Trabalho de Aprendizado Descritivo
Pipeline Descritivo

Guilherme Namen Pimenta[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

Resumo. *Este trabalho apresenta um sistema de descoberta de conhecimento para descrever o espaço urbano de Belo Horizonte usando uma arquitetura de dutos e filtros com algoritmos de mineração de dados. A partir de um banco de dados de imóveis, foram aplicados algoritmos como FPClose, OPUS Miner, FHM Freq, FHN e o software Cortana para identificar padrões no uso e ocupação do solo. Os resultados mostram uma distinção entre as regiões centrais e periféricas: as áreas centrais têm maior adensamento e presença de comércio, enquanto as periferias predominam em casas e menor verticalização. Também foi observada a importância de vagas de garagem para imóveis comerciais de alto padrão e a influência da área construída. Além disso, foi analisado o potencial da mineração de dados para identificar áreas vulneráveis à gentrificação na Região Centro-Sul. A abordagem demonstra a capacidade da mineração de dados em apoiar políticas públicas de planejamento urbano.*

1. Introdução

Os impressionantes avanços em tecnologia de aprendizado de máquina estão cada vez mais presentes no cotidiano das pessoas [Doshi-Velez and Kim 2017], frequentemente superando o desempenho humano em algumas tarefas [Silver et al. 2016]. No entanto, muitas ferramentas de aprendizado de máquina não permitem a auditoria de seus métodos, pois não produzem regras plenamente interpretáveis, dificultando o processo de verificação pelos administradores. Neste trabalho, o termo interpretação é definido como a capacidade de gerar termos plenamente compreensíveis por humanos. Para uma definição mais completa e discussão mais profunda, recomenda-se o trabalho de Doshi-Velez et al [Doshi-Velez and Kim 2017].

Uma forma de gerar regras compreensíveis é utilizar técnicas de mineração de dados [Zaki and Meira 2014], que possuem a capacidade de processar grandes quantidades de dados e gerar modelos descritivos de interesse, facilmente interpretáveis. Atualmente, há uma vasta gama de algoritmos e ferramentas [Mikut and Reischl 2011] de mineração de dados, desenvolvidos ao longo dos últimos 25 anos [Luna et al. 2019].

A análise do espaço urbano exige a compreensão de padrões complexos e multifacetados, frequentemente ocultos em grandes volumes de dados. A interoperabilidade dos métodos utilizados nesse processo é crucial, pois permite aos planejadores urbanos e gestores públicos compreenderem as forças que moldam a cidade, embasando a formulação de políticas públicas mais eficazes e socialmente justas. Nesse contexto, a mineração de dados, com sua capacidade de gerar regras compreensíveis por humanos, surge como uma ferramenta poderosa para desvendar tais padrões e relações. A identificação de áreas vulneráveis à gentrificação [Andrade and Mendonça 2020], por exemplo, demanda a análise

conjunta de variáveis socioeconômicas e características dos imóveis, e a capacidade de interpretar os resultados obtidos pode ser decisiva para a criação de políticas públicas que promovam o desenvolvimento urbano sustentável e inclusivo.

Dada a quantidade de informações geradas e a diversidade de algoritmos propostos, este trabalho tem como objetivo adaptar a arquitetura de dutos e filtros de descoberta de conhecimento em banco de dados [Nwagu et al. 2017], utilizando um conjunto mínimo de algoritmos de mineração de dados.

2. Desenvolvimento

Para guiar o desenvolvimento do pipeline de algoritmos de mineração de dados, foi escolhido o Princípio da Descrição de Comprimento Mínimo (DCM) da mesma forma que foi utilizado na tese de Proença [Proença 2021]. O princípio é definido como: ”DCM é baseado no seguinte pensamento: qualquer regularidade em um certo conjunto de dados pode ser utilizada para compressão de dados, para descrevê-lo usando menos símbolos que o necessário para descrever os dados literalmente”[Grünwald 2007]. Pode-se assemelhar este conceito com o processo de compressão de dados [Proença 2021].

