Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogr
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

Otimização da Integração de Topônimos por Similaridade Lexical

Optimization of Integration of Toponyms by Lexical Similarity

Lanna Kallen Parreiras [1], Fredy Sales Ribeiro [2] e Vagner Braga Nunes Coelho [3]

1 Universidade Federal de Minas Gerais, Graduanda em Engenharia Civil, Belo Horizonte, Brasil. lanpna@hotmail.com.
ORCID: https://orcid.org/0000-0002-5983-6991
2 Universidade Federal de Minas Gerais, Graduando em Engenharia Civil, Belo Horizonte, Brasil. fredysales1@gmail.com.
ORCID: https://orcid.org/0000-0002-4255-6790
3 Universidade Federal de Minas Gerais, Departamento de Cartografia, Belo Horizonte, Brasil. vagnercoelho@hotmail.com.
ORCID: https://orcid.org/0000-0002-7512-2024

Recebido: 12.2021 | Aceito: 05.2022

Resumo: Feições identificáveis do mundo real são, por intermédio de funções de mapeamento, instanciadas em um Banco de Dados Geográfico (BDG) como representações dessa realidade. Essas representações são individualizadas pelos atributos especificadores da classe mapeada. Entre esses atributos estão pelo menos uma geometria e um nome identificador (topônimo) associado à chave primária. No entanto, diferentes produtores de dados interpretam a realidade com pequenas discrepâncias, tornando algumas representações de características mapeadas semelhantes, mas não idênticas. Em particular, os topônimos têm pequenas diferenças resultantes de modificações ao longo dos anos, da forma como são soletrados ou, também, devido a erros humanos no registro dos dados. Portanto, ao tentar integrar diferentes BDGs, por meio de topônimos, eles não favorecem um pareamento total, uma vez que os registros não são identificados como sendo representativos da mesma realidade. No caso específico da classe toponímia, isso ocorre principalmente devido a erros de digitação decorrentes do processo de inserção de dados, especialmente pela inversão no posicionamento dos caracteres dentro da palavra. Nesta pesquisa, foi desenvolvida uma melhoria no Coeficiente de Dados e comparada com o método original aplicado em três BDGs distintos. A análise foi baseada nas frequências de caracteres e bigramas existentes nessas bases. A melhoria proposta baseou-se na hipótese de que bigramas invertidos, como 'αβ' e 'βα', podem, segundo certos critérios, ser admitidos como semelhantes. A análise identificou os caracteres mais comuns e os bigramas mais frequentes nas bases, cuja associação com uma análise da distância normalizada em um teclado padrão, permitiu a identificação de uma série de pares de bigramas considerados semelhantes. Essa proposta permitiu um aumento médio de 0,58% no total de instâncias pareadas nos BDGs testados.
Palavras-chave: Banco de Dados Geográfico. Similaridade lexical. Toponímia. Teclado.

Abstract: Real-world identifiable features are, through mapping functions, instantiated in a Geographic Database (GD) as representations of this reality. These representations are individualized by the specifier attributes of the mapped class. Among these attributes are at least one geometry and an identifier name (toponym) associated with the primary key. However, different data producers interpret reality with slight discrepancies, making some representations of mapped features similar but not identical. In particular, toponyms have small differences resulting from modifications over the years, the way they are spelled or, also, due to human errors in the recording of the data. Therefore, when trying to integrate different GDs, through toponyms, they do not favor a total pairing, since the records are not identified as being the same reality. In the particular case of the toponymy class, this occurs mainly due to typos arising from the data insertion process, especially by inversion in the positioning of the characters within the word. In this research, an improvement in the Dice Coefficient was developed and compared with the original method applied in three distinct GDs. The analysis was based on the frequencies of characters and bigrams existing in those bases. The proposed improvement was based on the hypothesis that inverted bigrams, like 'αβ' and 'βα', may, according to certain criteria, be admitted as similar. The analysis identified the most common characters and the most frequent bigrams in the bases whose association with a distance analysis on a standard keyboard allowed the identification of a series of pairs of bigrams to be considered similar. This proposal allowed an average increase of 0.58% in the total paired instances in the GDs tested.
Keywords: Geographic Database. Lexical similarity. Toponymy. Keyboard.

1 INTRODUÇÃO

A elaboração de um produto cartográfico clássico, segundo Pivac e Roić (2020), pressupõe uma série de atividades com o propósito de representar o mundo em um mapa. Dentre os diversos processos de produção, há o que se denomina reambulação (SILVA, 2020). Essa atividade é a responsável pela correta identificação de nomes que adequadamente identificam as feições no terreno, cuja informação apresentada no mapa recebe a denominação de toponímia (DIEDRICH; MACHADO, 2020). Com o advento da computação e, em particular dos Bancos de Dados, houve uma constante automação do processo ao ponto de se construir virtualmente uma Base Cartográfica Contínua, devidamente armazenada em um Banco de Dados Geográficos (BDG) (ELMASRI; NAVATHE, 2016). Esta base contínua é composta por diversos atributos, dos quais destacam-se a geometria e os identificadores das feições que servem como chave primária (topônimos).

As feições existentes no mundo podem ser instanciadas em Bancos de Dados Geográficos (BDGs) a partir de uma função de mapeamento específica. Entretanto, os diferentes produtores de dados, ou seja, os responsáveis pela inserção desses dados, podem representá-los de maneiras diferentes, fazendo com que existam registros similares, mas não idênticos (COELHO, 2010). Com isso, a partir dos modelos individuais dos vários produtores de dados, obtém-se uma multiplicidade de representações espaciais, caracterização por meio de instâncias distintas nos atributos e topônimos registrados com semelhança, porém distintos (FIELDING; VERD, 2021). Tais topônimos instanciam o BDG por meio de *strings* (cadeia de caracteres) que nomeiam/identificam a feição (ALDANA-BOBADILLA et al., 2020).

Ao se analisar as *strings* inseridas em um BDG, deve-se considerar possíveis equívocos provenientes do processo de inserção dos topônimos associados às feições, advindos de fatores como erros de grafia, convenções de escrita e duplicidade de representações. Com isso, instâncias referentes a uma mesma feição passam a ser consideradas distintas ou associadas a mais de um elemento, fazendo com que o BDG tenha informações ambíguas e registros duplicados (MACEDO et al., 2020).

