Análise massiva de dados na gestão pública: Uma proposta para identificação de *outliers* no cadastro de imóveis da prefeitura de São Paulo

Luiz Fernando Cavalcante Silva (Universidade de São Paulo)
lfcavalcante@usp.br

Renato de Oliveira Moraes (Universidade de São Paulo)
remo@usp.br

Hugo Martinelli Watanuki (Universidade de São Paulo)
hwatanuki@usp.br

Leandro Ramos da Silva (Universidade de São Paulo)
leandro.rsilva@usp.br

A quantidade de dados disponibilizados pelos governos são maiores a cada dia, gerando diversos conjuntos de dados com padrões e velocidades de atualização diferentes. Esses dados em conjunto entram no conceito de Big Data e a análise desses poderia ser benéfica para diversas áreas. O objetivo deste artigo é descrever um procedimento para identificar imóveis atípicos da cidade de São Paulo. Para tanto usou-se dados do cadastro de imóveis da cidade usado para o cálculo e gestão da arrecadação do imposto territorial e urbano. Os dados foram processados na plataforma HPCC Systems (High Performance Computing Cluster) da LexisNexis Risk Solutions. Os resultados mostram que as dificuldades estão mais na concepção do modelo e na ingestão dos dados e menos no processamento do grande volume de dados.

Palavras-chave: Dados Públicos, Big Data, Análise Multivariada, e-Governo.

1. Introdução

Diversos tipos de dados são disponibilizados pelo poder público com o intuito de aumentar a transparência dos governos em relação às suas políticas e práticas administrativas, esses dados caracterizam-se por quantidades massivas, oriundo de diferentes fontes, que nem sempre estão conectadas ou seguindo os mesmos padrões de dados. Isso os enquadra no conceito de *big data* que, segundo Laney (2001, apud Victorino, et al, 2017), é caracterizado por grande "volume" de dados, adquiridos em alta "velocidade" e com grande "variedade" de formatos, ou "3Vs". O acesso e a usabilidade desses dados não são, geralmente, simplificados o suficiente para que se possa realizar diretamente alguma análise sobre eles, ou realizar consultas. Essas limitações implicam na baixa utilidade dessa grande quantidade de dados tanto para a sociedade civil, quanto para as organizações empresariais.

A prefeitura do município de São Paulo disponibiliza em seu site dados do cadastro de imóveis da cidade. Esta base contém informações relacionadas ao terreno e à edificação e são usados para o cálculo e gestão da arrecadação do imposto territorial e urbano (IPTU). Ele contém cerca de 3 milhões de registros de imóveis residenciais, comerciais e industriais. Esta base de dados possui potencial de geração de valor público e também privado. Na área pública, a análise poderia identificar imóveis com características atípicas, diferentes do padrão geral de sua categoria, que seriam candidatas a uma análise individual mais criteriosa como, por exemplo, inspeção por um fiscal da prefeitura.

O objetivo deste artigo é descrever um procedimento para identificar imóveis atípicos da cidade de São Paulo. Isso foi feito através análise de cluster, e dentro de cada cluster foram identificados e ordenados os *outliers* – valores extremos, elementos atípicos. O capítulo seguinte apresenta a revisão da literatura, a terceiro descreve a metodologia adotada, o quarto capítulo apresenta os resultados observados e o último traz considerações finais sobre o estudo.

2. Revisão da literatura

De acordo com Alves (2018), sobre questões de órgãos de Inteligência governamentais, a crescente quantidade de dados disponíveis gera uma sobrecarga de informações ao analista com poder de decisão e exemplifica a questões de agências possuírem dados e não conseguirem fazer correlações entre eles para prevenir ações adversas, afirmando que a solução para isso é a utilização de análise de b*ig data* com auxílio de aplicações de inteligência artificial para facilitar o trabalho do analista.

Extrapolando as afirmações para fora do meio de Inteligência, Victorino (2017) sugere que não é trivial processar o volume de dados gerado pelo governo com a finalidade de gerar informações e ideias úteis para interessados, mas que a solução consiste na estruturação de um ecossistema de b*ig data* para análise desses dados governamentais.

Maciejewski (2017) afirma que não há muitas publicações detalhadas sobre aplicação de *big data* no setor público, mas que há casos reais de utilização com resultados significativos, o que justifica a análise do uso de *big data* em relação a possíveis aplicações no setor público.