Para o desenvolvimento do trabalho foi utilizada a metodologia Goal Question Metric (GQM) [Caldiera and Rombach 1994], metodologia muito utilizada no campo da Engenharia de Software [Sommerville 2011].

Sendo assim a meta é: desenvolver um sistema em arquitetura de dutos e filtros (*pipeline*) para melhor descrever os dados de uma base de dados imobiliário urbanos, utilizando o princípio DCM.

Os questionamentos são:

P1. Quais os algoritmos de mineração de dados são úteis para descrever uma base de dados urbanos?

P2. Quais informações eles podem gerar?

P3. Qual é a melhor sequência de algoritmos para descrever os dados e formar a arquitetura de dutos?

As métricas são:

M1. Número de resultados produzidos pelo algoritmo.

M2. Hiper parâmetros dos algoritmos.

M3. Qualidade do resultado.

3. Descrição dos dados

Para este trabalho foi utilizado um base de dados dos imóveis da Capital Mineira Belo Horizonte, os dados são providos pelo Portal de Dados Abertos da Prefeitura de Belo Horizonte. A escolha desta base deve-se ao seu tamanho, importância para as Políticas de Uso e Ocupação do Solo do município e a facilidade para comprovar os resultados, uma vez que os autores residem nesta localidade.

O Portal de Dados Abertos possui a série histórica dos dados imobiliários de Belo Horizonte, divididos pelas regionais, que se constitui de uma divisão geográfica baseada em critérios únicos de cada regional. Ao todo a capital está dividida nas seguintes

nove regionais: A divisão de Belo Horizonte em nove regionais administrativas (Bar-

Regional	População[a]	Área (km²)	Bairros
Barreiro	282.156	53,6	73
Centro-Sul	282.286	31,85	49
Leste	228.986	27,98	47
Nordeste	281.507	39,46	69
Noroeste	271.143	30,17	52
Norte	214.967	32,67	48
Oeste	316.908	36,06	67
Pampulha	266.859	51,21	63
Venda Nova	230.339	29,27	44
TOTAL	2.375.151	332,27	487[b]

Tabela 1. Regionais

[a]IBGE Censo 2010
[b]Alguns bairros estão em mais de uma regional

reiro, Centro-Sul, Leste, Nordeste, Noroeste, Norte, Oeste, Pampulha e Venda Nova), como apresentado na tabela 1, será utilizada como base para a segmentação dos dados e análise do espaço urbano. O trabalho investiga se as particularidades de cada região, como população, área e características socioeconômicas, refletem em padrões distintos de uso e ocupação do solo detectáveis via a arquitetura proposta. A segmentação permitirá uma análise mais granular e contextualizada, revelando nuances que poderiam passar despercebidas em uma abordagem agregada para toda a cidade. Adicionalmente, a população e a densidade demográfica de cada região servirão como importantes fatores para a interpretação dos resultados da mineração de dados, contextualizando a dinâmica urbana.

Ao todo foram coletados 876.817 registros imobiliários publicados na data de 03/06/2024 com as seguintes características: Nome da regional, índice cadastral, frequência da coleta de lixo, existência de meio fio, existência de via pavimentada, existência de arborização, existência de galeria pluvial, existência de iluminação pública, existência de rede de esgoto, existência de rede de água, existência de rede telefônica, área do terreno, área construtiva, tipo de construção, tipo de ocupação, padrão de acabamento do imóvel, quantidade de economias, fração ideal, tipo do logradouro, nome do logradouro, número do imóvel, CEP, zona homogenia (classificação do imóvel para critérios de cobrança de imposto), tipologia do imóvel, geometria do terreno, coordenadas de latitude e longitude do centroide da geometria do terreno.

