INTEGRAÇÃO DE TÉCNICAS MULTIVARIADAS E ESPACIAIS PARA A ANÁLISE DE DADOS CENSITÁRIOS E IMOBILIÁRIOS DE CADASTROS MUNICIPAIS

Integrating Multivariate and Spatial Techniques for Analyzing Census and Municipal Property Cadastre Data

Antônio Augusto Ferreira de Oliveira
Secretaria Municipal das Finanças de Fortaleza
Gabinete do Secretário
augusto.oliveira@sefin.fortaleza.ce.gov.br

Marco Aurélio Stumpf González
Universidade do Vale do Rio dos Sinos
Programa de Pós-Graduação em Engenharia Civil
mgonzalez@unisinos.br

Claudia Monteiro De Cesare
CMDeCesare & Associados
Responsável Técnica
cdcesare@uol.com.br

Luan Victor Vasconcelos Noberto
Secretaria Municipal das Finanças de Fortaleza
Célula de Gestão de Cadastros
luan.noberto@sefin.fortaleza.ce.gov.br

Klinsman Gledson Guimarães de Araújo
Secretaria Municipal das Finanças de Fortaleza
Célula de Gestão de Cadastros
klinsman.araujo@sefin.fortaleza.ce.gov.br

RESUMO:

Este estudo apresenta uma metodologia para qualificar o cadastro imobiliário de Fortaleza, utilizando dados do Censo IBGE e do Cadastro Imobiliário Municipal (CIM), por meio de técnicas multivariadas e espaciais. A Análise Fatorial por Componentes Principais (PCA) reduziu a dimensionalidade dos dados, resultando em quatro fatores principais que explicaram 70,78% da variância, relacionados à ocupação urbana, densidade populacional, valor de mercado e disponibilidade de terrenos vagos. Observou-se uma correlação elevada entre o número de domicílios do IBGE e as inscrições residenciais do CIM, embora o IBGE registre mais domicílios, sugerindo a desatualização do CIM devido à dificuldade de inclusão de imóveis informais ou erros no cadastro de imóveis formais. A técnica de regionalização SKATER foi utilizada para criar 12 zonas homogêneas, combinando diferentes agrupamentos espaciais e fatores principais. A análise revelou uma alta correlação entre a área de lotes vagos e a extensão dos setores censitários, além de uma relação inversa entre o valor de mercado dos imóveis e a vulnerabilidade social. A metodologia mostrou-se eficiente na detecção de inconsistências cadastrais e identificação de padrões territoriais, contribuindo para uma melhor gestão territorial, planejamento

urbano e tributação imobiliária.

Palavras-chave: PCA, Regionalização, Cadastro Imobiliário, Censo IBGE, Setores Censitários.

ABSTRACT:

This study presents a methodology to improve the property registry of Fortaleza by utilizing data from the IBGE Census and the Municipal Property Cadastre (CIM) through multivariate and spatial techniques. Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data, resulting in four main factors that explained 70.78% of the variance related to urban occupation, population density, market value, and the availability of vacant lots. A strong correlation was observed between the number of households IBGE recorded and the CIM's residential registrations, although IBGE registers more households. This suggests that the CIM is outdated, likely due to difficulties in including informal properties or errors in registering formal properties. The SKATER regionalization technique created 12 homogeneous zones, combining different spatial groupings and principal factors. The analysis revealed a high correlation between the area of vacant lots and the size of census tracts, as well as an inverse relationship between property market value and social vulnerability index. The methodology proved efficient in detecting cadastre inconsistencies and identifying territorial patterns, contributing to better territorial management, urban planning, and property taxation.

Keywords: PCA, Regionalization, Real Estate Registry, IBGE Census, Census Tracts.

1 INTRODUÇÃO

No Brasil, o censo é realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), que utiliza uma delimitação espacial, denominado setor censitário, como a menor unidade territorial para fins de controle, execução de pesquisa e divulgação dos dados estatísticos referentes à população, domicílios, renda, entre outras informações. Esses dados são a principal fonte de referência para o conhecimento das condições de vida da população, fornecendo importantes subsídios à Administração Pública e ao planejamento social e econômico do país (IBGE, 2024b). Além disso, estes dados contribuem com a gestão territorial e urbana dos municípios brasileiros.

Por outro lado, a administração dos cadastros imobiliários é de competência municipal na medida em que é a base para o ordenamento territorial, planejamento urbano, controle de uso, parcelamento e ocupação do solo urbano, conforme previsto no art. 30, incisos I e VIII, da Constituição Federal de 1988 (Brasil, 1988). Além disso, os cadastros imobiliários são a base para a tributação dos impostos municipais como o imposto predial e territorial urbano (IPTU) e o imposto de transmissão *intervivos* (ITBI).

Os dados disponibilizados pelo IBGE viabilizam realizar um amplo espectro de análises e estudos técnicos, integrando dados sobre a população e domicílios com as informações do cadastro imobiliário municipal (CIM), que abrangem imóveis e seus atributos. Esse cruzamento de dados facilita a identificação de padrões, a detecção de inconsistências e a proposição de novos recortes de zonas homogêneas, regiões fiscais e macrozoneamento, aprimorando a qualidade do CIM e a compreensão da dinâmica urbanística do município.

Nesse contexto, este trabalho propõe a utilização de técnicas multivariadas, como a Análise Fatorial por Componentes Principais (PCA), para reduzir a dimensionalidade dos dados e identificar os fatores mais relevantes que influenciam a ocupação e a organização territorial. Além disso, foi aplicada a técnica de regionalização SKATER, visando agrupar setores censitários de forma otimizada e

homogênea, facilitando o planejamento e a implementação de políticas tributárias e urbanísticas.