Os dados presentes no registro de imóveis da cidade de São Paulo (GEOSAMPA, 2019), por exemplo, podem ser utilizados para estimar o valor venal desses imóveis. Esse valor é utilizado para aplicações das alíquotas do IPTU e representa, de algum modo, o valor intrínseco do imóvel, não equivalendo ao valor de mercado. A Figura 1 ilustra como o é determinado o valor venal do imóvel para efeito de cálculo do IPTU.

Figura 1 – Cálculo dos valore venais (A, B e C)

Fonte: Secretaria Municipal da Fazenda de São Paulo (2021)

A análise de tais dados pode ser feita, por exemplo, por meio da análise de *clusters*, ou agrupamentos, corroborada por Jain (2010, apud Gonçalves, et al, 2018), que afirma que a organização de dados em agrupamentos é o método mais fundamental de aprendizagem e compreensão. Essa análise, de acordo com Heil e Volpi (2013), visa separar os registros similares em grupos, colocando registros não semelhantes em outros grupos.

Dentro dessa classe de análise, diferentes algoritmos podem ser utilizados. Um algoritmo simples e eficiente bastante utilizado com grande quantidade de dados numéricos dimensionais é o *Kmeans* (Gonçalves et al., 2018). Trata-se de um algoritmo não hierárquico que depende da entrada do usuário para definição do número de cluster ou grupos, sendo que essa entrada consiste em um registro qualquer contendo as variáveis de modo a representar um centroide representativo de um grupo em questão.

Entretando, de acordo com Xu et al. (2019), como resultado do rápido crescimento de tamanho e velocidade de geração de dados, algoritmos como o Kmeans tendem a enfrentar desafios para

processar conjuntos de dados na memória principal de um único computador. Isso de deve ao fato de o algoritmo principal do KMeans depender de hiperparâmetros para identificar o modelo mais ideal, o que requer a execução do algoritmo várias vezes com valores diferentes para o hiperparâmetros. No caso do KMeans, um valor pré-definido do hiperparâmetro K, representando o número de centróides, e os valores iniciais para os centróides, são necessários para treinar o modelo; e para enfrentar esses desafios o KMeans pode ser executado em sistemas de computação paralela e distruída de alto desempenho.

Nesse sentido, uma alternativa viável é a plataforma HPCC (*High Performance Computing Cluster*) Systems cujo sistema de código aberto para computação intensiva de dados foi desenvolvido pela LexisNexis Risk Solutions em 2000 (MIDDLETON; CHALA, 2011). Trata-se de um sistema de computação distribuída altamente escalonável com base em hardware comum e que fornece um ambiente de processamento paralelo capaz de tornar o algoritmo principal do Kmeans operacional ao lidar com volumes massivos de dados (XU et al., 2019).

3. Metodologia

A pesquisa tem caráter exploratório voltado à análise da viabilidade de manipulação de dados públicos de modo eficiente, considerando o conceito de b*ig data*, gerando resultados úteis para gestores públicos e outros interessados.

3.1 Obtenção do conjunto de dados

Na primeira abordagem, escolheu-se o conjunto de dados, do ano de 2019, referentes ao IPTU da cidade de São Paulo, disponibilizados pela Prefeitura de São Paulo. Essa escolha deu-se por motivos de maior facilidade de extração dos dados, disponibilização em período anual, maior estruturação e completude, com ano anterior ao do início do projeto.

Esse conjunto foi baixado da plataforma GEOSAMPA, mantida pela prefeitura de São Paulo (GEOSAMPA, 2019), onde estão disponíveis diversos dados. O conjunto de registro de imóveis encontra-se separado por ano na plataforma e cada um deles pode ser baixado no formato CSV.

3.2 *Upload* na plataforma HPCC *Systems*

O *upload* na plataforma é simples, não há necessidade de conversão de formato ou edição do arquivo CSV, a única ação necessária é a sinalização da codificação do arquivo na plataforma, UTF-8, em caso contrário, alguns caracteres latinos do registro não serão reconhecidos e certos nomes aparecerão com símbolos trocados.

3.3 Definição do *Layout* dos registros

Os dados já são estruturados, assim, pela plataforma, houve a necessidade de nomeação de cada campo ou coluna dos dados, essa nomeação foi idêntica a contida no registro original. Além disso, definiu-se o tipo de dado presente em cada campo, podendo o campo conter uma cadeia de caracteres ou um número, em suas diversas variações, como inteiros, decimais ou reais. A Tabela 1 mostra parte do *layout* dos registros disponíveis na base de dados da prefeitura.