Desse modo, ao integrar ou parear BDGs distintos, deve-se considerar relações de semelhança entre os topônimos, por meio da aplicação de uma função de similaridade (MACEDO et al., 2020). Nesta pesquisa, utilizou-se a métrica de *string* baseada no Coeficiente de Dice associado a um limiar de aceitação (MACEDO, 2020). Destarte, foi feita a análise dos bigramas eventualmente construídos para cada *string*, particularmente sua frequência de ocorrência em uma biblioteca de palavras e as relações métricas entre os caracteres dispostos em um teclado, visando inferir sobre os possíveis erros provenientes do processo de digitação, tornando-se fundamental para o aprimoramento da integração de BDGs distintos.

Nesse sentido, o presente artigo tem como objetivo aprimorar a integração lexical entre diferentes BDGs por meio de uma análise da frequência de caracteres e de bigramas e da distância entre os caracteres de um teclado, visando inferir com maior qualidade quais topônimos, instanciados individualmente, podem ser considerados similares.

2 SIMILARIDADE

Embora existam várias formas de se obter uma integração, conforme Bento, Zouaq e Gagnon (2020); Shoaib, Daud e Amjad (2020); Subathra e Umarani (2021); Zambon et al. (2020), nesta pesquisa entende-se similaridade de toponímias como a semelhança entre *strings* obtida por meio da análise comparativa de suas cadeias de caracteres, da frequência de bigramas e de caracteres e pela disposição espacial das letras em um teclado. Para isso, aplica-se uma função de similaridade, baseada no Coeficiente de Dice (C_D) (DICE, 1945), responsável por quantificar o quanto uma *string* é semelhante à outra.

Nesse viés, o Coeficiente de Dice (C_D) baseia-se na contagem de bigramas construídos a partir de caracteres adjacentes (DICE, 1945). Desse modo, dadas duas *strings* quaisquer φ_1 e φ_2, o C_D entre elas é dado pela Eq. (1).

$$C_D(\varphi_1, \varphi_2) = \frac{2 \,.\, |n_1 \cap n_2|}{|n_1| + |n_2|} \tag{1}$$

em que n_1e n_2 referem-se aos bigramas das representações φ_1 e φ_2, respectivamente. Por sua vez, $n_1 \cap n_2$ representa o conjunto dos bigramas comuns entre as duas *strings*, enquanto $|n_1|, |n_2|$ e $|n_1 \cap n_2|$ representam a cardinalidade dos bigramas referentes às representações φ_1, φ_2 e os em comum entre elas, respectivamente.

Sejam as *strings* abaixo:

- φ_1: 'a cinco';
- φ_2: 'cinco a'.

A partir das *strings*, obtém-se os seguintes conjuntos dos bigramas com respectiva cardinalidade:

- n_1:{'a_', '_c', 'ci', 'in', 'nc', 'co'} $\implies |n_1| = 6$
- n_2:{'ci', 'in', 'nc', 'co', 'o_', '_a'} $\implies |n_2| = 6$
- $n_1 \cap n_2$:{'ci', 'in', 'nc', 'co'} $\implies |n_1 \cap n_2| = 4$

Consequentemente, tem-se (Eq. (2)):

$$C_D(\varphi_1, \varphi_2) = \frac{2\ .\ 4}{6+6} = \frac{8}{12} = 0{,}67 \tag{2}$$

Com base no resultado obtido acima, percebe-se que, embora as *strings* sejam muito similares, o valor obtido no Coeficiente de Dice é muito baixo, o que permite inferir que os dois textos não possuem um valor para a similaridade que permita inferir se tratarem de uma mesma toponímia (MACEDO et al., 2020).

3 METODOLOGIA

Diante da problemática de inversão de caracteres no processo de inserção de dados em um BDG, cabe analisar quando bigramas distintos podem ser considerados similares. Para isso, seguem-se os seguintes passos:

- levantamento de dados;
- análise da frequência de caracteres e sua divisão em *clusters*;
- determinação dos bigramas de referência e os invertidos;
- análise da frequência dos bigramas de referência e os invertidos e suas divisões em *clusters*;
- aplicação do Coeficiente de Dice aprimorado e comparação com Dice original.

3.1 Levantamento de dados

Para realizar a análise dos dados, deve-se, primeiramente, ter disponíveis BDGs distintos e, preferencialmente, com cardinalidades diferentes, a fim de que se possa obter um resultado mais preciso e válido para qualquer BDG no Brasil.

Para esta pesquisa, utilizou-se os BDGs dos municípios de Contagem, Belo Horizonte e São Paulo, sendo o primeiro disponibilizado pela Secretaria de Planejamento e os dois últimos disponíveis em domínio público nos sites das respectivas prefeituras (CNEFE, 2021).

3.2 Análise da frequência de caracteres e sua divisão em *clusters*

A análise da frequência de caracteres permite determinar as regiões, no teclado, de maior influência no processo de inserção dos dados em um BDG, o que permite dividir os caracteres em *clusters*.

Para isso, desenvolveu-se algoritmos implementados na linguagem *Python*, de modo a quantificar a frequência dos caracteres em cada um dos BDGs utilizando o método de Jenks (JENKS, 1967). O espaço amostral foi dividido em três *clusters* em função da frequência: baixa, média e alta.

3.3 Determinação dos bigramas de referência e os invertidos

Para identificar bigramas invertidos que podem ser considerados similares, foi necessário determinar,

dados os bigramas invertidos $'\alpha\beta'$ e $'\beta\alpha'$, qual será denominado bigrama de referência e qual será o invertido. Para isso, utilizou-se o padrão de codificação *UCS Transformation Format* 8 (UTF-8), Keiser e Lemire (2021), de cada caractere.

3.4 Análise da frequência dos bigramas de referência e os invertidos e suas divisões em *clusters*

Identificados os bigramas de referência e os invertidos, passou-se a determinar a frequência de cada um deles, mediante o mesmo algoritmo desenvolvido, dividindo o espaço amostral em *clusters*, assim como realizado com os caracteres.

3.5 Aplicação do Coeficiente de Dice aprimorado e comparação com Dice original

Após analisar os bigramas invertidos que podem ser considerados semelhantes, deve-se otimizar o Coeficiente de Dice, a fim de que ele possa considerar os bigramas potencialmente invertidos, contribuindo para a otimização da integração.

