Aplicando algoritmos de clusterização para encontrar inconsistências em bases de dados fiscais

Virginia Queiroz[1], **Lara Sucupira Furtado** [2], **Vládia Celia Pinheiro**[1,3]

[1]Universidade de Fortaleza (UNIFOR)
Caixa Postal 60.811–905 – Fortaleza – CE.

[2]Departamento de Pós-Graduação em Engenharia de Transportes
Universidade Federal do Ceará – Fortaleza, CE.

[3]Secretaria Municipal das Finanças de Fortaleza
Fortaleza, CE.

Abstract. *Advancements in Geographic Information Systems and in digital governance have enabled many cities to implement digital and geocoded property databases. However, property registers have diverging information since decades old data was automatically fed into digital systems and remain in conflict with incoming more standardized registers. Such is the case of the property database in Fortaleza, where this study is based. An estimated 2048 registers on apartments buildings are currently inconsistent and require cleaning or normalizing. This paper presents how clustering algorithms can help find inconsistencies in property registries.*

Resumo. *Ainda que os dados sobre a propriedade estejam cada vez mais digitalizados, os registros ainda incluem informações desatualizadas advindas de bases de dados historicamente inconsistentes. Os registros de propriedade antigos são automaticamente inseridos em novos formatos digitais, ficando em conflito com campos e bases atuais mais padronizadas. Tal é o caso do banco de imóveis de Fortaleza, onde este estudo se baseia. Esse artigo apresenta como algoritmos de agrupamento podem ajudar a encontrar inconsistências nas bases prediais. Os resultados estimam que 2.048 registros prediais estejam inconsistentes, necessitando limpeza e correção.*

1. Introdução

Os impostos sobre a propriedade são um importante mecanismo de financiamento para serviços urbanos e infraestrutura pública [Medda 2011]. O cálculo de impostos exige uma análise de múltiplos fatores que determinam o preço da propriedade, um processo que requer uma base de dados atualizada. Embora muitas cidades tenham adotado sistemas digitais para carregar, monitorar e atualizar seus bancos de dados, ainda existem muitos valores cadastrais desatualizados, o que dificulta a avaliação e o cálculo de impostos [Eguino et al. 2020]. Por outro lado, avanços na área de Ciência de Dados tem sido fundamentais para incorporar ferramentas computacionais em órgãos públicos [Jordan and Mitchell 2015]. Algoritmos de Aprendizado de Máquina podem auxiliar na tomada de decisões, detectando padrões e estimando valores das propriedades (Chang et al., 2014).

Os algoritmos de agrupamento (*clustering/clusterização*) são uma poderosa ferramenta de aprendizado de máquina não supervisionado para explorar o grupamento de eventos e objetos. São também aceitos como uma metodologia para detectar anomalias, agrupando pontos similares em grupos e isolando aqueles distintos. Como exemplo, a clusterização foi aplicada em pesquisas para ajudar os auditores a avaliar reivindicações de seguro de vida e encontrar instâncias fraudulentas para direcionar os esforços de avaliação [Thiprungsri and Vasarhelyi 2011]. Embora falsos positivos ou negativos sejam uma preocupação, os autores consideram que o agrupamento é uma abordagem útil para sinalizar características suspeitas.

A clusterização espacial olha para esses processos dentro de um contexto geográfico específico contando com duas entradas de dados: atributos em estrutura de banco de dados e a localização espacial de cada objeto [Grubesic et al. 2014]. Por exemplo, [Geyer et al. 2017] aplicou métodos de agrupamento para interpretar dados sobre edifícios para identificar quais responderiam de forma semelhante à obras de *retrofit* devido às suas características construtivas. Essas informações foram usadas para implementar medidas de modernização que poderiam maximizar os impactos ambientais positivos, como a redução das emissões de carbono, com menor investimento.

Neste artigo, propomos um método baseado em clusterização espacial para conduzir análises na base de dados do Sistema de Informações Territoriais de Fortaleza (SITFOR), o cadastro predial e territorial de Fortaleza. O objetivo foi identificar informações inconsistentes sobre as características dos prédios de apartamentos em Fortaleza, de modo a auxiliar a Secretaria Municipal das Finanças de Fortaleza (SEFIN). A base do SITFOR reúne uma série de características sobre a construção e porte do edifício, totalizando 8005 prédios. Como problema, encontramos que algumas informações para unidades habitacionais localizadas em um mesmo edifício vertical estão estão diferentes na base de dados. Por exemplo, um apartamento localizado no segundo andar de um prédio não pode estar registrado no SITFOR como ”tendo elevador”enquanto outro apartamento no mesmo edifício tenha ”Não possui elevador”. Em suma, algumas características construtivas não devem variar entre as unidades, visto que muitas informações para uma unidade são constantes para todos os apartamentos de um edifício.