2 REFERENCIAL TEÓRICO

2.1 Dados censitários

Conforme o IBGE (2024b), os limites dos setores censitários são estabelecidos a partir de elementos facilmente reconhecíveis em campo, visando facilitar o trabalho do recenseador em identificar as unidades dentro de sua área de atuação, evitando omissões ou sobreposições de entrevistas e assegurando a qualidade estatística das pesquisas. Dessa forma, as regras de definição desses setores são elaboradas com o objetivo de garantir o êxito operacional dos censos e das pesquisas que utilizam esses setores como referência. O IBGE (2024b) ressalta que o tamanho do setor leva em consideração sua extensão territorial para que o recenseador possa realizar sua pesquisa dentre do prazo de coleta preestabelecido. Por exemplo, a área do setor em áreas urbanas de alta densidade de edificações deve vialibizar o levantamento de dados em entre 250 e 400 domicílios. Nessas áreas extremamente adensadas, o setor censitário tende a coincidir com os limites dos bairros do município. Entretanto, os setores censitários são mais desagregados do que os bairros, permitindo a análise mais detalhada e precisa da distribuição populacional, das condições de vida e do uso do solo. Portanto, os dados censitários são de grande interesse para a administração pública em nível municipal.

2.2 Análise fatorial por componentes principais

A análise fatorial pode ser entendida com a técnica estatística exploratória e multivariada de redução de dimensionalidade, onde, a partir de um conjunto com *k* variáveis intercorrelacionadas, possa se extrair *k* fatores (F_i) que sejam uma combinação linear daquelas, de tal forma que sejam ortogonais, isto é, não correlacionados entre si (Hair et al., 2018; Fávero; Belfiore, 2024). A equação 1 descreve a fórmula de cálculo da análise fatorial.

$$F_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ik}X_k \quad (1)$$

Onde: F_i é o i-ésimo fator extraído e $a_{i1} \ldots a_{ik}$ são os coeficientes de carga fatorial (loadings), representando a contribuição de cada variável $X_1, X_2, \ldots, X_k$ no fator F_i.

Os fatores são extraídos da matriz de correlações por meio da decomposição em autovalores e autovetores. A matriz de correlações **R** descreve as correlações lineares entre todas as variáveis do conjunto de dados. A sua decomposição em autovalores e autovetores é, na sua essência, uma exploração da estrutura interna de covariância dos dados, identificando as direções principais em que as variáveis estão correlacionadas e a magnitude dessas correlações. A matriz **R** é representada por:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \cdots & 1 \end{pmatrix} \quad (2)$$

Onde: os elementos ρ_{ij} da matriz **R** são os coeficientes de correlação de Pearson entre os pares de variáveis X_i e X_j. Por definição, a matriz **R** tem diagonal principal

com todos elementos iguais a 1, representando a correlação de uma variável com ela mesma. A decomposição da matriz **R** resulta da resolução da equação 3, também conhecida como equação característica:

$$\mathbf{R} \cdot \boldsymbol{v_i} = \lambda_i \cdot \boldsymbol{v_i} \quad (3)$$

Onde: $\boldsymbol{v_i}$ é o autovetor associado ao i-ésimo fator (F_i) e λ_i é o autovalor correspondente ao autovetor $\boldsymbol{v_i}$. Os autovalores indicam a quantidade de variância explicada por cada fator, de modo que os fatores são ordenados de acordo com a variância explicada. O primeiro fator, associado ao maior autovalor (λ_1) explica a maior parte da variância total presente no conjunto de dados, sendo seguido pelos demais fatores em ordem decrescente de variância explicada ($\lambda_1 \geq \lambda_2 \geq \ ... \geq \ \lambda_k \rightarrow F_1 \geq F_2 \geq \ ... \geq \ F_k$). Por outro lado, os autovetores ($\boldsymbol{v_i}$) representam as direções principais dos fatores extraídos dos dados.

A Análise Fatorial por Componentes Principais (PCA) é aplicada na regionalização de variáveis para reduzir a dimensionalidade dos dados e identificar padrões subjacentes. Ao transformar as variáveis originais em um conjunto menor de componentes principais, a PCA facilita a identificação de regiões homogêneas, otimizando a interpretação e a visualização dos dados, essencial para a segmentação espacial e planejamento regional (He, 2024). A PCA vem sendo aplicada em diversos trabalhos na área de regionalização (ver Ibebuchi et al., 2024; Rahman e Rahman, 2020; e Sudam et al. 2021; entre outros).

Esse processo permite reduzir as variáveis originais em um número menor de fatores, mantendo o máximo de informação possível e facilitando a análise dos dados. Por conta da autocorrelação intrínseca das dados censitários de 2022 e das variáveis do CIM, os fatores serão extraídos de uma base de dados única que integra as variáveis oriundas de ambas as fontes para depois serem utilizados numa análise de agrupamento espacial (regionalização) como forma de melhor entendimento da ocupação e organização territorial em Fortaleza.

2.3 Regionalização

A regionalização pode ser entendida como uma análise de agrupamento de dados na qual se impõe uma restrição espacial para a formação de *clusters* espaciais. Além da similaridade das variáveis dentro do grupo, na regionalização se exige que as unidades espaciais estejam conectadas por algum critério de vizinhança previamente definido. Esse critério é estabelecido pela criação de matrizes espaciais de vizinhança como as tradicionalmente usadas em econometria espacial. A abordagem da regionalização é amplamente usada, principalmente, em trabalhos de economia urbana e planejamento urbano, com a definição de zonas homogêneas definidas conforme critérios socioeconômicos, ambientais ou funcionais, permitindo a aplicação de políticas públicas eficientes aplicadas à distribuição de serviços urbanos, ao planejamento de transporte e à gestão de recursos territoriais. Existem diversas técnicas para regionalizar variáveis (ver Anselin, 2023a e 2023b).

