UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ADMINISTRAÇÃO

DEPARTAMENTO DE CIÊNCIAS ADMINISTRATIVAS

RICARDO GASTAL SORUCO

DIMENSIONAMENTO DE MERCADOS E PREVISÃO DE DEMANDA NO SETOR ELÉTRICO BRASILEIRO

Porto Alegre
2018

RICARDO GASTAL SORUCO

DIMENSIONAMENTO DE MERCADOS E PREVISÃO DE DEMANDA NO SETOR ELÉTRICO BRASILEIRO

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciências Administrativas da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de Bacharel em Administração.

Orientador: Vinícius Brei

Porto Alegre
2018

RICARDO GASTAL SORUCO

DIMENSIONAMENTO DE MERCADOS E PREVISÃO DE DEMANDA NO SETOR ELÉTRICO BRASILEIRO

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Ciências Administrativas da Universidade Federal do Rio Grande do Sul, como requisito parcial para a obtenção do grau de Bacharel em Administração.

Trabalho de Conclusão de Curso defendido e aprovado em:

Banca examinadora:

_____

Prof. Doutor. Vinícius Brei
Orientador
UFRGS

_____

AGRADECIMENTOS

À minha família e minha namorada Marina, meu agradecimento pelo apoio ao longo de toda a minha trajetória na Universidade Federal do Rio Grande do Sul. Ao Grupo de Pesquisa de Marketing e Consumo, em especial ao professor Brei, Carla e Vinicius. Brei e Carla, obrigado pelos ensinamentos e por exigir sempre o melhor de mim. Vinicius, obrigado pelo suporte no R Studio. A todos os meus amigos, obrigado pelos momentos vividos juntos e que fazem a nossa vida mais especial.

RESUMO

Este trabalho foi desenvolvido com o objetivo de identificar as regiões com maior potencial de demanda para materiais elétricos, bem como o desenvolvimento de um método de cálculo de previsão de demanda e dimensionamento de mercados. A monografia calculou o gasto padronizado com materiais elétricos para cada um dos 96 distritos para a cidade de São Paulo, a partir da projeção feita para os 18 mil setores censitários da localidade. O estudo trouxe cinco variáveis para classificar a demanda por meio de Máquinas de Suporte Vetorial no software livre R Studio. As bases de dados utilizadas são públicas e fornecidas pelo IBGE. A variável renda se mostrou determinante para a demanda.

Palavras-chave: Previsão de demanda**;** georreferenciamento**;** dimensionamento de mercados; máquinas de suporte vetorial; aprendizado em máquina.

ABSTRACT

This study aims to identifie the regions with the greatest potential of demand for electrical materials, as well as the development of a method for calculating demand forecast and market size. The research calculated the standardized spending on electrical materials for each of the 96 districts from the city of São Paulo, based on the projection made for the 18,000 census tracts of the locality. The study brought five variables to classify demand through Support Vector Machines in free software R Studio. The databases used are public and provided by IBGE. The income variable was determinant for demand.

Keywords: demand forecast; georeferencing; market size; support vector machines, machine learning

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

AM – Aprendizado em Máquina

CPE – Cesta de Produtos Elétricos

EQM – Erro Quadrático Médio

GIS – *Geographical Information Systems*

IBGE – Instituto Brasileiro de Geografia e Estatística

INPC – Índice Nacional de Preços ao Consumidor

IPCA – Índice Nacional de Preços ao Consumidor Amplo

POF – Pesquisa de Orçamentos Familiares

SVM – *Support Vector Machines*

SUMÁRIO

1. INTRODUÇÃO

A incerteza e as informações imperfeitas sempre foram e sempre serão uma característica da conjuntura vivida pelos seres humanos. Os cenários e previsões não servem para prever o amanhã, mas fazem com que os indivíduos antecipem e planejem ações estratégicas, que serão aplicadas de acordo com o rumo dos acontecimentos (SORUCO, 2016). A previsão de demanda propicia um preparo para o futuro e ajuda os negócios a terem uma melhor performance, seja na otimização de estoques, reduzindo custos de armazenamento, como também no aumento das receitas, se houver a integração com o geomarketing. A organização busca estar próxima dos clientes com maior potencial de demanda. Portanto, sabendo onde estão estes consumidores, pode se posicionar melhor geograficamente no mercado – local onde ocorrem as trocas de bens e serviços, bem como a venda de materiais elétricos.

O mercado de varejo de materiais elétricos é constituído por uma ampla gama de produtos (interruptores de parede, tomadas, disjuntores, conduletes, etc.). Uma empresa que atua no setor elétrico é capaz de oferecer mais de 1600 itens diferentes - o que demonstra a extensa lista de objetos deste mercado. No ano de 2016, houve uma queda expressiva na indústria de Materiais Elétricos de instalação no Brasil de, aproximadamente, 7%. (ABINEE, 2017).

O presente trabalho busca entender e identificar os locais com maior potencial de demanda de materiais elétricos nos distritos do município de São Paulo, a partir de dados sobre o comportamento do consumidor na Pesquisa de Orçamentos Familiares (POF). Uma importante contribuição da monografia é que a metodologia aplicada para previsão de demanda serve para diversos produtos, não apenas materiais elétricos, já que a base de dados conta com mais de 13 mil produtos cadastrados. A POF é realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e não tem periodicidade constante, dificultando um pouco a atualização dos dados. Ela tem como objetivo produzir informações para atualizar a ponderação de indicadores de preços ao consumidor, os quais são publicados mensalmente pelo IBGE. Ademais, serve para atualizar a participação das despesas das famílias, assim, atualizando as Contas Nacionais. Ainda, a pesquisa do IBGE permite que seja estudada a evolução dos hábitos de consumo das famílias brasileiras (os dados da POF começam em 1974 e a nova Pesquisa de Orçamentos Familiares, em andamento, será a 6ª pesquisa realizada); assim como diferentes estudos sobre distribuição, concentração e desigualdade de

renda, aspectos demográficos e socioeconômicos e, por fim, a quantidade adquirida "per capita" (IBGE, 2017).

Através da extrapolação dos dados disponíveis na pesquisa sobre os hábitos de consumo das famílias, será possível entender o comportamento de consumo em diferentes regiões na localização selecionada. Como consequência, o trabalho de muitas empresas do setor, tais quais varejistas de produtos elétricos, será facilitado e otimizado.

Na conjuntura atual, com um nível crescente de competitividade, as empresas buscam superar as suas concorrentes a todo o momento. Levando em consideração os quatro P's do marketing (produto, preço, praça e promoção), há múltiplas maneiras de ofertar um produto diferente do seu competidor. As organizações podem oferecer, por exemplo, um preço mais atrativo; um produto mais durável; realizar a melhor publicidade ou simplesmente estar em uma localização mais próxima de seus clientes. A demanda de um produto ou serviço diminui conforme aumenta a distância entre o mercado consumidor e a localização deste consumidor (BROWN, 1992, BEAVON, 1977 apud ARANHA, 2001). Este raciocínio parte da teoria do Lugar Central desenvolvida por Lösch (1954) e Christaller (1933). Portanto, saber a localização onde existe uma maior demanda por determinado produto ou serviço é uma informação relevante para uma empresa.

O fato de a localização do negócio estar em uma região abastada de consumidores, circunstancialmente, acarretará em um acréscimo de receita para a empresa, visto que um maior número de produtos será vendido. Sabendo quais são as regiões ideais da cidade, assim levando a um encontro de vendedores e compradores com maior facilidade, os negócios têm uma maior propensão a alcançarem o sucesso.

A cidade de São Paulo é a cidade mais populosa do país e ela, sozinha, representava 11,4% do PIB nacional de 2010. Este foi o ano do último Censo Demográfico, quando a cidade possuía 11.253.503 milhões de habitantes e um total de 3.571.928 milhões de domicílios. Em 2017, o IBGE estima que a população atingiu a marca de 12.106.920 milhões de pessoas – representando uma variação de 7%, aproximadamente. De acordo com a fundação SEADE, haviam 3.951.074 milhões de domicílios ocupados em julho de 2017 e a projeção deles indica que haverá mais de 4 milhões de domicílios em 2018. Portanto, a São Paulo é a cidade mais representativa da economia brasileira e por isso foi escolhida como a região a ser estudada nesta pesquisa.

A partir desta situação apresentada, o problema de pesquisa desenvolvido levanta o seguinte questionamento: "**Quais seriam os locais de maior potencial de demanda para materiais**

elétricos no município de São Paulo?". O estudo tem o intuito de dimensionar mercados e prever demanda a partir de uma metodologia que é derivada do aprendizado em máquina, mais especificamente, Máquinas de Suporte Vetorial.

O campo de estudo de Máquinas de Suporte Vetorial é ainda limitado no Brasil, pois há poucos trabalhos que fizeram um modelo para dimensionar mercados e previsão de demanda. Até pouco tempo, extrair dados mais relevantes sobre a POF e o Censo era um grande desafio e muitos trabalhos realizavam muita análise descritiva e pouco informativa. As características que fazem com que Máquinas de Suporte Vetorial tenham um bom grau de profundidade em suas análises são: boa capacidade de generalização; robustez em grandes dimensões; convexidade da função objetivo e fundamentação teórica bem definida. (SMOLA et al., 1999 *apud* SILVA, 2014).

O tema escolhido é de importante compreensão, pois o modelo de previsão de demanda indicará os locais com maior potencial de consumo. O modelo poderá indicar quais os locais de São Paulo que deverão sofrer maior aumento de consumo, indicando os melhores e piores distritos para se atuar. Com a futura retomada econômica, aumentando o PIB em aproximadamente 10% até 2021, segundo projeções do Boletim Focus de 29 de junho de 2018, há uma oportunidade para aqueles que souberem onde se posicionar no mercado. Portanto, a divulgação deste trabalho trará informações relevantes tanto para as empresas já existentes quanto para os indivíduos que procuram abrir o seu próprio negócio. Vale ressaltar que o programa utilizado para análise é *open source* e outros interessados poderão realizar a pesquisa para outras localidades e até mesmo outros segmentos de mercado.

1.1 OBJETIVOS

Desenvolver um método para cálculo da previsão de demanda e dimensionamento do mercado de materiais elétricos de São Paulo.

1.1.1 Objetivos Específicos

I. Analisar a disponibilidade e quais são os melhores dados secundários para previsão de demanda no setor elétrico em São Paulo

II. Analisar a viabilidade do uso do método de Support Vector Machines para estimar demanda e dimensionamento de mercado em São Paulo

III. Estimar a demanda com base nos dados secundários selecionados

IV. Estimar o potencial de mercado com base nos dados secundários selecionados.4

V. Identificar, dentro do mercado escolhido, quais regiões tem maior potencial de demanda.

2. CONCEITOS GERAIS E REVISÃO DA LITERATURA

Nesta seção, será analisada a teoria que auxiliará no desenvolvimento da pesquisa. A seção foi dividida em duas partes: a primeira delas trata sobre diferentes metodologias e desafios encontrados em modelos de previsão de demanda. Depois, serão aprofundados conceitos, teorias e pesquisas realizadas na área de geomarketing, assim como as tecnologias encontradas no ambiente de negócios atual. Parar começar este capítulo, é importante distinguir os conceitos de previsão de demanda e de dimensionamento de mercados.

2.1 PREVISÃO DE DEMANDA E DIMENSIONAMENTO DE MERCADOS

A demanda é a representação de uma vontade ou a ação de demandar, segundo o dicionário Michaelis. A ação de demandar é comandada por um sujeito, o qual possui a vontade de obter qualquer matéria física ou química. Os mercados são formados pela representação da ação de troca entre aqueles que fornecem qualquer tipo de objeto ou experiência, por exemplo, e os que demandam este objeto ou esta experiência. Nos dias de hoje, é muito comum que haja uma troca financeira entre os indivíduos ou pessoas jurídicas. Portanto, um mercado é composto pela soma de todas as transações envolvendo um determinado produto ou serviço. Comumente, se mede os mercados por localidades e segmentos, como o de materiais elétricos, e delimita-se um espaço de tempo com início e fim para se obter a produção ou venda total durante o período de um ano. Em 2018, a ABINEE mediu o mercado de materiais elétricos e o volume (financeiro) movimentado de 1º de janeiro de 2017 a 31 de dezembro de 2017 somou R$ 7,5 bilhões no Brasil.

As previsões de demanda, realizadas por economistas, estatísticos, cientistas sociais e/ou de dados buscam explicar, através de um modelo matemático, as relações causais de algum acontecimento. Isto é, explicar as causas e os efeitos de algum fenômeno por meio de dados concretos, a fim de criar modelos com um alto grau de confiabilidade. Também existe a possibilidade de a previsão de demanda simplesmente prever algum acontecimento - sem explicá-lo ou descrevê-lo.

Outro conceito interessante é o de dimensionamento de mercados. Através da previsão de demanda, é possível medir quanto dinheiro um segmento de atuação da economia tem disponível para gastar e quanto dinheiro os moradores desse lugar estão dispostos a gastar com os produtos

ou serviços desse segmento. O dimensionamento de mercados procura projetar qual o valor (em quantidades monetárias) que será movimentado em um lugar e intervalo de tempo determinado. Usando como exemplo esta monografia, a previsão de demanda pode apontar quais os distritos da cidade de São Paulo que possuem um potencial de vendas maior – sem necessariamente dizer de quantos reais será essa movimentação financeira. O dimensionamento busca trazer o somatório de todas as movimentações financeiras ocorridas no período e, então, medir o mercado em alguma moeda.

Existem numerosos modelos para prever demanda e dimensionar mercados. As previsões com base em séries temporais são compostas por um conjunto de valores observados e medidos durante períodos de tempo em ordem sucessiva. A projeção dos valores futuros é feita com base nos valores passados, sem sofrer influência de outras variáveis (TUBINO, 2007 *apud* NUNES et al., 2009). Médias móveis (exponenciais e ponderadas) e análise de tendências (linear, polinomial, exponencial, logarítmica, etc.) são os principais recursos para o estudo de séries temporais.

A média móvel exponencial é representada pela equação matemática abaixo:

$$M_t = M_{t-1} + \alpha(D_{t-1} - M_{t-1})$$

Onde: **Mt** = média prevista para o período t; **Mt-1** = média prevista para o período **t-1**; **α** = coeficiente de ponderação; **Dt-1** = demanda prevista para o período t-1.

O coeficiente de ponderação α (alfa) varia de 0 a 1. Quanto maior o coeficiente, maior será a sua reação a uma variação de demanda. Na média exponencial móvel, a previsão do ano anterior sofre um ajuste através de um coeficiente de ponderação alfa (NUNES et al., 2009).

A média móvel ponderada é representada na seguinte equação:

$$Mm_n = \frac{\sum_{i=1}^{n} D_i}{n}$$

Onde **Mmn** = média móvel para o período n; **Σ** representa o somatório da demanda encontrada no período i; **n** representa o número de períodos; **i** o índice do período (i=1,2,3,...).

A média móvel ponderada utiliza os valores anteriores para formar a média. Segundo Moreira (1998 *apud* NUNES et al. 2009), existe a possibilidade de se utilizar pesos maiores para os valores mais recentes da série de tempo, com o objetivo de deixar a previsão mais próxima da realidade.

Previsões de demanda com base em séries temporais são bastante importantes para a tomada de decisão de gestores e são uma ferramenta bastante importante para muitos sistemas de suporte à decisão. O planejamento para a produção, por exemplo, depende diretamente dos dados sobre a demanda esperada. Diferentes quantidades de matéria-prima alteram o custo de um produto ou serviço ofertado - portanto, saber quanto é a demanda é extremamente importante para uma empresa. Entretanto, o desempenho dessas previsões com base em séries temporais está longe de alcançar a perfeição, especialmente quando o padrão recente das vendas demonstra volatilidade (CHOI et al., 2011).

Os conceitos econômicos de endogeneidade e heterogeneidade são relevantes para a realização de um modelo de previsão de demanda. Nos estudos da área em questão, estes dois princípios costumam ser um problema. A endogeneidade é tipicamente encontrada nos preços dos produtos estudados, já a heterogeneidade é vista entre as preferências dos consumidores. Caso essas características sejam ignoradas, é possível que as estimativas estejam tendenciosas e inconsistentes para o efeito de atividades ligadas a marketing (CHINTAGUNTA, 2001).

Esses tópicos acabaram ganhando notoriedade com a popularidade dos modelos *logit* (regressão logística), os quais caracterizam a demanda a partir de dados agregados. Na regressão logística, a variável dependente é categórica, ou seja, uma variável pertencente a um grupo ou categoria (exemplo: feminino e masculino). Quando se trabalha com previsão de demanda de algum produto ou serviço, os resultados da regressão devem ficar entre "0" e "1", sendo "0" com uma probabilidade nula de compra e "1" como 100% de chance de compra. Em uma típica regressão linear, os valores podem ultrapassar o intervalo de "0" a "1". Nos modelos *logit*, a regressão respeita os limites da variável dependente. Caso a variável dependente seja binária - dois valores únicos são possíveis - o resultado será a "compra" ou "não compra" do produto ou serviço.

Em algumas pesquisas na área de marketing, na esfera de comportamento do consumidor, o modelo *logit* busca responder a incidência de compra e a escolha de marca. Um problema apontado por Chintagunta (2001) é o fato do modelo *logit* incluir na pesquisa uma opção de "não comprar" como alternativa de resposta para algum consumidor. Inicialmente, parece ser um raciocínio inteligente, pois um consumidor pode não consumir e ou gostar das marcas que estavam dentre as alternativas. No entanto, um indivíduo pode, ao mesmo tempo, não escolher nenhuma marca ou produto preferido e ter uma ordem de preferência entre as opções (ex: pode haver uma marca que o consumidor odeia mais do que as outras). Este dilema social é chamado de

Independência de Alternativas Irrelevantes (CHINTAGUNTA, 2001). A Independência de Alternativas Irrelevantes é proveniente do Teorema da Impossibilidade de Arrow, que surgiu através da tese de doutorado do economista Kenneth Arrow em 1950.

Segundo Albuquerque e Bronnenberg (2009), a combinação de dados agregados de diferentes fontes de pesquisa ou a combinação com dados de *market share* (fatia de mercado) ajuda a melhorar o resultado proposto pelo modelo de previsão de demanda, assim como a heterogeneidade dos consumidores. São utilizados principalmente dois tipos de dados. Primeiro, o número de diferentes marcas que o indivíduo consumiu nos últimos 12 meses. Em segundo, os dados sobre penetração de mercado (*market penetration*). É importante lembrar que *market share* e *market penetration* são termos diferentes. A fatia de mercado é um indicador que mede o volume de vendas de uma empresa em determinado período de tempo sobre o volume total de vendas de um mercado em determinado período de tempo. A penetração de mercado descreve até que ponto um produto ou serviço é conhecido por potenciais clientes e o número de consumidores que realmente compram o produto ou serviço (BIZFLUENT, 2017).

A primeira monografia que realizou um método de previsão de demanda com dados da POF e Máquinas de Suporte Vetorial no Brasil, a qual servirá de base para ser aprimorada no presente estudo, foi elaborada em 2014 por Camila de Araújo e Silva, da Universidade Federal de Brasília. A pesquisa realizada por ela foi sobre o mercado de comida japonesa no Distrito Federal e buscava analisar as oportunidades de negócios por meio do geomarketing e máquinas de suporte vetorial. A previsão, através de máquinas, estimou a demanda por comida japonesa em cada um dos 4.690 setores censitários do Distrito Federal. A predição por meio da máquina foi possibilitada depois que foram encontrados padrões de gastos nos consumidores de comida japonesa. Os dados socioeconômicos destes consumidores foram extraídos da Pesquisa de Orçamentos Familiares, mas para a obtenção de uma amostra de consumidores maior, foi analisado o consumo a nível Brasil. Para que, mais tarde, os padrões encontrados fossem aplicados para o Distrito Federal, a região em que foram analisadas as oportunidades de negócio. O estudo pode indicar os locais de Brasília e região que possuíam maior potencial de consumo, dado as características dos moradores de cada setor censitário (SILVA, 2014).

2.2 GEOMARKETING

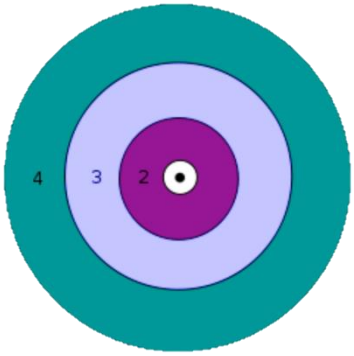
O geomarketing é uma área que, segundo alguns autores, surgiu na primeira metade do século XX nos Estados Unidos. Considerado como precursor da área por alguns, William Applebaum definiu a área como um campo que delimita e mede mercados, assim como os canais de distribuição – desde o produtor até o consumidor final (ARANHA e FIGOLI, 2001).

O americano desenvolveu e aplicou métodos quantitativos para selecionar pontos comerciais. Um dos primeiros estudos, feito na cidade de Cincinnati, abordou diferenças funcionais, nos formatos e na localização dos centros. Nos anos 1930, em pleno crescimento populacional e das cidades, as lojas começaram a se preocupar onde abririam suas próximas lojas. Nessa época, em uma de suas pesquisas, Applebaum utilizou o método que é conhecido por análogo. Ele estava buscando determinar, através de um mapeamento, a área de influência primária de algumas lojas e fez isso com uma série de entrevistas para descobrir onde moravam os consumidores. Depois disso, cruzava os dados das vendas com a localização de moradia dos clientes; consequentemente, conseguiu estimar possíveis localizações futuras para as próximas lojas (ARANHA e FIGOLI, 2001).

Outros autores definem que a origem do geomarketing vem da combinação da Geografia e da Economia. As primeiras teorias econômicas que remetem a geografia surgiram no início do século XIX. Von Thünen, em 1826, lançou a sua Teoria do "Estado Isolado", a qual é ensinada nos cursos superiores de ciências econômicas ainda hoje e cujo objetivo era auxiliar na maximização de resultados na atividade agrícola. A teoria original levava em consideração a distância entre o centro consumidor, a fertilidade dos solos e as diferentes culturas de alimentos (SILVA, 2013).

Um pressuposto central era o de que os custos com transporte aumentam à medida que a distância entre o agricultor e o centro comercial se eleva. Conforme a Figura 1, a região mais próxima do centro urbano (centro urbano representado pelo círculo branco) seria a ideal para produzir produtos mais perecíveis (em roxo); depois, mais afastada, seria a área (em lilás) destinada para a produção de alimentos menos perecíveis (grãos em geral), e, por fim, a área número 4 seria de pastagens para animais.

Figura 1: Anéis de Thünen



Fonte: Adaptado de Thünen (1966)

É interessante que esta teoria pode ser aplicada não apenas para a agricultura. O modelo criado por ele, que tem por objetivo maximizar o lucro do agricultor, é aplicável ao mundo atual de negócios. As empresas querem estar localizadas estrategicamente no local que lhe gere um resultado financeiro melhor. Seja pelo aumento da sua receita ou pela diminuição dos seus custos. Cada localidade é diferente da outra e, para tal circunstância, estudos relacionados ao geomarketing são fundamentais para o entendimento das características socioeconômicas de determinado lugar. Compreendendo estes atributos, um modelo econômico poderá ser criado e apontar os melhores locais para se abrir uma loja de materiais elétricos ou uma concessionária de veículos, por exemplo.