Tal problemática é importante pois impacta diretamente na valorização do imóvel. De acordo com a fórmula de cálculo do Imposto sobre a Propriedade Predial e Territorial Urbana (IPTU) de Fortaleza, o valor venal das unidades é calculado com base em diversos fatores, como tamanho do terreno, do prédio, sua localização, o tipo de acabamento da edificação e os tipos de equipamentos da habitação e urbanos existentes no logradouro. Logo, discrepâncias na base de dados que descaracterizem o padrão das unidades podem impactar negativamente na valorização do imóvel e assim na arrecadação de impostos. O método proposto identificou que cerca de 2.048 entradas de prédios de apartamentos apresentam discrepâncias, o que equivale a mais de 77.873 apartamentos individuais.

2. Revisão de Literatura

A Clusterização de Dados consiste numa técnica de agrupamento de objetos em subgrupos, ou *clusters*, baseado em critérios de similaridade ou diferença [Han et al. 2011]. Sua utilização visa obter *insigths* sobre dados não rotulados, apontando semelhanças ou discrepâncias em seus conteúdos como também visa a organização e resumo dos

destes dados. O agrupamento, por ser um métodos não supervisionado, não requer a identificação das propriedades e característica dos dados para o processo de treinamento, pois extraem informações e agrupam objetos com base em suas correlações e similaridades[Pu et al. 2020]. Em outras palavras, objetos com atributos semelhantes localizados dentro de uma localização geográfica específica são agrupados para extrair *outliers*.

Em [Pu et al. 2020], os autores usaram agrupamento e métodos de aprendizado de máquina baseados em *clustering outliers* (Sub-Space Clustering e One Class Support Vector Machine) para detectar anomalias manifestadas na forma de ameaças cibernéticas. Da mesma forma, [Leung and Leckie 2005] usaram um algoritmo de detecção de intrusão chamado detecção de anomalia não supervisionada para detectar anomalias. Em suma, o objetivo do agrupamento é "unir objetos em classes semelhantes, geralmente com o objetivo de minimizar a variação dentro do grupo e maximizar a variação entre grupos [Grubesic et al. 2014]."

As principais abordagens para clusterização são os métodos hierárquicos e particionais. No primeiro, o conjunto de dados é subdivido hierarquicamente de acordo com a sua confluência. Já nos métodos particionais, o conjunto de dados é subdividido em *n* número de grupos pré definidos que são formados segundo o critério de similaridade adotado.

[Zhang et al. 1996] pontuam que os métodos hierárquicos, tanto de aglomeração quanto de divisão, apresentam estimativa de complexidade de algoritmo de valor O (n^2), fazendo com que o processamento tenha menor desempenho na medida em que o valor de n (dados) aumenta, inviabilizando a utilização do método para base de dados muito grandes.

Quando falamos de algoritmos de agrupamento particionais, destacam-se o DBSCAN, Modelo de Mistura Gaussiana (GMM) e K-Means. O DBSCAN (Density-Based Spatial Clustering of Applications with Noise), que é um algoritmo de agrupamento baseado na densidade de agrupamento [Ester 1996], considera um "cluster"como uma série máxima de objetos ligados pela similaridade entre suas densidades. Qualquer ponto não incluído nessas séries é considerado ruído, o que minimiza o impacto de "outliers"na análise. Para a aplicação deste algoritmo, é necessário pré-definir o raio (geralmente se utiliza a distância Euclidiana) para o cálculo da densidade, bem como um número mínimo de "clusters"a serem formados. Contudo, devido à importância do conceito de densidade para este modelo, o DBSCAN pode não ser tão eficaz no agrupamento de conjuntos de dados de grande dimensão [Kriegel 2011].

Por outro lado, o Modelo de Mistura Gaussiana (GMM) também é muito utilizado em análise de dados [Bishop 2007]. Este método foi desenvolvido com o objetivo de corrigir a falta de reprodutibilidade dos resultados, característica do método K-Means. O GMM estima tanto o centroide quanto a forma geométrica dos "clusters", detetada por meio da posição e orientação da variância Gaussiana de cada "cluster"formado. Para a aplicação deste método, é necessário pré-definir o número de "clusters"[Ficklin et al. 2017]. Entre suas vantagens, destaca-se a capacidade de suportar grupos de geometria oval ou elíptica e fornecer probabilidades de um objeto pertencer a cada um dos possíveis agrupamentos, o que é extremamente importante para a análise

de dados. Contudo, o algoritmo não permite imprecisão na atribuição de objetos aos ”clusters”, o que limita sua eficácia no agrupamento de dados categóricos. Isto ocorre pois o algoritmo presume uma distribuição normal em todas as ”features”e é sensível a anomalias [Ranalli and Rocci 2014].

