ISSN: 2447-3359

REVISTA DE GEOCIÊNCIAS DO NORDESTE

Northeast Geosciences Journal

v. 11, nº 2 (2025)

https://doi.org/10.21680/2447-3359.2025v11n2ID39805

Aprimorando a interoperabilidade entre dados geoespaciais: métricas de alinhamento de similaridade semântica PLN com IA entre dados de cobertura e uso da terra

Enhancing interoperability between geospatial data: NPL semantic similarity alignment metrics with AI between land cover and land use data

Vitor Silva de Araujo [1]; Silvana Phillipi Camboim [2]; Naíssa Batista da Luz [3]

[1] Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas/Departamento de Geomática, Curitiba/PR, Brasil. Email: vitorsilvadearaujo@ufpr.br
ORCID: http://orcid.org/0000-0003-4880-3016

[2] Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas/Departamento de Geomática, Curitiba/PR, Brasil. Email: silvanacamboim@ufpr.br
ORCID: https://orcid.org/0000-0003-3557-5341

[3] Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas/Departamento de Geomática, Curitiba/PR, Brasil. Email: naissa@ufpr.br
ORCID: https://orcid.org/0000-0001-9803-9170

Resumo: A evolução das fontes de dados geoespaciais e seus diversos sistemas de classificação impõem desafios à integração e interoperabilidade de dados. Esta pesquisa aborda esses desafios introduzindo uma metodologia baseada em Inteligência Artificial (IA) que utiliza Processamento de Linguagem Natural (PLN) para medir a similaridade semântica entre o uso da terra, os sistemas de classificação da vegetação e o banco de dados topográfico nacional. Aproveitando técnicas de PLN, como as do ChatGPT-4.0, esta abordagem automatiza o processo de alinhamento semântico, reduzindo o trabalho manual. O estudo teve como objetivo alinhar o mapeamento topográfico brasileiro ET-EDGV com sistemas de classificação nacionais (Manuais de Vegetação e Uso da Terra do IBGE) e internacionais (Dynamic World, Avaliações Globais de Recursos Florestais - FRA) mais abrangentes. Ao aplicar coeficientes de similaridade semântica, a pesquisa buscou criar uma estrutura harmonizada para a integração de dados geoespaciais. A metodologia combinou medidas de similaridade semântica baseadas em IA garantindo um alinhamento consistente dos dados. Os resultados mostraram fortes alinhamentos para classes como "Vegetação Cultivada" e "Culturas Agrícolas" e identificaram desafios para ecossistemas brasileiros únicos, como a "Campinarana". A classe "Manguezal" destacou a necessidade de definições específicas para cada contexto. O estudo conclui que o PNL (Processamento de Linguagem Natural) pode contribuir para o alinhamento semântico automatizado, aprimorando a integração e a interoperabilidade de dados geoespaciais. Embora focada em dados brasileiros, essa metodologia é adaptável globalmente, permitindo uma representação mais precisa da paisagem e uma tomada de decisão mais informatizada. Pesquisas futuras devem integrar modelos avançados de IA (Inteligência Artificial) e ecossistemas mais amplos para refinar o processo.

Palavras-chave:: Mapa topográfico; Processamento de linguagem natural; Similaridade semântica.

Abstract: The evolution of geospatial data sources and their diverse classification systems poses challenges to data integration and interoperability. This research addresses these challenges by introducing an AI-driven methodology that utilises Natural Language Processing (NLP) to measure semantic similarity between land use, vegetation classification systems, and the national topographic database. Leveraging NLP techniques, such as those in ChatGPT-4.0, this approach automates the semantic alignment process, reducing manual work. The study aimed to align the Brazilian ET-EDGV topographic mapping with broader national (IBGE Vegetation and Land Use Manuals) and international (Dynamic World, Global Forest Resources Assessments (FRA)) classification systems. By applying semantic similarity coefficients, the research sought to create a harmonized framework for integrating geospatial data. The methodology combined AI-based semantic similarity measures, ensuring consistent data alignment. Results showed strong alignments for classes like "Cultivated Vegetation" and "Crops" and identified challenges for unique Brazilian ecosystems such as "Campinarana". The "Mangrove" class highlighted the need for context-specific definitions. The study concludes that NLP can contribute to automated semantic alignment, enhancing geospatial data integration and interoperability. Although focused on Brazilian data, this methodology is adaptable globally, supporting more accurate landscape representation and informed decision-making. Future research should integrate advanced AI models and broader ecosystems to refine the process.

Keywords: Topographic Map; Natural language processing; Semantic similarity.

Recebido: 09/04/2025; Aceito: 03/10/2025; Publicado: 25/12/2025.

1. Introdução

Descrever a paisagem é uma função fundamental em mapear, sendo que o mapeamento topográfico aborda especificamente a representação da paisagem. Fremlin e Robinson (1998) afirmam que o mapeamento topográfico representa a Terra como uma entidade composta, onde a paisagem reflete sua aparência. No entanto, elementos importantes da paisagem, como a vegetação, têm consistentemente apresentado desafios devido a sua tendência rápida à obsolescência (Gersmehl, 1981; Langran, 1985). O advento das tecnologias de sensoriamento remoto demonstrou precocemente seu valor para o mapeamento da paisagem (Doyle, 1973). Inicialmente, a classificação de uso e cobertura da terra (UCT) aderiu a dois princípios significativos: hierarquização baseada em escala e compatibilidade semântica com outras fontes de dados confiáveis (Andeson et al., 1976).

Ao longo do tempo, muitas definições semânticas surgiram de várias iniciativas nacionais de mapeamento sistemático e de inúmeros projetos de monitoramento de uso e cobertura da terra (UCT). A paisagem atual exige a integração de dados de múltiplas fontes, tornando a compatibilidade semântica essencial para alcançar a interoperabilidade completa. A integração da semântica no mapeamento — seja para fins de uso e cobertura da terra (UCT) ou topográficos — requer um alinhamento cuidadoso com as classificações empregadas por diversas regulamentações. A integração de fontes heterogêneas requer compatibilidade semântica, especialmente para conceitos vagos como "floresta" (Bennett, 2001; Mallenby, 2008; Varzi, 2001).

A integração e a reutilização de dados de diversas fontes, escalas e usos dependem da interoperabilidade dos dados. Ballatore et al. (2013) e Robinson et al. (2017) enfatizam que a clareza, a confiabilidade e a aplicabilidade dos itens cartográficos podem ser afetadas negativamente se os modelos não apresentarem alinhamento conceitual. Contextos digitais, interativos e orientados ao usuário agravam especialmente esse problema. Outro problema são as raras atualizações dos mapas topográficos brasileiros; muitos ainda se baseiam em dados da década de 1990. Isso contradiz as rápidas mudanças regionais em biomas como a Amazônia, o Cerrado e a Caatinga (Souza, 2020) e enfatiza a necessidade de respostas integradas e flexíveis. Pesquisas revelam que a integração de dados geoespaciais exige compatibilidade estrutural e geométrica. Mais importante ainda, exige também coerência conceitual entre os modelos de dados (Kuhn, 2003; Yu et al., 2018; Machado e Camboim, 2024). Isso é particularmente crucial ao lidar com categorias temáticas complexas, como vegetação, terreno ou cobertura do solo.

No Brasil, há uma clara falta de metodologias para a integração automatizada de dados entre instituições, como o Departamento de Serviços Geográficos do Exército Brasileiro (DSG) e o Instituto Brasileiro de Geografia e Estatística (IBGE), que mapeiam conceitos similares ou equivalentes. Como destacam Souza et al. (2025), melhorias em análise textual, processamento de linguagem natural (PLN) e tecnologias de inteligência artificial poderiam ajudar a solucionar alguns problemas de alinhamento semântico entre diferentes estruturas conceituais. As definições dessas categorias têm variado entre instituições, regiões e disciplinas técnicas (Bravo, 2014; Brown et al., 2022). Com decisões subjetivas sobre os conceitos, esse alinhamento tornou-se um processo manual e demorado. Atualmente, não existem ferramentas especializadas para medir ou quantificar a equivalência entre entidades mapeadas em diferentes modelos.