Desse modo, os resultados foram avaliados por meio de uma comparação entre o método de Dice original e o método aprimorado, sendo possível quantificar e qualificar a eficiência da otimização do processo.

4 ESPAÇO AMOSTRAL

A fim de basear as análises, utilizou-se três BDGs contendo endereços de municípios distintos: Contagem (MG), Belo Horizonte (MG) e São Paulo (SP). A escolha do espaço amostral deu-se pelo acesso aos BDGs e pela diferença de cardinalidade entre eles (Tabela 1), sendo um deles o maior BDG do país.

Tabela 1 – Cardinalidade dos BDGs analisados.

Município	**Instâncias no BDG**
Contagem	4.681
Belo Horizonte	20.248
São Paulo	65.522

Elaboração: Os autores (2021).

Segundo Barbosa (2004), os topônimos são formados por um atributo genérico que denomina o acidente geográfico ou ação antrópica (rua, praça, alameda, rio, etc) e por um atributo particular que singulariza a feição (SANTOS, 2008). Nesse sentido, os BDGs analisados possuem redes viárias compostas por diversos atributos, como tipo, logradouro e bairro. No entanto, para a análise em questão, utilizou-se apenas um atributo da classe toponímia, a parte da *string* que nomeia e particulariza um determinado local, o logradouro.

5 ANÁLISE DE CARACTERES

5.1 Frequência de caracteres

Para determinar a frequência de cada caractere em um BDG, elaborou-se um algoritmo implementado na linguagem *Python* (Algoritmo 1), a fim de analisar os caracteres mais utilizados no processo de inserção de dados em um BDG e, consequentemente, sendo os mais suscetíveis a erros.

Algoritmo 1 – Pseudocódigo da frequência de caracteres.

Input: B_1: lista de logradouros de um BDG.
Output: $frequencia$: dicionário contendo os caracteres e suas frequências.
begin
    $frequencia = \{\}$
    for $logradouro$ **in** B_1 **do**
        for $caracter$ **in** $logradouro$ **do**
            if $caracter$ **in** $frequencia.keys()$ **then**
                $frequencia[caracter] \leftarrow +1$
            end
            else $frequencia[caracter] \leftarrow 1$
        end
    end
end
return $frequencia$

Elaboração: Os autores (2021).

O Algoritmo 1 lê um conjunto de dados em formato *string* e realiza a contagem da frequência dos caracteres a partir dos logradouros de um BDG analisado. Para essa contagem, despreza-se a existência de sinais diacríticos, caracteres especiais e a diferença entre caracteres maiúsculos e minúsculos. A Tabela 2 mostra o total de caracteres em cada um dos BDGs em estudo.

Tabela 2 – Quantidade de caracteres.

Município	Quantidade de caracteres
Contagem	64.291
Belo Horizonte	295.450
São Paulo	840.055

Elaboração: Os autores (2021).

5.2 Classificação em *clusters*

Para a análise da distribuição de frequência dos caracteres, dividiu-se estes em *clusters* com base no método de Jenks (*natural breaks*). Nesse sentido, os caracteres foram divididos em três classes: baixa, média e alta frequência, identificadas como *clusters* 1, 2 e 3, respectivamente. O método de otimização de Jenks foi utilizado para dividir o conjunto de dados de modo que eles possam ser agrupados (*clustering*) no melhor arranjo, reduzindo a variância dentro das classes e maximizando a variância entre elas. Assim, a Tabela 3 apresenta os limites dos *clusters* para os caracteres.

Tabela 3 – Limites superiores e inferiores dos *clusters* dos caracteres.

Município	Limites					
	***Cluster* 1**		***Cluster* 2**		***Cluster* 3**	
	Inferior	**Superior**	**Inferior**	**Superior**	**Inferior**	**Superior**
Contagem	1	1.069	1.069	4.158	4.158	8.448
Belo Horizonte	2	4.643	4.643	18.241	18.241	38.665
São Paulo	10	10.078	10.078	43.010	43.010	122.487

Elaboração: Os autores (2021).

6 ANÁLISE DE BIGRAMAS

A análise dos bigramas tem como objetivo identificar a possível inversão de caracteres durante o processo de inserção de dados em um BDG. Nesse sentido, cabe avaliar bigramas invertidos que podem ser considerados iguais. Para isso, utilizou-se um algoritmo implementado na linguagem *Python* (Algoritmo 2),

cujo objetivo é analisar a frequência dos bigramas.

Algoritmo 2 – Pseudocódigo da frequência de bigramas.

```
Input: B_1: lista de logradouros de um BDG.
Output: bigramas: dicionário contendo os bigramas e suas frequências.
begin
    freq_bigramas = {}
    for logradouro in B_1 do
        big ← BIG(logradouro) /* BIG(logradouro) = bigramas gerados a partir da string lida. */
        for bigrama in big do
            if bigrama in freq_bigramas.keys() then
                freq_bigramas[bigrama] ← +1
            end
            else freq_bigramas[bigrama] ← 1
        end
    end
end
return freq_bigramas
```

Elaboração: Os autores (2021).

O Algoritmo 2 recebe uma lista de logradouros, em formato *string*, e realiza a contagem dos bigramas existentes nessas cadeias de caracteres, armazenando a informação em um dicionário *Python*, cuja chave é o próprio bigrama e o valor da sua respectiva frequência.

6.1 Bigramas invertidos

Bigramas com caracteres invertidos, por exemplo ‘AR’ e ‘RA’, foram ordenados com base no padrão de codificação *UCS Transformation Format* 8 (UTF-8) de cada caractere - codificação binária universal para comprimento de caractere - ao se associar um número inteiro correspondente a cada caractere (função ORD).

Assim, sejam dois bigramas invertidos ‘$\alpha\beta$’ e ‘$\beta\alpha$’. Dessa forma, se ORD(α) > ORD(β), então ‘$\alpha\beta$’ é considerado o bigrama considerado e ‘$\beta\alpha$’ o bigrama invertido. Caso contrário, tem-se ‘$\beta\alpha$’ como bigrama de referência enquanto ‘$\alpha\beta$’ passa a ser invertido.

Considere os bigramas ‘AT’ e ‘TA’. Assim, sabendo que ORD(‘A’) = 97 e ORD(‘T’) = 116, ORD(‘A’) > ORD(‘T’) e, consequentemente, considera-se ‘AT’ como o bigrama de referência e ‘TA’ como o bigrama invertido.