O modelo K-Means, um dos algoritmos de agrupamento mais conhecidos e amplamente utilizados, baseia-se na distribuição dos dados em um número específico de ”clusters”(k), escolhido pelo usuário. Este método procura minimizar a distância entre cada objeto e o centroide do respectivo ”cluster”. A seleção inicial do centroide é aleatória, o que pode causar variações nos resultados entre diferentes execuções do algoritmo e tornar o modelo mais sensível a ”outliers”[Xu 2005]. No entanto, para aumentar a eficiência, é possível recorrer a algoritmos para determinar o número ideal de grupos a serem formados, como o Método do Cotovelo, o Índice Davies-Bouldin ou o Método da Silhueta de Rousseeuw, entre outros.

No estudo realizado por [Işeri and Gursel Dino 2022], o K-Means, juntamente com os métodos do Cotovelo e da Silhueta, é empregado para prever o uso de energia e o conforto térmico de edifícios com características distintas. O método alcança pontuações entre 10 e 5%, um desempenho considerado bom para a técnica utilizada. Em outro estudo, [Aprilia and Agustiani 2021] aplica o K-Means para explorar a melhor implementação de impostos sobre terras em Butal, na Índia. Utilizando dados de mais de 3,3 milhões de edifícios (abrangendo cerca de 72 aldeias), foram formados grupos com base na eficiência da coleta de impostos, identificando 36 aldeias com menor probabilidade de adimplência. Neste caso, os pesquisadores optaram por definir os centroides iniciais do algoritmo, em vez de determinar o número de agrupamentos a serem formados.

O K-Means se sobressai em relação aos modelos hierárquicos por produzir agrupamentos simples [Ankerst 1999], que proporcionam melhor desempenho na execução. Sua principal vantagem é ser simples e de fácil implementação e chegar a resultados de fácil analise. Devido ao volume expressivo de dados a serem analisados e a necessidade de detectar discrepâncias entre apartamentos de uma mesma localidade, este estudo considera o potencial do método K-means, um modelo de Clusterização do tipo particionado, para detecção de discrepâncias entre informações de imóveis de um mesmo edifício.

3. Contexto da Pesquisa

A presente pesquisa tem como objeto de estudo os dados obtidos para Fortaleza, capital do estado do Ceará no Nordeste brasileiro. Em 2010, a (SEFIN) implementou o SITFOR, um banco de dados digital que usava Sistemas de Informações Geográficas para armazenar dados de propriedade em uma base georreferenciada. O Sistema permite que cada propriedade tenha suas feições associadas a coordenadas geográficas em um sistema de cadastro interativo.

No entanto, a transição para o SITFOR ainda está em andamento e lida com alguns entraves. Para atualizar a base, necessita-se administrar o legado de dados históricos coletados ao longo de 40 anos, antes dos avanços digitais. Esse legado de décadas anteriores foi automaticamente adicionado ao sistema SIG e agora gera informações divergentes de novos registros mais padronizados. Além disso, a dinamicidade dos dados imobiliários leva a constantes alterações, que também podem gerar discrepâncias que devem ser monitoradas pela equipe de avaliação imobiliária.

Como exemplo de discrepância, destaca-se o caso das divergências na base para construções verticais. Um prédio de apartamentos pode ter vários andares, mas todas as unidades de apartamentos, independentemente do andar, devem ter características idênticas no banco de dados para campos como: ”O prédio de apartamentos tem elevador?”ou ”Quantos andares tem o prédio de apartamentos?”. Essas características são muito importantes, já que o valor venal e o IPTU cobrado são diretamente proporcionais ao padrão da edificação. Infraestruturas como a presença de elevador, de garagem e de piscina, por exemplo, aumentam o valor do fator multiplicador que é utilizado na fórmula de cálculo do IPTU. Em suma, as características da construção recebem uma pontuação ponderada de acordo com seu indicativo de ”luxo”que é então utilizada para calcular o imposto.

Assim, resolver as inconsistências é imprescindível para garantir uma tributação justa e ter uma representação precisa dos imóveis em Fortaleza.

4. Metodologia

Essa pesquisa se utiliza de um método de agrupamento espacial para processar dados e melhorar o desempenho e a consistência do banco de dados de propriedades de Fortaleza [Grubesic et al. 2014]. O método consiste em três estapas: primeiramente, selecionamos quais características do edifício deveriam ser analisadas, visto que a base reúne cerca 47 atributos; depois os dados foram tratados e normalizados; por fim, aplicamos o algoritmo de agrupamento aos dados do SITFOR.

4.1. Variáveis estudadas

Após examinar os dados brutos, consultamos especialistas da SEFIN na célula do Registro de Imóveis de Fortaleza para compreender melhor seus desafios em manter a base do SITFOR. Realizamos duas reuniões quinzenais onde os especialistas explicaram o problema das discrepâncias no registro das edificações e nos ajudaram a avaliar todas as 47 características dos edifícios para identificar quais deveriam permanecer consistentes para um mesmo empreendimento vertical.