Quantificar a equivalência, expressa como similaridade semântica, entre entidades mapeadas em diferentes modelos de dados deve permitir o alinhamento de classes com base em seus valores. Assim, definições semânticas similares devem apresentar métricas que indiquem alto grau de equivalência entre duas entidades. Da mesma forma, será possível analisar métricas gerais entre os modelos, quantificando a equivalência entre eles e destacando as entidades menos e mais equivalentes com base em análises individuais, o que pode indicar quais características dos dados podem promover a interoperabilidade quando adaptadas. Além disso, indica um caminho para automatizar esse processo de alinhamento.

Isso não apenas aborda a lacuna na cobertura de mapas topográficos nacionais, sugerindo o uso de dados de mapeamentos mais frequentes e com similaridade métrica, mas também abre a possibilidade de aprimorar o modelo atual, usando indicadores métricos semelhantes ou diferentes entre as entidades. Isso abre a possibilidade de aprofundar a interoperabilidade entre instituições, onde, em um cenário benéfico, essas instituições nacionais teriam maior grau de interoperabilidade em suas produções, promovendo políticas públicas baseadas em evidências.

Nesse contexto, o Processamento de Linguagem Natural (PLN) surge como uma ferramenta crítica para formalizar a geosemântica (Kuhn, 2005), permitindo a comparação de definições e aprimorando a interoperabilidade, o que é fundamental para integrar diversos conjuntos de dados geoespaciais (Elavarasi et al., 2014; Martinez-Gil, 2022; Meng et al., 2013; Yuhua Li et al., 2003). À medida que novas fontes de dados se tornam disponíveis, as técnicas de PLN oferecem uma abordagem poderosa para processar, compreender e conciliar os extensos componentes textuais das informações geoespaciais. Este estudo utiliza PNL e IA para alinhar o banco de dados topográfico brasileiro (ET-EDGV) com classificações nacionais e internacionais, combinando assim dados atualizados com mais frequência e aprimorando a interoperabilidade e a qualidade das informações geoespaciais.

O objetivo principal é alcançar o alinhamento semântico por meio da similaridade entre o banco de dados topográfico brasileiro e bancos de dados nacionais e internacionais, com potencial para automatizar a operação. A metodologia é ilustrada por um estudo de caso brasileiro, cujo modelo de dados topográficos é caracterizado por alta precisão posicional, mas não é suficientemente atualizado (Silva & Camboim, 2020). Em contraste, outros bancos de dados de uso e cobertura da terra (UCT), atualizados com mais frequência usando imagens de satélite, fornecem dados mais atuais. A integração desses conjuntos de dados em um nível semântico proporcionaria benefícios significativos para a conservação e a tomada de decisões, combinando os pontos fortes de ambos os tipos de dados.

Os resultados sugerem a possibilidade de alinhar semanticamente diferentes modelos de dados, mensurar operações por similaridade e viabilizar a automação futura de processos. Portanto, além dos alinhamentos, foi possível analisar os agrupamentos de classes e suas diferenças, indicando a possibilidade de adaptar os dados para promover a interoperabilidade.

Esta pesquisa avança nas práticas de integração de dados e aprimora a qualidade da informação geográfica. Embora focada em dados geoespaciais brasileiros, a metodologia proposta tem ampla aplicabilidade. Ela fornece uma estrutura robusta para medir a similaridade semântica, que pode orientar a integração de diversas fontes de dados geoespaciais globalmente, promovendo assim a interoperabilidade de dados e aprimorando a qualidade geral da informação geoespacial.

2. Método

2.1 Seleção e aplicação de métodos de similaridade semântica

Para medir a similaridade semântica em dados geoespaciais, métodos tem sido desenvolvidos: métodos baseados em conhecimento, métodos baseados em corpus, métodos baseados em redes neurais profundas e métodos híbridos (Gorman & Curran, 2006; Rada et al., 1989; Sánchez et al., 2012; Wang et al., 2017; Zhu & Iglesias, 2017). Entre estes, os métodos baseados em corpus foram selecionados por sua capacidade de aproveitar grandes volumes de texto e capturar relações contextuais entre termos geoespaciais, com base na hipótese distribucional (Ali et al., 2018; Chandrasekaran & Mago, 2022; Martinez-Gil, 2022; Sitikhu et al., 2019; Gorman & Curran, 2006). Usando essa hipótese, os métodos baseados em corpus constroem representações vetoriais que capturam efetivamente as relações semânticas dentro da terminologia geoespacial.

Entre as técnicas baseadas em corpora disponíveis, os embeddings de palavras ganharam destaque. Diversos métodos, como redes neurais e matrizes de coocorrência de palavras, têm sido utilizados para gerar esses embeddings, com modelos populares incluindo word2vec, GloVe, fastText e BERT (Bojanowski et al., 2017; Devlin et al., 2019; Levy & Goldberg, 2014; Mikolov et al., 2013; Pennington et al., 2014; Schnabel et al., 2015). Avaliando a eficácia desses modelos (Dharma et al., 2022), selecionamos o modelo Generative Pre-trained Transformer (GPT), o avanço mais recente em PNL, conhecido por sua capacidade de capturar nuances semânticas complexas em grandes conjuntos de dados. Essa escolha reflete a eficácia e a robustez da arquitetura Transformer (Vaswani et al., 2023), que lida com relações semânticas no domínio geoespacial.

Uma vez gerados os vetores de palavras, o próximo passo crítico é medir com precisão a distância entre eles. Dentre as diversas medidas de similaridade, a similaridade de cosseno é a mais eficaz para conceitos geoespaciais complexos (Ali et al., 2018; Chandrasekaran & Mago, 2022; Machado-García et al., 2014; Sitikhu et al., 2019). Ela é amplamente utilizada em PNL devido à sua capacidade de comparar vetores de diferentes comprimentos e capturar relações direcionais, tornando-a ideal para avaliar similaridades semânticas em dados geoespaciais.

A similaridade por cosseno calcula o cosseno do ângulo entre dois vetores, com valores que variam de -1 (vetores opostos) a 1 (vetores idênticos). Essa medida envolve a normalização de vetores e o cálculo de seu produto escalar, fornecendo uma indicação confiável de relações textuais no espaço vetorial (Manning et al., 2009; Wilson & Schakel, 2015). Essa equação foi originalmente usada em um contexto de recuperação de informação e agora é adaptada para comparar definições semânticas.

A Equação 1 mostra a fórmula para a similaridade de cosseno:

$$sim(d_1, d_2) = \frac{\overrightarrow{V}(d1).\ \overrightarrow{V}(d2)}{\left|\overrightarrow{V}(d1).\ |x|\ \overrightarrow{V}(d2)\right|}$$

Equação 1 – Equação de semelhança por cosseno.
Fonte: Manning et al. (2009).

$\overrightarrow{V}$(d1) e $\overrightarrow{V}$(d2) representam as representações vetoriais dos documentos d1 e d2, $\left|\overrightarrow{V}(d1)\right|$ e $\left|\overrightarrow{V}(d2)\right|$ são seus comprimentos euclidianos. Essa normalização garante que a similaridade por cosseno se concentre na direção dos vetores, desconsiderando sua magnitude absoluta, tornando-a particularmente eficaz para comparar definições de diferentes fontes de dados geoespaciais. A Figura 1 ilustra os componentes utilizados para determinar a similaridade entre $\overrightarrow{V}$(d1) e $\overrightarrow{V}$(d2), onde $\overrightarrow{V}(q)$ representa o vetor de consulta e, θ é o ângulo entre eles.

