UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

FILIPE SOUZA DOS SANTOS

Sistema para descobrimento e visualização de informações utilizando índices sociais de bairros e georreferenciamento no município de Porto Alegre

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Engenharia da Computação

Orientador: Prof. Viviane Pereira Moreira

Porto Alegre
2018

AGRADECIMENTOS

À minha mãe, Maria Regina, pela vida, pela amizade mais verdadeira do mundo, pelo amor e por sempre me dizer que tudo vai dar certo. E tudo sempre dá certo. Ao meu pai, Nilton, por sempre me demonstrar que a academia não é a fonte máxima de conhecimento e que a observação da natureza, dos animais e das plantas é o que realmente nos torna mais humanos.

Aos meus 5 irmãos que me acompanham desde os meus primeiros minutos nesta vida. Diego, obrigado por sempre cuidar de mim e me oferecer todo apoio necessário sem medir esforços para que eu completasse esta graduação. Rodrigo, as nossas conversas me fazem ter muito orgulho de quem sou hoje e de como eu vejo o mundo. Lisandra e Luciane, minhas segundas mães, amigas para todo o sempre, eu não seria nada sem vocês. E, Márcio, o mais velho, aquele que traz segurança a todos, que sempre na medida do possível se fez presente para que, todos juntos, pudéssemos dar boas risadas sobre o passado. Como sou grato por vocês cinco. Ninguém teve sorte maior que a minha.

Aos meus 9 sobrinhos, Guilherme, Tales, Vitória, Evelin, Francielly, Pedro, Tainá, Tiago e Luísa por sempre mostrarem que as gerações mais novas carregam com si o novo. E por me lembrarem que é a elas a quem pedimos o mundo emprestado.

À Nicoli Bonalume, meu amor, que me acompanha desde o início e que conheceu todos os desafios e momentos de ansiedade. Como sou grato por compartilhar a vida contigo. Olhando para trás, nós evoluímos tanto como pessoas. Agradeço por cada momento ao teu lado.

A todos meus amigos, sejam eles de infância, graduação ou mundo. Eu seria injusto em tentar citá-los, pois certamente esqueceria de alguém. Mas gostaria de, em nome de todos que não citei, agradecer ao meu querido amigo João Paulo Perbiche, por todo o apoio durante nossa jornada na França, em 2014. Obrigado, meu amigo.

À Prof. Viviane, por muito me acompanhar durante o TG1 e TG2 e que, mesmo à distância, ainda continuava compartilhando todo seu conhecimento. Sempre buscou me apresentar para novas pessoas e sempre compreendeu mudanças de escopo deixando-me explorar livremente os assuntos de minha curiosidade.

Ao Prof. Valter Roesler, por muitos anos compartilhando sua experiência com tantos jovens no laboratório PRAV ao qual tenho muito orgulho de ter feito parte. E, por fim, ao INF UFRGS, pela excelência de ensino. A universidade me transformou e todos os funcionários participaram disso.

RESUMO

A visualização de informações é uma ferramenta importante para a tomada de decisões no setor público. O uso de mapas é um dos exemplos de como um conjunto de dados pode ser analisado comparativamente, ter sua visualização sobreposta a regiões conhecidas por determinada caraterística e, como consequência, oferecer uma melhor compreensão por parte do analisador. Por outro lado, devido às características específicas de cada bairro de uma cidade, o número de variáveis como, por exemplo, renda per capita, número de postos de saúde, taxa de arborização, acesso à saneamento básico, dentre muitos outros fatores, tende a um número muito grande dificultando-se, assim, o processo de análise. O sistema proposto visa dado qualquer conjunto de pontos georreferenciados, extrair automaticamente relações entre o conjunto de entrada e indicadores socioeconômicos de bairros de Porto Alegre, sendo capaz, ao fim da análise, de mostrar, no mapa, se os pontos de entrada possuem uma alta correlação com determinado indicador utilizando algoritmos de data mining presentes no software WEKA. A fim de se testar e aprimorar o sistema, um caso prático de utilização é discutido ao longo do documento buscando-se elucidar visualmente no mapa a relação entre índices socioeconômicos e estabelecimentos previamente classificados que comercializam alimentos predominantemente ultraprocessados, predominantemente *in natura*, além de feiras e feiras orgânicas. Para feiras e feiras orgânicas - que significam acesso a alimentos frescos para as populações dos bairros - constatou-se uma correlação média de $0,4441$ e $0,483808$, respectivamente, para indicadores ligados ao número de habitantes indicando uma forte tendencia a zonas Centrais e de $0,2893$ e $0,3183$, respectivamente, para fatores como renda per capita média e escolaridade indicando uma forte relação com bairros socioeconomicamente mais favorecidos.

Palavras-chave: Visualização de informações. Descoberta de conhecimento. WEKA.

ABSTRACT

Information visualization is an important tool for the public sector. The use of maps is one example of how a set of data can be comparatively analyzed having their view superimposed on well known regions and, as a consequence, offering a better point of view to the analyzer. On the other hand, due to the specific characteristics of each neighborhood of a city the number of variables such as per capita income, number of health services, afforestation rate, access to basic sanitation, among many others factors, tends to a very large number, thus making harder the analysis process. The proposed system aims at giving any set of geo-referenced points to extract relations between it and socioeconomic indexes of Porto Alegre city neighborhoods being able at the end of the analysis to show on the map whether the points have a high correlation with a given index using data mining algorithms available on WEKA software. In order to test it and to improve the system, a practical case of use is discussed throughout the document seeking to elucidate visually on the map the relationship between socioeconomic indices and food shops which sells predominantly ultraprocessed foods, in natura ones, besides fairs and organic fairs. For fairs and organic fairs - which means access to fresh food - a mean correlation of 0.4441 and 0.483808, respectively, was found linking population indexes mainly indicating a strong tendency of these classes being part of central zones. Also, a correlation of $0,2893$ and $0,3183$, respectively, was found for factors such as average per capita income and schooling indicating a strong relationship with socioeconomically favored neighborhoods.

Keywords: Information Visualization. Data Mining. WEKA.

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
PNUD	Programa das Nações Unidas para o Desenvolvimento
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Aplicações que utilizam *data mining* - mineração de dados, em português - voltadas ao bem público e social são cada vez mais comuns e exploram um grande número de áreas tais quais saúde, educação, finanças públicas, dentre outras. Especificamente, as aplicações vão desde predições relacionadas ao diabetes (KAVAKIOTIS et al., 2017), passando pela dimensão das cidades através da predição da necessidade de manutenção em redes elétricas (RUDIN et al., 2012) até a dimensão de bairros na previsão de características da população através do Google Street View (GEBRU et al., 2017), por exemplo. A popularização de ferramentas comerciais auxilia no desenvolvimento destas novas aplicações como, por exemplo, o Watson [1], da IBM, ou o compilado de ferramentas da plataforma Google AI [2]. Do ponto de vista de ferramentas gratuitas para resolução de problemas, softwares como o WEKA (FRANK; HALL; WITTEN, 2016) [3] e bibliotecas como o Scikit-learn[4] ganham espaço dentro da comunidade acadêmica.

O volume de dados exigido para tais procedimentos é, em geral, bastante expressivo. Porém, revisões sistemáticas apontam que, assim como em outros países, o acesso a dados públicos no Brasil carrega em si o ônus para o pesquisador de um grande esforço prévio de tratamento dos dados e sua aquisição antes mesmo da pesquisa ser iniciada (OLIVEIRA; SILVEIRA, 2018). Assim, de um lado há uma necessidade de um grande volume de informações, porém, de outro, há uma imensa dificuldade em se ter acesso às mesmas.

Uma outra área importante nesse contexto é a visualização de informações. Conforme sugere Paula et al. (2011), há vários mecanismos para que a informação tenha um sentido de forma mais rápida e clara, mostrando-se que a visualização através de mapas, por exemplo, pode ser melhor entendida e mais bem aceita do que uma visualização em gráficos de colunas, dependendo do contexto em que se está sendo feita uma análise. Atualmente, várias ferramentas comerciais e abertas dão vazão à visualização dos dados por meio de mapas. A popularização de ferramentas como o Google Maps [5] e o fornecimento de APIs - *Application Programming Interface* - para utilização destes serviços colaboram nesse desenvolvimento. Na prática, várias iniciativas de mapeamentos - inclusive muitas vindas de cidadãos comuns - têm surgido na cidade de Porto Alegre como, por exemplo, o

[1] https://www.ibm.com/watson/
[2] https://ai.google/
[3] https://www.cs.waikato.ac.nz/ml/weka/
[4] https://scikit-learn.org/stable/
[5] https://www.google.com/maps

mapeamento de feiras orgânicas [6], de pontos de maior violência na cidade [7], ou de locais propícios para se deixar uma bicicleta em segurança [8]. Já em um aspecto mais voltado à visualização de indicadores sociais através de mapas, outras tantas iniciativas - estas mais institucionalmente fortalecidas - também foram criadas nos últimos anos como, por exemplo, o Atlas Brasil (FJP; IPEA; PNUD, 2013) que traz indicadores bairro a bairro das principais metrópoles do país.

O presente trabalho se propõe a reunir as áreas citadas acima - *data mining*, dados públicos e visualização de informações através de mapas - na tentativa de se descobrir novos conhecimentos a partir de dados públicos e compreendê-los de forma interativa através de um sistema web disponível para o acesso irrestrito em navegadores web. *Objetivamente, dado um conjunto de entrada de pontos georreferenciados e classificados previamente por um pesquisador, descobrir relações entre os pontos e indicadores sociais dos bairros da cidade de Porto Alegre*. As classes dependem do interesse do pesquisador. Os indicadores são adquiridos através da base do Atlas Brasil[9] executado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD)[10] que expõe mais de 260 variáveis como renda per capita média do bairro, taxa de analfabetismo, IDH, dentre outras. Alguns casos de utilização serão descritos como a análise automática de mais de 15.000 estabelecimentos que comercializam alimentos em Porto Alegre com o objetivo de descobrir relações entre, por exemplo, estabelecimentos que comercializam predominantemente alimentos *in-natura* em contraste com renda per capita ou qualquer outro indicador social dos bairros da cidade.