Em primeira instância, as 18 seguintes variáveis foram destacadas: Idade da edificação; posicionamento em rua, posicionamento em lote de terreno, material de cobertura, material de forro, material de piso, acabamento interno, acabamento externo, esgoto, iluminação pública, tipo de janela, vidro de janela, estado de conservação, jardim, pomar, piscina, garagem, elevador. Após uma análise de dados iniciais, percebeu-se que maior parte das variações se dava devido aos campos de material de piso e acabamento interno, o que é justificado já que tais características podem ser modificadas na unidade devido à reformas, etc.

Ao apresentarmos esses resultados para a SEFIN, conduzimos mais uma reunião de trabalho onde reavaliamos os fatores e apresentamos a metodologia de análise para validação dos especialistas. Concordou-se que a análise final deveria remover esses fatores internos já que cada apartamento pode ter tais informações diferentes, o que não indica discrepância na base. Restou um total de 16 variáveis que são aquelas onde todas as unidades de um mesmo prédio de apartamentos devem ter características idênticas, uma vez que são construídas com os mesmos materiais e padrões construtivos e estão localizadas em uma rua com igual acesso a serviços públicos. A Tabela 1 apresenta, para

cada variável, as respostas possíveis e o percentual de cada uma delas no banco de dados de imóveis do SITFOR. A quantidade de respostas possíveis mostra que a amostra tem uma alta variabilidade.

A distribuição de apartamentos por opção de resposta por *feature* está representada na Figura 1. Quanto maior a quantidade cores a coluna possuir, mais possibilidade de respostas distintas encontramos. Assim, é possível perceber que a heterogeneidade das possibilidades de respostas das *features* selecionadas e a presença de *outliers*, principalmente em relação a faixa de idade e acabamento externo. Isto justifica ainda mais a utilização algoritmo de *clusterização* sensíveis a dados divergentes.

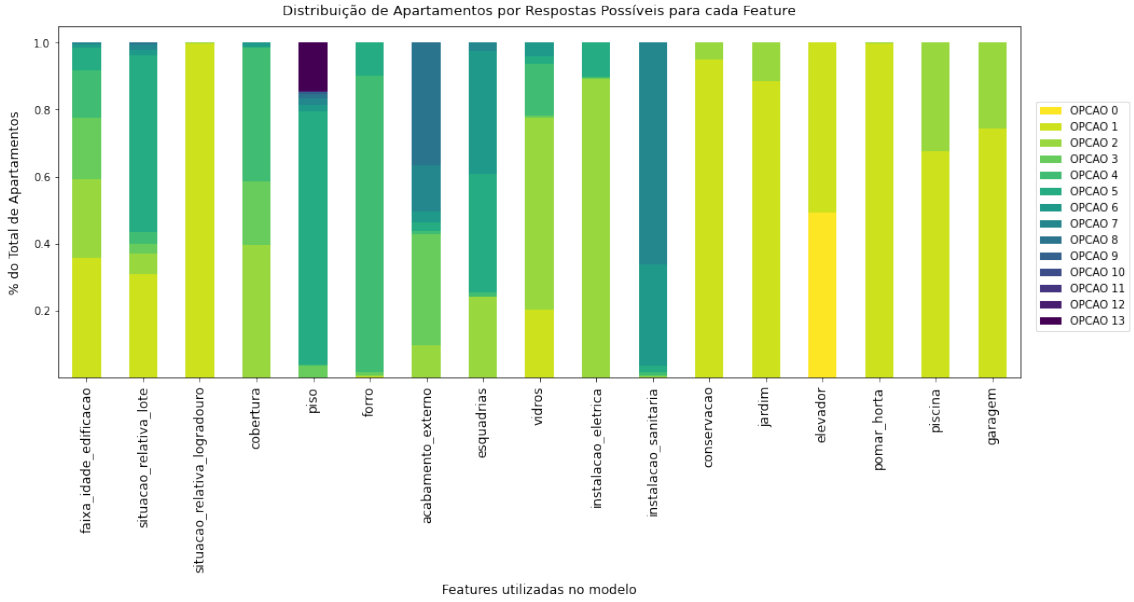

Figura 1. Distribuição de Apartamentos por Respostas Possíveis para cada *Feature*

4.2. Tratamento e análise dos dados

Iniciamos o processo normalizando a amostra para converter os dados em categóricos. Os valores numéricos foram codificados em categorias de intervalos. Para a análise, empregamos o método de agrupamento não hierárquico k-means, uma estratégia de otimização estruturada destinada a identificar *K "clusters"*. A premissa é que a soma dos objetos em K grupos deve ser igual ao número total de objetos, n. Em uma base de dados sem anomalias, esperamos que todas as inscrições de um mesmo empreendimento vertical formem apenas um agrupamento. Se um empreendimento apresenta apartamentos com características distintas, a possibilidade de combinações aumenta, gerando mais de um grupo. Assim, a indicação de anomalia ocorre quando inscrições prediais localizadas no mesmo espaço geográfico se encontram em mais de um grupo.