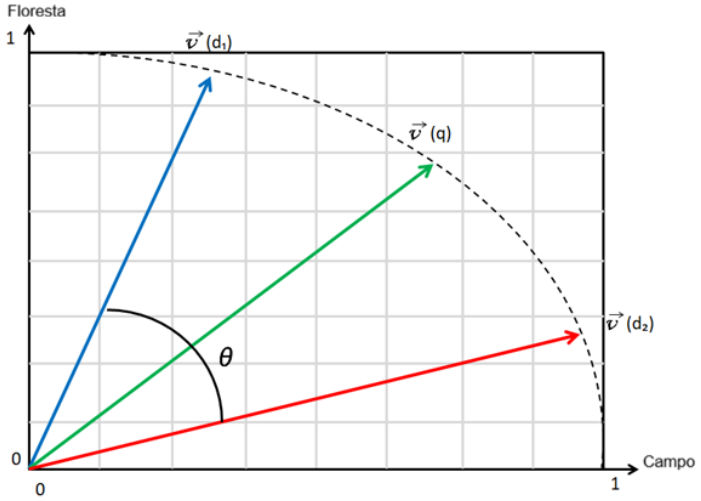


Figura 1 - Componentes de similaridade por cosseno ilustrados entre d_1 e d_2. sim(d1, d2) = cos θ.
Fonte: Adaptado de MANNING et al., 2009.

Para aplicar esse método, uma consulta é tratada como uma "caixa de palavras", e a similaridade por cosseno é usada para medir a pontuação de uma definição em relação a essa consulta. Essa abordagem permite a seleção das correspondências com maior pontuação com base em sua similaridade (Manning et al., 2009). A Equação 2 mostra como a pontuação (Score) de similaridade de cosseno para uma determinada consulta e documento é calculada:

$$Score\,(q,d) = \frac{\overrightarrow{V}(q).\,\overrightarrow{V}(d)}{\left|\overrightarrow{V}(q).\,\overrightarrow{V}(d)\right|}$$

Equação 2 – Score da equação de similaridade do cosseno.
Fonte: MANNING et al. (2009).

A seleção de métodos de similaridade semântica baseados em corpus, especificamente empregando embeddings gerados por GPT e similaridade por cosseno, foi examinada e validada como uma abordagem apropriada para medir a similaridade semântica em dados geoespaciais. Essa combinação fornece uma solução robusta que pode lidar eficazmente com as complexidades inerentes à terminologia geoespacial, aprimorando significativamente a interoperabilidade de dados e a qualidade geral da informação geográfica.

2.2 Definição do estudo de caso e coleta de dados

Neste estudo de caso, a metodologia foi aplicada para aprimorar o banco de dados topográfico nacional na escala 1:25.000, que atende às necessidades de mapeamento de propósito geral de forma diferente dos mapas temáticos projetados para usos específicos (Anderson et al., 1976; Doyle, 1973; Fremlin & Robinson, 1998). Consequentemente, esta seção justifica a seleção de classes de outros modelos, que são usadas como entradas para o método, priorizando a legenda do mapeamento topográfico. Além disso, é apresentada uma proposta para harmonizar as definições de conceitos de várias fontes, visando estabelecer definições semânticas mais precisas para o uso e cobertura da terra (UCT) nas classes do mapa topográfico e reduzir a ambiguidade e o sentido vago desses conceitos geográficos.

Esta metodologia foi desenvolvida para alinhar semanticamente as definições da classificação de UCT com as do mapeamento topográfico brasileiro. O processo envolve três etapas principais: seleção de fontes de dados compatíveis, coleta de dados semânticos e realização do processamento e alinhamento de dados. Cada etapa é detalhada no fluxograma metodológico da Figura 2, que descreve a rotina computacional e os dados de entrada utilizados.

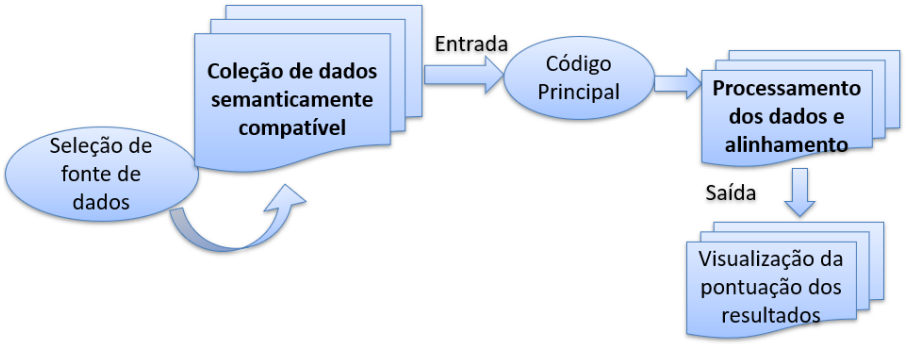


Figura 2 – Fluxograma metodológico.

O primeiro passo foi identificar e selecionar fontes de dados que fossem representadas na mesma escala de mapeamento, e tivessem definições semânticas das classes de dados e se enquadrassem no âmbito do uso e cobertura da terra. A escala escolhida foi a de 1:25.000, a menor escala do mapeamento sistemático nacional, permitindo a generalização para as escalas maiores estipuladas na legislação (DSG, 2017).

Para garantir o uso de padrões amplamente adotados, as agências nacionais de mapeamento do Brasil (DSG e IBGE) foram as principais fontes das definições nacionais para este estudo. Essas fontes incluem o modelo conceitual de mapeamento topográfico ET-EDGV 3.0, homologado pela Comissão Nacional de Cartografia (CONCAR), bem como os modelos temáticos apresentados no Manual de Uso e Cobertura da Terra e no Manual Brasileiro de Vegetação, ambos do IBGE. Para complementar esses dados com fontes globais, também foram incluídos o Relatório de Avaliação dos Recursos Florestais (FRA) e o Dynamic World da Organização das Nações Unidas para Alimentação e a Agricultura (FAO) (Brown et al., 2022; DSG, 2017; FRA, 2015; IBGE, 2012, 2013).

Algumas considerações devem ser feitas em relação às classes selecionadas dos manuais do IBGE para gerar um arquivo de entrada para o código principal. No manual de uso e cobertura da terra, as classes referentes a áreas antropizadas, como esse não é o objetivo dos mapas, não foram consideradas no arquivo de entrada, visto que essa classe específica, no mapeamento topográfico, apresenta grande complexidade em termos de modelagem e conceitualização de dados, sendo sugerida uma abordagem específica para a mesma. As classes de dados relacionadas à água e seus usos não foram consideradas pelo mesmo motivo.

Classes relacionadas a sistemas de transição de vegetação não foram consideradas, pois recomenda-se uma abordagem específica com a classe "Vegetação de contato" no mapa topográfico. Todas as demais definições de classes de dados foram organizadas em uma tabela estruturada, acessível por meio de um link[1]. Essa tabela contém definições para todas as classes e subclasses nos respectivos modelos. Essa tabela serve como nossa "caixa de palavras" e fornece a base para consultas subsequentes. Esse conjunto de dados estruturado garante consistência e fornece um recurso centralizado para comparar e alinhar classes de dados geoespaciais em diferentes fontes.

2.3 Processamento e alinhamento de dados

Utilizamos Python, uma linguagem de programação versátil e amplamente usada, para processar e alinhar definições de dados de forma eficiente. O ecossistema consolidado do Python facilitou a integração perfeita com as bibliotecas da OpenAI, proporcionando um ambiente de pesquisa colaborativo e reproduzível. A implementação foi realizada utilizando um Notebook do Google Collaboratory, escolhido por sua acessibilidade e capacidade de hospedar todas as ferramentas de processamento de linguagem natural (PLN) necessárias para este estudo. O código completo, dividido em duas seções principais (busca semântica e comparação de similaridade), está disponível publicamente no link[2]. Este repositório inclui notas detalhadas do autor e uma versão simplificada contendo apenas o código-fonte para facilitar a replicação.

2.3.1 **Fase de Busca Semântica**

O primeiro componente da metodologia concentra-se na busca semântica, que é crucial para identificar definições com a maior similaridade semântica a um determinado termo de busca. O processo é detalhado no pseudocódigo abaixo:

Algoritmo de Busca_Semântica Pseudocódigo

INPUT: OpenAI_API_key, "words_PT1_frmt.csv"
OUTPUT: Lista classificada de definições com base na similaridade semântica.

IMPORT bibliotecas de PNL necessárias
PROVIDE OpenAI_API_key para autenticar o acesso.

LOAD "word_box_01.csv" Contendo definições de classes de cobertura e uso do solo

[1] https://anonymous.4open.science/r/word_box-4177.