Tabela 1 – Layout parcial dos dados

referencia_do_imovel	cep_do_imovel	quantidade_de_esquinas_frentes	fracao_ideal	area_do_terreno	area_construida	area_ocupada	valor_do_m2_do_terreno	valor_do_m2_de_construcao
CD PLACE DES VOSGES	05614-040	4	0,0080	17960	529	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0080	17960	529	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0080	17960	529	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0080	17960	529	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0081	17960	533	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0081	17960	533	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0081	17960	533	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0081	17960	533	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0082	17960	540	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0082	17960	540	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0082	17960	540	13824	2852	2213
CD PLACE DES VOSGES	05614-040	4	0,0082	17960	540	13824	2852	2213

3.4 Cálculo do valor venal

O cálculo do valor venal tem como base alguns parâmetros quantitativos diretos, como área do terreno e da construção, e outros qualitativos, como o tipo do terreno, presença do imóvel em condomínio e subdivisão da zona urbana que o imóvel está localizado, que recebem valores numéricos de acordo com a especificação. Existem três tipos de procedimentos para obtenção dos valores relacionados aos parâmetros. O primeiro procedimento é a retirada direto dos dados do registro, como o tipo do terreno, e necessitam apenas de programação para que os coeficientes sejam aplicados corretamente no lugar da informação qualitativa (Tabela 2).

Tabela 2 – Coeficientes para o tipo de terreno

TABELA III - FATORES DIVERSOS
(Tabela III, anexa à lei nº 10.235, de 16/12/86)

1	Fator terreno encravado	0,50
2	Fator terreno de fundo	0,60
3	Fator terreno interno	0,70

Fonte: Secretaria Municipal da Fazenda de São Paulo (2021)

O segundo procedimento depende de cálculos realizados no próprio conjunto de dados, em que os resultados servem de índice para outras tabelas disponibilizadas como anexo em leis específicas, casos esses em que se monta um CSV e se realiza os procedimentos de integração na plataforma de modo a permitir a relação dessa tabela com o conjunto de registros principal.

Como exemplo, o fator de profundidade, em que se tem a profundidade equivalente dividindo a área do terreno pela testada, dados disponíveis no conjunto de dados, e se associa um fator a essa profundidade (Tabela 3).

Tabela 3 – Coeficientes para o tipo de terreno

Tabela I - Fatores de Profundidade

Profundidade Equivalente	Fator	Profundidade Equivalente	Fator
até 10	0,7071	69	0,7614
11	0,7416	70	0,7559
12	0,7746	71	0,7506
13	0,8062	72	0,7454
14	0,8367	73	0,7402
15	0,8660	74	0,7352
16	0,8944	75	0,7303
17	0,9220	76	0,7255
18	0,9487	77	0,7207
19	0,9747	78	0,7161
de 20 a 40	1,0000	79	0,7116

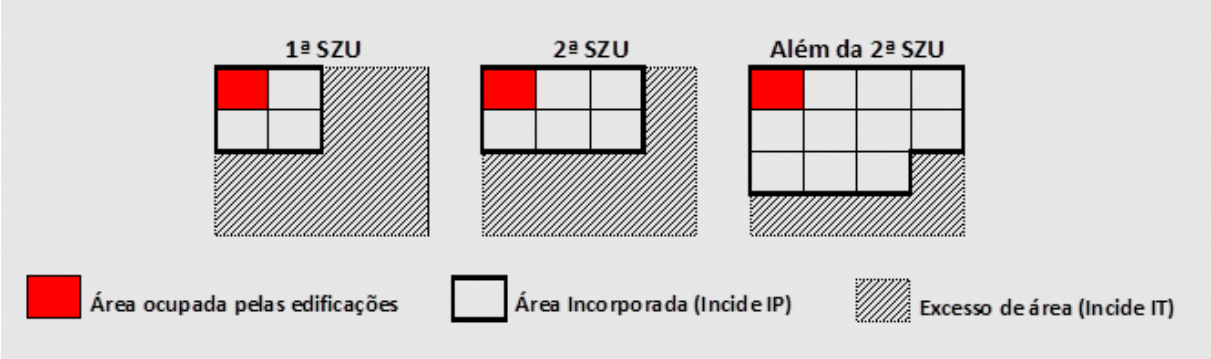
Fonte: Secretaria Municipal da Fazenda de São Paulo (2021)

O terceiro procedimento é semelhante ao segundo, mas a lógica dos dados é diferente. Para a obtenção da subdivisão da zona do imóvel, necessária para a relação de coeficientes de outros parâmetros (Figura 2), deve-se utilizar a tabela que relaciona valores de metro quadrado de construção com cada padrão de construção, disponibilizada em legislação. Os valores devem ser atualizados de acordo com decreto vigente e os procedimentos de *upload* e integração na plataforma devem ser refeitos para comparação com os dados de registro originais (Tabela 4).