Quando há a necessidade de definir previamente o número de grupos para o processamento do algoritmo de agrupamento (o número de dobras K), utilizamos métricas como o Escore de Davies-Bouldin (DBS) e o Método do Cotovelo *"V-measures"* para avaliar o desempenho dos agrupamentos. O DBS é calculado levando-se em conta a média de similaridade (razão entre as distâncias entre objetos de um mesmo

Tabela 1. Features do edifício de apartamentos

Features dos Prédios	Opções de respostas
Faixa de Idade Edificação	Faixas a cada 10 anos
Situação Relativa Logradouro	Nulo (9.4%); Frente (88,5%); Fundos (4.1%); Galeria (0.02%); Vila (1,6%)
Situação Relativa Lote	Nulo (9.5%); Isolado recuado (26.2%); Isolado alinhado (6.3%); Recuado sem espaço lateral (17.1%); Alinhado sem espaço lateral (15.9%); Isolado superposto (21.6%); Isolado superposto alinhado (1.2%); Superposto sem espaço lateral alinhado (1.7%); Superposto sem espaço lateral recuado (0.4%)
Cobertura	Nulo (9.4%); Palha (0,02%); Cerâmica (62.6%); Fibra/Cimento (9.1%); Laje (16.2%); Metálica (1.5%); Especial (1.2%)
Forro	Nulo (9.4%); Sem forro (24.8%); Madeira (0.8%); Estuque (1.0%); Laje (58.8%); Especial (5.2%)
Acabamento Externo	Nulo (9.4%); Sem acabamento externo (1.9%); Caiação (28.9%); Pintura impermeável (27.9%); Pintura óleo (0.9%); Pintura plástica (1.6%); Aparente Rústico (1.8%); Aparente Luxo (6.1%); Especial (21.5%)
Instalação Sanitária	Nulo (9.4%), Sem instalação sanitária (2.2%); Externa com fossa (5.8%); Externa com rede de esgoto (5%); Interna com fossa simples (15.9%); Interna com fossa completa e rede de esgoto (2.9%); Interna com mais de um com fossa (18%); Interna com mais de uma rede de esgoto (40.7%)
Instalação Elétrica	Nulo (9.4%); Sem instalação elétrica (2.4%); Embutida (62%); Semi embutida (13.7%); Aparente Simples (7.3%); Aparente luxo (5.1%)
Esquadrias	Nulo (9.4%); Sem esquadrias (1.4%); Madeira (47%); Rústicas (1.5%); Ferro (6.7%); Alumínio (12.2%); Mista (19.9%); Especial (2%)
Vidros	Nulo (9.4%); Sem vidros (53.6%); Vidro comum (26.2%); Vitrais (0.8%); Vidro fumê (0.8%); Mista (1%); Especial (2.8%)
Estado de Conservação	Nulo (9.4%); Bom (62.9%); Regular (24.6%); Ruim (3%)
Jardim	Sim (7.2%); Não (92.8%)
Pomar	Sim (0.2%); Não (99.8%)
Piscina	Sim (12%); Não (88%)
Garagem	Sim (14.3%); Não (85.7%)
Elevador	Sim (18%); Não (82%)

grupo e objetos de outros agrupamentos) de cada grupo com o seu par mais similar [Davies and Bouldin 1979]. Logo, objetos menos dispersos resultam em pontuações mais baixas.

O valor do escore pode atingir até 2.09 para grandes conjuntos de dados, conforme a convenção aceita pela comunidade científica. Valores elevados para esses tipos de conjuntos de dados podem ser explicados pelo seu alto volume de objetos e dimensões, o que facilita uma maior dispersão entre os pontos do conjunto [Carusi and Bianchi 2019].

5. Resultados

Seguindo a metodologia previamente apresentada, executou-se uma série de experimentos de agrupamento no conjunto de dados completo de 234.476 objetos (inscrições prediais). O número máximo de grupos encontrados foi 8, sendo que cerca de 33% do conjunto de dados foi agrupado em mais de um grupo (ver Figura 2). A análise aponta um total de 2.048 prédios e 77.873 apartamentos que apresentam informações divergentes, ou seja, cujas características formaram mais de um grupo. A maior variação deve-se a 6 características: Faixa de Idade Edificação, Acabamento Externo, Situação Relativa Lote, Esquadrias, Vidros e Instalação Sanitária, apresentando, respectivamente 6, 6, 5, 5, 5 e 5 respostas distintas entre imóveis de um mesmo prédio.