[2] https://anonymous.4open.science/r/MainCode-C56E

```
    COMPUTE embeddings para todas as definições em "word_box_01.csv"

    PROMPT usuário INPUT termo de busca (e.g., "floresta")
    COMPUTE embedding para o termo de busca

    FOR para cada definição em "word_box_01.csv"
       CALCULATE similaridade semântica entre embeddings de termos de pesquisa embeddings de definições
    END FOR

    SORT Definições em ordem decrescente de pontuação de similaridade
    DISPLAY a lista hierarquizada de definições
END Algorithm
```

Esta fase começa com a importação das bibliotecas de PNL necessárias e a autenticação da aplicação com uma chave da API da OpenAI. O arquivo de dados, “word_box_01.csv”, contendo várias definições de classes de cobertura e uso do solo, é então carregado e transformado em vetores de embeddings. Um termo de busca, como “floresta”, é solicitado ao usuário, e seu embedding é calculado. A similaridade entre o embedding do termo de busca e o embedding de cada definição de classe é calculado, classificado e exibido em ordem decrescente de similaridade. Essa abordagem sistemática identifica e classifica as definições mais relevantes, facilitando o alinhamento semântico em todo o conjunto de dados.

2.3.2 Semelhança entre definições de classe

A segunda fase da metodologia concentra-se na comparação da similaridade semântica entre todas as definições de classe. Este processo é detalhado no seguinte pseudocódigo:

```
Algoritmo de similaridade_entre_classes Pseudocódigo

    INPUT: "word_box_01.csv"
    OUTPUT Matriz de similaridade e representações visuais (mapa de calor e gráficos)
    IMPORT Bibliotecas necessárias de PNL e visualização
    DEFINE Modelo de transformação para geração de embeddings de palavras

    LOAD definições de classes de dados de "words.csv"
    COMPUTE embeddings para todas as definições de classe

    INITIALIZE uma matriz nula "sim" com dimensões (N x N), onde N é o número de definições

    FOR cada par de definições (i, j) no conjunto de dados
       CALCULATE similaridade semântica entre embedding da definição_i e definição_j
       STORE valor de similaridade na matriz "sim" na posição (i, j)
    END FOR

    DISPLAY matriz "sim" como um mapa de calor para inspeção visual
    GENERATE grafos relacionais baseados em valores de similaridade
END Algorithm
```

Esta fase gera uma matriz de similaridade que mapeia as relações entre as definições de classe. Cada entrada na matriz representa a pontuação de similaridade por cosseno entre pares de definições de classe, fornecendo informações sobre o alinhamento semântico dentro do conjunto de dados. A matriz resultante é então visualizada como um mapa de calor para aprimorar a interpretabilidade das relações semânticas. Essas ferramentas visuais facilitam uma análise mais profunda de como as classes de dados se relacionam semanticamente em diferentes fontes. Essas etapas metodológicas delineiam uma estrutura clara e reproduzível para o alinhamento de classes de dados geoespaciais usando ferramentas avançadas de PNL

(Processamento de Linguagem Natural) e IA (Inteligência Artificial), permitindo a interoperabilidade semântica em aplicações geoespaciais.

3. Resultados

Os resultados deste estudo demonstram o alinhamento semântico bem-sucedido das definições entre o mapeamento topográfico brasileiro e outras fontes, bem como a extração de métricas de similaridade entre os conceitos mapeados. Esta seção está dividida em subseções para apresentar os resultados e suas implicações.

3.1 Alinhamento de definições e métricas de similaridade

O resultado do processo de alinhamento é uma tabela abrangente que pode ser acessada e baixada por meio de um link fornecido[3]. Em fundo cinza, esta tabela lista os nomes das classes de mapas topográficos brasileiros e seus atributos na primeira coluna, seguidos por suas respectivas definições ET-EDGV na segunda coluna. Da terceira à sexta colunas, encontram-se definições análogas de outras fontes, escolhidas com base nos maiores índices de similaridade calculados entre o mapa topográfico e as definições de cada fonte.

This tabular layout visually represents semantic alignment between the topographic mapping definitions and those from other sources. Additionally, the last column includes a harmonized definition generated by ChatGPT-4.0. This approach highlights the potential of artificial intelligence to refine and enhance the clarity and completeness of geographic concept definitions. Table 1 exemplifies how definitions from different sources have been aligned, emphasizing the potential for automation and machine-readable processing in AI-based applications.

Este layout tabular representa visualmente o alinhamento semântico entre as definições de mapeamento topográfico e as de outras fontes. Além disso, a última coluna inclui uma definição harmonizada gerada pelo ChatGPT-4.0. Essa abordagem destaca o potencial da inteligência artificial para refinar e aprimorar a clareza e a abrangência das definições de conceitos geográficos. A Tabela 1 exemplifica como as definições de diferentes fontes foram alinhadas, enfatizando o potencial de automação e processamento legível por máquina em aplicações baseadas em IA.

Tabla 1- Tabela de alinhamento das classes de dados com as maiores pontuações entre as definições de dados

.

	ET-EDGV 3.0-DEFINITION	FRA – FAO DEFINITION	IBGE-Manual Uso e Cobertura-DEFINITION	IBGE-Manual Vegetação DEFINITION	Dynamic World -DEFINITION	CHAT GPT HARMONIZATION
Grassland / Campo	Campo é uma forma particular de ocorrência (normalmente circunstancial) de uma vegetação …/ Grassland is a particular form of occurrence (usually circumstantial) of vegetation …	Toda a terra que não seja classificada como floresta ou outra terra arborizada. /All land that is not classified as forest or other wooded land.	Entendem-se como áreas campestres as diferentes categorias de vegetação fisionomicamente bem diversa da florestal…/ Grassland areas are understood as the different categories of vegetation that are physiognomically very different from forest vegetation…	A Estepe Gramíneo-Lenhosa é o tipo mais representativo dos campos do sul do Brasil…/ The Grassy-Woody Steppe is the most representative type of grassland in southern Brazil…	Áreas abertas cobertas por gramíneas homogêneas com pouca ou nenhuma vegetação alta.../ Open areas covered by homogeneous grasses with little or no tall vegetation…	Campo é uma área de terra que não é classificada como floresta ou outra terra arborizada.../ Grassland is an area of land that is not classified as forest or other wooded land,

[3] https://anonymous.4open.science/r/Analogous-definitions_output-7362.

Clean Grassland / Campo Limpo	Vegetação predominantemente herbácea, com raros arbustos e ausência de árvores. /Predominantly herbaceous vegetation, with rare shrubs and no trees.	Toda a terra que não seja classificada como floresta ou outra terra arborizada./All land that is not classified as forest or other wooded land.	Entende-se como áreas campestres as diferentes categorias de vegetação fisionomicamente bem diversa da florestal…/Grassland areas are understood as the different categories of vegetation that are physiognomically very different from forest vegetation…	A Estepe Gramíneo-Lenhosa é o tipo mais representativo dos campos do sul do Brasil… /The Grassy-Woody Steppe is the most representative type of grassland in southern Brazil…	Áreas abertas cobertas por gramíneas homogêneas com pouca ou nenhuma vegetação alta... /Open areas covered by homogeneous grasses with little or no tall vegetation…	O termo "Campo Gramíneo-lenhoso" refere-se a áreas de terra que não são classificadas como florestas ou terras arborizadas…/ The term "Grassland-Woodland" refers to areas of land that are not classified as forest or wooded land.
Dirty Grassland /Campo Sujo	Vegetação com fisionomia herbácea e arbustiva, com arbustos e subarbustos espaçados entre si…/ Vegetation with herbaceous and shrubby features, with shrubs and subshrubs spaced apart…	Terrenos definidos como "Outros terrenos Florestados", com mais de 0,5 hectares; com árvores com mais de 5 metros de altura e.../ Land defined as "Other Forested Land" over 0.5 hectares; with trees...	Entende-se como áreas campestres as diferentes categorias de vegetação fisionomicamente bem diversa da florestal, ou seja, aquelas que se caracterizam por …/ Countryside areas are understood as the different categories of vegetation that are physiognomically very different from forest vegetation, that is, those that …	As maiores extensões de Estepe Parque foram observadas na parte leste do Planalto das Araucárias, na porção central do Planalto …/ The largest extensions of Steppe Park were observed in the eastern part of the Araucárias Plateau, in the central portion of the Rio …	Áreas abertas cobertas por gramíneas homogêneas com pouca ou nenhuma vegetação alta. Outras áreas homogêneas de vegetação semelhante a gramíneas.../ Open areas covered by homogeneous grasses with little or no tall vegetation. Other homogeneous areas of grass-like vegetation…	O termo "Campo Parque" refere-se a terrenos que não são predominantemente florestados, agrícolas ou urbanos, com características específicas de vegetação.../ The term "Campo Parque" refers to land that is not predominantly forested, agricultural or urban, with specific vegetation characteristics.