Ao analisar a área construída dos imóveis, foi constado que, ela varia muito em função dos tipos de imóveis. Sendo observado a existência de muitos valores fora do padrão (*outliers*). Sendo assim, uma nova dimensão binária foi adicionada aos dados utilizando o intervalo I definido da seguinte forma: IQR o valor interquartil e q_n o valor do enésimo quartil $I = [q_{0,25} - 1.5IQR; q_{0,75} + 1.5IQR]$. Ao todo foram classificados imóveis 42.849 como sendo *outliers* e 833.968 não.

4. Arquitetura de Dutos e Filtros

4.1. Primeira Fase: Limpeza e Seleção

Na primeira fase da arquitetura os dados coletados do Portal de Dados Abertos são limpos e corrigidos da seguinte forma: Os campos nulos de área construída foram ajustados para zero caso sejam um lote vago e caso não seja foi utilizada a média da vizinhança. Dados nulos referentes à meio fio, pavimentação e iluminação foram corrigidos por imagens do Google Street View®. Os endereços fora de Belo Horizonte foram removidos.

4.2. Segunda Fase: Pré-processamento

A ferramenta GritBot da empresa RuleQuest Research é uma ferramenta automática que detecta anomalias nos dados, sendo considerada uma precursora dos algoritmos de mineração de dados. A publicação de Bay e Schwabacher [Bay and Schwabacher 2003] propuseram uma abordagem focada no desempenho para grandes bases de dados e realiza uma análise comparativa com a ferramenta, como conclusão, afirmam que ela apresenta bons resultados para bases não muito grandes. Sendo assim, ela foi utilizada nos dados de cada regional de forma separada para gerar uma lista de registros anômalos.

4.3. Terceira Fase: Transformação dos Dados

Para um melhor desempenho dos algoritmos de mineração de dados, a base de dados foi segmentada em nove bases referente aos imóveis de cada regional. As primeiras aplicações de mineração de dados, processavam as informações já transformadas em bancos transacionais. Para os dados geográficos as bases não têm esta propriedade. Sendo assim foi realizado uma avaliação de como transformar o banco de dados em um banco de transações.

4.3.1. Transações

Para analisar os dados através dos algoritmos de mineração de itemsets frequentes, inicialmente deve-se escolher uma forma de agrupar os itens em transações. Para esse estudo foi utilizado o agrupador do CEP, que é o código postal dos Correios do Brasil. Ele é utilizado para facilitar o encaminhamento e a entrega das correspondências aos destinatários. O código está relacionado indiretamente pelo uso e ocupação do espaço urbano, pois quanto mais correspondências um local possui, mais o logradouro terá CEPs distintos. Por exemplo, condomínios muito grandes possuem um código próprio. Ele também possui a vantagem de estar relacionado ao bairro, pois grandes avenidas ou ruas podem estar em bairros diferentes e apresentar vários códigos distintos para cada bairro.

A base de dados possui localidades que não têm CEP (valor do campo igual a zero), nestes casos o agrupamento utilizado foi o nome e tipo do logradouro e o resultado concatenado aos demais.

4.3.2. Itens

Para definir o conjunto dos itens, inicialmente, foi utilizado o produto cartesiano entre os valores das colunas de qualidade do acabamento do imóvel, o tipo construtivo e se a área construída é um *outlier*, totalizando 101 itens. Os tipos construtivos possuem a informação do uso do imóvel sendo residencial ou não, sendo assim não há necessidade de utilizar a coluna do tipo de ocupação do imóvel.

4.3.3. Utilidade

Os imóveis do espaço urbano podem variar muito em função da área construída, não somente pelos tipos e pela qualidade de acabamento. Como a área é um valor numérico em metros quadrados, ela foi adotada como valor e utilidade. O problema é que existem terrenos em que a área construída é zero. Para tal será feita duas abordagens, uma em que os terrenos sem construção são removidos da análise dos algoritmos, e uma abordagem em que será utilizada a área do terreno sem imóvel como uma utilidade negativa.

