do Instituto Brasileiro de Geografia e Estatística visitam todos os domicílios do país aplicando questionários que visam medir a densidade populacional e conhecer o perfil da população Brasileira.

Considerando que esta pesquisa tem como objetivo a elaboração de um método quantitativo que auxilie o processo de decisão locacional para uma rede de franquias em expansão, ou seja, que queira abrir uma nova filial, é importante destacar que a escolha da franquia a ser estudada possui uma importância secundária dentro da construção dos resultados, dado que o intuito é o de construir um método que possa ser reproduzido para qualquer instituição que se encontre em um processo semelhante ao que está sendo apresentado.

Dito isso, o ramo de atividade escolhido para a aplicação do método foi o da comercialização de chocolates, que possui como uma de suas principais franquias a rede Cacau Show. Esta escolha se deu pelo fato de a franquia ter conquistado o posto de maior franquia de chocolates finos do mundo, tendo como objetivo, segundo a Associação Brasileira de Franchising – ABF, ampliar os seus negócios, mesmo em meio a um período de recessão, como o atravessado atualmente. Fica claro, portanto, que este estudo é pertinente e está alinhado com a missão e os objetivos da empresa.

A **população** foi então definida como sendo todas as lojas, localizadas no Distrito Federal, que têm como seu produto principal o chocolate fino, bem como todos os indivíduos que residem no Distrito Federal e consomem chocolate. Quanto à **amostra**, foram selecionadas 44 lojas de chocolates que se encontram no do

Distrito Federal e todos os indivíduos que foram identificados como consumidores de chocolate na Pesquisa de Orçamentos Familiares 2008-2009.

É necessário também que se faça uma segmentação do mercado de forma precisa. Como visto anteriormente, o Aprendizado de Máquina, segundo Monard e Baranauskas (2003), utiliza da inferência indutiva para generalizar modelos e predizer eventos futuros partindo de um conjunto de exemplos fornecidos ao indutor. Desta forma, os dados obtidos com base na POF 2008-2009 e no censo de 2010 do IBGE servirão como entrada para que o Sistema Especialista (SE) possa reconhecer padrões entre os consumidores de chocolate e então consiga prever a demanda do produto dentro da região estudada.

Definida a **demanda** pelo produto desejado e a **amostra** de franquias e empresas dentro do Distrito Federal que comercializam produtos semelhantes, ou que possam ser considerados concorrentes, será possível a identificação de oportunidades de negócio para a instalação de uma nova franquia.

3.2 Elaboração da Cesta de Produtos

Para a elaboração da cesta de produtos, foram selecionados todos os produtos relacionados ao consumo de chocolates finos encontrados na POF 2008-2009. Esta seleção não considerou itens tais como granulados ou chocolates em pó, visto que o consumo destes tipos de produto não pode estar diretamente relacionado com o consumo dos produtos da Cacau Show.

Feita a seleção, a cesta final de produtos relacionados ao consumo de chocolates finos obteve um total de 62 itens, como mostra o **Tabela 1**:

CADASTRO DE PRODUTOS POF 2008-2009	QUADRO	GRUPO	CÓDIGO	PRODUTO
	63 a 69	69	00701	TABLETE DE CHOCOLATE
			00702	BARRA DE CHOCOLATE
			00703	CHOCOLATE EM TABLETE
			00704	CHOCOLATE EM BARRA
			00705	CHOCOLATE BISS (TABLETE)
			00707	CHOCOLATE BATOM
			00708	BATON CHOCOLATE
			00709	TUBETE DE CHOCOLATE
			00901	BOMBOM DE QUALQUER MARCA
			00902	BOMBOM CARAMELIZADO DE QUALQUER MARCA
			00903	BOMBONS SORTIDOS DE QUALQUER MARCA
			00904	TRUFA
			00905	BOBOM CASEIRO

			05601	CHOCOLATE EM CREME
			05605	CREME DE CHOCOLATE
			05801	OVO DE PASCOA
			05802	COELHINHO DA PASCOA DE CHOCOLATE
			05803	CHOCOLATE COELHINHO DA PASCOA
			05804	KINDER OVO
			10001	TABLETE DE CHOCOLATE LIGHT
			10002	BARRA DE CHOCOLATE LIGHT
			10003	CHOCOLATE EM TABLETE LIGHT
			10004	CHOCOLATE EM BARRA LIGHT
			10005	CHOCOLATE BISS (TABLETE) LIGHT
			10101	TABLETE DE CHOCOLATE DIET
			10102	BARRA DE CHOCOLATE DIET
			10103	CHOCOLATE EM TABLETE DIET
			10104	CHOCOLATE EM BARRA DIET
			10105	CHOCOLATE BISS (TABLETE) DIET
			10107	TABLETE DE CHOCOLATE DIETETICO
			10108	BARRA DE CHOCOLATE DIETETICO
			10109	CHOCOLATE EM TABLETE DIETETICO
			10110	CHOCOLATE EM BARRA DIETETICO
			10111	CHOCOLATE BISS (TABLETE) DIETETICO
			10401	BOMBOM DE QUALQUER MARCA LIGHT
			10402	BOMBOM CARAMELIZADO DE QUALQUER MARCA LIGHT
			10403	BOMBONS SORTIDOS DE QUALQUER MARCA LIGHT
			10404	TRUFA LIGHT
			10405	BOMBOM LIGHT
			10501	BOMBOM DE QUALQUER MARCA DIET
			10502	BOMBOM CARAMELIZADO DE QUALQUER MARCA DIET
			10503	BOMBONS SORTIDOS DE QUALQUER MARCA DIET
			10504	TRUFA DIET
			10505	BOMBOM DE QUALQUER MARCA DIETETICO
			10506	BOMBOM CARAMELIZADO DE QUALQUER MARCA DIETETICO
			10507	BOMBONS SORTIDOS DE QUALQUER MARCA DIETETICO
			10508	TRUFA DIETETICA
			12101	CHOCOLATE EM CREME LIGHT
			12201	CHOCOLATE EM CREME DIET
			12204	CHOCOLATE EM CREME DIETETICO
			12301	OVO DE PASCOA LIGHT
			12302	COELHINHO DA PASCOA DE CHOCOLATE LIGHT
			12303	CHOCOLATE COELHINHO DA PASCOA LIGHT
			12304	KINDER OVO LIGHT
			12401	OVO DE PASCOA DIET
			12402	COELHINHO DA PASCOA DE CHOCOLATE DIET
			12403	CHOCOLATE COELHINHO DA PASCOA DIET
			12404	KINDER OVO DIET
			12405	OVO DE PASCOA DIETETICO
			12406	COELHINHO DA PASCOA DE CHOCOLATE

				DIETETICO
			12407	CHOCOLATE COELHINHO DA PASCOA DIETETICO
			12408	KINDER OVO DIETETICO

Tabela 1: Cesta de Produtos relacionados ao chocolate fino.
Fonte: Elaborada pelo autor a partir da POF 2008-2009.

3.3 Seleção das Variáveis

Após a elaboração da Cesta de Produtos, o próximo passo foi selecionar as variáveis que podem influenciar o consumo de chocolate. A empresa Ipsos realizou em 2015 uma pesquisa cujo objetivo foi identificar as principais motivações de compra dos consumidores em relação ao chocolate. Seguindo o que foi dito em sua descrição institucional, encontrada no site da empresa, a Ipsos é a terceira maior instituição de pesquisa e inteligência de mercado do mundo, foi fundada em 1975 e, atualmente, possui filiais em 97 países. Ela possui, como uma de suas especialidades, criar uma maior conexão entre as marcas e as motivações humanas dos seus consumidores.

Em sua pesquisa, realizada com o foco para o chocolate, constatou-se que a idade é um fator bastante relevante quando se tenta entender o consumo de chocolate no Brasil. Seus dados mostram que 89% dos entrevistados, entre 13 e 19 anos, afirmam consumir chocolate, enquanto que apenas 42% dos entrevistados com mais de 60 anos tiveram uma resposta positiva quanto ao consumo deste produto.

Além disso, foi verificado que o consumo de chocolate também varia de acordo com o gênero do indivíduo. O levantamento mostra que 71% das mulheres responderam afirmativamente quanto ao consumo de chocolate, enquanto que apenas 64% dos homens responderam da mesma forma. As mulheres também consomem o produto com mais frequência, visto que 35% das entrevistadas disseram comer chocolate pelo menos uma vez por semana e, para os homens, esse número cai para 30%.

Outra variável que pode influenciar o consumo de chocolate, segundo pesquisa realizada pelo IBOPE (Instituto Brasileiro de Opinião Pública e Estatística), é a renda. Essa pesquisa realizou 18.884 entrevistas no período compreendido entre agosto de 2009 e julho de 2010, nas regiões metropolitanas de São Paulo, Rio de

Janeiro, Porto Alegre, Curitiba, Belo Horizonte, Salvador, Recife, Fortaleza e Brasília. Os resultados da **Figura 5** mostram que, dentre os entrevistados pertencentes às Classes AB, 69% afirmaram ter consumido chocolate nos últimos 7 dias. Este número diminui à medida que a Classe dos entrevistados cai para C e DE, atingindo os valores de 66% e 57% respectivamente. Embora essa redução quanto a classe social não seja muito expressiva, os dados apresentados abaixo confirmam os resultados apresentados pela pesquisa da Ipsos em 2015 e ratificam a hipótese de que o consumo de chocolate se relaciona positivamente com a renda do indivíduo.

Chocolate: penetração por classe, sexo e idade.	elementos	total	2010
			Consumidores de chocolate*
total	% horz	100%	66%
Brasil - AB	% horz	100%	69%
Brasil - C	% horz	100%	66%
Brasil - DE	% horz	100%	57%
Masculino	% horz	100%	61%
Feminino	% horz	100%	71%
12 - 19 anos	% horz	100%	77%
20 - 24 anos	% horz	100%	73%
25 - 34 anos	% horz	100%	70%
35 - 44 anos	% horz	100%	61%
45 - 54 anos	% horz	100%	60%
55 - 64 anos	% horz	100%	53%

*consumiram nos últimos 7 dias.

TARGET GROUP INDEX

Figura 5: Mapeamento do consumo de chocolate no Brasil
Fonte: Target Group Index – IBOPE 2009 e 2010.

Sendo assim, as variáveis que serão consideradas como influenciadoras do consumo de chocolate, segundo as pesquisas do Instituto Ipsos (2015) e pelo IBOPE (2010), ficam assim definidas:

VARIÁVEL		DESCRIÇÃO				
Idade	24 variáveis	0 ano - idade	1 ano de idade	2 anos de idade	3 anos anos de idade	4 anos anos de idade
		5 anos de idade	6 anos de idade	7 anos de idade	8 anos de idade	9 anos de idade
		10 anos de idade	11 anos de idade	12 anos de idade	13 anos de idade	14 anos de idade
		15 anos de idade	16 anos de idade	17 anos de idade	18 anos de idade	19 anos de idade
		20 anos de idade	21 anos de idade	22 anos de idade	23 anos de idade	24 anos de idade
	16 faixas de idade	25 a 29 anos de idade				
		30 a 34 anos de idade				
		35 a 39 anos de idade				
		40 a 44 anos de idade				
		45 a 49 anos de idade				
		50 a 54 anos de idade				
		55 a 59 anos de idade				
		60 a 64 anos de idade				
		65 a 69 anos de idade				
		70 a 74 anos de idade				
		75 a 79 anos de idade				
		80 a 84 anos de idade				
		85 a 89 anos de idade				
		90 a 94 anos de idade				
		95 a 99 anos de idade				
		100 anos de idade ou mais				
Gênero		Homem				
		Mulher				
Renda*		Renda 1: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de 1/8 salário mínimo				
		Renda 2: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de 1/8 a 1/4 salário mínimo				
		Renda 3: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de 1/4 a 1/2 salário mínimo				
		Renda 4: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de 1/2 a 1 salário mínimo				
		Renda 5: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de 1 a 2 salários mínimos				

	Renda 6: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de 2 a 3 salários mínimos
	Renda 7: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de 3 a 5 salários mínimos
	Renda 8: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de 5 a 10 salários mínimos
	Renda 9: Domicílios particulares com rendimento nominal mensal domiciliar *per capta* de mais de 10 salários mínimos
* Salário mínimo utilizado: R$ 510,00	
** Inclusive as pessoas que recebiam somente em benefícios	

Tabela 2: Descrição das variáveis sociodemográficas que podem influenciar o consumo de chocolate
Fonte: Elaborado pelo autor com base no Censo Demográfico 2010, realizado pelo IBGE.

3.4 Tratamento de dados e aplicação do método

3.4.1 Tratamento dos dados

Feita a seleção das variáveis que podem impactar o consumo de chocolate, o próximo passo foi iniciar o processo de Aprendizado de Máquina, com o objetivo de conseguir, com base nos dados obtidos com a POF 2008-2009, prever o consumo dos indivíduos encontrados no Censo Demográfico de 2010, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

Segundo o IBGE, o censo 2010 compreendeu um levantamento minucioso de todos os domicílios do país. Foram visitados 67,6 milhões de domicílios nos 5.565 municípios brasileiros para responder questões tais como: quem são esses indivíduos, quantos são, onde vivem e como vivem.

Tendo isso em mente, inicia-se o processo de treinamento da máquina, que tem como primeira etapa a leitura dos dados da POF 200-2009. O trabalho será feito inicialmente com as seguintes bases de dados: **T_MORADOR_S.txt** e **T_CADERNETA_DESPESA_S.txt**, ambas disponibilizadas no site do IBGE.

Essa leitura será feita utilizando o s*oftware* RStudio e o Excel. No RStudio foi efetuada, como primeira etapa, a leitura da base de dados T_MORADOR_S.txt, programando o *sofware* para criar um novo arquivo contendo apenas as informações necessárias para a análise que está sendo realizada. Ou seja, foram selecionadas apenas as variáveis que podem influenciar o consumo de chocolate, definidas na seção anterior, e as informações de identificação do indivíduo e do seu domicílio. Dessa forma, o novo arquivo, gerado a partir da base T_MORADOR_S.txt, contém apenas as seguintes variáveis:

VARIÁVEL	FORMATO	TAMANHO	DECIMAIS	POSIÇÃO INICIAL
CÓDIGO DA UF	Numérico	2		3
NÚMERO SEQUENCIAL	Numérico	3		5
DV DO SEQUENCIAL	Numérico	1		8
NÚMERO DO DOMICÍLIO	Numérico	2		9
NÚMERO DA UC	Numérico	1		11
FATOR DE EXPANSÃO 1 (DESENHO AMOSTRAL)	Numérico	14	8	16
FATOR DE EXPANSÃO 2 (AJUSTADO P/ ESTIMATIVAS)	Numérico	14		30
IDADE CALCULADA EM ANOS	Numérico	3		60
SEXO	Numérico	2		76
RENDA PER CAPITA DA UC	Numérico	16	2	197

Tabela 3: Variáveis selecionadas a partir da base de dados T_MORADOR_S.txt.
Fonte: Elaborado pelo autor com base na POF 2008-2009.

Em seguida, o mesmo procedimento foi feito para a base de dados T_CADERNETA_DESPESA_S.txt. Seguindo exatamente o mesmo raciocínio, as variáveis selecionadas para a esta segunda base foram:

VARIÁVEL	FORMATO	TAMANHO	DECIMAIS	POSIÇÃO INICIAL
CÓDIGO DA UF	Numérico	2		3
NÚMERO SEQUENCIAL	Numérico	3		5
DV DO SEQUENCIAL	Numérico	1		8
NÚMERO DO DOMICÍLIO	Numérico	2		9
NÚMERO DA UC	Numérico	1		11
ESTRATO GEOGRÁFICO	Numérico	2		14
NÚMERO DO QUADRO	Numérico	2		44
CÓDIGO DO ITEM	Numérico	5		46
VALOR DA DESPESA / AQUISIÇÃO	Numérico	11	2	53

Tabela 4: Variáveis selecionadas a partir da base de dados T_CADERNETA_DESPESA_S.txt.
Fonte: Elaborado pelo autor com base na POF 2008-2009.

Selecionadas as variáveis da base T_CADERNETA_DESPESA_S.txt, deve-se criar um novo arquivo, contendo apenas aqueles indivíduos que consumiram pelo

menos um dos itens da Cesta de Produtos. Filtradas as bases, o próximo passo é unir as suas informações, utilizando como referência as seguintes variáveis em comum: CÓDIGO DA UF; NÚMERO SEQUENCIAL; DV DO SEQUENCIAL; NÚMERO DO DOMICÍLIO e NÚMERO DA UC. Isto significa que o mesmo indivíduo que respondeu às questões que compõem a base de dados T_MORADOR_S, respondeu também às questões da base T_CADERNETA_DESPESA_S. Logo, é preciso que essas informações se unam e formem um único banco de dados mais completo e que forneça o necessário para a análise da escolha locacional de uma nova franquia. É importante ressaltar que o que se deseja obter é o valor total do gasto com chocolate por pessoa, portanto é preciso que o R some todos os gastos que tenham um mesmo indivíduo como responsável pela compra.

Quanto às variáveis de renda e idade, foram criadas variáveis *dummy* representando as faixas de idade e renda definidas na **Tabela 2.** Dessa forma, a variável “Faixa etária” foi dividida em 80 variáveis, onde a primeira delas, por exemplo, foi denominada de V073 e receberá o valor 1 caso o indivíduo tenha 0 anos de idade e seja homem, caso contrário receberá 0. Reforçando a ideia, a última variável criada para “Faixa etária” recebeu do nome de V154 e terá valor igual a 1 caso o indivíduo tenha 100 anos de idade ou mais e seja mulher, caso contrário receberá 0. Tratando-se da variável “Renda”, houve uma divisão de 09 variáveis, onde aquela que englobar a renda monetária mensal da UC receberá 1 e as demais receberão 0.

Este procedimento é necessário para que o conjunto de dados, oriundos das bases T_MORADOR_S e T_CADERNETA_DESPESA_S, possam se unir com a base de dados do Censo 2010, visto que as duas devem estar organizadas da mesma forma.

3.4.2 Treinamento da Máquina

O objetivo aqui é fazer com que a máquina utilize os dados da POF 2008-2009 para estabelecer um padrão de consumo para o mercado de chocolate. Este processo se divide em duas etapas: a primeira é o treinamento da máquina e a segunda é a validação do classificador.

Essa divisão visa evitar o *underfitting*, que ocorre quando poucos exemplos representativos são oferecidos ao sistema de aprendizado, fazendo com que a

hipótese se ajuste muito pouco ao conjunto de treinamento, e o *overfitting*, que se caracteriza pelo excesso de ajuste da hipótese em relação ao conjunto de treinamento. Sendo assim, a partir da manipulação da amostra utilizada na etapa de treinamento, é possível induzir hipóteses que se ajustem mais ao conjunto de treinamento, o que compromete o seu desempenho em novos exemplos. Seguindo o mesmo raciocínio, uma hipótese na situação de *underfitting* possui um mal desempenho em um conjunto de teste e apresenta uma melhora de desempenho muito pequena no conjunto de treinamento. Para evitar estes dois extremos, a base de dados foi divida em duas partes, sendo 70% destinados ao treinamento da máquina e 30% à validação do classificador. (Monard; Baranauskas, 2003, p. 46)

Explicando brevemente o método, temos um exemplo $(x_{i1}, x_{i2}, \dots, x_{im}, y_i) = (\vec{x_i}, y_i)$ que possui m atributos x_{ij} e cada atributo corresponde à uma coordenada no espaço de descrição, onde $1 \leq i \leq n$ e n representa o número de exemplos. Além disso, cada coordenada definida por um atributo x_{ij} está inserida em uma região do espaço de descrição, que foi associada, pelo classificador, à uma classe y_i. (Monard; Baranauskas, 2003, p. 45)

Outro ponto importante em um treinamento de máquina é a definição da taxa de erro de um classificador h, bem como a sua precisão. A taxa de erro nada mais é do que uma comparação entre a classe verdadeira do exemplo e rótulo imputado pelo classificador, geralmente representada pela fórmula $Err(h) = \frac{1}{n}\sum_{i=1}^{n}\|y_i \neq h(x_i)\|$, em que o operador $\|E\|$ retorna o valor 1 quando a condição E é verdadeira e 0 no caso contrário. Esta relação informa a porcentagem de erro do classificador. Logo, para se definir o nível de precisão basta calcular $Acc(h) = 1 - Err(h)$. (Monard; Baranauskas, 2003, p. 45)

Neste trabalho, por se tratar de um problema de regressão, a comparação entre a classe real do exemplo e a atribuição dada pela máquina foi feita por meio do Erro Médio Quadrado, dado por (3.1):

$$mse_err(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h = x_i))^2 \qquad (3.1)$$

É importante destacar que o que se deseja, segundo Weiss & Kulikowski (1991), é a construção de classificadores com uma baixa taxa de erro (3.1) em relação ao conjunto teste. Portanto, o cálculo da precisão (ou erro) da hipótese teve como foco os 30% da amostra destinados à etapa de validação.

Feita a divisão da base de dados, o próximo passo foi a criação da lista parâmetros, C e σ, que será utilizada no aprendizado da máquina. Ambos os parâmetros são estabelecidos pelo usuário do método. O parâmetro de regularização C representa, em maior ou menor grau, o nível de importância dos erros de classificação gerados pelo classificador, enquanto que o parâmetro σ se relaciona com a precisão do classificador.

Para a definição de cada par de parâmetros, utilizou-se aquele que possuísse o menor Erro Médio Quadrado ($\mathrm{mse_err(h)}$). Isto causa uma redução da diferença entre o valor real e aquele atribuído pelo classificador, fazendo com que a máquina se torne mais precisa e próxima da realidade.

3.4.3 Aplicação do método para o Distrito Federal

Finalizado o aprendizado da máquina, inicia-se a previsão do comportamento de consumo de chocolate no Distrito Federal. Como visto na sessão de tratamento dos dados, percebe-se que a POF 2008-2009 e o Censo Demográfico de 2010 do IBGE (Distrito Federal) estão organizados de tal forma que, em ambas as bases, as variáveis de idade, gênero e renda se comportam da mesma maneira, sendo que apenas a POF 2008-2009 possui informações relacionadas ao gasto (despesa) com chocolate. Com isso, a máquina, de maneira simplória, irá comparar as duas bases, aplicando o padrão de consumo estabelecido a partir da POF 2008-2009 no Censo 2010. Ou seja, indivíduos que possuem características semelhantes de idade, gênero e renda, hipoteticamente, terão padrões de consumo também semelhantes.