3.2 Análise quantitativa do alinhamento semântico

O grau de similaridade foi avaliado alinhando definições semânticas de diferentes fontes com as classes do mapeamento topográfico brasileiro, representadas pelo valor = S. Os maiores escores de similaridade alinharam-se com classes análogas, permitindo a criação de diagramas de alinhamento que mapeiam as relações entre os modelos de dados. As cardinalidades das relações de dados, como 1...* (uma correspondência em um modelo para muitas em outro) e *...1 (muitas correspondências em um modelo para uma em outro), ilustram o nível de detalhamento dos dados alinhados. Por exemplo, no diagrama que compara o ET-EDGV e o Manual de Vegetação do Brasil, observa-se que as classes de floresta no mapa topográfico são representadas por uma única classe. Em contraste, esse conceito é representado no mapeamento da vegetação por seis classes principais, como ilustrado no diagrama, e mais 26 subclasses a partir destas. Todas as definições das subclasses foram inseridas no arquivo de entrada do código principal para todas as classes dos modelos. No caso da cardinalidade florestal 1...*, o maior valor de escore foi considerado para representá-la no diagrama. A Figura 3 ilustra o alinhamento semântico entre os mapas topográficos brasileiros e as fontes de dados nacionais e internacionais. Observe que os termos das classes de dados foram mantidos em seus idiomas originais.

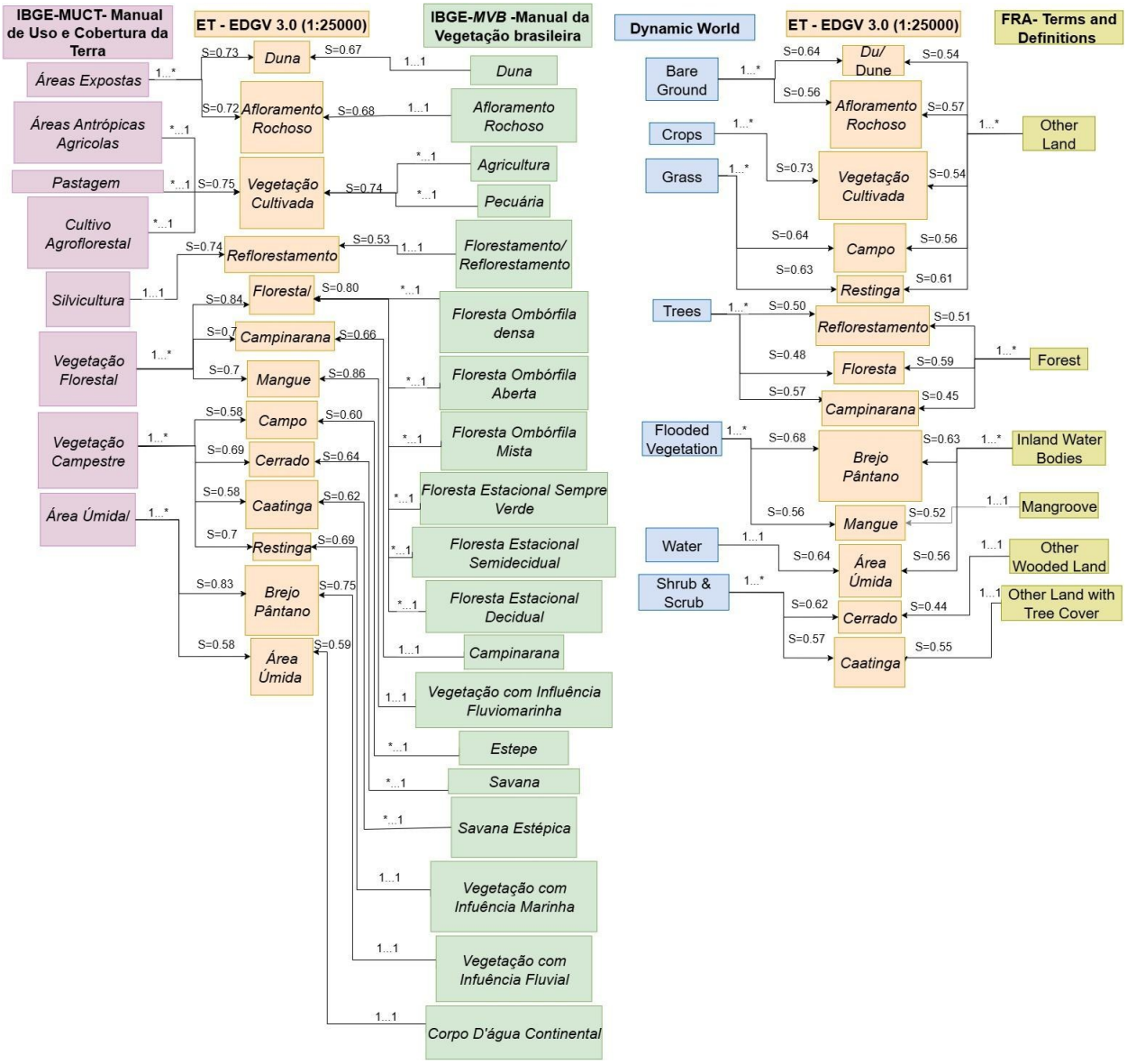


Figura 3 – Diagrama de alinhamento entre o Mapa Topográfico Brasileiro e outras fontes.

3.3 Análise de Correlação

Uma matriz de correlação foi construída para analisar as relações semânticas entre definições de classes de diferentes fontes. Essa matriz representa os índices de similaridade semântica para todos os pares de definições, com valores variando de 0 (nenhuma similaridade) a 1 (definições idênticas). A diagonal principal da matriz contém o valor 1, pois cada definição é comparada consigo mesma. A matriz revela o quão próximas as definições de diferentes fontes estão do padrão brasileiro de mapeamento topográfico (ET-EDGV 3.0). A Tabela 2 mostra um subconjunto desses valores, fornecendo informações sobre o alinhamento semântico entre os conjuntos de dados. Notavelmente, índices de similaridade mais altos foram observados para definições mais detalhadas de fontes do IBGE.