4.4. Quarta Fase: Mineração de Dados

Para minerar os dados a ferramenta SPMF [Fournier-Viger 2024] foi utilizada pois apresenta uma grande quantidade de algoritmos de mineração de dados com os mais diversos propósitos. Para descoberta de subgrupos a ferramenta Cortana Subgroup Discovery [Meeng and Knobbe 2011] por prover diversas formas de configuração de descoberta de subgrupos.

4.4.1. Análise dos Itensets frequentes

Para descrever as regiões geográficas foi escolhido os algoritmos de mineração de itens frequentes pela sua capacidade interpretativa. O trabalho de Luna et al[Luna et al. 2019], traz um excelente retrospectiva. Com base no princípio DCM foram filtrados apenas algoritmos mineração de itemset fechados. Um itemset é considerado fechado se e somente se não existe nenhum super conjunto com o mesmo suporte [Lucchese et al. 2004]. A partir deles é possível gerar todos os itemsets com suporte maior ou igual ao limite informado, porém sem a informação do suporte. Para a arquitetura, foi escolhido o algoritmo FPClose [Grahne and Zhu 2005] pela eficiência em gerar apenas candidatos válidos. O suporte mínimo de 40% foi utilizado para reduzir ao máximo o número de resultados.

Resultados As regionais Norte, Venda Nova, Barreiro, Nordeste, Leste e Noroeste apresentarm o perfil predominante de moradia baseado em casas com o padrão de acabamento variando do 1 ao 3 em seus CEPs, podendo ter ou não a presença de lotes vagos, exceto pela regional Norte na qual os lotes vagos apresentam-se de forma isolada. A regional Pampulha também possui o perfil predominante de moradia com casas com o padrão de acabamento variando do 2 ao 4 em seus CEPs podendo ter ou não a presença de lotes vagos. A regional Oeste apresentou perfil de casas com padrões variando do 2 ao 3 sem a presença de lotes vagos e lotes vagos isolados.

As regionais Leste e Noroeste apresetaram lojas com padrão de qualidade variando do 2 ao 3. A lojas situam-se em locais isolados, indicando centros comerciais e em conjunto com residências indicando a presença de comércio local.

Somente a regional Leste apresentou o tipo construtivo barracão em conjunto com casas de padrão 2.

A regional Centro-Sul apresentou todos os itemsets com apenas um item, o que pode ser uma característica do planejamento urbano inicial da Capital Mineira, em que as ruas foram muito bem definidas e segmentadas. A regional possui muitos imóveis com acabamento alto e lojas bem segmentadas por CEP, indicando a presença de centro comerciais, o que vai de encontro com a alta intensidade de coleta de lixo, indicando uma intensa

atividade humana na área. Somente neste local foram encontrados imóveis do tipo apartamento, o que indica um grande número de edifícios ligados á moradia.

A dinâmica da Capital não foge à dinâmica das grandes cidades em que os subúrbios fornecem moradias e a região central fornece trabalho, o que o algoritmo traz de novo é que as regiões mais próximas ao centro possuem uma quantidade significativa de comércio local e central com moradias de baixo padrão construtivo. Ao contrário das regiões mais distantes, que predominam apenas casas como moradia, distinguindo-se apenas os padrões de acabamento mais frequentes.

4.4.2. Análise dos itemsets estatisticamente relevantes

O trabalho de Webb e Vreeken [Webb and Vreeken 2013] propõe um novo algoritmo, chamado OPUS Miner, para encontrar associações interessantes em dados. Ao contrário de métodos tradicionais que retornam muitos padrões redundantes, o OPUS Miner visa encontrar um conjunto menor de associações autossuficientes que são mais propensas a serem úteis para o usuário.

Uma Associações Autossuficiente é definida da seguinte forma:

- Produtiva: Seus itens ocorrem juntos com mais frequência do que o esperado se fossem independentes.
- Não Redundante: Sua frequência não pode ser explicada pela frequência de seus subconjuntos.
- Independentemente Produtiva: Sua frequência não pode ser explicada pela frequência de seus superconjuntos.