Uma das técnicas de regionalização mais utilizadas é SKATER, acrônimo para *Spatial 'K'luster Analysis by Tree Edge Removal,* desenvolvida por Assunção et al. (2006). A técnica SKATER é utilizada para a regionalização de dados espaciais, segmentando áreas geográficas em regiões homogêneas com base em variáveis selecionadas, a partir da criação de um grafo mínimo, no qual as arestas representam similaridades entre áreas. Em seguida, são removidas as arestas para formar clusters

espacialmente contíguos e homogêneos. É uma técnica importante para analisar padrões espaciais, utilizada em planejamento urbano e estudos de mobilidade, entre outros (Furtado, 2014). Esta técnica tem sido utilizada para diversas aplicações que envolvem a regionalização (ver Ghiffary et al., 2023; Wu et al, 2024; Zareba et al., 2023). Por exemplo, Kim e Yoon (2021) estudaram a delimitação de áreas de Mobilidade Aérea Urbana (UAM), visando o gerenciamento de tráfego de aeronaves não tripuladas (UAS). Partindo da regionalização e análise de correspondência em áreas altamente urbanizadas nas cidades de São Francisco e Manhattan, foram incorporados dados populacionais e do espaço aéreo urbano em 3D para delinear limites regionais. Em seguida, avaliaram a viabilidade operacional e econômica de implementar UAM em cada uma das regiões classificadas em cinco categorias utilizando SKATER.

3 METODOLOGIA

3.1 Fonte de dados

Os dados do IBGE foram obtidos do “Censo Demográfico 2022: Malha de Setores Censitários preliminares” (IBGE, 2024b) que disponibilizam os arquivos vetoriais da malha de setores com dados de população e domicílios referentes ao Município de Fortaleza, cujos principais campos são apresentados na Tabela 1.

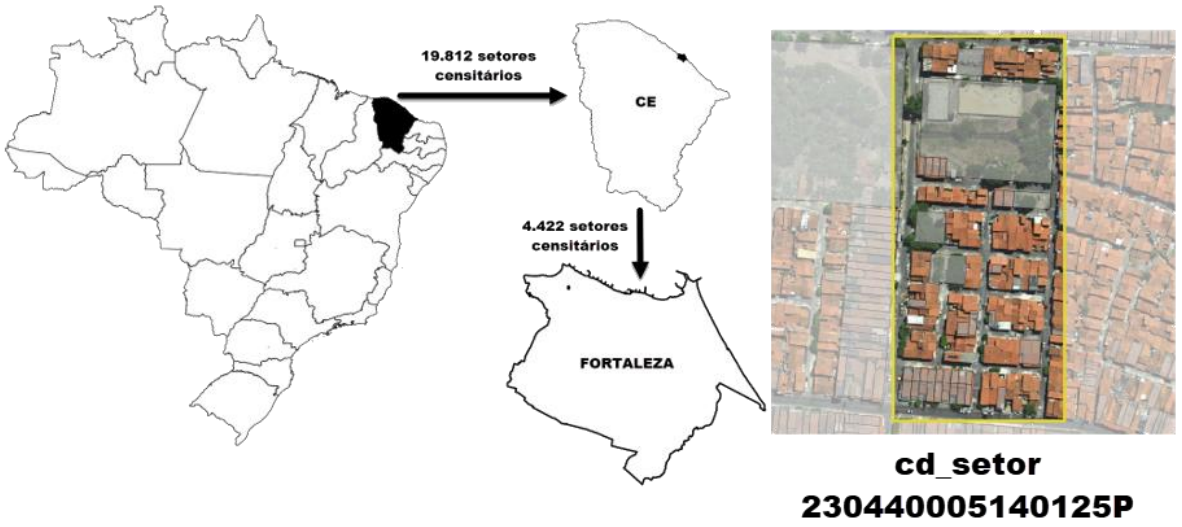
Tabela 1 – Principais campos e variáveis da malha dos setores censitários com dados de população e domicílios

Variável	Descrição
id	Geocódigo de setor censitário
a_set	Área do setor censitário em quilômetros quadrados
n_pes	Total de pessoas
n_d	Total de domicílios
n_dp	Total de domicílios particulares
n_dc	Total de domicílios coletivos
dpo_mean	Média de moradores em domicílios particulares ocupados
p_dpoi	Percentual de domicílios particulares ocupados imputados
n_dpo	Total de domicílios particulares ocupados
n_pes_a	Total de pessoas por km² de área do setor
n_dp_a	Total de domicílios particulares por km² de área do setor

Fonte: IBGE (2024b), com campos renomeados pelos autores.

Ao todo, foram utilizados 4.422 setores censitários, conforme apresentado na Figura 1.

Figura 1 - Setores censitários do Município de Fortaleza



Fonte: elaboração própria.

Os dados do CIM foram obtidos da Secretaria de Finanças do Município de Fortaleza (2024) e publicados na sua infraestrutura de dados espaciais. Foram agrupados os campos relevantes em relação aos dados do IBGE, organizados por setor censitário, conforme Tabela 2.