6.2 Classificação em *clusters*

Do mesmo modo como aplicado aos caracteres, foi desenvolvido a classificação dos bigramas em *clusters*. Assim, as Tabelas 4 e 5 apresentam os limites dos *clusters* para os bigramas de referência e os invertidos.

Tabela 4 – Limites superiores e inferiores dos *clusters* dos bigramas de referência.

Município	**Limites**					
	***Cluster* 1**		***Cluster* 2**		***Cluster* 3**	
	Inferior	**Superior**	**Inferior**	**Superior**	**Inferior**	**Superior**
Contagem	0	201	201	629	629	1.472
Belo Horizonte	0	1.044	1.044	3.324	3.324	8.296
São Paulo	0	2.115	2.115	7.620	7.620	22.594

Elaboração: Os autores (2021).

Tabela 5 – Limites superiores e inferiores dos *clusters* dos bigramas invertidos.

Município	Limites					
	***Cluster* 1**		***Cluster* 2**		***Cluster* 3**	
	Inferior	**Superior**	**Inferior**	**Superior**	**Inferior**	**Superior**
Contagem	0	121	121	434	434	1.073
Belo Horizonte	0	503	503	1.648	1.648	5.267
São Paulo	0	2.002	2.002	6.662	6.662	15.458

Elaboração: Os autores (2021).

A distribuição de *clusters* dos bigramas de referência e invertido ocorrem de modo que se tem um arranjo simples da forma (Eq. 3):

$$A_{3,2} = \frac{3!}{(3-2)!} = 6 \tag{3}$$

Logo, tem-se os seguintes arranjos para os *clusters*: {1,1}; {1,2}; {1,3}; {2,2}; {2,3}; {3,3}.

6.3 Bigramas semelhantes

Pressupostos os erros inerentes ao processo de inserção de dados nos BDGs pelos produtores, pode-se presumir que a inversão de bigramas é um potencial causador de divergências nesses dados, isso porque um operador pode, equivocadamente, inverter os caracteres dos bigramas ao realizar o processo de digitação dos dados. Por exemplo, ao invés de representar a instância "Rua Administradores" este pode inserir a *string* "Rua Administardores" erroneamente, invertendo os bigramas 'RA' por 'AR'. Essa inversão cria uma nova instância no BDG cuja representação no mundo real possivelmente não existe, ou pode gerar duplicidade de representações.

A fim de otimizar o Coeficiente de Dice para o pareamento de instâncias de BDGs, deve-se assumir a semelhança entre os bigramas de referência e os bigramas invertidos que apresentam maior potencial de erros. Para isso, considera-se os bigramas com maior frequência nos BDGs, visto a grande repetição desses no processo de inserção. Portanto, foram considerados os bigramas que se enquadram nos seguintes arranjos de *clusters*: {3,3} e {2,3}, ou seja, bigramas de referência e invertidos que se enquadram nesses arranjos podem ser considerados semelhantes, uma vez que, por terem média e alta frequência, podem ter sido invertidos erroneamente.

Já os arranjos: {1,1}, {1,2}, {1,3} e {2,1}, foram, automaticamente, classificados como não semelhantes e não passíveis de pareamento, uma vez que pelo menos um dos bigramas pertence a uma classe de baixa frequência. Por outro lado, os bigramas que apresentaram o arranjo {2,2}, passaram por uma inspeção cuja análise levou em consideração a distância entre as teclas dos caracteres em um teclado ABNT-2 (Figura 1). Para essa análise, arbitrou-se a geometria quadrada para as teclas, tomando como referência a distância normalizada de 1 unidade do centro de uma tecla a outra. Por exemplo, considerar-se-á a distância entre as teclas A e S como 1, entre A e D como 2, entre A e W como 1,2 e, assim, sucessivamente.

Figura 1 – Teclado modelo ABNT-2.

Elaboração: Os autores (2021).

7 MÉTODO DE DICE APRIMORADO

Identificados os bigramas de maior potencial de inversão no BDG analisado, deve-se otimizar o Coeficiente de Dice de modo que este inclua em sua comparação o espelhamento desses bigramas. Assim, o Algoritmo 3 apresenta o pseudocódigo desse aprimoramento do método.

Algoritmo 3 – Pseudocódigo do aprimoramento da métrica de Dice.

Input: T_1, T_2 *e* B: BDG's de comparação e lista de bigramas com potencial inversão.
Output: $resposta$: Matriz contendo as tuplas construídas a partir dos pares id_1e id_2 com os respectivos valores de similaridade otimizado.

```
begin
    resposta = [ ]
    n ← |T1|
    m ← |T2|
    for id1 to T1 do
        bigramas_id1 = BIG(id1)   /* BIG(id1) = bigramas gerados a partir da string lida. */
        for big to bigramas_id1 do
            if ordenar(big) in B do bigramas_id1[big] = ordenar(big)
        end
        for id2 to T2 do
            bigramas_id2 = BIG(id2)
            for big to bigramas_id2 do
                if ordenar(big) in B do bigramas_id2[big] = ordenar(big)
            end
            max ← 0
            if |T1[id1]| and |T2[id2]| ≠ 1 then
                valor ← Dice(bigramas_id1, bigramas_id2)
                if valor ≥ max then max ← valor
            end
            resposta.append(id1, id2, max)
        end
    end
end
return resposta
```

Elaboração: Os autores (2021).

O Algoritmo 3 recebe como parâmetro de comparação dois BDGs com suas instâncias armazenadas em listas e uma lista contendo os bigramas com potenciais inversões. Dessa forma, realiza-se a comparação entre essas instâncias através da métrica de Dice, registrando e armazenando o valor de similaridade entre estes. Porém, conforme Algoritmo 3, toda vez que for identificado um bigrama potencialmente invertido na cadeia de caracteres da *string* do logradouro analisado, esse será tomado como semelhante ao seu invertido, ou seja, '$\alpha\beta$' = '$\beta\alpha$'. Dessa forma, ao realizar esse espelhamento de bigramas, tem-se um aumento no valor do coeficiente de Dice para os logradouros que contiverem os bigramas potencialmente invertidos.

Assim, considere os exemplos:

- φ_1: 'a cinco';
- φ_2: 'cinco a';
- φ_3: 'administradores';
- φ_4: 'administardores'.