O OPUS Miner é um algoritmo branch-and-bound que explora o espaço de busca de itemsets e usa limites eficientes para podar o espaço de busca e acelerar a descoberta. As suas limitações são: dificuldade em buscar itemsets muito grandes devido ao rigor estatístico e considera apenas associações positivas. Para análise o algoritmo foi configurado para as 10 regras de associação de maior medida de qualidade para atender o princípio DCM. Os parâmetros *check independency (filter)*, *search by lift* [Lift (data mining) 2024], *check redundancy* e *correction for multicompare* foram definidos como verdadeiros. A métrica de lift foi utilizada para que itemsets de baixo suporte não sejam penalizados, pois caso contrário o resultado seria muito próximo ao resultado do algoritmo anterior.

Resultados Este algoritmo demonstrou um grande poder para detectar padrões em imóveis de alta qualidade de acabamento, tanto moradias quanto comerciais. Estes imóveis sofrem grandes penalizações ao serem analisados por algoritmos de análise de suporte porque são muito infrequentes.

Em todas as regionais, exceto a regional Norte apresentaram itemsets com vaga de garagem não residencial de medio a alto padrão associado a imóveis do tipo sala de médio a alto padrão, demonstrando a importância de se ter estacionamento próximo aos centros de prestação de serviços.

Uma outra questão é que imóveis com as área construtiva excepcionais tendem a estarem no mesmo CEP que os imóveis de mesmo padrão construtivo com padrão de acabamento

ou superior. A regional Centro-Sul destaca-se apresentando 34 CEPs com sala padrão de acabamento 5 associadas a garagem comercial de padrão de acabamento 5. Demonstrando a tendência de que os imóveis comerciais de alto padrão devem ter o serviço de garagem próximo por padrão. Os resultados também mostram uma tendência, da regional Nordeste, de se adaptar imóveis de moradia para o comércio quando eles estão próximos aos imóveis comerciais.

4.4.3. Análise da utilidade em metros quadrados construídos

Ao analisar os imóveis, somente a frequência não é suficiente devido às características dos edifícios. Algumas regiões podem ter poucos deles, abrigando uma quantidade muito grande de pessoas, tanto com o intuito de moradia como comércio. Sendo assim o algoritmo FHM Freq, que é uma adaptação do algoritmo FHM [Fournier-Viger et al. 2014] para limitar os itemsets pelo suporte mínimo. No experimento foi utilizada a seguinte configuração: utilidade mínima como 1 Km² e no mínimo 30% de suporte.

Resultados Ao cruzar a informação da área construída com o suporte, obtém-se uma análise detalhada dos imóveis. A análise é complentar ao algoritmo FPMax. A mesma revela a presença de comércio local nas regionais Norte, Barreiro, Leste, Venda Nova, Noroeste, Nordeste e Oeste associados a casas. A regional Pampulha não apresentou imóeis comericais, indicando uma escassez, que pode ser tanto pela baixa quantidade ou pela pouca área construída.

As regionais Norte, Barreiro, Leste, Venda Nova, Noroeste e Nordeste apresentaram moradias do tipo barracão associadas a moradias do tipo casa.

Somente as regionais Pampulha, Oeste e Centro-Sul apresentaram moradias do tipo apartamento em localidades separadas e em conjunto com moradias do tipo casa. Em todos os caso o padrão de acabamento apresentado foi o três, tanto da casa quanto do apartamento. As regional Oeste apresenta uma grande área de moradias do tipo apartamentos com acabamento 3, totalizando mais de 4 km². Divergindo da análise anterior que sugeriu baixa verticalização. Acredita-se que a edificação da região se concentrou em poucos prédios de muitos andares com perfil de moradia, pois a regional abriga a maior populacão da capital. A regional Pampulha possui características de edifícios de apartamento semelhantes à regional Oeste e destaca-se por suas casas de acabamento 4, totalizando mais de 1,2 km² em 779 CEPs.