Tabela 4 – Subdivisões da zona urbana por valor e padrão de construção

Tipop	Valor	Zona
Residencial horizontal - padrão A	1137	1SZU
Residencial horizontal - padrão B	1383	1SZU
Residencial horizontal - padrão C	1755	1SZU
Residencial horizontal - padrão D	2374	1SZU
Residencial horizontal - padrão E	2732	1SZU
Residencial horizontal - padrão F	3153	1SZU
Residencial vertical - padrão A	1259	1SZU
Residencial vertical - padrão B	1631	1SZU
Residencial vertical - padrão C	2139	1SZU

Figura 2 – Guia de cálculo do excesso de área e da área incorporada, relacionado da subdivisão da zona urbana



Fonte: Secretaria Municipal da Fazenda de São Paulo (2021)

3.5 Definição das variáveis a serem utilizadas e transformação

As variáveis escolhidas para análise foram as quantitativas existentes, necessárias direta ou indiretamente ao cálculo venal, e variáveis qualitativas julgadas relevante pelos autores da pesquisa, sendo essas últimas transformadas em quantitativas para que possam ser analisadas nos algoritmos de agrupamento (TABELA 5). Manteve-se o registro de cada imóvel e o CEP para identificação, mas essas variáveis não entrarão na análise.

Tabela 5 – Variáveis parciais utilizadas para análise

id	quantidade_de_esquinas_frentes	area_do_terreno	area_construida	area_ocupada	ano_da_construcao_corrigido	quantidade_de_pavimentos	testada_para_calculo	zona	condominio	area_exc
1	0	3350	25	978	1986	17	50	1	1	0
2	0	8625	200	6234	2006	30	75	1	1	0
3	0	2280	14	527	1990	19	40	1	1	172
4	0	5500	147	1239	2002	24	39	1	1	544
5	1	5272	93	4861	2007	25	64.6	1	1	0
6	0	8701	12	7210	2010	30	60.2	1	1	0
7	0	3350	49	978	1986	17	50	1	1	0
8	0	1710	198	1708	1998	17	30	1	1	0
9	0	1946	164	1849	2009	21	28.4	1	1	0
10	0	5750	17	4964	2005	26	50	1	1	0
11	0	3250	302	1010	2002	21	48.3	1	1	0
12	0	5750	117	4964	2004	22	50	1	1	0

3.6 Padronização dos dados

Para evitar que as análises de agrupamentos fossem afetadas pelo efeito de escala das variáveis, o valor de todas foi padronizado, como recomendado por Yu, et al (2017), utilizando a fórmula z-score, em que se considera o valor da variável no registro, subtrai-se a média e se divide pelo desvio padrão.

3.7 Análise de agrupamentos

A plataforma utilizada possui dois algoritmos para agrupamentos, o KMeans e o DBSCAN. O KMeans é um algoritmo não hierárquico que depende da entrada do usuário para definição do número de cluster ou grupos, sendo que essa entrada consiste em um registro qualquer contendo as variáveis de modo a representar um centroide representativo de um grupo em questão. O DBSCAN não tem classificação quanto à hierarquia e não depende de entrada direta para o número de grupos, mas possui entradas quanto à quantidade mínima de registros em um grupo

e a distância mínima entre registros para que sejam considerados do mesmo grupo, com isso, ele define um número de grupos para os registros e os classifica em cada um deles.

Idealmente, seria útil a utilização de um método hierárquico para definição de um número de grupos e posterior utilização de um método não hierárquico para definição de cada grupo, mas, devido à falta do primeiro na plataforma, escolheu-se utilizar o DBSCAN com o parâmetro de número mínimo de registros fixo em 2 e o valor de distância variando, com início de 0.1 e passo do mesmo valor, até que o número de clusters estivesse menor que 20, números maiores seriam difíceis de terem significados atribuídos, e o número de *outliers* razoável (menos de 10% do valor total de registros). Assim, seriam selecionados um registro de cada grupo gerado e eles seriam utilizados como centroides dos agrupamentos a serem calculados pelo Kmeans, que tornaria os grupos mais heterogêneos entre si e homogêneos em seus interiores. Essa opção resolveria as desvantagens do Kmeans que, de acordo com Yu, et al (2017), referem-se a dificuldade de determinar o número de clusters pelos centroides iniciais, que também afetam os clusters resultantes. O método foi testado com um número menor de imóveis, todos do mesmo CEP. Inicialmente sendo bem-sucedido, mas não foi funcional com números maiores forçando uma outra abordagem.