Tabela 2 - Variáveis do CIM testadas no trabalho

Variável	Descrição	Variável	Descrição
n_lot	Quantidade de lotes (parcela territorial) de acordo com o CIM	a_ed_nres	Área edificada total não residencial (m²)
a_lot_vg	Área dos lotes vagos (sem edificação) (m²)	a_t_med	Área mediana do lote (m²)
p_lot_vg	Percentual de lotes vagos (sem edificação) no setor (quant. de lotes vagos/quant. de lotes no setor).	i_med	Idade mediana das edificações (em anos)
p_ass	Percentual de assentamento precário	te_med	Testada mediana do lote (m)
p_pres	Percentual de área do setor dentro de Zona de Proteção Ambiental dos Recursos Hídricos	vv_med	Valor venal mediano das inscrições imobiliárias de acordo com lançamento do IPTU do exercício fiscal de 2024 (R$)
p_zeis	Percentual de Zonas Especiais de Interesse Social (ZEIS) no setor	imp_med	Valor mediano do IPTU de acordo com lançamento do exercício fiscal de 2024 (R$)
n_i	Quantidade de inscrições municipais	n_itbi	Quantidade de guias de ITBI pagas de 1º de janeiro de 2014 a 31 de julho de 2024.
n_i_res	Quantidade de inscrições municipais residenciais	vm_med	Valor unitário mediano da parcela territorial de acordo modelos de aprendizado de máquina (R$/m²)
n_i_nres	Quantidade de inscrições municipais não residenciais	ia	Índice aproveitamento ponderado[1]de acordo com o plano diretor
n_i_t	Quantidade de inscrições municipais territoriais	n_i_a	Quantidade de inscrições por km²

Variável	Descrição	Variável	Descrição
a_ed	Área edificada total (m²)	n_i_a_res	Quantidade de inscrições residenciais por km²
a_ed_res	Área edificada total residencial (m²)		

Nota: 1) o índice de aproveitamento foi ponderado pelas áreas de interseção, considerando as diferentes zonas dentro de um mesmo setor censitário.
Fonte: elaboração própria.

Além das variáveis das Tabelas 1 e 2, foi utilizado o rendimento nominal médio por responsável no setor censitário, expresso em salários mínimos (IBGE, 2011). Também foi incluído o Índice de Vulnerabilidade Social (IVS), calculado conforme a metodologia de Costa e Marguti (2015) para o IPEA, baseado na média aritmética de 16 indicadores relacionados a três dimensões: infraestrutura urbana, capital humano e renda/trabalho. Complementando o Índice de Desenvolvimento Humano Municipal (IDHM), o IVS oferece um mapeamento da exclusão e vulnerabilidade social nos municípios brasileiros. A divulgação do IVS foi baseada nos resultados do censo de 2010 relacionados a sua respectiva malha de setores (Costa e Marguti, 2015). Por conta disso, foi necessário adaptar aquele índice à malha do censo de 2022. Quando não havia dados disponíveis na nova malha, imputou-se o índice do setor mais próximo. O mesmo procedimento foi realizado para o rendimento nominal médio por responsável.

3.2 Etapas e procedimentos executados

Inicialmente, foi realizada a integração das variáveis obtidas das fontes de dados mencionadas em uma única base por meio do cruzamento espacial entre a malha dos setores censitários e a malha de lotes do CIM. Quando não havia espacialização do lote no CIM, utilizou-se do centroide da quadra onde este estava localizado para associação ao setor respectivo. Em seguida, calculou-se a correlação de Pearson entre os pares de variáveis de cada fonte, com destaque para aquelas que apresentaram as maiores correlações positivas ou negativas em valores absolutos. As variáveis com maior correlação foram consideradas candidatas para serem utilizadas na PCA para melhorar a compreensão das relações entre elas.

Finalizada a escolha das variáveis relevantes, verificou-se a adequação global da análise fatorial pelo teste de esfericidade de Bartlett que consistiu na elaboração da matriz de correlações **R** e na aplicação do teste da hipótese de que ela seja diferente de matriz identidade de mesma ordem (hipótese alternativa) para um determinado número de graus de liberdade e nível de significância (Fávero e Belfiore, 2024). O teste avalia se as variáveis são suficientemente correlacionadas para justificar a aplicação da análise fatorial.

Após a verificação dos critérios de adequação, procedeu-se à padronização das variáveis e à execução da PCA, que resultou na extração dos fatores com base nos autovalores da matriz de correlação e no cálculo das variâncias explicadas individualmente e acumuladas. Para definir o número de fatores a serem retidos, utilizou-se inicialmente o critério da raiz latente, que considera apenas os fatores cujos autovalores são superiores a 1, indicando que esses fatores explicam mais variância do que uma variável original individual. Constatou-se que quatro fatores explicavam quase 71% da variância acumulada. Além disso, utilizou-se o gráfico *scree plot* para validar a escolha dos fatores, reforçando a decisão de extrair exclusivamente aqueles que explicavam uma parcela significativa da variância. Desta feita, os quatro fatores

oferecem uma consolidação de todas as variáveis escolhidas para a análise, propiciando maior facilidade na interpretação dos resultados do estudo. Com os fatores selecionados, foram calculados os scores fatoriais, permitindo determinar o valor correspondente de cada fator nos diferentes setores censitários, o que permitirá identificar discrepâncias individualizadas por setor.