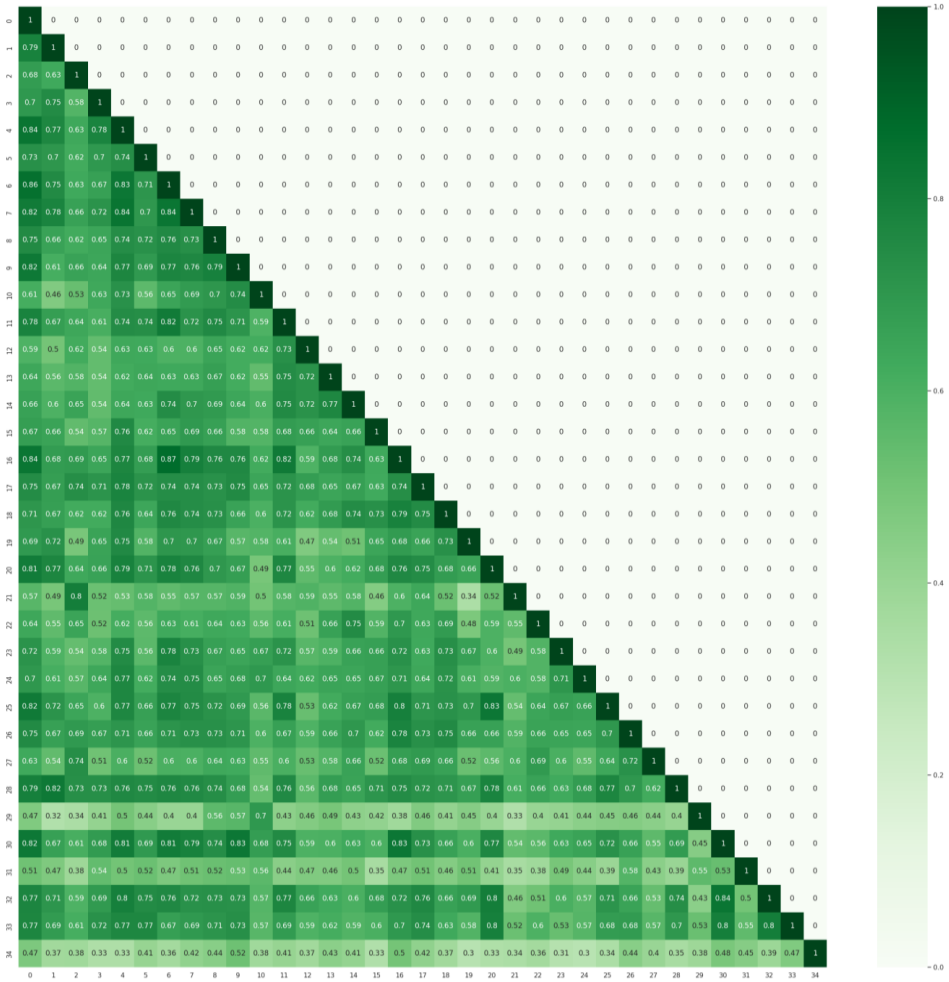
Tabla 2 - Valores de correlação de classe de dados alinhados (pontuação mais alta) e nomenclatura.

ET-EDGV 3.0 Classe	**FRA - FAO Classe/SCORE**	**IBGE MUCT Classe/SCORE**	**IBGE MVB Classe/SCORE**	**Dynamic World Classe/SCORE**
Grassland/ *campo*	Other Land/ *Outras Áreas*/**0.56**	Grassland Vegetation/ *Vegetação Campestre*/**0.58**	Steppe / *Estepe*/**0.6**	Grass/*Grama*/**0.64**
Cultivated Vegetation/ *Vegetação Cultivada*	Other Land/ *Outras Áreas*/**0.54**	Agricultural Area/*Área Agrícola*/**0.75**	Agriculture/ *Agricultura*/**0.74**	Crops/Colheita/ **0.73**
Mangrove/ *Mangue*	Mangrove/ *Mangue*/**0.52**	Forest Vegetation/ *Vegetação Florestal*/**0.7**	Fluvio Marine Influence Vegetation / *Vegetação com influência Fluviomarinha*/ **0.86**	Flooded Vegetation/ *Vegetação* Submersa/ **0.56**
Forest/ *Floresta*	Forest/ *Floresta/* **0.59**	Forest Vegetation/ *Vegetação Florestal*/ **0.84**	Dense rainforest / *Floresta Ombrófila Densa/* 0.8	Trees/ Árvores/**0.48**
Wetland/ *Área Úmida*	In Land Water Bodies/*corpos de água terrestre/* **0.56**	Wetland / *Área Úmida* **0.58**	Continental Water Body/ *Corpos D'água Continentais*/**0.59**	Water/Água/**0.64**
Steppe Savannah /*Caatinga*	Other Land with Tree Cover/ *Outras terras com cobertura arbórea/* **0.55**	Grassland Vegetation/*Vegetação Campestre*/**0.58**	Steppe Savannah/ *Savana Estépica/* **0.62**	Shrub & Scrub/ *Arbusto e Matagal* /**0.57**
Savannah /*Cerrado*	Other Wooded Land/ *Outras terras arborizada/* **0.44**	Grassland Vegetation/ *Vegetação Campestre* /**0.69**	Savannah/ *Savan/* **0.64**	Shrub & Scrub/ *Arbusto e Matagal* / **0.62**
Campinarana	Forest / *Floresta*/**0.45**	Forest Vegetation/ *Vegetação Florestal* /**0.7**	*Campinarana*/**0.66**	Trees/ Árvores/ **0.57**
Reforestation/ *Reflorestamento*	Forest / *Floresta*/**0.51**	Forestry/ *Reflorestamento*/**0.74**	Reforestation/ *Reflorestamento* /**0.53**	Trees/ Árvores/ **0.5**
Restinga	Other Land/ *Outra Terra/* **0.61**	Wetland / *Área Úmida* **0.7**	Marine Influence Vegetation / *Vegetação com influência Marinha/* **0.69**	Grass/ *Grama/* **0.63**
Marsh or Swamp/ *Brejo ou Pântano*	In Land Water Bodies/ *corpos de água terrestre*/**0.63**	Wetland / *Área Úmida* 0.83	Rriver Influence Vegetation/ *Vegetação com influência Fluvial*/**0.75**	Flooded Vegetation / *Vegetação* Submersa/**0.68**
Dune/*Duna*	Other Land/ *Outra Terra*/**0.54**	Exposed Areas/*Áreas Expostas*/**0.73**	Dune / *Duna/* 0.67	Bare Ground/ *Terra nua*/**0.66**
Rocky Outcrop/ *Afloramento Rochoso*	Other Land/ *Outra Terra*/**0.57**	Exposed Areas/ *Áreas Expostas*/**0.72**	Rocky Outcrop/ *Afloramento Rochoso*/ 0.68	Bare Ground/ *Terra nua*/**0.56**

3.4 Visualizando resultados com mapas de calor

Para melhor visualizar essas tendências, foi criado um mapa de calor para representar as correlações de cada elemento de dados com todos os outros elementos. Essa visualização ajuda a identificar a similaridade entre definições de diferentes fontes. A similaridade média entre duas classes pode ser calculada agrupando as definições da mesma classe de dados. Por exemplo, a similaridade entre todas as definições de "Floresta" em diferentes fontes foi determinada, permitindo uma discussão mais aprofundada sobre a utilidade desse coeficiente. A Figura 4 ilustra o mapa de calor aplicado a essas definições de classe. Pode-se observar que definições de maior generalidade apresentam baixos valores de correlação entre si, resultando em uma linha de tom mais claro, como nas linhas 34 e 31. Em contraste, altas correlações ocorrem por meio de pontos escuros ou manchas mais escuras.