Do ponto de vista de organização do documento, a Seção 2 trará uma análise mais profunda das referências pesquisadas previamente para que se estabelecessem as bases deste trabalho. A Seção 3 evidencia a metodologia aplicada envolvendo os diversos processos necessários para se atingir o objetivo final do sistema: aquisição dos dados, tratamento, utilização das funções da ferramenta de mapa escolhida, e processo a ser executado para o descoberta de conhecimento. Na Seção 4 do documento, é proposto, primeiramente, um modelo genérico para que, dado qualquer conjunto de dados georreferenciados e classificados, se possa realizar a descoberta e visualização de informações. Num segundo momento, o modelo é então aplicado a um caso real para que validações e melhorias possam ser discutidas. A Seção 4.3 faz uma discussão sobre o caso prático que

[6] https://feirasorganicas.org.br/
[7] http://www.ondefuiroubado.com.br/porto-alegre/RS
[8] https://www.bikedeboa.com.br/
[9] http://www.atlasbrasil.org.br/2013/
[10] http://atlasbrasil.org.br/2013/

foi aplicado à descoberta de conhecimento a partir de uma base de estabelecimentos comerciais que foram classificados de acordo com seu perfil de venda em cinco classes: (i) predominantemente ultraprocessados, (ii) predominantemente in natura, (iii) mistos, (iv) feiras e (v) feiras orgânicas. A Seção também aborda tentativas que foram feitas ao longo do desenvolvimento da aplicação e também sobre a qualidade dos dados que foram escolhidos. Por fim, conclui-se o presente trabalho evidenciando-se dificuldades, propondo-se melhorias e estabelecendo-se possibilidades de trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 Indicadores Socioeconômicos

A distribuição social da renda é extremamente concentrada em uma parcela da população mundial (YUNES, 2000). De acordo com uma pesquisa produzida pela Universidade das Nações Unidas, em 2006, 10% de pessoas no mundo detinham 85% da riqueza mundial, enquanto que a metade mais desfavorecida detinha menos de 1% do total da riqueza (ONU-WINDER, 2006). Essa concentração de renda, aumenta as desigualdades socioeconômicas entre os grupos, provocando situações que implicam em algum grau de injustiça, isto é, desigualdades que são injustas porque estão associadas a características sociais que colocam alguns grupos em desvantagem (BARATA, 2009).

A desigualdade socioeconômica é um dos grandes determinantes do processo de saúde-doença das populações, estando inclusive associada ao risco de morte com uma redução na expectativa de vida, independentemente dos fatores de risco convencionais, como alta ingestão de álcool, inatividade física, tabagismo, hipertensão, diabetes e obesidade (STRINGHINI, 2017). Portanto, estudar os indicadores socioeconômicos é essencial para o planejamento de políticas públicas nas mais diversas áreas e setores do governo, principalmente na área da saúde. Além disso, áreas que antes eram pautadas sob um ponto de vista exclusivamente clínico e de culpabilização individual passam a compreender a complexidade que envolve os indivíduos, desde os fatores econômicos, sociais, ambientais, culturais, psicológicos e comportamentais (GEORGE, 2011; PORTRAIT; LINDEBOOM; DEEG, 2001; WHITEHEAD, 2000). Um exemplo disso é o novo Guia Alimentar para a População Brasileira (SAÚDE, 2014) que aborda questões socioambientais e tenta dar menos importância à quantidade de calorias ou proteínas dos alimentos e um foco maior em seus tipos de acordo com a classificação NOVA criada por Monteiro et al. (2016): ultraprocessados, processados, minimamente processados ou *in natura*. De acordo com a classificação, entende-se ultraprocessados por alimentos constituídos por formulações industriais com normalmente mais de cinco ingredientes e com frequente adição de estabilizantes e conservantes como, por exemplo, refrigerantes, biscoitos, bolos, alimentos congelados prontos, salsicha e hambúrguer. Já processados normalmente contém até três ingredientes e adição de sal, açúcar ou vinagre. Exemplos deste tipo são queijo, pães e conservas. Por último, *in natura* ou minimamente processados são alimentos frescos, consumidos logo após sua retirada da natureza. Partes comestíveis de plantas

(frutos, folhas, caules, raízes) e de animais entram neste tipo. Todos estes conceitos serão utilizados ao longo do texto e, assim, se fez importante suas respectivas definições.

Ainda neste contexto, a desigualdade socioeconômica também influencia na aquisição e consumo de alimentos. Embora possa haver muitas opções de alimentos nas cidades, muitas pessoas não possuem acesso físico ou econômico a esses alimentos (HLPE, 2017). Além disso, alimentos nutritivos não são acessíveis em muitos bairros de baixa renda, denominando-se estes locais de desertos ou pântanos alimentares. Em uma pesquisa acerca do ambiente alimentar realizada por Junior et al. (2018) é sugerida uma classificação para os estabelecimentos do Rio de Janeiro de acordo com o Guia Alimentar para a População Brasileira e, a partir disso, os autores tentam associar a renda per capita à taxa de concentração de estabelecimentos de um tipo ou de outro. Em linhas gerais, a classificação sugere, por exemplo, que açougues vendem predominantemente alimentos *in natura* enquanto que bares e lancherias vendem predominantemente alimentos ultra-processados e assim sucessivamente para outros tipos de atividades. A partir disso, com dados obtidos a partir de alvarás, o referido trabalho realiza a classificação de cerca de 9.000 estabelecimentos e, manualmente, são verificadas relações entre a renda dos bairros e a concentração de estabelecimentos por tipos. É sugestivo, portanto, que outras descobertas podem ser feitas com base em pontos georreferenciados (isto é, todo ponto que possua latitude e longitude) quando associado a indicadores socioeconômicos.

2.2 Bases de dados de Indicadores Socioeconômicos

Para a obtenção de indicadores separados por bairros, duas fontes de dados destacaram-se em relação às demais. Em se tratando de dados demográficos, o Instituto Brasileiro de Geografia e Estatística (IBGE)[1] apresenta uma enorme quantidade de informações. O próprio website apresenta uma série de mapas interativos, dos quais, visualmente, é possível que se extraia conhecimento. Como pontos negativos, pode-se citar uma difícil exportação dos dados e uma difícil exportação das coordenadas geográficas que limitam cada bairro. Em se tratando de dados demográficos aliados a dados mais voltados à qualidade de vida humana nas metrópoles como IDH (Índice de Desenvolvimento Humano) ou GINI (utilizado para medir desigualdade dentro de um bairro), a base de mais de 200 variáveis do PNUD, mostrou-se bastante completa contemplando as delimitações geográficas de cada bairro. O levantamento desta base de dados foi feito em 16 regiões metropolitanas con-

[1] https://censo2010.ibge.gov.br/

tando com mais de 300 pessoas envolvidas. Desta forma, torna-se viável explorar-se estas fontes e associá-las a pontos geográficos para a obtenção de conhecimento relacionado a estilo de vida, hábitos de consumo, qualidade de vida, dentre outras características dos bairros em questão.

2.3 *Data Mining e Descoberta de Conhecimento*

Conforme Fayyad, Piatetsky-Shapiro and Smyth (1996), o descobrimento de conhecimento em base de dados - do inglês *Knowledge Discovery in Database* (KDD) - consiste no processo de fazer um volume grande de informações primeiramente ter sentido seja para geração de um relatório ou para criação de um modelo mais sofisticado de predição, por exemplo. Para o autor, *data mining* é um passo a ser utilizado no processo de KDD com a aplicação de algoritmos de modo a produzir um conjunto de padrões em cima dos dados extraídos ou modelos.

Para a automatização dos processo de descoberta de conhecimento, três tipos de algoritmos - classificação, regressão, e seleção de atributos - e duas ferramentas - Scikit-learn e WEKA - foram mais profundamente estudados.

2.3.1 Algoritmos Mineração de Dados

A primeira distinção importante é entre métodos supervisionados e não supervisionados. Conforme Luxburg and Schoelkopf (2008) métodos supervisionados utilizam-se da observação de dados que foram previamente classificados para que então com novos dados tenha-se a capacidade de se definir a qual classificação estes pertencem. Já métodos não supervisionados não contém uma classificação previa e, portanto, buscam agrupar os dados analisados em grupos próximos extraindo-se, então, informações de proximidade do conjunto de entrada em questão.

Uma segunda distinção importante é entre algoritmos de classificação, regressão e seleção de atributos. Classificação, segundo (SINGHAL; JENA, 2013), é um método utilizado para, a partir de dados de entrada que contém *labels*, ou seja, marcações, possa-se ser possível predizer quais marcações devem receber dados que ainda não estão marcados. Quando os valores de predição não são mais do domínio discreto, porém contínuo, não temos mais um problema de classificação, mas sim de regressão. Com a regressão,

portanto, não se é do interesse classificar entre N pontos discretos, mas, sim, predizer-se um valor contínuo em função das variáveis de entrada. Em linhas gerais, dado um conjunto $X_1, X_2...X_n$, uma vez aprendida uma função $f(X_1, X_2...X_n) \rightarrow Y$ se Y é discreta, então temos um problema de classificação. Se Y é contínua, então temos um problema de regressão. Por fim, métodos de Seleção de Atributos, também conhecidos em inglês por *Feature Selection*, não tem como resultado final um modelo de predição, mas sim o conjunto de variáveis de entrada, como citado no exemplo acima - $X_1, X_2...X_n$ - que são mais relevantes. Esse processo auxilia na redução do número de variáveis eliminando ruídos e simplificando modelos posteriores construídos a partir desta seleção (SHEENA; KUMAR; KUMAR, 2016)

Devido ao fato deste trabalho não buscar um modelo de predição, seja ele de classificação ou de regressão, mas sim uma busca pelas variáveis que têm maior relação com a classificação que foi previamente estabelecida - leia-se aqui: encontrar quais índices socioeconômicos tem maior relação com os pontos classificados e georreferenciados -, é natural que o interesse deste documento gira em torno de algoritmos de Seleção de Atributos.

2.3.2 Ferramentas de Mineração de Dados

No domínio de ferramentas gratuitas, duas alternativas foram analisadas. A primeira, *Scikit-learn*[2], é uma biblioteca contendo algoritmos supervisionados e não supervisionados para linguagem Python. A ferramenta é extremamente completa contendo algoritmos de classificação, regressão, *clustering* e de seleção de atributos. É mantida através de código aberto e possibilita a comercialização de aplicações desenvolvidas com base nela. A documentação é bastante vasta, tal qual o suporte da comunidade possuindo, inclusive, uma API facilitando a integração com outras aplicações.

Já o WEKA (FRANK; HALL; WITTEN, 2016) contém exatamente as mesmas funcionalidades, porém é um software executado em Java. Apesar de poder ser também executado através de linhas de comando, sua integração não é tão facilmente executada quanto à API fornecida pela Scikit-learn. Sua vantagem reside em uma comunidade bastante forte e participativa, além do fato de, em poucos minutos, já possibilitar a um usuário, sem escrever nenhuma linha de código, realizar análises e executar algoritmos de pré-processamento de dados, classificação, associação e clustering.