Posteriormente, foi realizada a regionalização de três maneiras distintas utilizando: (i) todas as variáveis originais, sem a redução dimensional; (ii) todos fatores calculados em substituição às variáveis originais; e (iii) apenas os 4 fatores extraídos da análise fatorial. Em todas as abordagens, aplicou-se a técnica SKATER. Para a definição da proximidade entre as regiões, foi escolhida uma matriz de pesos espaciais baseada na contiguidade do tipo *queen*, que considera como vizinhos todos os setores que compartilham uma borda ou um vértice. Essa matriz foi padronizada por linha. Para avaliar objetivamente a eficiência de cada regionalização, foram calculados o WSS (*Within-Cluster Sum of Squares*), que representa a soma total das variações dentro de cada agrupamento, e o BSS (*Between-Cluster Sum of Squares*), que mede a variabilidade entre as médias dos agrupamentos. O cenário ideal ocorre quando o WSS é minimizado, refletindo maior homogeneidade dentro dos grupos, enquanto o BSS é maximizado, indicando maior distinção entre eles.

4 ANÁLISE E DISCUSSÃO DOS PRINCIPAIS RESULTADOS

Entre as variáveis descritas nas tabelas 1 e 2, foram selecionadas aquelas que apresentaram os maiores coeficientes de correlação de Pearson entre si para a extração dos respectivos fatores na etapa posterior da PCA, quais sejam: a_lot_vg, vm_med, n_i_res, a_ed_res, n_i, a_ed, n_lot, n_i_t, n_i_a_res, n_i_a, ia, n_itbi, a_set, renda_sm, n_d, n_dp, n_dpo, ivs, n_pes, n_dp_a, n_pes_a. A Tabela 3 mostra os 10 maiores valores de correlação de Pearson obtidos:

Tabela 3 -Maiores correlações de Pearson entre as duas fontes de dados

Variáveis do CIM	Variáveis do IBGE	Corr.
a_lot_vg	a_set	0,87
vm_med	renda_sm	0,78
n_i_res	n_d	0,70
a_ed_res	renda_sm	0,68
vm_med	ivs	-0,64
n_i_res	n_pes	0,61
a_ed_res	n_d	0,60
n_lot	n_pes	0,55
n_lot	n_d	0,53
a_ed_res	n_pes	0,53

Fonte: elaboração própria.

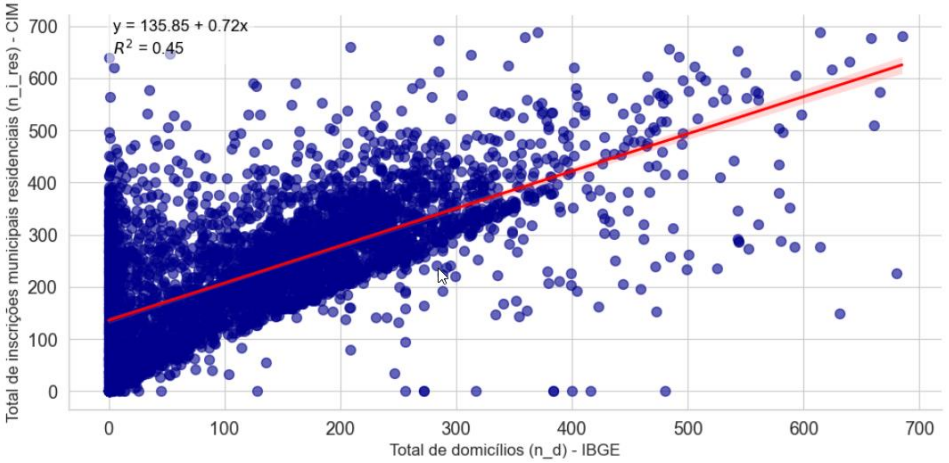
Observa-se uma alta correlação positiva de Pearson (0,87) entre a_lot_vg (área de lotes vagos - CIM) e a_set (área do setor censitário - IBGE), indicando que setores censitários maiores tendem a conter mais lotes sem construção. A correlação de Pearson de 0,70 entre *n_i_res* (número de inscrições municipais residenciais do CIM) e *n_d* (número de domicílios do IBGE) indica a forte relação entre essas variáveis. Cabe salientar que o IBGE considera a população residente "de direito", ou seja, os indivíduos são contabilizados em seu local de residência habitual na data de

referência, incluindo todos os moradores de domicílios particulares, permanentes e improvisados, e coletivos. A correlação com as inscrições de uso residencial do CIM reflete o alinhamento entre o número de unidades residenciais cadastradas e a quantidade de domicílios apurados pelo censo, embora o número de inscrições seja menor devido às limitações e dificuldades do CIM em acompanhar o surgimento de novas unidades informais. A correlação moderadamente forte de 0,68 entre *a_ed_res* (área edificada total residencial do CIM) e *renda_sm* (renda domiciliar do responsável do IBGE) sugere que setores mais densificados em termos de edificações tendem a estar associados às áreas nas quais residem famílias com maior poder aquisitivo. A correlação negativa (-0,64) entre as variáveis *vm_med* (valor unitário de mercado mediano do terreno) e *IVS* (índice de vulnerabilidade social) sugere que setores integrados por lotes de maior valor unitário apresentam menor vulnerabilidade social. Em outras palavras, áreas com lotes mais caros tendem a estar associadas a menores níveis de vulnerabilidade, possivelmente devido a suas melhores condições sociais, econômicas e urbanas.

A Em vários setores observou-se um maior número de inscrições residenciais do que domicílios. Analisando esses dados, observou-se que isso ocorreu devido ao lotes não espacializados no CIM serem enquadrados pelo seu centroide em outro setor que o correto e também, na minoria das vezes, inscrições que deveriam ter sido canceladas no CIM, por não mais existirem de fato.