Bucklin et al. (2008) analisou a intensidade de distribuição de concessionárias de veículos em um centro urbano. Neste caso, foram concebidas três medidas para se chegar à distribuição ideal de concessionárias. A primeira delas envolve a acessibilidade da loja, depois vem a medida de concentração de lojas e, por fim, a dispersão entre as lojas. O comportamento de consumo é diferente entre cada consumidor e se sabe que as características sócio demográficas afetarão da decisão de compra, mas é importante entender que cada mercado tem suas peculiaridades. No estudo de Bucklin et al. (2008), foi diagnosticado que o consumo deve ser sensibilizado pela acessibilidade, concentração e dispersão entre as concessionárias já existentes. Para que as pesquisas de previsão de demanda de um ponto de vendas sejam mais ricas, deverão ser ponderadas as particularidades de cada mercado e espaço geográfico.

Na conjuntura empresarial atual, tecnologias e sistemas (*Geographical Information Systems* – GIS) a respeito de georreferenciamento, geoprocessamento e dados georreferenciados

vem sendo amplamente utilizados. Um sistema de informação geográfica (SIG) permite visualizar, indagar, analisar e interpretar dados para entender relacionamentos, padrões e tendências.

Para que estas tecnologias consigam atender as demandas das organizações, é devidamente necessário que haja um *input* de um grande número de informações, que são adquiridas através de pesquisas e grandes bases de dados. A metodologia utilizada por Silva (2014), que foi a primeira monografia a utilizar dados da Pesquisa de Orçamentos Familiares com *Support Vector Machines*, realizou uma operação semelhante, importando as linhas de código para depois gerar informações de grande relevância.

O nível de complexidade e competitividade do ambiente vivido pelas empresas faz com que o uso deste tipo de ferramenta, seja adquirida de terceiros, ou criada na própria empresa, esteja presente em vários setores de atuação empresarial. Sanati e Sanati (2013) realizaram um estudo que mostrou o aumento da procura por *Geographical Information Systems* no setor da saúde. Na plataforma PubMed, de procura de artigos de medicina, o número de artigos relacionados a GIS triplicou. A pesquisa inclusive recomendou a alocação de futuros recursos para a coleta de dados que possibilitem a geocodificação em um futuro próximo. O entendimento geográfico se tornou uma ferramenta de extrema relevância no processo decisório de uma empresa (TAKETA, 1993).

3. MÉTODO

Este capítulo descreverá as bases de dados utilizadas pela pesquisa e como foi feita a elaboração da cesta de produtos elétricos. Também será fundamentado o método para se chegar aos resultados finais sobre a previsão de demanda.

3.1 A PESQUISA

O presente trabalho visa realizar um cálculo de previsão de demanda, através de métodos quantitativos, de materiais elétricos na cidade de São Paulo. A estimativa da demanda é possível de ser quantificada com as informações contidas nas bases de dados do Instituto Brasileiro de Geografia e Estatística. A renda per capita, valor gasto no produto específico, estrato geográfico, forma de aquisição, local de aquisição, dentre outras variáveis são exemplos de informações que existem nas bases de dados, que serão úteis para o estudo. Estas bases de dados são públicas e serão mais exploradas na seção 3.3 desta monografia (Coleta de Dados).

Para realizar a análise destas informações, será utilizado um método analítico conhecido por *Support Vector Machines*, que é uma área do Aprendizado em Máquina. De uma maneira sucinta, ocorre uma importação de informações (variáveis) que são consideradas determinantes no consumo de materiais elétricos. A máquina buscará estabelecer um padrão de consumo com base nestas informações, as quais foram fornecidas pelas bases de dados do trabalho. Desta forma, a máquina conseguirá predizer o consumo de materiais elétricos para os diferentes distritos de São Paulo, onde as características socioeconômicas variam de um distrito para o outro.

A presente metodologia desta monografia foi desenvolvida a partir de uma pesquisa bibliográfica. A base inicial de pesquisa para o Aprendizado em Máquina com dados da POF e do Censo provém do estudo de SILVA (2014).

3.2 UNIVERSO E AMOSTRA

É importante frisar que, apesar dos resultados do trabalho serem focalizados na cidade de São Paulo, o comportamento de consumo será analisado com base nas informações disponíveis sobre todos aqueles que consumiram os produtos da Cesta de Produtos Elétricos (Tabela 2). Para

que o comportamento de consumo pudesse ser interpretado da melhor maneira, o universo da pesquisa compreende não apenas os paulistanos, mas todos os brasileiros, de todas as regiões do Brasil. As informações relevantes sobre estes indivíduos para esta pesquisa podem ser encontradas nas amostras da Pesquisa de Orçamentos Familiares (2008-2009) e no Censo Demográfico (2010). Sendo que a POF traz as informações sobre as despesas dos brasileiros com materiais elétricos e o Censo traz as características sociodemográficas dos moradores da cidade de São Paulo. Portanto, o universo da base de dados é composto pelos 59 mil domicílios que responderam a POF e todos os brasileiros representados no Censo 2010. A amostra é representada pelos domicílios que consumiram qualquer produto da Cesta de Materiais Elétricos (3.3.1) e, também, todos os moradores da cidade de São Paulo.

3.2.1 A Pesquisa de Orçamentos Familiares (2008-2009)

A Pesquisa de Orçamentos Familiares é uma pesquisa domiciliar de abrangência nacional e é realizada pelo Instituto Brasileiro de Geografia e Estatística. O intuito da POF é de analisar diversos aspectos relacionados aos rendimentos das famílias. A atualização de indicadores monetários, como o IPCA e o INPC, é possibilitada através das informações adquiridas com as famílias brasileiras.

É importante mencionar que a Pesquisa de Orçamentos Familiares funciona por amostragem. Diferentemente do Censo Demográfico Nacional, onde todos os 67 milhões de domicílios foram verificados. Na amostragem da POF, foram selecionados 59.548 domicílios para a realização de entrevistas.

A POF tem duração de 12 meses, devido ao grande número de domicílios e, também, porque as despesas mudam ao longo do ano, como no verão e no inverno. Existe uma divisão do ano em períodos, e as famílias serão selecionadas para serem analisadas durante um desses períodos, que duram nove dias, em média.

A Pesquisa de Orçamentos Familiares não possui uma periodicidade fixa, mas o IBGE visa realizar uma POF completa de cinco em cinco anos (IBGE, 2008). A POF já foi realizada em cinco ocasiões, sendo que a primeira ocorreu entre 1974 e 1975. A pesquisa que está em andamento (durante 2017 e 2018) é a sexta edição da Pesquisa de Orçamentos Familiares.

O foco da pesquisa é trazer dados sobre os gastos do domicílio e gastos individuais. No entanto, como uma pesquisa do tamanho da POF tem custos elevados, são pesquisadas outras informações sobre os moradores, tais como: segurança e condições de vida (bem-estar), número de bens duráveis, nível educacional, dentre outros. Os dados são minuciosos, pois as famílias realizam um autopreenchimento de todas as despesas (e respectivas quantidades adquiridas) do 2º ao 8º dia da coleta, assim, diminuindo a probabilidade de coletar informações menos confiáveis. Caso a pesquisa tivesse um dia de duração, as famílias teriam que estimar muitos dos seus gastos, pois seria impossível de lembrar com precisão de todas as despesas e quantidades dos últimos sete dias.

Os resultados da Pesquisa de Orçamentos Familiares servem para o cruzamento de dados, geração de novas informações, servir de base a pesquisas (públicas e privadas) ou então, simplesmente, para se compreender os hábitos e as características das famílias brasileiras. Além disso, estes resultados servem para a formulação de políticas públicas.

A amostragem da POF é definida pela amostragem aleatória simples, uma técnica estatística. Atualmente, a pesquisa é dividida em diferentes regiões do país. Nas pesquisas que antecederam 2003, a POF pesquisava somente nas seguintes capitais do Brasil: Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Curitiba e Porto Alegre, Goiânia e no Distrito Federal. A amostra foi redesenhada para propiciar a publicação de resultados para o Brasil. A Pesquisa de Orçamentos Familiares (2008-2009), junto com o Censo Demográfico (2010), são as duas principais fontes de dados da pesquisa e ambas são realizadas pelo IBGE.

3.2.2 Censo Demográfico 2010

O Censo Demográfico é uma pesquisa estatística realizada pelo IBGE que busca recolher um grande número de informações sobre a população brasileira. No Brasil, assim como na maioria dos países, ele é realizado a cada 10 anos. O Censo é uma pesquisa diferente da Pesquisa de Orçamentos Familiares quando nos referimos ao tipo de amostra, pois a amostra do Censo, além de ser muito maior, é uma amostra estratificada - ao invés da amostragem aleatória simples na POF, com 59.548 domicílios. No Censo de 2010, foram visitados 10,7% do total de domicílios brasileiros, totalizando 6.192.332 domicílios (IBGE, 2010).

Na tabela abaixo, percebe-se que as cidades são estratificadas em cinco categorias através do número de habitantes.

Tabela 1: Amostragem por tamanho de município

Habitantes por município	Fração Amostral
Até 2,5 mil	50%
2,5 mil até 8 mil	33%
8 mil a 20 mil	20%
20 mil a 500 mil	10%
500 mil ou mais	5%

Fonte: Elaborado pelo autor a partir de dados do IBGE (2010)

É importante destacar que o Censo Demográfico possui uma abrangência nacional e, portanto, muito maior que a da POF, possibilitando expandir e predizer os resultados encontrados nos padrões de consumo da Pesquisa de Orçamentos Familiares. Isto é possível porque há cinco importantes variáveis socioeconômicas em comum nas duas pesquisas. Elas são: "Idade", "Gênero", "Renda", "Alfabetização" e "Raça".

3.3 COLETA DE DADOS

Os dados coletados, que foram selecionados para poder explicar os padrões de consumo, provêm dos microdados da POF (2008-2009). Estes microdados são separados em 16 bases diferentes e nem todas elas serão utilizadas, pois cada escopo de pesquisa demanda diferentes bases. A partir da seleção dos produtos na Cesta de Produtos Elétricos, pode-se partir para a coleta dos dados que interessam o trabalho a ser executado.

3.3.1 Elaboração da Cesta

Para identificar o padrão de consumo de produtos elétricos, foram selecionados os produtos que estão relacionados diretamente com a instalação de energia elétrica nos domicílios. Cabe ressaltar que estes foram escolhidos da lista que contém os produtos cadastrados na POF 2008-2009.

Na Tabela 2, estão os 16 produtos compõem a chamada Cesta de Produtos Elétricos (CPE). Quatro desses materiais representam o mesmo produto, mas que foram coletados em dois diferentes quadros (Despesas de 90 dias; Despesas de 12 Meses) e por isso aparecem duas vezes na cesta.

Tabela 2: Cesta de Produtos Elétricos (CPE)

Quadro	Grupo de Despesa	Código do Item	Produto
DESPESAS DE 90 DIAS		*00201*	*ENERGIA ELÉTRICA KWH*
		01201	FIO E MATERIAL ELÉTRICO EM GERAL
		01202	MATERIAL PARA INSTALAÇÃO ELÉTRICA
		01203	FIO PARA INSTALAÇÃO ELÉTRICA
		03801	CAIXA DE LUZ (PADRÃO)
DESPESAS DE 12 MESES		01201	FIO E MATERIAL ELÉTRICO EM GERAL
		01202	MATERIAL PARA INSTALAÇÃO ELÉTRICA
		01203	FIO PARA INSTALAÇÃO ELÉTRICA
		03801	CAIXA DE LUZ (PADRÃO)
~~OUTRAS DESPESAS~~		~~05601~~	~~REGULADOR DE VOLTAGEM~~
DESPESAS INDIVIDUAIS		*01001*	*ENERGIA ELÉTRICA DE OUTROS IMOVEIS (KWH)*
CADERNETA DE DESPESA	86	05501	LÂMPADA DE QUALQUER TIPO
		05502	LÂMPADA
		05503	LÂMPADA FLUORESCENTE
		05504	LÂMPADA FRIA
		05505	LÂMPADA BRANCA
		05506	LÂMPADA ELETRÔNICA
		05507	LÂMPADA TIPO ECONÔMICA

Fonte: Elaboração própria a partir da Pesquisa de Orçamentos Familiares 2008-2009 (IBGE, 2010)

Durante a seleção, foram procurados os produtos através de palavras-chave ou "partes" de palavras-chave, visto que as palavras da lista estão sem acentuação e podem aparecer no feminino, masculino e palavras derivadas do seu substantivo. A lista compreende quase 14 mil produtos e por isso se utilizou esta técnica. As palavras utilizadas foram: "eletric"; "lâmpad"; "luz"; "tomada"; "interruptor"; "disjuntor"; "ferrage"; "fio"; "material"; "volt"; "inst"; "resist". Também foram verificados os códigos referentes a despesas com habitações.

É importante destacar que os itens em amarelo foram adicionados para fortalecer o modelo de previsão, pois não se tratam realmente de materiais elétricos. Eles foram adicionados para aumentar o número de informações encontrados na POF, a fim de tornar o modelo mais confiável. Por outro lado, o regulador de voltagem foi removido da cesta de produtos elétricos porque apresentou um ínfimo número de observações na amostra.

3.3.2 Seleção de Bases de Dados e Variáveis

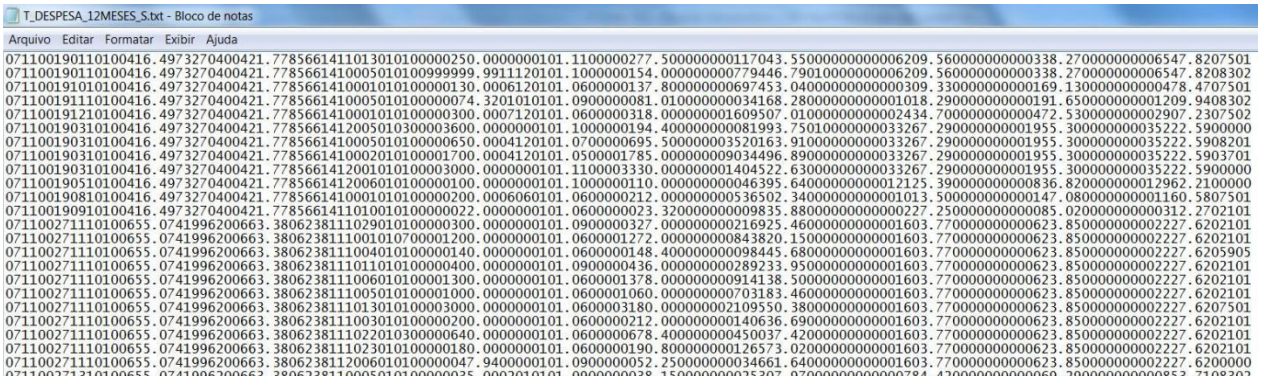
Para entender o comportamento de consumo dos itens compreendidos na Cesta de Produtos Elétricos, procedeu-se a seleção de variáveis que possam explicar o comportamento dos indivíduos que consumiram pelo menos um dos produtos da cesta. A identificação destes consumidores será

realizada no programa R Studio, um programa com linguagem de programação *open source*, em vista disso, facilitando ainda mais o acesso a este tipo de pesquisa, a qual usa dados públicos. Ainda, apenas 1 dessas bases já supera o limite de linhas do Microsoft Excel. Serão utilizadas, primeiramente, quatro bases de dados da Pesquisa de Orçamentos Familiares.

A primeira delas é sobre as "despesas de 90 dias", onde estão as informações sobre o consumo de "Fios e materiais elétricos em geral; Material para instalação elétrica; Fio para instalação elétrica; e Caixa de luz (padrão)". Nas "despesas de 90 dias", cada linha do código (em formato ".txt") possui 27 informações dispostas em 180 caracteres.

As "despesas de 12 meses" formam a segunda base a ser utilizada, onde estão as informações sobre o consumo dos mesmos produtos que estão sendo analisados na base "despesas de 90 dias", só que agora é sobre o consumo em outro intervalo de tempo – 12 meses. Nas "despesas de 12 meses", cada linha do código (em formato ".txt") possui 24 informações dispostas em 154 caracteres. Em um primeiro momento, é impossível concluir algo sobre o conteúdo dos códigos – como pode ser visto na figura 2.

Figura 2: Exemplo de disposição dos microdados da POF no formato .txt



Fonte: IBGE (2010)

Após o estudo das regras dos códigos, pode-se gerar uma planilha de linhas e colunas, a qual será utilizada no R Studio e também no Microsoft Excel, como na Figura 3:

Figura 3: Exemplo de disposição dos microdados da POF no formato .xls

	K	L	M	N	O	P	Q	R	S
3	NÚMERO_QUADRO	CÓDIGO_ITEM	FORMA_AQUISIÇÃO	VALOR_DESPESA_AQUISIÇÃO	MÊS_ÚLTIMA_DESPESA	NÚMERO_MESES	FATOR_ANUALIZAÇÃO	DEFLATOR_FATOR	VALOR_DESPESA_DEFLACIONADO
4	11	01301	01	00000250.00	00	00	01	01.11	00000277.50
5	10	00501	01	00999999.99	11	12	01	01.10	00000154.00
6	10	00101	01	00000130.00	06	12	01	01.06	00000137.80
7	10	00501	01	00000074.32	01	01	01	01.09	00000081.01
8	10	00101	01	00000300.00	07	12	01	01.06	00000318.00
9	12	00501	03	00003600.00	00	00	01	01.10	00000194.40
10	10	00501	01	00000650.00	04	12	01	01.07	00000695.50
11	10	00201	01	00001700.00	04	12	01	01.05	00001785.00
12	12	00101	01	00003000.00	00	00	01	01.11	00003330.00
13	12	00601	01	00000100.00	00	00	01	01.10	00000110.00
14	10	00101	01	00000200.00	06	06	01	01.06	00000212.00
15	11	01001	01	00000022.00	00	00	01	01.06	00000023.32
16	11	02901	01	00000300.00	00	00	01	01.09	00000327.00
17	11	00101	07	00001200.00	00	00	01	01.06	00001272.00
18	11	00401	01	00000140.00	00	00	01	01.06	00000148.40
19	11	01101	01	00000400.00	00	00	01	01.09	00000436.00
20	11	00601	01	00001300.00	00	00	01	01.06	00001378.00
21	11	00501	01	00001000.00	00	00	01	01.06	00001060.00

Fonte: Elaboração própria a partir de dados do IBGE (2010)

A Figura 3 retrata as primeiras 34 linhas de um total de 2.625.052 linhas existentes nos 4 arquivos. Sendo que cada uma dessas linhas possui uma média de 160 caracteres (para as 4 bases). Ou seja, são mais de 420 milhões de caracteres analisados. Se todos os caracteres estivessem dispostos em uma linha reta, o caminho percorrido com todos os caracteres seria de 650 quilômetros (é possível ligar Paris, na França, até Zurique, na Suíça – já considerando todas as curvas das estradas). Os 160 caracteres ocupam 25 centímetros em uma tela de computador (Figura 2). Cada linha de código da figura 2 possui 154 caracteres, os quais resultarão em 24 colunas de diferentes informações (figura 3).

No registro de "outras despesas", a terceira base, se encontram os dados sobre o consumo de "Regulador de voltagem". Nas "outras despesas", cada linha do código (em formato ".txt") possui 23 informações dispostas em 151 caracteres.

A quarta base é a "caderneta de despesa", onde estão as informações que se buscam sobre o consumo das lâmpadas da CPE. Na "caderneta de despesa", cada linha do código (em formato ".txt") possui 28 informações dispostas em 182 caracteres.

Tendo definido os produtos a serem buscados dentro das já referidas bases de dados, pode-se programar a leitura no programa R Studio para gerar o consumo observado nos indivíduos que consumiram da cesta de produtos elétricos. Há alguns detalhes que precisam ser levados em consideração para que a programação gere os resultados mais aproximados da realidade.

Primeiramente, é necessário realizar um cálculo da inflação monetária sobre o consumo observado para corrigir os valores encontrados. Assim, a renda e as despesas ficarão com valores atualizados. Também é essencial multiplicar os valores de consumo observado pelo fator de

expansão, o qual é encontrado na mesma linha de código na base de dados. O fator de expansão, ou peso amostral, serve de ajuste para compensar a não resposta de unidades investigadas pela pesquisa do IBGE. "O peso foi calculado para cada domicílio e atribuído a cada unidade de consumo e pessoa desse domicílio" (IBGE, 2011).

A partir do consumo real observado, foram estabelecidas as seguintes variáveis para se explicar o comportamento do consumo.

a) Idade: as informações sobre "idade" são encontradas na base "T_MORADOR_S.txt".

b) Gênero: as informações sobre "gênero" são encontradas na base "T_MORADOR_S.txt".

c) Renda: as informações sobre "renda" são encontradas na base "T_MORADOR_S.txt".

d) Alfabetização: As informações sobre "alfabetização" são encontradas na base "T_MORADOR_S.txt".

e) Raça: as informações sobre "raça" são encontradas na base "T_MORADOR_S.txt".

f) Consumo de energia elétrica: as informações sobre o "consumo de energia elétrica" são encontradas na base "T_DESPESA_90DIAS_S.txt".

g) Consumo (em reais): o consumo real observado (em reais) servirá de base para entender o comportamento de consumo a partir das variáveis delimitadas acima.

Para a validação dos dados, serão utilizados dados do Censo 2010 da cidade de São Paulo. Mais especificamente, os "Resultados do Universo" e "Agregados_por_setores_censitários" (IBGE, 2016). Portanto, a validação ocorrerá em cada um dos 18.362 setores censitários que compõem São Paulo, exceto aqueles que apresentarem informações incompletas (foram removidos 158 setores censitários com informações incompletas). O programa R Studio irá gerar um "Gasto Padronizado", que será explicado na sequência da metodologia. Em um segundo momento, será feito o agrupamento dos setores censitários nos seus respectivos distritos através da média simples. A Figura 4 ilustra uma das seis bases, no formato "xls", utilizadas.

Figura 4: Exemplo de disposição dos microdados do Censo no formato .xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Cod_setor	Situacao_seto	V001	V002	V003	V004	V005	V006	V007	V008	V009	V010	V011	V012	V013	V014
2	355030801000001	1	2	907777	903817	3960	0	0	14	51	94	49	42	12	5	1
3	355030801000002	1	0	846525	846525	0	0	2	28	87	94	45	30	12	4	4
4	355030801000003	1	0	505662	505662	0	0	3	11	38	50	19	16	10	3	39
5	355030801000004	1	0	446011	446011	0	0	3	8	39	57	14	14	9	1	36
6	355030801000005	1	0	615215	615215	0	0	2	20	68	76	37	18	8	3	8
7	355030801000006	1	0	507028	507028	0	0	2	12	34	63	20	25	6	1	49
8	355030801000007	1	0	447486	447486	0	0	3	21	66	62	26	15	6	1	49
9	355030801000008	1	0	465060	465060	0	0	5	16	42	49	15	20	7	2	70
10	355030801000009	1	0	627150	627150	0	0	6	19	63	78	33	30	11	1	4

Fonte: Desenvolvido pelo autor

No caso do "DomicílioRenda_SP1" (acima), as colunas G até O representam o número de pessoas que pertencem a uma faixa de renda – dentro de um setor censitário (coluna A). Sendo a coluna G retratando a faixa 1 (até 1/8 de salário mínimo) de renda e as colunas subsequentes caracterizando as seguintes faixas de renda, as quais terminam na coluna O, que é a faixa 9 de renda (acima de 10 salários mínimos).