A regional Centro-Sul foi a única a apresentar itens com área construída muito acima da média total, especialmente lojas com acabamento 3 e apartamentos com acabamento 3 e 4. A região demonstra intensa edificação tanto vertical quanto horizontal. A associação de casas e apartamentos e a presença de lotes vagos pela análise anterior, indica que ainda há potencial para a criação de mais edifícios na região.

4.4.4. Análise da utilidade negativa

Para avaliar o impacto da existência de lotes vagos em cada regional, o algoritmo FHN [Fournier-Viger 2014] foi utilizado de duas formas, um com a mesma base de transações do algoritmo anterior e uma base com item de lote vago com a utilidade negativa. Para analisar todas as bases a utilidade mínima foi de 1Km². Nesta abordagem o que foi

avaliado foram o número de itemsets resultantes das duas execuções, caso o impacto de se ter muitos lotes vagos seja significativo o número de itemset da segunda execução será menor.

Resultados Nesta abordagem foi verificado que somente as regionais Pampulha e Norte tiveram um impacto negativo. Indicando que elas possuem CEPs com alternância de terrenos vagos com terrenos com imóveis construídos. [Conurbação 2024].

4.4.5. Análise de subgrupo

Para uma abordagem mais profunda dos imóveis da Capital, foi escolhido a análise do tema social de gentrificação [dos Santos Veloso and Teixeira de Andrade 2019, Solla 2019, Andrade and Mendonça 2020]. Os estudos citados levam em consideração regiões urbanas limitadas a bairros, para esta análise o objetivo é levantar os imóveis que podem estar em situação social vulnerável dentro de uma região urbana mais extensa, no caso a regional. Para avaliar esta questão, foi utilizado a abordagem de descoberta de subgrupos [Atzmueller 2015]. Ela é uma técnica descritiva que gera subgrupos de um banco de dados que possuem elementos que demonstram comportamentos interessantes em relação a um valor alvo. Para tal o valor alvo escolhido foi o padrão de acabamento 5, o mais alto. A partir do valor alvo apenas a regional Centro-Sul foi escolhida, pois ela contém a grande maioria dos imóveis de alto acabamento, diversidade de tipos construtivos e ocupação.

Criação do Banco de Dados Para a análise de subgrupo um novo banco de dados foi criado. Apenas os imóveis destinados a moradia que não fossem uma vaga de garagem foram escolhidos. Desta forma eles foram agrupados pelas coordenadas geográficas (utilizando o Sistema de Referência Geocêntrico para a América do Sul 2000, SIRGAS 2000) do centroide do terreno. As seguintes informações foram sumarizadas: área construída foi somada e para o caso de edifícios ela foi dividida pelo total de economias para obter a área construída média por núcleo familiar, a área do terreno, o tipo construtivo e o padrão de acabamento foram sumarizados pelo valor máximo. Ao todo foram coletados 14.956, sendo 1.567 registros de imóveis com padrão 5.

Configuração do Cortana Para configurar o Cortana foi escolhida a variável padrão de acabamento com o valor alvo P5; com a medida de qualidade de Jaccard com valor mínimo de 0,2; com a profundidade 5; estratégia best first; estratégia numérica best.

Resultados Foram gerados 2.993 subgrupos todos eles com a medida de qualidade variando de 0,44 à 0,2. Sendo assim iremos analisar o terceiro subgrupo, por ter sido menos redundante.

Cobertura	Qualidade	Probabilidade	Positivos	Regra
2.209	0,44	0,52	1.159	Construção >= 423m² ∧ Terreno >= 441m² ∧ Logitude <= 7796076,5 ∧ 608975,7 <= Latitude >= 614227

Tabela 2. Subgrupo escolhido

Padrão Acabamento	Registros Imobiliários
1 e 2	31
3	303
4 e 5	3.815

Tabela 3. Regra aplicada aos dados originais

Para analizar os fenômeno da gentrificação o filtro de área construída foi removido para obter residências menores, os registros foram agrupados pela zona homogênia e sumarizado pelo total de núcleos familiares. Os seguintes bairros com maior potencial foco de gentrificação foram: São Lucas, São Pedro, Vila Acaba Mundo e Funcionários.