Figura 2 destaca uma relação, embora imperfeita, entre o total de domicílios apurados pelo censo e o total de inscrições municipais residenciais no cadastro imobiliário (CIM). Observa-se que, enquanto o censo apurou 1.034.611 domicílios em 2022, o CIM possuía apenas 849.151 inscrições em julho de 2024, incluindo tanto as inscrições não residenciais quanto terrenos vagos. Essa discrepância evidencia a desatualização do CIM, que está provavelmente relacionada à dificuldade de inclusão de imóveis informais ou mesmo a falhas na inclusão dos imóveis formais. Além disso, a falta de uma delimitação física em ocupações territoriais desordenadas inviabiliza a inserção destes imóveis no cadastro, considerando os critérios padronizados de cadastramento dos imóveis. Em vários setores observou-se um maior número de inscrições residenciais do que domicílios. Analisando esses dados, observou-se que isso ocorreu devido ao lotes não espacializados no CIM serem enquadrados pelo seu centroide em outro setor que o correto e também, na minoria das vezes, inscrições que deveriam ter sido canceladas no CIM, por não mais existirem de fato.

Figura 2 - Diagrama de dispersão com reta de regressão entre o número total de domicílios (n_d) e o total de inscrições municipais residenciais do CIM (n_i_res) por setor censitário.



Nota: para fins de melhor visualização, os dados de domicílios e de inscrições foram limitados até 700. Fonte: elaboração própria.

As cargas fatoriais apresentadas na Tabela 4 representam os coeficientes de correlação de Pearson entre as variáveis e os fatores, facilitando a identificação das variáveis que contribuem mais para cada um deles. Para cada coluna dessa tabela, se destacou o maior valor de carga fatorial para fins de auxiliar na interpretação dos resultados, apresentada na sequência. Na parte final dessa tabela, apresentam-se a os autovalores, todos superiores a 1, e as variância explicadas individuais e acumuladas.

Tabela 4 - Cargas fatoriais

Fonte	Variável	F1	F2	F3	F4
IBGE	n_d	0,93	0,06	0,08	0,06
IBGE	n_pes	0,92	-0,04	0,08	0,04
CIM	n_i_res	0,78	0,45	0,05	0,05
CIM	n_lot	0,68	-0,02	-0,20	0,27
CIM	a_ed_res	0,63	0,68	-0,04	0,07
CIM	n_itbi	0,41	0,48	-0,04	0,26
CIM	n_i_t	0,22	-0,04	-0,10	0,67
CIM	n_i_a_res	0,14	0,25	0,65	-0,06
IBGE	a_set	0,13	0,02	-0,10	0,93
CIM	ia	0,12	0,58	0,14	-0,22
IBGE	renda_sm	0,06	0,87	0,03	0,05
IPEA	ivs	0,02	-0,79	0,00	-0,01
CIM	a_lot_vg	-0,01	0,00	-0,01	0,92
CIM	vm_med	-0,03	0,89	0,11	-0,05
IBGE	n_pes_a	-0,06	-0,09	0,95	-0,09
IBGE	n_dp_a	-0,07	-0,01	0,95	-0,07
CIM	p_ass	-0,32	-0,44	0,28	-0,09
Autovalor		4,99	3,07	2,22	1,75
Variância %		29,35	18,08	13,06	10,28

Fonte	Variável	F1	F2	F3	F4
Variância Acumulada %		29,35	47,43	60,50	70,78

Fonte: elaboração própria.

O Fator 1 está relacionado à ocupação urbana e densidade populacional residencial nos setores censitários, focando em variáveis que indicam o número de domicílios, pessoas, número de lotes e imóveis residenciais. Esse fator é especialmente relevante na análise dos “vazios cadastrais”, apontando áreas onde a Administração Pública deve priorizar os esforços de atualização do cadastro imobiliário. O Fator 2 captura características associadas ao valor de mercado de imóveis e renda média do responsável, refletindo a condição socioeconômica dos setores. A participação da variável de percentual de assentamento precário (*p_ass*) predomina nesse fator, embora com um coeficiente de correlação negativa moderado (-0,44). O fator 3 se relaciona com a densidade populacional e inscrições por área, refletindo setores com maior número de pessoas e domicílios por km². O fator 4, que reflete a disponibilidade de terrenos vagos e o uso do solo, indica setores com grande quantidade de áreas não edificadas, sugerindo um potencial para desenvolvimento urbano ou mesmo ausência de cadastro de novas áreas edificadas.

A Figura 4 mostra uma aplicação prática da detecção dos “vazios cadastrais” por meio do Fator 1 que tem alta carga de correlação com o número de domicílios (IBGE) e o número de inscrições municipais (CIM). Valores muito discrepantes em relação a esse fator indicam setores com disparidades entre essas variáveis. Na Figura 3, as divisas dos setores estão representadas em amarelo, as divisas dos lotes georreferenciados estão na cor branca e a restituição das edificações obtidas da segmentação do Google estão em preto.

Observa-se que os setores destacados em vermelho apresentam valores negativos elevados e, em muitos deles, não há sequer delimitação dos lotes no CIM. Em contraste, o valor projetado para o setor destacado em azul é positivo (0,77), sendo que a maioria dos seus lotes estão espacializados no CIM. Isso demonstra como a PCA pode funcionar como uma ferramenta poderosa para identificar inconsistências entre os dados censitários e municipais, auxiliando na detecção de áreas propensas à investigação adicional ou atualização no cadastro municipal.

Figura 4 - Recorte dos setores censitários com Fator Principal 1 (F1) calculados.

Nota: valores negativos, marcados em vermelho, demonstram desatualização cadastral; enquanto os setores em azul apresentam F1 positivo, indicando que o número de imóveis no cadastro está atualizado. Fonte: elaboração própria.