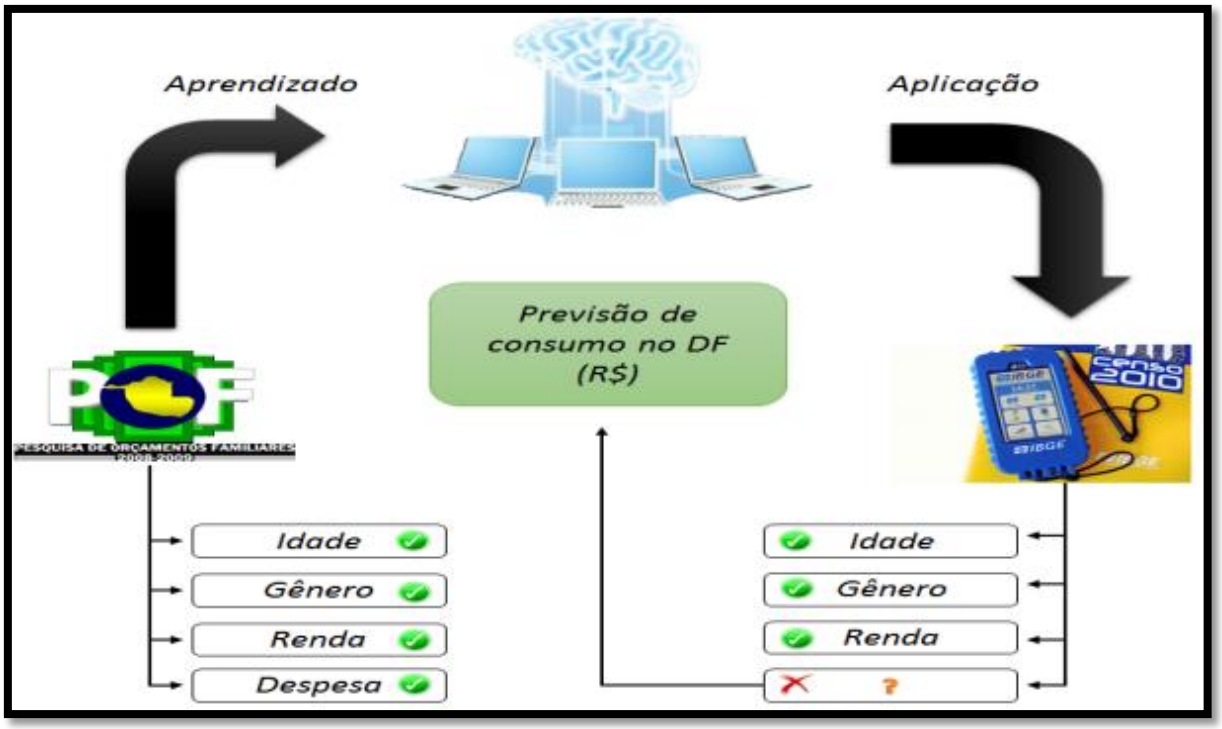


Figura 6: Esquematização do método de aprendizagem de máquina.
Fonte: Elaborada pelo autor.

A POF 2008-2009 é uma pesquisa de âmbito nacional e foi utilizada em sua totalidade para o estabelecimento de um padrão de consumo. Quanto ao censo 2010, que também se trata de uma pesquisa com abrangência nacional, foram utilizados apenas os dados referentes ao Distrito Federal.

Finalizada a etapa de aplicação, será possível determinar o Gasto Padronizado (GP) para cada setor censitário do Distrito Federal. O GP segue a mesma lógica do *score padronizado* e foi definido aqui como sendo a diferença entre a previsão do Gasto (G), atribuído a um determinado setor censitário, e o Gasto Médio $(\overline{G})$ registrado no Distrito Federal, divido pelo desvio-padrão (σ), como mostra a equação (3.2) a seguir:

$$GP = \frac{(G - \overline{G})}{\sigma} \tag{3.2}$$

Essa equação permite a comparação da posição relativa de cada setor censitário, no quesito despesa com chocolate, em relação aos demais. Assim, quanto mais próximo de zero for o GP do setor censitário, mais próximo ele estará do Gasto Médio do Distrito Federal $(\overline{G})$.

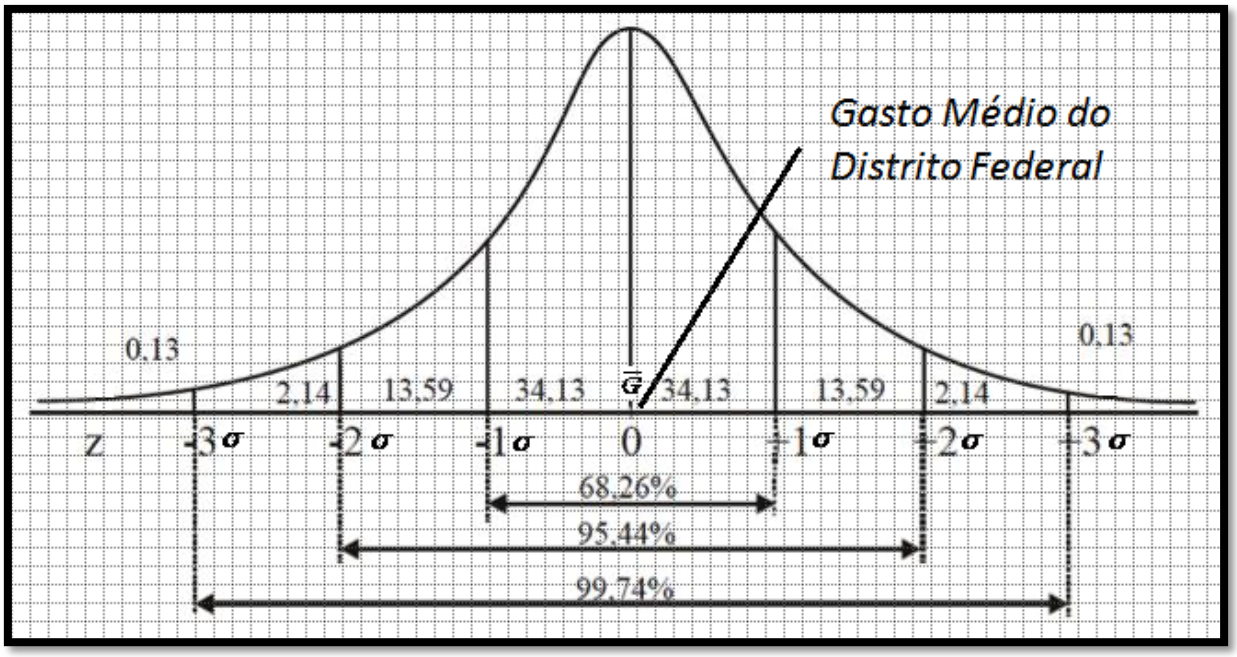


Figura 7: Curva de distribuição normal e o Desvio-Padrão.
Fonte: Elaborada pelo autor.

Um conceito importante para a análise dos resultados aqui pretendida é a definição do multicitado setor censitário. O IBGE (2010) define um setor censitário da seguinte forma:

> *"O setor censitário é a unidade territorial estabelecida para fins de controle cadastral, formado por área contínua, situada em um único quadro urbano ou rural, com dimensão e número de domicílios que permitam o levantamento por um recenseador."* (IBGE, 2010).

Sendo assim, Por meio do índice (GP), os setores censitários foram classificados de acordo com o seu nível de demanda pelo chocolate, como mostra a tabela abaixo:

N	COR	INTERVALO DE GP	CLASSIFICAÇÃO
1		0.00000 - 0.08522	Demanda Extremamente Baixa
2		0.08523 - 0.08617	Demanda Muito Baixa
3		0.08618 - 0.08685	Demanda Baixa
4		0.08686 - 0.08738	Demanda Média Baixa
5		0.08739 - 0.08787	Demanda Média
6		0.08788 - 0.08838	Demanda Média Alta
7		0.08839 - 0.08904	Demanda Alta
8		0.08905 - 0.08997	Demanda Muito Alta
9		0.08998 - 0.09285	Demanda Extremamente alta

Tabela 5: Classificação dos setores censitários por demanda.
Fonte: Elaborada pelo autor.

Como mostra a **Tabela 5**, a classificação dos setores censitários foi construída considerando nove níveis de demanda. No nível mais baixo, os setores censitários que apresentaram um Gasto Padronizado entre 0.00000 e 0.08522 possuem uma demanda por chocolate muito baixa se comparados ao restante do Distrito Federal. Para o nível mais alto, aqueles que tiveram seu GP entre 0.08998 e 0.09285 foram considerados como regiões de demanda muito alta pelo produto.

4 RESULTADOS

4.1 Oportunidades de negócios

Finalizada a etapa de treinamento da máquina e de obtenção do consumo padronizado para cada setor censitário do Distrito Federal, inicia-se o processo de reconhecimento dos pontos que poderão ser classificados como oportunidades de negócio. Esta etapa consiste em identificar as áreas que apresentam uma predisposição para o cosumo de chocolate, levando em consideração as informações geradas pelo método Aprendizado de Máquina e a aplicação dos padrões de consumo, oriundos da POF 2008-2009, na base de dados do censo demográfico do IBGE (2010) para o Distrito Federal.

Além das informações, obtidas na sessão anterior, sobre qual o perfil de quem consome o chocolate e quais são as características demográficas desses indivíduos, foram definidos alguns critérios para que um ponto qualquer possa ser classificado como uma oportunidade de negócio. Estabeleceu-se, portanto, que uma oportunidade de negócio deve respeitar os seguintes critérios:

I. CRITÉRIOS ELIMINATÓRIOS

a. estar dentro de uma região classificada, pelo IBGE, como urbana;
b. estar localizado em um ponto comercial;
c. estar inserido em um setor censitário que apresente, de acordo com o seu GP, uma classificação de demanda alta, muito alta ou extremamente alta.

II. CRITÉRIOS CLASSIFICATÓRIOS

a. quanto mais alto a previsão de GP melhor será a oportunidade de negócio em relação às demais.

Relacionando-se ao supracitado item “b” dos critérios eliminatórios, que trata da restrição de concorrência em um mesmo estabelecimento comercial, existe a necessidade de uma definição detalhada sobre quais lojas, presentes no Distrito Federal, serão consideradas como concorrência para uma nova franquia da Cacau Show.

Dessa forma, prosseguiu-se com a busca, através do *Google Maps*, por lojas que possuam uma proposta parecida com aquela apresentada pela Cacau Show. O

resultado dessa pesquisa foi uma lista com 44 lojas, onde 15 são da própria Cacau Show, visto que a alocação dessas lojas próximas uma das outras pode gerar um processo de “canibalismo” dentro da rede. Além disso, 9 lojas pertencem à franquia Brasil Cacau, 10 lojas à rede Kopenhagen, 3 lojas à Kaebisch Chocolate e as marcas Aguimar Ferreira Bombons, Baby Chocolates, Brigadeirando, Chocolataria Gramado Brasília, Dulce Patagonia, Fábrica de Chocolate e Stans Chocolate possuem uma loja cada.

No próprio site do *Google Maps* foi possível obter as coordenadas de cada uma dessas lojas, que serão necessárias para a criação dos mapas. Para obter a Latitude e Longitude de um ponto pelo *Google Maps*, basta clicar com o botão direito do mouse no local desejado do mapa e, em seguida, selecionar a opção *“o que há aqui?”.* Essas informações foram organizadas na **Tabela 6** a seguir:

RELAÇÃO DAS LOJAS DE CHOCOLATE NO DISTRITO FEDERAL

n	Franquias	Latitude	Longitude
1	Aguimar Ferreira Bombons	-15.794664	-47.932744
2	Baby Chocolate	-15.721464	-47.885227
3	Brigadeirando	-15.797283	-47.918500
4	Cacau Show	-15.764850	-47.885547
5	Cacau Show	-15.771168	-47.883658
6	Cacau Show	-15.780587	-47.886290
7	Cacau Show	-15.791697	-47.883233
8	Cacau Show	-15.811305	-47.897661
9	Cacau Show	-15.826655	-47.906740
10	Cacau Show	-15.826903	-47.923091
11	Cacau Show	-15.826650	-47.982081
12	Cacau Show	-15.836614	-47.867390
13	Cacau Show	-15.794612	-48.113322
14	Cacau Show	-15.803363	-48.068285
15	Cacau Show	-15.815585	-48.101630
16	Cacau Show	-15.836694	-48.020073
17	Cacau Show	-15.863376	-48.079490
18	Cacau Show	-15.649792	-47.791864
19	Chocolataria Gramado Brasília	-15.740263	-47.894282
20	Chocolates Brasil Cacau	-15.789035	-47.883114
21	Chocolates Brasil Cacau	-15.756334	-47.882702

22	Chocolates Brasil Cacau	-15.768630	-47.887177
23	Chocolates Brasil Cacau	-15.791218	-47.883093
24	Chocolates Brasil Cacau	-15.818981	-47.895625
25	Chocolates Brasil Cacau	-15.831197	-47.915881
26	Chocolates Brasil Cacau	-15.799862	-47.923055
27	Chocolates Brasil Cacau	-15.811279	-48.063955
28	Chocolates Brasil Cacau	-15.836139	-48.057459
29	Dulce Patagonia	-15.762371	-47.889152
30	Fábrica di Chocolate	-15.733239	-47.899765
31	Kaebisch Chocolate - Chocolate Belga	-15.744838	-47.886888
32	Kaebisch Chocolate - Chocolate Belga	-15.751507	-47.885517
33	Kaebisch Chocolate - Chocolate Belga	-15.801581	-47.926417
34	Kopenhagen	-15.720829	-47.886469
35	Kopenhagen	-15.786741	-47.889016
36	Kopenhagen	-15.791218	-47.883374
37	Kopenhagen	-15.796127	-47.891996
38	Kopenhagen	-15.799137	-47.887729
39	Kopenhagen	-15.803945	-47.940756
40	Kopenhagen	-15.809641	-47.943461
41	Kopenhagen	-15.833039	-47.956108
42	Kopenhagen	-15.841277	-48.044323
43	Kopenhagen	-15.810034	-47.943877
44	Stans Chocolates	-15.818474	-47.896079

Tabela 6: Identificação da concorrência para uma franquia da Cacau Show.
Fonte: Elaborado pelo autor com o auxílio do *Google Maps.*

Em seguida, utilizou-se o programa *Qgis* para a criação dos mapas. Este *software* permite a junção, de forma prática, entre o Gasto Padronizado (GP), concorrência (**Tabela 6**) e as malhas digitais do Distrito Federal. Como resultado desse cruzamento de dados, foram criados uma série de mapas coloridos de acordo com o GP (**Tabela 5**) e sinalizados com a localização de cada uma das 44 lojas que vendem chocolates finos do Distrito Federal. Estes mapas estarão disponíveis, por questões de organização, no **apêndice A** deste trabalho.

Com isso, foi feita uma análise individual para cada um dos 19 subdistritos do DF, com o intuito de listar, de acordo com os critérios elencados, todos os possíveis pontos de oportunidade para o mercado de chocolates finos do Distrito Federal.

4.1.1 Subdistrito Brasília

O subdistrito Brasília apresentou um bom índice de demanda, considerando que a maior parte do seu território apresenta uma classificação entre “Média Alta” e

“Muito Alta”. Apesar disso, não parece ser uma boa região para a instalação de uma nova franquia da Cacau-Show, pois existe uma concentração muito alta de unidades consideradas como concorrência, incluindo lojas da própria Cacau-Show. Ainda assim, caso haja interesse na instalação de uma loja nesse subdistrito, indica-se o setor hoteleiro norte, visto que ainda não possui concorrência próxima e apresentou um bom índice de GP.

4.1.2 Subdistrito Brazlândia

O subdistrito Brazlândia não apresenta um bom indicativo de GP. Observando o mapa percebe-se que a maior parte do seu território está classificado entre as faixas “Extremamente Baixa” e “Média Baixa”. Dessa forma, apesar de não existir concorrência, não é indicada a abertura de uma franquia de Cacau-Show nessa região.

4.1.3 Subdistrito Candangolândia

O subdistrito Candangolândia não possui nenhum setor censitário com índices de GP suficientes para a identificação de uma oportunidade de negócio. A classificação de demanda está aproximadamente entre “Extremamente Baixa” e “Baixa”, sendo possível afirmar que não se trata de um bom local para a instalação de uma empresa de chocolates finos.

4.1.4 Subdistrito Ceilândia

O subdistrito Ceilândia apresenta uma previsão de demanda predominante entre “Extremamente Baixa” e “Média”. Apesar disso, existem pontos dentro do seu território que se encontram dentro da faixa necessária para a identificação de uma oportunidade de negócio. Sendo assim, é uma opção viável a instalação de uma unidade da Cacau-Show. Além disso, existem apenas duas lojas concorrentes e estas se encontram em regiões com níveis baixos de GP, o que as tornam menos competitivas.

4.1.5 Subdistrito Cruzeiro

O subdistrito Cruzeiro possui seis lojas concorrentes em seu território e apresenta uma demanda bem equilibrada, apresentando um GP entre “Muito Baixa” e “Extremamente Alta”, o que o classifica como sendo um bom local para a instalação de uma loja da Cacau-Show. Apesar disso, se trata de uma região basicamente residencial, possuindo poucos estabelecimentos comerciais. Considerando que se trata de uma região pequena e que possui uma quantidade relativamente alta de lojas concorrentes, pode ser classificada com uma região arriscada para investimento, porém com possibilidade de atingir bons níveis de demanda.

4.1.6 Subdistrito Gama

O subdistrito Gama apresenta níveis baixos de GP, atingindo níveis de demanda entre “Extremamente baixa” e “Média”. Este fato já exclui a possibilidade de identificação de oportunidades de negócio na região. Além disso, não foi identificada nenhuma concorrência em seu território, o que pode ser explicado pela pouca demanda pelo produto no local.

4.1.7 Subdistrito Guará

O subdistrito do Guará possui duas lojas concorrentes em seu território ao mesmo tempo que não apresenta bons níveis de GP. Apesar da maior parte dos seus setores censitários apresentarem níveis de demanda entre “Extremamente Baixa” e “Média alta”, existem alguns locais com uma classificação aceitável para a identificação de uma oportunidade de negócio. Portanto, há mercado para uma franquia da Cacau-Show, porém a escolha locacional deve ser feita de maneira estratégica.

4.1.8 Subdistrito Lago Norte

O subdistrito Lago Norte possui uma demanda abaixo do desejado, variando majoritariamente o seu nível entre as faixas “Muito Baixa” e “Média Alta”. O único ponto que apresenta bons níveis de demanda, alcançando a classificação “Muito Alta”, está abastecido com duas lojas concorrentes próximas uma da outra. Isto

significa que não é interessante, para a Cacau-Show, abrir uma loja nesse subdistrito, visto que o único local que apresenta um nível de demanda aceitável possui muita oferta do produto.

4.1.9 Subdistrito Lago Sul

O mapa do subdistrito Lago Sul mostra que os moradores dessa região possuem níveis de GP bastante satisfatórios, tendo sua grande maioria variando entre “Média” e “Extremamente Alta”. Além disso, existe apenas uma loja concorrente em todo o seu território, que está inserida em uma região de demanda média e um pouco distante de locais classificados com uma demanda “Extremamente Alta”. Isso faz do Lago Sul um local muito bom para a abertura de uma loja de chocolates finos, considerando que se trata de uma região nobre do Distrito Federal, ou seja, as pessoas possuem um elevado nível de renda, e que apresenta uma previsão de consumo estimada como sendo “Extremamente alta”. Para melhorar a situação, a única concorrente inserida no subdistrito não se localiza perto das regiões de muita demanda, abrindo caminho para que uma nova entrante aproveite um mercado com um bom poder aquisitivo e que consome muito chocolate.

4.1.10 Subdistrito Núcleo Bandeirante

O subdistrito Núcleo Bandeirante não apresenta bons resultados de demanda estimada, tendo a maioria do seu território classificado entre os níveis “Muito Baixa” e “Média”. Entretanto, existem dois subdistritos com demanda entre “Alta” e “Muito Alta”, abrindo espaço para uma franquia da Cacau-Show. Outro ponto positivo é que não existe concorrência nesse subdistrito. Portanto, entende-se que no Núcleo Bandeirante existem oportunidades de negócio para o mercado de chocolates finos.

4.1.11 Subdistrito Paranoá

O subdistrito Paranoá possui dois grandes setores censitários, sendo que o primeiro possui uma demanda “Média” e o segundo uma demanda “Extremamente Baixa”. Os demais setores censitários da região possuem níveis de GP variando entre “Média Baixa” e “Muito Alta”. As regiões de demanda alta devem atrair o

mercado de chocolates finos pelo fato de não apresentarem concorrência. Talvez o problema dessa região talvez seja encontrar pontos comerciais, visto que a sua grande maioria é formada por áreas residenciais.

4.1.12 Subdistrito Planaltina

O subdistrito Planaltina apresenta uma grande variação de GP entre os seus setores censitários, que variam de “Extremamente Baixa” a “Extremamente Alta”. O fato de não possuir concorrência e ter vários locais com GP nos níveis “Muito Alta” e “Extremamente Alta” faz com que o subdistrito seja um bom alvo para o mercado de chocolates finos.

4.1.13 Subdistrito Recanto das Emas

O subdistrito Recanto das Emas possui a maioria do seu território classificado entre os níveis “Muito Baixa” e “Média Alta”. Os setores censitários que se enquadram em níveis suficientes para representarem oportunidades de negócio são regiões muito pequenas e, geralmente, rodeadas por setores com baixa estimativa de GP. Isto torna o investimento arriscado e limitado. Sendo assim, pode não ser aconselhável a abertura de uma loja de chocolates finos do Recanto das Emas.

4.1.14 Subdistrito Riacho Fundo

Para a região do Riacho Fundo, apresento as mesmas críticas feitas para o subdistrito Recanto das Emas. As regiões que apresentam um GP entre os níveis “Alta” e “Extremamente Alta” são muito pequenas e se encontram isoladas por regiões com estimativa de demanda inferior, aumentando o risco envolvido no empreendimento. Dessa forma, entende-se que não é vantajoso para o investidor abrir uma loja de chocolates finos no Riacho Fundo.

4.1.15 Subdistrito Samambaia

O subdistrito Samambaia possui em seu território uma loja concorrente e uma predominância na classificação de GP entre os níveis “Muito Baixa” e “Média”. Observa-se, a partir do mapa, que a loja existente se encontra distante dos maiores setores que apresentam uma previsão de demanda “Muito Alta” e “Extremamente

Alta”, fazendo com que essas regiões se tornem possíveis oportunidades de negócio.

4.1.16 Subdistrito Santa Maria

O Subdistrito Santa Maria possui setores censitários grandes e com uma previsão de GP majoritariamente entre “Alta” e “Extremamente Alta”. Além disso, a região não possui concorrência para esse mercado, o que representa um ponto positivo para a análise. Entretanto, a maior parte do subdistrito é formada por regiões residenciais, o que compromete um pouco a análise de oportunidades de negócios, visto que só é possível abrir uma loja da Cacau-Show em pontos comerciais.

4.1.17 Subdistrito São Sebastião

O Subdistrito São Sebastião possui sua maioria classificada como “Muito Alta” e “Extremamente Alta”. Entretanto, esses grandes setores com uma boa estimativa de demanda são regiões em que não existem centros comerciais e, portanto, não podem ser caracterizados como regiões com oportunidade de negócio. Os setores que possuem esses comércios apresentam uma classificação entre “Muito Baixa” e “Média”, o que já exclui a possibilidade de identificação bons pontos comerciais.

4.1.18 Subdistrito Sobradinho

O subdistrito Sobradinho não é muito vasto, porém apresenta uma grande variabilidade na classificação do GP, variando de “Extremamente Baixa” até “Extremamente Alta”. Além disso, a região possui uma loja concorrente que se encontra rodeada por setores com classificação “Média”. As regiões com demanda “Extremamente Alta” não possuem concorrência, o que faz delas boas opções para a instalação de um negócio como esse.

4.1.19 Subdistrito Taguatinga

O subdistrito de Taguatinga é uma região muito grande que pode ser dividida em blocos de setores que apresentam comportamentos de consumo semelhantes em relação ao chocolate fino. Existem dois “blocos” de setores que apresentam

níveis de GP muito satisfatórios, variando basicamente entre “Muito Alta” e “Extremamente Alta”, enquanto que os demais “blocos” possuem uma estimativa de demanda abaixo do desejável, variando entre “Muito baixa” e “Média”, com alguns pontos isolados, portanto arriscados, com demanda “Extremamente Alta”. Além disso, o subdistrito contém cinco lojas concorrentes, sendo que todas elas estão localizadas nas regiões com elevadas projeções de GP. Este cenário faz de Taguatinga um bom lugar para a instalação de uma loja da Cacau-Show, visto que, apesar da concorrência, o mercado da região é muito grande e, pelo fato de existirem esses dois blocos com altos níveis de GP, os potenciais consumidores se encontram agrupados, o que pode potencializar os resultados do negócio.