[2] scikit-learn.org/stable/modules/svm.html#svm-classification

Conforme exposto na Seção 2.3.1 que evidencia o interesse em métodos de Seleção de Atributos, torna-se necessária uma análise em termos do que a biblioteca Scikit-learn e a ferramenta WEKA apresentam. Do lado da Scikit-learn, há cinco abordagens[3]: *Removing features with low variance*, *Univariate feature selection*, *Recursive feature elimination*, *Feature selection using SelectFromModel*, *Feature selection as part of a pipeline*. A primeira abordagem basicamente verifica a variância $Var[X] = p(1-p)$ de um atributo. Caso ela seja menor do que um limiar especificado pelo programador, então o atributo é eliminado. Portanto, toda variável que tem seus valores repetidos diversas vezes e, logo, apresenta baixa variância, é eliminada. A segunda abordagem permite que seja passada uma função simples que tratará apenas uma variável por vez e retorna um número escolhido - *KBests* - de melhores variáveis de acordo com a função que foi passada. Já a terceira função, *Recursive feature elimination*, utiliza um processo igual à regressão linear, dando pesos inicialmente a um grande número de variáveis e, recursivamente, diminuindo este número conforme se aproxima de um modelo melhor. Por fim, retorna os n - tal que n é um valor escolhido - atributos que ainda não foram eliminados. A penúltima permite que se passe qualquer classificador ou algoritmo de regressão e que, a partir do modelo gerado, se obtenha os atributos mais importantes. Por fim, a última abordagem não se trata exatamente de um método de seleção, mas sim de um *pipeline* - conjunto de passos sequenciais - que podem ser descritos pelo programador, ou seja, pode-se, por exemplo, concatenar qualquer um dos quatro métodos citados anteriormente para se chegar em um conjunto de atributos finais.

Em se tratando do WEKA, um número também razoável de algoritmos de seleção de atributos é disponibilizado. São eles: *CfsSubsetEval*, *ClassifierAttributeEval*, *CorrelationAttributeEval* e *InfoGainAttributeEval*. O primeiro tenta escolher um conjunto de atributos que apresente alta correlação, $\rho_{x,y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, com a classe - ou seja, se há um indicativo de que dado o crescimento ou decrescimento do atributo há também variação na classe - e baixa correlação entre os demais atributos. O segundo algoritmo permite que um classificador seja passado como parâmetro e, com base nos melhores resultados, os melhores atributos são devolvidos. O terceiro também leva em consideração a correlação com a classe, porém sem eliminar atributos que são fortemente correlacionados entre si. Por último, o algoritmo de *InfoGainAttributeEval* verifica qual atributo traz o maior ganho de informação, ou seja, reduz a entropia - desinformação - dos dados.

[3]https://scikit-learn.org/stable/modules/feature_selection.html

2.4 Ferramentas para visualização de informações em Mapas

Em linhas gerais, segundo Peterson (2015), uma ferramenta, ou API, para visualização de mapas normalmente se preocupará em, primordialmente, fornecer pedaços - ou *tilings* em inglês - do mapa como um todo, ou seja, aos poucos, conforme a necessidade do usuário, são enviados pedaços que possam ser combinados ao chegar no destino para dar a impressão de que o tempo de carga da imagem, ou seja, do trecho de mapa, foi rápido. O texto ainda avalia 8 APIs bastante utilizadas na data de escrita, sendo Google Maps a mais utilizada e conhecida segundo os autores. Dentre estas, Bing Maps API, Nokia HERE API, Map Quest API, Baidu Map API e Mapstraction aparentemente estão descontinuadas ou possuem documentação bastante escassa e ratificam um ponto discutido pelos autores referente à viabilidade a longo prazo das API devido ao custo de atualizações dos mapas.

Assim, Google Maps, Open Street Maps e Leaflet aparecem como alternativas para novas aplicações com o porém da API Open Street Maps apresentar menor performance dentre elas. Das que restam, Google Maps API tem caráter comercial enquanto que Leaflet tem caráter de código aberto. Uma terceira alternativa, mais recente e comercial, é a API MapBox, criada pela empresa UBER[4]. Todas apresentam funcionalidades que são fundamentais para a visualização de informações em mapas: alta performance em navegadores convencionais, acesso às camadas - ou *layers* em inglês - para coloração e outras estilizações, possibilidade de adicionar marcadores, polígonos e multipolígonos e, por fim, algoritmos embutidos de clusterização de pontos para que, mesmo ao se trabalhar com muitos pontos - por exemplo, 400 mil pontos, ainda assim se obtenha uma alta performance devido ao processo de agrupamento.

Outra distinção que se faz importante é o preço. Tanto Google Maps API quanto MapBox possibilitam milhares de carregamentos de mapas, ou seja, a primeira vez em que a página é carregada não importando navegações, de forma gratuita. Apenas como ponto de referência, a API MapBox oferece 50.000 carregamentos de mapa gratuitamente[5] enquanto que a API Google Maps oferece U$$200,00 dólares iniciais em créditos tendo um custo de aproximadamente U$$2,00 dólares a cada mil carregamentos. Todas as três plataformas - incluindo-se Leaflet que é gratuita - demonstram um custo muito próximo de zero cabendo ao programador a escolha de sua preferência dependendo, por exemplo, da

[4] https://www.mapbox.com/
[5] https://www.mapbox.com/pricing/

linguagem ou *framework* utilizado na implementação do aplicação.

3 METODOLOGIA

O presente trabalho descreve o desenvolvimento de um sistema *web*, baseado em algoritmos de análise de dados, para a visualização da correlação entre indicadores socioeconômicos e dados georreferenciados do município de Porto Alegre. Para testar o sistema, este foi aplicado ao ambiente alimentar da cidade em questão. O trabalho foi dividido em duas partes: (A) desenvolvimento de um modelo genérico e (B) aplicação do sistema a um caso específico.

3.1 Desenvolvimento de um modelo genérico

O desenvolvimento do sistema genérico foi dividido em três etapas: (1) fonte de dados, (2) linguagens de programação e (3) ferramentas.

3.1.1 Fonte de dados

Para os indicadores socioeconômicos, optou-se pela base de indicadores sociais provenientes do Atlas Brasil em detrimento da base oferecida pelo IBGE[1] uma vez que a primeira não se concentra apenas em informações demográficas, mas também em questões mais subjetivas como IDH e marcadores de vulnerabilidade além de granularizar os bairros da cidade de Porto Alegre de maneira a expor regiões menos favorecidas como vilas por mais que façam parte de bairros socioeconomicamente mais favorecidos. Foi consultado o próprio site do Atlas Brasil na aba de Downloads[2]. Um conjunto de planilhas foi então sanitarizada e salva em um banco MySQL em um servidor de dados RDS da Amazon[3].

3.1.2 Linguagens de programação

Para o *frontend* da aplicação foi escolhida a biblioteca ReactJs[4] em Javascript que possibilita uma forma simples e com alta performance para manipulação de páginas em

[1] https://censo2010.ibge.gov.br/
[2] http://atlasbrasil.org.br/2013/pt/download/
[3] https://aws.amazon.com/pt/rds/
[4] https://reactjs.org/

HTML. Além disso, a biblioteca trabalha com o conceito de componentes - diferentemente do conceito de classes, por exemplo, - e todas as três APIs de mapas apresentadas na Seção 2 possuem seus próprios componentes facilmente integráveis.

Já o *backend* foi feito em NodeJs[5] e contém basicamente chamadas para consultas ao banco MySQL além de algumas sanitarizações de dados. Ambos os serviços estão hospedados em servidores da Amazon, o que possibilita um rápido acesso aos recursos - fator importante em visualizações interativas, e disponíveis para qualquer usuário interessado.

Todos os serviços foram hospedados na AWS - Amazon Web Services por apresentar uma alta taxa de transferência de dados - levando-se em conta que o número de pontos pode ser grande dependendo da base de dados de pontos georreferenciados.

3.1.3 Ferramentas

Para a aplicação dos algoritmos de *Feature Selection* foi utilizado o software WEKA pelo fato do *backend* ter sido feito em NodeJs e a ferramenta em questão permitir chamadas através de linha de comando. O processo consistiu em gerar um arquivo .arff - esperado pelo WEKA - de modo dinâmico através da biblioteca *ARFF parsing and formating*[6], executar o software WEKA especificando o algoritmo de Seleção a ser utilizado e o arquivo .arff gerado e, por fim, coletar os resultados.

Para a visualização do mapa, foi escolhida a API Mapbox em detrimeto da API Google Maps e Leaflet. Cabe lembrar que todas apresentam um custo próximo ou igual à zero. Um fator de desempate foi o fato da API Mapbox ter um componente especial para Javascript utilizando a biblioteca ReactJs - já citada anteriormente e utilizada para o desenvolvimento geral do *frontend* - feita pela própria equipe da Mapbox. Além disso, não necessitava de uma hospedagem própria tal qual necessita a Leaflet.

3.2 Aplicação do sistema a um caso específico

Baseado na tese de Junior et al. (2018) que estudou o sistema alimentar do município do Rio de Janeiro e dada a inexistência de um estudo parecido no município de Porto Alegre e também pela relevância social do tema, optou-se pela base proveniente da Cia de

[5]https://nodejs.org/en/
[6]https://www.npmjs.com/package/arff

Processamento de Dados do Município de Porto Alegre (PROCEMPA)[7] que contém mais de 15.000 estabelecimentos que comercializam alimentos. Para aquisição dos dados, uma série de e-mails foram trocados com a instituição de modo que toda a base pudesse ser acessada através de uma planilha. Uma etapa de limpeza de dados foi feita manualmente a fim de retirar todos os estabelecimentos que não tinham ligações com alimentação com base em seus tipos de atividade.

Posteriormente, seguindo a classificação específica para estabelecimentos comerciais proposta pelo autor da tese em questão (Tabela 3.1) com base na classificação de alimentos NOVA criada em 2016 por Monteiro et al. (2016) foram aplicados *labels* para cada estabelecimento e salvos na base de dados RDS da Amazon para consultas posteriores. Uma diferença importante foi aplicada: além de estabelecimentos que comercializam alimentos predominantemente *in natura* adicionaram-se duas novas marcações: feiras e feiras orgânicas. O objetivo é segmentar ainda mais a base para que *insights* sobre o acesso a alimentos frescos pudessem ser explorados.