A análise de subgrupo apresenta a vantagem de permitir que o processo social da gentrificação seja analisado em diversos níveis, não somente entre os extremos como foi feita esta análise. E em localidades geralmente associados a imóveis de alto padrão.

4.5. Quarta Fase: Conhecimento

As fases da arquitetura demonstraram ser capazes de extrair conhecimento mais amplos até os mais detalhados, como por exemplo encontrar padrões de imóveis que podem estar mais sujeitos a certos processos sociais. Esta fase constitui-se do relatório final contendo todo o conhecimento gerado. Para mais detalhe acessar os resutlados no repositório https://github.com/guinamen/aprendizado

5. Respostas

P1 Os algoritmos de mineração de dados mais úteis para descrever uma base de dados foram: FPMax, OpusMiner, FHM Freq, FHN e Descoberta de Subgrupo. Esta sequência de algoritmos descrevem os dados de forma mais genérica à mais específica. Com um mínimo de hiper parâmetros e um mínimo de resultados, todos os algoritmos geraram resultados de extrema importância para descrever o espaço urbano da Capital Mineira.

P2 Eles conseguem, na sequência proposta, gerar um conjunto de informações em vários níveis, do mais amplo ao mais expecífico.

P3 A melhor sequência de execução dos algoritmos foi a apresentada no trabalho, a grande importância do trabalho foi que para utilizar o algoritmo de descoberta de subgrupos de forma mais eficiente, a melhor forma é compreender a base de dados para segmentar muito bem os registros que serão processados. Diminuído o tempo de execução e melhorando a qualidade do resultado.

6. Conclusão

O trabalho demonstrou a eficácia de uma arquitetura de dutos e filtros com algoritmos de mineração de dados para descrever o espaço urbano de Belo Horizonte de forma compacta, revelando padrões e tendências no uso e ocupação do solo a partir dos dados imobiliários. A clara distinção entre as regiões centrais e periféricas, a importância de fatores como a presença de vagas de garagem para imóveis comerciais de alto padrão e a identificação de áreas potencialmente vulneráveis à gentrificação ilustram o potencial da abordagem.

A segmentação da análise por regionais administrativas permitiu observar nuances na dinâmica urbana, evidenciando a importância de se considerar as particularidades de cada

área na formulação de políticas públicas. A utilização de algoritmos como FPClose, OPUS Miner, FHM Freq, FHN além da análise de subgrupos com o Cortana, demonstrou a versatilidade da mineração de dados para responder a diferentes perguntas de pesquisa no contexto do planejamento urbano.

Apesar dos resultados promissores, a pesquisa apresenta limitações. A base de dados utilizada, embora extensa, representa um retrato estático da cidade e não captura a dinâmica temporal das transformações urbanas, os resultados devem ser avaliados por especialistas e outras bases deveriam ser utilizadas. A análise se concentrou em variáveis relacionadas aos imóveis, sendo necessária, também, a incorporação de outras dimensões, como dados socioeconômicos e de infraestrutura, para uma compreensão mais abrangente do espaço urbano.

Como trabalhos futuros, destaca-se a incorporação de dados espaçotemporais para análise de séries históricas, permitindo identificar tendências e prever cenários futuros de uso e ocupação do solo. A integração com outras fontes de dados, como redes sociais e imagens de satélite, também apresenta grande potencial para enriquecer a análise e gerar conhecimento ainda mais aprofundado sobre a dinâmica urbana. E uso de outras técnicas de mineração de dados como a mineração de dados de mesma localidade, para evitar a segmentação por regionais.