5 CONCLUSÃO E RECOMENDAÇÕES

5.1 Conclusão

O método, aqui proposto, consistiu em construir uma ferramenta quantitativa capaz de identificar oportunidades de negócio no Distrito Federal. Para isso, foi escolhido o mercado de chocolates finos e iniciou-se o tratamento dos dados da Pesquisa de Orçamentos Familiares de 2008-2009, criando as variáveis desejadas e eliminando as informações desnecessárias para esta análise. Dentro da POF 2008-2009, foram selecionados produtos relacionados com o chocolate fino. Em seguida, foram identificados quais os indivíduos consumiram pelo menos um dos itens selecionados, obtendo o gasto total de cada respondente com chocolate, bem como suas informações de renda, gênero, idade e localização. A partir dessas informações, o classificador foi capaz de estabelecer um padrão para o consumo do produto, tendo, como pré-requisito para a previsão, as características obtidas na POF – 2008-2009. Após o Aprendizado de Máquina, ocorreu a etapa de aplicação na base de dados do Censo do IBGE de 2010, tendo com o resultado final o mapeamento da demanda de chocolates finos para todo o Distrito Federal.

A partir da análise dos resultados e do mapeamento bem-sucedido das oportunidades de negócio para o mercado de chocolates finos no Distrito Federal, pode-se dizer, com um embasamento prático, que o *Geomarketing*, aliado às técnicas de aprendizado de máquina, representa uma ferramenta com recursos extremamente valiosos para a delimitação de estratégias organizacionais, seja no auxílio de ações de *Marketing* ou em questões de escolha locacional e expansão dos negócios, com possibilidade de foco e previsões de consumo para um público alvo específico.

É possível afirmar, também, que o objetivo geral e os específicos, traçados no início deste trabalho, foram atingidos. Isto se justifica pela construção, por meio por meio de cada objetivo específico listado, de uma ferramenta quantitativa capaz de auxiliar o processo de decisão locacional de uma rede de franquias. O problema de pesquisa, que questionava se a ferramenta poderia representar uma vantagem competitiva sustentável para uma rede de franquias, foi solucionado, considerando que o método atingiu todos os requisitos, baseados na definição de Barney (1991,1996), sejam eles:

I. **Valioso:** Foi capaz de explorar ou neutralizar as ameaças do ambiente, gerando uma redução de custos ou aumento das receitas da empresa;
II. **Raro:** Considerando que se trata de uma área de conhecimento pouco explorada, pode ser considerado um recurso raro entre os concorrentes atuais e potenciais da empresa;
III. **Difícil de ser imitado:** Trata-se de um recurso intelectual, portanto não pode ser facilmente copiado;
IV. **Difícil de ser substituído:** Não foram encontradas estratégias equivalentes e que seja, também, rara e difícil de ser imitada entre os concorrentes.

Um dos maiores êxitos deste estudo foi a sua contribuição intelectual para uma área do conhecimento que ainda é pouco difundida, principalmente no Brasil, e que pode trazer muitos benefícios, tanto para o setor privado quanto para a esfera pública. O diferencial dessa metodologia é a possibilidade de se fazer estimativas de demanda, ou aceitação de determinado produto, para novos mercados, ou seja, regiões que ainda não possuem um histórico de consumo em relação ao objeto de análise. Além disso, o resultado final da análise, como mostra o apêndice I deste estudo, oferece ao gestor mapas georreferenciados contendo informações detalhadas sobre o funcionamento do mercado em relação ao seu setor de atividade. Exemplo disso é a classificação de cada subdistrito por faixas de Gasto Padrão (GP) e a localização da concorrência. Um produto final como este reduz os riscos envolvidos na decisão e permite que o gestor, tendo o relatório em mãos, consiga traçar suas estratégias de uma forma rápida e eficiente.

5.2 Limitações e recomendações

Quanto às limitações, foram identificadas algumas observações e recomendações para trabalhos futuros na área de *Geomarketing* e Aprendizado e Máquina. O objetivo desse estudo foi, de maneira objetiva, identificar oportunidades de negócios, dentro do Distrito federal, para o mercado de chocolates finos. Entretanto, seria interessante que fosse feita, após a identificação dos pontos classificados como oportunidades de negócios, um “ranqueamento” com base na estimativa de faturamento de cada ponto comercial. A intenção seria encontrar o local mais vantajoso para a abertura do negócio. Além disso, uma das limitações identificadas ao longo do estudo foi o fato de a base utilizada na aplicação – Censo

Demográfico do IBGE – conter informações coletadas no ano de 2010. Certamente as características demográficas do Distrito Federal sofreram modificações ao longo desses seis anos, o que acaba comprometendo, de certa forma, a previsão do classificador. A mesma crítica é feita para os dados de treinamento e validação – Pesquisa de Orçamentos Familiares – considerando que foram coletados durante os anos de 2008 e 2009 e podem não refletir, de forma fidedigna, o comportamento dos consumidores em 2016.

Outro ponto que requer uma atenção especial é o fato de a POF 2008-2009 não conseguir abarcar, dependendo do objeto escolhido para previsão, alguns produtos ou serviços. Exemplo disso são os mercados de luxo, que talvez não pudessem se utilizar desse conjunto de dados para mapear o seu mercado, considerando que a POF 2008-2009, pelo que foi observado, não possui em suas listas especificações de produto para este tipo de demanda. O mercado de chocolates finos, utilizado na aplicação do método, é uma área que se enquadra, de certa forma, nesse mercado de luxo. Quem consome um chocolate da Cacau-Show não é, necessariamente, a mesma pessoa que compra uma barra de chocolate no supermercado, o que acaba comprometendo a previsão do classificador. Apesar dessas limitações, que se concentraram basicamente na base de dados utilizada, o método se mostrou extremamente sólido e provou a sua utilidade, tanto para o setor privado quanto para a esfera pública.

REFERÊNCIAS

ARANHA, F. Sistemas de informação geográfica: Uma arma estratégica para o Database Marketing. São Paulo: v. 36, n. 2, p. 12-16, Abr/Mai/Jun. 1996.

ARNOFF, S. Geographic Information Systems: A Management Perspective. WDL, Ottawa, 1989.

BARNEY, J. Commentary: a hierarchy of corporate resources (A. L. Brumagin). In P. Shrivastava, A. S. Huff, & J. E. Dutton (Eds.),Advances in strategic management, v. 10, p. 113-125, 1994.

BARNEY, J. Firm resources and sustained competitive advantage. Journal of Management, p. 99-120, 1991.

BARNEY, J. Gaining and sustaining competitive advantage. Reading, MA: Adisson Wesley, 1996.

BARNEY, J. Integrating organizational behavior and strategy formulation research: a resource- based analysis. In P. Shrivastava, A. S. Huff, & J. E. Dutton (Eds.), Advances in strategic management, v. 8, p. 203-227, 1992.

BARNEY, J.; HESTERLY, W. Organizational economics: understanding the relationship between organizations and economics analysis. London: Handbook of organizations studies, p. 115-147, 1996.

BOSMAN, A. Structured problems and information systems (mimeographed). Groningen: Groningen State University, 1973.

BURROUGH, P. A; MC DONNELL, R. A. Principles of geographical information system for land resource assessment. Oxford: Claderon Press, 1998.

BUSCHIOLI, R. da R.; IAHN, J. F. Centralidade de campo: uma análise das conclusões de Cleonice Bourlegat através das teorias da localização e aglomeração. Disponível em: <http://egal2009.eadyplanners.info/area01/1248_Buscioli_Roberson_da_Rocha.pdf>. Acesso em: 25 abr. 2015.

CÂMARA G.; QUEIROZ, G. R. Arquiteturas de sistemas de informação geográfica. In: Câmara, G.; Davis, C.; Monteiro, A. (Org.). Introdução à Ciência da Geoinformação. São José dos Campos: INPE, 2004. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/introd/cap3-arquitetura.pdf>. Acesso em: 23 abr. 2015.

CÂMARA, G.; QUEIROZ, G. R. Arquitetura de Sistemas de Informaçãos Geográficas. In G. Câmara , A. M. Monteiro , & C. Davis (Eds). Introdução a ciência da geoinformação, 2006.

CARNASCIALI, A. M.; DELAZARI, L. S. A localização Geográfica como Recurso Organizacional: Utilização de Sistemas Especialistas para Subsidiar a Tomada de Decisão Locacional do Setor Bancário. Curitiba: RAC, v. 15, n. 1, art. 6, p. 103-125, 2011.

CLIQUET, G. Geomarketing: Methods and Strategies in Spacial Marketing. 1 ed. London: ISTE Limited, 2006.

COWEN, D. J. GIS versus CAD versus DBMS: what are the differences. Photogrammetric Engineering and Remote Sensing, v. 54, p. 1551-1554, 1988.

DAENZER, W. F. Systems engineering Leitfaden zur methodischen Durchfurung umfangreicher Plamungsvorhaben. Germany: Peter Hanstein, v. 2, 1978.

Eisgruber, L. M.; FORSTER, D. L. Development in the economic theory of information (plus discussion paper). American Journal of Agricultural Economics, p. 901, 1978.

ELDRANDALY, K.; ELDIN, N; SUI, D. A COM-based spatial decision support system for industrial site selection, Journal of Geographic Information and Decision Analysis, v. 7, p. 72-92, 2003.

ERIK DE MAN, W. H. Establishing a geographical information system in relation to its use: A process of strategic choices. International Journal of Geographical Information Systems. Vol 2, Issue 3, 1998.

FRIEDHEIM, A. Os termos de Franchising mais utilizados por quem é desse ramo. 2013. Disponível em <http://www.portaldofranchising.com.br/artigos-sobre-franchising/os-termos-de-franchising-mais-utilizados-por-quem-e-desse-ramo> Acesso em 15 abr. 2015.

FURLAN, A. A. Geoprocessamento: estudos de Geomarketing e as possibilidades de sua aplicação no planejamento do desenvolvimento socioeconômico. Espaço e Tempo, São Paulo, n. 29, p. 97-105, 2011.

GENARO, S. Sistema Especialista: o conhecimento artificial. Rio de Janeiro: Livros Técnicos e Científicos Editora S. A., 1986.

GRANT, R. M. Contemporary strategy analysis: concepts, techniques, applications. Cambrige, MA: Blackwell Business, 1996.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. Censo Demográfico de 2010: características da população e dos domicílios. Rio de Janeiro: IBGE, 2011.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. Pesquisa de Orçamentos Familiares 2008-2009: despesas, rendimentos e condições de vida. Rio de Janeiro: IBGE, 2010

JUNIOR, A. C. Módulo geomarketing. Apostila. NGeo-DECiv-UFSCAR, 2007.

LOCH, R. E. N. Cartografia: representação, comunicação e visualização de dados espaciais. Florianópolis: Editora da UFSC, 2006.

MAGUIRE, D.; GOODCHILD, M.; RHIND, D. (ed.). Geographical information systems. 2a. ed. : John Wiley and Sons, 1993. 2v.

MONARD, M. C., BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. Sistemas Inteligentes-Fundamentos e Aplicações, 2003.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. Rezende, V. 1, p. 89-114, 2003.

PAREDES, E. A. Sistemas de Informação Geográfica: Princípios e Aplicações (Geoprocessamento). São Paulo: Érica, 674 p., 1994.

PETERAF, M. A. The cornerstones of competitive advantage: resource-based view. Strategic Management Journal, p. 179-191, 1993.

PORTER, M. Towards a dynamic theory of strategy. Strategic Management Journal, v. 12, p. 95-117, 1991.

RODRIGUE, J. P. Von Thunen's regional land use model. Disponível em: < https://people.hofstra.edu/geotrans/eng/ch6en/conc6en/vonthunen.html>. Acesso em: 25 abr. 2015.

SALOMONSSON, O. Data gaps in the planning process: Na application to environmental planning. Proceedings of the Workshop on Information Requirements for Development Planning in Developing Countries. Enschede: International Institute for Aerospace Survey and Earth Sciences, 1980.

SEEMANN, C. C.; PRADO, K. Geoprocessamento como prática biopolítica no governo municipal. Rio de Janeiro: Revista de Adminsitração Pública, v. 46, 2012.

SHEPARD, D. et al. Database Marketing: O novo marketing direto. São Paulo: Makron Books, 1993.

SIMON, H. A. Administrative Behaviour. New York: The Free Press, n. 3, 1976.

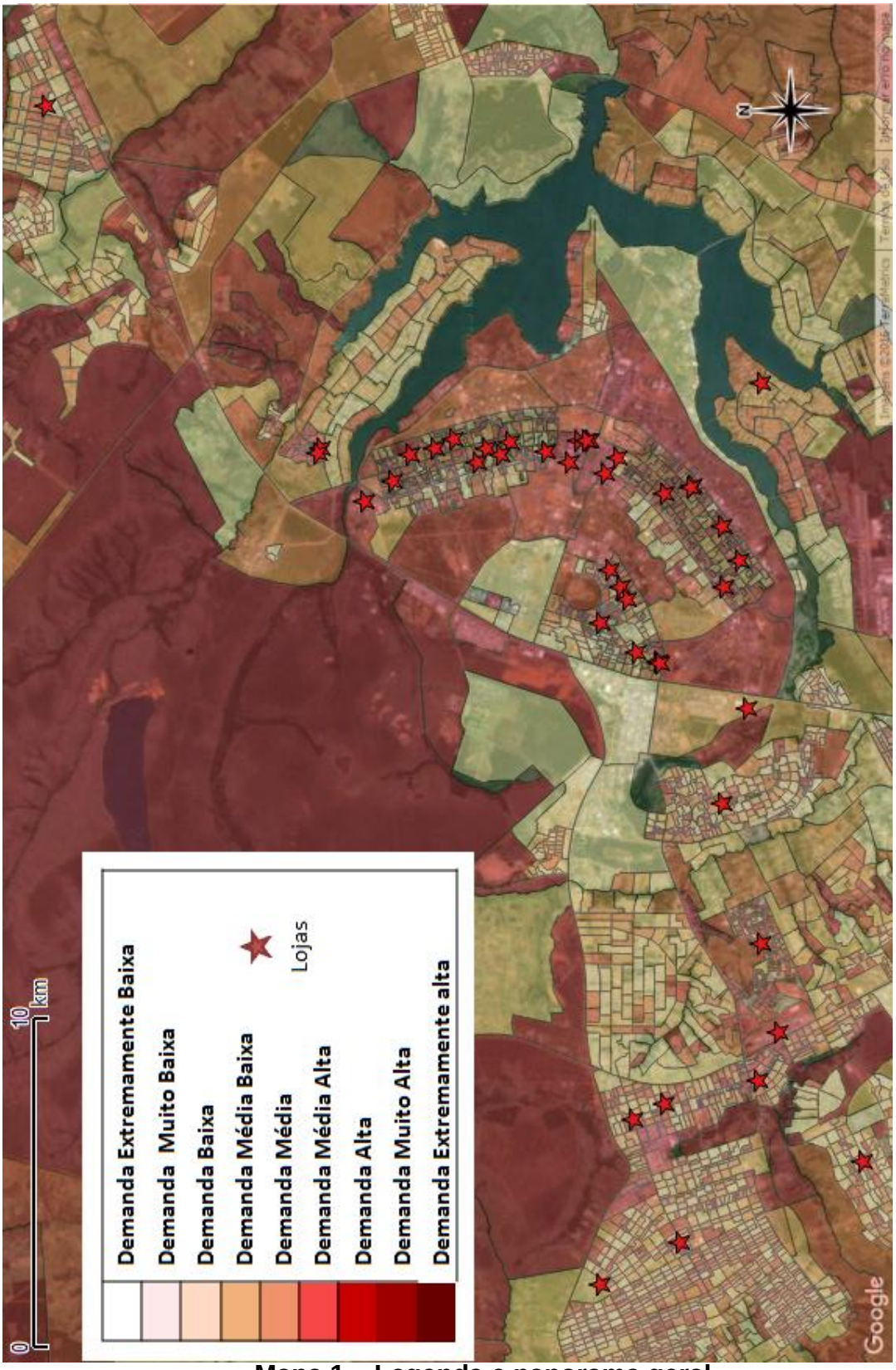
VIEIRA, C.; ALBERT, C.; BAGOLIN, I. P. A produção vitivinícola da Serra Gaúcha - Brasil e de Mendoza – Argentina: uma análise comparativa a luz das teorias de desenvolvimento regional. Disponível em: <http://online.unisc.br/seer/index.php/cepe/ article/viewFile/405/236>. Acesso em: 25 abr. 2015.

WATERMAN, D. A. A guide to expert systems. Canadá: Addison-Wesley Publishing Company, 1986.

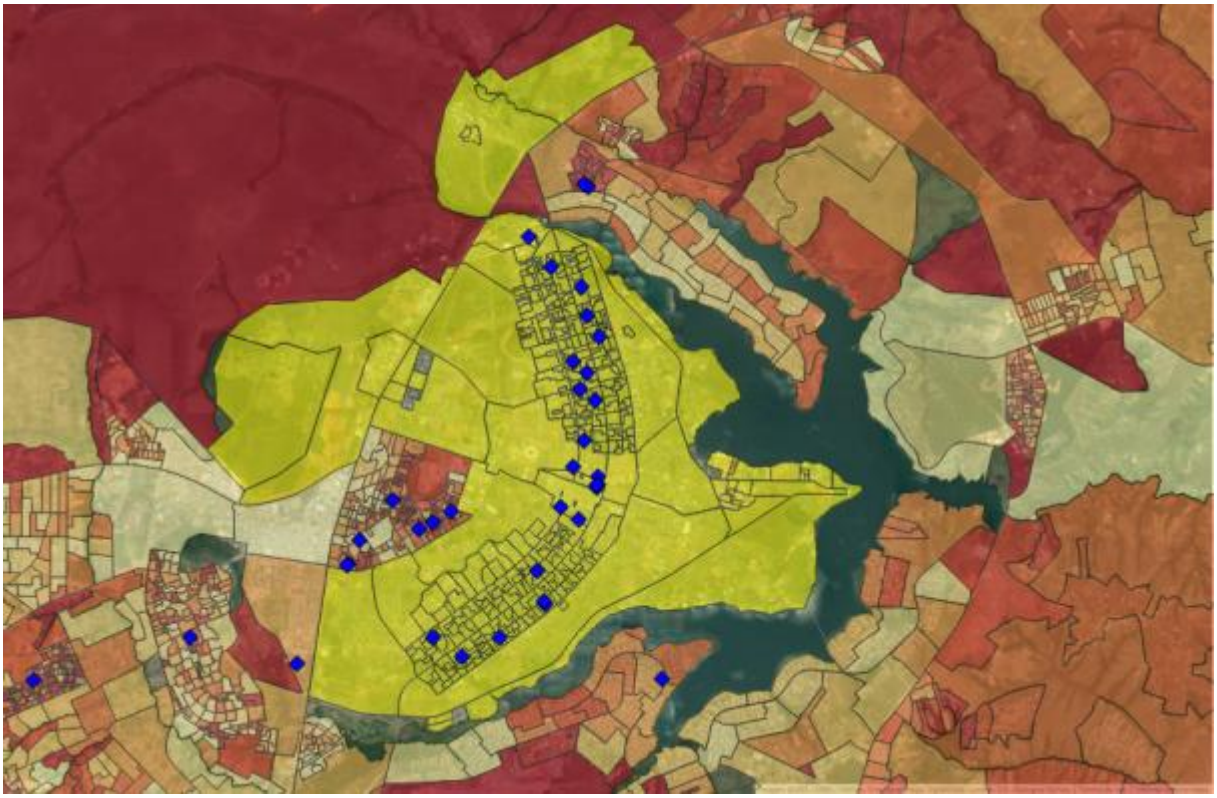
WATERMAN, D. A. Building expert systems. Canadá: Addison-Wesley Publishing Company, 1983.

WENERFELT, B. A resource-based view of the firm. Strategic Management Journal, v. 5, p. 171-80, 1984.

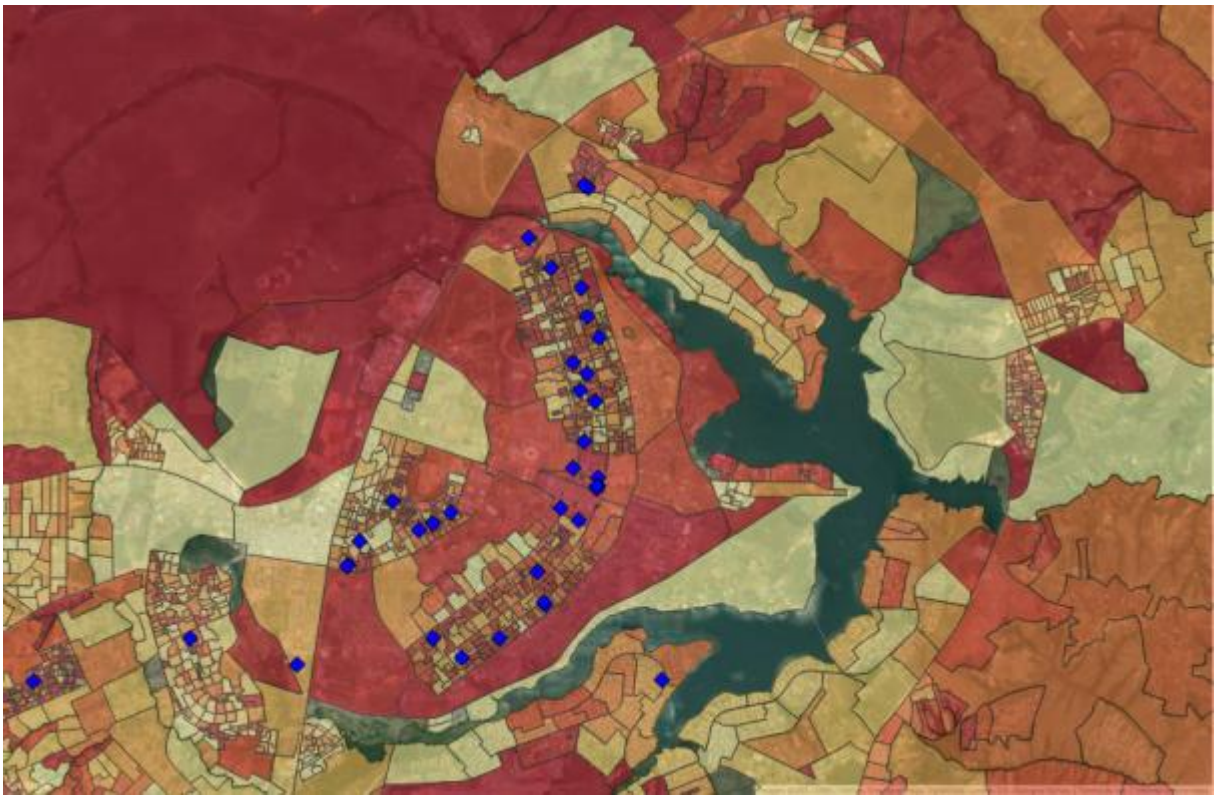
Apêndice A – Mapas desenvolvidos no QGIS