Tabela 3.1: Categorização de Estabelecimentos

GRUPO	EXEMPLOS	LABEL
Predomin. *in natura*	Açougue, fruteiras e peixarias	NATURA
Padrão Misto	Bar, churrascaria, mercado, mercearia, padaria	MIX
Predomin. Ultraprocessados	Cantina, confeitaria, lanchonete, pizzaria	ULTRA
Feiras	Feiras convencionais, feiras modelo	FEIRA
Feiras Orgânicas	Feiras que não utilizam agrotóxicos	FEIRA_ORGÂNICA

Fonte: adaptação de Junior et al. (2018)

Ao todo, 18.271 estabelecimentos foram classificados, incluindo os repetidos devido a apresentarem mais de um tipo de atividade comercial. É importante destacar que não foi possível definir se os estabelecimentos estavam atualmente ativos ou não com base nas planilhas obtidas junto à PROCEMPA. Porém, como de forma intuitiva é possível deduzir que o perfil dos bairros do município de Porto Alegre não se alterou significativamente ao longo dos últimos anos, este fator tende a poder ser desconsiderado. Como todos os estabelecimentos continham apenas endereço em texto livre - não em coordenadas geográficas - foi necessário um processo de tradução. O serviço utilizado foi

[7] https://alvaraweb.procempa.com.br/alvara/

o Geocode.xyz[8] que é gratuito desde que cada endereço entre em uma fila compartilhada com o mundo inteiro para tradução de endereços. Foi criado então um *script* que rodou durante horas para traduzir, um a um, cada endereço. A taxa de sucesso das traduções foi de aproximadamente 90%, ou seja, cerca de 16.400 estabelecimentos estavam prontos para serem utilizados.

[8] http://geocode.xyz

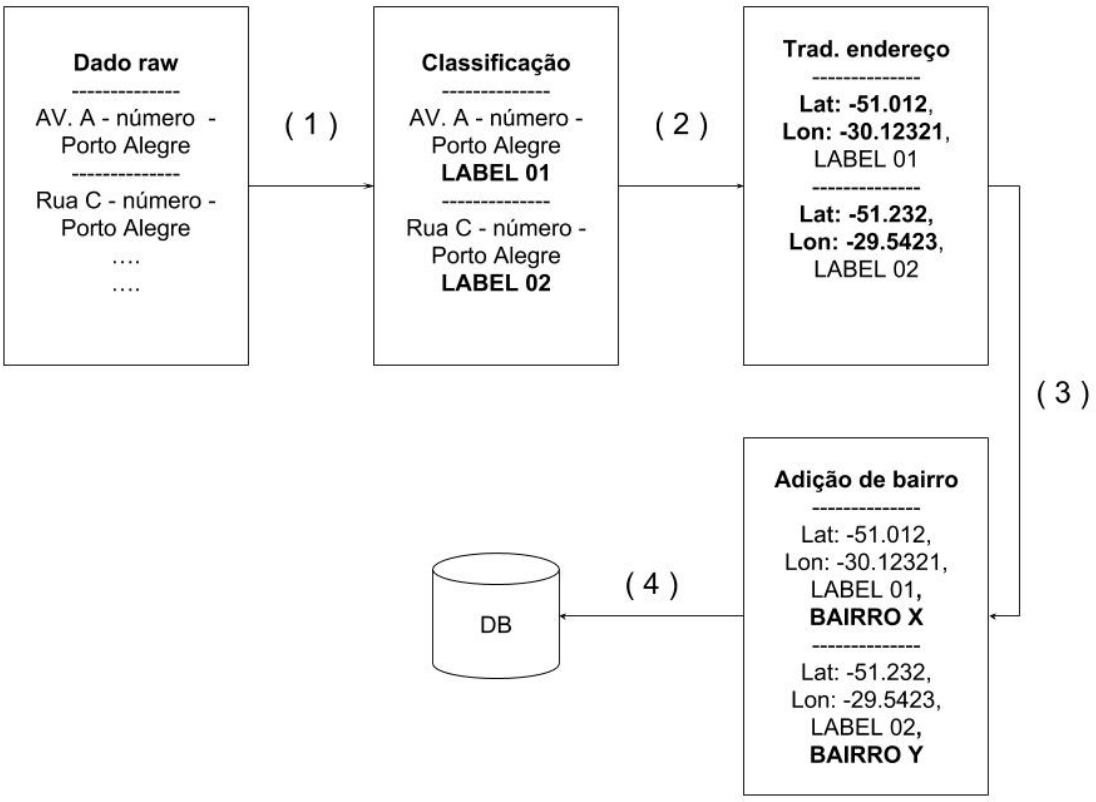
4 RESULTADOS E DISCUSSÃO

Nesta Seção serão descritos os resultados de acordo com as partes estabelecidas na Metodologia e, por fim, é feita uma discussão à respeito de todos os temas.

4.1 Sistema Genérico

O sistema genérico consiste em basicamente dois fluxos. O primeiro, conforme mostra a Figura 4.1, diz respeito a como ocorre o processamento dos dados desde sua obtenção, passando pela classificação (1), tradução de endereço convencional em coordenadas geográficas (2), atrelamento entre ponto e bairro da base PNUD (3) e, por fim, o *upload* no banco de dados MySQL. Já o segundo, mostrado na Figura 4.2, é relacionado à arquitetura, evidenciando a maneira como um usuário acessa o sistema e como este está organizado em nível de aplicação.

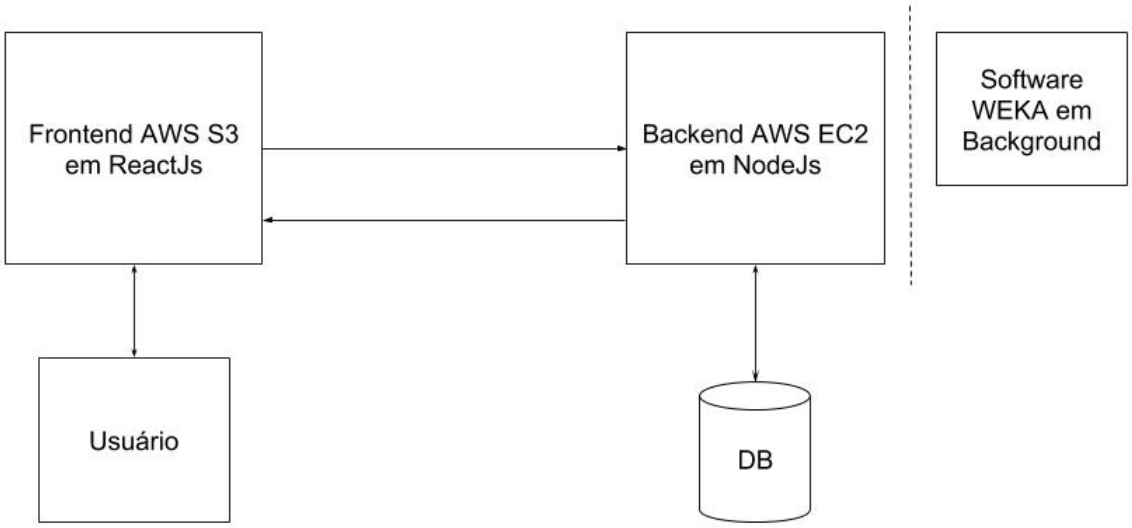
Figura 4.1: Processamento até que se atinja pontos georreferenciados e classificados



Fonte: Autor (2018)

O banco de dados foi organizado em duas *tables* principais. A *table markers*

Figura 4.2: Visão geral da arquitetura do sistema



Fonte: Autor (2018)

agrupa todos os pontos georreferenciados juntamente com suas classificações conforme mostra a Tabela 4.1 de forma representativa. Já a *table pnud* agrupa todos os bairros e seus indicadores conforme mostra a Tabela 4.2 de forma também representativa.

Tabela 4.1: Representação da tabela *markers* tal que *pnudId* é uma *foreign key* para fazer a junção das tabelas

lat	lgn	pnudId	type
-51.0231	-30.43221	8	LABEL 01
-51.1203	-31.3920	7	LABEL 02
-51.4223	-30.4231	6	LABEL 03

Fonte: Autor (2018)

Tabela 4.2: Representação da tabela *pnud*

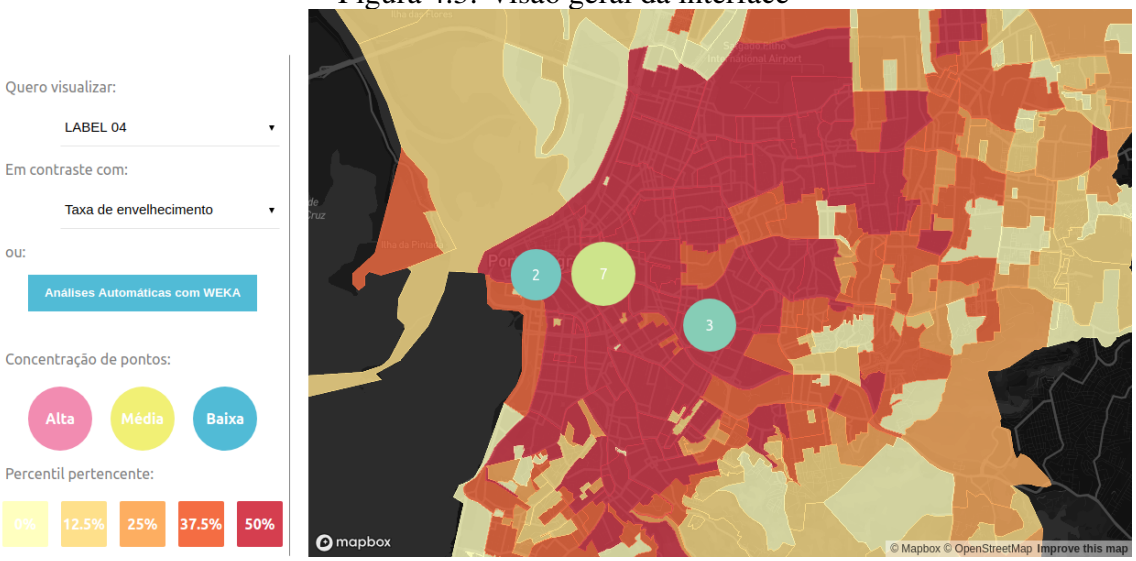
NOME	INDIC_01	INDIC_02	...	INDIC_N
RESTINGA	1.25	0.001	...	12.4
BOM FIM	0.34	0.006	...	4.7
CENTRO HISTÓRICO	0.31	0.012	...	9.1

Fonte: Autor (2018)

Em termos de interface, a Figura 4.3 mostra cinco partes principais: seleção de tipo de ponto georreferenciado (*label* "Quero Visualizar:"), seleção de indicador socioeconômico (*label* "Em contraste com:"), seleção de *insights* proporcionados pelo WEKA (*label* "Análises Automáticas com WEKA"), legendas (*labels* "Concentração de pontos e "Percentis") e, por fim, o mapa em si. Dessa forma, é possível contrastar um das classificações de ponto georreferenciado com um indicador escolhido de maneira a, visualmente,

verificar se há alguma relação entre ambos. Caso se deseje não mais um processo de verificação visual, mas sim através do WEKA utilizando-se um dos algoritmos de *Feature Selection* expostos na Seção 2, pode-se então clicar no botão "*Análises Automáticas com WEKA*". Caso se deseje informações adicionais de uma região, basta que esta seja clicada para que maiores informações, como nome do bairro e valor do indicador selecionado, sejam exibidas na forma de uma *popup*.