Aplicando o Coeficiente de Dice nas *strings* acima, encontra-se $C_D(\varphi_1, \varphi_2) = 0{,}67$ e com o método de Dice otimizado ($C_{D'}$) o valor encontrado é de $C_{D'}(\varphi_1, \varphi_2) = 0{,}83$. Já para φ_3 e φ_4, tem-se $C_D(\varphi_3, \varphi_4) = 0{,}78$ e $C_{D'}(\varphi_3, \varphi_4) = 0{,}85$. Na primeira situação, considerando as *strings* φ_1 e φ_2 com o método de Dice tradicional o pareamento entre as instâncias não seria considerado com um limiar de aceitação de 0,80; já com o método aprimorado, devido ao aumento de 25,76%, ele seria identificado como instâncias semelhantes referentes à mesma feição. Já nas *strings* φ_3 e φ_4, teve-se um aumento de 8,97% na similaridade, além de garantir a efetividade de um limiar de aceitação maior que 0,80 (MACEDO et al., 2020). Embora pareça um aumento insignificante, a depender do limiar de aceitação utilizado, esse método acarreta no pareamento de

novas instâncias, contribuindo para melhorar a quantidade e qualidade dos registros tidos como semelhantes, o que permite observar a validade do método de Dice otimizado.

8 RESULTADOS E DISCUSSÕES

De modo a obter os parâmetros necessários para a validação dos bigramas invertidos que podem ser considerados semelhantes, a princípio, deve-se analisar a frequência dos caracteres em cada um dos três BDGs analisados, conforme mostrado na Tabela 6.

Tabela 6 – Frequência de caracteres. *A representação ‘ ’ identifica a *string* referente à barra de espaço.

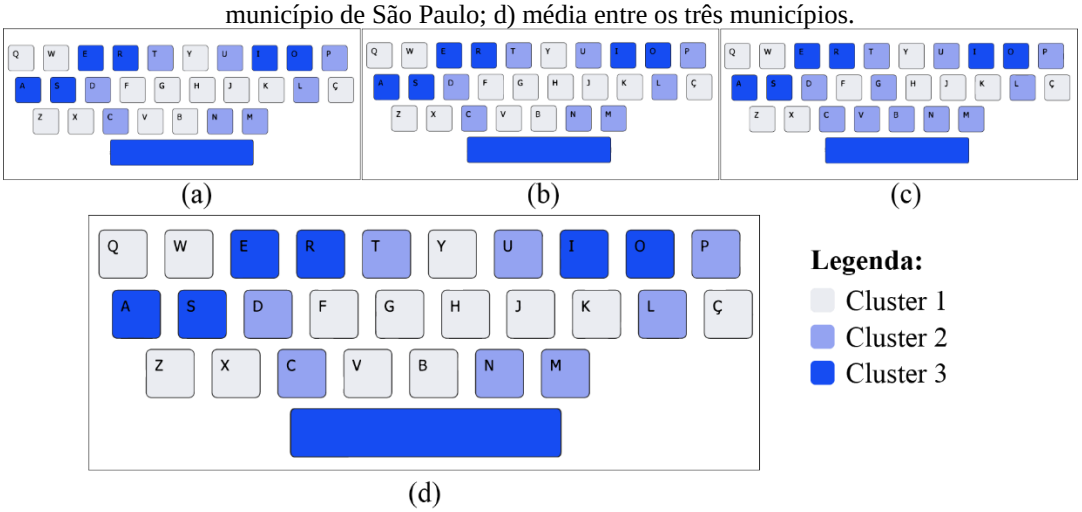
Caractere	Contagem (%)	Belo Horizonte (%)	São Paulo (%)	Média (%)
A	13,14	13,09	14,58	13,6
O	9,71	10,02	9,64	9,79
E	8,83	9,38	8,85	9,02
_*	8,45	9,11	7,85	8,47
I	8,07	7,66	8,4	8,04
R	7,21	7,36	7,97	7,51
S	6,47	6,17	5,12	5,92
N	5,45	5,31	4,68	5,15
D	4,23	4,41	4,01	4,22
T	4,23	4	3,23	3,82
C	3,5	3,24	3,02	3,25
L	3,26	4,26	4,77	4,1
U	2,9	2,66	2,76	2,77
M	2,81	3,07	2,92	2,93
P	1,66	1,57	2,01	1,75
V	1,55	1,54	1,2	1,43
G	1,41	1,39	2,15	1,65
B	1,32	1,36	2,76	1,81

Elaboração: Os autores (2021).

A análise da Tabela 6 permite observar que a frequência dos caracteres entre os três BDGs em estudo é análoga, uma vez que se trata das letras mais usuais na Língua Portuguesa, criando parâmetros para a otimização entre as instâncias.

Além disso, estabelecidos os limites inferiores e superiores de cada um dos *clusters*, desenvolveu-se o mapa de calor (Figura 2) dos três BDGs, em que é possível analisar as regiões de maior frequência. Por meio da Figura 2, observa-se a semelhança entre eles, o que permite afirmar que o volume de dados não altera, significativamente, a frequência média dos caracteres.

Figura 2 – Mapa de calor da frequência de caracteres: a) município de Contagem; b) município de Belo Horizonte; c) município de São Paulo; d) média entre os três municípios.



Elaboração: Os autores (2021).

Com a determinação dos limites inferiores e superiores de cada um dos *clusters*, acrescenta-se à Tabela 6 o *cluster* correspondente a cada um dos caracteres em cada um dos municípios, como é mostrado na Tabela 7.

Tabela 7 – Frequência e divisão em *clusters* dos caracteres. *A representação '_' identifica a *string* referente à barra de espaço.

Caractere	Contagem (%)	Cluster	Belo Horizonte (%)	Cluster	São Paulo (%)	Cluster	Média (%)
A	13,14	3	13,09	3	14,58	3	13,6
O	9,71	3	10,02	3	9,64	3	9,79
E	8,83	3	9,38	3	8,85	3	9,02
_*	8,45	3	9,11	3	7,85	3	8,47
I	8,07	3	7,66	3	8,4	3	8,04
R	7,21	3	7,36	3	7,97	3	7,51
S	6,47	3	6,17	3	5,12	3	5,92
N	5,45	2	5,31	2	4,68	2	5,15
D	4,23	2	4,41	2	4,01	2	4,22
T	4,23	2	4	2	3,23	2	3,82
C	3,5	2	3,24	2	3,02	2	3,25
L	3,26	2	4,26	2	4,77	2	4,1
U	2,9	2	2,66	2	2,76	2	2,77
M	2,81	2	3,07	2	2,92	2	2,93
P	1,66	2	1,57	2	2,01	2	1,75
V	1,55	1	1,54	1	1,2	2	1,43
G	1,41	1	1,39	1	2,15	2	1,65
B	1,32	1	1,36	1	2,76	2	1,81

Elaboração: Os autores (2021).