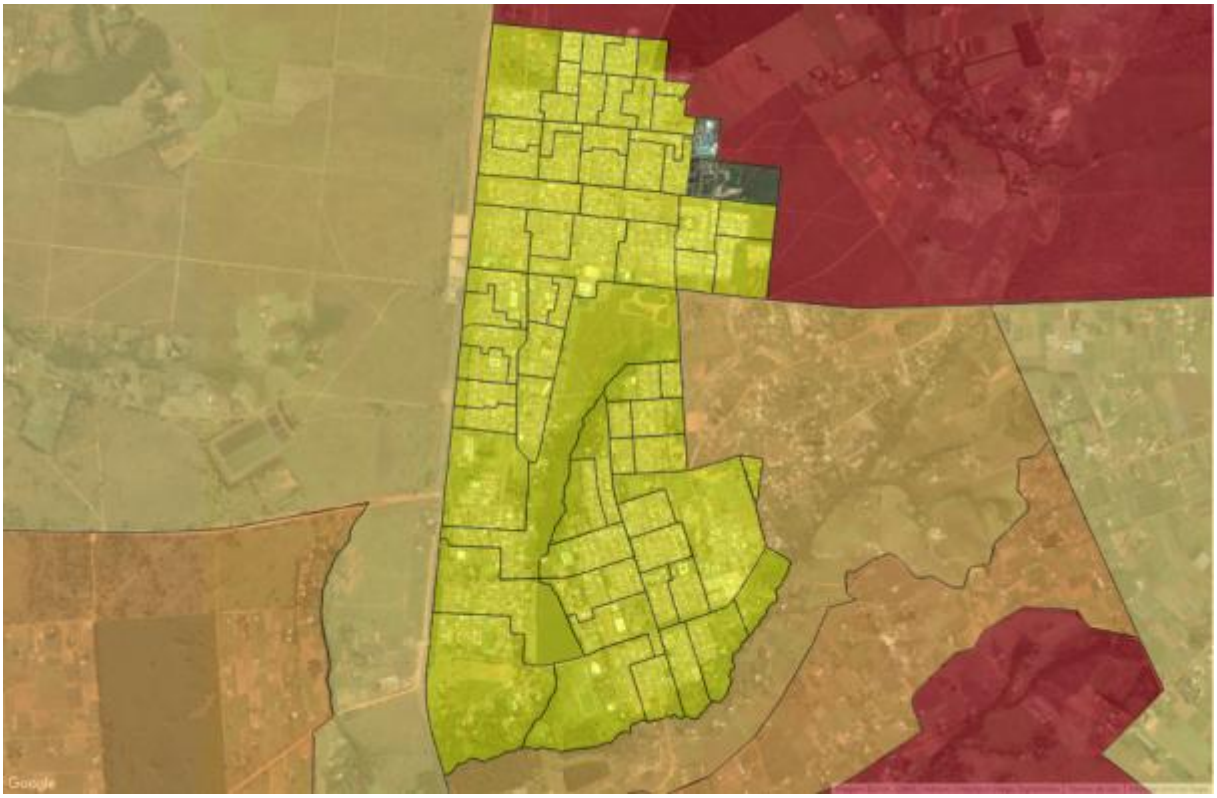
Mapa 1 – Legenda e panorama geral



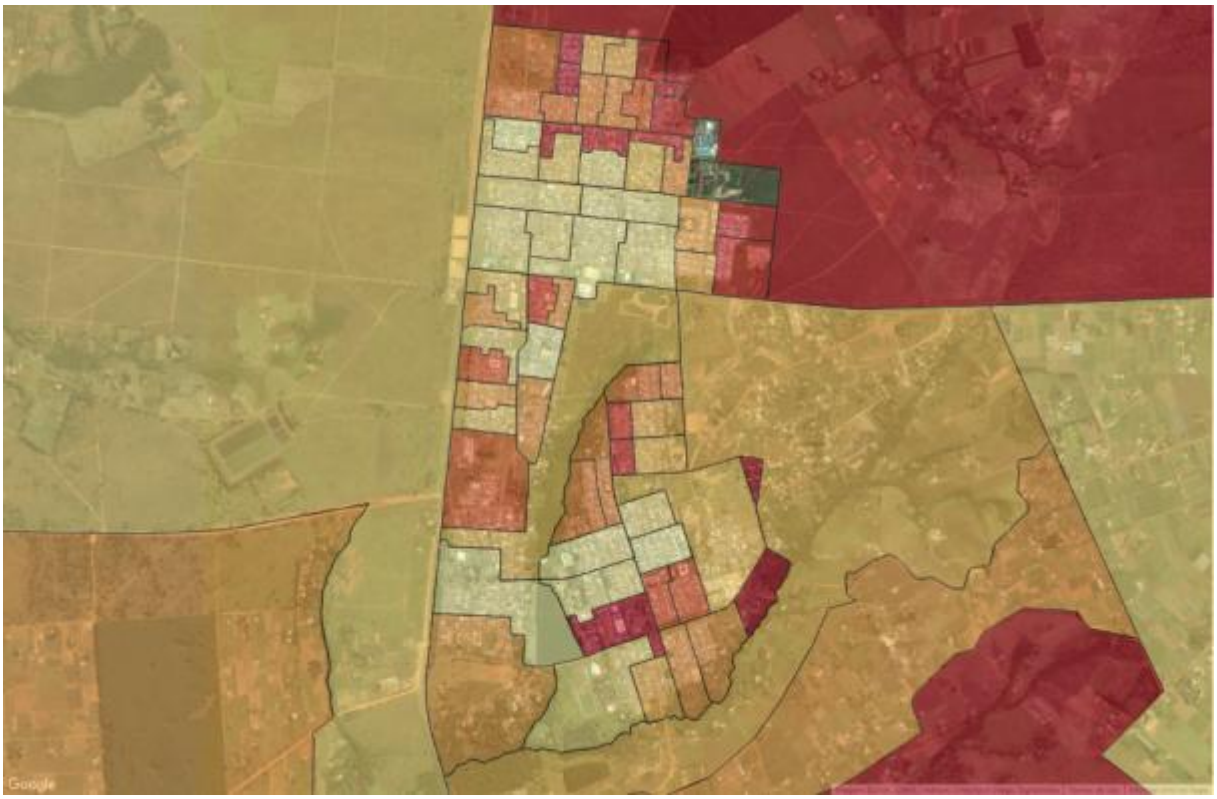
Mapa 2 – Subdistrito Brasília: Indicação urbana



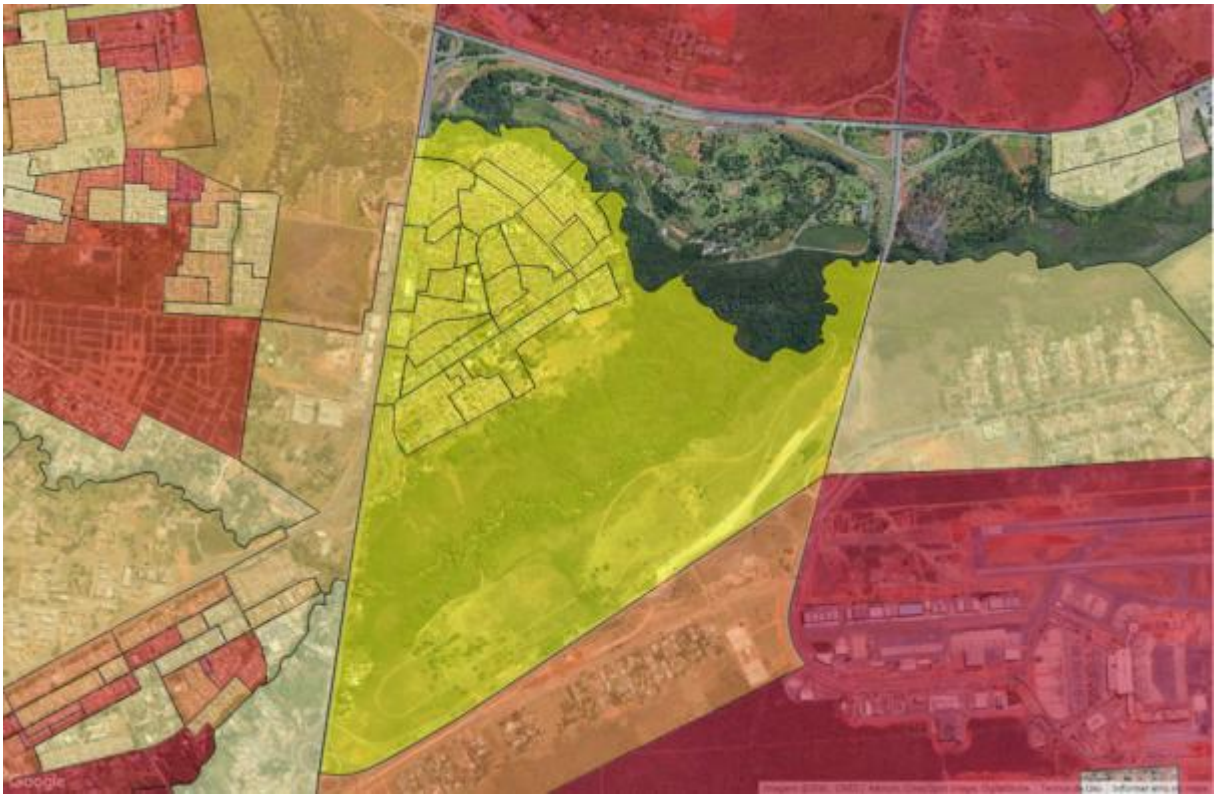
Mapa 3 – Subdistrito Brasília: Análise de demanda



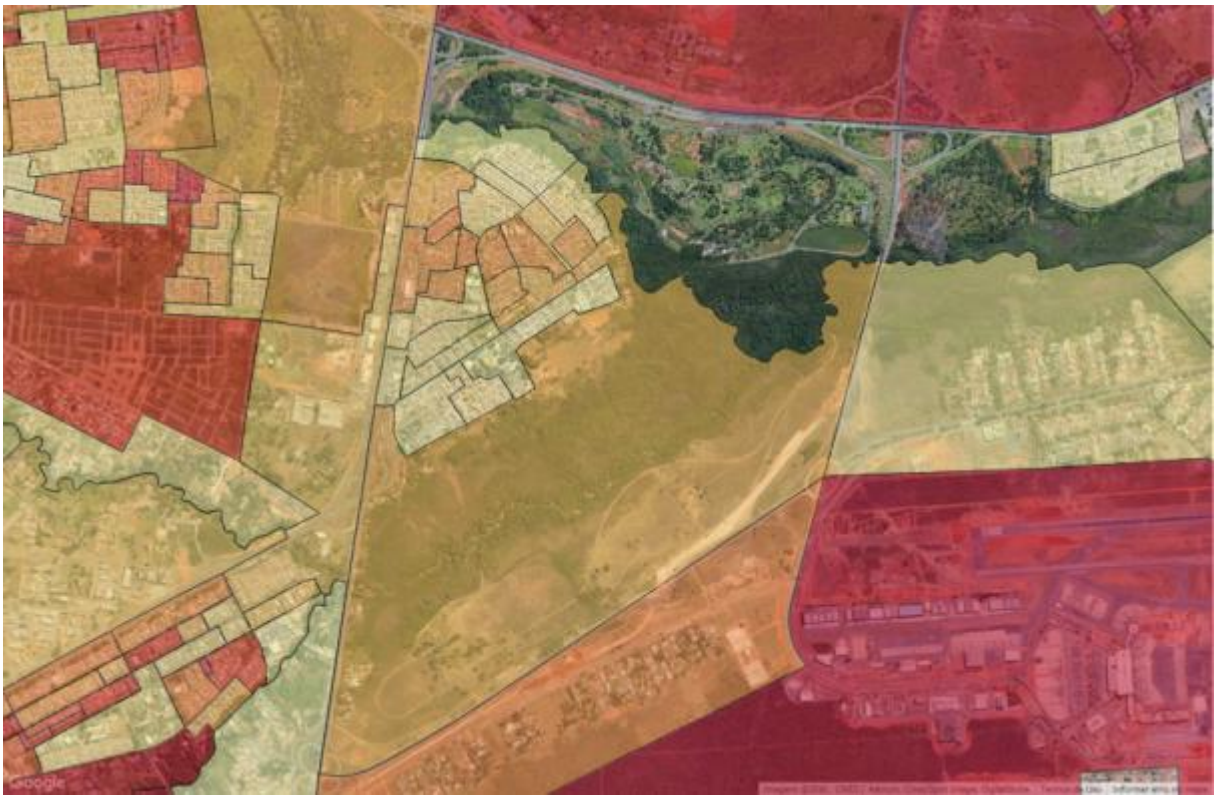
Mapa 4 – Subdistrito Brazlândia: Indicação urbana



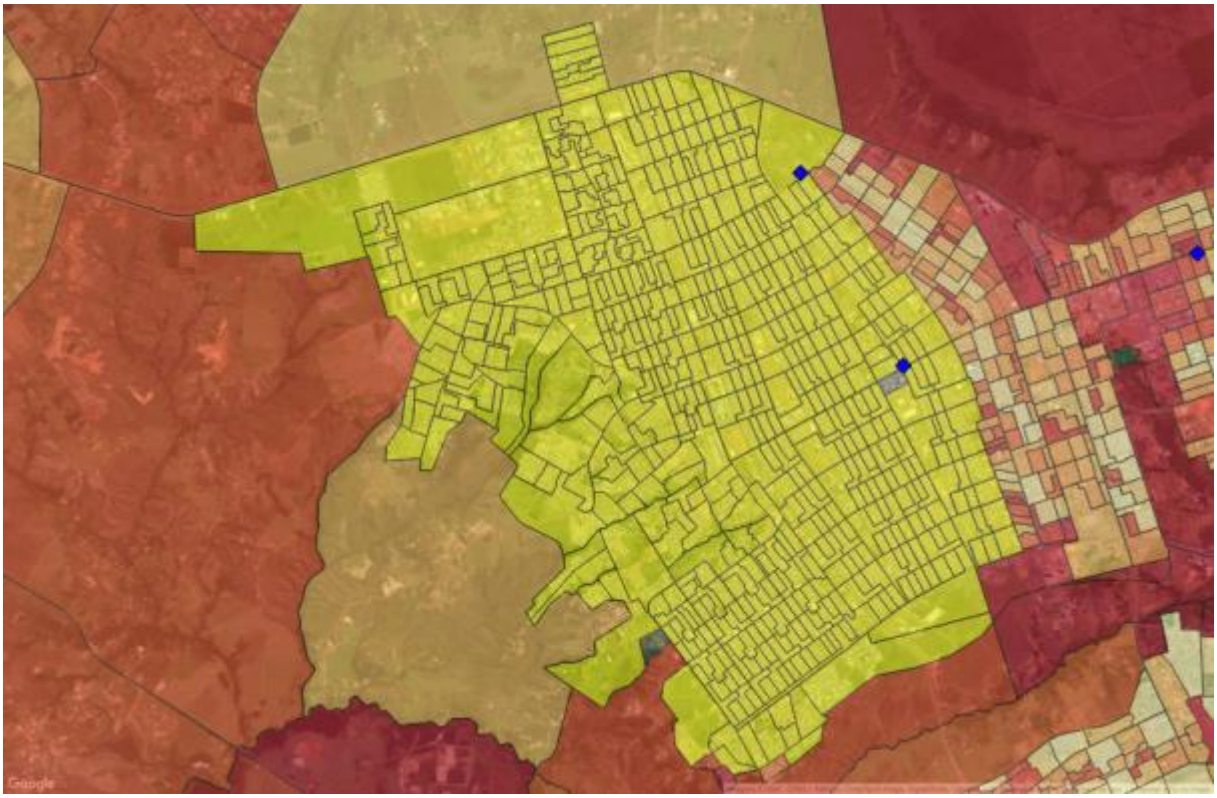
Mapa 5 – Subdistrito Brazlândia: Análise de demanda



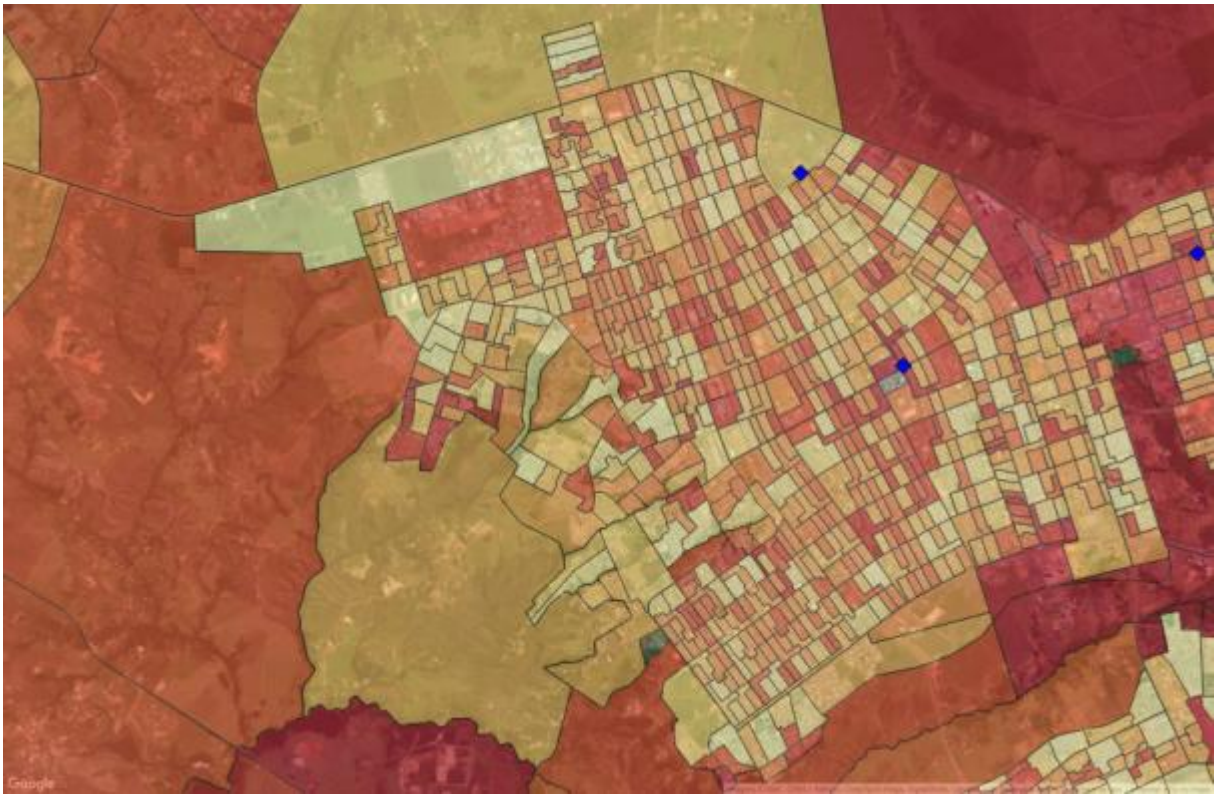
Mapa 6 – Subdistrito Candangolândia: Indicação urbana



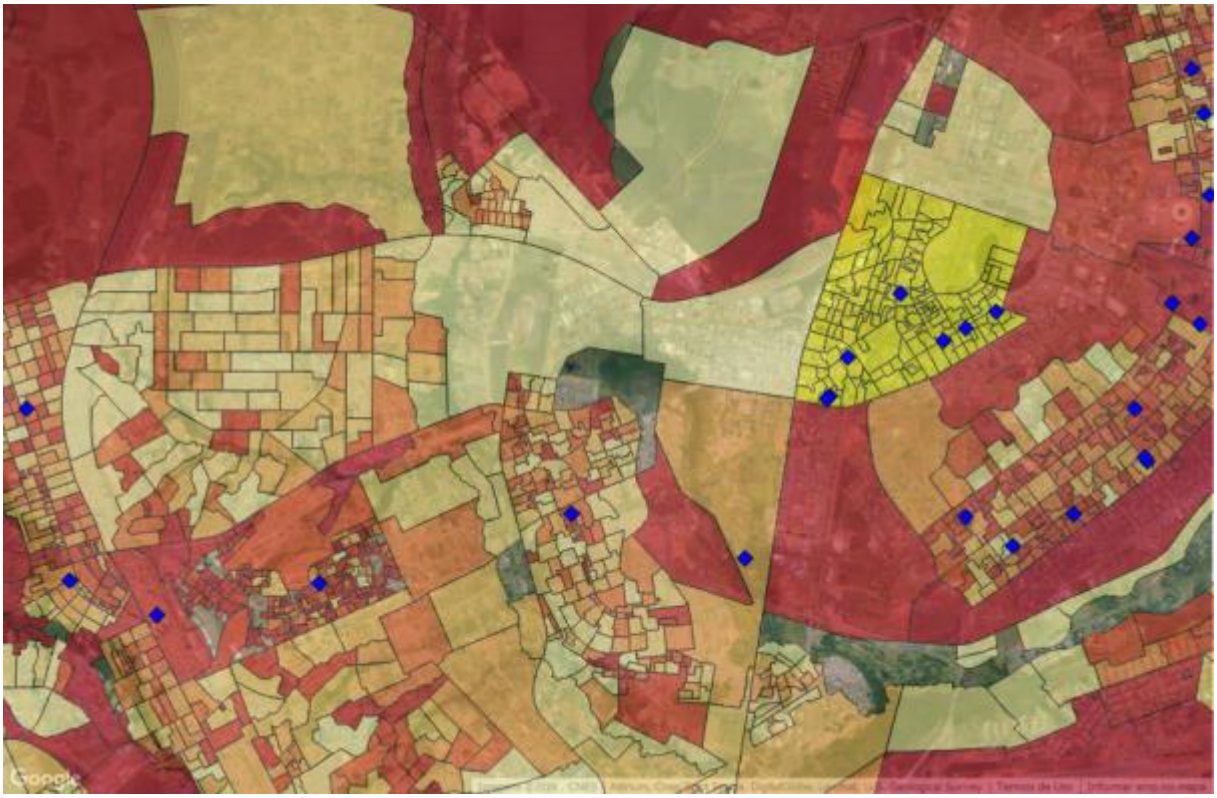
Mapa 7 – Subdistrito Candangolândia: Análise de demanda