3.8 Análise fatorial

A não utilização do DBSCAN criou a necessidade de utilizar outro método para definição do número de grupo e seus centroides. Assim, ao considerar a possibilidade de forte correlação entre certas variáveis, fez-se a matriz de similaridade entre as variáveis (TABELA 6).

Tabela 6 – Matriz de similaridade

	ano_da_construcao_corrigido	area_construida	area_do_terreno	area_exc	area_ocupada	condominio	construcao	excesso	quantidade_de_esquinas_frentes	quantidade_de_pavimentos	terreno	testada_para_calculo	zona
ano_da_construcao_corrigido	1												
area_construida	0.008313	1											
area_do_terreno	0.105726	0.12631	1										
area_exc	0.004819	0.158131	0.690417	1									
area_ocupada	0.219784	0.119954	0.769134	0.294714	1								
condominio	0.2839	-0.03722	0.18187	0.055018	0.29688	1							
construcao	0.040314	0.84863	0.082666	0.09119	0.097554	0.005606	1						
excesso	-0.00205	0.388574	0.176893	0.389562	0.04519	-0.004033	0.227416	1					
quantidade_de_esquinas_frentes	0.248424	0.023759	0.308524	0.187693	0.356305	0.377082	0.034202	0.010293	1				
quantidade_de_pavimentos	0.346166	-0.009309	0.15653	0.064092	0.329733	0.788462	0.032766	-0.002396	0.363468	1			
terreno	-0.024376	0.742376	0.125705	0.178927	0.08681	-0.042298	0.617491	0.483841	0.020126	-0.024831	1		
testada_para_calculo	0.230385	0.041443	0.473992	0.133929	0.508527	0.361743	0.037241	0.060959	0.410281	0.30428	0.027169	1	
zona	0.182196	-0.016838	0.06068	0.005486	-0.024971	-0.417692	-0.039716	-0.001779	-0.024832	-0.486439	-0.04056	0.030349	1

A matriz apresentou resultados que corroboravam com a possibilidade de correlação, assim optou-se por fazer uma análise fatorial exploratória.

De acordo com Rossoni et al (2016), a análise fatorial é uma técnica multivariada que visa a obtenção de um número mínimo de fatores que contenham, combinados, o máximo de informações contidas nas variáveis originais. Além disso, Rossoni et al (2016) vai dizer que a redução no número de variáveis é desejável quando há intenção de uso de outra técnica de análise variada no conjunto de dados. Assim, como há diversas variáveis nos registros de imóveis e pretende-se submeter o conjunto novamente ao Kmeans, a técnica mostra-se adequada.

Para a análise fatorial, a solução utilizada foi o *software* R, pois esse possui diversas funções relacionadas à análise fatorial, incluindo extração, rotação e determinação do número de fatores, além de ser de código aberto e gratuito. (BEAUJEAN, 2013)

Inicialmente, foi verificado à adequação da amostra com a análise, que foi considerada válida com KMO > 0,5 e teste de Barlett com $p < 0,05$.

A escolha do número de fatores foi pautada pelo maior número de componentes com autovalor maior do que 1, o que resultou em 4 componentes.

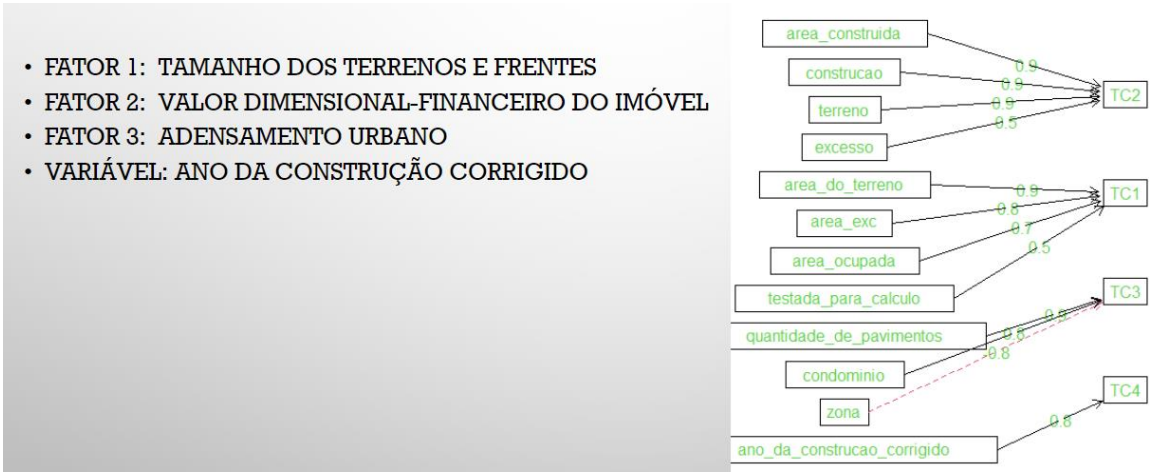
Foi aplicado aos componentes uma rotação oblíqua, permitindo uma certa correlação entre os fatores, com o objetivo de obter interpretações melhores para os resultados, o que ocorre devido