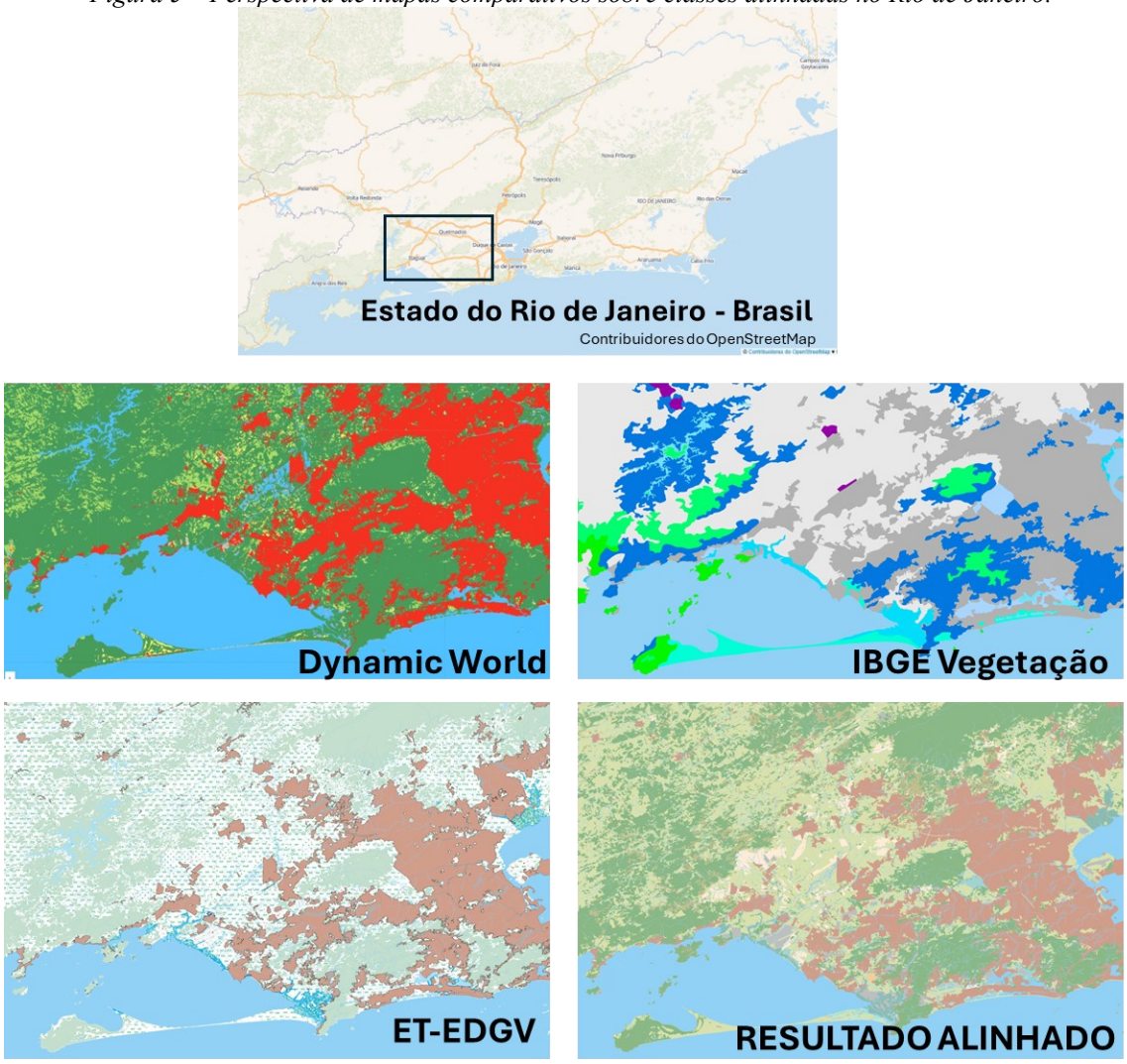
Figura 4 – Mapa de calor das classes de floresta com as pontuações classe por classe.



Foram gerados mapas para fornecer uma perspectiva comparativa das classes alinhadas dentro da mesma área geográfica. Os mapas gerados utilizaram as classes alinhadas para cada modelo, com o objetivo de alcançar uma apresentação uniforme. A região selecionada faz parte do estado do Rio de Janeiro, e todas as imagens foram representadas em uma escala aproximada de 1:250.000. As legendas para o mapa de uso e cobertura da terra e para o mapa topográfico são apresentadas de forma unificada. O mapa de vegetação foi criado de acordo com as normas estabelecidas nos manuais técnicos de mapas pertinentes. A legenda foi obtida do Banco de Informações Ambientais do IBGE, que pode ser acessado

através do link[4]. O mapa Dynamic World, em conjunto com sua respectiva legenda, foi obtido da plataforma oficial do projeto. O mapa de vegetação da FAO foi excluído devido à falta de dados comparáveis. No entanto, os mapas comparativos e suas respectivas legendas podem ser acessados através do seguinte link[5].

Figura 5 – Perspectiva de mapas comparativos sobre classes alinhadas no Rio de Janeiro.



4. Discussão

Este estudo forneceu informações valiosas sobre o alinhamento entre diferentes sistemas de classificação de uso da terra e vegetação, especificamente o Manual de Uso e Cobertura da Terra e Vegetação do IBGE, o Mapeamento Topográfico ET-EDGV e padrões internacionais como o Dynamic World e a Avaliação Global de Recursos Florestais (FRA). Essas informações ajudam a elucidar tanto o potencial de integração quanto os desafios enfrentados com os diversos padrões de mapeamento. Técnicas de inteligência artificial e PNL, como as implementadas no ChatGPT-4.0, também

[4] https://bdiaweb.ibge.gov.br/#/consulta/vegetacao

[5] https://anonymous.4open.science/r/Comparative_Maps-3247/README.md

geraram definições harmonizadas. Essas abordagens podem representar um caminho promissor para melhorar a clareza e a consistência dos sistemas de classificação individuais, embora não ofereçam uma solução definitiva.

A integração de diferentes sistemas de classificação de cobertura e uso da terra (LULC) revela a complexidade de alinhar conjuntos de dados com diferentes níveis de detalhamento e foco temático. A ET-EDGV, como padrão de mapeamento topográfico, fornece um nível de detalhamento mais alto devido à sua dependência de fotografia aérea e verificação em campo. Isso contrasta com os padrões globais, que tendem a ser mais genéricos e podem abranger várias categorias da ET-EDGV. Notavelmente, o Manual de Vegetação do Brasil do IBGE demonstra um alto grau de detalhamento, muitas vezes alinhando-se estreitamente com as classes do ET-EDGV, às vezes em uma correspondência quase 1:1. A integração de dados detalhados de mapeamento topográfico em conjuntos de dados UCT mais amplos tende a ser semanticamente mais precisa do que o inverso, embora alguma perda de informação seja inevitável. Por exemplo, o foco da FRA em florestas ilustra a divergência semântica decorrente de diferentes objetivos e necessidades do usuário.

Uma contribuição central deste trabalho é o uso de inteligência artificial (IA) para quantificar e apoiar conexões semânticas entre definições de classificação. A aplicação de valores de similaridade (valores S) permite medir o alinhamento entre classes, muitas vezes refletindo a interpretação humana. Esses valores revelaram fortes alinhamentos em casos como “Vegetação Cultivada” e “Culturas Agrícolas”, destacando o potencial para uma integração de dados perfeita. No entanto, surgiram discrepâncias, como entre “Pastagem” no IBGE UCT, “Reflorestamento” no Manual de Vegetação do IBGE e a ET-EDGV, indicando a necessidade de investigação adicional. Classes regionais únicas, como “Campinarana” da Amazônia, apresentaram baixos valores S em comparação com as classificações internacionais, refletindo o desafio de alinhar tipos específicos de cada região com categorias globais mais amplas. Da mesma forma, discrepâncias como os baixos valores S para “Manguezal” entre o ET-EDGV e o FRA sugerem diferenças nos critérios ou escopo da classificação.

Outra descoberta importante relaciona-se com o conceito de cardinalidade no alinhamento semântico. Relações de classe um-para-um (1:1) geralmente resultaram em valores S mais altos e alinhamento semântico mais forte do que relações um-para-muitos (1...*). Isso ressalta a maior facilidade de alcançar precisão semântica quando as classificações mapeiam diretamente, em vez de exigir agregação ou desagregação. Esses desafios são especialmente evidentes ao traduzir categorias nacionais detalhadas, como "Área Úmida" do ET-EDGV, em classes internacionais mais generalizadas, muitas vezes exigindo simplificações que reduzem a especificidade e podem levar a desalinhamentos.

Apesar desses desafios, a pontuação de similaridade assistida por IA é valiosa na identificação de alinhamentos ideais e na promoção da interoperabilidade de dados. Trabalhos futuros poderiam explorar o refinamento de padrões internacionais para acomodar melhor ecossistemas específicos de cada região, como "Campinarana", melhorando a representação de biomas únicos. Investigar valores S baixos com maior profundidade também pode ajudar a ajustar modelos de IA para uma melhor correspondência semântica. Os esforços para aprimorar as metodologias de IA devem se concentrar em lidar melhor com relacionamentos um-para-muitos e com as características sutis dos dados temáticos. Além disso, a adoção de grandes modelos de linguagem (LLMs) de código aberto, como o Llama, poderia reduzir a dependência de tecnologias proprietárias, como as da OpenAI.