Mapa 8 – Subdistrito Ceilândia: Indicação urbana

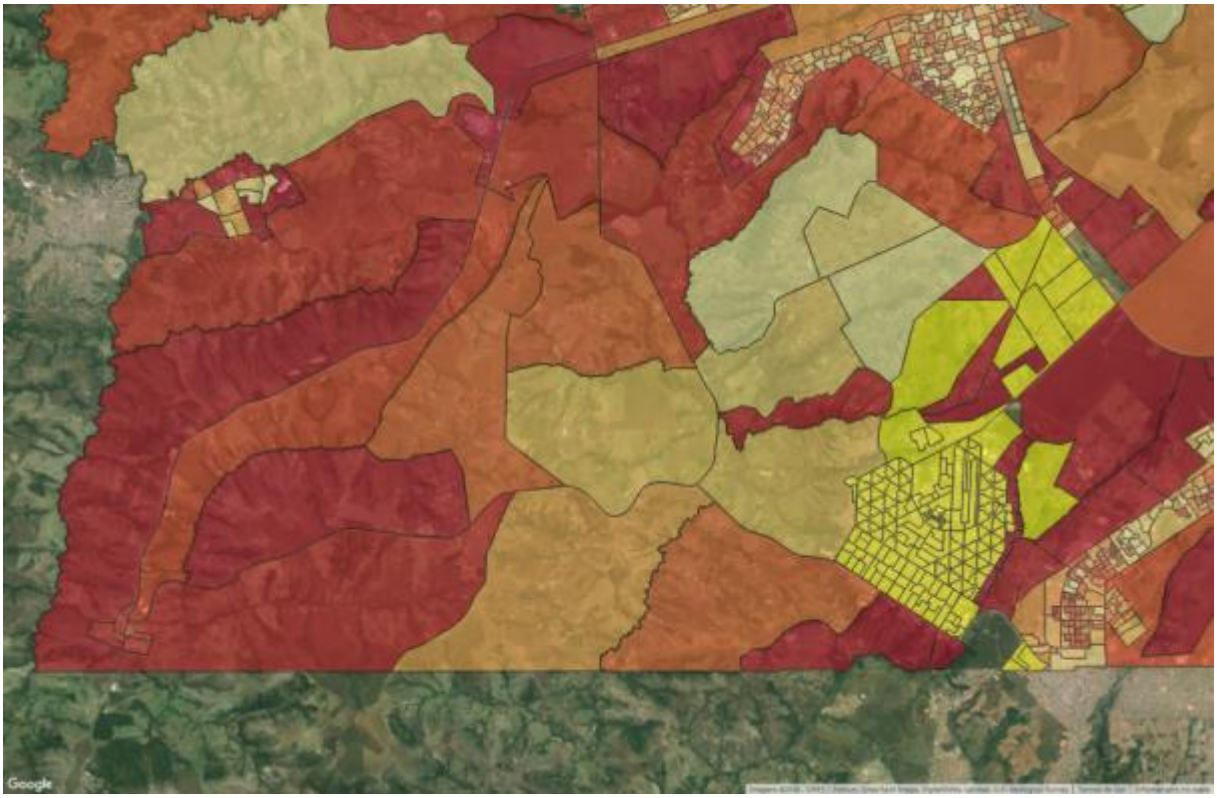


Mapa 9 – Subdistrito Ceilândia: Análise de demanda

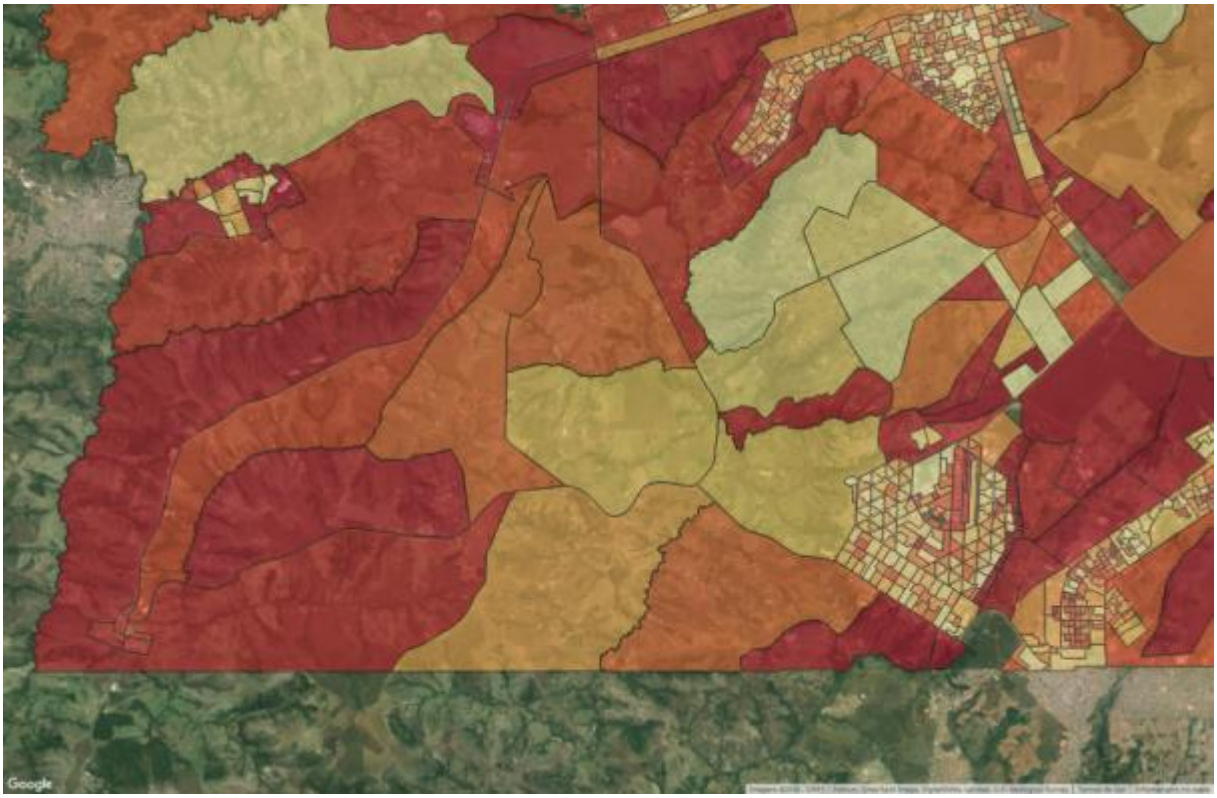


Mapa 10 – Subdistrito Cruzeiro: Indicação urbana

Mapa 11 – Subdistrito Cruzeiro: Análise de demanda



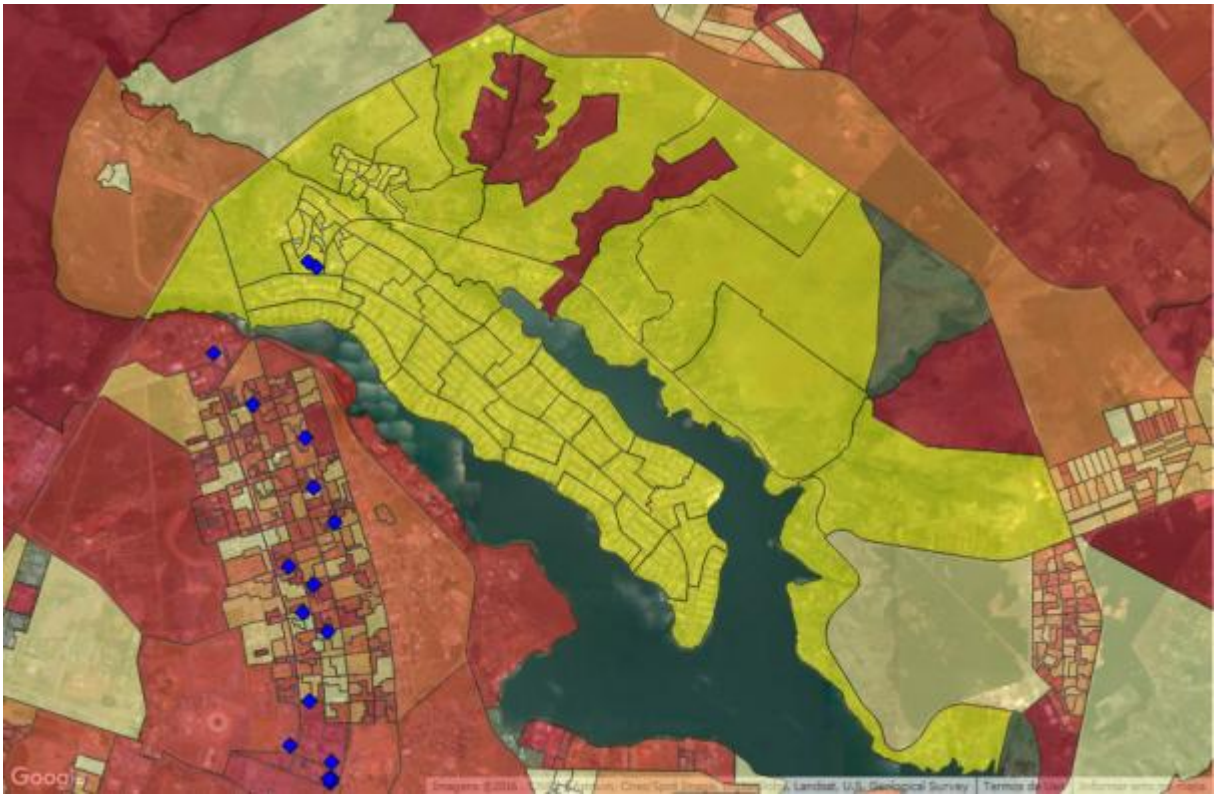
Mapa 12 – Subdistrito Gama: Indicação urbana



Mapa 13 – Subdistrito Gama: Análise de demanda

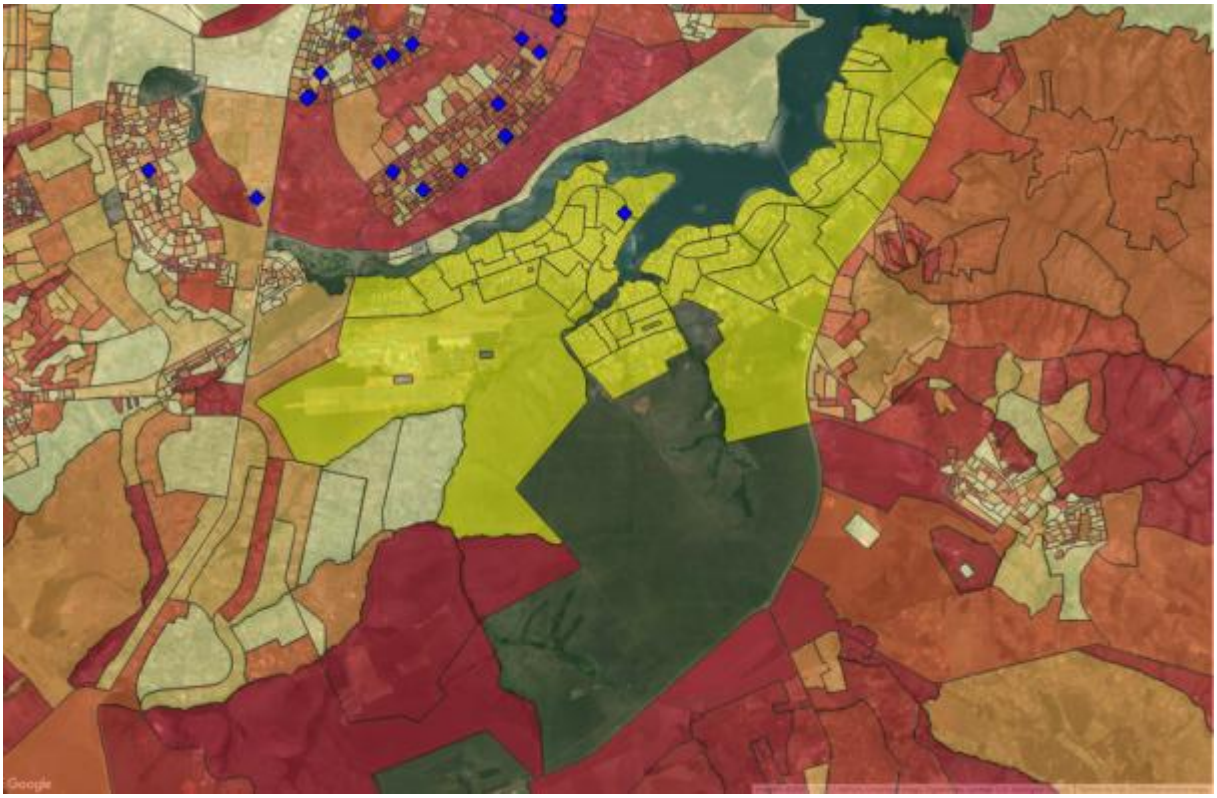
Mapa 14 – Subdistrito Guará: Indicação urbana

Mapa 15 – Subdistrito Guará: Análise de demanda



Mapa 16 – Subdistrito Lago Norte: Indicação urbana

Mapa 17 – Subdistrito Lago Norte: Análise de demanda

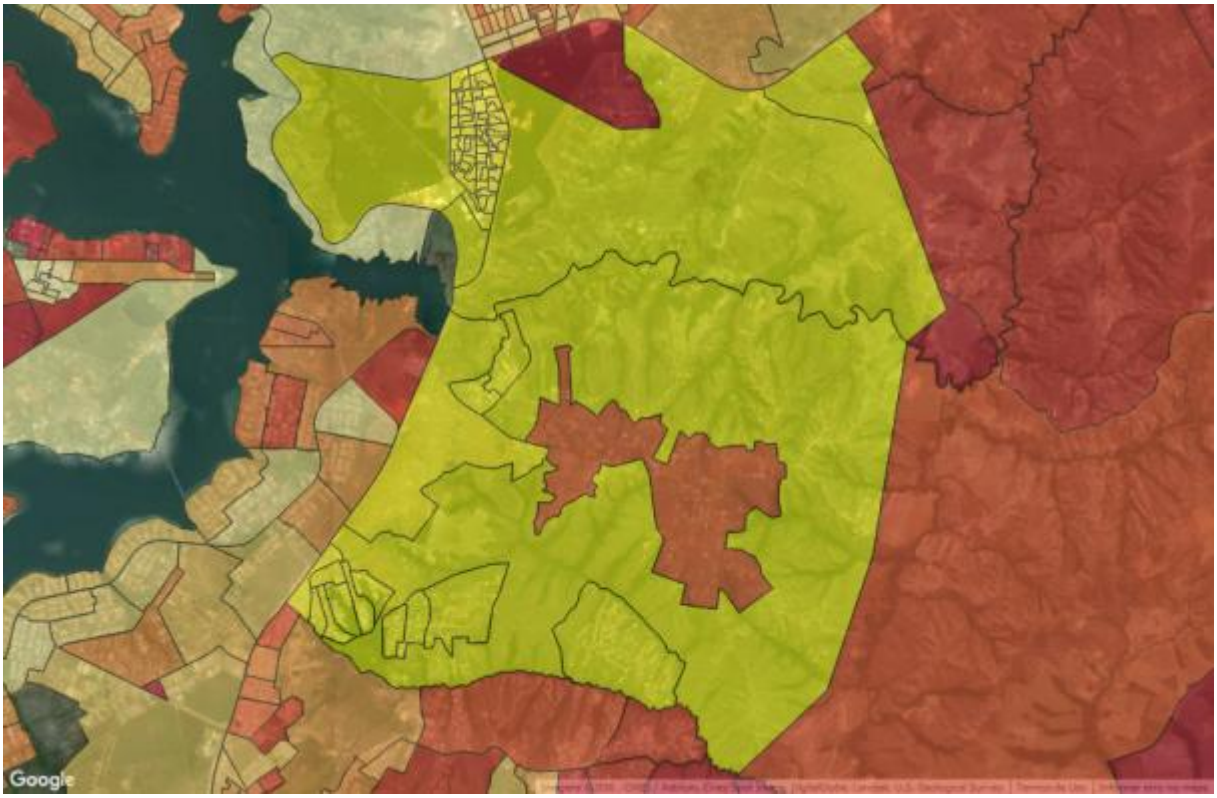


Mapa 18 – Subdistrito Lago Sul: Indicação urbana

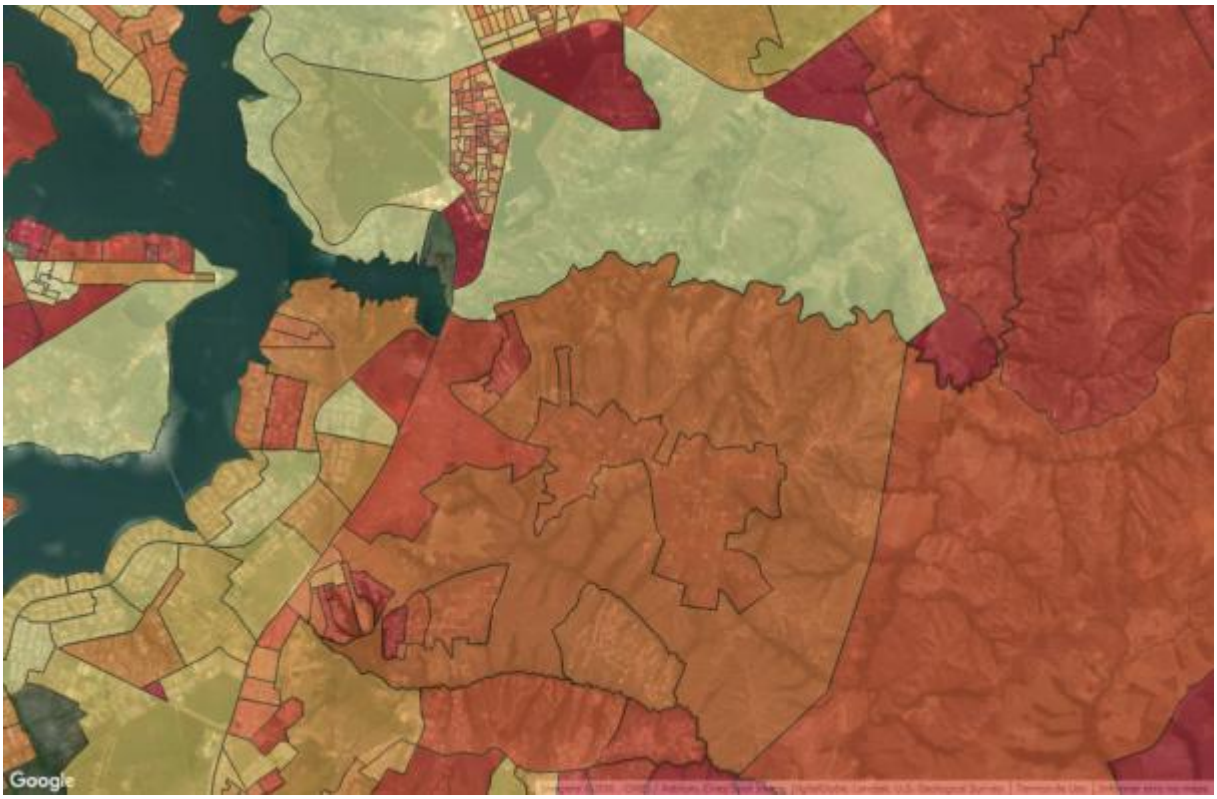
Mapa 19 – Subdistrito Lago Sul: Análise de demanda

Mapa 20 – Subdistrito Núcleo Bandeirante: Indicação urbana

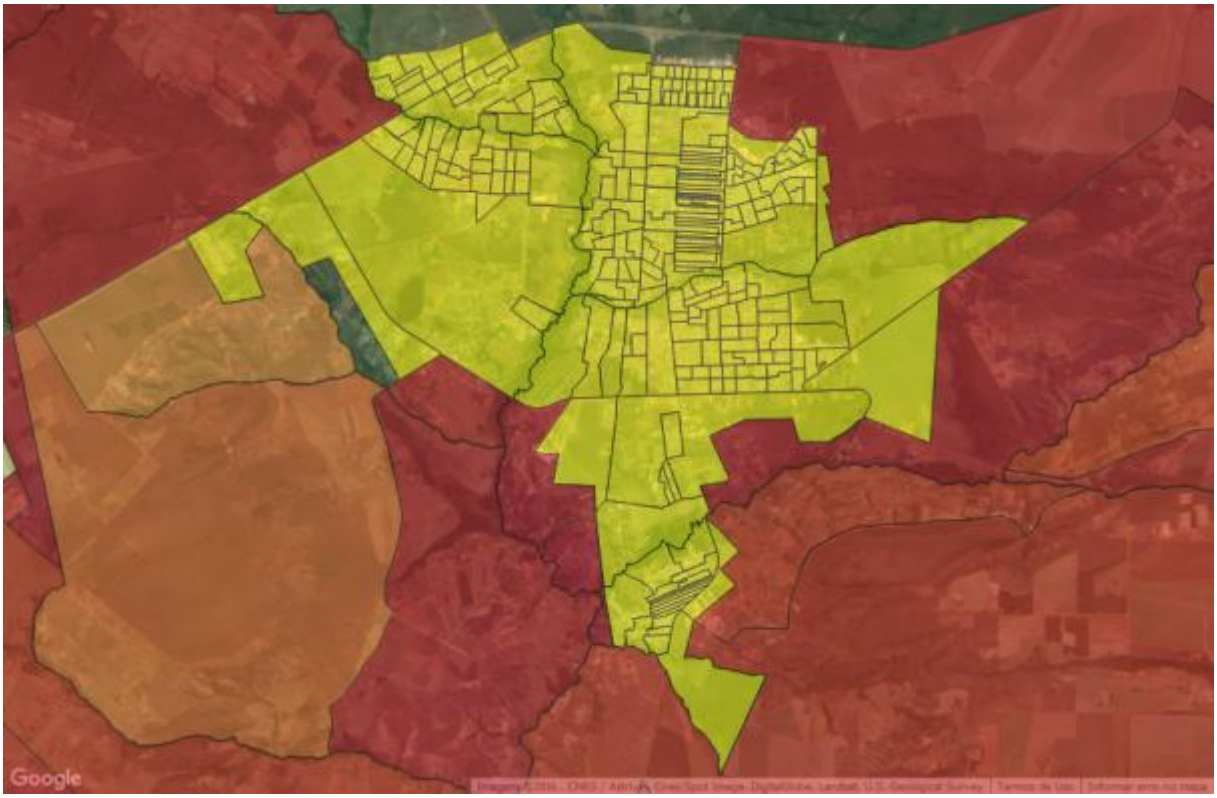
Mapa 21 – Subdistrito Núcleo Bandeirante: Análise de demanda



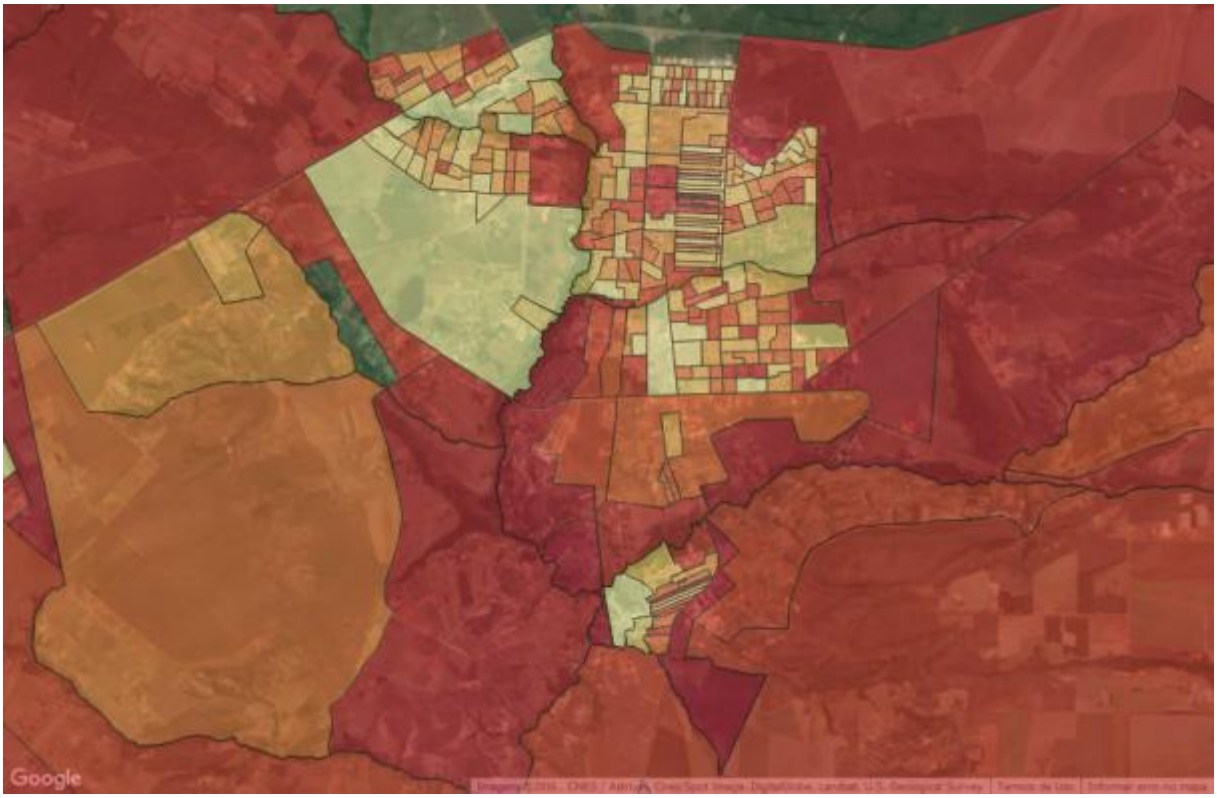
Mapa 22 – Subdistrito Paranoá: Indicativo urbano