Referências

[Andrade and Mendonça 2020] Andrade, L. T. d. and Mendonça, J. G. d. (2020). Urban policies, mobility and gentrification in two neighbourhoods of belo horizonte 1. *Sociologia & Antropologia*, 10:561–586.

[Atzmueller 2015] Atzmueller, M. (2015). Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49.

[Bay and Schwabacher 2003] Bay, S. D. and Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 29–38.

[Caldiera and Rombach 1994] Caldiera, V. R. B. G. and Rombach, H. D. (1994). The goal question metric approach. *Encyclopedia of software engineering*, pages 528–532.

[Conurbação 2024] Conurbação (2024). Conurbação — Wikipedia, the free encyclopedia. [Online; acessado 17-Julho-2024].

[dos Santos Veloso and Teixeira de Andrade 2019] dos Santos Veloso, C. and Teixeira de Andrade, L. (2019). Sapucaí street: entertainment hub and commercial gentrification in belo horizonte. *International Journal of the Sociology of Leisure*, 2:43–61.

[Doshi-Velez and Kim 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[Fournier-Viger 2014] Fournier-Viger, P. (2014). Fhn: efficient mining of high-utility itemsets with negative unit profits. In *Advanced Data Mining and Applications: 10th Inter-*

national Conference, ADMA 2014, Guilin, China, December 19-21, 2014. Proceedings 10, pages 16–29. Springer.

[Fournier-Viger 2024] Fournier-Viger, P. (2024). Spmf an open-source data mining library. [Online; acessado 17-Julho-2024].

[Fournier-Viger et al. 2014] Fournier-Viger, P., Wu, C.-W., Zida, S., and Tseng, V. S. (2014). Fhm: Faster high-utility itemset mining using estimated utility co-occurrence pruning. In *Foundations of Intelligent Systems: 21st International Symposium, ISMIS 2014, Roskilde, Denmark, June 25-27, 2014. Proceedings 21*, pages 83–92. Springer.

[Grahne and Zhu 2005] Grahne, G. and Zhu, J. (2005). Fast algorithms for frequent itemset mining using fp-trees. *IEEE transactions on knowledge and data engineering*, 17(10):1347–1362.

[Grünwald 2007] Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

[Lift (data mining) 2024] Lift (data mining) (2024). Lift (data mining) — Wikipedia, the free encyclopedia. [Online; acessado 17-Julho-2024].

[Lucchese et al. 2004] Lucchese, C., Orlando, S., and Perego, R. (2004). Mining frequent closed itemsets without duplicates generation. *ISTI-CNR Technical Report*, 13.

[Luna et al. 2019] Luna, J. M., Fournier-Viger, P., and Ventura, S. (2019). Frequent itemset mining: A 25 years review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1329.

[Meeng and Knobbe 2011] Meeng, M. and Knobbe, A. (2011). Flexible enrichment with cortana–software demo. In *Proceedings of BeneLearn*, pages 117–119.

[Mikut and Reischl 2011] Mikut, R. and Reischl, M. (2011). Data mining tools. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(5):431–443.

[Nwagu et al. 2017] Nwagu, C. K., Omankwu, O. C., and Inyiama, H. (2017). Knowledge discovery in databases (kdd): an overview. *Int J Comput Sci Inf Secur (IJCSIS)*, 15(12):13–16.

[Proença 2021] Proença, H. M. (2021). *Robust rules for prediction and description*. PhD thesis, PhD thesis, Leiden University, https://hdl. handle. net/1887/3220882.

[Silver et al. 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

[Solla 2019] Solla, L. F. S. (2019). Resistência à gentrificação?: Estudo de caso do bairro bonfim em belo horizonte. Master's thesis, Universidade Federal de Minas Gerais.

[Sommerville 2011] Sommerville, I. (2011). Software engineering (ed.). *America: Pearson Education Inc*.

[Webb and Vreeken 2013] Webb, G. I. and Vreeken, J. (2013). Efficient discovery of the most interesting associations. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):1–31.

[Zaki and Meira 2014] Zaki, M. J. and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.