à rotação promover maior concentração dos pesos de certas variáveis em um fator particular. (Rossoni et al, 2016).

Após a análise, as 4 componentes explicavam, conjuntamente, 71% da variação dos dados, valor relevante considerando a grande quantidade de registros. Porém, um dos fatores relacionava o aumente do ano da construção com o aumento da quantidade de esquinas de um imóvel, essa associação é pouco significativa no cenário real, além disso, ao se analisar às comunalidades, ou seja, o quanto os 4 fatores explicam a variação de cada variável, percebeu-se que a comunalidade do número de esquinas de um imóvel era de menos de 50%, então, resolveu-se tirar essa variável da análise.

A retirada do número de esquinas de um imóvel gerou outra análise com 75% da variação dos dados explicada por 4 componentes. O fator que anteriormente relacionava idade do imóvel com esquinas tornou-se um fator de apenas uma variável, o que indica que não há necessidade desse fator, e que se pode utilizar a variável diretamente, gerando 4 fatores relevantes (FIGURA 3).

Figura 3 – Novas variáveis pela análise fatorial



A partir do novo número de variáveis, os centroides iniciais foram criados artificialmente com os valores dos primeiros e terceiros quartis, de cada variável, representando, nessa ordem, valores menores e maiores. Assim, pôde-se assumir que a classificação inicial de clusters fosse binária, em que cada uma das 4 variáveis poderia assumir dois valores qualitativos, gerando 16 arranjos diferentes de centroides. Esse procedimento auxiliou a análise resolvendo o problema do número de agrupamentos inicial e a dificuldade de significação.

Com o novo conjunto, pode-se utilizar diretamente o Kmeans com entrada dos centroides artificiais, gerando resultados significativos.

4. Análise dos dados e resultados

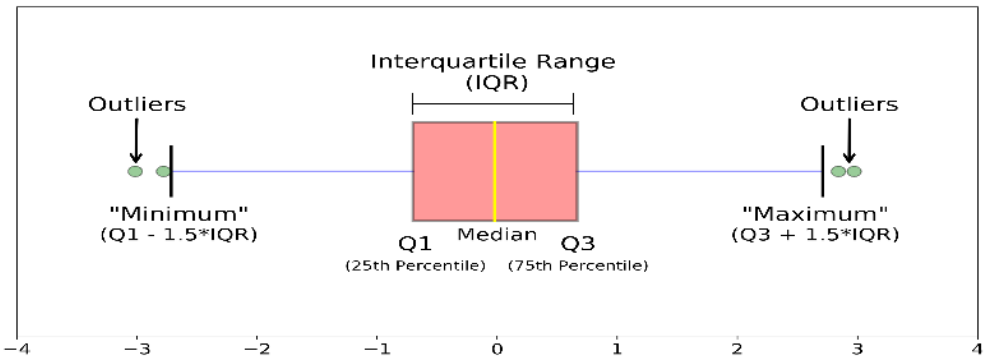
Pela definição de b*ig data* de Laney (2001, apud Victorino, et al, 2017), pode-se dizer que o conjunto de dados selecionados, individualmente, não seria considerado como *big data*, pois não possui variedade de formatos, já que é único e estruturado, e também não é adquirido em grande velocidade, por ser anual, mas a escolha prevalece pelo fato do conjunto ser massivo, com mais de 3 milhões de registros, relacionado com políticas urbanas e, em conjunto com outros dados, faria parte da definição de b*ig data* e teria que ser processado para que se obtivesse informações relevantes. Assim, caso os procedimentos e análises sejam favoráveis nesse conjunto, há possibilidade de estender a pesquisa a outros conjuntos disponibilizados.