Algumas dessas características detalha aspectos da construção de pequena escala e internas, as quais são muito difíceis de avaliar em grande quantidade. Essas descobertas podem indicar que a equipe de avaliação imobiliária deve reconsiderar os recursos usados para calcular os impostos sobre a propriedade, pois o conjunto de dados pode já estar configurado de maneira a perpetuar discrepâncias e resultados imprecisos.

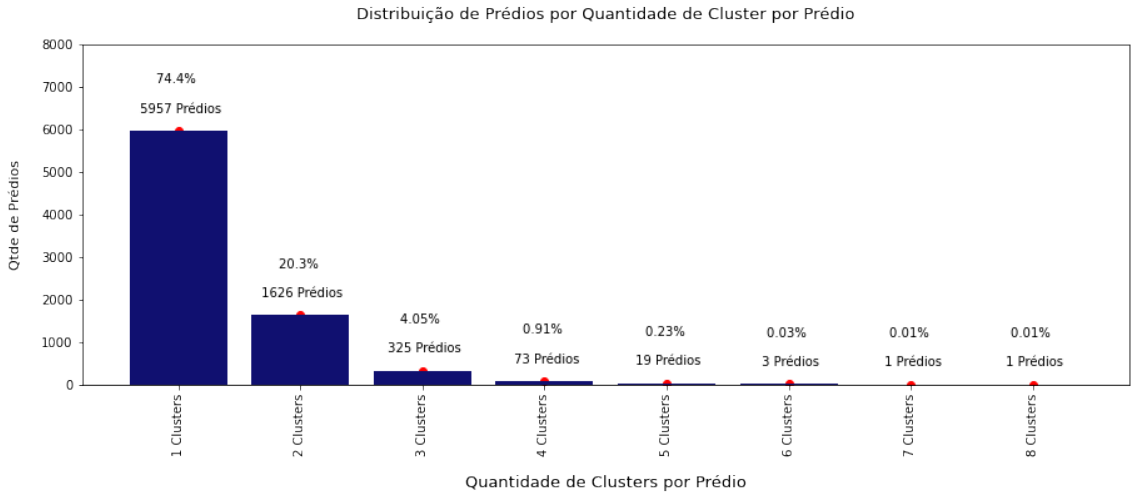

Figura 2. Gráfico de barras com o número de grupos gerados pelo algoritmo e quantos prédios ficaram em cada grupo

O mapa da Figura 3 ilustra a localização de todas as inscrições de apartamentos na cidade de Fortaleza (em cinza) e destaca aqueles onde o algoritmo de agrupamento apontou mais de um grupo. O gradiente de cores mostra um maior número de grupos em cores mais escuras e maior concentração de inscrições na área norte da cidade próxima ao Oceano Atlântico.