Outro parâmetro importante para a validação da otimização do método, é a determinação dos *clusters* para os bigramas. Assim, as Tabela 8 e 9 identificam os bigramas de referência pertencentes aos *clusters* 3 e 2 (de maior probabilidade de pareamento) e seus respectivos bigramas invertidos. Vale afirmar que os bigramas presentes no *cluster* 1 serão eliminados da análise, uma vez que possuem baixa frequência dentro do BDG, sendo também baixa a probabilidade de se obter pareamentos confiáveis.

Tabela 8 – Frequência e *cluster* dos bigramas de referência.

Bigrama de Referência	Contagem		Belo Horizonte		São Paulo	
	Cluster	Frequência	Cluster	Frequência	Cluster	Frequência
O_	3	1472	3	8296	3	22594
A_	3	1303	3	5479	3	11372
E_	3	1134	3	5333	3	12834
AR	3	1071	3	5291	3	14954
AN	3	964	3	4449	3	14685
IN	3	907	2	3188	2	5479
OS	3	892	3	4037	3	8851
ES	3	836	3	3902	3	9637
NT	3	831	3	3693	2	6522
DO	3	806	3	4425	3	7620
DE	3	756	3	3720	3	11340
AS	3	725	2	2406	2	6611
S_	3	699	3	3324	3	8649
IR	3	629	2	2741	3	10166
AL	2	626	2	3000	3	12340
OR	2	625	2	2854	3	7960
EI	2	620	2	2724	3	9553
CO	2	580	2	2006	2	5750
EN	2	565	2	3205	2	7300
IO	2	537	2	2343	3	9160
ER	2	515	2	2668	3	10493
IS	2	476	2	1875	2	4668

continua

conclusão

Bigrama de Referência	Contagem		Belo Horizonte		São Paulo	
	Cluster	Frequência	Cluster	Frequência	Cluster	Frequência
EL	2	441	2	2943	3	8140
JO	2	431	2	1644	2	5015
NO	2	402	2	1931	2	4119
CI	2	367	2	1430	2	2227
AO	2	332	2	1716	2	3377
AD	2	330	2	1513	2	4353
IT	2	329	2	1236	2	3032
QU	2	323	2	1332	1	1999
LO	2	317	2	1904	2	3446
ST	2	301	2	1687	2	3979
AT	2	278	2	1109	2	3461
IM	2	269	2	1166	2	4617
L_	2	267	2	1378	2	3781
GU	2	263	1	956	2	3863
IL	2	253	2	2006	2	3934
ET	2	248	1	791	1	1306
AM	2	241	2	1174	2	3353
AC	2	217	1	1012	2	4079
DI	2	217	2	1044	2	4064
R_	2	210	2	1343	2	2375
HO	2	201	1	928	2	2521

Elaboração: Os autores (2021).

Tabela 9 – Frequência e *cluster* dos bigramas invertidos.

Bigrama Invertido	Contagem		Belo Horizonte		São Paulo	
	Cluster	Frequência	Cluster	Frequência	Cluster	Frequência
_O	1	105	2	549	1	1353
_A	2	364	3	2186	2	5830
_E	2	285	2	1505	1	945
RA	3	1073	3	5277	3	15458
NA	2	389	3	2181	2	6317
NI	2	361	2	1299	2	4183
SO	2	214	2	1488	2	2465
SE	3	585	3	2146	2	5146
TN	1	0	1	0	1	0
OD	1	65	1	477	1	1129
ED	2	138	2	1112	1	1036
SA	3	434	3	2203	2	5876
_S	2	371	3	2016	2	3709
RI	3	857	3	3512	3	12345
LA	2	211	2	1259	2	5228
RO	3	693	3	3258	3	9460
IE	1	47	1	326	1	1008
OC	2	131	1	395	1	826
NE	2	250	2	1284	2	4600
OI	2	219	1	495	1	298
RE	3	790	3	3239	3	7314
SI	2	197	2	1152	2	3066
LE	2	198	2	952	2	3127
OJ	1	2	1	3	1	905
ON	3	621	3	2541	2	5131
IC	2	171	2	880	2	2640
OA	2	227	2	837	2	2529
DA	3	607	3	2373	3	7207
TI	2	329	2	1362	2	4704
UQ	1	8	1	28	1	40
OL	2	190	2	973	2	2364

continua

conclusão

Bigrama Invertido	Contagem		Belo Horizonte		São Paulo	
	Cluster	Frequência	Cluster	Frequência	Cluster	Frequência
TS	1	5	1	50	1	0
TA	3	644	3	2918	2	5794
MI	2	139	2	1289	2	2085
_L	2	205	2	997	2	4858
UG	1	49	1	177	1	368
LI	2	371	3	2247	3	9105
TE	3	512	3	2104	2	4745
MA	3	706	3	3134	3	10062
CA	3	730	3	2965	3	10607
ID	2	139	2	746	1	1187
_R	2	190	2	1176	2	5553
OH	1		1	11	1	0

Elaboração: Os autores (2021).

As Tabelas 8 e 9 permitem avaliar que essa distribuição de *clusters* apresenta significativa semelhança nos três BDGs analisados, destacando maior semelhança nos bigramas de alta frequência, como o bigrama ‘AR’ e ‘RA’, cujo arranjo de classes {3,3} é encontrada em todos os municípios, o que permite inferir que existe uma alta probabilidade de esses bigramas estarem invertidos por erros no processo de digitação dos dados.

Para a análise dos bigramas pertencentes ao arranjo de *clusters* {2,2}, levou-se em consideração a distância normalizada entre os caracteres do teclado. Assim, a Tabela 10 apresenta essas distâncias para os BDGs de Contagem, Belo Horizonte e São Paulo.

Tabela 10 – Distância entre as teclas dos bigramas no arranjo {2,2}.