As bases de dados do Censo 2010 relevantes para a pesquisa são:

a) "DomicilioRenda_SP1" -> Representa a variável "Renda"
b) "Pessoa01_SP1" -> Representa a variável "Alfabetização"
c) "Pessoa03_SP1" -> Representa a variável "Raça"
d) "Pessoa11_SP1" -> Representa a variável "Homem"
e) "Pessoa12_SP1" -> Representa a variável "Mulher"
f) "Pessoa13_SP1" -> Representa a variável "Idade"

3.4 ANÁLISE DOS DADOS

Utiliza-se a metodologia do Aprendizado em Máquina para realizar a extrapolação das informações. Como a base de microdados da POF é relativamente pequena, principalmente quando comparamos com o tamanho da amostra do Censo. Assim, se utilizam suas informações de despesas com produtos (não existentes no Censo) para ensinar o padrão observado de consumo. Assim, os resultados do modelo são expansíveis através do alto volume de informações que o Censo disponibiliza. Isto só é possível, pois há algumas variáveis em comum entre as duas pesquisas.

3.4.1 Aprendizado em máquina e máquinas de suporte vetorial

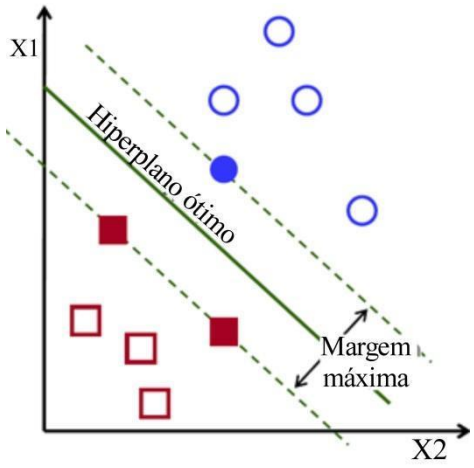
O aprendizado em máquina (AM) é uma classe de inteligência artificial que proporciona aos computadores e *softwares* a habilidade de aprender a preverem resultados. Além de não necessitar de uma programação explícita. No campo de análise de dados, o aprendizado em máquina serve como um método para calcular modelos matemáticos e algoritmos complexos e, consequentemente, análises preditivas de diferentes mercados se tornarão mais confiáveis e robustas. "Usando algoritmos que aprendem iterativamente com os dados, o aprendizado em máquina permite que os computadores encontrem informações ocultas sem serem explicitamente programados para procurarem estas informações" (SAS, 2016). Estas informações que eram previamente "ocultas" surgiram através do aprendizado sobre a relação de dados históricos e tendências.

Um conceito básico para o entendimento de Aprendizado de Máquina é o da indução, pois se trata de um aprendizado indutivo, dividido entre supervisionado e não supervisionado. Este princípio foi fundamentado por Angluin e Smith em 1983 e serve para formular uma hipótese geral a partir das presentes premissas ou dados (SILVA, 2014).

O aprendizado supervisionado faz o uso de um professor externo, que agrega conhecimento sobre o ambiente e utiliza exemplos para ensinar a máquina (com entradas e saídas). Assim, o Aprendizado de Máquina cria um algoritmo com o conhecimento computado pelo professor. No caso do aprendizado não supervisionado, o algoritmo aprende a agrupar as entradas computadas a partir de um padrão de qualidade. Estas técnicas ajudam a encontrar padrões/tendências e no entendimento dos dados (LORENA e CARVALHO, 2007).

As máquinas de suporte vetorial são uma das diversas aplicações de aprendizado em máquina. Sua fundação teórica está na Teoria do Aprendizado Estatístico, elaborada por Vapnik em 1995. Simplificadamente, as *Support Vector Machines* estabelecem princípios que deverão ser seguidos para se atingir uma boa generalização na classificação de dados pertencentes a classes distintas (SILVA, 2014). Como pode ser observado na Figura 5.

Figura 5: Representação do hiperplano



Fonte: Adaptado de BHALLA (2017)

Na figura acima, podemos ver que o plano traçado pela máquina busca separar as duas categorias de comportamentos distintos (não comprador em vermelho; comprador em azul). É preciso salientar que o plano traçado ótimo tenta explicar da melhor maneira possível o comportamento observado pelos consumidores.

Para as SVM, o tipo de aprendizado sempre será supervisionado. Isso ocorre porque os dados agregados fornecidos serão anteriormente organizados e corretamente rotulados antes de ingressarem no sistema; no caso de aprendizado não supervisionado, os dados são latentes, ou seja, podem não ser rotulados ou são desconhecidos. Sendo um dado latente, o aprendizado pela máquina será realizado por *clustering*, que é uma outra aplicação do aprendizado em máquina. Um exemplo é a mineração de dados (*data mining*), a qual manipula com dados não estruturados ou semiestruturados.

Um dos primeiros estudos sobre previsão de demanda envolvendo *Support Vector Machines* foi realizado em 2001 durante a competição EUNITE. O trabalho vencedor usou SVM para prever a carga de eletricidade máxima diária dos próximos 31 dias. Como a previsão do tempo, ao menos no início no século XXI, não era muito precisa para espaços de tempo superiores a 14 dias, os cientistas tiveram que utilizar séries temporais sobre a previsão do clima, já que a temperatura influencia diretamente no consumo energético. Através do aprendizado em máquina, a equipe vencedora descobriu que a série de dados (entrada) do modelo causava altíssima distorção nos resultados (saída) se ela estivesse equivocada.

3.4.2 Treinamento e Validação

Como um dos objetivos do trabalho é prever a demanda de materiais elétricos na cidade de São Paulo, a base de dados do Censo de 2010 será utilizada. O Censo Demográfico possui informações que também estão presentes na POF (Idade, Gênero, Renda, Alfabetização e Raça) e estas informações são sobre toda a população da região – não sendo uma amostra, o que traz um alto grau de detalhamento geográfico. Tendo finalizado a análise sobre as características dos indivíduos que consumiram produtos da cesta delimitada, podem-se utilizar dados sobre os cidadãos de São Paulo (município) e agregar estas informações aos distritos da cidade.

Como comentado anteriormente, o aprendizado em máquina visa classificar os resultados corretamente, a partir dos dados que lhe foram fornecidos. Para que o algoritmo utilizado pela ferramenta classifique o "consumo" ou "não consumo" precisamente, é necessário aperfeiçoar os classificadores através da validação, assim possibilitando ajustes. Segundo Silva (2014), a parte de treinamento da máquina corresponde por 70% dos dados da base, enquanto que a validação corresponde por 30% dos dados da base. Ou seja, 70% das linhas de código serão destinadas para o treinamento e 30% das linhas de código para a validação.

Já que um dos objetivos do trabalho é encontrar a demanda a partir de classificadores que apresentam baixo Erro Quadrático Médio. Para isso, a pesquisa utilizará dois parâmetros para medir o treinamento e a validação. O primeiro deles é o parâmetro de regularização (Parâmetro *C*), que mede o grau de importância dos erros de classificação gerados. O segundo é o parâmetro sigma (Parâmetro σ), que verifica o nível de qualidade da previsão feita pelo modelo. Entretanto, existem dois fenômenos estatísticos (*overfitting* e *underfitting*) que podem acontecer com a função de aproximação do aprendizado em máquina. Quando estes fenômenos acontecem, o resultado poderá levar a conclusões equivocadas.

O *overfitting* ocorre quando um modelo treina a base de dados demasiadamente. Os "ruídos" ou flutuações aleatórias podem ser interpretados e aprendidos como um conceito pelo modelo. O problema é que nem sempre esses conceitos se aplicam aos novos dados, afetando negativamente a capacidade de generalização da máquina. No *undefitting*, o modelo de predição da máquina não é capaz de modelar o treinamento e nem generalizar para os novos dados (BROWNLEE, 2016).

De acordo com Silva (2014), o *overfitting* ocorre quando existe um alto grau de importância dos erros junto com um baixo grau na qualidade da previsão. O *underfitting* acontece quando se encontra um baixo grau de importância dos erros conjuntamente com um alto grau na qualidade de previsão – conforme o quadro abaixo.

Quadro 1:*Overfitting* e *Underfitting*

Grau de importância dos erros	Qualidade da previsão	
Alto C	Baixo σ	**Overfitting**
Baixo C	Alto σ	**Underfitting**

Fonte: Elaborado pelo autor a partir de SILVA (2014)

Para equilibrar o modelo de previsão, será selecionado o par de parâmetros que apresentar o menor Erro Quadrático Médio (EQM). Sendo que yi representa o valor previsto pela máquina; ŷi representa o valor real observado; e n representa o número de observações feitas – conforme a equação abaixo.

$$EQM = \sum_{i} \frac{(y_i - \hat{y}_i)^2}{n}$$

"Minimizar o erro quadrático médio fará com que ŷ seja o estimador de máxima verossimilhança para y" (FACURE, 2017). Ou seja, quanto menor for o EQM, mais próximo da realidade será o resultado previsto pela máquina.

3.5 Aplicação do método na cidade de São Paulo

Dentro da programação utilizada, as variáveis foram agrupadas em classes, conforme as classes que o IGBE utilizou como padrão no Censo 2010 (IBGE, 2010). O Quadro 2 representa as classes utilizadas para cada variável.

Quadro 2: Variáveis e classes

VARIÁVEL	CLASSE
Alfabetização	Pessoas Alfabetizadas
Idade	Pessoas de 0 a 9 anos de idade
	Pessoas de 10 a 19 anos de idade
	Pessoas de 20 a 29 anos de idade
	Pessoas de 30 a 39 anos de idade
	Pessoas de 40 a 49 anos de idade
	Pessoas de 50 a 59 anos de idade
	Pessoas de 60 a 69 anos de idade
	Pessoas de 70 anos de idade ou mais
Gênero	Homens
	Mulheres
Renda	Renda 1: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de até 1/8 salário mínimo (R$ 63,75)
	Renda 2: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de mais de 1/8 a 1/4 salário mínimo (R$ 63,76 até R$ 127,50)
	Renda 3: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de mais de 1/4 a 1/2 salário mínimo (R$ 127,51 até R$ 255,00)
	Renda 4: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de mais de 1/2 a 1 salário mínimo (R$ 255,01 até R$ 510,00)
	Renda 5: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de mais de 1 a 2 salários mínimos (R$ 510,01 até R$ 1.020,00)
	Renda 6: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de mais de 2 a 3 salários mínimos (R$ 1.020,01 até R$ 1.530,00)
	Renda 7: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de mais de 3 a 5 salários mínimos (R$ 1.530,01 até R$ 2.550,00)
	Renda 8: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de mais de 5 a 10 salários mínimos (R$ 2.550,00 até R$ 5.100,00)
	Renda 9: Domicílios particulares com rendimento nominal mensal domiciliar *per capita* de mais de 10 salários mínimos (acima de R$ 5.100,00)
Raça	Branca
	Preta
	Amarela
	Parda
	Indígena

Fonte: Adaptado de SILVA (2014) a partir do Censo Demográfico 2010

Cabe ressaltar que o IBGE tomou R$ 510,00 o valor de 1 salário mínimo, já que este era o valor em 2010. Em 2018, o salário mínimo é de R$ 954,00 (G1, 2018). Todas as 26 classes na tabela acima são as variáveis de cada um dos 18.362 setores censitários de São Paulo.

Após rodar o código no programa R Studio, a validação irá retornar em valores muito pequenos, principalmente quando se pensa na demanda total de um produto para uma cidade com população superior a 10 milhões de habitantes. Mas estes valores serão referentes ao gasto padronizado (GP), o qual é definido pela seguinte função:

$$GP = \frac{\left(G - \overline{G}\right)}{\sigma}$$

O gasto padronizado é igual a diferença entre o gasto projetado para o setor censitário e o gasto médio projetado para todos os setores censitários, ponderados pelo desvio padrão dos gastos projetados. Os quantis, representados no quadro 3, foram criados a partir de uma função do R Studio chamada "*classIntervals*" e estão no estilo "*quantile*".

Quadro 3: Intervalos de valores e classificação de demanda

COR	INTERVALO DE VALORES	CLASSIFICAÇÃO
	-0,4595 até -0,1389	Demanda Muito Baixa
	-0,1389 até -0,0024	Demanda Baixa
	-0,0024 até +0,1725	Demanda Média
	+0,1725 até +0,3494	Demanda Alta
	+0,3494 até +0,7823	Demanda Muito Alta

Fonte: Elaboração Própria

Em vermelho, caracterizando uma demanda muito baixa, estão os distritos com GP entre "-0,4595 até -0,1389"; em laranja, caracterizando uma demanda baixa, estão os distritos com GP entre "-0,1389 até -0,0024"; em amarelo, caracterizando uma demanda média, estão os distritos com GP entre "-0,0024 até +0,1725"; em verde, caracterizando uma demanda alta, estão os distritos com GP entre "+0,1725 até +0,3494"; em azul, caracterizando uma demanda muito alta, estão os distritos com GP entre +0,3494 até +0,7823".

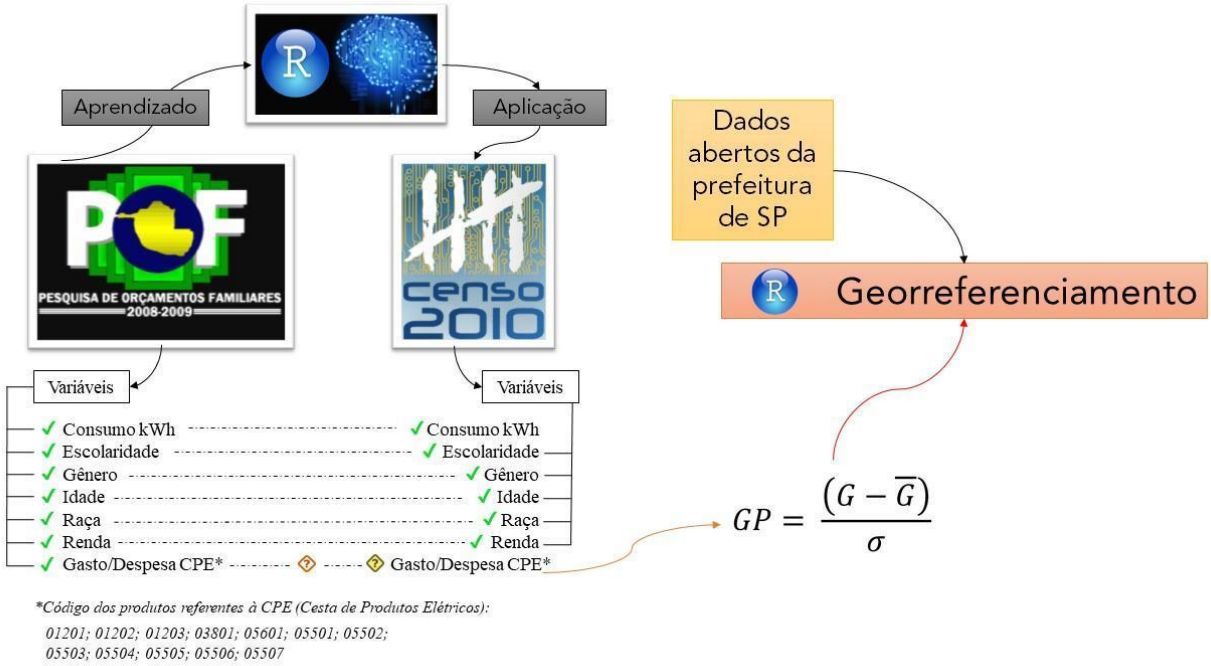
4. ANÁLISE DOS RESULTADOS

Na presente seção, serão analisados os resultados do trabalho. Eles serão divididos em resultados por variável de pesquisa e depois os resultados gerais para os distritos de São Paulo.

4.1 RESULTADOS POR VARIÁVEIS

Os resultados apresentados neste trabalho foram obtidos através da programação no programa R Studio, a qual está disponível no apêndice. O modelo proposto pode ser ilustrado na Figura 6.

Figura 6: Modelo metodológico da pesquisa



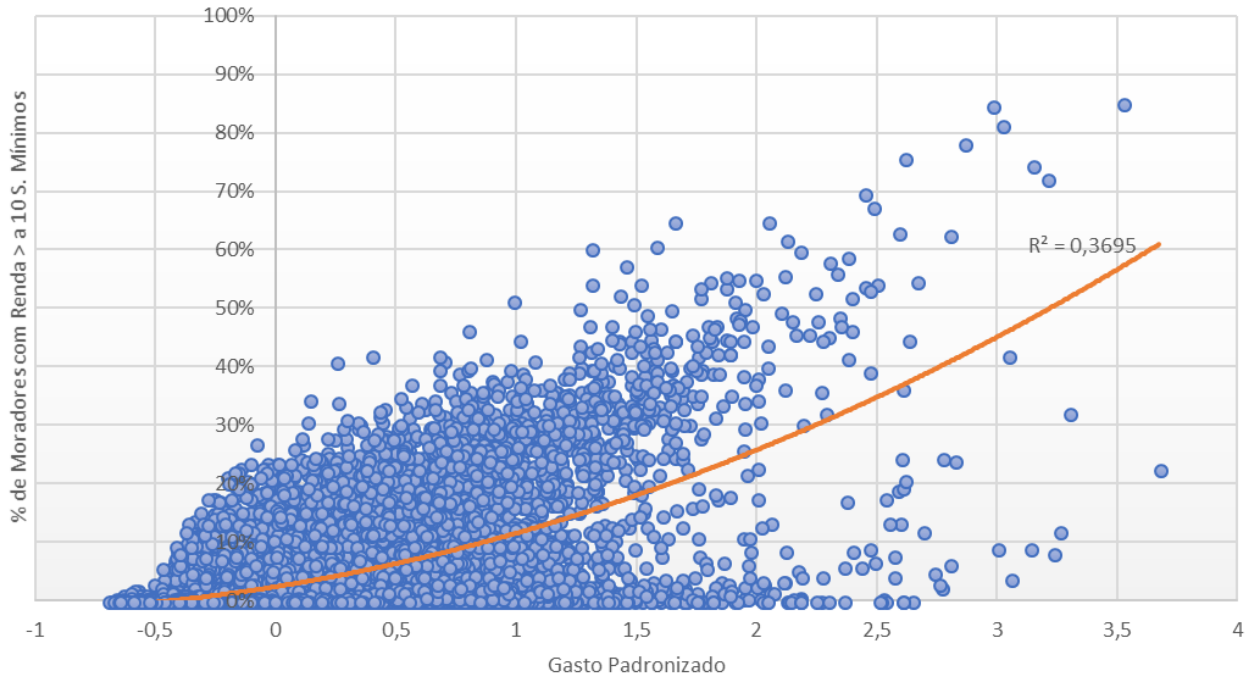
Fonte: Desenvolvido pelo autor a partir de SILVA (2014)

O padrão de compra dos produtos da CPE é definido por cinco variáveis: escolaridade, gênero, idade, raça e renda. Estas mesmas variáveis são encontradas no Censo 2010, mas em uma escala muito superior e que representa o universo populacional de São Paulo. A máquina foi, portanto, ensinada através das informações dispostas na POF e executou a previsão de gastos com

produtos elétricos com os dados do Censo – através de uma função chamada "ksvm", do pacote "kernlab".

No hiperplano (Figura 5), apresentado anteriormente, pode-se perceber que a máquina busca distinguir padrões no comportamento de consumo e classificá-los como potenciais em duas categorias distintas. O gráfico 1, representado abaixo, em duas dimensões (Gasto por Renda), ilustra um hiperplano relativo a pesquisa realizada.

Gráfico 1: Gasto padronizado por Concentração de renda acima de 10 salários mínimos



Fonte: Desenvolvido pelo autor

O eixo vertical do gráfico está representado pela porcentagem de moradores (dentro de cada setor censitário) que recebem 10 ou mais salários mínimos – a categoria de renda mais elevada da pesquisa. O eixo horizontal mostra o gasto padronizado para o consumo de materiais da cesta elétrica e, quanto maior for o valor deste gasto padronizado, maior será a demanda por produtos elétricos. A partir dos dados apresentados por dispersão, nota-se que, quanto maior a quantidade de indivíduos com renda superior a 10 salários mínimos, maior tende a ser o gasto com produtos elétricos. Ainda, uma menor quantidade de pessoas com renda superior a 10 salários mínimos

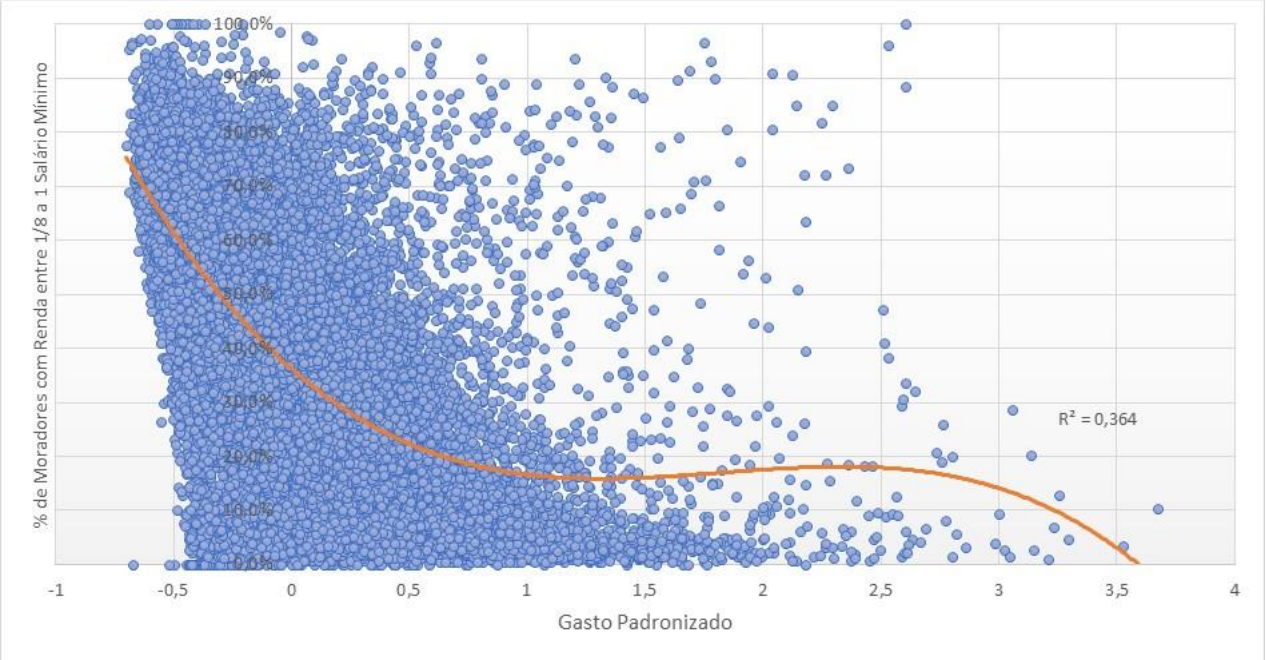
dentro de um setor censitário indica que o gasto/demanda será menor. A linha de tendência (em laranja) confirma este padrão.