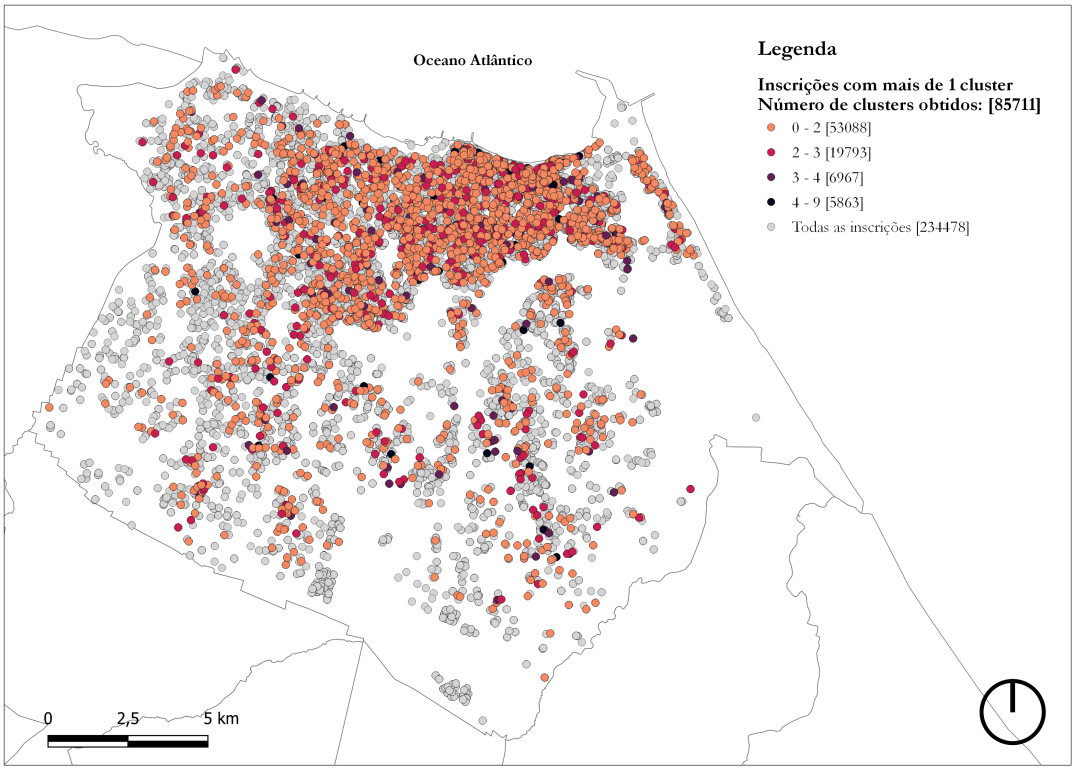

Figura 3. Mapa de Fortaleza com apartamentos georreferenciados e coloridos de acordo com a quantidade de *clusters* gerados

Percebe-se que o maior número de clusters (de 3 a 6) está localizado na parte norte do mapa, que, por sua proximidade com a orla, concentram os domicílios de maior renda. Tal distribuição fica ainda mais clara quando agregamos os dados a nível de bairro. A Figura 4 ilustra os bairros de Fortaleza coloridos de acordo com a contagem de inscrições em cada bairro que retornaram mais de um grupo.

Ainda que exista também um número alto em bairros mais periféricos (afastados da orla), há uma tendência de concentração de discrepância em bairros onde a população possui melhores condições socio-econômicas. Assim, os impostos calculados para edifícios de maior padrão têm maior probabilidade de serem imprecisos, o que prejudica o objetivo de promover justiça fiscal. Essa constatação está de acordo com declarações de funcionários do governo de Fortaleza, que apontam que os impostos prediais são proporcionalmente mais baixos nos bairros ricos.

6. Conclusão

Esta pesquisa contribui com estudos de agrupamento espacial, mostrando como o método pode ser usado com sucesso para detectar anomalias, observando o contexto dos dados sobre a propriedade. Descobrimos que cerca de 36% do conjunto de dados foi agrupado em mais de um *cluster*, ou seja, que possuem valores inconsistentes no banco de dados. Esta variação é causada principalmente por 4 características do edifício: Faixa de Idade Edificação, Acabamento Externo, Situação Relativa ao Lote e Esquadrias.

Reavaliar e estudar essas características individuais mais a fundo é de suma im-

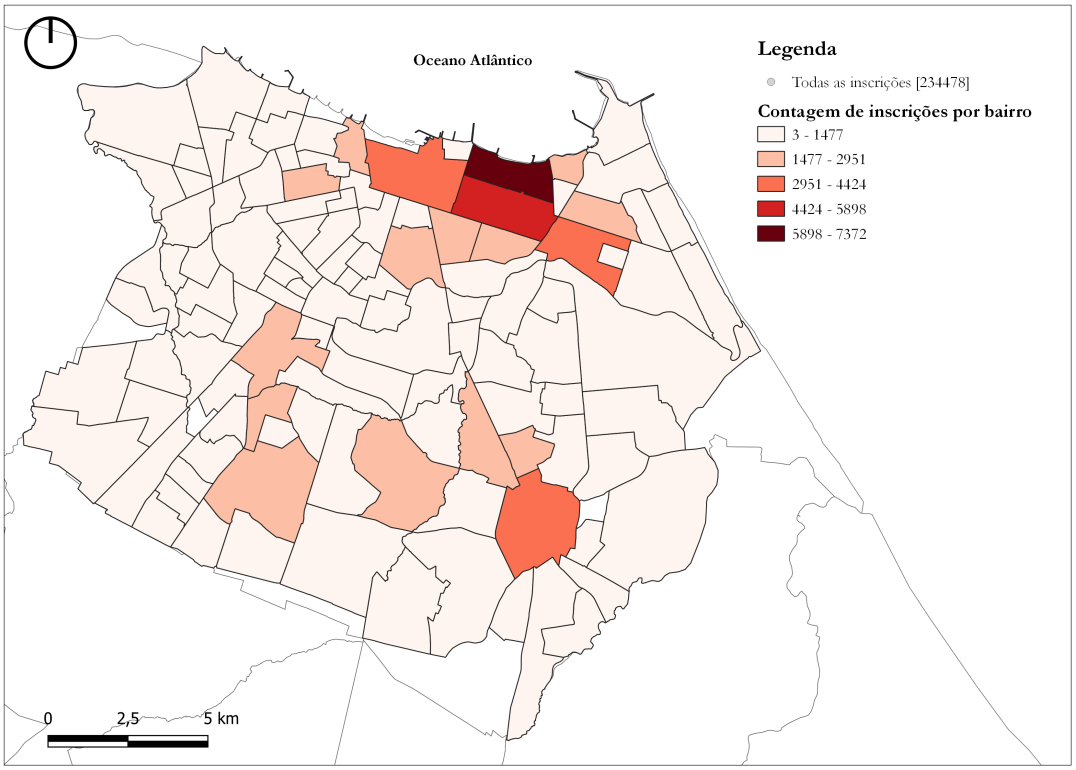

Figura 4. Bairros de Fortaleza categorizados de acordo com a contagem de inscrições que retornaram mais de um grupo após as análises