Contagem		Belo Horizonte		São Paulo	
Bigrama de Referência	Distância	Bigrama de Referência	Distância	Bigrama de Referência	Distância
AL	8,0	AL	8,0	AP	8,8
AO	7,8	AO	7,8	AO	7,8
EL	6,3	EL	6,3	IS	5,8
IS	5,8	IS	5,8	EM	5,2
CO	5,6	DI	4,8	CI	4,7
DI	4,8	CI	4,7	EN	4,3
CI	4,7	EN	4,3	AT	3,9
EN	4,3	R_	3,0	GO	3,9
IT	3,0	IT	3,0	R_	3,0
R_	3,0	L_	2,5	IT	3,0
L_	2,5	IN	2,3	NO	3,0
MO	2,3	IM	2,0	L_	2,5
IM	2,0	LO	1,0	IN	2,3
IL	1,6	-	-	IM	2,0
LO	1,0	-	-	LO	1,0
IO	1,0	-	-	OP	1,0
-	-	-	-	AS	1,0
-	-	-	-	RR	0,0

Elaboração: Os autores (2021).

Percebe-se pela Tabela 10 que os únicos bigramas que estão, simultaneamente, nos três BDGs no arranjo {2,2} são ‘CI’, ‘IS’, ‘IT’, ‘EN’, ‘R_’, ‘AO’, ‘L_’, ‘LO’ e ‘IM’. Dessa forma, cabe analisá-los individualmente. Em relação aos bigramas ‘CI’, ‘IS’, ‘IT’, ‘EN’, ‘R_’, ‘AO’, ‘L_’ e ‘IM’, percebe-se que eles possuem uma distância significativa, o que diminui a probabilidade de que tenha ocorrido uma inversão acidental das teclas no momento da digitação. Em muitos desses casos, inclusive, as teclas são acessadas por mãos distintas do produtor de dados. Assim, há a necessidade de inspeção, uma vez que há potenciais similaridades cujo procedimento desenvolvido não foi capaz de garantir. Porém, o bigrama ‘LO’ possui uma distância normalizada de apenas 1,0 unidade e, somada com a alta frequência de seus caracteres (Figura 2),

pode-se incluí-lo no grupo de bigramas tomados como semelhantes para os casos de potencial inversão.

Além disso, com o objetivo de validar o método de Dice otimizado, considerou-se o BDG do município de Contagem, cujo cadastro geral de endereços (*trechok*) do município foi fornecido pela Secretaria de Planejamento. Essa escolha deu-se devido à análise em pesquisa anterior (MACEDO et al., 2020) e pela possibilidade de conferência *in loco*. Desse modo, realizou-se um pareamento via identidade lexical, considerando o método de Dice tradicional e o otimizado. O número de instâncias pareadas para coeficiente de Dice maior que 0,60 são mostrados na Tabela 11.

Tabela 11 – Pareamento do BDG de Contagem via identidade lexical.

Instâncias pareadas pelo método de Dice	Instâncias com aumento pelo método de Dice otimizado
884	27

Elaboração: Os autores (2021).

Além disso, nas 27 instâncias que tiveram aumento em seu Coeficiente de Dice pelo método de Dice otimizado, obteve-se um aumento mínimo no coeficiente de 3,57% e um aumento máximo de 16,67%, com uma média de 6,11% nos resultados. Em BDGs com alto volume de instâncias, isso significaria no pareamento de muitos registros, além da possibilidade de utilizar um maior valor para o limiar de aceitação, garantindo maior número de registros, concomitante a uma maior verossimilhança entre eles.

9 CONCLUSÃO

O processo de inserção de dados em um BDG está suscetível a vários erros de origem humana, principalmente pela inversão de letras durante o processo de digitação e erros ortográficos. Dessa forma, faz-se necessário ter métodos capazes de identificar as instâncias semelhantes, de modo a se obter uma base de dados mais completa em relação a cada uma tomada individualmente. Assim, diz-se que a integração é benéfica por permitir um acréscimo de dados e a eliminação de redundâncias.

Para isso, o presente trabalho desenvolveu uma metodologia para analisar a frequência de caracteres e bigramas de três BDGs distintos, com a finalidade de estabelecer bigramas invertidos que podem ser considerados como semelhantes. Nesse sentido, percebeu-se que o caractere “A” é o de maior frequência nos três BDGs analisados, com uma frequência média de 13,60%; outrossim, percebe-se que, dos 5 caracteres mais utilizados, 4 correspondem às vogais (A, O, E e I), devido à construção lexical existente na Língua Portuguesa. Com posse dessas informações, dividiu-se os caracteres em três *clusters* e elaborou-se um mapa de calor, possibilitando a identificação das regiões de maior utilização.

Além da análise da frequência dos caracteres, determinou-se quais bigramas seriam considerados de referência e quais seriam os invertidos, por meio de uma codificação padrão, e estabeleceu-se as frequências de ambos. Análogo ao que foi desenvolvido para os caracteres, distribuiu-se os bigramas em *clusters* e analisou-se o par de bigramas de referência e invertido em diferentes tipos de arranjos. Assim, desconsiderou-se os arranjos em que pelo menos um dos bigramas fosse pertencente ao *cluster* 1, devido à baixa frequência e, consequentemente, menor probabilidade de pareamento. Sob esse ponto de vista, os arranjos {3,3} e {2,3} foram considerados semelhantes (‘A_’ e ‘_A’; ‘AR’ e ‘RA’; ‘NA’ e ‘NA’, por exemplo) e o {2,2} necessitou de uma análise da distância entre os caracteres em um teclado, o que mostrou que apenas o bigrama ‘LO’ poderia ser acrescentado aos casos de potencial inversão.

Construída a metodologia, o algoritmo de Dice foi melhorado e implementado. Os resultados mostraram uma melhora em 27 instâncias das 884 pareadas no BDG (3,05%) de Contagem-MG, confirmadas com inspeção de campo. Nessas *strings* melhoradas, obteve-se um acréscimo médio de 6,11% do valor para o Coeficiente de Dice.

De posse da metodologia otimizada, constatou-se sua importância no processo de reambulação. Nesse sentido, ao ser aplicada em um BDG com grande número de instâncias, como a base de dados de todo o país, por exemplo, pode resultar no pareamento de um número significativo de instâncias, demonstrando sua validade e aplicabilidade em identificar os topônimos.