Figura 4.3: Visão geral da interface



Fonte: Autor (2018)

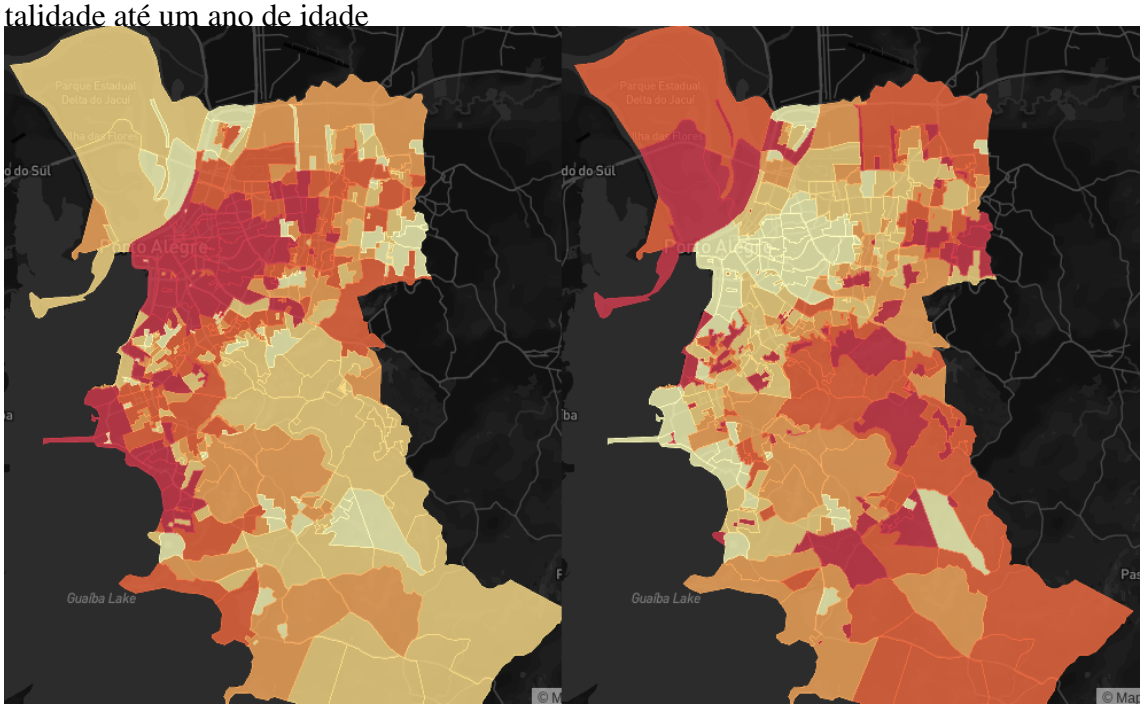
Conforme exposto anteriormente, duas legendas tornaram-se importantes. A primeira, "Concentração de pontos", serve para que o usuário, além da visualização do número de pontos agrupados, possa também observar uma diferenciação na cor, sendo ROSA para alto agrupamento, AMARELO para médio agrupamento e AZUL CLARO para baixo agrupamento. O mesmo vale para a coloração dos bairros, ou seja, se o indicador socioeconômico do bairro pertence ao primeiro percentil, este recebe cor AMARELO CLARO e assim, sucessivamente, até o pertencimento ao quinto percentil recebendo coloração BORDÔ. Isso significa que quanto mais escura a coloração do bairro, maior é o índice naquela região. As escalas foram gerenciadas pela biblioteca *d3-scale*[1].

Uma análise visual dos indicadores socioeconômicos foi realizada com o objetivo de se entender de maneira geral o comportamento dos bairros frente a diferentes indicadores socieconômicos. Um resultado importante foi o de que há, claramente, uma dualidade entre os bairros das regiões centrais e da Zona Sul da cidade (CENTRO HISTÓRICO, BOM FIM, BELA VISTA, MENINO DEUS, CIDADE BAIXA, IPANEMA, TRISTEZA, dentre outros) quando comparados aos demais bairros que não necessaria-

[1]https://github.com/d3/d3-scale

mente estão nestas duas regiões (Figura 4.4 e Figura 4.5). Outro ponto importante, e que será o foco da Seção 4.3, é a aparente forte correlação entre os dados socioeconômicos. De maneira prática, é possível visualmente, verificar-se que o indicador de RENDA PER CAPITA MÉDIA apresenta uma distribuição de cores no mapa muito parecida com diversos outros indicadores positivos como, por exemplo, índice de escolarização, baixa taxa de desemprego, acesso a saneamento, dentre outros.

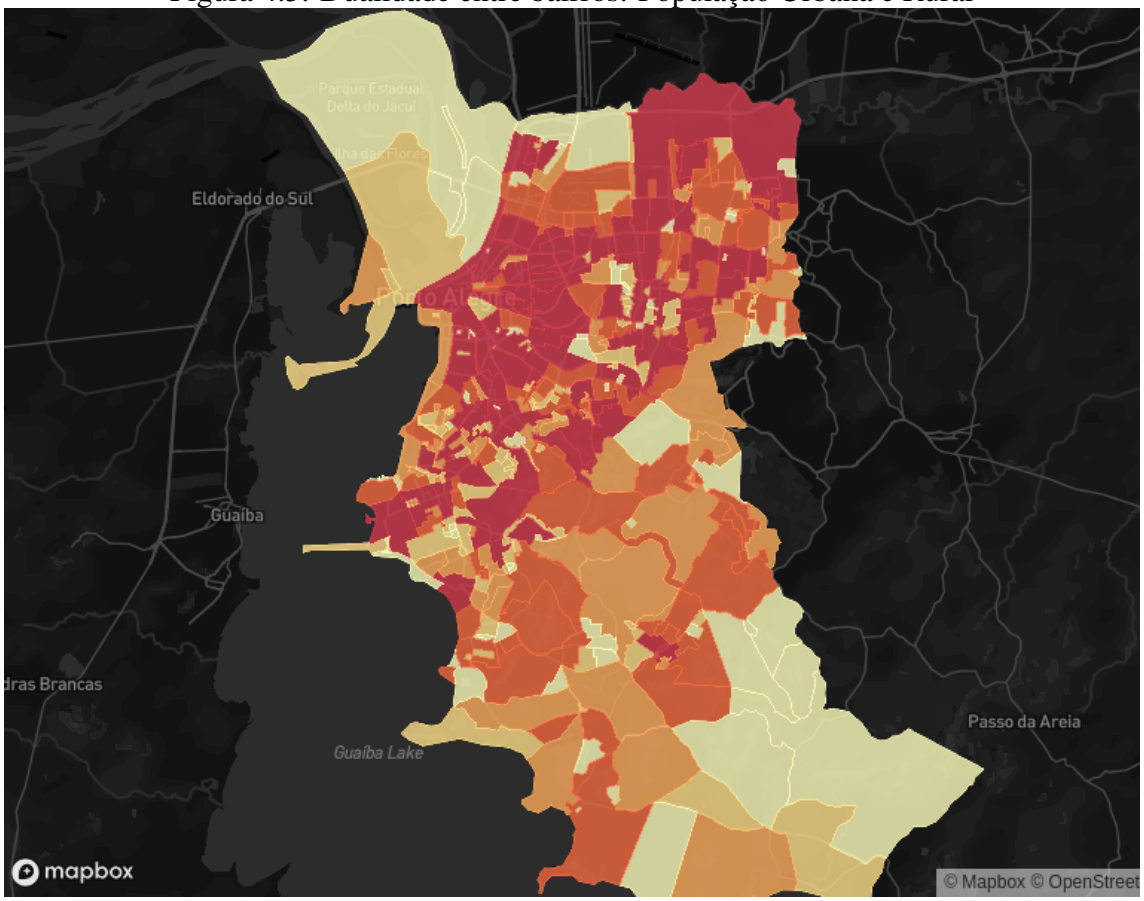
Figura 4.4: Dualidade entre bairros: à esquerda, Renda per Capita Média, à direita, Mortalidade até um ano de idade



Fonte: Autor (2018)

Especificamente para a utilização do WEKA, foram então estabelecidas algumas relações, chamadas de agrupamentos, conforme mostra a Tabela 4.3 visando a geração dos arquivos *.arff*. O objetivo é fomentar a discussão sobre com qual dos dois os algoritmos de Seleção de Atributos funcionariam melhor: de um lado, no AGRUPAMENTO_01, seriam analisadas N linhas tal que N é o número total de pontos georreferenciados da base – o que pode ser muito grande na ordem de milhares. De outro, no AGRUPAMENTO_02, M linhas seriam analisadas tal que M é o número de bairros registrados pelo PNUD - portanto, em torno de 360 linhas. Como o caráter de decisão sobre qual agrupamento utilizar tem uma tendência a ser muito mais empírico, novas discussões são feitas nas próximas Seções.

Figura 4.5: Dualidade entre bairros: População Urbana e Rural



Fonte: Autor (2018)

Tabela 4.3: Conjunto de agrupamentos

RELAÇÃO	DESCRIÇÃO
AGRUPAMENTO_01	TODOS os estabelecimentos com uma coluna adicional TEM ou NAO TEM *LABEL X* com seus bairros de pertencimento
AGRUPAMENTO_02	TODOS os bairros com uma coluna adicional de TEM ou NAO TEM estabelecimentos *LABEL X*

Fonte: Autor (2018)

4.2 Aplicação do sistema a um caso específico

Conforme descrito na Seção 3.2 foi feita a tradução de endereços, classificação e o upload na *table markers* de 18.271 estabelecimentos comerciais classificados em seu tipo de venda: alimentos predominantemente ultraprocessados (ULTRA), predominantemente *in natura* (NATURA), de padrão misto (MIX), feiras (FEIRA) e feiras orgânicas (FEIRA_ORGANICA). Com o objetivo de testar o sistema de forma prática - verificando principalmente a visualização dos dados -, uma série de resultados e análises são apresen-

tados abaixo.

4.2.1 Análise Visual

Inicialmente, uma análise apenas de caráter visual se fez importante para que se pudesse compreender melhor os dados. Do ponto de vista geral, destaca-se, aqui, a alta concentração de todos os tipos de estabelecimentos na região central. Uma segunda região, esta não central, que cobre os perímetros do bairro ANCHIETA, chama bastante atenção pela quantidade de estabelecimentos de tipo MIX e NATURA, o que já se era esperado uma vez que trata-se da Central de Abastecimento do Rio Grande do Sul (CEASA-RS)[2], local bastante conhecido pela comercialização e distribuição de muitos produtos, principalmente hortifrutigranjeiros.