A esfera renda foi, com certeza, decisiva no comportamento de compra. Claramente, não é a única variável que explica este fenômeno. A relação da porcentagem de moradores no setor censitário com renda na faixa salarial 9 (acima de 10 salários mínimos) mostra que os indivíduos são mais propensos ao consumo. Contudo, há uma quantidade considerável de indivíduos com um consumo similar, mesmo com renda inferior.

Nove dos 15 distritos mais ricos estão entre os que apresentaram os 15 maiores gastos padronizados, sendo que nenhum dos outros seis distritos possuíam renda inferior a renda mediana dos distritos. Entre os distritos que apresentaram as 15 menores demandas por produtos elétricos, 11 distritos estão na lista dos 15 distritos mais pobres em renda. Além disso, todos os outros quatro distritos possuem renda abaixo a mediana dos distritos de São Paulo (cidade).

O gráfico de dispersão (Gráfico 2) mostra a relação entre a porcentgem de moradores no setor censitário que possuem um rendimento entre as faixas salariais 1 e 4 (de 1/8 de salário mínimo até 1 salário mínimo).

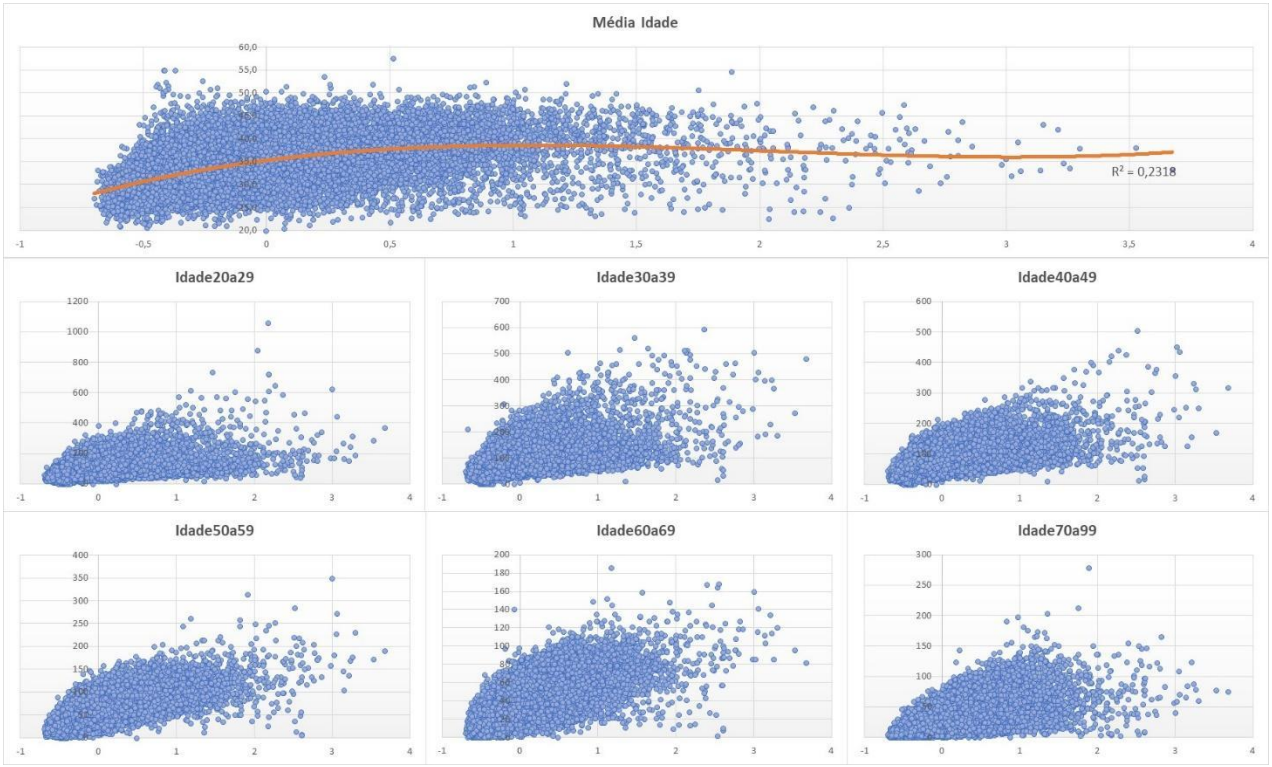
Gráfico 2: Gasto padronizado por Concentração de renda nas faixas de 1/8 a 1 salário mínimo



Fonte: Elaborado pelo autor

Percebe-se que houve um aumento no gasto padronizado conforme a porcentagem de renda entre as faixas 1 e 4 foi diminuindo. Quanto maior o número de indivíduos que possuem renda inferior a 1 salário mínimo, menor será a demanda. Também, quanto menor for a quantidade de moradores que possuem renda menor do que 1 salário mínimo, maior será a demanda pelos produtos, pois os moradores desses distritos pertencem a faixas de renda superiores. Mesmo assim, há uma certa quantidade de compradores um gasto padronizado relativamente alto e, ao mesmo tempo, pertencem a faixas de renda que não são consideradas elevadas.

Gráfico 3: Dispersão de Idade por Gasto

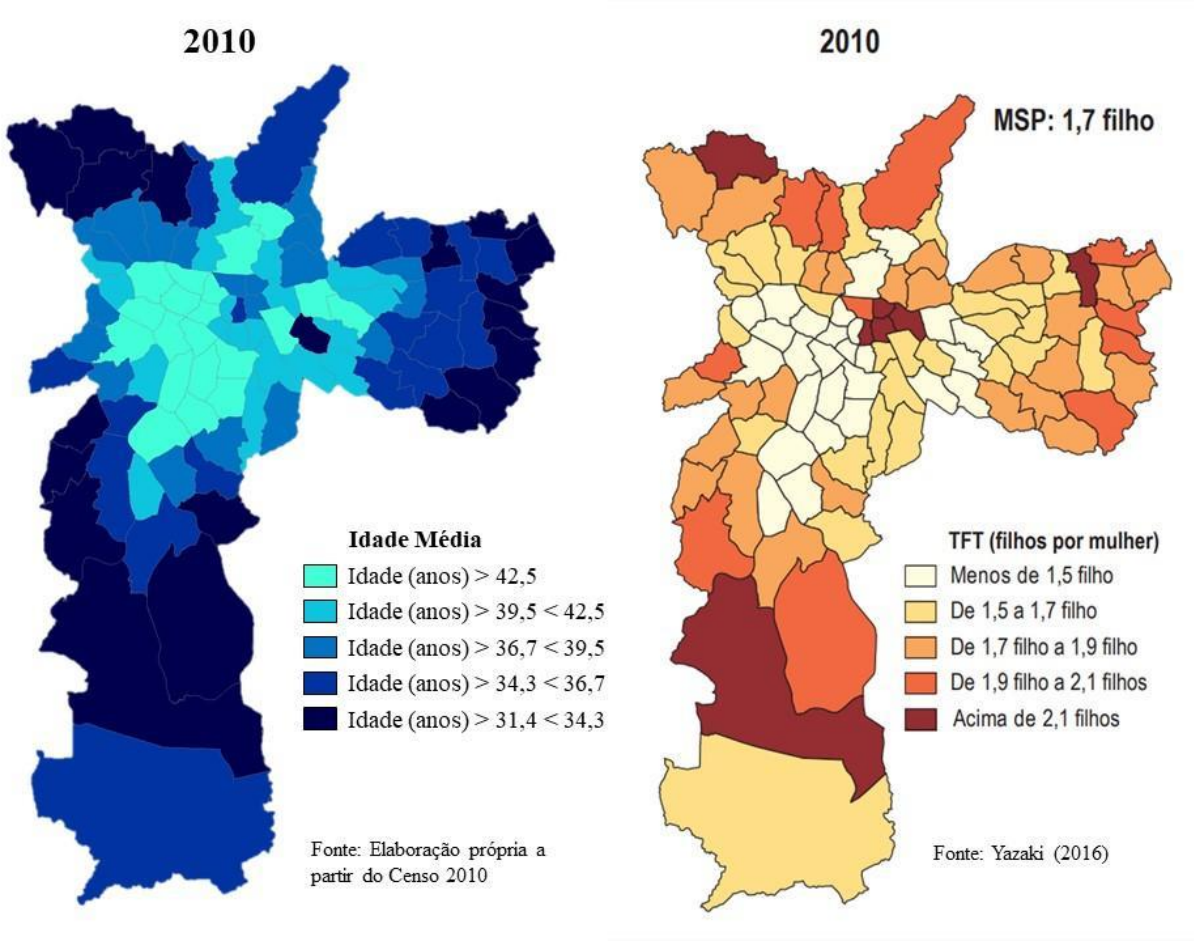


Fonte: Elaborado pelo autor

Observando o gráfico 3, percebe-se que os setores censitários com idade média mais alta apresentam uma tendência maior ao consumo dos produtos da cesta de materiais elétricos. Os distritos e setores censitários com idade média menor são, em média, locais onde a renda per capita é mais baixa. Este fato pode ser explicado pela taxa de fecundidade, que é número de filhos por mulher. Yazaki, em 2016, publicou um estudo sobre as diferenças regionais de fecundidade em

São Paulo e em 2010, no ano do Censo, as taxas de fecundidade eram maiores nos distritos de menor renda per capita, conforme a figura 7.

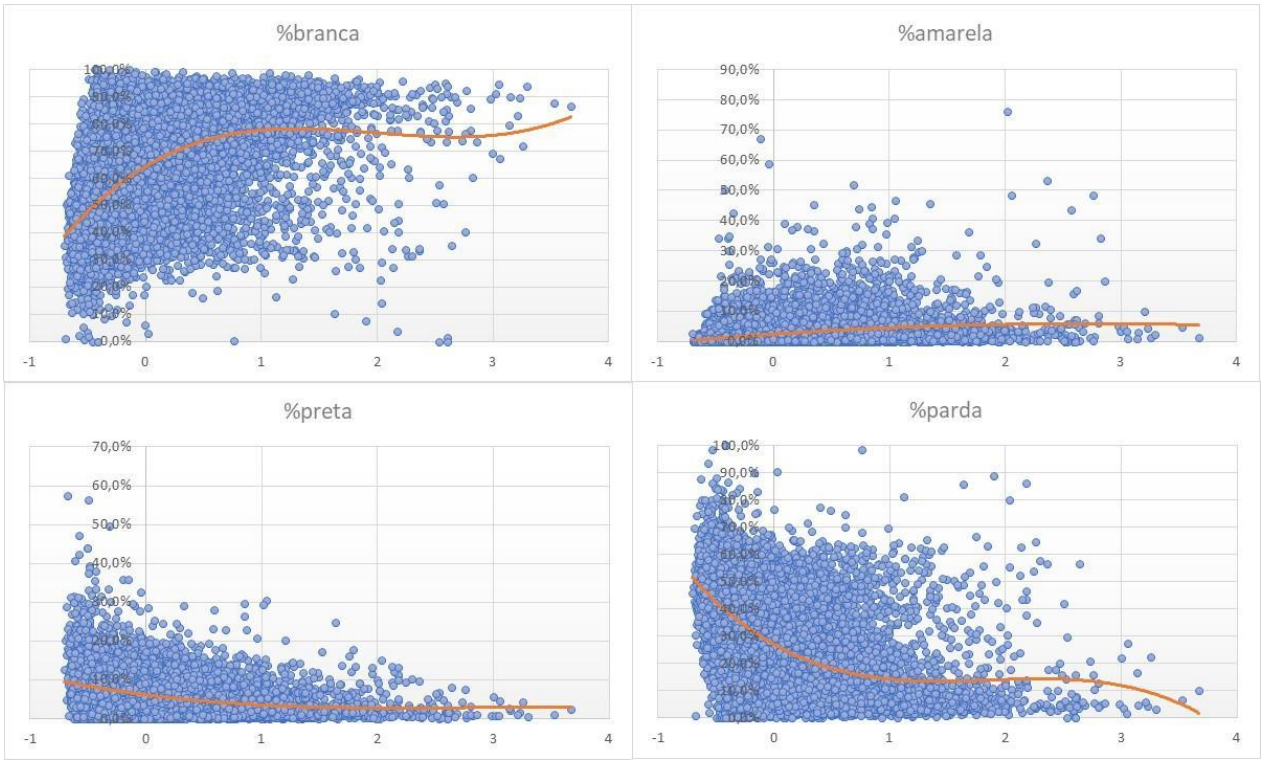
Figura 7: Idade média e taxa de fecundidade em São Paulo



Fonte: Elaboração própria a partir do Censo 2010 e Yazaki (2016)

É perceptível na figura 7 que os distritos com idade média superior são aqueles que possuem menor taxa de fecundidade. Consequentemente, estes são os mesmos distritos que possuem uma renda per capita mais alta e, portanto, maior demanda por produtos elétricos – como visto anteriormente, a renda é um fator determinante no consumo dos produtos da cesta.

Gráfico 4: Dispersão de Raça por Gasto

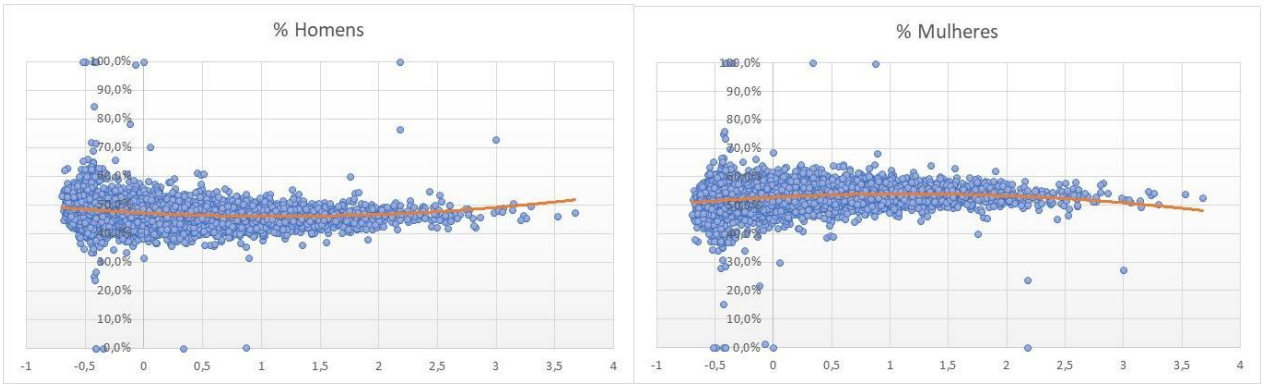


Fonte: Elaborado pelo autor

A variável raça é dividida em cinco categorias: branca, parda, preta, amarela e indígena. A raça indígena não foi representada no gráfico 4 porque há um número muito pequeno de observações e apresenta muitos *outliers*. Pode-se perceber que, quanto maior a porcentagem de brancos em um setor censitário, maior tende a ser a demanda pelos produtos determinados. Os brancos concentram uma renda média maior e este cenário pode ser percebido por todo o território brasileiro, onde a desigualdade entre as raças ainda é muito grande. Esta desigualdade existe desde o processo de colonização do Brasil.

Ainda, o gráfico 4 mostra que o gasto com materiais elétricos vai aumentando a medida que a proporção de pretos e pardos diminui. Por último, a variação na proporção de amarelos não indica muitas mudanças no gasto esperado, mas existe um fato interessante no que diz respeito ao consumo previsto para a raça amarela. Se a quantidade de amarelos for relativamente alta, a tendência é que o consumo esperado fique nas faixas médias de demanda. No entanto, se a proporção de amarelos for baixa, o consumo projetado tende a baixo ou alto, diferentemente das outras raças.

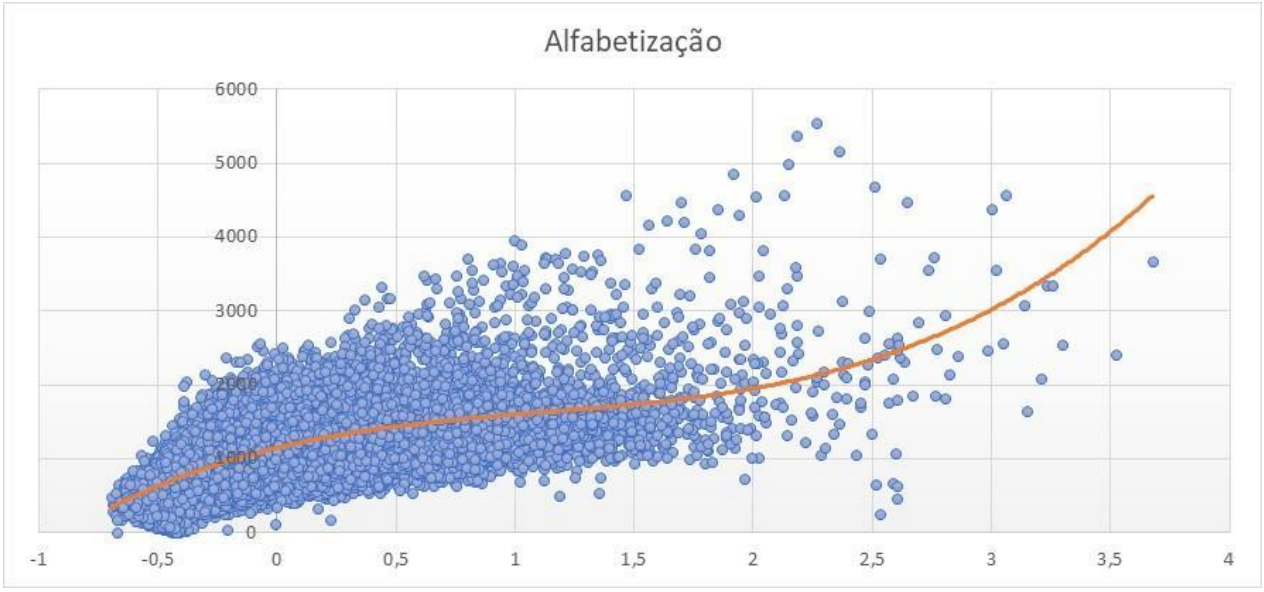
Gráfico 5: Dispersão de Sexo por Gasto



Fonte: Elaborado pelo autor

Após analisar o gráfico 5, percebe-se que o sexo não é um fator determinante na demanda por materiais elétricos. As linhas de tendência (em laranja), mostram que a demanda seria maior para as mulheres, mas isso pode ser explicado pelo número total de mulheres em relação a homens. Em São Paulo, 52,6% da população de 2010 era do sexo feminino – segundo dados do resultado do universo do Censo 2010.

Gráfico 6: Dispersão de Alfabetização por Gasto



Fonte: Elaborado pelo autor

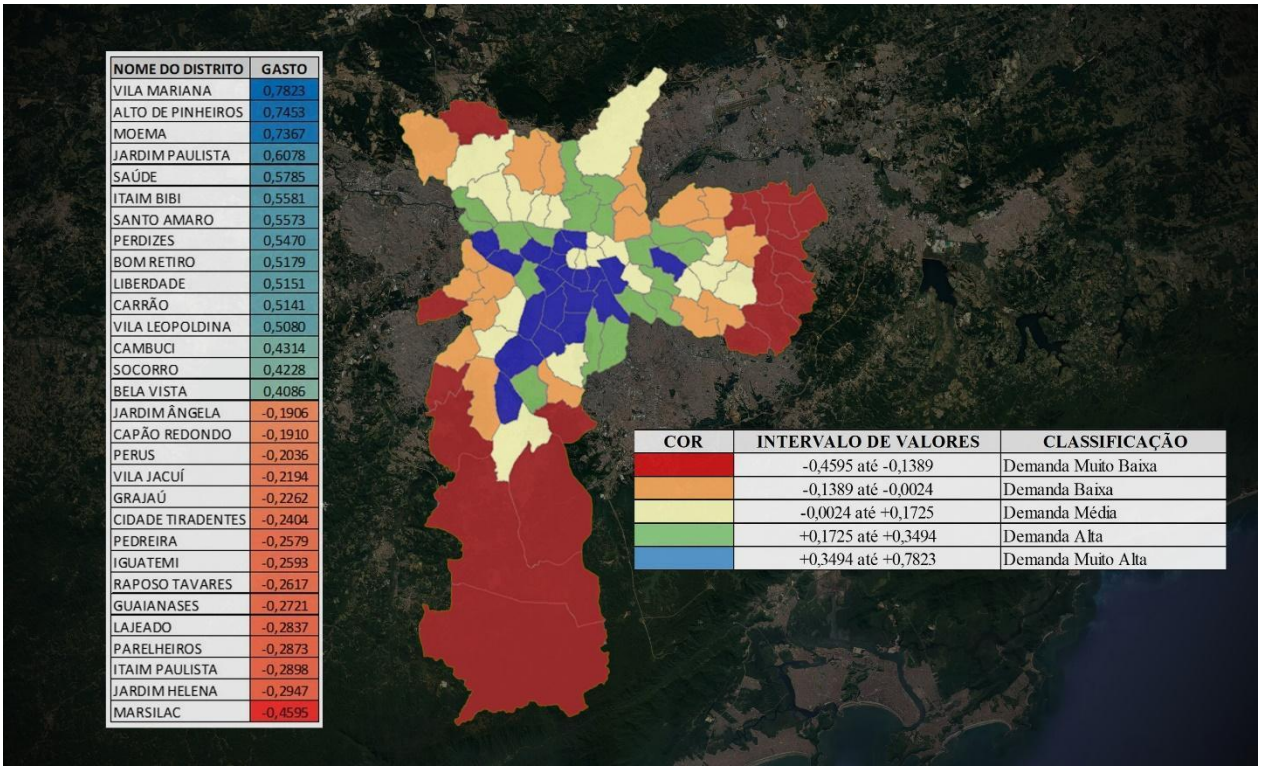
A variável alfabetização está fortemente correlacionada com a renda, pois quanto maior a renda per capita, maior também será a taxa de alfabetização. Portanto, os distritos e setores censitários com maior número de alfabetizados terão a demanda mais alta por materiais elétricos, uma vez que regiões com maior índice de analfabetos terão uma demanda mais baixa. Este era um resultado já esperado pela pesquisa, pois indivíduos com maior escolarização possuem uma renda per capita maior e isso pode ser percebido ao redor do mundo.

4.2 RESULTADOS GERAIS

Foi obtido um gasto padronizado entre -0,4595 e +0,7823 para cada distrito da cidade de São Paulo. Estes valores representam a média entre todos os setores censitários pertencentes a um mesmo distrito. Os setores censitários, por sua vez, apresentam valores entre -0,6990 e +3,6730.

Na Figura 8, representada abaixo, estão representados os 96 distritos da cidade de São Paulo. Vale ressaltar que os valores do gasto padronizado são representados pela diferença entre o Gasto Registrado no Setor Censitário e o Gasto Médio dos Setores Censitários, ponderada pelo Desvio Padrão dos Gastos.

Figura 8 – Gasto padronizado para os distritos



Fonte: Elaborado pelo autor

O distrito de Vila Mariana foi o que apresentou o maior potencial de demanda, seguido por Alto de Pinheiros e Moema – todos com GP superior a 0,7. Na ponta de baixo, o distrito de Marsilac (mais ao sul de São Paulo) é o local com menor potencial de demanda. A Tabela 3 mostra a média de todos os distritos individualmente.