A compatibilização dos dados com a plataforma baseou-se em transformações numéricas e escritas, de modo que fosse possível a manipulação daqueles nas variáveis apresentadas (área do terreno, ano da construção, testada, etc.)

O cálculo dos valores venais, apesar da simplicidade da fórmula e do conceito, não necessitando de qualquer análise, mostrou-se complicado no quesito dependência de diversas fontes externas, contidas em legislações, sendo algumas dependentes da procura e leitura de decretos para atualização de valores em tabelas base, caso da tabela de subdivisão da zona, em que há atualização periódica, em porcentagem, por decreto municipal, dos valores por metro quadrado construído.

O agrupamento inicial, feita com um número reduzido de imóveis, CEP único com cerca de 4.000 imóveis, gerou resultados satisfatórios, no sentido de criar agrupamentos pertinentes, com possibilidade de investigação mais detalhada (como um grupo de Imóveis de 3 a 4 andares, grandes e valorizados), além de mostrar os imóveis destoantes das características desse agrupamento após a aplicação de um *boxplot* e verificação de imóveis nos extremos (Figura 4).

Figura 4 – *Boxplot* e *outliers*



Fonte: Galarnyk (2021)

Porém, o método de agrupamento utilizado, ao ser aplicado em um número relativamente maior de dados, aumentou consideravelmente o tempo de processamento necessário na fase do *DBSCAN*, o que tornou inviável o procedimento quando o número de imóveis processados é acima de 10 mil, valor extremamente baixo quando se considera os 3 milhões do total de registros. Além disso, ao tornar aleatório os registros, variando os CEPS, percebia-se uma crescente dificuldade em dar significado aos agrupamentos gerados, assim, após reunião com funcionários da empresa responsável pela plataforma *HPCC*, junto com os autores, o projeto entrou em fase de testes com novos procedimentos para que se pudesse processar mais dados em tempo hábil, e, paralelamente aos testes, discutiu-se a possibilidade de novas abordagens, como a redução da análise para áreas de responsabilidade das Subprefeituras, ou a mudança das variáveis presentes na análise. Por fim, foi descoberto um possível problema com o algoritmo *DBSCAN* o que impossibilitava a continuação de seu uso.

Após a análise fatorial, para diminuição das variáveis e criação direta dos centroides iniciais para o *Kmeans*, sem a utilização do *DBSCAN* como intermediário, foi realizada a análise de clusters com o *Kmeans*. Nessa etapa, o processamento de todos os registros simultaneamente foi possível em tempo hábil, e isso gerou agrupamentos razoáveis em relação à uma primeira análise sobre os imóveis contidos neles, com exceção de 2 dos 16 agrupamentos que, por suas peculiaridades, devem ser investigados para melhor compreensão do significado deles e possível validação do novo procedimento realizado. Essa investigação é o patamar atual da pesquisa, que avançará conforme os resultados das investigações nas próximas semanas.

5. Considerações Finais

Considerando os processos realizados, pode-se verificar que, apesar de promissora, a análise em b*ig data* possui diversas características que a deixam muito complexa. Apesar do conjunto de dados ser massivo, ele é relativamente simples e, ainda assim, foram necessários diversos tipos de manipulações e discussões para que ele se tornasse utilizável.

Primeiro há questões de coleta, que foram fáceis de serem resolvidas, mas não se aplicam a outros conjuntos pesquisados nas fases iniciais, que muitas vezes necessitam da criação de processos de coleta específicos para que se tornem acessíveis, o que poderia complicar em caso de integração de outros conjuntos. Há também questões de estruturação, os registros do IPTU possuem estruturação padrão, mas outros registros governamentais, como os das licitações de compra em órgão públicos, possuem estruturação variável em relação aos diversos setores e abrangências de seus conjuntos. Também há diferença de tempo de atualização de conjuntos de registros, outro fator que atrapalharia a problematização.

Em questões relacionadas ao próprio conjunto escolhido, a quantidade de informações utilizadas era maior do que as 13 variáveis exemplificada, sendo 10 delas diretas e 3 calculadas, contendo cálculos dependentes de tabelas externa de fácil acesso, mas que necessitam de coleta manual para utilização. A escolha das variáveis é um processo subjetivo e depende de diversas discussões envolvendo interessados. Quanto ao processamento, há possibilidade de problemas técnicos e adaptações forçadas por limitações computacionais ou de análise, o que exige mais discussão e pesquisa sobre métodos melhores. Como exemplo, pode-se mostrar a dificuldade da definição da quantidade de grupos a serem formados para facilitar a análise e a questão da diminuição das variáveis para facilitar a significação.