De maneira mais específica, uma vez que é do interesse a obtenção de informações referentes ao acesso das populações a alimentos ULTRA, NATURA, FEIRA e FEIRAS-ORGÂNICAS, chega-se, como resultado a partir da análise visual, que, além das regiões centrais concentrarem o maior número de estabelecimentos de todos estes tipos, as categoria de ULTRA e NATURA estão presentes em todas as regiões com exceção do extremo da Zona Leste da cidade que, aparentemente, concentra pouquíssimos alvarás cadastrados (Figura 4.6). Quando colocada à vista a categoria FEIRAS, o padrão muda bastante: a Zona Leste passa a não ter nenhum ponto em sua totalidade de território e o número de pontos diminui drasticamente em todos os outros territórios ficando a região Centro e Norte mais bem abastecidas. Por fim, pontos classificados como FEIRAS_ORGÂNICAS mostram-se ainda mais discrepantes entre as regiões com uma total concentração na Zona Central (14 pontos georreferenciados) com uma pequena exceção na Zona Sul (dois pontos).

4.2.2 Análise com WEKA

A fim de contemplar o objetivo principal do trabalho (*insights* para o usuário), as análises utilizando-se o WEKA são descritas abaixo em torno dos seus resultados. Uma série de nomes de variáveis da base do PNUD serão apresentadas em forma de códigos das quais algumas - as mais relevantes - serão explicadas, já outras serão abordadas na

[2]http://www.ceasa.rs.gov.br/

discussão.

4.2.2.1 Visão Geral

Para as análises de Seleção de Atributos, os dados foram agrupados conforme previsto na Tabela 4.3. Primeiramente, levantou-se a hipótese de que o AGRUPAMENTO_01 faz mais sentido quando os pontos georreferenciados estão mais bem distribuídos (caso que ocorre nos classificados como ULTRA e NATURA) pois caso contrário, basicamente todos os bairros teriam seus campos TEM - indicativo de que há um estabelecimento deste tipo - preenchidos causando ruído na seleção. De maneira contrária, as relações do tipo AGRUPAMENTO_02 hipoteticamente fazem mais sentido no caso dos pontos classificados como FEIRA e FEIRA_ORGANICA, dada a baixa distribuição dos pontos. As duas hipóteses foram confirmadas. Como caso de teste da hipótese, escolheu-se estabelecimentos do tipo FEIRA_ORGÂNICA para que se pudesse comparar seu desempenho utilizando-se o algoritmo *CorrelationAttributeEval* com ambos os AGRUPAMENTOS_01 e 02 conforme mostra a Tabela 4.4. Desta forma, ratifica-se o que foi exposto por Sheena, Kumar and Kumar (2016), que menos atributos de entrada podem colaborar para melhores resultados uma vez que a ordem de grandeza das correlações do AGRUPAMENTO_01 da tabela são extremamente inferiores quando comparadas às da segunda coluna sabendo-se que a correlação pertence ao intervalo de -1 a 1.

Tabela 4.4: Verificação do ruído causado devido ao tipo de agrupamento

Grupo	Agrupamento 01	Agrupamento 02
n_instâncias TEM = 0	18.004	322
n_instâncias TEM = 1	18	13
1º Correlação mais forte - Atributo	0.236004-T_SUPER25M	0.50264-MULHER80
2º Correlação mais forte - Atributo	0.0234156-T_FBSUPER	0.47237-MUL75A79
3º Correlação mais forte - Atributo	0.233565-T_VULNERA..	0.47230-PESO65

Fonte: Autor

Tornou-se importante também, para fins de simplificação da própria interface, que se escolha um ou mais algoritmos de *Feature Selection* como padrão. Como o objetivo final é apresentar *insights* para o usuário relacionando-se os índices socioeconômicos com os pontos georreferenciados e não a criação de um modelo de classificação automática, entendeu-se que dois algoritmos fornecidos pelo WEKA fossem os mais interessantes e mais facilmente compreensíveis: *CfsSubsetEval* e *CorrelationAttributeEval* uma vez que estes não envolvem classificadores nem definições estatísticas mais complexas. Conforme levantado na Seção 4.1 de que há uma correlação interna muito forte entre os indicado-

res socioeconômicos da base do PNUD, o algoritmo *CfsSubsetEval* acabou se mostrando inapropriado, pois, ao fim de sua análise, quando elimina os atributos que estão correlacionados entre si, faz com que o conjunto de atributos selecionados fique muito parecido em todas as análises (Tabela 4.5) mesmo quando aplicado um limiar para eliminar-se correlações negativas. Taxas de analfabetismo (T_ANALF) e razão de pessoas que dependem da população ativa (RAZDEP) - a exemplo de idosos e crianças - foram correlacionadas com NATURA, FEIRA e FEIRA_ORGANICA, fato este que não é plausível quando se acompanha o comparativo diretamente no mapa, por exemplo, contrastando-se as cores com os indicadores.

Tabela 4.5: Algoritmo *CfsSubsetEval*. Em negrito índices que não condizem com a realidade

	NATURA	FEIRA	FEIRA_ORGANICA
Atributo 01	**RAZDEP**	**RAZDEP**	**RAZDEP**
Atributo 02	**T_ANALF25A29**	T_ENV	T_MED18M
Atributo 03	T_FBBAS	**T_ANALF15A17**	HOMEM75A79
Atributo 04	T_FBFUND	**T_ANALF18A24**	HOMENS80

Fonte: Autor (2018)

4.2.2.2 Resultados específicos

Conforme apresentado na Tabela 4.6, alguns resultados - especialmente o conjunto de indicadores socioeconômicos - são esperados com base no conjunto de entrada exposto (Tabela 4.7). Nitidamente, devido ao menor número de pontos georreferenciados para os conjuntos FEIRA e FEIRA_ORGANICA, há a obtenção de valores maiores, ou seja, maior correlação. Para os conjuntos ULTRA e NATURA o valor final é bastante baixo, indicando uma provável dificuldade em se atrelar o indicador socioeconômico com os pontos.

Considerando que o objetivo da análise possa ser verificar o acesso a alimentos frescos (NATURA, FEIRA E FEIRA_ORGÂNICA), é interessante verificar que os atributos referentes à População Feminina estão bastante presentes (MULH40A44, MULH35A39, e, assim, sucessivamente). Porém, é importante aqui que se possa compreender realmente o que estes indicadores significam cabendo a Seção de Discussão esse aprofundamento. A Figura 4.7 traz, em cada quadrante, a associação entre o primeiro indicador de cada um dos conjuntos de entrada juntamente com os seus respectivos tipos de ponto.

Dado que FEIRA e FEIRA_ORGANICA são os tipos de pontos georreferenciados que aprensentam resultados significativos, ou seja, com correlações mais próximas de 1, e

Tabela 4.6: Conjuntos de entrada para utilização no WEKA

TIPO	RELAÇÃO	ALGORITMO	N_ATRIBUTOS_SAIDA
ULTRA	AGRUPAMENTO_01	CorrelationAttributeEval	5
NATURA	AGRUPAMENTO_01	CorrelationAttributeEval	5
FEIRA	AGRUPAMENTO_02	CorrelationAttributeEval	5
FEIRA_ORGANICA	AGRUPAMENTO_02	CorrelationAttributeEval	5

Fonte: Autor (2018)

Tabela 4.7: Resultados obtidos dado o conjunto de entrada da Tabela 4.6

TIPO	ATRIBUTO	VALOR
ULTRA	T_RMAXIDOSO	0.110122
	T_SUPER25M	0.107483
	T_FBSUPER	0.106412
	T_FORA0A5	0.105758
	P_SUPER	0.105642
NATURA	T_RMAXIDOSO	0.051554
	T_FREQFUND1824	0.043612
	PREN40	0.42541
	REN3	0.042424
	PREN20	0.041399
FEIRA	MULH40A44	0.44848
	MULH35A39	0.44578
	MULH60A64	0.44565
	PESOM25M	0,44163
	PESO25	0.43916
FEIRA_ORGANICA	MULHER80	0.50264
	MULH75A79	0.48237
	PESO65	0.4723
	HOMEM75A79	0.47187
	MULH70A74	0.48986

Fonte: Autor (2018)

com o objetivo de se aprofundar mais sobre estes dois tipos específicos na Seção Discussão uma vez que consistem no acesso a alimentos frescos, buscou-se, na Tabela 4.8 trazer não apenas os cinco primeiros atributos selecionados pelo WEKA, mas os demais 10 atributos não relacionados à demografia populacional - que aparece entre os 46 primeiros atributos no tipo FEIRA_ORGANICA, por exemplo.

4.3 Discussão

Primeiramente, torna-se importante pontuar, do ponto de vista do Sistema Genérico proposto, a visualização dos dados. O principal desafio foi como representar duas dimensões de dados sendo estas de indicadores socioeconômicos e pontos georreferenci-

Tabela 4.8: Ranking específico dos tipos FEIRA E FEIRA_ORGANICA eliminando atributos de demografia populacional

TIPO	ATRIBUTO	VALOR
	T_FUND25M	0.29171
	T_MED18A24	0.29152
	T_FUND18M	0.29091
	P_FUND	0.2907
FEIRA	I_ESCOLARIDADE	0.28978
	T_FBMED	0.28918
	T_MED18M	0.28792
	T_FBMED_tudo	0.28758
	T_MED25M	0.28701
	T_FUND18A24	0.28632
	CORTE4	0.32268
	RDPC4	0.32177
	RDPC5	0.32089
	RDPC10	0.32032
FEIRA_ORGANICA	CORTE9	0.32019
	RDPCT	0.31893
	RDPC	0.31892
	CORTE3	0.31807
	RDPC3	0.31282
	T_FBSUPER	0.30854

Fonte: Autor (2018)

ados do cotidiano das populações. Como mostra Plewe (2007), o desenvolvimento dos mapas foi sempre guiado por grandes empresas na tentativa - muitas vezes bem sucedida - de explorar mercados até então inexplorados e lucrativos deixando-se, assim, a visualização mais "social"de lado. Isso somado à dificuldade de acesso aos dados públicos torna a tarefa mais complicada. Porém, de um modo geral, a aplicação parece ter conseguido representar de forma facilmente compreensível as relações as quais se propõe alinhando-se com representações atuais de dados em cima de mapas tais quais o conceito de mapas jornalísticos que trazem uma abordagem bastante estética em suas visualizações (WEBER; RALL, 2012).