Conforme mencionado anteriormente, a regionalização foi conduzida de três formas distintas (ver 3.2). Em todas as abordagens, foi aplicado o método SKATER com 12 agrupamentos definidos aprioristicamente. Cabe salientar que Fortaleza está dividida em 12 regionais administrativas. A Tabela 5 apresenta indicadores de homogeneidade intraclusters (WSS) e interclusters (BSS), bem como a relação entre eles. Espera-se com tais cálculos ter um critério mais objetivo para escolha e aferição da melhor regionalização.

Tabela 5 - Análise de Variância Intra (WSS) e Interclusters (BSS) nas Diferentes Regionalizações

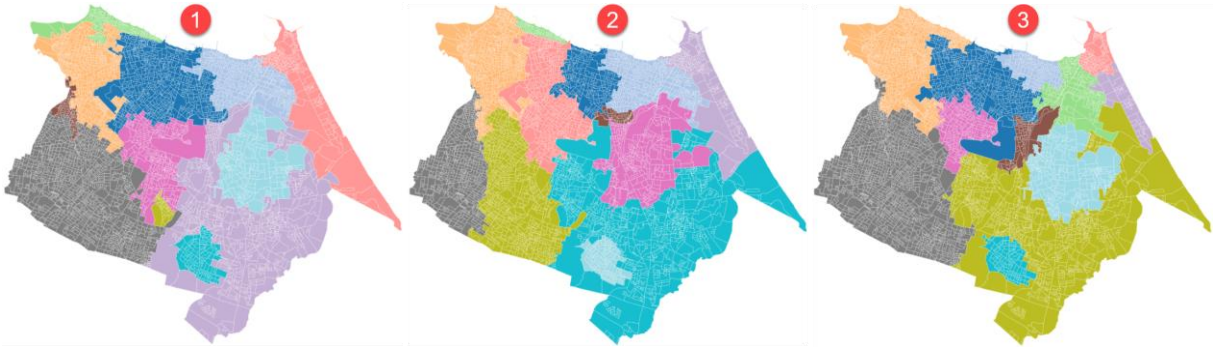
Regionalização	WSS	BSS	Razão BSS/TSS
1	128.956,01	21.391,99	0,14
2	68.065,25	7.108,75	0,09
3	14.239,45	3.448,55	0,19

Nota: regionalização 1 é aquela onde se utilizou todas as variáveis originais sem a redução dimensional; 2 é a regionalização com todos os fatores e 3 é a regionalização com os 4 fatores principais extraídos. Fonte: elaboração própria.

Entre as três regionalizações feitas com o método SKATER, a regionalização 1 tem a maior variabilidade interna (WSS) e uma diferenciação moderada em relação às outras regiões. A regionalização 2 apresenta menos variabilidade interna, mas também é menos diferenciada das demais. A regionalização 3 é a mais homogênea (menor WSS) e tem a maior proporção de variabilidade explicada pelas diferenças entre regiões (maior BSS/TSS). Isso sugere que a terceira regionalização, a que utiliza apenas 4 fatores, é a que melhor separa os dados, com a maior proporção de

variabilidade explicada pelas diferenças entre as regiões, embora os valores absolutos de BSS sejam relativamente baixos. A Figura 5 apresenta os resultados das regionalizações.

Figura 5 - Mapas das regionalizações com 12 agrupamentos (*clusters*)



Fonte: elaboração própria.

5 CONCLUSÃO

O estudo revelou algumas divergências significativas entre os dados censitárias e as variáveis do cadastro imobiliário de Fortaleza (CIM), indicando a necessidade de atualização do CIM, principalmente nas áreas mais ao sul da cidade e também nas áreas com predominância de assentamentos precários. A correlação positiva entre o tamanho dos setores censitários e o número de lotes vagos destaca o potencial de desenvolvimento urbano nos setores com maior extensão territorial, com ressalvas para as áreas de proteção ambiental. Como esperado, fatores como a renda e o valor de mercado dos imóveis mostraram correlações inversas com a vulnerabilidade social, indicando que as áreas mais valorizadas tendem a ser menos vulneráveis. A combinação de PCA e regionalização SKATER proporcionou uma visão clara dos padrões territoriais e inconsistências entre os dados censitários e imobiliários, permitindo a identificação de áreas prioritárias para intervenção pública. Essa abordagem é uma ferramenta valiosa para a delimitação de zonas homogêneas e regiões fiscais utilizadas para múltiplos fins no âmbito da administração municipal.

Entre as limitações do trabalho, destaca-se a sensibilidade dos parâmetros do SKATER, que resulta em agrupamentos bastante distintos dependendo das variáveis utilizadas. Além disso, o caráter periodico do Censo e da produção do IVS inviabiliza que os dados sejam comparados na mesma data de referência.. Pesquisas futuras podem explorar os novos dados a serem divulgados pelo IBGE, principalmente aqueles relacionados a rendimento e emprego, bem como testar outras técnicas de regionalização, aplicação de modelos econométricos espaciais e aprendizado de máquina, capazes de prever divergências entre dados censitários e imobiliários. Por fim, cabe enfatizar que a metodologia desenvolvida no presente estudo apresenta um extraordinário potencial de contribuir com a qualificação do cadastro imobiliário dos município, melhorando o planejamento e a implementação de políticas urbanas e tributárias.

AGRADECIMENTOS

Os autores agradecem a Secretaria de Finanças do Município de Fortaleza pelos dados utilizados nesse trabalho. O segundo autor agradece o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de

Financiamento 001.