Por fim, é perceptível a quantidade de esforço necessário para análise de um conjunto simples e estável, com modificação lenta e, por isso, pode-se inferir que esse esforço seria ainda maior considerando integração de vários conjuntos de dados governamentais. Porém, a análise é possível e apenas um processo de agrupamento, que ainda assim tem uma complexidade razoável até seu uso efetivo, já mostra informações interessantes sobre as características compartilhadas entre os imóveis da maior cidade do país. Portanto, a utilização dessas análises pode trazer informações úteis para gestão pública e outros interessados, desde que os processos de processamento e análise sejam modificados e pensados de modo a otimizar as etapas necessárias.

REFERÊNCIAS

ALVES, Paulo M. M. R.. O impacto de big data na atividade de inteligência. Revista Brasileira de Inteligência, [S. L.], v. 13, p. 25-44, dez. 2018. Disponível em: https://rbi.enap.gov.br/index.php/RBI/issue/view/14/37. Acesso em: 10 maio 2021.

BEAUJEAN, A. Alexander. Factor Analysis using R. Practical Assessment, Research & Evaluation. [S. L.], p. 1-12. fev. 2013. Disponível em: https://www.researchgate.net/publication/236963001_Factor_Analysis_using_R. Acesso em: 31 maio 2021.

GALARNYK, Michael. Understanding Boxplots. Disponível em: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51. Acesso em: 31 maio 2021.

GEOSAMPA (São Paulo). IPTU_2019. [S. l.], 2019. Disponível em: http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx#. Acesso em: 28 maio 2021.

GONÇALVEZ, A. L.; FARACO, F. M; DE SOUZA, J. A.; TODESCO, J. L.; NUNES, R. C. T. Análise de agrupamentos sobre textos: um estudo dos resumos do banco de teses e dissertações da CAPES. Anais do Congresso Internacional de Conhecimento e Inovação – ciki, [S. l.], v. 1, n. 1, 2018. Disponível em: https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/589. Acesso em: 29 maio. 2021.

HEIL, Jonilson; VOLPI, Neida Maria Patias. Emprego da estatística multivariada como proposta para o cálculo do valor venal e tributação imobiliária. Junho de 2013. educapes.capes.gov.br. Disponível em: http://educapes.capes.gov.br/handle/1884/30480. Acesso em: 28 maio 2021.

MACIEJEWSKI, Mariusz. To do more, better, faster and more cheaply: using big data in public administration. International Review Of Administrative Sciences, [S.L.], v. 83, n. 1, p. 120-135, 9 jul. 2016. SAGE Publications. http://dx.doi.org/10.1177/0020852316640058.

MIDDLETON, Anthony; CHALA, Arjuna. Introduction to HPCC (High Performance Computing Cluster). 2011. Disponivel em: http://cdn.hpccsystems.com/whitepapers/wp introduction HPCC.pdf. Acesso em: 28 maio 2021.

ROSSONI, Luciano; ENGELBERT, Ricardo; BELLEGARD, Ney Luiz. Normal science and its tools: reviewing the effects of factor analysis in management. Revista de Administração, [S.L.], p. 198-211, 2016. Business Department, School of Economics, Business & Accounting USP. http://dx.doi.org/10.5700/rausp1234.

SECRETARIA MUNICIPAL DA FAZENDA DE SÃO PAULO. Cartela de IPTU. 2021. Disponível em: https://web1.sf.prefeitura.sp.gov.br/CartelaIPTU. Acesso em: 28 maio 2021.

VICTORINO, Marcio de Carvalho et al.. Uma proposta de ecossistema de big data para a análise de dados abertos governamentais. Inf. & Soc.:Est., João Pessoa, v.27, n.1, p. 213-230, jan./abr. 2017. Disponível em: < https://www.proquest.com/docview/1894716714/fulltextPDF>. Acesso em: 07 maio 2021.

XU, Lili et al. Massively Scalable Parallel KMeans on the HPCC Systems Platform. In: 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS). IEEE, 2019. p. 1-8.

YU, Shyr-Shen; CHU, Shao-Wei; WANG, Chuin-Mu; CHAN, Yung-Kuan; CHANG, Ting-Cheng. Two improved k-means algorithms. Applied Soft Computing, [S.L.], v. 68, p. 747-755, jul. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.asoc.2017.08.032.