É recorrente durante o documento e também na literatura, expressões frente à dificuldade no acesso às informações públicas como as restrições de acesso e restrições de compartilhamento *online* (KITCHIN, 2014). No exemplo da obtenção dos dados deste trabalho foram necessárias inúmeras horas de sanitarização em planilhas e troca de e-mails. Esforço este que poderia ter sido evitado caso fossem fornecidas APIs, por exemplo. Porém, salienta-se, aqui, também, que as entidades governamentais parecem estar caminhando para uma maneira mais transparente e acessível de apresentar seus dados

compreendendo que, além de fins de fiscalização, pode-se utilizar da tecnologia para o desenvolvimento e descoberta de novos conhecimentos. Dois exemplos bastante significativos são o Portal Brasileiro de Dados Abertos[3] que fornecesse seus dados através de download sobre inúmeros temas do interesse social e o Portal de Sistemas e Consultas do Tribunal de Contas do Estado do Ceará (TCM-CE)[4] que fornece uma API para consultas retornando formatos bastante atualizados e rotineiros para programadores como JSON e XML, o que é realmente um avanço uma vez que permite a automatização da leitura das informações.

Ainda relativo aos dados socieconômicos e fazendo-se uma relação com algoritmos de mineração, a Seção Resultados já indicava uma possível forte correlação entre os atributos de bairro que poderia vir a causar ruídos tanto para um caso de métodos de classificação quanto para seleção de atributos. De certa forma, isso é explicável e esperado, uma vez que os indicadores - se não estão ligados diretamente uns aos outros devido às suas próprias fórmulas - estão socialmente conectados. Em linhas gerais, é extremamente natural e bastante difundido que bairros com, por exemplo, RENDA PER CAPITA MÉDIA alta, obterão altíssimos escores em outros indicadores tais como educação e saúde. A Tabela 4.5 deixou claro este fato ao eliminar de forma brusca muitos atributos correlacionados sobrando outros tantos que não faziam tanto sentido.

Apesar do presente trabalho tentar aproximar um usuário comum a *insights* automatizados através do software WEKA, uma série de simplificações foram feitas e, portanto, não exploradas. Por exemplo, ao deter-se a apenas um tipo de algoritmo dentre três - Classificação, Regressão e Seleção de Atributos - deixa-se de ser explorada uma série de possibilidades como, por exemplo, a predição de pontos georreferenciados futuros. Outra simplificação que foi feita de maneira a facilitar a compreensão do usuário final - considerando-o leigo na área - foi a seleção de apenas um método de Seleção de Atributos dentre muitos outros que o WEKA proporciona, indicando-se aqui, uma fraqueza do presente trabalho e, naturalmente, uma porta aberta a se explorar maneiras mais didáticas de apresentar outros métodos ao usuário. Outro ponto importante é de que, apesar da atrativa promessa de seleção automática de atributos, ainda assim foram necessárias análises não automatizadas criando-se uma forte necessidade de um mínimo conhecimento prévio na área a qual os pontos georreferenciados estão inseridos. A Tabela 4.8, por exemplo, necessitou ser criada uma vez que o WEKA sugeria, dentre as primeiras 40 posições, apenas atributos relacionados a população absoluta, o que, apesar de fazer sentido lógico, uma

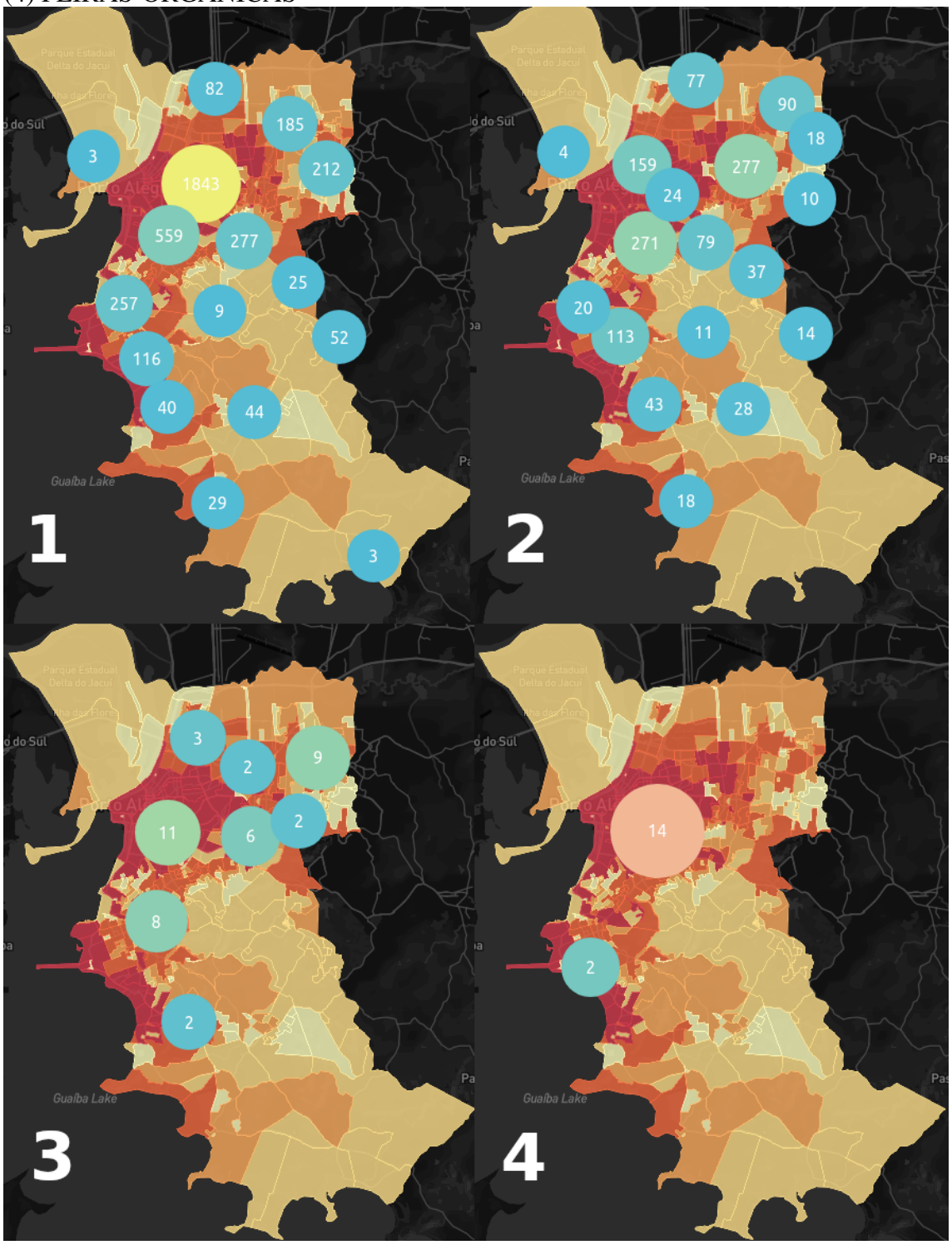
[3] http://dados.gov.br/
[4] http://api.tce.ce.gov.br/

vez que pontos do tipo FEIRA_ORGANICA estão concentrados em regiões mais densamente populosas, não faz sentido analítico, pois é trivial que estas regiões destacam-se por este tipo de indicador.

Uma vez que o caso específico ao qual foi aplicado o Sistema baseia-se amplamente nos conhecimentos desenvolvidos por Junior et al. (2018) e utiliza sua classificação para os estabelecimentos, fomentam-se alguns questionamentos frente a qualidade da classificação. Por exemplo, será que todas as lancherias são necessariamente do tipo ULTRA? E quanto às cafeterias, não poderiam haver algumas especializas em produtos minimamente processados? Um primeiro ponto de partida para corrigir possíveis mal interpretações poderia ser o próprio sistema aqui proposto permitir a mudança de forma dinâmica das classificações. Ainda neste contexto, e especificamente sobre alvarás, outra discussão importante é frente aos bairros socialmente menos favorecidos dado o conhecimento intuitivo de que estes apresentam estabelecimentos que não necessariamente foram formalmente cadastrados junto às prefeituras de suas respectivas cidades. Fica claro, assim, que apenas dados secundários - aqueles que já existem nas bases de dados governamentais - podem não refletir de maneira ampla a realidade fazendo-se necessários, quando possível, dados primários - coletados individualmente em regiões ultra específicas.

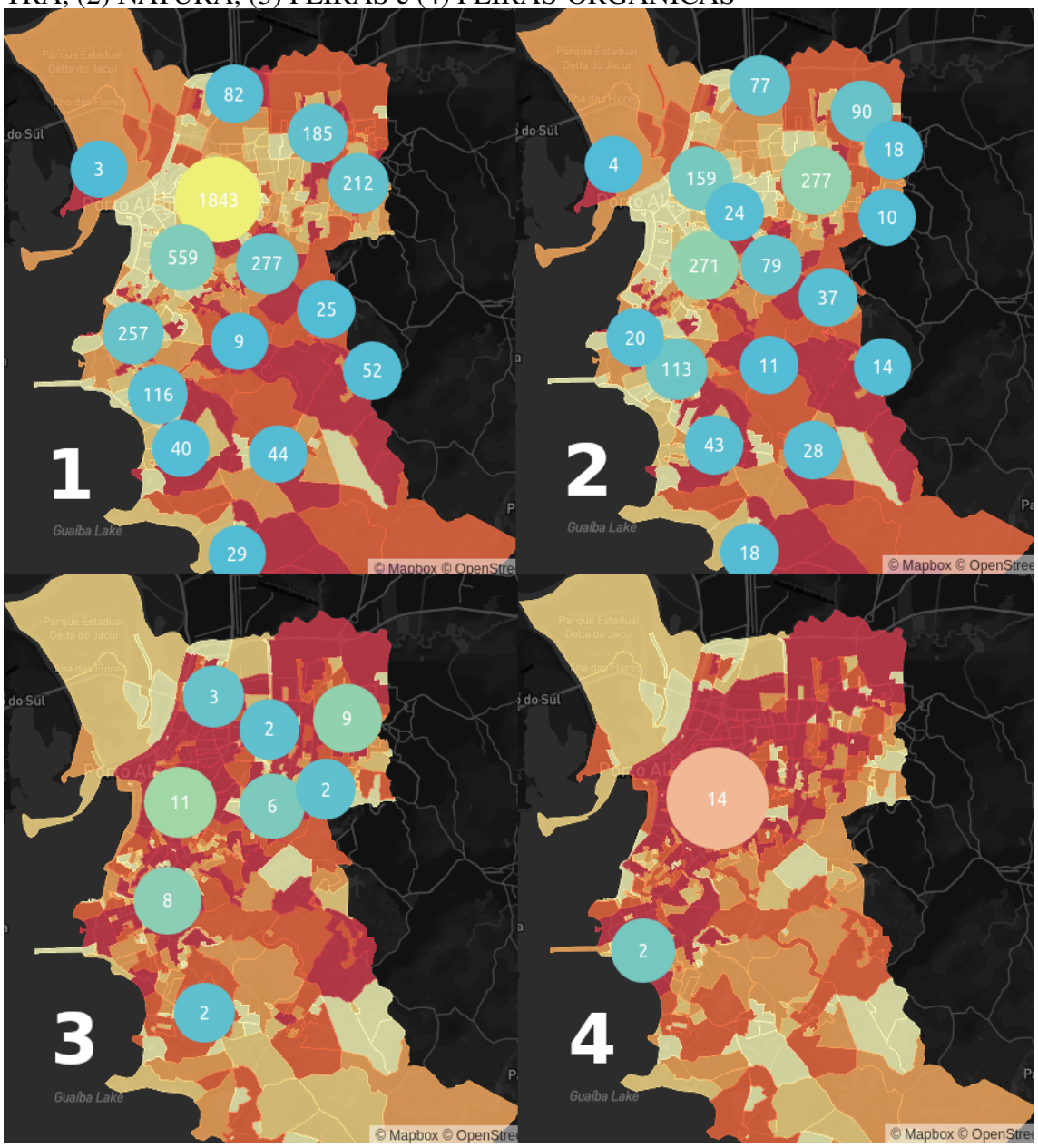
Por fim, conforme exposto na Tabela 4.8 foi possível, com um grau discutível, porém aceitável, que o acesso a alimentos frescos no Município de Porto Alegre está fortemente relacionado com os indicadores como renda, saúde e educação. Discutível pois apresentam como média de correlação dado o algoritmo *CorrelationAttributeEval* o valor de 0.2893 para o tipo FEIRA e 0.3183 para o tipo FEIRA_ORGANICA, o que está relativamente longe de 1.0. Aceitável pois refletem bem as características dos bairros aos quais pertencem.