Para que os valores em "Reais" sejam encontrados, será necessário multiplicar o valor projetado pelo Gasto Padronizado pelo desvio-padrão dos gastos e, ainda, somar a média do gasto. No apêndice B, pode-se ver que a média e o desvio-padrão são calculados antes do treinamento dos dados. A média e o desvio-padrão são feitos sobre o "treina.df.scaled", que pertence as bases de dados da POF.

Os valores gerados em reais pelo modelo, com a remoção dos pesos amostrais, alcançaram 2,9 milhões de reais (valores correntes de 2009), o que representaria R$ 5 milhões a valores de 2018, corrigido pelo IPCA. Quando multiplicamos os pesos amostrais médios (fatores de expansão) da amostra com o resultado encontrado – o mercado teria um volume anual de R$ 3,6 bilhões, o que superestima o tamanho do mercado em São Paulo (cerca de R$ 900 milhões). No Brasil, o mercado de materiais elétricos de instalação movimentou cerca de R$ 7,5 bilhões em 2017

(ABINEE, 2018). Os fatores de expansão calculados pelo IBGE não mostraram a realidade vista no mercado de materiais elétricos.

Tabela 3: Gastos padronizados dos distritos

DISTRITO	NOME DO DISTRITO	GASTO	DISTRITO	NOME DO DISTRITO	GASTO
355030890	VILA MARIANA	0,7823	355030856	PARI	0,1044
355030802	ALTO DE PINHEIROS	0,7453	355030829	FREGUESIA DO Ó	0,0973
355030832	MOEMA	0,7367	355030850	LIMÃO	0,0860
355030845	JARDIM PAULISTA	0,6078	355030842	JARAGUÁ	0,0482
355030877	SAÚDE	0,5785	355030883	VILA ANDRADE	0,0404
355030835	ITAIM BIBI	0,5581	355030838	JABAQUARA	0,0397
355030871	SANTO AMARO	0,5573	355030857	PARQUE DO CARMO	0,0231
355030860	PERDIZES	0,5470	355030854	MORUMBI	0,0063
355030809	BOM RETIRO	0,5179	355030881	TREMEMBÉ	-0,0017
355030849	LIBERDADE	0,5151	355030810	BRÁS	-0,0024
355030820	CARRÃO	0,5141	355030812	BUTANTÃ	-0,0049
355030888	VILA LEOPOLDINA	0,5080	355030813	CACHOEIRINHA	-0,0111
355030814	CAMBUCI	0,4314	355030889	VILA MARIA	-0,0164
355030879	SOCORRO	0,4228	355030837	ITAQUERA	-0,0197
355030807	BELA VISTA	0,4086	355030876	SAPOPEMBA	-0,0394
355030869	SANTA CECÍLIA	0,4073	355030818	CANGAIBA	-0,0395
355030826	CONSOLAÇÃO	0,3994	355030839	JAÇANÃ	-0,0444
355030853	MOOCA	0,3966	355030803	ANHANGUERA	-0,0643
355030834	IPIRANGA	0,3622	355030873	SÃO MATEUS	-0,0691
355030801	ÁGUA RASA	0,3620	355030894	VILA SÔNIA	-0,0694
355030815	CAMPO BELO	0,3494	355030822	CIDADE ADEMAR	-0,0719
355030870	SANTANA	0,3484	355030846	JARDIM SÃO LUÍS	-0,0775
355030880	TATUAPÉ	0,3452	355030841	JAGUARÉ	-0,0845
355030806	BARRA FUNDA	0,3340	355030817	CAMPO LIMPO	-0,0881
355030885	VILA FORMOSA	0,3340	355030892	VILA MEDEIROS	-0,0971
355030848	LAPA	0,3066	355030864	PONTE RASA	-0,1065
355030886	VILA GUILHERME	0,2974	355030828	ERMELINO MATARAZZO	-0,1224
355030893	VILA PRUDENTE	0,2963	355030811	BRASILÂNDIA	-0,1310
355030872	SÃO LUCAS	0,2922	355030867	RIO PEQUENO	-0,1389
355030816	CAMPO GRANDE	0,2792	355030875	SÃO RAFAEL	-0,1507
355030840	JAGUARA	0,2745	355030847	JOSÉ BONIFÁCIO	-0,1517
355030891	VILA MATILDE	0,2515	355030874	SÃO MIGUEL	-0,1556
355030862	PINHEIROS	0,2459	355030884	VILA CURUÇÁ	-0,1819
355030827	CURSINO	0,2047	355030843	JARDIM ÂNGELA	-0,1906
355030851	MANDAQUI	0,2027	355030819	CAPÃO REDONDO	-0,1910
355030895	SÃO DOMINGOS	0,1985	355030861	PERUS	-0,2036
355030882	TUCURUVI	0,1871	355030887	VILA JACUÍ	-0,2194
355030859	PENHA	0,1830	355030830	GRAJAÚ	-0,2262
355030868	SACOMÃ	0,1750	355030825	CIDADE TIRADENTES	-0,2404
355030821	CASA VERDE	0,1725	355030858	PEDREIRA	-0,2579
355030808	BELÉM	0,1677	355030833	IGUATEMI	-0,2593
355030804	ARICANDUVA	0,1445	355030865	RAPOSO TAVARES	-0,2617
355030805	ARTUR ALVIM	0,1439	355030831	GUAIANASES	-0,2721
355030878	SÉ	0,1410	355030896	LAJEADO	-0,2837
355030866	REPÚBLICA	0,1240	355030855	PARELHEIROS	-0,2873
355030863	PIRITUBA	0,1128	355030836	ITAIM PAULISTA	-0,2898
355030823	CIDADE DUTRA	0,1113	355030844	JARDIM HELENA	-0,2947
355030824	CIDADE LIDER	0,1103	355030852	MARSILAC	-0,4595

Fonte: Elaborado pelo autor

Tabela 4: Gastos padronizados dos distritos x Renda

NOME DO DISTRITO	GASTO	RENDA 1 A 4	RENDA 8 A 9	NOME DO DISTRITO	GASTO	RENDA 1 A 4	RENDA 8 A 9
VILA MARIANA	0,7823	5%	55%	PARI	0,1044	35%	8%
ALTO DE PINHEIROS	0,7453	6%	59%	FREGUESIA DO Ó	0,0973	31%	8%
MOEMA	0,7367	3%	71%	LIMÃO	0,0860	33%	8%
JARDIM PAULISTA	0,6078	5%	67%	JARAGUÁ	0,0482	52%	2%
SAÚDE	0,5785	10%	43%	VILA ANDRADE	0,0404	39%	29%
ITAIM BIBI	0,5581	6%	62%	JABAQUARA	0,0397	31%	17%
SANTO AMARO	0,5573	9%	45%	PARQUE DO CARMO	0,0231	47%	4%
PERDIZES	0,5470	6%	54%	MORUMBI	0,0063	16%	53%
BOM RETIRO	0,5179	30%	11%	TREMEMBÉ	-0,0017	47%	5%
LIBERDADE	0,5151	13%	34%	BRÁS	-0,0024	27%	9%
CARRÃO	0,5141	25%	13%	BUTANTÃ	-0,0049	13%	34%
VILA LEOPOLDINA	0,5080	11%	49%	CACHOEIRINHA	-0,0111	46%	4%
CAMBUCI	0,4314	18%	25%	VILA MARIA	-0,0164	41%	6%
SOCORRO	0,4228	22%	17%	ITAQUERA	-0,0197	48%	2%
BELA VISTA	0,4086	11%	41%	SAPOPEMBA	-0,0394	54%	1%
SANTA CECÍLIA	0,4073	12%	33%	CANGAIBA	-0,0395	45%	3%
CONSOLAÇÃO	0,3994	6%	53%	JAÇANÃ	-0,0444	44%	5%
MOOCA	0,3966	13%	27%	ANHANGUERA	-0,0643	55%	1%
IPIRANGA	0,3622	23%	22%	SÃO MATEUS	-0,0691	48%	2%
ÁGUA RASA	0,3620	23%	16%	VILA SÔNIA	-0,0694	25%	27%
CAMPO BELO	0,3494	10%	50%	CIDADE ADEMAR	-0,0719	51%	4%
SANTANA	0,3484	13%	31%	JARDIM SÃO LUÍS	-0,0775	50%	2%
TATUAPÉ	0,3452	11%	32%	JAGUARÉ	-0,0845	33%	18%
BARRA FUNDA	0,3340	12%	43%	CAMPO LIMPO	-0,0881	45%	5%
VILA FORMOSA	0,3340	29%	13%	VILA MEDEIROS	-0,0971	41%	3%
LAPA	0,3066	11%	37%	PONTE RASA	-0,1065	41%	5%
VILA GUILHERME	0,2974	25%	13%	ERMELINO MATARAZZO	-0,1224	47%	2%
VILA PRUDENTE	0,2963	27%	12%	BRASILÂNDIA	-0,1310	58%	1%
SÃO LUCAS	0,2922	32%	6%	RIO PEQUENO	-0,1389	33%	18%
CAMPO GRANDE	0,2792	19%	25%	SÃO RAFAEL	-0,1507	62%	1%
JAGUARA	0,2745	29%	8%	JOSÉ BONIFÁCIO	-0,1517	49%	1%
VILA MATILDE	0,2515	32%	8%	SÃO MIGUEL	-0,1556	51%	2%
PINHEIROS	0,2459	7%	56%	VILA CURUÇÁ	-0,1819	59%	1%
CURSINO	0,2047	25%	20%	JARDIM ÂNGELA	-0,1906	64%	1%
MANDAQUI	0,2027	23%	15%	CAPÃO REDONDO	-0,1910	54%	2%
SÃO DOMINGOS	0,1985	37%	9%	PERUS	-0,2036	59%	1%
TUCURUVI	0,1871	21%	14%	VILA JACUÍ	-0,2194	57%	2%
PENHA	0,1830	29%	9%	GRAJAÚ	-0,2262	61%	1%
SACOMÃ	0,1750	36%	8%	CIDADE TIRADENTES	-0,2404	65%	0%
CASA VERDE	0,1725	29%	13%	PEDREIRA	-0,2579	57%	2%
BELÉM	0,1677	22%	18%	IGUATEMI	-0,2593	65%	0%
ARICANDUVA	0,1445	40%	5%	RAPOSO TAVARES	-0,2617	43%	6%
ARTUR ALVIM	0,1439	37%	3%	GUAIANASES	-0,2721	59%	2%
SÉ	0,1410	25%	6%	LAJEADO	-0,2837	67%	0%
REPÚBLICA	0,1240	16%	22%	PARELHEIROS	-0,2873	67%	1%
PIRITUBA	0,1128	35%	9%	ITAIM PAULISTA	-0,2898	64%	1%
CIDADE DUTRA	0,1113	46%	3%	JARDIM HELENA	-0,2947	65%	1%
CIDADE LIDER	0,1103	44%	3%	MARSILAC	-0,4595	73%	1%

Fonte: Elaborado pelo autor

Visto que o percentual de moradores que se encontram nas faixas de renda entre 5 e 7 não consegue explicar precisamente as variações do gasto padronizado, ele não foi incluído na tabela acima. Isso ocorre porque tanto os distritos com rendas elevadas como os com rendas mais baixas apresentam concentrações similares. Por exemplo, o distrito de Moema, cuja concentração das rendas 8 e 9 é a mais elevada, possui uma concentração das rendas 5, 6 e 7 muito similar com o distrito de Marsilac (um dos distritos mais pobres). As concentrações das rendas 1, 2, 3, 4, 8 e 9 são totalmente opostas quando comparamos estes dois distritos. Os setores censitários/distritos que apresentam um grande agrupamento de moradores entre as faixas 1 e 4 terão, muito provavelmente, um GP baixo, enquanto que os setores censitários/distritos que apresentam grande agrupamento nas faixas 8 e 9 de renda terão um GP alto. Os gráficos de dispersão 1 e 2 confirmam esta lógica.

É perceptível que há uma grande relação entre o gasto padronizado com a variável renda. A formatação condicional das cores (azul, verde, amarelo, laranja e vermelho) foi a mesma que foi utilizada na classificação da demanda, conforme o Quadro 3.

5. CONSIDERAÇÕES FINAIS

Através dos resultados da presente pesquisa, é possível gerar um diferencial competitivo capaz de tornar as decisões dos gestores mais qualificadas, visto que o modelo indica as áreas geográficas que possuem um maior potencial de compra dos produtos selecionados. O processo para a obtenção de tais dados começa pela escolha do mercado a ser estudado e a consequente verificação dos produtos que têm informações coletadas na Pesquisa de Orçamentos Familiares. O IBGE divulga duas planilhas que mostram os 13.778 produtos gerais e a segunda planilha com os 1631 produtos relacionados ao consumo alimentar. Dentro desses produtos, há uma imensa variedade de mercados que, se estudadas, podem elencar múltiplas oportunidades de negócios. Há informações sobre o mercado de produtos elétricos, educação, televisão, manutenção doméstica, alimentação, transportes, saúde, lazer, financeiro/crédito, imóveis, dentre outras áreas.

Os resultados obtidos pelo modelo comprovam que é possível utilizar Máquinas de Suporte Vetorial e dados secundários. Sejam estes dados coletados pelo governo, diferentes empresas ou até mesmo pela própria empresa e disponíveis para consulta, é sim viável realizar um estudo dentro desse escopo. A possibilidade de modificar o código dos produtos escolhidos para análise é um grande benefício para próximos trabalhos, pois o método serve para calcular mais de 13 mil produtos diferentes (número de produtos cadastrados na Pesquisa de Orçamentos Familiares), já que o R Studio buscará pelo produto ou produtos selecionados em todas as bases importadas.

Os dados utilizados na pesquisa vêm de 2008 a 2009, portanto é importante ressaltar que eles estão um pouco defasados. No entanto, como o código utilizado no programa R Studio é de fácil adaptação, o mesmo poderá ser utilizado como base quando o IBGE divulgar os resultados sobre a nova Pesquisa de Orçamentos Familiares e o novo Censo Demográfico. A POF está em andamento durante 2017/2018 e terá uma amostra ainda maior - 75 mil domicílios em mais de 1900 municípios. A previsão para a divulgação dos resultados é em meados de 2019. Já o próximo Censo, terá a sua coleta em 2020 e com previsão de divulgação dos resultados para o final do mesmo ano. Pode-se fazer uma atualização dos dados de 2008/2009 através de indicadores relacionados a inflação para se chegar em um valor mais próximo da realidade de 2018. Como o trabalho gerou um gasto padronizado, baseado no padrão de consumo dos moradores, é válido utilizar estes gastos padronizados, pois eles representam o hábito de consumo dos indivíduos. Esse comportamento é exemplo clássico da teoria da demanda, cuja ideia discute as escolhas de compras dentre diversos

bens que o orçamento permite adquirir. Pode-se realmente ver que o orçamento familiar impactou diretamente da demanda de cada um dos indivíduos da cidade de São Paulo e os distritos mais ricos apresentaram uma demanda claramente superior quando comparamos aos distritos mais pobres.

No Brasil, a previsão de demanda feita através de *Support Vector Machines* ainda não foi amplamente explorada. Mas há estudos que concluem que os modelos de SVM possuem uma acurácia superior a outros modelos, principalmente no que tange ao erro médio absoluto da previsão, além de ser bastante confiável na previsão (FILHO et al. 2017).

Conforme visto na revisão da literatura sobre geomarketing, as organizações querem estar localizadas estrategicamente no local que lhe gere um resultado financeiro melhor. Os resultados levantados por esta pesquisa conseguem indicar a localização dos clientes com maior potencial de consumo. Estes locais estão representados na tabela 3 e os clientes com maior potencial de consumo são moradores da região central da cidade de São Paulo. Essa é uma informação relevante para o setor de materiais elétricos. No entanto, para o setor de materiais elétricos, a conveniência da localização para o consumo, isto é, estar próximo dos consumidores, não chega a ser uma relação positiva clara, segundo Silva (2017). Para tal, seria importante um estudo complementar para determinar qual o grau de influência das lojas do setor e o quão dispostos a se deslocar estariam os consumidores para efetuarem suas compras.

Com relação as limitações do trabalho, apesar da POF disponibilizar um grande número de produtos, há alguns mercados em que possam faltar objetos específicos da pesquisa. Neste trabalho, tomando como exemplo, a pesquisa cadastrou "Fios e material elétrico **em geral**". Portanto, há produtos que poderiam ser acrescentados na pesquisa do IBGE e suas omissões acabam por diminuir a riqueza de informações. Interruptores, plugues, tomadas são alguns dos produtos que poderiam estar dentro da Pesquisa de Orçamentos Familiares e trariam informações ainda mais valiosas.

REFERÊNCIAS

ABINEE. **Desempenho setorial.** 2017 Disponível em: <http://www.abinee.org.br/abinee/decon/decon15.htm> Acesso em: 08/10/2017.

ABINEE. **Desempenho setorial**. 2018. Disponível em: < http://www.abinee.org.br/abinee/decon/decon15.htm > Acesso em: 17/05/2018

ALBUQUERQUE e BRUNNENBERG. **Marketing Science. Estimating Demand Heterogeneity Using Aggregated Data: An Application to the Frozen Pizza Category.** Marketing Science Disponível em: <http://dx.doi.org/10.1287/mksc.1080.0403> Acesso em 08/10/2017.

ARANHA, Francisco; FIGOLI, Susana. **Geomarketing: memórias de viagem.** São Paulo, p. 1-73, 2001.

BIZ FLUENT. **Market share vs market penetration.** Disponível em: <https://bizfluent.com/info-8588206-market-share-vs-market-penetration.html.> Acesso em: 18/11/2017.

BHALLA, Deepanshu. Support Vector Machine Simplified using R. 2017. Disponível em: <http://www.listendata.com/2017/01/support-vector-machine-in-r-tutorial.html> Acesso em: 03/01/2018

BROWNLEE, Jason. **Overfitting and Underfitting With Machine Learning Algorithms**. 2016. Disponível em: <https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/> Acesso em: 03/01/2018

CHEN et al. **Load forecasting using support vector Machines: a study on EUNITE competition 2001**. 2004.

CHINTAGUNTA, Pradeep K. **Endogeneity and Heterogeneity in a Probit Demand Model: Estimation Using Aggregated Data.** 2001 **.**

CHOI, YU e AU. **A hybrid SARIMA wavelet transform method for sales forecasting**. 2010.

FACURE, Matheus. **Funções custo para regressão**: Entendendo as funções custo ou objetivo e como AM difere de otimização pura. Disponível em: <https://matheusfacure.github.io/2017/03/03/func-custo-regr/#EQM> Acesso em: 02/01/2018

FILHO et al. **Uma comparação de técnicas de regressão para a previsão de consumo de energia residencial no cenário nacional**. 2017. Disponível em: <http://www.sbpo2017.iltc.br/pdf/169301.pdf> Acesso em: 10/05/2018

G1. **Salário mínimo em 2018: veja o valor**. 2018. Disponível em: <https://g1.globo.com/economia/noticia/salario-minimo-em-2018-veja-o-valor.ghtml>. Acesso em: 23/04/2018

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Análise do consumo alimentar pessoal no Brasil**. 2011. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv50063.pdf> Acesso em 03/01/2018

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Características gerais da população, religião e pessoas com deficiência**. 2012

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Censo Demográfico de 2010**: Características da população e dos domicílios. 2011

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Comitê de Estatísticas Sociais.** Disponível em:<https://censo2010.ibge.gov.br/200-comite-de-estatisticas-sociais/base-de-dados/1145-pesquisa-de-orcamentos-familiares> Acesso em: 23/10/2017.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Estatísticas (2016)**. Disponível em: <https://downloads.ibge.gov.br/downloads_estatisticas.htm> Acesso em: 29/11/2017

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Estudo da Modalidade de Censo Demográfico Contínuo – EMCD**. 2010. Disponível em: <https://ww2.ibge.gov.br/home/estatistica/populacao/censo_continuo/modelo_operacional/amostra.shtm> Acesso em: 18/12/2017

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Malha Territorial.** Disponível em: <https://ww2.ibge.gov.br/home/presidencia/noticias/imprensa/ppts/0000000483.pdf> Acesso em: 16/11/2017.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Notas técnicas Censo Demográfico 2010**. 2017. Disponível em: <https://www.ibge.gov.br/estatisticas-novoportal/multidominio/genero/9662-censo-demografico-2010.html?&t=notas-tecnicas> Acesso em: 03/01/2018

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Pesquisa de Orçamentos Familiares.** Disponível em: <https://ww2.ibge.gov.br/home/estatistica/pesquisas/pesquisa_resultados.php?id_pesquisa=25> Acesso em: 08/10/2017.

LORENA e CARVALHO. **Uma Introdução às Support Vector Machines.** 2007.

NUNES et.al. **Aplicação dos conceitos de previsão de demandas baseadas em séries temporais em uma concessionária de motocicletas.** 2009.

SANATI, Negin A.; SANATI, Mehri. **Growing interest in use of geographic information systems in health and healthcare research: a review of PubMed from 2003 to 2011**. 2013

SAS. Machine Learning: **What it is and why it matters**. 2016. Disponível em: <https://www.sas.com/it_it/insights/analytics/machine-learning.html> Acesso em: 05/11/2017

SILVA, Camila. **Mercado de comida japonesa no Distrito Federal: Análise das oportunidades de negócio por meio do geomarketing e máquinas de suporte vetorial**. 2014.

SILVA, Luciana. **Análise de clientes de uma siderúrgica em Goiás a partir de princípios do Geomarketing:** um estudo de caso com 173 observações. 2013.

SILVA, Rodrigo Ledur. **Modelo de atratividade geográfica no setor de materiais elétricos**. 2017

SORUCO, Ricardo Gastal. **Impactos do aquecimento global em eventos extremos no meio urbano**. 2016

Tramontina - **Reforma e Construção.** Disponível em: <http://www.tramontina.com.br/6-reforma-e-construcao/376-materiais-eletricos?ajax=0&page=1&ajax=0> Acesso em: 06/10/2017.

TAKETA, Richard. **Management and the Geographer: The Relevance of Geography in Strategic Thinking**. 1993. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.0033-0124.1993.00465.x/abstract> Acesso em: 22/12/2017

VON THÜNEN, J. H. **The isolated state**. Oxford: Pergamon Press, 1966.