REFERÊNCIAS

ANSELIN, Luc. Local Indicators of Spatial Association—LISA. **Geographical Analysi**s, v. 27, n. 2, p. 93-115, abr. 1995. DOI: https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

ANSELIN, Luc. **An introduction to spatial data science with GeoDa**, V.1: Exploring spatial data, 2023a. Disponível em: https://lanselin.github.io/introbook_vol1/index.html. Acesso em: 21 ago. 2024.

ANSELIN, Luc. **An introduction to spatial data science with GeoDa**, V.2**:** Spatially Constrained Clustering - Hierarchical Methods, 2023b. Disponível em: https://geodacenter.github.io/workbook/9c_spatial3/lab9c.html. Acesso em: 21 ago. 2024.

ASSUNÇÃO, R. M.; NEVES, M. C.; CÂMARA, G.; FREITAS, C. da C. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. **International Journal of Geographical Information Science,** v. 20, n. 7, p. 797-811, ago. 2006. DOI: https://doi.org/10.1080/13658810600665111.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Senado Federal, 1988.

COSTA, Marco Aurélio; MARGUTI, Bárbara Oliveira (eds.). **Atlas da vulnerabilidade social nos municípios brasileiros**. Brasília: Ipea, 2015. 77 p. Disponível em: https://repositorio.ipea.gov.br/handle/11058/4381. Acesso em: 21 ago. 2024.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de análise de dados**: estatística e machine learning com excel®, SPSS®, Stata®, R® e python®. 2.ed. Rio de Janeiro: LTC, 2024. 1255 p.

FURTADO, Victor Marinho. **Agrupamento de Conjuntos de Instâncias: Uma Aplicação ao ENEM**. 2014. Dissertação (Mestrado). Programa de Pós-graduação em Engenharia de Sistemas e Computação. Universidade Federal do Rio de Janeiro, 2014. Disponível em: https://www.cos.ufrj.br/uploadfile/1417018604.pdf. Acesso em: 21 ago. 2024.

GHIFFARY, Ghardapaty G. et al. Application of spatial ‘K’luster analysis by tree edge removal (SKATER) on infrastructure inequality mapping. In: AIP Conference Proceedings. AIP Publishing, 2023. DOI: https://doi.org/10.1063/5.0181062

HAIR, Joseph F.; ANDERSON, Rolph; BLACK, William; BABIN, Barry. **Multivariate data analysis**. 8. ed. Cengage Learning Emea, 2018. 832 p.

HE, X. (2024). Principal Component Analysis (PCA). In: **Geographic Data Analysis Using R**. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-97-4022-2_8

IBEBUCHI, Chibuike Chiedozie; OBAREIN, Omon A.; ABU, Itohan-Osa. Application of autoencoders artificial neural network and principal component analysis for pattern extraction and spatial regionalization of global temperature data. **Machine Learning: Science and Technology**, v. 5, n. 1, p. 015009, 2024. DOI: 10.1088/2632-2153/ad1c34

IBGE. **Censo 2022:** informações de população e domicílios por setores censitários auxiliam gestão pública. Agência de Notícias IBGE, 12 jun. 2023. Disponível em:

https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/39525-censo-2022-informacoes-de-populacao-e-domicilios-por-setores-censitarios-auxiliam-gestao-publica. Acesso em: 21 ago. 2024.

IBGE. **Censo demográfico 2022** : registros de nascimento : resultados do universo. Rio de Janeiro: IBGE, 2024a. 61 p. Disponível em: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=73110. Acesso em: 21 ago. 2024.

IBGE. **Malha de Setores Censitários preliminares**. Rio de Janeiro: IBGE, 2024b. 43 p. Disponível em: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2102072. Acesso em: 21 ago. 2024.

IBGE. **Sinopse do censo demográfico**. Rio de Janeiro: IBGE, 2011. 265 p.. Disponível em: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=249230. Acesso em: 21 ago. 2024.

KIM, Namwoo; YOON, Yoonjin. Regionalisation for urban air mobility application in metropolitan areas: case studies in San Francisco and New York. **International Journal of Traffic and Transportation Management**, v. 3, n. 2, p. 1-8, 2021. DOI: 10.5383/JTTM.03.02.001.

RAHMAN, Ayesha S.; RAHMAN, Ataur. Application of principal component analysis and cluster analysis in regional flood frequency analysis: a case study in New South Wales, Australia. **Water**, v. 12, n. 3, p. 781, 2020. DOI: https://doi.org/10.3390/w12030781.

SECRETARIA DA FINANÇAS DE FORTALEZA. **Infraestrutura de Dados Espaciais da SEFIN - PMF**. Fortaleza: PMF, 2024. Disponível em: https://ide.sefin.fortaleza.ce.gov.br/visualizador. Acesso em: 21 ago. 2024.

SUDAM, G.V., CHATURVEDI, A., JAYAKUMAR, K.V. (2023). Regionalization of Precipitation in Andhra Pradesh and Telangana State by Using PCA. In: Szymanski, J.R., Chanda, C.K., Mondal, P.K., Khan, K.A. (eds) **Energy Systems, Drives and Automatio**ns. ESDA 2021. Lecture Notes in Electrical Engineering, vol 1057. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-99-3691-5_46

WU, Chao et al. Urban Green Space Assessment: Spatial Clustering Method Based on Multisource Data to Facilitate Zoning Planning. Journal of Urban Planning and Development, v. 150, n. 4, p. 04024032, 2024. DOI: https://doi.org/10.1061/JUPDDM.UPENG-486

ZAREBA, Mateusz et al. Big-data-driven machine learning for enhancing spatiotemporal air pollution pattern analysis. Atmosphere, v. 14, n. 4, p. 760, 2023. DOI: https://doi.org/10.3390/atmos14040760