Agradecimentos

Os autores agradecem a Pró-Reitoria de Pesquisa da Universidade Federal de Minas Gerais (UFMG) e a Fundep (Fundação de Desenvolvimento da Pesquisa) que, por meio do processo de nº 28359*76, viabilizaram e fomentaram o desenvolvimento desta pesquisa. Além disso, agradecem o apoio do município de Contagem – MG pela disponibilização do cadastro geral de endereços para desenvolvimento e aplicação da metodologia.

Contribuição dos Autores

Lanna Kallen Parreiras foi responsável pela investigação, conceptualização, curadoria dos dados, análise formal, investigação, metodologia, visualização e redação. Fredy Sales Ribeiro contribuiu com as etapas de investigação, conceptualização, curadoria dos dados, análise formal, investigação, metodologia, software e redação. E Vagner Braga Nunes Coelho foi responsável pela investigação, conceptualização, curadoria dos dados, análise formal, investigação, metodologia, aquisição de financiamento, administração do projeto, supervisão e validação.

Conflitos de Interesse

Os autores declaram que não há conflito de interesses.

Referências

ALDANA-BOBADILLA, E.; MOLINA-VILLEGAS, A.; LOPEZ-AREVALO, I.; REYES-PALACIOS, S.; MUÑIZ-SANCHEZ, V.; ARREOLA-TRAPALA, J. **Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text**. Remote Sensing, v. 12, n. 18, p. 3041, 2020. DOI. 10.3390/rs12183041.

BARBOSA, R. P. **Glossário de termos genéricos de nomes geográficos.** Prefácio de Orlando Valverde. IBGE. Rio de Janeiro, 2004.

BENTO, A.; ZOUAQ, A.; GAGNON, M. **Ontology matching using convolutional neural networks**. In: Proceedings of the 12th language resources and evaluation conference, 2020, Marseille, France. Proceedings. Marseille, France: European Language Resources Association, 2020. p. 5648-5653.

COELHO, V. B. N. **Processamento de consultas em banco de dados geográficos ambíguos.** Rio de Janeiro: UFRJ/COPPE, 2010. Disponível em: < http://objdig.ufrj.br/60/teses/coppe_d/VagnerBragaNunesCoelho.pdf >. Acesso em: 15 maio 2022.

CNEFE. **Cadastro Nacional de Endereços para Fins Estatísticos, relativo à cidade de São Paulo.** Disponível em: <http://dados.prefeitura.sp.gov.br/pt_PT/dataset/cadastro-nacional-de-enderecos-para-fins-estatisticos-censo-2010>. Acesso: 12 de maio de 2021.

DICE, L. R. **Measures of the amount of ecologic association between species**. Ecology, v. 26, n. 3, p. 297-302, 1945. DOI. 10.2307/1932409.

DIEDRICH, M. H.; MACHADO, N. T. G. **Toponímia: cultura e patrimônio do Rio Grande do Sul.** Caderno Prudentino de Geografia, v. 1, n. 42, p. 98-117, 2020.

ELMASRI, R.; NAVATHE, S. B. **Fundamentals of Database Systems**. 7. ed. Boston: Pearson, 2016.

FIELDING, N. G.; VERD, J. M. **Geographic Information Systems as Mixed Analysis.** *In*: ONWUEGBUZIE, A. J.; JOHNSON, R. B. **The Routledge Reviewer's Guide to Mixed Methods Analysis**. New York: Routledge, 2021. cap. 21, p. 227-237. DOI. 10.4324/9780203729434-21.

JENKS, G. F. **The data model concept in statistical mapping.** International yearbook of cartography, v. 7, n. 1, p. 186-190, 1967.

KEISER, J.; LEMIRE, D. **Validating UTF-8 in less than one instruction per byte.** Software: Practice and Experience, v. 51, n. 5, p. 950-964, 2021. DOI. 10.1002/spe.2920.

MACEDO, D. R.; COELHO, V. B. N.; NASCIMENTO, G. H.; RIBEIRO, F. S.; KALLEN, L. P. **Relatório técnico: Integração do Cadastro Único com Cadastro Nacional de Endereços para Fins Estatístico através da modelagem de um banco de dados espacial.** CHAMADA CNPq/Ministério da Cidadania n° 30/2019. Belo Horizonte, 2020.

PIVAC, D.; ROIĆ, M. **Systematic monitoring of cadastral resurveys.** Geodetski list, v. 74, n. 2, p. 221-238, 2020.

SANTOS, C. J. B. **Geonímia do Brasil: A padronização os nomes geográficos num estudo de caso dos municípios fluminenses.** Rio de Janeiro: UFRJ, 2008.

SILVA, R. B. **Cadastro técnico multifinalitário: da atualização cartográfica a desafios de gestão urbana em Várzea Grande – MT**. Revista Mato-Grossense de Geografia, v. 18, n. 1, p. 3-22, 2020.

SHOAIB, M.; DAUD, A.; AMJAD, T. **Disambiguation in Bibliographic Databases: A Survey**. arXiv preprint arXiv:2004.06391, 2020. DOI. 10.48550/arXiv.2004.06391.

SUBATHRA, M.; UMARANI, V. **AHP based feature ranking model using string similarity for resolving name ambiguity.** International Journal of Nonlinear Analysis and Applications, v. 12, n. Special Issue, p. 1745-1751, 2021. DOI. 10.22075/IJNAA.2021.5862.

ZAMBON, G.; MUCHETTI, S. S.; SALVI, D.; ANGELINI, F.; BRAMBILLA, G.; BENOCCI1, R. **Analysis of noise annoyance complaints in the city of Milan, Italy**. In: Journal of Physics: Conference Series. Rome, Italy: IOP Publishing, 2020. v. 1603, p. 012029. DOI. 10.1088/1742-6596/1603/1/012029.

Biografia do autor principal

Lanna Kallen Parreiras é natural de Belo Horizonte – MG, é técnica em Edificações pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) e graduanda em Engenharia Civil na Universidade Federal de Minas Gerais (UFMG). Foi bolsista de Iniciação Científica em dois projetos do Departamento de Cartografia do Instituto de Geociências da UFMG e, atualmente, atua como estagiária de Geoprocessamento na Unidade Regional de Geoinformação de Minas Gerais. Futura mestre, doutora e professora universitária, acredita na pesquisa e no desenvolvimento científico e tecnológico como propulsores para um mundo melhor.