YAZAKI, Lúcia. **Diferenciais regionais de fecundidade no município de São Paulo**. 2016. Disponível em: <http://www.seade.gov.br/produtos/midia/2016/06/N.2_jun2016-final.pdf> Acesso em: 19/05/2018

APÊNDICES

Apêndice A - Programação utilizada no RStudio

```
# Limpa o workspace
  rm(list=ls())

## Instala os pacotes necessários
install.packages(c("maptools",
                   "rgdal",
                   "RColorBrewer",
                   "classInt",
                   "shapefiles",
                   "ggplot2"))

  # Define o local dos arquivos
setwd("C:/Users/Ricardo/Documents/UFRGS/Trabalho de Conclusão de
Curso/Dados POF")

# Define os produtos a serem estudados
# códigos de produtos do estudo
codigosdoestudo <-
c("0600201","0801201","0801202","0801203","0801304","0803801","1101201","1
101202","1101203",
                      "1103801", "1201301","4701001",
"8605501","8605502","8605503", "8605504",
                      "8605505", "8605506", "8605507")

  # NO - T_DOMICILIO_S.txt
    #domicilio.df <- read.fwf("T_DOMICILIO_S.txt",
    #
width=c(2,2,3,1,2,2,14,14,4,4,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,1,1,1,16,16,
16,1,1,1,1,1,1,1,1,1,1,1,1,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
    #
colClasses=c('factor','factor','factor','factor','factor','factor','numeri
c','numeric','factor','numeric','numeric','numeric','factor','factor','fac
tor','factor','numeric','numeric','factor','factor','numeric','factor','fa
ctor','numeric','factor','factor','factor','factor','factor','numeric','nu
meric','numeric','factor','factor','factor','factor','factor','factor','fa
ctor','factor','factor','factor','factor','factor','factor','factor','fact
or','factor','factor','factor','factor','factor','factor','factor','factor
','factor','factor','factor','factor','factor','factor','factor','factor',
'factor'))
    #colnames(domicilio.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','FATOR_EXPANSAO1','8','9','QT_MO
RADORES','11','QT_FAMILIAS','TIPO_DOMICILIO','14','15','16','QT_COMODOS','
18','AGUA_CANALIZADA','AGUA','QT_BANHEIROS','22','23','24','25','RUA_PAVIM
ENTADA','27','28','29','30','31','32','CORREIOS','AUTO_ESTRADA','AREA_INDU
STRIAL','ESTRADA_FERRO','FIOS_ALTATENSAO','GASODUTO_OLEODUTO','LIXAO','ESG
OTO_CEU_ABERTO','RIO_POLUIDO','AREA_DESLIZAMENTO','SEPARACAO_LIXO','44','4
5','ENERGIA_ELETRICA','FONTEPROPRIA_ENERGIAELETRICA','DIESELGASOLINA_ENERG
IAELETRICA','ENERGIA_SOLAR','ENERGIA_EOLICA','AGUA_ENERGIAELETRICA','BIODI
```

```
ESEL_ENERGIAELETRICA','SMISTO_ENERGIAELETRICA','OUTRAFONTE_ENERGIAELETRICA
','AQUECIMENTO_ENERGIAELETRICA','AQUECIMENTO_GAS','AQUECIMENTO_ENERGIASOLA
R','AQUECIMENTO_ENERGIAELETRICA','AQUECIMENTO_ENERGIAELETRICA','FOGAO_GAS'
,'FOGAO_LENHA','FOGAO_CARVAO','FOGAO_ENERGIAELETRICA','FOGAO_OUTRAFONTE')
    #head(domicilio.df)
    # remover colunas desnecessárias

  # NO - T_CONDICOES_DE_VIDA_S.txt
    #condicoes_vida.df <- read.fwf("T_CONDICOES_DE_VIDA_S.txt",
    #
width=c(2,2,3,1,2,1,2,2,14,14,1,1,1,16,16,16,11,11,1,1,1,1,1,1,1,1,1,1,1,1
,1,1,1,1,1,1,1,1,1,1,1,1,1),
    #
colClasses=c('factor','factor','factor','factor','factor','factor','factor
','factor','numeric','numeric','factor','numeric','factor','numeric','nume
ric','numeric','numeric','numeric','factor','factor','factor','factor','fa
ctor','factor','factor','factor','factor','factor','factor','factor','fact
or','factor','factor','factor','factor','factor','factor','factor','factor
','factor','factor','factor','factor'))
    #colnames(condicoes_vida.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','8','FATOR_EXPANSAO1','10','
11','QT_ALIMENTO','TIPO_ALIMENTO','14','15','16','17','18','19','SERVICO_A
GUA','SERVICO_COLETALIXO','SERVICO_ILUMINACAORUA','SERVICO_LIMPEZAURBANA',
'SERVICO_ESCOAMENTOAGUA','SERVICO_ENERGIAELETRICA','SERVICO_TRANSPCOLETIVO
','SERVICO_EDUCACAO','SERVICO_SAUDE','SERVICO_LAZERESPORTE','SERVICO_ESGOT
AMENTOSANITARIO','POUCO_ESPACO','RUAVIZINHOS_BARULHENTOS','CASA_ESCURA','T
ELHADO_GOTEIRA','UMIDADE','36','PROBLEMAS_AMBIENTAIS','VIOLENCIA_VANDALISM
O','INUNDACOES','CONDICOES_MORADIA','ATRASO_ALUGUEL','ATRASO_DESPESAS','AT
RASO_PRESTACOES')

  # NO - T_INVENTARIO_S.txt
    #inventario.df <- read.fwf("T_INVENTARIO_S.txt",
    #
width=c(2,2,3,1,2,1,2,14,14,2,5,2,4,1,2,2,16,16,16),
    #
colClasses=c('factor','factor','factor','factor','factor','factor','factor
','numeric','numeric','factor','factor','numeric','numeric','factor','fact
or','factor','numeric','numeric','numeric'))
    #colnames(inventario.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','FATOR_EXPANSAO1','9','NUM_Q
UADRO','COD_ITEM','QT_ITEM','13','14','15','16','17','18','19')

  # T_DESPESA_90DIAS_S.txt
    despesa_90dias.df <- read.fwf("T_DESPESA_90DIAS_S.txt",

width=c(2,2,3,1,2,1,2,14,14,2,5,2,11,2,5,11,16,2,16,16,16,4,5,5,14,2,5),

colClasses=c('factor','factor','factor','factor','factor','factor','factor
','numeric','numeric','factor','factor','factor','numeric','numeric','nume
ric','numeric','numeric','factor','numeric','numeric','numeric','numeric',
'factor','factor','numeric','factor','factor'))
    colnames(despesa_90dias.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','FATOR_EXPANSAO1','9','NUM_Q
```

```
UADRO','COD_ITEM','FORMA_AQUISICAO','VALOR_DESPESA_AQUISICAO','14','15','1
6','17','18','19','20','21','QT_ITEM','23','24','QT_FINAL','26','27')
    despesa_90dias.df <- despesa_90dias.df[,c(2,3,4,8,10,11,13)]
    despesa_90dias.df$VALOR_DESPESA_AQUISICAO <-
despesa_90dias.df$VALOR_DESPESA_AQUISICAO * 4 # Transforma o gasto
trimestral em anual
    despesa_90dias.df$codigoagregado <-
paste(despesa_90dias.df$NUM_QUADRO,despesa_90dias.df$COD_ITEM,sep = "")
    despesa_90dias.df$gastodoestudo <-
ifelse(despesa_90dias.df$codigoagregado %in%
codigosdoestudo,despesa_90dias.df$VALOR_DESPESA_AQUISICAO *
despesa_90dias.df$FATOR_EXPANSAO1,0)
    despesa_90dias.df <- despesa_90dias.df[,c(1,2,3,8,9)]
    head(despesa_90dias.df)

    #Remove Outliers
    boxplot(despesa_90dias.df$gastodoestudo)
    despesa_90dias.df.clean <- subset(despesa_90dias.df, gastodoestudo <
25000000)
    boxplot(despesa_90dias.df.clean$gastodoestudo)


  # T_DESPESA_12MESES_S.txt
    despesa_12meses.df <- read.fwf("T_DESPESA_12MESES_S.txt",

width=c(2,2,3,1,2,1,2,14,14,2,5,2,11,2,2,2,5,11,16,2,16,16,16,5),

colClasses=c('factor','factor','factor','factor','factor','factor','factor
','numeric','numeric','factor','factor','factor','numeric','factor','numer
ic','numeric','numeric','numeric','numeric','factor','numeric','numeric','
numeric','factor'))
    colnames(despesa_12meses.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','FATOR_EXPANSAO1','9','NUM_Q
UADRO','COD_ITEM','FORMA_AQUISICAO','VALOR_DESPESA_AQUISICAO','14','15','1
6','17','18','19','20','21','22','23','24')
    despesa_12meses.df <- despesa_12meses.df[,c(2,3,4,8,10,11,13)]
    despesa_12meses.df$codigoagregado <-
paste(despesa_12meses.df$NUM_QUADRO,despesa_12meses.df$COD_ITEM,sep = "")
    despesa_12meses.df$gastodoestudo <-
ifelse(despesa_12meses.df$codigoagregado %in%
codigosdoestudo,despesa_12meses.df$VALOR_DESPESA_AQUISICAO *
despesa_12meses.df$FATOR_EXPANSAO1,0)
    despesa_12meses.df <- despesa_12meses.df[,c(1,2,3,8,9)]
    head(despesa_12meses.df)

    #Remove Outliers
    boxplot(despesa_12meses.df$gastodoestudo)
    despesa_12meses.df.clean <- subset(despesa_12meses.df, gastodoestudo <
10000000)
    boxplot(despesa_12meses.df.clean$gastodoestudo)


  # T_OUTRAS_DESPESAS_S.txt
    outras_despesas.df <- read.fwf("T_OUTRAS_DESPESAS_S.txt",
```

```
width=c(2,2,3,1,2,1,2,14,14,2,5,2,11,1,2,5,11,16,2,16,16,16,5),

colClasses=c('factor','factor','factor','factor','factor','factor','factor
','numeric','numeric','factor','factor','factor','numeric','factor','numer
ic','numeric','numeric','numeric','factor','numeric','numeric','numeric','
factor'))
    colnames(outras_despesas.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','FATOR_EXPANSAO1','9','NUM_Q
UADRO','COD_ITEM','FORMA_AQUISICAO','VALOR_DESPESA_AQUISICAO','14','15','1
6','17','18','19','20','21','22','23')
    outras_despesas.df <- outras_despesas.df[,c(2,3,4,8,10,11,13)]
    outras_despesas.df$codigoagregado <-
paste(outras_despesas.df$NUM_QUADRO,outras_despesas.df$COD_ITEM,sep = "")
    outras_despesas.df$gastodoestudo <-
ifelse(outras_despesas.df$codigoagregado %in%
codigosdoestudo,outras_despesas.df$VALOR_DESPESA_AQUISICAO *
outras_despesas.df$FATOR_EXPANSAO1,0)
    outras_despesas.df <- outras_despesas.df[,c(1,2,3,8,9)]
    head(outras_despesas.df)

    #Remove Outliers
    boxplot(outras_despesas.df$gastodoestudo)
    outras_despesas.df.clean <- subset(outras_despesas.df, gastodoestudo <
10000000)
    boxplot(outras_despesas.df.clean$gastodoestudo)


  # T_SERVICO_DOMS_S.txt
    servico_doms.df <- read.fwf("T_SERVICO_DOMS_S.txt",

width=c(2,2,3,1,2,1,2,14,14,2,5,2,11,5,11,1,2,2,2,5,11,11,16,16,2,2,16,16,
16),

colClasses=c('factor','factor','factor','factor','factor','factor','factor
','numeric','numeric','factor','factor','factor','numeric','factor','numer
ic','factor','factor','numeric','numeric','numeric','numeric','numeric','n
umeric','numeric','factor','factor','numeric','numeric','numeric'))
    colnames(servico_doms.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','FATOR_EXPANSAO1','9','NUM_Q
UADRO','COD_ITEM','FORMA_AQUISICAO','VALOR_DESPESA_AQUISICAO','14','15','1
6','17','18','19','20','21','22','23','24','25','26','27','28','29')
    servico_doms.df <- servico_doms.df[,c(2,3,4,8,10,11,13)]
    servico_doms.df$codigoagregado <-
paste(servico_doms.df$NUM_QUADRO,servico_doms.df$COD_ITEM,sep = "")
    servico_doms.df$gastodoestudo <- ifelse(servico_doms.df$codigoagregado
%in% codigosdoestudo,servico_doms.df$VALOR_DESPESA_AQUISICAO *
servico_doms.df$FATOR_EXPANSAO1,0)
    servico_doms.df <- servico_doms.df[,c(1,2,3,8,9)]
    head(servico_doms.df)

    #Remove Outliers
    boxplot(servico_doms.df$gastodoestudo)
```

```
    servico_doms.df.clean <- subset(servico_doms.df, gastodoestudo <
10000000)
    boxplot(servico_doms.df.clean$gastodoestudo)


  # T_ALUGUEL_ESTIMADO_S.txt
    aluguel_estimado.df <- read.fwf("T_ALUGUEL_ESTIMADO_S.txt",

width=c(2,2,3,1,2,1,2,14,14,2,5,2,11,2,2,2,5,11,16,2,16,16,16),

colClasses=c('factor','factor','factor','factor','factor','factor','factor
','numeric','numeric','factor','factor','factor','numeric','factor','numer
ic','numeric','numeric','numeric','numeric','factor','numeric','numeric','
numeric'))
    colnames(aluguel_estimado.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','FATOR_EXPANSAO1','9','NUM_Q
UADRO','COD_ITEM','FORMA_AQUISICAO','VALOR_DESPESA_AQUISICAO','14','N_MESE
S','16','17','18','19','20','21','22','23')
    aluguel_estimado.df <- aluguel_estimado.df[,c(2,3,4,8,10,11,13,15)]
    aluguel_estimado.df$codigoagregado <-
paste(aluguel_estimado.df$NUM_QUADRO,aluguel_estimado.df$COD_ITEM,sep =
"")
    aluguel_estimado.df$gastodoestudo <-
ifelse(aluguel_estimado.df$codigoagregado %in%
codigosdoestudo,aluguel_estimado.df$VALOR_DESPESA_AQUISICAO *
aluguel_estimado.df$FATOR_EXPANSAO1 * aluguel_estimado.df$N_MESES,0)
    aluguel_estimado.df <- aluguel_estimado.df[,c(1,2,3,9,10)]
    head(aluguel_estimado.df)

    #Remove Outliers
    boxplot(aluguel_estimado.df$gastodoestudo)
    aluguel_estimado.df.clean <- subset(aluguel_estimado.df, gastodoestudo
< 10000000)
    boxplot(aluguel_estimado.df.clean$gastodoestudo)


  # T_CADERNETA_DESPESA_S.txt
    caderneta_despesa.df <- read.fwf("T_CADERNETA_DESPESA_S.txt",

width=c(2,2,3,1,2,1,2,14,14,2,2,5,2,11,2,5,11,16,2,16,16,16,2,8,5,10,5,5),

colClasses=c('factor','factor','factor','factor','factor','factor','factor
','numeric','numeric','factor','factor','factor','factor','numeric','numer
ic','numeric','numeric','numeric','factor','numeric','numeric','numeric','
factor','numeric','factor','numeric','factor','factor'))
    colnames(caderneta_despesa.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','FATOR_EXPANSAO1','9','NUM_Q
UADRO','NUM_GRUPO','COD_ITEM','FORMA_AQUISICAO','VALOR_DESPESA_AQUISICAO',
'15','16','17','18','19','20','21','22','23','24','25','QT_ITEM','27','28'
)
    caderneta_despesa.df <- caderneta_despesa.df[,c(2,3,4,8,10,11,12,14)]
```

```
    caderneta_despesa.df$codigoagregado <-
paste(caderneta_despesa.df$NUM_GRUPO,caderneta_despesa.df$COD_ITEM,sep =
"")
    caderneta_despesa.df$gastodoestudo <-
ifelse(caderneta_despesa.df$codigoagregado %in%
codigosdoestudo,caderneta_despesa.df$VALOR_DESPESA_AQUISICAO *
caderneta_despesa.df$FATOR_EXPANSAO1,0)
    caderneta_despesa.df <- caderneta_despesa.df[,c(1,2,3,9,10)]
    head(caderneta_despesa.df)

    #Remove Outliers
    boxplot(caderneta_despesa.df$gastodoestudo)
    caderneta_despesa.df.clean <- subset(caderneta_despesa.df,
gastodoestudo < 45000)
    boxplot(caderneta_despesa.df.clean$gastodoestudo)


  # T_DESPESA_INDIVIDUAL_S.txt
    despesa_individual.df <- read.fwf("T_DESPESA_INDIVIDUAL_S.txt",

width=c(2,2,3,1,2,1,2,2,14,14,2,5,2,11,2,5,11,16,2,16,16,16,2,5,2,2),

colClasses=c('factor','factor','factor','factor','factor','factor','factor
','factor','numeric','numeric','factor','factor','factor','numeric','numer
ic','numeric','numeric','numeric','factor','numeric','numeric','numeric','
factor','factor','factor','factor'))
    colnames(despesa_individual.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','8','FATOR_EXPANSAO1','10','
NUM_QUADRO','COD_ITEM','FORMA_AQUISICAO','VALOR_DESPESA_AQUISICAO','15','1
6','17','18','19','20','21','22','23','24','25','26')
    despesa_individual.df <- despesa_individual.df[,c(2,3,4,9,11,12,14)]

    despesa_individual.df$codigoagregado <-
paste(despesa_individual.df$NUM_QUADRO,despesa_individual.df$COD_ITEM,sep
= "")
    despesa_individual.df$gastodoestudo <-
ifelse(despesa_individual.df$codigoagregado %in%
codigosdoestudo,despesa_individual.df$VALOR_DESPESA_AQUISICAO *
despesa_individual.df$FATOR_EXPANSAO1,0)
    despesa_individual.df <- despesa_individual.df[,c(1,2,3,8,9)]
    head(despesa_individual.df)

    #Remove Outliers
    boxplot(despesa_individual.df$gastodoestudo)
    despesa_individual.df.clean <- subset(despesa_individual.df,
gastodoestudo < 4000000)
    boxplot(despesa_individual.df.clean$gastodoestudo)


  # T_DESPESA_VEICULO_S.txt
    despesa_veiculo.df <- read.fwf("T_DESPESA_VEICULO_S.txt",

width=c(2,2,3,1,2,1,2,2,14,14,2,5,2,11,1,2,5,11,16,2,16,16,16,5),
```

```
colClasses=c('factor','factor','factor','factor','factor','factor','factor
','factor','numeric','numeric','factor','factor','factor','numeric','facto
r','numeric','numeric','numeric','numeric','factor','numeric','numeric','n
umeric','factor'))
    colnames(despesa_veiculo.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','8','FATOR_EXPANSAO1','10','
NUM_QUADRO','COD_ITEM','FORMA_AQUISICAO','VALOR_DESPESA_AQUISICAO','15','1
6','17','18','19','20','21','22','23','24')
    despesa_veiculo.df <- despesa_veiculo.df[,c(2,3,4,9,11,12,14)]

    despesa_veiculo.df$codigoagregado <-
paste(despesa_veiculo.df$NUM_QUADRO,despesa_veiculo.df$COD_ITEM,sep = "")
    despesa_veiculo.df$gastodoestudo <-
ifelse(despesa_veiculo.df$codigoagregado %in%
codigosdoestudo,despesa_veiculo.df$VALOR_DESPESA_AQUISICAO *
despesa_veiculo.df$FATOR_EXPANSAO1,0)
    despesa_veiculo.df <- despesa_veiculo.df[,c(1,2,3,8,9)]
    head(despesa_veiculo.df)

    #Remove Outliers
    boxplot(despesa_veiculo.df$gastodoestudo)
    despesa_veiculo.df.clean <- subset(despesa_veiculo.df, gastodoestudo <
4000000)
    boxplot(despesa_veiculo.df.clean$gastodoestudo)


    consumo.df <-
rbind(despesa_90dias.df.clean,despesa_12meses.df.clean,outras_despesas.df.
clean,servico_doms.df.clean,aluguel_estimado.df.clean,caderneta_despesa.df
.clean,despesa_individual.df.clean,despesa_veiculo.df.clean)
    str(consumo.df)
    gastodoestudo.df <- aggregate(consumo.df$gastodoestudo,by =
list(consumo.df$COD_UF,consumo.df$NUM_SEQ,consumo.df$NUM_DV), sum)
    colnames(gastodoestudo.df) <-
c('COD_UF','NUM_SEQ','NUM_DV','gastodoestudo')
    head(gastodoestudo.df)


  # Cria as variáveis de característica do Setor Censitário com base no
arquivo de morador

    # T_MORADOR_S.txt
  morador.df <- read.fwf("T_MORADOR_S.txt",

width=c(2,2,3,1,2,1,2,2,14,14,2,2,2,2,2,2,4,3,6,7,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,16,16,16,2,5,5,5,5,5,5,5,16,8,2,2,2,2,2,2,2,2,2,2,2,2),

colClasses=c('factor','factor','factor','factor','factor','factor','factor
','factor','numeric','numeric','factor','factor','factor','factor','factor
','factor','factor','numeric','numeric','numeric','factor','factor','facto
r','factor','numeric','factor','factor','numeric','factor','factor','numer
ic','factor','numeric','numeric','factor','factor','factor','factor','nume
ric','numeric','numeric','factor','numeric','numeric','numeric','numeric',
```

```
'numeric','numeric','numeric','numeric','factor','factor','factor','numeri
c','factor','factor','factor','numeric','factor','factor','factor','numeri
c','factor','factor'))
  colnames(morador.df) <-
c('1','COD_UF','NUM_SEQ','NUM_DV','5','6','7','8','FATOR_EXPANSAO1','10','
11','12','CONDICAO_FAMILIA','14','15','16','17','IDADE','19','20','SEXO','
ALFABETIZACAO','ESCOLA_CRECHE','CURSO_FREQUENTA','25','26','ESCOLARIDADE',
'28','29','30','ANOS_ESTUDO','COR_PELE','33','34','CARTAO_CREDITO','36','3
7','38','39','40','41','GRAVIDEZ','43','44','45','46','47','48','49','REND
A','RELIGIAO','PLANO_SAUDE','53','54','55','USA_MEDICAMENTO','USA_SERVICOS
AUDE','58','59','60','61','62','63','64')
  morador.df <-
morador.df[,c(2,3,4,9,13,18,21,22,23,24,27,31,32,35,42,50,51,52,56,57)]
  str(morador.df)

  # Faz as primeiras observações
  head(morador.df)

  # Guarda a base na memória
  attach(morador.df)

  # Cria as variáveis
  FATOR_EXPANSAO1.morador <- as.numeric(morador.df$'FATOR_EXPANSAO1')
  IDADE_ANOS <- morador.df$IDADE
  idade0a9 <- ifelse(IDADE_ANOS>=0 &
IDADE_ANOS<9,1,0)*FATOR_EXPANSAO1.morador
  idade10a19 <- ifelse(IDADE_ANOS>=10 &
IDADE_ANOS<19,1,0)*FATOR_EXPANSAO1.morador
  idade20a29 <- ifelse(IDADE_ANOS>=20 &
IDADE_ANOS<29,1,0)*FATOR_EXPANSAO1.morador
  idade30a39 <- ifelse(IDADE_ANOS>=30 &
IDADE_ANOS<39,1,0)*FATOR_EXPANSAO1.morador
  idade40a49 <- ifelse(IDADE_ANOS>=40 &
IDADE_ANOS<49,1,0)*FATOR_EXPANSAO1.morador
  idade50a59 <- ifelse(IDADE_ANOS>=50 &
IDADE_ANOS<59,1,0)*FATOR_EXPANSAO1.morador
  idade60a69 <- ifelse(IDADE_ANOS>=60 &
IDADE_ANOS<69,1,0)*FATOR_EXPANSAO1.morador
  idade70acima <- ifelse(IDADE_ANOS>=70 &
IDADE_ANOS<99,1,0)*FATOR_EXPANSAO1.morador

  # Gênero
  COD_SEXO <- morador.df$SEXO
  homem <- ifelse(COD_SEXO=="01",1,0)*FATOR_EXPANSAO1.morador
  mulher <- ifelse(COD_SEXO=="02",1,0)*FATOR_EXPANSAO1.morador

  # Renda
  RENDA_PER_CAPITA <- as.numeric(morador.df$RENDA)
  renda1 <- ifelse(RENDA_PER_CAPITA>=63.75,1,0)*FATOR_EXPANSAO1.morador
  renda2 <- ifelse(RENDA_PER_CAPITA>63.75 & RENDA_PER_CAPITA
<=127.5,1,0)*FATOR_EXPANSAO1.morador
  renda3 <- ifelse(RENDA_PER_CAPITA>127.5 & RENDA_PER_CAPITA
<=255,1,0)*FATOR_EXPANSAO1.morador
```

```
  renda4 <- ifelse(RENDA_PER_CAPITA>255 & RENDA_PER_CAPITA <=
510,1,0)*FATOR_EXPANSAO1.morador
  renda5 <- ifelse(RENDA_PER_CAPITA>510 & RENDA_PER_CAPITA <=
1020,1,0)*FATOR_EXPANSAO1.morador
  renda6 <- ifelse(RENDA_PER_CAPITA>1020 & RENDA_PER_CAPITA <=
1530,1,0)*FATOR_EXPANSAO1.morador
  renda7 <- ifelse(RENDA_PER_CAPITA>1530 & RENDA_PER_CAPITA <=
2550,1,0)*FATOR_EXPANSAO1.morador
  renda8 <- ifelse(RENDA_PER_CAPITA>2550 & RENDA_PER_CAPITA <=
5100,1,0)*FATOR_EXPANSAO1.morador
  renda9 <- ifelse(RENDA_PER_CAPITA>5100,1,0)*FATOR_EXPANSAO1.morador

  # Alfabetização
  COD_SABE_LER <- morador.df$ALFABETIZACAO
  alfabetizado <- ifelse(COD_SABE_LER=="01",1,0)*FATOR_EXPANSAO1.morador
  analfabeto <- ifelse(COD_SABE_LER=="02",1,0)*FATOR_EXPANSAO1.morador

  # Cor da Pele
  COD_COR_PELE <- morador.df$COR_PELE
  branca <-ifelse(COD_COR_PELE=="01",1,0)*FATOR_EXPANSAO1.morador
  preta <-ifelse(COD_COR_PELE=="02",1,0)*FATOR_EXPANSAO1.morador
  amarela <-ifelse(COD_COR_PELE=="03",1,0)*FATOR_EXPANSAO1.morador
  parda <-ifelse(COD_COR_PELE=="04",1,0)*FATOR_EXPANSAO1.morador
  indigena <-ifelse(COD_COR_PELE=="05",1,0)*FATOR_EXPANSAO1.morador
  naosaberaca <-ifelse(COD_COR_PELE=="09",1,0)*FATOR_EXPANSAO1.morador

  # Remove a base da memória
  detach(morador.df)

  # Agrega os resultados para as variáveis criadas
  setores.df <-
aggregate(cbind(idade0a9,idade10a19,idade20a29,idade30a39,idade40a49,idade
50a59,idade60a69,idade70acima,homem,mulher,RENDA_PER_CAPITA,renda1,renda2,
renda3,renda4,renda5,renda6,renda7,renda8,renda9,alfabetizado,analfabeto,b
ranca,preta,amarela,parda,indigena,naosaberaca),by=list(morador.df$COD_UF,
morador.df$NUM_SEQ,morador.df$NUM_DV),sum)
  head(setores.df,10)

  # Nomeia as colunas
  colnames(setores.df) <-
c("COD_UF","NUM_SEQ","NUM_DV","Idade0a9","Idade10a19","Idade20a29","Idade3
0a39","Idade40a49","Idade50a59","Idade60a69","Idade70a99","Homem","Mulher"
,"RENDA_PERCAPITA","Renda1","Renda2","Renda3","Renda4","Renda5","Renda6","
Renda7","Renda8","Renda9","alfabetizado","analfabeto","branca","preta","am
arela","parda","indigena","naosaberaca")

  # Une as bases
  tudo.df <-
merge(gastodoestudo.df,setores.df,by=c("COD_UF","NUM_SEQ","NUM_DV"))

  # Retira quem naosaberaca, RENDA_PERCAPITA e analfabeto
  str(tudo.df)
  tudo.df <- tudo.df[,-c(15,26,32)]
```

```
  # Observa a base criada
  head(tudo.df)

  # Habilita o pacote kernlab
  library(kernlab)

  #Divide o data frame criado em 2 partes: 70% Treinamento e 30% Validação
  # Fixa a semente
  set.seed(23238)

  # Cria a dummy para separar a base
  # Cria um novo vetor onde 70% das linhas terão o valor "1" (que serão as
linhas utilizadas na base de treinamento)
  sorteio <- rbinom(nrow(tudo.df),1,0.7)
  head(sorteio)

  # Retira as variáveis de controle (COD_UF, NUM_SEQ, NUM_DV)
  tudoLimpa.df <- tudo.df[,-c(1:3)]

##Transformar os valores em escala. Salvar valores da média e desvio-
padrão antes de transformar em escala
  # Define a Base de treinamento
  treina.df <- tudoLimpa.df[which(sorteio==1),] # Seleciona apenas as
linhas nas quais o valor do vetor "sorteio" é 1
  treina.df.scaled <- scale(treina.df)
  treina.df_mean <- attributes(treina.df.scaled)$'scaled:center' #the mean
  treina.df_std <- attributes(treina.df.scaled)$'scaled:scale'   #the
standard deviation
  head(treina.df.scaled,10)


  # Cria a base de Validação com as linhas não inclusas na base de
treinamento
  valida.df <- tudoLimpa.df[which(sorteio!=1),]
  valida.df <- as.data.frame((scale(valida.df)))

  # Cria a lista de parâmetros para a busca
  listaC <- seq(0.01, 10, length.out = 50)
  listaSigma2 <- c(seq(0.01,10, length.out = 25), seq(0.01,10,length.out =
25))
  parms <- expand.grid(C = listaC, sigma = listaSigma2)


  # Para cada par de parâmetros pegue aquele com o menor EQM
  apply_pb <- function(X, MARGIN, FUN, ...)
  {
    env <- environment()
    pb_Total <- sum(dim(X)[MARGIN])
    counter <- 0
    pb <- txtProgressBar(min = 0, max = pb_Total, style = 3)

    wrapper <- function(...)
    {
      curVal <- get("counter", envir = env)
```

```
      assign("counter", curVal +1 , envir = env)
      setTxtProgressBar(get("pb", envir = env), curVal +1)
      FUN(...)
    }
    res <- apply(X, MARGIN, wrapper, ...)
    close(pb)
    res
  }



  # Realiza o treinamento da máquina e compara os resultados reais com os
preditivos
  head(treina.df.scaled)

  EQM <- apply_pb(parms,1,function(x)
  {
    svm <- ksvm(gastodoestudo ~.,data = treina.df.scaled,
kernel="rbfdot",type="eps-svr",

kpar=list(sigma=as.numeric(x[2])),C=as.numeric(x[1]),scaled
                =FALSE);mean((valida.df$gastodoestudo-
predict(svm,valida.df)^2))
  }
  )

  # Junta todos os resultados
  parms$EQM <- EQM

  # Obtém os parâmetros de menor valor
  summary(parms$EQM)
  hist(parms$EQM)
  iMin <- which(parms$EQM==min(parms$EQM))


  # Resultado final
  parms[iMin[1],]
  C <- as.numeric(parms[iMin[1],][1])
  sigma <- as.numeric(parms[iMin[1],][2])

  # C <- 1
  # sigma <- 0.01

  # Realiza o treinamento com os valores ótimos a partir dos parâmetros
criados.
  library(kernlab)
  svm <- ksvm(gastodoestudo~.,
              data = treina.df.scaled,
              kernel="rbfdot",
              type="eps-svr",
              kpar=list(sigma=sigma),
              C=C,
              scaled=FALSE)
```

```
  # Teste

  rmse <- function(error)
  {
    sqrt(mean(error^2))
  }

  error <- EQM$residuals  # same as data$Y - predictedY
  predictionRMSE <- rmse(error)

  # Define o local da base de dados do Censo
  setwd("C:/Users/Ricardo/Documents/UFRGS/Trabalho de Conclusão de
Curso/Censo 2010/SP_Capital_20171016/SP Capital/Base informaçoes
setores2010 universo SP_Capital/CSV")
  library(readxl)

  # Cria a base agrupadora para previsão da escala de despesas
  censo.df <- data.frame()

  # Função para conversão de character para numeric
  asNumeric <- function(x) as.numeric(as.character(x))
  charNumeric <- function(d) modifyList(d, lapply(d[,sapply(d,
is.character)], asNumeric))
  factorsNumeric <- function(d) modifyList(d, lapply(d[,sapply(d,