portância para que os funcionários do governo possam se organizar estrategicamente para corrigir discrepâncias mais sobressalentes, economizando recursos públicos. A implementação de uma solução que identifica e corrige discrepâncias em uma base de dados de valores de propriedades privadas tem um impacto significativo na promoção da justiça fiscal e na melhoria da arrecadação de impostos. Uma base de dados precisa e atualizada garante que todos os proprietários de imóveis são tributados de forma justa e equitativa, evitando a subavaliação ou a sobretaxação de propriedades. Assim, a aplicação correta de taxas de imposto baseada em valores de propriedades corretos pode aumentar a arrecadação de impostos, otimizando os recursos financeiros disponíveis para serviços urbanos e infraestrutura pública. Além disso, a transparência e a precisão nos cálculos de impostos promovem a confiança do público no sistema fiscal, incentivando a conformidade e a prontidão para o pagamento de impostos. Portanto, uma base de dados imobiliários precisa e confiável é um instrumento fundamental para uma gestão fiscal eficiente e justa.

Também enfatizamos a importância do envolvimento dos especialistas dos órgãos públicos em relatar as suas circunstâncias e problemáticas reais. Os trabalhos futuros buscarão testar outros algoritmos de agrupamento, aplicar esta metodologia a outras tipologias de edifícios, como casas, e estimar os impactos monetários de tais discrepâncias no cálculo final dos valores de imposto predial. Também planejamos criar um pipeline para ajudar os agentes governamentais a replicar essa metodologia e usar processos automatizados para otimizar seus recursos e gerar impactos positivos em maior escala.

Referências

Ankerst, Mihael ; M. Breunig, M. . K. H.-P. . S. J. (1999). Optics: ordering points to identify the clustering structure. In *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99)*, pages 49—-60, Philadelphia, PA.

Aprilia, H. and Agustiani, D. (2021). Application of data mining using the k-means algorithm in rural and urban land and building tax (pbb-p2) receivables data in bantul regency. *Journal of Physics: Conference Series*, 1823:012063.

Bishop, C. M. (2007). Pattern recognition and machine learning.

Carusi, C. and Bianchi, G. (2019). Scientific community detection via bipartite scholar/journal graph co-clustering. *Journal of Informetrics*, 13(1):354–386.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1(2):224–227.

Eguino, H., Erba, D., Da Silva, E., De Oliveira, A., Piumetto, M., Iturre, T., and Rodríguez, A. (2020). Catastro, valoración inmobiliaria y tributación municipal: Experiencias para mejorar su articulación y efectividad. *Informe del Banco Interamericano de Desarrollo (BID)*.

Ester, Martin ; Kriegel, H.-P. S. J. . X. X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, 96(34):226—-231.

Ficklin, S., Dunwoodie, L., Poehlman, W., Watson, C., Roche, K., and Feltus, F. (2017). Discovering condition-specific gene co-expression patterns using gaussian mixture models: A cancer case study. *Scientific Reports*, 7:5.

Geyer, P., Schlüter, A., and Cisar, S. (2017). Application of clustering for the development of retrofit strategies for large building stocks. *Advanced Engineering Informatics*, 31:32–47.

Grubesic, T. H., Wei, R., and Murray, A. T. (2014). Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers*, 104(6):1134–1156.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Işeri, O. and Gursel Dino, I. (2022). *Building Archetype Characterization Using K-Means Clustering in Urban Building Energy Models*, pages 222–236.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Kriegel, Hans-Peter; Kröger, P. S. J. Z. A. (2011). Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, 26.

Leung, K. and Leckie, C. (2005). Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38*, pages 333–342.

Medda, F. R. (2011). 04land value finance: Resources for public transport. *Innovative land and property taxation*, page 42.

Pu, G., Wang, L., Shen, J., and Dong, F. (2020). A hybrid unsupervised clustering-based anomaly detection method. *Tsinghua Science and Technology*, 26(2):146–153.

Ranalli, M. and Rocci, R. (2014). Mixture models for ordinal data: a pairwise likelihood approach. *Statistics and Computing*, 26.

Thiprungsri, S. and Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. *International Journal of Digital Accounting Research*, 11.

Xu, Rui; Wunsch, D. (2005). Survey of clustering algorithms. *Kdd*, 16(3):645—-678.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: An efficient data clustering method for very large databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, SIGMOD '96, page 103–114, New York, NY, USA. Association for Computing Machinery.