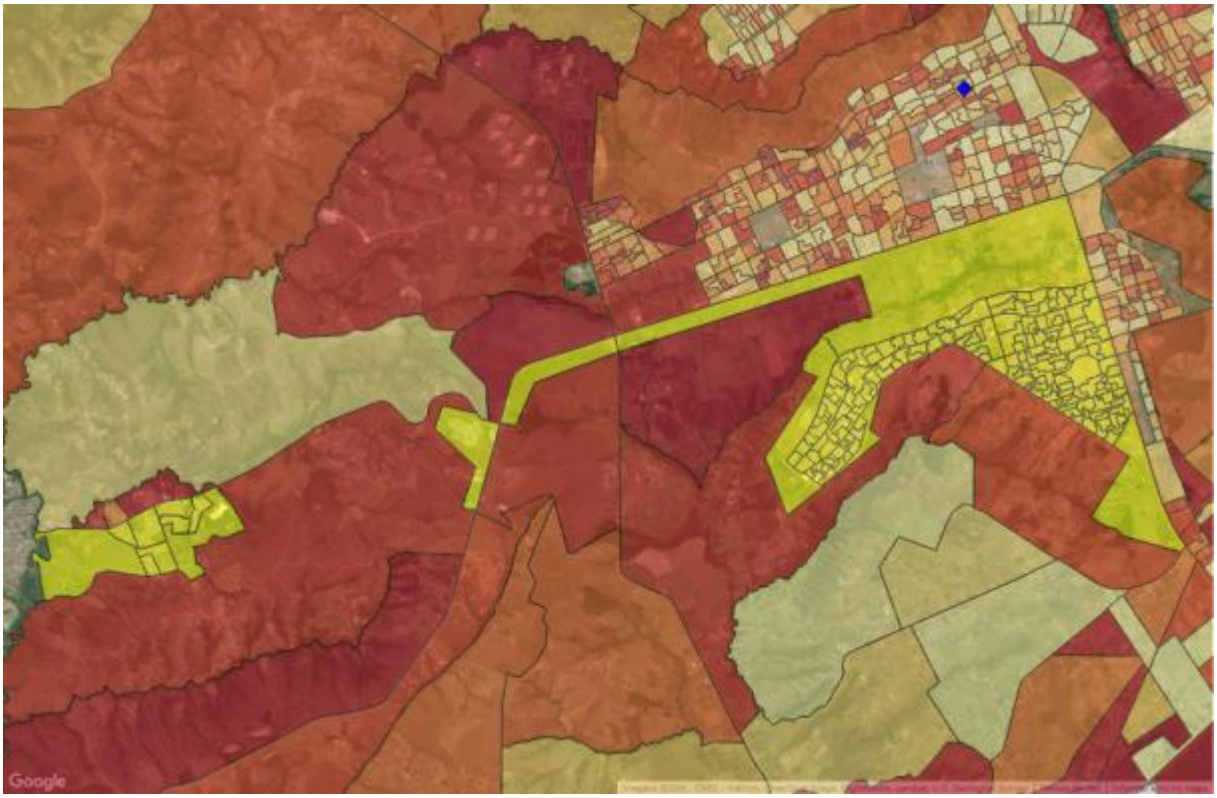
Mapa 23 – Subdistrito Paranoá: Análise de demanda



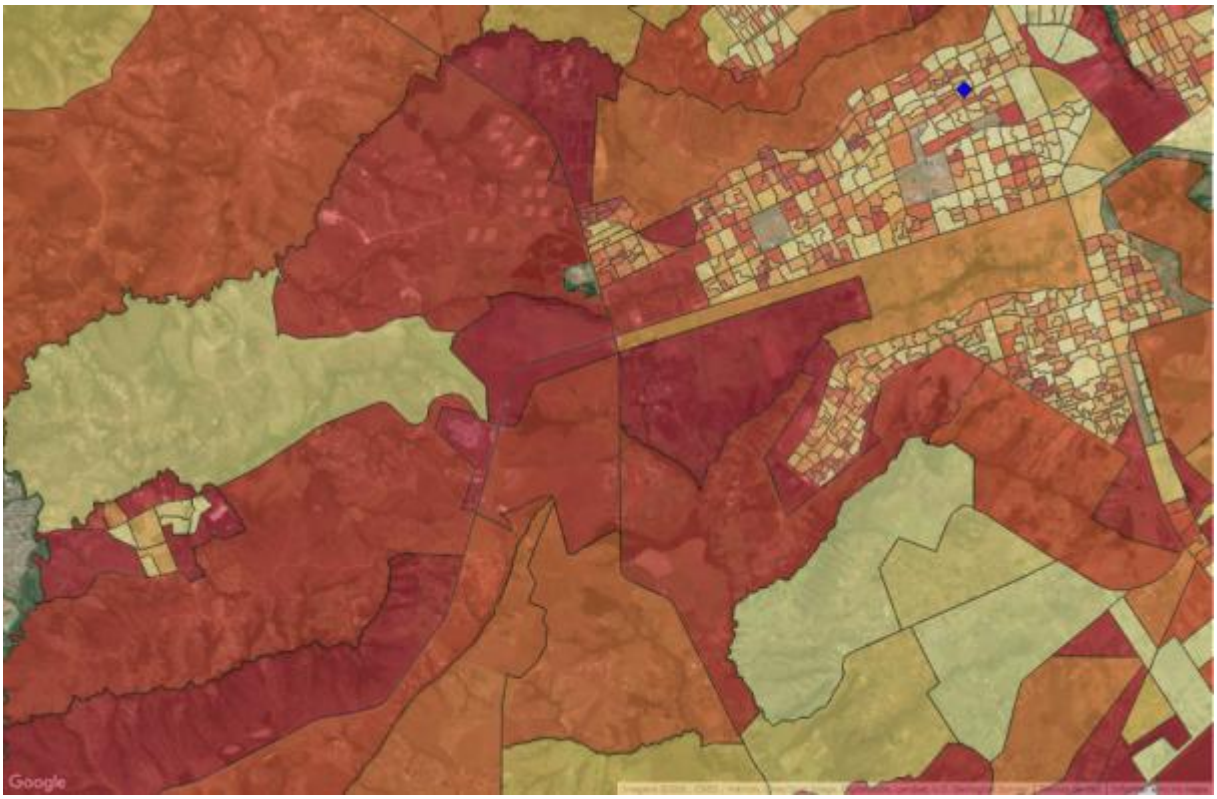
Mapa 24 – Subdistrito Planaltina: Indicativo urbano



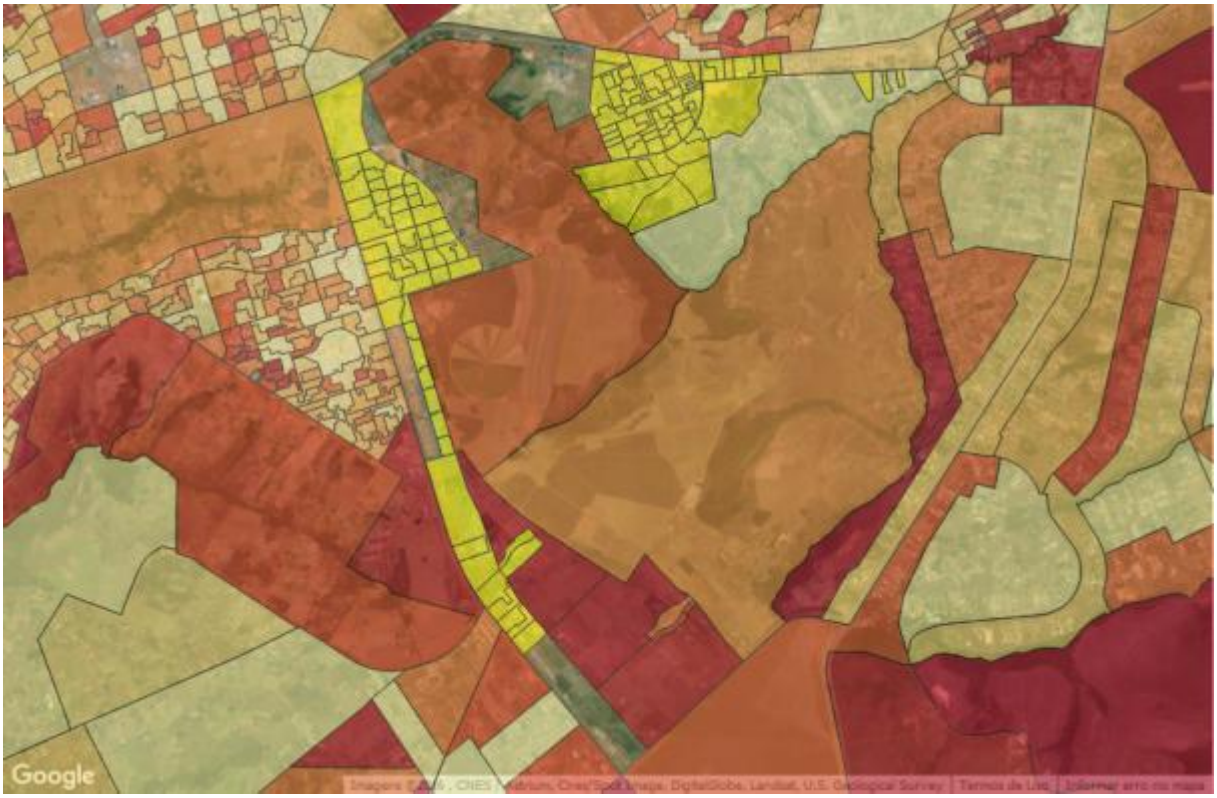
Mapa 25 – Subdistrito Planaltina: Análise de demanda



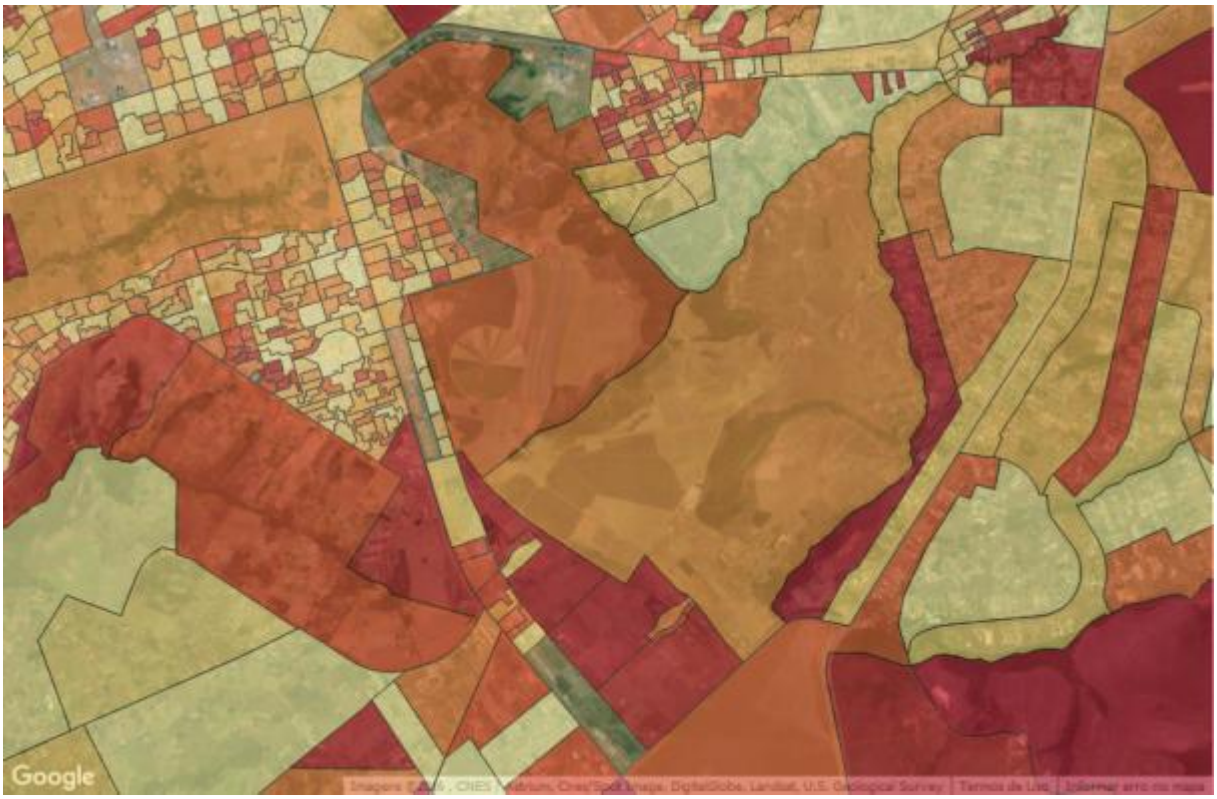
Mapa 26 – Subdistrito Recanto das Emas: Indicativo urbano



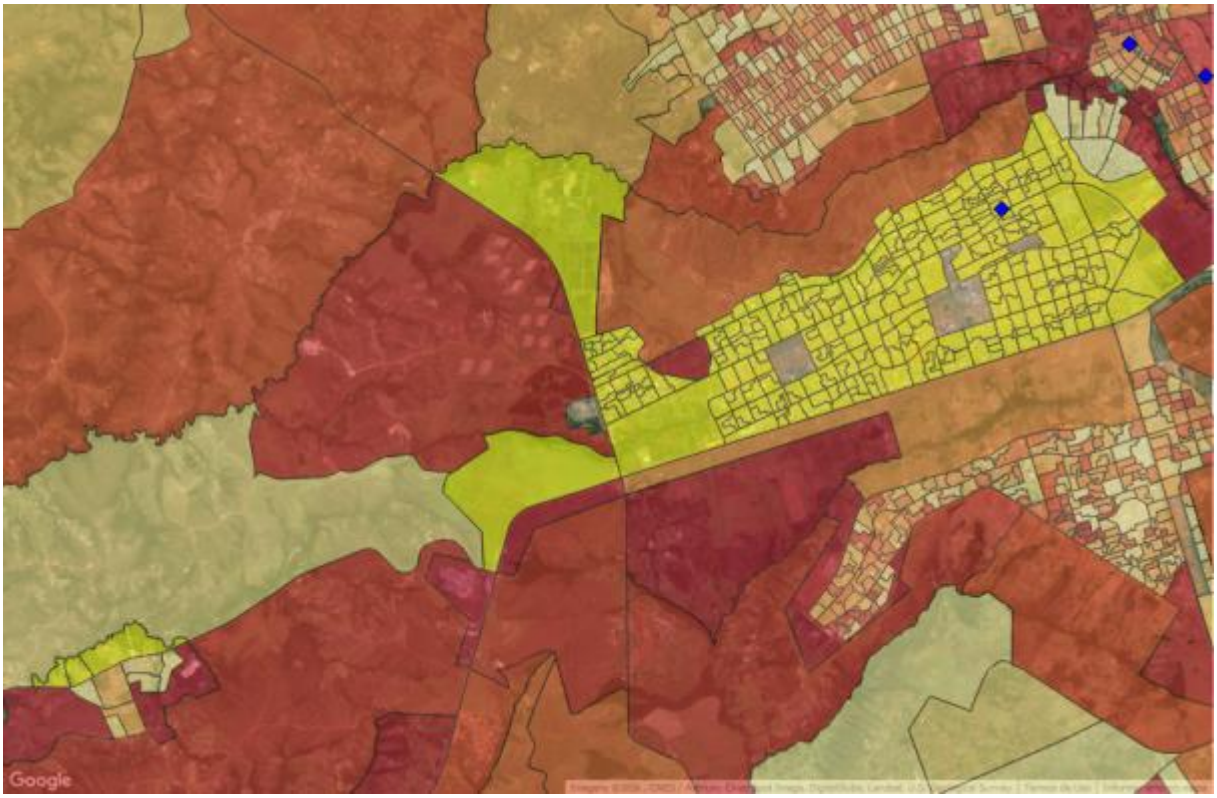
Mapa 27 – Subdistrito Recanto das Emas: Análise de demanda



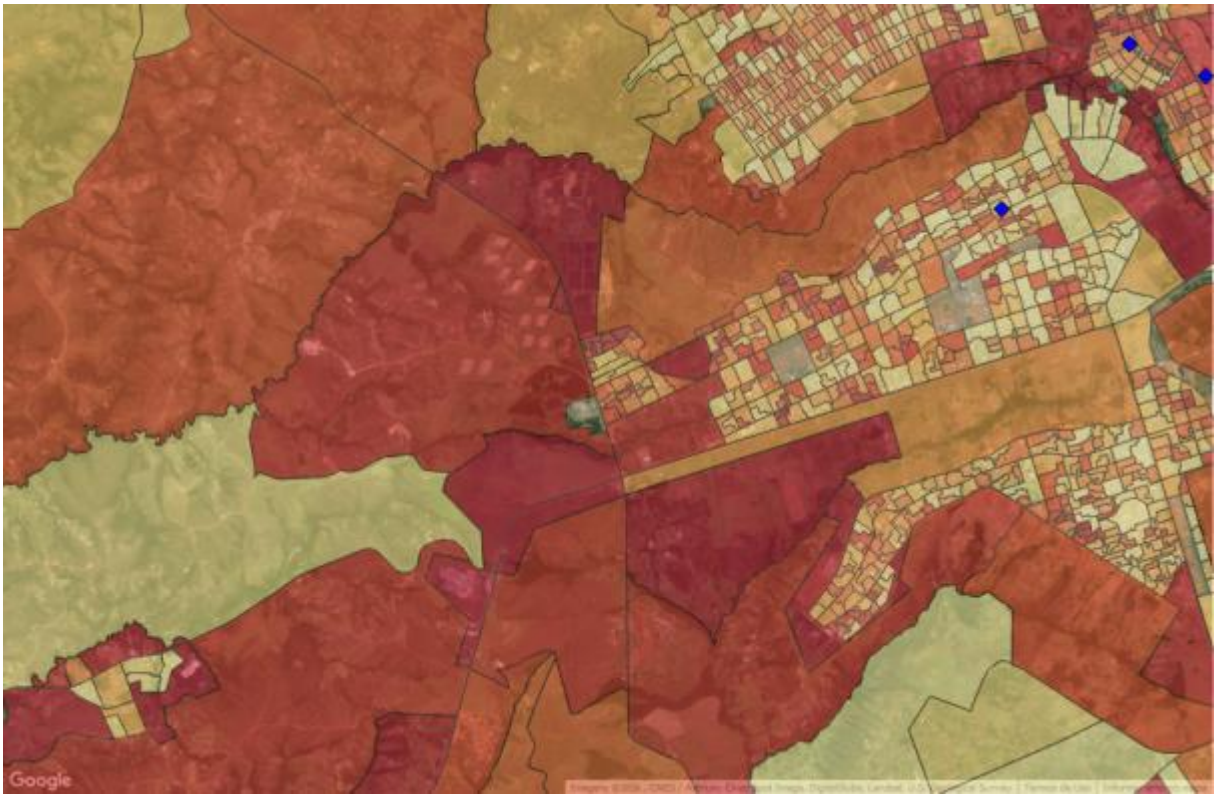
Mapa 28 – Subdistrito Riacho Fundo: Indicação urbana



Mapa 29 – Subdistrito Riacho Fundo: Indicação urbana



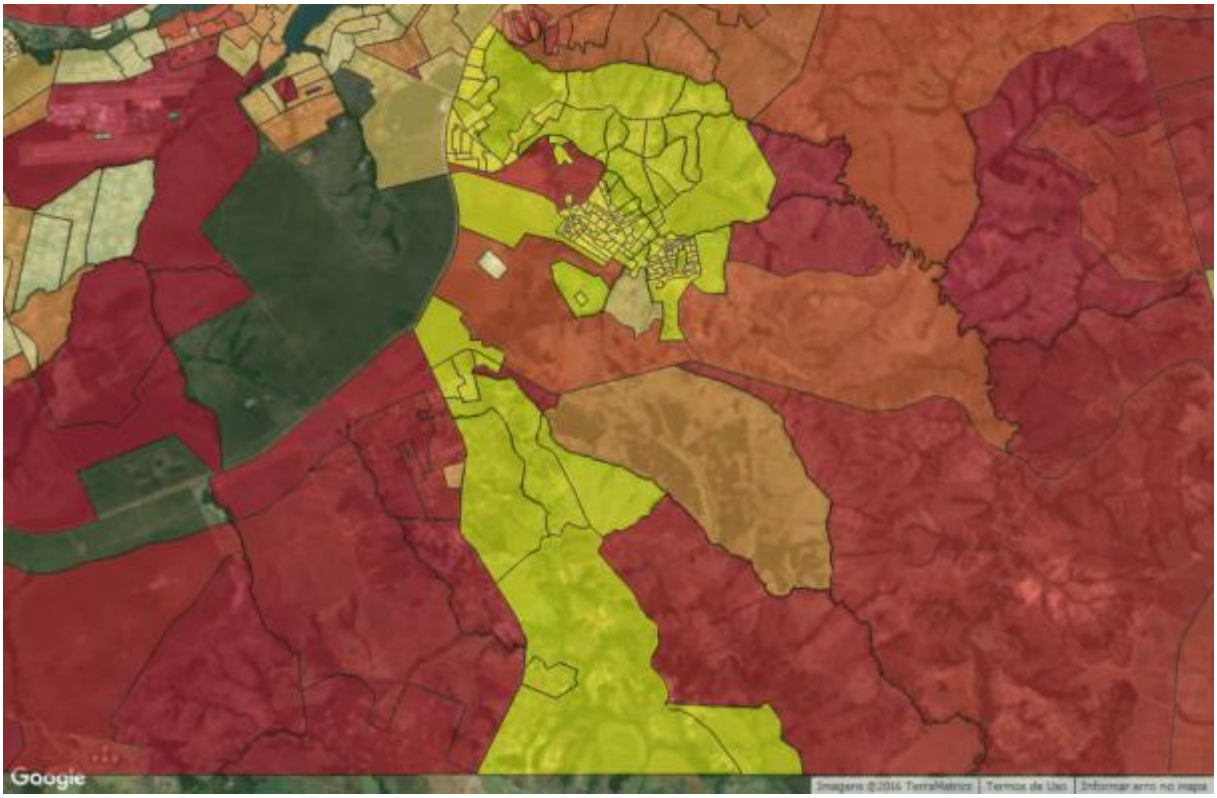
Mapa 30 – Subdistrito Samambaia: Indicação urbana



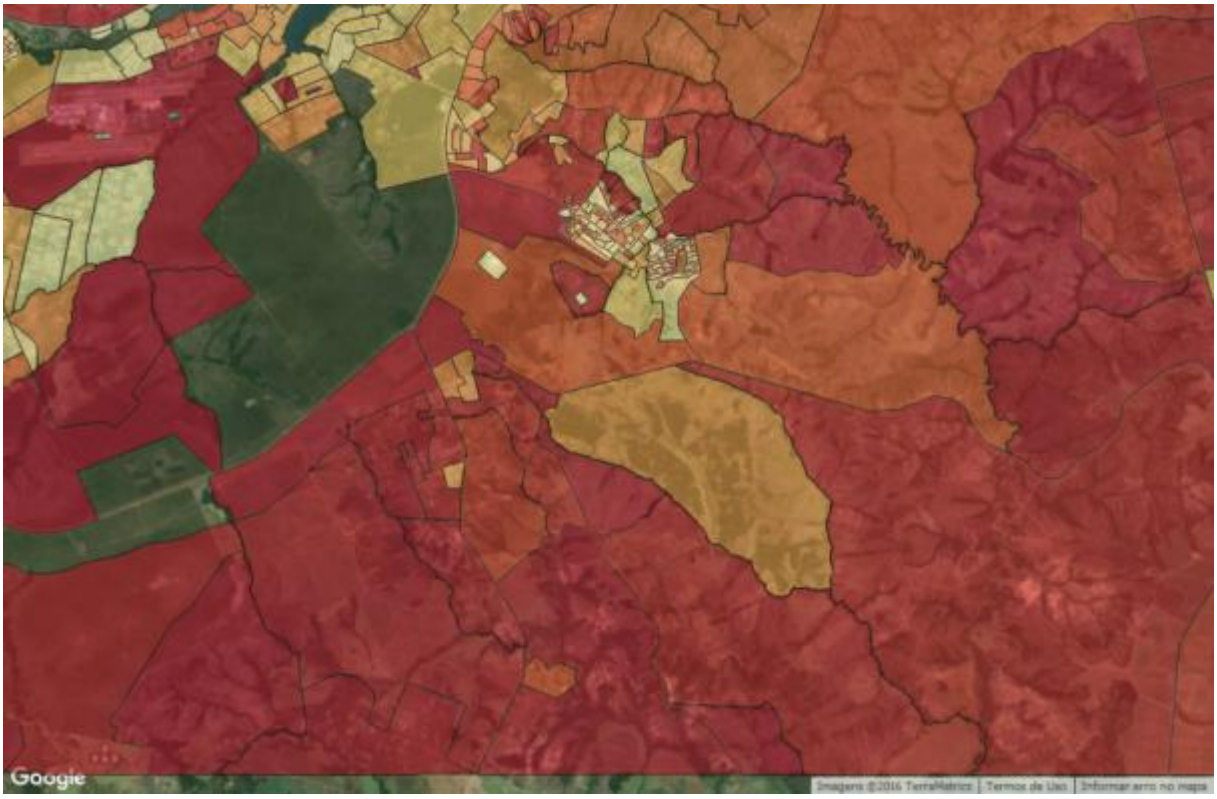
Mapa 31 – Subdistrito Samambaia: Análise de demanda