is.factor)], asNumeric))

  # Importa os dados de Idade
  i_AC <- read_excel("Pessoa13_AC.xls", col_types = "numeric")
  i_AL <- read_excel("Pessoa13_AL.xls", col_types = "numeric")
  i_AM <- read_excel("Pessoa13_AM.xls", col_types = "numeric")
  i_AP <- read_excel("Pessoa13_AP.xls", col_types = "numeric")
  i_BA <- read_excel("Pessoa13_BA.xls", col_types = "numeric")
  i_CE <- read_excel("Pessoa13_CE.xls", col_types = "numeric")
  i_DF <- read_excel("Pessoa13_DF.xls", col_types = "numeric")
  i_ES <- read_excel("Pessoa13_ES.xls", col_types = "numeric")
  i_GO <- read_excel("Pessoa13_GO.xls", col_types = "numeric")
  i_MA <- read_excel("Pessoa13_MA.xls", col_types = "numeric")
  i_MG <- read_excel("Pessoa13_MG.xls", col_types = "numeric")
  i_MS <- read_excel("Pessoa13_MS.xls", col_types = "numeric")
  i_MT <- read_excel("Pessoa13_MT.xls", col_types = "numeric")
  i_PA <- read_excel("Pessoa13_PA.xls", col_types = "numeric")
  i_PB <- read_excel("Pessoa13_PB.xls", col_types = "numeric")
  i_PE <- read_excel("Pessoa13_PE.xls", col_types = "numeric")
  i_PI <- read_excel("Pessoa13_PI.xls", col_types = "numeric")
  i_PR <- read_excel("Pessoa13_PR.xls", col_types = "numeric")
  i_RJ <- read_excel("Pessoa13_RJ.xls", col_types = "numeric")
  i_RN <- read_excel("Pessoa13_RN.xls", col_types = "numeric")
  i_RO <- read_excel("Pessoa13_RO.xls", col_types = "numeric")
  i_RR <- read_excel("Pessoa13_RR.xls", col_types = "numeric")
  i_RS <- read_excel("Pessoa13_RS.xls", col_types = "numeric")
  i_SC <- read_excel("Pessoa13_SC.xls", col_types = "numeric")
  i_SE <- read_excel("Pessoa13_SE.xls", col_types = "numeric")
```

```
  i_SP1 <- read_excel("~/UFRGS/Trabalho de Conclusão de Curso/Censo 
2010/SP_Capital_20171016/SP Capital/Base informaçoes setores2010 universo 
SP_Capital/EXCEL/Pessoa13_SPCapital.xlsx", 
                      col_types = c("numeric", "numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric", 
```

```
                                      "numeric", "numeric", "numeric",
"numeric", "numeric",
                                      "numeric", "numeric", "numeric"))
  i_SP2 <- read_excel("Pessoa13_SP2.xls", col_types = "numeric")
  i_TO <- read_excel("Pessoa13_TO.xls", col_types = "numeric")
  idade <- rbind(i_SP1)

remove(i_AC,i_AL,i_AM,i_AP,i_BA,i_CE,i_DF,i_ES,i_GO,i_MA,i_MG,i_MS,i_MT,i_
PA,i_PB,i_PE,i_PI,i_PR,i_RJ,i_RN,i_RO,i_RR,i_RS,i_SC,i_SE,i_SP2,i_TO)
  str(idade)

  # Define Código do Setor Censitário
  SETOR <- idade[,1]
  SETOR$Cod_setor <- as.character(idade$Cod_setor)

  # Cria as classes de idade da mesma forma que foi feito com os dados da
POF
  idade0a9 <- rowSums(charNumeric(idade[,c(24:45)]), na.rm = T)
  idade10a19 <- rowSums(charNumeric(idade[,c(46:55)]), na.rm = T)
  idade20a29 <- rowSums(charNumeric(idade[,c(56:65)]), na.rm = T)
  idade30a39 <- rowSums(charNumeric(idade[,c(66:75)]), na.rm = T)
  idade40a49 <- rowSums(charNumeric(idade[,c(76:85)]), na.rm = T)
  idade50a59 <- rowSums(charNumeric(idade[,c(86:95)]), na.rm = T)
  idade60a69 <- rowSums(charNumeric(idade[,c(96:105)]), na.rm = T)
  idade70acima <- rowSums(charNumeric(idade[,c(106:136)]), na.rm = T)

  # Agrupa todas as classes de idade num data frame temporário
  temp <- as.data.frame(SETOR)
  temp$idade0a9 <- idade0a9
  temp$idade10a19 <- idade10a19
  temp$idade20a29 <- idade20a29
  temp$idade30a39 <- idade30a39
  temp$idade40a49 <- idade40a49
  temp$idade50a59 <- idade50a59
  temp$idade60a69 <- idade60a69
  temp$idade70acima <- idade70acima

  # Agrupa tudo na base censo.df
  censo.df <- as.data.frame(SETOR)
  censo.df <- merge(censo.df, temp, by = c("Cod_setor")) # Usa a coluna do
setor como base para a mescla dos dados
  str(censo.df) # Observa a estrutura

  # Importa os dados dos Homens
  h_AC <- read_excel("Pessoa11_AC.xls", col_types = "numeric")
  h_AL <- read_excel("Pessoa11_AL.xls", col_types = "numeric")
  h_AM <- read_excel("Pessoa11_AM.xls", col_types = "numeric")
  h_AP <- read_excel("Pessoa11_AP.xls", col_types = "numeric")
  h_BA <- read_excel("Pessoa11_BA.xls", col_types = "numeric")
  h_CE <- read_excel("Pessoa11_CE.xls", col_types = "numeric")
  h_DF <- read_excel("Pessoa11_DF.xls", col_types = "numeric")
  h_ES <- read_excel("Pessoa11_ES.xls", col_types = "numeric")
  h_GO <- read_excel("Pessoa11_GO.xls", col_types = "numeric")
  h_MA <- read_excel("Pessoa11_MA.xls", col_types = "numeric")
```

```
  h_MG <- read_excel("Pessoa11_MG.xls", col_types = "numeric")
  h_MS <- read_excel("Pessoa11_MS.xls", col_types = "numeric")
  h_MT <- read_excel("Pessoa11_MT.xls", col_types = "numeric")
  h_PA <- read_excel("Pessoa11_PA.xls", col_types = "numeric")
  h_PB <- read_excel("Pessoa11_PB.xls", col_types = "numeric")
  h_PE <- read_excel("Pessoa11_PE.xls", col_types = "numeric")
  h_PI <- read_excel("Pessoa11_PI.xls", col_types = "numeric")
  h_PR <- read_excel("Pessoa11_PR.xls", col_types = "numeric")
  h_RJ <- read_excel("Pessoa11_RJ.xls", col_types = "numeric")
  h_RN <- read_excel("Pessoa11_RN.xls", col_types = "numeric")
  h_RO <- read_excel("Pessoa11_RO.xls", col_types = "numeric")
  h_RR <- read_excel("Pessoa11_RR.xls", col_types = "numeric")
  h_RS <- read_excel("Pessoa11_RS.xls", col_types = "numeric")
  h_SC <- read_excel("Pessoa11_SC.xls", col_types = "numeric")
  h_SE <- read_excel("Pessoa11_SE.xls", col_types = "numeric")
  h_SP1 <- read_excel("~/UFRGS/Trabalho de Conclusão de Curso/Censo 
2010/SP_Capital_20171016/SP Capital/Base informaçoes setores2010 universo 
SP_Capital/EXCEL/Pessoa11_SPCapital.xlsx",
                      col_types = c("numeric", "numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
                                    "numeric", "numeric", "numeric", 
"numeric", "numeric",
```

```
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric"))

  h_SP2 <- read_excel("Pessoa11_SP2.xls", col_types = "numeric")
  h_TO <- read_excel("Pessoa11_TO.xls", col_types = "numeric")
  homem <- rbind(h_SP1)

remove(h_AC,h_AL,h_AM,h_AP,h_BA,h_CE,h_DF,h_ES,h_GO,h_MA,h_MG,h_MS,h_MT,h_
PA,h_PB,h_PE,h_PI,h_PR,h_RJ,h_RN,h_RO,h_RR,h_RS,h_SC,h_SE,h_SP2,h_TO)
  str(homem)

  # Define Código do Setor Censitário
  SETOR <- homem[,1]
  SETOR$Cod_setor <- as.character(idade$Cod_setor)

  # Soma a quantidade de homens do Censo
  homem <- rowSums(charNumeric(homem[,24:136]), na.rm = T)

  # Agrupa todos os valores num data frame temporário
  temph <- as.data.frame(SETOR)
  temph$homem <- homem

  # Agrupa tudo na base censo.df
  censo.df <- merge(censo.df, temph, by=c("Cod_setor")) # Usa a coluna do
setor como base para a mescla dos dados
  str(censo.df)

  # Importa os dados das Mulheres
  m_AC <- read_excel("Pessoa12_AC.xls", col_types = "numeric")
  m_AL <- read_excel("Pessoa12_AL.xls", col_types = "numeric")
  m_AM <- read_excel("Pessoa12_AM.xls", col_types = "numeric")
  m_AP <- read_excel("Pessoa12_AP.xls", col_types = "numeric")
  m_BA <- read_excel("Pessoa12_BA.xls", col_types = "numeric")
  m_CE <- read_excel("Pessoa12_CE.xls", col_types = "numeric")
  m_DF <- read_excel("Pessoa12_DF.xls", col_types = "numeric")
  m_ES <- read_excel("Pessoa12_ES.xls", col_types = "numeric")
```

```
  m_GO <- read_excel("Pessoa12_GO.xls", col_types = "numeric")
  m_MA <- read_excel("Pessoa12_MA.xls", col_types = "numeric")
  m_MG <- read_excel("Pessoa12_MG.xls", col_types = "numeric")
  m_MS <- read_excel("Pessoa12_MS.xls", col_types = "numeric")
  m_MT <- read_excel("Pessoa12_MT.xls", col_types = "numeric")
  m_PA <- read_excel("Pessoa12_PA.xls", col_types = "numeric")
  m_PB <- read_excel("Pessoa12_PB.xls", col_types = "numeric")
  m_PE <- read_excel("Pessoa12_PE.xls", col_types = "numeric")
  m_PI <- read_excel("Pessoa12_PI.xls", col_types = "numeric")
  m_PR <- read_excel("Pessoa12_PR.xls", col_types = "numeric")
  m_RJ <- read_excel("Pessoa12_RJ.xls", col_types = "numeric")
  m_RN <- read_excel("Pessoa12_RN.xls", col_types = "numeric")
  m_RO <- read_excel("Pessoa12_RO.xls", col_types = "numeric")
  m_RR <- read_excel("Pessoa12_RR.xls", col_types = "numeric")
  m_RS <- read_excel("Pessoa12_RS.xls", col_types = "numeric")
  m_SC <- read_excel("Pessoa12_SC.xls", col_types = "numeric")
  m_SE <- read_excel("Pessoa12_SE.xls", col_types = "numeric")
  m_SP1 <- read_excel("~/UFRGS/Trabalho de Conclusão de Curso/Censo
2010/SP_Capital_20171016/SP Capital/Base informaçoes setores2010 universo
SP_Capital/EXCEL/Pessoa12_SPCapital.xlsx",
                      col_types = c("numeric", "numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
```

```
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric"))

  m_SP2 <- read_excel("Pessoa12_SP2.xls", col_types = "numeric")
  m_TO <- read_excel("Pessoa12_TO.xls", col_types = "numeric")
  mulher <- rbind(m_SP1)

remove(m_AC,m_AL,m_AM,m_AP,m_BA,m_CE,m_DF,m_ES,m_GO,m_MA,m_MG,m_MS,m_MT,m_
PA,m_PB,m_PE,m_PI,m_PR,m_RJ,m_RN,m_RO,m_RR,m_RS,m_SC,m_SE,m_SP2,m_TO)
  str(mulher)

  # Define Código do Setor Censitário
  SETOR <- mulher[,1]
  SETOR$Cod_setor <- as.character(idade$Cod_setor)

  # Soma a quantidade de mulheres do Censo
  mulher <- rowSums(charNumeric(mulher[,24:136]), na.rm = T)

  # Agrupa todos os valores num data frame temporário
  tempm <- as.data.frame(SETOR)
  tempm$mulher <- mulher

  # Agrupa tudo na base censo.df
  censo.df <- merge(censo.df, tempm, by="Cod_setor") # Usa a coluna do
setor como base para a mescla dos dados
  str(censo.df)

  warnings()

  # Importa os dados de Renda
  r_AC <- read_excel("DomicilioRenda_AC.XLS", col_types = "numeric")
  r_AL <- read_excel("DomicilioRenda_AL.XLS", col_types = "numeric")
  r_AM <- read_excel("DomicilioRenda_AM.XLS", col_types = "numeric")
  r_AP <- read_excel("DomicilioRenda_AP.XLS", col_types = "numeric")
```

```
  r_BA <- read_excel("DomicilioRenda_BA.XLS", col_types = "numeric")
  r_CE <- read_excel("DomicilioRenda_CE.XLS", col_types = "numeric")
  r_DF <- read_excel("DomicilioRenda_DF.XLS", col_types = "numeric")
  r_ES <- read_excel("DomicilioRenda_ES.XLS", col_types = "numeric")
  r_GO <- read_excel("DomicilioRenda_GO.XLS", col_types = "numeric")
  r_MA <- read_excel("DomicilioRenda_MA.XLS", col_types = "numeric")
  r_MG <- read_excel("DomicilioRenda_MG.XLS", col_types = "numeric")
  r_MS <- read_excel("DomicilioRenda_MS.XLS", col_types = "numeric")
  r_MT <- read_excel("DomicilioRenda_MT.XLS", col_types = "numeric")
  r_PA <- read_excel("DomicilioRenda_PA.XLS", col_types = "numeric")
  r_PB <- read_excel("DomicilioRenda_PB.XLS", col_types = "numeric")
  r_PE <- read_excel("DomicilioRenda_PE.XLS", col_types = "numeric")
  r_PI <- read_excel("DomicilioRenda_PI.XLS", col_types = "numeric")
  r_PR <- read_excel("DomicilioRenda_PR.XLS", col_types = "numeric")
  r_RJ <- read_excel("DomicilioRenda_RJ.XLS", col_types = "numeric")
  r_RN <- read_excel("DomicilioRenda_RN.XLS", col_types = "numeric")
  r_RO <- read_excel("DomicilioRenda_RO.XLS", col_types = "numeric")
  r_RR <- read_excel("DomicilioRenda_RR.XLS", col_types = "numeric")
  r_RS <- read_excel("DomicilioRenda_RS.XLS", col_types = "numeric")
  r_SC <- read_excel("DomicilioRenda_SC.XLS", col_types = "numeric")
  r_SE <- read_excel("DomicilioRenda_SE.XLS", col_types = "numeric")
  r_SP1 <- read_excel("~/UFRGS/Trabalho de Conclusão de Curso/Censo
2010/SP_Capital_20171016/SP Capital/Base informaçoes setores2010 universo
SP_Capital/EXCEL/DomicilioRenda_SPCapital.xlsx",
                      col_types = c("numeric", "numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric"))

  r_SP2 <- read_excel("DomicilioRenda_SP2.XLS", col_types = "numeric")
  r_TO <- read_excel("DomicilioRenda_TO.XLS", col_types = "numeric")
  renda <- rbind(r_SP1)

remove(r_AC,r_AL,r_AM,r_AP,r_BA,r_CE,r_DF,r_ES,r_GO,r_MA,r_MG,r_MS,r_MT,r_
PA,r_PB,r_PE,r_PI,r_PR,r_RJ,r_RN,r_RO,r_RR,r_RS,r_SC,r_SE,r_SP2,r_TO)
  str(renda)

  # Define Código do Setor Censitário
  SETOR <- renda[,1]
  SETOR$Cod_setor <- as.character(idade$Cod_setor)

  # Agrupa todos os valores num data frame temporário
  temp <- as.data.frame(SETOR)

  # Cria as classes de renda da mesma forma que foi feito com os dados da
POF
  renda1 <- rowSums(charNumeric(renda[,7]), na.rm = T)
  renda2 <- rowSums(charNumeric(renda[,8]), na.rm = T)
  renda3 <- rowSums(charNumeric(renda[,9]), na.rm = T)
  renda4 <- rowSums(charNumeric(renda[,10]), na.rm = T)
  renda5 <- rowSums(charNumeric(renda[,11]), na.rm = T)
  renda6 <- rowSums(charNumeric(renda[,12]), na.rm = T)
```

```
  renda7 <- rowSums(charNumeric(renda[,13]), na.rm = T)
  renda8 <- rowSums(charNumeric(renda[,14]), na.rm = T)
  renda9 <- rowSums(charNumeric(renda[,15]), na.rm = T)

  # Agrupa todas as classes de renda num data frame temporário
  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
  str(temp)

  # Agrupar classes de variáveis no data frame
  censo.df <- merge(censo.df, temp, by="Cod_setor")
  str(censo.df)

  # Importa os dados de Alfabetização
  a_AC <- read_excel("Pessoa01_AC.xls", col_types = "numeric")
  a_AL <- read_excel("Pessoa01_AL.xls", col_types = "numeric")
  a_AM <- read_excel("Pessoa01_AM.xls", col_types = "numeric")
  a_AP <- read_excel("Pessoa01_AP.xls", col_types = "numeric")
  a_BA <- read_excel("Pessoa01_BA.xls", col_types = "numeric")
  a_CE <- read_excel("Pessoa01_CE.xls", col_types = "numeric")
  a_DF <- read_excel("Pessoa01_DF.xls", col_types = "numeric")
  a_ES <- read_excel("Pessoa01_ES.xls", col_types = "numeric")
  a_GO <- read_excel("Pessoa01_GO.xls", col_types = "numeric")
  a_MA <- read_excel("Pessoa01_MA.xls", col_types = "numeric")
  a_MG <- read_excel("Pessoa01_MG.xls", col_types = "numeric")
  a_MS <- read_excel("Pessoa01_MS.xls", col_types = "numeric")
  a_MT <- read_excel("Pessoa01_MT.xls", col_types = "numeric")
  a_PA <- read_excel("Pessoa01_PA.xls", col_types = "numeric")
  a_PB <- read_excel("Pessoa01_PB.xls", col_types = "numeric")
  a_PE <- read_excel("Pessoa01_PE.xls", col_types = "numeric")
  a_PI <- read_excel("Pessoa01_PI.xls", col_types = "numeric")
  a_PR <- read_excel("Pessoa01_PR.xls", col_types = "numeric")
  a_RJ <- read_excel("Pessoa01_RJ.xls", col_types = "numeric")
  a_RN <- read_excel("Pessoa01_RN.xls", col_types = "numeric")
  a_RO <- read_excel("Pessoa01_RO.xls", col_types = "numeric")
  a_RR <- read_excel("Pessoa01_RR.xls", col_types = "numeric")
  a_RS <- read_excel("Pessoa01_RS.xls", col_types = "numeric")
  a_SC <- read_excel("Pessoa01_SC.xls", col_types = "numeric")
  a_SE <- read_excel("Pessoa01_SE.xls", col_types = "numeric")
  a_SP1 <- read_excel("~/UFRGS/Trabalho de Conclusão de Curso/Censo
2010/SP_Capital_20171016/SP Capital/Base informaçoes setores2010 universo
SP_Capital/EXCEL/Pessoa01_SPCapital.xlsx",
                      col_types = c("numeric", "numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
```

```
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric"))

  a_SP2 <- read_excel("Pessoa01_SP2.xls", col_types = "numeric")
  a_TO <- read_excel("Pessoa01_TO.xls", col_types = "numeric")
  alfabetizado <- rbind(a_SP1)

remove(a_AC,a_AL,a_AM,a_AP,a_BA,a_CE,a_DF,a_ES,a_GO,a_MA,a_MG,a_MS,a_MT,a_
PA,a_PB,a_PE,a_PI,a_PR,a_RJ,a_RN,a_RO,a_RR,a_RS,a_SC,a_SE,a_SP2,a_TO)
  str(alfabetizado)

  alfabetizado <- alfabetizado[,c(1,3:79)]

  # Define Código do Setor Censitário
  SETOR <- alfabetizado[,1]
  SETOR$Cod_setor <- as.character(alfabetizado$Cod_setor)

  # Agrupa todos os valores num data frame temporário
  temp <- as.data.frame(SETOR)

  #Definir classes de idade, gênero, renda, alfabetização e raça
  alfabetizado <- rowSums(charNumeric(alfabetizado[,2:78]), na.rm = T)
  temp$alfabetizado <- alfabetizado
  str(temp)

  # Agrupa tudo na base censo.df
```

```
  censo.df <- merge(censo.df, temp, by="Cod_setor")  # Usa a coluna do
setor como base para a mescla dos dados
  str(censo.df)

  # Importa os dados de Raça
  p_AC <- read_excel("Pessoa03_AC.xls", col_types = "numeric")
  p_AL <- read_excel("Pessoa03_AL.xls", col_types = "numeric")
  p_AM <- read_excel("Pessoa03_AM.xls", col_types = "numeric")
  p_AP <- read_excel("Pessoa03_AP.xls", col_types = "numeric")
  p_BA <- read_excel("Pessoa03_BA.xls", col_types = "numeric")
  p_CE <- read_excel("Pessoa03_CE.xls", col_types = "numeric")
  p_DF <- read_excel("Pessoa03_DF.xls", col_types = "numeric")
  p_ES <- read_excel("Pessoa03_ES.xls", col_types = "numeric")
  p_GO <- read_excel("Pessoa03_GO.xls", col_types = "numeric")
  p_MA <- read_excel("Pessoa03_MA.xls", col_types = "numeric")
  p_MG <- read_excel("Pessoa03_MG.xls", col_types = "numeric")
  p_MS <- read_excel("Pessoa03_MS.xls", col_types = "numeric")
  p_MT <- read_excel("Pessoa03_MT.xls", col_types = "numeric")
  p_PA <- read_excel("Pessoa03_PA.xls", col_types = "numeric")
  p_PB <- read_excel("Pessoa03_PB.xls", col_types = "numeric")
  p_PE <- read_excel("Pessoa03_PE.xls", col_types = "numeric")
  p_PI <- read_excel("Pessoa03_PI.xls", col_types = "numeric")
  p_PR <- read_excel("Pessoa03_PR.xls", col_types = "numeric")
  p_RJ <- read_excel("Pessoa03_RJ.xls", col_types = "numeric")
  p_RN <- read_excel("Pessoa03_RN.xls", col_types = "numeric")
  p_RO <- read_excel("Pessoa03_RO.xls", col_types = "numeric")
  p_RR <- read_excel("Pessoa03_RR.xls", col_types = "numeric")
  p_RS <- read_excel("Pessoa03_RS.xls", col_types = "numeric")
  p_SC <- read_excel("Pessoa03_SC.xls", col_types = "numeric")
  p_SE <- read_excel("Pessoa03_SE.xls", col_types = "numeric")
  p_SP1 <- read_excel("~/UFRGS/Trabalho de Conclusão de Curso/Censo
2010/SP_Capital_20171016/SP Capital/Base informaçoes setores2010 universo
SP_Capital/EXCEL/Pessoa03_SPCapital.xlsx",
                      col_types = c("numeric", "numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
                                    "numeric", "numeric", "numeric",
"numeric", "numeric",
```

```
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
"numeric", "numeric", "numeric",
```

```
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric", "numeric", "numeric",
"numeric", "numeric",
                                        "numeric"))

  p_SP2 <- read_excel("Pessoa03_SP2.xls", col_types = "numeric")
  p_TO <- read_excel("Pessoa03_TO.xls", col_types = "numeric")
  cor_pele <- rbind(p_SP1)

remove(p_AC,p_AL,p_AM,p_AP,p_BA,p_CE,p_DF,p_ES,p_GO,p_MA,p_MG,p_MS,p_MT,p_
PA,p_PB,p_PE,p_PI,p_PR,p_RJ,p_RN,p_RO,p_RR,p_RS,p_SC,p_SE,p_SP2,p_TO)
  str(cor_pele)

  # Cria as classes de raça da mesma forma que foi feito com os dados da
POF
  branca <- rowSums(charNumeric(cor_pele[,4]), na.rm = T)
  preta <- rowSums(charNumeric(cor_pele[,5]), na.rm = T)
  amarela <- rowSums(charNumeric(cor_pele[,6]), na.rm = T)
  parda <- rowSums(charNumeric(cor_pele[,7]), na.rm = T)
  indigena <- rowSums(charNumeric(cor_pele[,8]), na.rm = T)

  # Define Código do Setor Censitário
  SETOR <- cor_pele[,1]
  SETOR$Cod_setor <- as.character(idade$Cod_setor)

  # Agrupa todas as classes de idade num data frame temporário
  temp <- as.data.frame(SETOR)
  temp$branca <- branca
  temp$preta <- preta
  temp$amarela <- amarela
  temp$parda <- parda
```

```
  temp$indigena <- indigena
  str(temp)

  # Agrupa tudo na base censo.df
  censo.df <- merge(censo.df, temp, by="Cod_setor") # Usa a coluna do
setor como base para a mescla dos dados
  str(censo.df)
  str(valida.df)

  # Verifica e corrige os nomes nas bases para ficarem iguais
  names(valida.df)
  names(censo.df)

  names(censo.df) <- c("SETOR",
                       "Idade0a9",
                       "Idade10a19",
                       "Idade20a29",
                       "Idade30a39",
                       "Idade40a49",
                       "Idade50a59",
                       "Idade60a69",
                       "Idade70a99",
                       "Homem",
                       "Mulher",
                       "Renda1",
                       "Renda2",
                       "Renda3",
                       "Renda4",
                       "Renda5",
                       "Renda6",
                       "Renda7",
                       "Renda8",
                       "Renda9",
                       "alfabetizado",
                       "branca",
                       "preta",
                       "amarela",
                       "parda",
                       "indigena")
  names(censo.df)

  #Limpar os setores censitários de SP com informações incompletas
  censo.df.limpo <- censo.df[-c(242,    251,  252,  273,  284,  291,  296,
      303,  589,  619,  1273,
                                1276,   1657, 1658, 1679, 2026, 2036, 2045,
      2319, 2361, 2420, 2491, 4242, 4332, 4339, 5408, 5439,
                                5484,   5541, 5555, 5556, 5561, 5562, 5573,
      5575, 5576, 5578, 5579, 5583, 5585, 5587, 5590, 5602,
                                5604,   5605, 5606, 5607, 5608, 5609, 5611,
      5612, 5613, 5615, 5616, 5617, 5618, 5622, 5623, 5624,
                                5626,   5627, 5633, 5635, 5647, 6288, 6402,
      6451, 6571, 7086, 7213, 7248, 7453, 7586, 8011, 8053,
                                8081,   8108, 8150, 8156, 8160, 8171, 8516,
      8520, 8524, 8785, 8815, 9400, 9413, 9482, 9601, 10187,
```

```
                              10199,  10200,      10201,      10209,
    10219,      10222,      10225,      10459,      10463,      10551,
    10631,      10644,
                              10654,  10675,      10677,      10681,
    10692,      10693,      10705,      10714,      10716,      10726,
    10729,      10733,
                              10735,  10797,      10869,      11094,
    11105,      11107,      11113,      11634,      11659,      12633,
    12859,      13440,
                              13606,  13689,      13736,      14970,
    15028,      15108,      15112,      15583,      15602,      15629,
    15630,      15637,
                              15638,  15643,      15648,      15659,
    15805,      15850,      16591,      16715,      16910,      16976,
    17006,      17207,
                              17435,  17542,      17882,      17992,
    18080,      18102), ]

  # Converte os missing values em 0
  censo.df.limpo[is.na(censo.df.limpo)] <- 0
  str(censo.df.limpo)

  # Transforma em escala todos os valores exceto a coluna "SETOR"
  temp <- scale(censo.df.limpo[,-1])

  # Cria o data frame com a informação SETOR + valores numéricos em escala
  censoNorm.df.limpo <- data.frame(censo.df.limpo$SETOR, temp)
  names(censoNorm.df.limpo)[1] <- "SETOR" # Corrige o título da coluna
Setor
  str(censoNorm.df.limpo)

  # A partir das variáveis criadas. Cria-se a coluna GASTO para predizer o
valor gasto por setor censitário
  censoNorm.df.limpo$GASTO <- predict(svm,censoNorm.df.limpo)
  summary(censoNorm.df.limpo$GASTO)
  censoNorm_final.df <- censoNorm.df.limpo[,c("SETOR","GASTO")]
  str(censoNorm_final.df)

### Apresentação dos Resultados: Etapa realizada utilizando a internet
(para o download dos arquivos shapefile) e a programaçao abaixo: ###

  # Exporta tabela dos resultados por setor censitário
  setwd("C:/Users/Ricardo/Documents/UFRGS/Trabalho de Conclusão de
Curso/Exportação dos Dados do Censo 2010")

  # Nomeia o arquivo de saída e exporta para o Excel
  nomearquivo <- "Gasto-MaterialEletrico-SPClean.csv"
  write.table(censoNorm_final.df, sep = ";",dec = ",", row.names = F,
file= nomearquivo)

  #Agregar os setores censitários em distritos de SP
  #censoNorm_final.df.setores <- substr(censoNorm_final.df$SETOR, 1 , 9)

  # Define o local dos arquivos shapefile
```

```
  setwd("C:/Users/Ricardo/Documents/UFRGS/Trabalho de Conclusão de
Curso/Exportação dos Dados do Censo 2010/Shapefiles SP")

  # Prepara os packages

  library(rgdal)
  library(rgdal)
  library(sp)
  library(maptools)
  library(maptools)
  library(ggmap)
  library(RColorBrewer)
  library(rgeos)
  library(RgoogleMaps)
  library(spdep)
  library(plyr)
  library(Hmisc)

  # Lê os arquivos shapefile da região
  #distrito <- readOGR(".", "DEINFO_DISTRITO")
  #ponderacao <- readOGR(".", "DEINFO_AREA_PONDERACAO_2010")
  setor <- readOGR(".","DEINFO_SETOR_CENSITARIO_2010")

  #plot(distrito)
  #plot(ponderacao)
  plot(setor)

  #Mesclar predição de gastos por setor censitário com o shapefile
  setor.spdf <- merge(setor,censoNorm_final.df, by.x= "CODSETOR", by.y=
"SETOR")
  names(setor.spdf)
  head(setor.spdf)

  # Seleciona apenas a informação do Gasto
  gasto <- setor.spdf$GASTO
  gasto[is.na(gasto)] <- 0

  # Define o número de classes
  nClasses <- 5

  # Define as cores do mapa
  library(RColorBrewer)
  library(RColorBrewer)
  tCores = brewer.pal(nClasses,"Spectral")

  # Gera os intervalos
  # Faz os gastos ficarem agrupados em 5 classes (para facilitar a
visualização nos mapas)
  library(classInt)
  library(classInt)
  classe <- classIntervals(gasto, nClasses, style = "quantile")

  # Atribui cada cor a sua respectiva classe
  colcode <- findColours(classe, tCores)
```

```
  # Pinta o mapa
  plot(setor.spdf, col = colcode, border = "NA", axes = TRUE, alpha = .5)
  plot(setor, col = NA, border = grey(.5), lwd = 0.01, add = TRUE)
  title(main = "Gasto-MaterialEletrico-SP", xlab = "Longitude", ylab =
"Latitude")
```