Figura 4.6: Distribuição dos estabelecimentos: (1) ULTRA, (2) NATURA, (3) FEIRAS e (4) FEIRAS-ORGANICAS



Fonte: Autor (2018)

Figura 4.7: Primeiro atributo selecionado pelo WEKA para cada tipo de ponto: (1) ULTRA, (2) NATURA, (3) FEIRAS e (4) FEIRAS-ORGANICAS



Fonte: Autor (2018)

5 CONCLUSÃO

Abordou-se três áreas distintas neste trabalho: acesso a dados públicos, algoritmos de *data mining* presentes no software WEKA e visualização de informações através de mapas. A primeira apresentou-se como um fator de dificuldade dado o enorme tempo gasto para pesquisa, aquisição e tratamento dos dados, fatores bastante explorados na literatura e que, na prática, mostram-se verdadeiros. A segunda área foi explorada de maneira não tão profunda aproveitando-se dela apenas os algoritmos que convinham para o objetivo final da aplicação. Já a terceira área foi explorada em termos de tecnologias mais recentes. Uma integração genérica entre elas foi sugerida e narrou-se implementações e arquitetura para ao fim, aplicar-se a um caso prático como objeto de estudo para fomentar possíveis melhorias e compreensão de como o sistema deve funcionar.

Escolheu-se como ferramentas e métodos para a criação de um sistema genérico, a partir de uma análise comercial e de literatura: MapBox API para mapas e WEKA para mineração dos dados com algoritmos de *Feature Selection* uma vez o objetivo final era identificar os melhores atributos dado um conjunto de entrada de pontos e bairros de pertencimento.

Um caso prático de utilização foi pautado na classificação de estabelecimentos e estudo do ambiente alimentar da cidade de Porto Alegre. Diferentes tipos de agrupamentos dos dados foram explorados e uma classificação adaptada da literatura foi utilizada. Para os tipos ULTRA e NATURA, os indicadores socioeconômicos não foram tão facilmente compreensíveis apresentando baixa correlação com os pontos. Já para os tipos FEIRA e FEIRA_ORGANICA obtiveram-se resultados mais factíveis e aplicáveis à realidade. Entendeu-se que depreendendo da maneira como os dados são agrupados e comparados pode-se obter diferentes resultados. Assim, foram testadas diferentes maneiras e chegou-se a conclusão de que: agrupar as classificações dos pontos georreferenciados com os índices socieconômicos dos bairros depende do número de pontos georreferenciados que a classificação possui e que granularizar as classificações, como foi o caso de FEIRA e FEIRA_ORGANICA, aumenta as chances de se compreender de forma mais acertiva o problema estudado. Conseguiu-se, ao fim, se chegar a uma conclusão sobre o ambiente alimentar de Porto Alegre.

Dentre as dificuldades encontradas, além do acesso a dados públicos citado anteriormente, pode-se ressaltar a elaboração de uma arquitetura genérica dado que são possíveis inúmeras classificações de pontos geográficos e, naturalmente, decorrem muitas in-

terpretações que são facilmente visualizáveis no software WEKA, porém não facilmente reproduzíveis em uma sistema web cujo objetivo é ser mais atrativo para um usuário leigo. Na mesma linha, compreender todos os algoritmos de *data mining* não é uma tarefa trivial e traduzir isso para o usuário final em uma interface web também é um ponto crítico e difícil.

Como melhorias e perspectivas para trabalhos futuros, compreender como traduzir em interface todas as funcionalidades do WEKA - ou qualquer outra ferramenta escolhida - para que um usuário leigo possa usufrui-las e aplicá-las de forma a embasar matematicamente suas observações através de modelos é um fator a ser explorado. Do ponto de vista de funcionalidades, fornecer uma forma de *upload* de arquivo *.csv* ou *.xls* com os pontos georreferenciados e suas classificações e permitir a classificação dinâmica dos pontos também teriam impacto na utilização. Do ponto de vista de arquitetura, uma possível migração da API MapBox - paga - para a Leaflet - código aberto - poderia ser interessante. Por fim, a validação em forma de questionários qualitativos e quantitativos tanto para especialistas em determinadas áreas às quais se estariam ligados novos conjuntos de pontos georreferenciados quanto para usuários leigos na função de apenas visualizadores, poderia elucidar pontos chaves na compreensão de dados públicos e indicadores socioeconômicos quando reproduzidos na forma de mapas.

REFERÊNCIAS

BARATA, R. **Como e por que as desigualdades sociais fazem mal saúde**. [S.l.]: Editora FIOCR, 2009.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, p. 37–54, 03 1996.

FJP; IPEA; PNUD. **Atlas Brasil IDH NO BRASIL 2010**. 2013. Available from Internet: <http://www.novomilenio.inf.br/baixada/bsfotos/IDHM-PNUD-2010-Brasil.pdf>.

FRANK, E.; HALL, M. A.; WITTEN, I. H. **The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Fourth Edition**. [S.l.]: Morgan Kaufmann, 2016.

GEBRU, T. et al. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 114, n. 50, p. 13108–13113, 2017. ISSN 0027-8424. Available from Internet: <http://www.pnas.org/content/114/50/13108>.

GEORGE, F. Sobre determinantes da saúde. September 2011. Available from Internet: <http://bit.ly/2vZqVke>.

HLPE. Nutrition and food systems. a report by the high level panel of experts on food security and nutrition of the committee on world food security. 2017.

JUNIOR, P. C. P. de C. et al. Desigualdades territoriais na disponibilidade de alimentos no municÍpio do rio de janeiro. 2018. Available from Internet: <https://www.arca.fiocruz.br/bitstream/icict/27009/2/paulo_cesar_pereira.pdf>.

KAVAKIOTIS, I. et al. Machine learning and data mining methods in diabetes research. **Computational and Structural Biotechnology Journal**, v. 15, p. 104 – 116, 2017. ISSN 2001-0370. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S2001037016300733>.

KITCHIN, R. The data revolution: Big data, open data, data infrastructures and their consequences. 2014.

LUXBURG, U. von; SCHOELKOPF, B. **Statistical Learning Theory: Models, Concepts, and Results**. 2008. Available from Internet: <https://arxiv.org/pdf/0810.4752.pdf>.

MONTEIRO, C. A. et al. Nova. the star shines bright. food classification. **World Nutrition**, v.7, n. 1-3, p. 28–38, July 2016. Available from Internet: <https://worldnutritionjournal.org/index.php/wn/article/view/5/4>.

OLIVEIRA, E. F. de; SILVEIRA, M. S. Open government data in brazil a systematic review of its uses and issues. In: **Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age**. New York, NY, USA: ACM, 2018. (dg.o '18), p. 60:1–60:9. ISBN 978-1-4503-6526-0. Available from Internet: <http://doi.acm.org/10.1145/3209281.3209339>.

ONU-WINDER. The world distribution of household wealth. **Helsinki,FI. United Nations University**, 2006.

PAULA, M. M. V. de et al. A visualização de informação e a transparência de dados públicos. **Portal VisPública - Modelo de Visualização de Dados Públicos**, p. 1–12, 2011. Available from Internet: <http://vispublica.gov.br/vispublica/resources/pdf/MMVPAULA-INFOVIS.pdf>.

PETERSON, M. Evaluating mapping apis. In: ____. [S.l.: s.n.], 2015. p. 183–197. ISBN 978-3-319-07925-7.

PLEWE, B. Web cartography in the united states. cartography and geographic information science. v. 34, n. 2, p. 133–136, 2007.

PORTRAIT, F.; LINDEBOOM, M.; DEEG, D. Life expectancies in specic health states: results from a joint model of health status and mortality of older persons. v. 38, n. 4, p. 525–536, 2001.

RUDIN, C. et al. Machine learning for the new york city power grid. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 34, n. 2, p. 328–345, Feb 2012. ISSN 0162-8828.

SAÚDE, M. da. **Guia alimentar para a população brasileira Promovendo a Alimentação Saudável**. 2nd. ed. [S.l.]: Ministério da Saúde, 2014.

SHEENA; KUMAR, K.; KUMAR, G. Article: Analysis of feature selection techniques: A data mining approach. **IJCA Proceedings on International Conference on Advances in Emerging Technology**, ICAET 2016, n. 1, p. 17–21, September 2016. Full text available.

SINGHAL, S.; JENA, M. A study on weka tool for data preprocessing , classification and clustering. In: . [s.n.], 2013. Available from Internet: <https://pdfs.semanticscholar.org/095c/fd6f1a9dc6eaac7cc3100a16cca9750ff9d8.pdf>.

STRINGHINI, S. et al. Socioeconomic status and the 25 x 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. **The Lancet**, v. 389, n. 10075, p. 1229–1237, March 2017. Available from Internet: <https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)32380-7/fulltext>.

WEBER, W.; RALL, H. Data visualization in online journalism and its implications for the production process. **International Conference on Information Visualisation**, IV, n. 349-356, 2012.

WHITEHEAD, M. The concepts and principles of equity and health. Geneva: WHO, 2000.

YUNES, R. C. **Mudanças no cenário econômico: Velhos e Novos Males da Saúde**. 2nd. ed. [S.l.]: Editora Hucite, 2000. 33-60 p.