Mapa 32 – Subdistrito Santa Maria: Indicação urbana

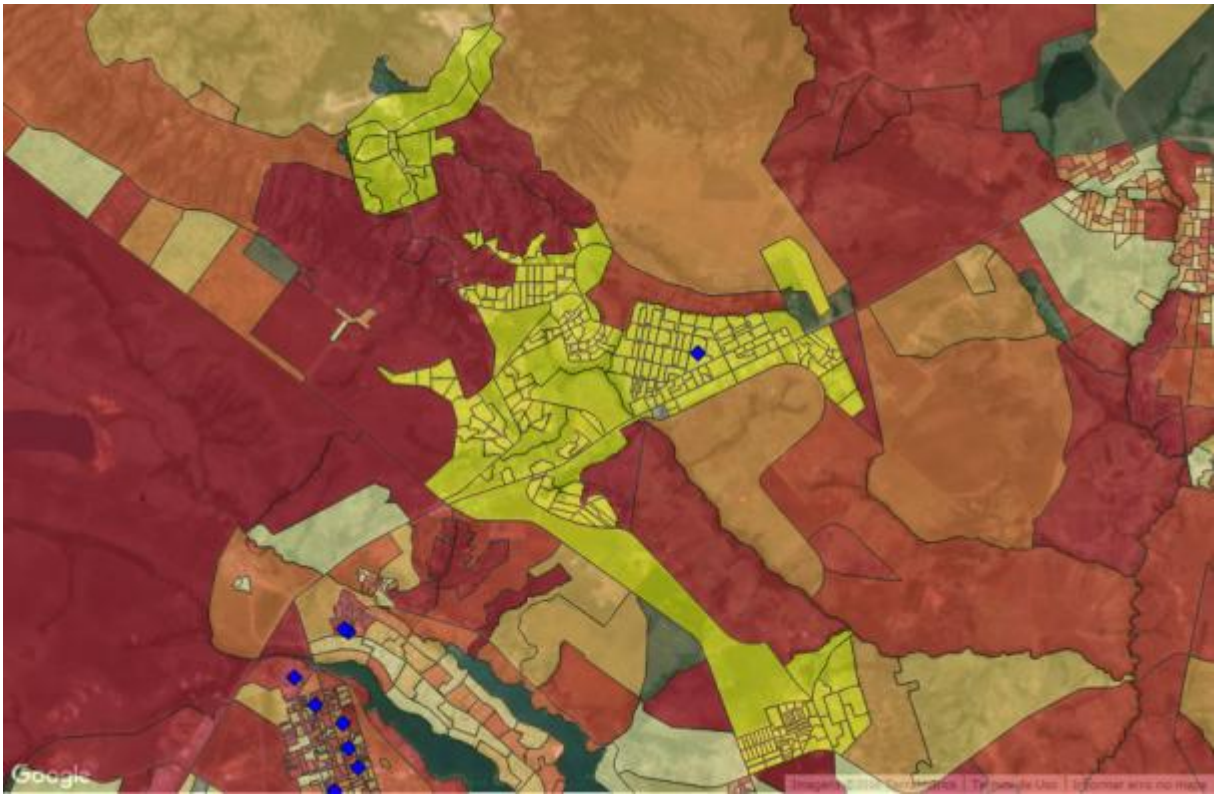
Mapa 33 – Subdistrito Santa Maria: Análise de demanda



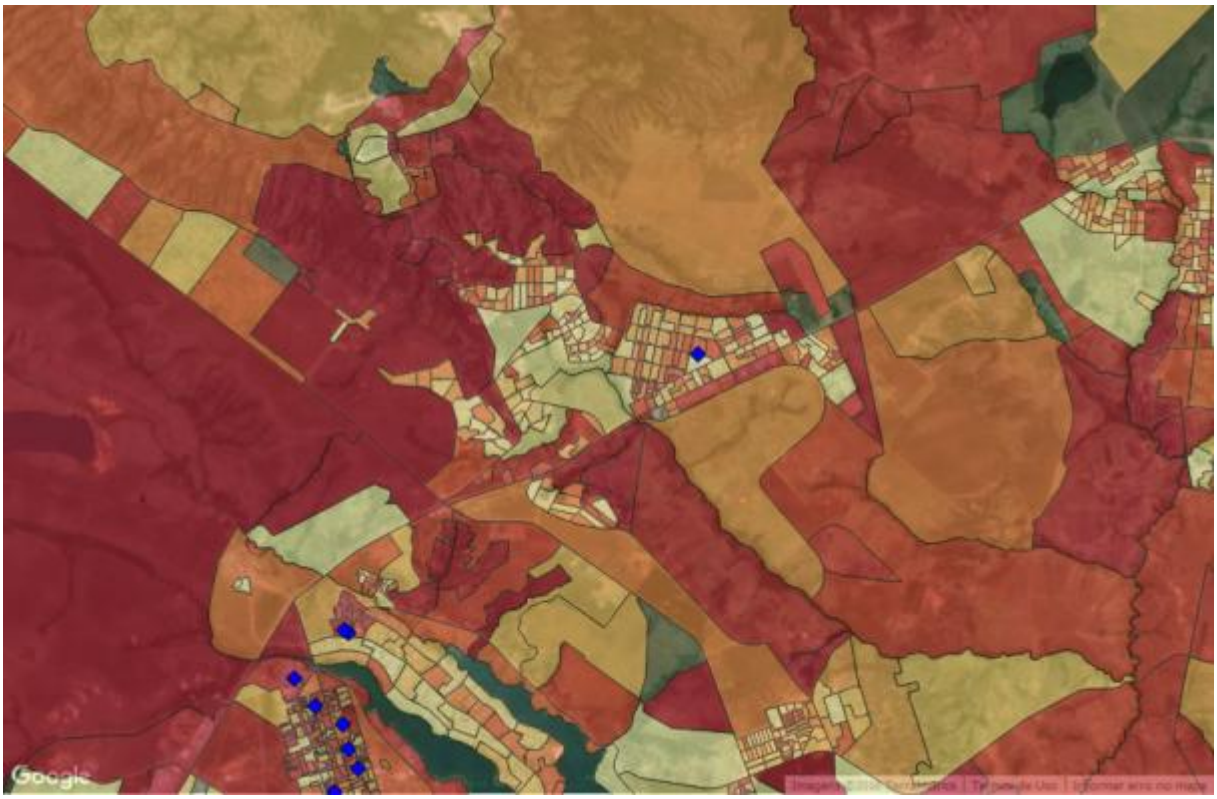
Mapa 34 – Subdistrito São Sebastião: Indicação urbana



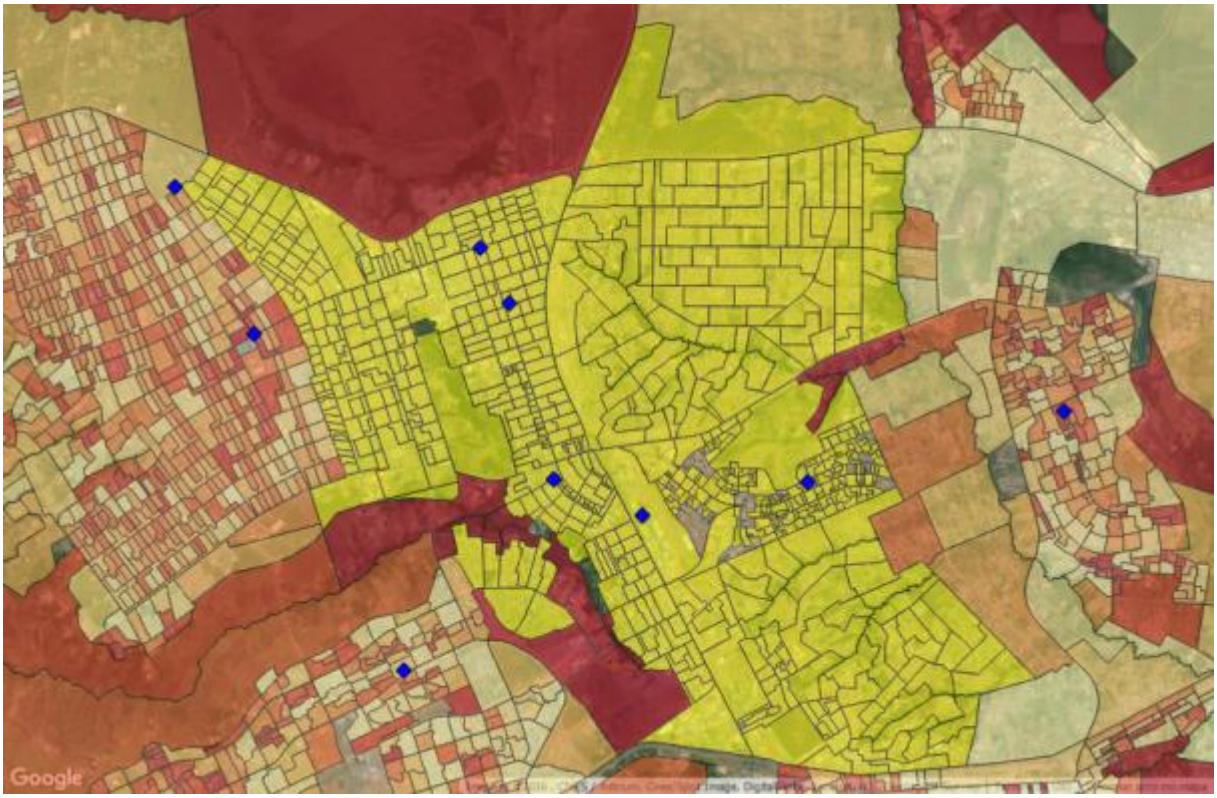
Mapa 35 – Subdistrito São Sebastião: Análise de demanda