5. Conclusão

A descrição da paisagem tem sido, há muito tempo, uma função essencial do mapeamento, especialmente para representações topográficas. Este estudo confirma que o mapeamento topográfico, caracterizado por alto nível de detalhamento e precisão, pode ser efetivamente alinhado com classificações temáticas e de uso e cobertura da terra mais amplas, utilizando metodologias baseadas em inteligência artificial (IA). Ao empregar técnicas de processamento de linguagem natural (PLN) e utilizar a similaridade semântica como medida-chave, esta pesquisa abordou o desafio de alinhar fontes de dados geoespaciais díspares, contribuindo para uma melhor interoperabilidade e práticas de integração mais robustas. A metodologia delineada demonstrou que a IA pode preencher lacunas semânticas, criando conexões entre fontes de dados que espelham a compreensão humana e se alinham aos princípios da geosemântica, conforme descrito por Kuhn (2005). Quando utilizada de forma criteriosa, a metodologia mostrou que os valores de similaridade semântica podem

orientar a harmonização de dados, reduzindo o esforço manual e minimizando o viés humano, ao mesmo tempo que garante consistência e relevância em diferentes estruturas de mapeamento. O principal resultado é caracterizado pelo alinhamento entre classes de dados por meio da similaridade semântica entre definições formais. Embora o foco tenha sido em dados geoespaciais brasileiros — destacando ecossistemas únicos como o de Campinarana —, a metodologia possui aplicabilidade global. Ele fornece uma estrutura escalável para integrar diversos conjuntos de dados geoespaciais. Para investigações futuras, recomenda-se a adoção de fatores locais, como características climáticas, para minimizar as especificidades de cada formação, especialmente entre modelos globais. Trabalhos futuros devem incorporar modelos de IA em evolução e expandir o método para incluir ecossistemas e tipos de dados adicionais. Espera-se que os avanços contínuos em PNL e IA aprimorem a precisão semântica da integração de dados, promovendo uma compreensão mais profunda das representações da paisagem e suas propriedades semânticas.

Referências

Ali, A., Alfayez, F., & Alquhayz, H. (2018). Semantic Similarity Measures Between Words: A Brief Survey.

Andeson, J. R., Hardy, E. E., Roach, J. T., & Witmer, R. E. (1976). Professional Paper (Professional Paper). https://books.google.com.br/books?hl=en&lr=&id=dE-ToP4UpSIC&oi=fnd&pg=PA7&dq=+A+Land+Use+and+Land+Cover+Classification+System+for+&ots=sZki0-_l5E&sig=A72-24e0caZeg-TGQYRtmQjRMcc

Bennett, B. (2001). What is a Forest? On the Vagueness of Certain Geographic Concepts. https://doi.org/10.1023/A:1017965025666

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. 5. https://doi.org/10.48550/arXiv.1607.04606 , Focus to learn more.

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., & Tait, A. M. (2022). Dynamic World, Near real-time global 10 m land use land cover mapping. Scientific Data, 9(1), 251. https://doi.org/10.1038/s41597-022-01307-4

Chandrasekaran, D., & Mago, V. (2022). Evolution of Semantic Similarity—A Survey. ACM Computing Surveys, 54(2), 1–37. https://doi.org/10.1145/3440755

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). https://doi.org/10.18653/v1/N19-1423

Dharma, E. M., Gaol, F. L., Warnars, H. L. H. S., & Soewito, B. (2022). The Accuracy Comparison Among Word2vec, Glove, And Fasttext Towards Convolution Neural Network (Cnn) Text Classification.pdf. 100(2). https://doi.org/10.29207/resti.v6i3.3711

Doyle, J. F. (1973). Can Satellite Photography Contribute To Topographic Mapping? p. 315–325.

BRASIL. (2017). Especificações Técnicas Para Estruturação De Dados Geoespaciais Vetoriais (Et-Edgv 3.0). Ministério Do Planejamento, Desenvolvimento e Gestão Comissão Nacional de Cartografia; PDF. https://www.bdgex.eb.mil.br/portal/index.php?option=com_content&view=article&id=81&Itemid=353&lang=pt

Elavarasi, S. A., Akilandeswari, D. J., & Menaga, K. (2014). A Survey on Semantic Similarity Measure.

FRA. (2020). FRA 2020 Terms and Definitions. Viale delle Terme di Caracalla Rome 00153, Italy. https://www.fao.org/3/I8661EN/i8661en.pdf

Fremlin, G., & Robinson, A. H. (1998). What Is It That Is Represented on a Topographical Map? Cartographica: The International Journal for Geographic Information and Geovisualization, 35(1–2), 13–19. https://doi.org/10.3138/CP64-0LM7-0P51-PT77

Gersmehl, P. J. (1981). Maps In Landscape Interpretation. Cartographica: The International Journal for Geographic Information and Geovisualization, 18(2), 79–115. https://doi.org/10.3138/Q508-5316-U142-R6G3

Gorman, J., & Curran, J. R. (2006). Scaling distributional similarity to large corpora. Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL - ACL '06, 361–368. https://doi.org/10.3115/1220175.1220221

IBGE (Ed.). (2012). Manual técnico da vegetação brasileira (2a edição revista e ampliada). Instituto Brasileiro de Geografia e Estatística-IBGE. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=263011

IBGE (Ed.). (2013). Manual técnico de uso da terra (3a edição). Instituto Brasileiro de Geografia e Estatística-IBGE. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=232440

Kuhn, W. (2005). Geospatial Semantics: Why, of What, and How? In S. Spaccapietra & E. Zimányi (Eds.), Journal on Data Semantics III (Vol. 3534, pp. 1–24). Springer Berlin Heidelberg. https://doi.org/10.1007/11496168_1

Langran, G. (1985). Map Design for Computer Processing: Literature Review and DMA Product Critique.

Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings.

Machado-García, N., González-Ruiz, L., & Balmaseda-Espinosa, C. (2014). Recuperación de objetos geoespaciales utilizando medidas de similitud semántica. 8(2).

Mallenby, D. (2008). Handling Vagueness in Ontologies of Geographical Information. https://academiccommons.columbia.edu/doi/10.7916/D8D50KPG

Manning, C., Raghavan, P., & Schuetze, H. (2009). Introduction to Information Retrieval.

Martinez-Gil, J. (2022). A comprehensive review of stacking methods for semantic similarity measurement. Machine Learning with Applications, 10, 100423. https://doi.org/10.1016/j.mlwa.2022.100423

Meng, L., Huang, R., & Gu, J. (2013). A Review of Semantic Similarity Measures in WordNet. International Journal of Hybrid Information Technology, 6(1).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.3115/v1/D14-1162

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1), 17–30. https://doi.org/10.1109/21.24528

Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications, 39(9), 7718–7728. https://doi.org/10.1016/j.eswa.2012.01.082

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/D15-1036

Silva, L. S. L., & Camboim, S. P. (2020). Brazilian Nsdi Ten Years Later: Current Overview, New Challenges And Propositions For National Topographic Mapping. Boletim de Ciências Geodésicas, 26(4), e2020018. https://doi.org/10.1590/s1982-21702020000400018

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019). A Comparison of Semantic Similarity Methods for Maximum Human Interpretability. 2019 Artificial Intelligence for Transforming Business and Society (AITB), 1–4. https://doi.org/10.1109/AITB48515.2019.8947433

Varzi, A. C. (2001). Vagueness in geography. Philosophy & Geography, 4(1), 49–65. https://doi.org/10.1080/10903770124125

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Wang, Z., Mi, H., & Ittycheriah, A. (2017). Sentence Similarity Learning by Lexical Decomposition and Composition (arXiv:1602.07019). arXiv. http://arxiv.org/abs/1602.07019

Wilson, B. J., & Schakel, A. M. J. (2015). Controlled Experiments for Word Embeddings (arXiv:1510.02675). arXiv. http://arxiv.org/abs/1510.02675

Yuhua Li, Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering, 15(4), 871–882. https://doi.org/10.1109/TKDE.2003.1209005

Zhu, G., & Iglesias, C. A. (2017). Computing Semantic Similarity of Concepts in Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering, 29(1), 72–85. https://doi.org/10.1109/TKDE.2016.2610428