Apêndice B - Programação utilizada no RStudio com alterações para gerar valores em reais

```
# Modificações

# códigos de produtos do estudo
codigosdoestudo <-
c("0801201","0801202","0801203","0803801","1101201","1101202","1101203",
                    "1103801", "8605501","8605502","8605503", "8605504",
                    "8605505", "8605506", "8605507")

#(...)

## Tarefa: Transformar os valores em escala. Média = 0 ##
# Define a Base de treinamento
treina.df <- tudoLimpa.df[which(sorteio==1),] # Seleciona apenas as linhas
nas quais o valor do vetor "sorteio" é 1
treina.df.scaled <- scale(treina.df)
treina.df.mean <- attributes(treina.df.scaled)$'scaled:center' #the mean
treina.df.std <- attributes(treina.df.scaled)$'scaled:scale'   #the
standard deviation
head(treina.df.scaled,10)

#(...)

## Tarefa: Aplicar o treinamento na base de validação ##
# Cria a base de Validação com as linhas não inclusas na base de
treinamento
# Essa base serve para testar se o treinamento está correto e aperfeiçoá-
lo
valida.df <- tudoLimpa.df[which(sorteio!=1),]
valida.df <- as.data.frame((scale(valida.df)))

# Cria a lista de parâmetros para a busca
listaC <- seq(0.01, 10, length.out = 50)
listaSigma2 <- c(seq(0.01,10, length.out = 25), seq(0.01,10,length.out =
25))
parms <- expand.grid(C = listaC, sigma = listaSigma2)

# Para cada par de parâmetros pegue aquele com o menor EQM
apply_pb <- function(X, MARGIN, FUN, ...)
{
  env <- environment()
  pb_Total <- sum(dim(X)[MARGIN])
  counter <- 0
  pb <- txtProgressBar(min = 0, max = pb_Total, style = 3)

  wrapper <- function(...)
  {
    curVal <- get("counter", envir = env)
    assign("counter", curVal +1 , envir = env)
    setTxtProgressBar(get("pb", envir = env), curVal +1)
    FUN(...)
  }
```

```
  res <- apply(X, MARGIN, wrapper, ...)
  close(pb)
  res
}

# Realiza o treinamento da máquina e compara os resultados reais com os
preditivos
head(treina.df.scaled)

EQM <- apply_pb(parms,1,function(x)
{
  svm <- ksvm(gastodoestudo ~.,data = treina.df.scaled,
kernel="rbfdot",type="eps-svr",
              kpar=list(sigma=as.numeric(x[2])),C=as.numeric(x[1]),scaled
              =FALSE);mean((valida.df$gastodoestudo-
predict(svm,valida.df)^2))
}
)

# Junta todos os resultados
parms$EQM <- EQM

# Obtém os parâmetros de menor valor
summary(parms$EQM)
hist(parms$EQM)
iMin <- which(parms$EQM==min(parms$EQM))

# Resultado final
parms[iMin[1],]
C <- as.numeric(parms[iMin[1],][1])
sigma <- as.numeric(parms[iMin[1],][2])

# C <- 1
# sigma <- 0.01

# Realiza o treinamento com os valores ótimos a partir dos parâmetros
criados
library(kernlab)
svm <- ksvm(gastodoestudo~.,
            data = treina.df.scaled,
            kernel="rbfdot",
            type="eps-svr",
            kpar=list(sigma=sigma),
            C=C,
            scaled=FALSE)

# Teste

rmse <- function(error)
{
  sqrt(mean(error^2))
}

error <- EQM$residuals  # same as data$Y - predictedY
```

```
predictionRMSE <- rmse(error)

#(...)

## Tarefa: Aplicar nos dados do Censo o resultado do treinamento realizado
anteriormente
# A partir das variáveis criadas. Cria-se a coluna GASTO para predizer o
valor gasto por setor censitário
censoNorm.df.limpo$GASTO <- predict(svm,censoNorm.df.limpo) *
treina.df.std [1] + treina.df.mean [1]
summary(censoNorm.df.limpo$GASTO)
censoNorm_final.df <- censoNorm.df.limpo[,c("SETOR","GASTO")]
str(censoNorm_final.df)
```

Apêndice C - Programação utilizada no RStudio para gerar gráfico dos distritos

```
# Prepara os packages

  library(rgdal)
  library(sp)
  library(maptools)
  library(maptools)
  library(ggmap)
  library(RColorBrewer)
  library(rgeos)
  library(RgoogleMaps)
  library(spdep)
  library(plyr)
  library(Hmisc)

  #Importa Dataframe com médias dos distritos
  library(readxl)
  Gasto_Distritos_Clean <- read_excel("~/UFRGS/Trabalho de Conclusão de
Curso/Exportação dos Dados do Censo 2010/Gasto Distritos Clean.xlsx",
                                      range = "B1:C97", col_types =
c("numeric",

"numeric"))

  # Lê os arquivos shapefile da região
  distrito <- readOGR(".", "DEINFO_DISTRITO")
  ponderacao <- readOGR(".", "DEINFO_AREA_PONDERACAO_2010")
  setor <- readOGR(".","DEINFO_SETOR_CENSITARIO_2010")

  plot(distrito)

  # Agrupa os dados dos gastos de materiais elétricos, distribuídos no
mapa de São Paulo
  setor.spdf <- merge(distrito,Gasto_Distritos_Clean, by.x= "COD_DIST",
by.y= "COD_DISTRITO")
  names(setor.spdf)
  head(setor.spdf)


  # Seleciona apenas a informação do Gasto
  gasto <- setor.spdf$GASTOCLEAN
  gasto[is.na(gasto)] <- 0

  # Define o número de classes
  nClasses <- 5

  # Define as cores do mapa
  library(RColorBrewer)
  tCores = brewer.pal(nClasses,"Spectral")

  # Gera os intervalos
```

```
  # Faz os gastos ficarem agrupados em 5 classes
  library(classInt)
  library(classInt)
  classe <- classIntervals(gasto, nClasses, style = "quantile")

  # Atribui cada cor a sua respectiva classe
  colcode <- findColours(classe, tCores)

    # Pinta o mapa
  plot(setor.spdf, col = colcode, border = "NA", axes = TRUE, alpha = .5)
  title(main = "Gasto-MaterialEletrico-SP", xlab = "Longitude", ylab =
"Latitude")
```