Mapa 36 – Subdistrito Sobradinho: Indicação urbana



Mapa 37 – Subdistrito Sobradinho: Análise de demanda



Mapa 38 – Subdistrito Taguatinga: Indicação urbana

Mapa 39 – Subdistrito Taguatinga: Análise de demanda

Apêndice B – Programação no RStudio

```
## Limpa WorkSpace
rm(list=ls())
setwd("C:\\Users\\felix\\Documents\\Marcelo\\TCC\\Censo
2010\\Estatística\\DF_20150527\\DF\\Base            informaçoes
setores2010 universo DF\\CSV")
load("RESULTADOS.15062016.RData")
## Define WorkSpace
setwd("C:\\Users\\felix\\Documents\\Marcelo\\TCC\\WorkingDirec
tory")

### Lendo a base de dados T_MORADOR_S.txt

## Invoca a funcao LeBasesPosicaoFixa.R (disponivel na pagina
do Prof. Elias da UFPR)
source("LeBasesPosicaoFixa.R.txt")


## Cria um novo arquivo (MORADOR.txt), selecionando apenas as
variaveis desejadas.
DMORADOR<-read.table("T_MORADOR_S.txt")
fselpr<-function(x)x
rcsel.pfix(file.inp="T_MORADOR_S.txt",file.out="MORADOR.txt",
           first=c(3,5,8,9,11,16,30,60,76,197),
           last=c(4,7,8,10,11,29,43,62,77,212),
           fselpr)

## Le os dados do arquivo de interesse
DadosMorador<-read.table("MORADOR.txt")

## Deleta o arquivo MORADOR.txt
file.remove("MORADOR.txt")

## Coloca os nomes das variaveis
```

```
colnames(DadosMorador)<-
c("COD_UF","NUM_SEQ","DV_SEQ","NUM_DOM","NUM_UC",
                          "FAT_EXP1","FAT_EXP2","IDADE_ANOS",
                          "SEXO","REND_PERC")


## Coloca os labels nas variaveis
library(Hmisc)
label(DadosMorador$COD_UF)<-'CODIGO DA UF'
label(DadosMorador$NUM_SEQ)<-'NUMERO SEQUENCIAL'
label(DadosMorador$DV_SEQ)<-'DV DO SEQUENCIAL'
label(DadosMorador$NUM_DOM)<-'NUMERO DO DOMICILIO'
label(DadosMorador$NUM_UC)<-'NUMERO DA UC'
label(DadosMorador$FAT_EXP1)<-'FATOR DE EXPANSAO 1'
label(DadosMorador$FAT_EXP2)<-'FATOR DE EXPANSAO 2'
label(DadosMorador$IDADE_ANOS)<-'IDADE CALCULADA EM ANOS'
label(DadosMorador$SEXO)<-'SEXO'
label(DadosMorador$REND_PERC)<-'RENDA PER CAPTA DA UC'

### Lendo base de dados T_CADERNET_DESPESA_S
DCADERNETA<-read.table("T_CADERNETA_DESPESA_S.txt")

## Cria base de dados (CADERNETA) apenas com as variaveis
desejadas
rcsel.pfix(file.inp       ="T_CADERNETA_DESPESA_S.txt",file.out
="C_DESPESA.txt",
           first=c(3,5,8,9,11,14,28,42,44,46,53),
           last=c(4,7,8,10,11,27,41,43,45,50,63),
           fselpr)

CADERNETA<-read.table("C_DESPESA.txt")
## Exclui o arquivo criado
file.remove("C_DESPESA.txt")
```

```
## Da nomes para as variaveis
colnames(CADERNETA)<-
c("COD_UF","NUM_SEQ","DV_SEQ","NUM_DOM","NUM_UC","FAT_EXP1",

"FAT_EXP2","NUM_QUA","NUM_GRU","COD_ITEM","VAL_DESP")

## Coloca os labels nas variaveis
label(CADERNETA$COD_UF)<-'CODIGO DA UF'
label(CADERNETA$NUM_SEQ)<-'NUMERO SEQUENCIAL'
label(CADERNETA$DV_SEQ)<-'DV DO SEQUENCIAL'
label(CADERNETA$NUM_DOM)<-'NUMERO DO DOMICILIO'
label(CADERNETA$NUM_UC)<-'NUMERO DA UC'
label(CADERNETA$FAT_EXP1)<-'FATOR DE EXPANSAO 1'
label(CADERNETA$FAT_EXP2)<-'FATOR DE EXPANSAO 2'
label(CADERNETA$NUM_QUA)<-'NUMERO DO QUADRO'
label(CADERNETA$NUM_GRU)<-'NUMERO DO GRUPO'
label(CADERNETA$COD_ITEM)<-'CODIGO DO ITEM'
label(CADERNETA$VAL_DESP)<-'VALOR DA DESPESA'


### Seleciona somente quem consumiu os produtos da cesta
CADERNETA<-CADERNETA[which(CADERNETA$NUM_GRU==69                &
(CADERNETA$COD_ITEM%in%c(00701,00702,00703,00704,
00705,00707,00708,00709,
00901,00902,00903,00904,
00905,05601,05605,05801,
05802,05803,05804,10001,
10002,10003,10004,10005,
10101,10102,10103,10104,
10105,10107,10108,10109,
10110,10111,10401,10402,
10403,10404,10405,10501,
10502,10503,10504,10505,
10506,10507,10508,12101,
```

```
12201,12204,12301,12302,
12303,12304,12401,12402,
12403,12404,12405,12406,
12407,12408))),]

### Junta as bases T_MORADOR_S e T_CADERNETA_DESPESA_S
pof<-
merge(DadosMorador,CADERNETA,by=c("COD_UF","NUM_SEQ","DV_SEQ",
"NUM_DOM",
"NUM_UC"),all=TRUE)
### Quem nao consumiu recebe 0
pof$VAL_DESP[is.na(pof$VAL_DESP)]<-0

### Criando as variaveis de acordo com o censo
## Homens
pof$V073<-ifelse(pof$IDADE_ANOS==0&pof$SEXO==1,1,0)
pof$V074<-ifelse(pof$IDADE_ANOS==1&pof$SEXO==1,1,0)
pof$V075<-ifelse(pof$IDADE_ANOS==2&pof$SEXO==1,1,0)
pof$V076<-ifelse(pof$IDADE_ANOS==3&pof$SEXO==1,1,0)
pof$V077<-ifelse(pof$IDADE_ANOS==4&pof$SEXO==1,1,0)
pof$V078<-ifelse(pof$IDADE_ANOS==5&pof$SEXO==1,1,0)
pof$V079<-ifelse(pof$IDADE_ANOS==6&pof$SEXO==1,1,0)
pof$V080<-ifelse(pof$IDADE_ANOS==7&pof$SEXO==1,1,0)
pof$V081<-ifelse(pof$IDADE_ANOS==8&pof$SEXO==1,1,0)
pof$V082<-ifelse(pof$IDADE_ANOS==9&pof$SEXO==1,1,0)
pof$V083<-ifelse(pof$IDADE_ANOS==10&pof$SEXO==1,1,0)
pof$V084<-ifelse(pof$IDADE_ANOS==11&pof$SEXO==1,1,0)
pof$V085<-ifelse(pof$IDADE_ANOS==12&pof$SEXO==1,1,0)
pof$V086<-ifelse(pof$IDADE_ANOS==13&pof$SEXO==1,1,0)
pof$V087<-ifelse(pof$IDADE_ANOS==14&pof$SEXO==1,1,0)
pof$V088<-ifelse(pof$IDADE_ANOS==15&pof$SEXO==1,1,0)
pof$V089<-ifelse(pof$IDADE_ANOS==16&pof$SEXO==1,1,0)
pof$V090<-ifelse(pof$IDADE_ANOS==17&pof$SEXO==1,1,0)
pof$V091<-ifelse(pof$IDADE_ANOS==18&pof$SEXO==1,1,0)
```

```
pof$V092<-ifelse(pof$IDADE_ANOS==19&pof$SEXO==1,1,0)
pof$V093<-ifelse(pof$IDADE_ANOS==20&pof$SEXO==1,1,0)
pof$V094<-ifelse(pof$IDADE_ANOS==21&pof$SEXO==1,1,0)
pof$V095<-ifelse(pof$IDADE_ANOS==22&pof$SEXO==1,1,0)
pof$V096<-ifelse(pof$IDADE_ANOS==23&pof$SEXO==1,1,0)
pof$V097<-ifelse(pof$IDADE_ANOS==24&pof$SEXO==1,1,0)
pof$V098<-
ifelse(pof$IDADE_ANOS>=25&pof$IDADE_ANOS<=29&pof$SEXO==1,1,0)
pof$V099<-
ifelse(pof$IDADE_ANOS>=30&pof$IDADE_ANOS<=34&pof$SEXO==1,1,0)
pof$V100<-
ifelse(pof$IDADE_ANOS>=35&pof$IDADE_ANOS<=39&pof$SEXO==1,1,0)
pof$V101<-
ifelse(pof$IDADE_ANOS>=40&pof$IDADE_ANOS<=44&pof$SEXO==1,1,0)
pof$V102<-
ifelse(pof$IDADE_ANOS>=45&pof$IDADE_ANOS<=49&pof$SEXO==1,1,0)
pof$V103<-
ifelse(pof$IDADE_ANOS>=50&pof$IDADE_ANOS<=54&pof$SEXO==1,1,0)
pof$V104<-
ifelse(pof$IDADE_ANOS>=55&pof$IDADE_ANOS<=59&pof$SEXO==1,1,0)
pof$V105<-
ifelse(pof$IDADE_ANOS>=60&pof$IDADE_ANOS<=64&pof$SEXO==1,1,0)
pof$V106<-
ifelse(pof$IDADE_ANOS>=65&pof$IDADE_ANOS<=69&pof$SEXO==1,1,0)
pof$V107<-
ifelse(pof$IDADE_ANOS>=70&pof$IDADE_ANOS<=74&pof$SEXO==1,1,0)
pof$V108<-
ifelse(pof$IDADE_ANOS>=75&pof$IDADE_ANOS<=79&pof$SEXO==1,1,0)
pof$V109<-
ifelse(pof$IDADE_ANOS>=80&pof$IDADE_ANOS<=84&pof$SEXO==1,1,0)
pof$V110<-
ifelse(pof$IDADE_ANOS>=85&pof$IDADE_ANOS<=89&pof$SEXO==1,1,0)
pof$V111<-
ifelse(pof$IDADE_ANOS>=90&pof$IDADE_ANOS<=94&pof$SEXO==1,1,0)
```

```
pof$V112<-
ifelse(pof$IDADE_ANOS>=95&pof$IDADE_ANOS<=99&pof$SEXO==1,1,0)
pof$V113<-ifelse(pof$IDADE_ANOS>=100&pof$SEXO==1,1,0)
## Mulheres
pof$V114<-ifelse(pof$IDADE_ANOS==0&pof$SEXO==2,1,0)
pof$V115<-ifelse(pof$IDADE_ANOS==1&pof$SEXO==2,1,0)
pof$V116<-ifelse(pof$IDADE_ANOS==2&pof$SEXO==2,1,0)
pof$V117<-ifelse(pof$IDADE_ANOS==3&pof$SEXO==2,1,0)
pof$V118<-ifelse(pof$IDADE_ANOS==4&pof$SEXO==2,1,0)
pof$V119<-ifelse(pof$IDADE_ANOS==5&pof$SEXO==2,1,0)
pof$V120<-ifelse(pof$IDADE_ANOS==6&pof$SEXO==2,1,0)
pof$V121<-ifelse(pof$IDADE_ANOS==7&pof$SEXO==2,1,0)
pof$V122<-ifelse(pof$IDADE_ANOS==8&pof$SEXO==2,1,0)
pof$V123<-ifelse(pof$IDADE_ANOS==9&pof$SEXO==2,1,0)
pof$V124<-ifelse(pof$IDADE_ANOS==10&pof$SEXO==2,1,0)
pof$V125<-ifelse(pof$IDADE_ANOS==11&pof$SEXO==2,1,0)
pof$V126<-ifelse(pof$IDADE_ANOS==12&pof$SEXO==2,1,0)
pof$V127<-ifelse(pof$IDADE_ANOS==13&pof$SEXO==2,1,0)
pof$V128<-ifelse(pof$IDADE_ANOS==14&pof$SEXO==2,1,0)
pof$V129<-ifelse(pof$IDADE_ANOS==15&pof$SEXO==2,1,0)
pof$V130<-ifelse(pof$IDADE_ANOS==16&pof$SEXO==2,1,0)
pof$V131<-ifelse(pof$IDADE_ANOS==17&pof$SEXO==2,1,0)
pof$V132<-ifelse(pof$IDADE_ANOS==18&pof$SEXO==2,1,0)
pof$V133<-ifelse(pof$IDADE_ANOS==19&pof$SEXO==2,1,0)
pof$V134<-ifelse(pof$IDADE_ANOS==20&pof$SEXO==2,1,0)
pof$V135<-ifelse(pof$IDADE_ANOS==21&pof$SEXO==2,1,0)
pof$V136<-ifelse(pof$IDADE_ANOS==22&pof$SEXO==2,1,0)
pof$V137<-ifelse(pof$IDADE_ANOS==23&pof$SEXO==2,1,0)
pof$V138<-ifelse(pof$IDADE_ANOS==24&pof$SEXO==2,1,0)
pof$V139<-
ifelse(pof$IDADE_ANOS>=25&pof$IDADE_ANOS<=29&pof$SEXO==2,1,0)
pof$V140<-
ifelse(pof$IDADE_ANOS>=30&pof$IDADE_ANOS<=34&pof$SEXO==2,1,0)
```

```
pof$V141<-
ifelse(pof$IDADE_ANOS>=35&pof$IDADE_ANOS<=39&pof$SEXO==2,1,0)
pof$V142<-
ifelse(pof$IDADE_ANOS>=40&pof$IDADE_ANOS<=44&pof$SEXO==2,1,0)
pof$V143<-
ifelse(pof$IDADE_ANOS>=45&pof$IDADE_ANOS<=49&pof$SEXO==2,1,0)
pof$V144<-
ifelse(pof$IDADE_ANOS>=50&pof$IDADE_ANOS<=54&pof$SEXO==2,1,0)
pof$V145<-
ifelse(pof$IDADE_ANOS>=55&pof$IDADE_ANOS<=59&pof$SEXO==2,1,0)
pof$V146<-
ifelse(pof$IDADE_ANOS>=60&pof$IDADE_ANOS<=64&pof$SEXO==2,1,0)
pof$V147<-
ifelse(pof$IDADE_ANOS>=65&pof$IDADE_ANOS<=69&pof$SEXO==2,1,0)
pof$V148<-
ifelse(pof$IDADE_ANOS>=70&pof$IDADE_ANOS<=74&pof$SEXO==2,1,0)
pof$V149<-
ifelse(pof$IDADE_ANOS>=75&pof$IDADE_ANOS<=79&pof$SEXO==2,1,0)
pof$V150<-
ifelse(pof$IDADE_ANOS>=80&pof$IDADE_ANOS<=84&pof$SEXO==2,1,0)
pof$V151<-
ifelse(pof$IDADE_ANOS>=85&pof$IDADE_ANOS<=89&pof$SEXO==2,1,0)
pof$V152<-
ifelse(pof$IDADE_ANOS>=90&pof$IDADE_ANOS<=94&pof$SEXO==2,1,0)
pof$V153<-
ifelse(pof$IDADE_ANOS>=95&pof$IDADE_ANOS<=99&pof$SEXO==2,1,0)
pof$V154<-ifelse(pof$IDADE_ANOS>=100&pof$SEXO==2,1,0)
## Renda per capta
pof$RENDA1<-
ifelse(pof$REND_PERC>0&pof$REND_PERC<=(1/8)*510,1,0)
pof$RENDA2<-
ifelse(pof$REND_PERC>(1/8)*510&pof$REND_PERC<=(1/4)*510,1,0)
pof$RENDA3<-
ifelse(pof$REND_PERC>(1/4)*510&pof$REND_PERC<=(1/2)*510,1,0)
```

```
pof$RENDA4<-
ifelse(pof$REND_PERC>(1/2)*510&pof$REND_PERC<=510,1,0)
pof$RENDA5<-ifelse(pof$REND_PERC>510&pof$REND_PERC<=2*510,1,0)
pof$RENDA6<-
ifelse(pof$REND_PERC>2*510&pof$REND_PERC<=3*510,1,0)
pof$RENDA7<-
ifelse(pof$REND_PERC>3*510&pof$REND_PERC<=5*510,1,0)
pof$RENDA8<-
ifelse(pof$REND_PERC>5*510&pof$REND_PERC<=10*510,1,0)
pof$RENDA9<-ifelse(pof$REND_PERC>10*510,1,0)

### Resumindo os dados
library(dplyr)
by<-group_by(pof,COD_UF,NUM_SEQ,DV_SEQ,NUM_DOM)
summa.choc<-summarise(by,
                      V073=sum(V073*FAT_EXP2.x),
                      V074=sum(V074*FAT_EXP2.x),
                      V075=sum(V075*FAT_EXP2.x),
                      V076=sum(V076*FAT_EXP2.x),
                      V077=sum(V077*FAT_EXP2.x),
                      V078=sum(V078*FAT_EXP2.x),
                      V079=sum(V079*FAT_EXP2.x),
                      V080=sum(V080*FAT_EXP2.x),
                      V081=sum(V081*FAT_EXP2.x),
                      V082=sum(V082*FAT_EXP2.x),
                      V083=sum(V083*FAT_EXP2.x),
                      V084=sum(V084*FAT_EXP2.x),
                      V085=sum(V085*FAT_EXP2.x),
                      V086=sum(V086*FAT_EXP2.x),
                      V087=sum(V087*FAT_EXP2.x),
                      V088=sum(V088*FAT_EXP2.x),
                      V089=sum(V089*FAT_EXP2.x),
                      V090=sum(V090*FAT_EXP2.x),
                      V091=sum(V091*FAT_EXP2.x),
```

```
V092=sum(V092*FAT_EXP2.x),
V093=sum(V093*FAT_EXP2.x),
V094=sum(V094*FAT_EXP2.x),
V095=sum(V095*FAT_EXP2.x),
V096=sum(V096*FAT_EXP2.x),
V097=sum(V097*FAT_EXP2.x),
V098=sum(V098*FAT_EXP2.x),
V099=sum(V099*FAT_EXP2.x),
V100=sum(V100*FAT_EXP2.x),
V101=sum(V101*FAT_EXP2.x),
V102=sum(V102*FAT_EXP2.x),
V103=sum(V103*FAT_EXP2.x),
V104=sum(V104*FAT_EXP2.x),
V105=sum(V105*FAT_EXP2.x),
V106=sum(V106*FAT_EXP2.x),
V107=sum(V107*FAT_EXP2.x),
V108=sum(V108*FAT_EXP2.x),
V109=sum(V109*FAT_EXP2.x),
V110=sum(V110*FAT_EXP2.x),
V111=sum(V111*FAT_EXP2.x),
V112=sum(V112*FAT_EXP2.x),
V113=sum(V113*FAT_EXP2.x),
V114=sum(V114*FAT_EXP2.x),
V115=sum(V115*FAT_EXP2.x),
V116=sum(V116*FAT_EXP2.x),
V117=sum(V117*FAT_EXP2.x),
V118=sum(V118*FAT_EXP2.x),
V119=sum(V119*FAT_EXP2.x),
V120=sum(V120*FAT_EXP2.x),
V121=sum(V121*FAT_EXP2.x),
V122=sum(V122*FAT_EXP2.x),
V123=sum(V123*FAT_EXP2.x),
V124=sum(V124*FAT_EXP2.x),
V125=sum(V125*FAT_EXP2.x),
```

```
V126=sum(V126*FAT_EXP2.x),
V127=sum(V127*FAT_EXP2.x),
V128=sum(V128*FAT_EXP2.x),
V129=sum(V129*FAT_EXP2.x),
V130=sum(V130*FAT_EXP2.x),
V131=sum(V131*FAT_EXP2.x),
V132=sum(V132*FAT_EXP2.x),
V133=sum(V133*FAT_EXP2.x),
V134=sum(V134*FAT_EXP2.x),
V135=sum(V135*FAT_EXP2.x),
V136=sum(V136*FAT_EXP2.x),
V137=sum(V137*FAT_EXP2.x),
V138=sum(V138*FAT_EXP2.x),
V139=sum(V139*FAT_EXP2.x),
V140=sum(V140*FAT_EXP2.x),
V141=sum(V141*FAT_EXP2.x),
V142=sum(V142*FAT_EXP2.x),
V143=sum(V143*FAT_EXP2.x),
V144=sum(V144*FAT_EXP2.x),
V145=sum(V145*FAT_EXP2.x),
V146=sum(V146*FAT_EXP2.x),
V147=sum(V147*FAT_EXP2.x),
V148=sum(V148*FAT_EXP2.x),
V149=sum(V149*FAT_EXP2.x),
V150=sum(V150*FAT_EXP2.x),
V151=sum(V151*FAT_EXP2.x),
V152=sum(V152*FAT_EXP2.x),
V153=sum(V153*FAT_EXP2.x),
V154=sum(V154*FAT_EXP2.x),
RENDA1=sum(RENDA1*FAT_EXP2.x),
RENDA2=sum(RENDA2*FAT_EXP2.x),
RENDA3=sum(RENDA3*FAT_EXP2.x),
RENDA4=sum(RENDA4*FAT_EXP2.x),
RENDA5=sum(RENDA5*FAT_EXP2.x),
```

```
                        RENDA6=sum(RENDA6*FAT_EXP2.x),
                        RENDA7=sum(RENDA7*FAT_EXP2.x),
                        RENDA8=sum(RENDA8*FAT_EXP2.x),
                        RENDA9=sum(RENDA9*FAT_EXP2.x),
                        Price=sum(VAL_DESP*FAT_EXP2.x))


#### Treinamento da maquina

# Habilita o pacote kernlab
library(kernlab)

#Mantem somente quem consumiu chocolate
#summa.choc<-summa.choc[which(summa.choc$Price>0),]

## Standardized Data
summaZ<-as.data.frame(scale(summa.choc[,-c(1:4)]))

## Separa treinamento e validação
set.seed(3636)
iTrain<-sample(1:nrow(summaZ),size = trunc(0.7*nrow(summaZ)))
train<-summaZ[iTrain,]
valid<-summaZ[-iTrain,]

library(kernlab)
svm<-ksvm(Price~.,data=train,
          kernel="vanilladot",type="eps-svr",
          C=1,scaled=FALSE)


# salva os resultados
save.image("RESULTADOS.15062016.RData")

### CENSO
```

```
setwd("C:\\Users\\felix\\Documents\\Marcelo\\TCC\\Censo
2010\\Estatística\\DF_20150527\\DF\\Base            informaçoes
setores2010 universo DF\\CSV")

## Cria a base para previsão
censo.df<-data.frame()


#### Idade (Homem e Mulher)

## Verifica quantas variaveis tem na base (sinopse)
scan("Pessoa11_DF.csv",sep=';',what="character",
     nlines=1)
## Cria classe da variavel
class<-rep("character",137)

## Importa os dados
idadeH<-read.csv2("Pessoa11_DF.csv",colClasses = class)

## Código do setor
Cod_setor<-idadeH[,1]

## idade Homens
idadeH<-idadeH[,c(1,26:136)]
str(idadeH)
summary(idadeH)

# Conversao de character para numeric
asNumeric<-function(x)as.numeric(as.character(x))
charNumeric<-
function(d)modifyList(d,lapply(d[,sapply(d,is.character)],asNu
meric))
```

```
factorsNumeric<-
function(d)modifyList(d,lapply(d[,sapply(d,is.factor)],asNumer
ic))


# Cria variaveis
V073<-rowSums(charNumeric(idadeH[,c(26:36)]))
V098<-rowSums(charNumeric(idadeH[,c(35:39)]),na.rm = T)
V099<-rowSums(charNumeric(idadeH[,c(40:44)]),na.rm = T)
V100<-rowSums(charNumeric(idadeH[,c(45:49)]),na.rm = T)
V101<-rowSums(charNumeric(idadeH[,c(50:54)]),na.rm = T)
V102<-rowSums(charNumeric(idadeH[,c(55:59)]),na.rm = T)
V103<-rowSums(charNumeric(idadeH[,c(60:64)]),na.rm = T)
V104<-rowSums(charNumeric(idadeH[,c(65:69)]),na.rm = T)
V105<-rowSums(charNumeric(idadeH[,c(70:74)]),na.rm = T)
V106<-rowSums(charNumeric(idadeH[,c(75:79)]),na.rm = T)
V107<-rowSums(charNumeric(idadeH[,c(80:84)]),na.rm = T)
V108<-rowSums(charNumeric(idadeH[,c(85:89)]),na.rm = T)
V109<-rowSums(charNumeric(idadeH[,c(90:94)]),na.rm = T)
V110<-rowSums(charNumeric(idadeH[,c(95:99)]),na.rm = T)
V111<-rowSums(charNumeric(idadeH[,c(100:104)]),na.rm = T)
V112<-rowSums(charNumeric(idadeH[,c(105:109)]),na.rm = T)
V113<-as.numeric(idadeH[,110])
temp<-data.frame()
temp<-as.data.frame(Cod_setor)
idadeHH<-data.frame()
idadeHH<-as.data.frame(Cod_setor)
idadeHH$V073<-V073
idadeHH<-merge(idadeHH,idadeH[,-
c(2:12,35:110)],by="Cod_setor")
names(idadeHH)<-
c("Cod_setor","V073","V074","V075","V076","V077","V078","V079"
,"V080",
```

```
"V081","V082","V083","V084","V085","V086","V087","V088",
                  "V089","V090","V091","V092","V093","V094",
                  "V095","V096","V097")
temp<-merge(temp,idadeHH,by="Cod_setor")

temp$V098<-V098
temp$V099<-V099
temp$V100<-V100
temp$V101<-V101
temp$V102<-V102
temp$V103<-V103
temp$V104<-V104
temp$V105<-V105
temp$V106<-V106
temp$V107<-V107
temp$V108<-V108
temp$V109<-V109
temp$V110<-V110
temp$V111<-V111
temp$V112<-V112
temp$V113<-V113s
idade_H<-temp

## Idade mulheres

## Cria classe da variavel
class<-rep("character",137)

## Importa os dados
idadeM<-read.csv2("Pessoa12_DF.csv",colClasses = class)

## Código do setor
Cod_setor<-idadeM[,1]
```

```
## idade
idadeM<-idadeM[,c(1,26:136)]

# Cria variaveis
V114<-rowSums(charNumeric(idadeH[,c(26:36)]))
V139<-rowSums(charNumeric(idadeH[,c(35:39)]),na.rm = T)
V140<-rowSums(charNumeric(idadeH[,c(40:44)]),na.rm = T)
V141<-rowSums(charNumeric(idadeH[,c(45:49)]),na.rm = T)
V142<-rowSums(charNumeric(idadeH[,c(50:54)]),na.rm = T)
V143<-rowSums(charNumeric(idadeH[,c(55:59)]),na.rm = T)
V144<-rowSums(charNumeric(idadeH[,c(60:64)]),na.rm = T)
V145<-rowSums(charNumeric(idadeH[,c(65:69)]),na.rm = T)
V146<-rowSums(charNumeric(idadeH[,c(70:74)]),na.rm = T)
V147<-rowSums(charNumeric(idadeH[,c(75:79)]),na.rm = T)
V148<-rowSums(charNumeric(idadeH[,c(80:84)]),na.rm = T)
V149<-rowSums(charNumeric(idadeH[,c(85:89)]),na.rm = T)
V150<-rowSums(charNumeric(idadeH[,c(90:94)]),na.rm = T)
V151<-rowSums(charNumeric(idadeH[,c(95:99)]),na.rm = T)
V152<-rowSums(charNumeric(idadeH[,c(100:104)]),na.rm = T)
V153<-rowSums(charNumeric(idadeH[,c(105:109)]),na.rm = T)
V154<-as.numeric(idadeH[,110])
temp<-data.frame()
temp<-as.data.frame(Cod_setor)
idadeMM<-data.frame()
idadeMM<-as.data.frame(Cod_setor)
idadeMM$V114<-V114
idadeMM<-merge(idadeMM,idadeM[,-
c(2:12,35:110)],by="Cod_setor")
names(idadeMM)<-
c("Cod_setor","V114","V115","V116","V117","V118","V119","V120"
,
"V121","V122","V123","V124","V125","V126","V127","V128","V129"
,
```

```
"V130","V131","V132","V133","V134","V135","V136","V137","V138"
)
temp<-merge(temp,idadeMM,by="Cod_setor")

temp$V139<-V139
temp$V140<-V140
temp$V141<-V141
temp$V142<-V142
temp$V143<-V143
temp$V144<-V144
temp$V145<-V145
temp$V146<-V146
temp$V147<-V147
temp$V148<-V148
temp$V149<-V149
temp$V150<-V150
temp$V151<-V151
temp$V152<-V152
temp$V153<-V153
temp$V154<-V154
idade_M<-temp
file.remove("RESULTADOSFINAL.16062016")
save.image("RESULTADOSFINAL.16062016.RData")

## Junta
censo.df<-data.frame()
censo.df<-merge(idade_H,idade_M,by="Cod_setor")
View(censo.df)


#### Renda

## Verifica quantas variaveis tem na base
```

```
scan("DomicilioRenda_DF.csv",sep=';',what="character",nlines=1
)

## Cria classe da variavel
class<-rep("character",17)

## Importa os dados
renda<-read.csv2("DomicilioRenda_DF.csv",colClasses = class)

## Codigo do setor censitario
Cod_setor<-renda[,1]
temp<-as.data.frame(Cod_setor)

## Cria as variaveis
RENDA1<-renda[,7]
RENDA2<-renda[,8]
RENDA3<-renda[,9]
RENDA4<-renda[,10]
RENDA5<-renda[,11]
RENDA6<-renda[,12]
RENDA7<-renda[,13]
RENDA8<-renda[,14]
RENDA9<-renda[,15]

temp$RENDA1<-RENDA1
temp$RENDA2<-RENDA2
temp$RENDA3<-RENDA3
temp$RENDA4<-RENDA4
temp$RENDA5<-RENDA5
temp$RENDA6<-RENDA6
temp$RENDA7<-RENDA7
temp$RENDA8<-RENDA8
temp$RENDA9<-RENDA9
```

```
## Junta com o censo
censo.df<-merge(censo.df,temp,"Cod_setor")

## Converte as variaveis para numeric
censo.df$V073<-as.numeric(censo.df$V073)
censo.df$V074<-as.numeric(censo.df$V074)
censo.df$V075<-as.numeric(censo.df$V075)
censo.df$V076<-as.numeric(censo.df$V076)
censo.df$V077<-as.numeric(censo.df$V077)
censo.df$V078<-as.numeric(censo.df$V078)
censo.df$V079<-as.numeric(censo.df$V079)
censo.df$V080<-as.numeric(censo.df$V080)
censo.df$V081<-as.numeric(censo.df$V081)
censo.df$V082<-as.numeric(censo.df$V082)
censo.df$V083<-as.numeric(censo.df$V083)
censo.df$V084<-as.numeric(censo.df$V084)
censo.df$V085<-as.numeric(censo.df$V085)
censo.df$V086<-as.numeric(censo.df$V086)
censo.df$V087<-as.numeric(censo.df$V087)
censo.df$V088<-as.numeric(censo.df$V088)
censo.df$V089<-as.numeric(censo.df$V089)
censo.df$V090<-as.numeric(censo.df$V090)
censo.df$V091<-as.numeric(censo.df$V091)
censo.df$V092<-as.numeric(censo.df$V092)
censo.df$V093<-as.numeric(censo.df$V093)
censo.df$V094<-as.numeric(censo.df$V094)
censo.df$V095<-as.numeric(censo.df$V095)
censo.df$V096<-as.numeric(censo.df$V096)
censo.df$V097<-as.numeric(censo.df$V097)
censo.df$V098<-as.numeric(censo.df$V098)
censo.df$V099<-as.numeric(censo.df$V109)
censo.df$V100<-as.numeric(censo.df$V100)
censo.df$V101<-as.numeric(censo.df$V101)
censo.df$V102<-as.numeric(censo.df$V102)
```

```
censo.df$V103<-as.numeric(censo.df$V103)
censo.df$V104<-as.numeric(censo.df$V104)
censo.df$V105<-as.numeric(censo.df$V105)
censo.df$V106<-as.numeric(censo.df$V106)
censo.df$V107<-as.numeric(censo.df$V107)
censo.df$V108<-as.numeric(censo.df$V108)
censo.df$V109<-as.numeric(censo.df$V109)
censo.df$V110<-as.numeric(censo.df$V110)
censo.df$V111<-as.numeric(censo.df$V111)
censo.df$V112<-as.numeric(censo.df$V112)
censo.df$V113<-as.numeric(censo.df$V113)
censo.df$V114<-as.numeric(censo.df$V114)
censo.df$V115<-as.numeric(censo.df$V115)
censo.df$V116<-as.numeric(censo.df$V116)
censo.df$V117<-as.numeric(censo.df$V117)
censo.df$V118<-as.numeric(censo.df$V118)
censo.df$V119<-as.numeric(censo.df$V119)
censo.df$V120<-as.numeric(censo.df$V120)
censo.df$V121<-as.numeric(censo.df$V121)
censo.df$V122<-as.numeric(censo.df$V122)
censo.df$V123<-as.numeric(censo.df$V123)
censo.df$V124<-as.numeric(censo.df$V124)
censo.df$V125<-as.numeric(censo.df$V125)
censo.df$V126<-as.numeric(censo.df$V126)
censo.df$V127<-as.numeric(censo.df$V127)
censo.df$V128<-as.numeric(censo.df$V128)
censo.df$V129<-as.numeric(censo.df$V129)
censo.df$V130<-as.numeric(censo.df$V130)
censo.df$V131<-as.numeric(censo.df$V131)
censo.df$V132<-as.numeric(censo.df$V132)
censo.df$V133<-as.numeric(censo.df$V133)
censo.df$V134<-as.numeric(censo.df$V134)
censo.df$V135<-as.numeric(censo.df$V135)
censo.df$V136<-as.numeric(censo.df$V136)
```

```
censo.df$V137<-as.numeric(censo.df$V137)
censo.df$V138<-as.numeric(censo.df$V138)
censo.df$V139<-as.numeric(censo.df$V139)
censo.df$V140<-as.numeric(censo.df$V140)
censo.df$V141<-as.numeric(censo.df$V141)
censo.df$V142<-as.numeric(censo.df$V142)
censo.df$V143<-as.numeric(censo.df$V143)
censo.df$V144<-as.numeric(censo.df$V144)
censo.df$V145<-as.numeric(censo.df$V145)
censo.df$V146<-as.numeric(censo.df$V146)
censo.df$V147<-as.numeric(censo.df$V147)
censo.df$V148<-as.numeric(censo.df$V148)
censo.df$V149<-as.numeric(censo.df$V149)
censo.df$V150<-as.numeric(censo.df$V150)
censo.df$V151<-as.numeric(censo.df$V151)
censo.df$V152<-as.numeric(censo.df$V152)
censo.df$V153<-as.numeric(censo.df$V153)
censo.df$V154<-as.numeric(censo.df$V154)
censo.df$RENDA1<-as.numeric(censo.df$RENDA1)
censo.df$RENDA2<-as.numeric(censo.df$RENDA2)
censo.df$RENDA3<-as.numeric(censo.df$RENDA3)
censo.df$RENDA4<-as.numeric(censo.df$RENDA4)
censo.df$RENDA5<-as.numeric(censo.df$RENDA5)
censo.df$RENDA6<-as.numeric(censo.df$RENDA6)
censo.df$RENDA7<-as.numeric(censo.df$RENDA7)
censo.df$RENDA8<-as.numeric(censo.df$RENDA8)
censo.df$RENDA9<-as.numeric(censo.df$RENDA9)
censo.df[is.na(censo.df)]<-0
summary(censo.df)

### Faz a previsao do gasto
library(kernlab)
str(censo.df)
temp<-scale(censo.df[,-1])
```

```
GPchoc.df<-data.frame(censo.df$Cod_setor,temp)
GPchoc.df$Price<-predict(svm,GPchoc.df)
summary(GPchoc.df$Price)

save.image("RESULTADOS.16062016.RData")
save.image("RESULTADOS.15062016.RData")
View(GPchoc.df)
# Deixa apenas o Cod_setor e Price
Gasto_Padrão<-GPchoc.df[,c(1,93)]
help("write.table")

## Exporta a base Gasto_Padrão para construção dos mapas
setwd("C:\\Users\\felix\\Documents\\Marcelo\\TCC\\Censo
2010\\QGis")
write.table(Gasto_Padrão,file  =  "Gasto_Padrão(TDF).csv",sep  =
",",row.names=FALSE,col.names = FALSE)
```