Construindo um Dataset Georreferenciado para a Educação Superior no Brasil

Maria A. Ramos[1], **Diêgo de A. Correia**[1], **Rafael Luciano L. Silva**[1], **Fábio J. Coutinho**[1]

[1]Instituto de Computação – Universidade Federal de Alagoas (UFAL)

{marr, dac, rlls, fabio}@ic.ufal.br

Abstract. *INEP annually provides microdata from the Higher Education Census, which includes detailed information on undergraduate courses and educational institutions distributed throughout the country. Although these data are rich in content, they have structural and consistency limitations that hinder their full use, especially in comparative analyses of historical series. Given this scenario, this work describes the development of a dataset containing information from 2009 to 2023. The dataset was submitted to a standardization and enrichment process, during which geolocation data was added. The dataset is available in open format.*

Resumo. *O INEP disponibiliza anualmente os microdados do Censo da Educação Superior, que abrangem informações detalhadas sobre cursos de graduação e instituições de ensino distribuídas por todo o território nacional. Embora sejam ricos em conteúdo, esses dados apresentam limitações estruturais e de consistência que dificultam seu aproveitamento pleno, especialmente em análises comparativas de séries históricas. Diante desse cenário, este trabalho descreve o desenvolvimento de um dataset que compreende as informações do período de 2009 a 2023. O conjunto de dados foi submetido a um processo de padronização e enriquecimento, adicionando dados de geolocalização. O dataset está disponível em formato aberto.*

1. Introdução

No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é a instituição encarregada de conduzir estudos, pesquisas e avaliações regulares sobre o sistema de ensino, com a finalidade de fornecer suporte à elaboração e à execução de políticas públicas voltadas para a educação. Para tanto, a coleta de dados é uma etapa fundamental, sendo realizada anualmente através do Censo Escolar e do Censo da Educação Superior. Essas ações permitem o levantamento de dados estatísticos que possibilitem a compreensão e o monitoramento do sistema de educação no Brasil, viabilizando o planejamento e a implementação de políticas públicas alinhadas à realidade de cada região [Alves 2020].

Em se tratando do Censo da Educação Superior, são coletadas informações detalhadas sobre instituições de ensino superior, tipos de cursos ofertados e perfil dos estudantes de graduação em todo o território nacional. Neste contexto, são disponibilizados pelo INEP registros anonimizados em nível individual, denominados microdados, os quais permitem análises estatísticas mais granulares [da Fonseca and Namen 2016].

Entretanto, conforme observado por [Vizzotto 2020] e [Saraiva et al. 2023], a simples disponibilização dos microdados não garante, por si só, a sua utilização prática. No caso do Censo da Educação Superior, embora exista um volume significativo de informações, o formato original apresenta limitações que dificultam estudos longitudinais (análises históricas). Entre os principais problemas destacam-se as variações na estrutura dos arquivos — com colunas que surgem, desaparecem ou mudam de nome no decorrer dos anos — e a falta de padronização de valores em campos equivalentes.

Além da dificuldade com análises históricas mais robustas, outra questão importante diz respeito à capacidade de promover análises espaciais com maior precisão, visto que a localização dos cursos presenciais disponibilizada pelo INEP informa apenas o nome do município e seu respectivo estado. Além disso, o tratamento dos dados fornecidos pelo INEP possui maior complexidade devido ao grande volume de registros e à quantidade de variáveis disponíveis em cada edição do censo — são centenas de atributos disponibilizados a cada ano.

Para superar essas barreiras, este trabalho constrói um *dataset* que abrange os microdados da Educação Superior, referentes ao período de 2009 a 2023, fornecendo uma visão única e padronizada dos dados. O *dataset* também foi enriquecido com as coordenadas geográficas referentes à localização dos cursos presenciais. Essa informação provê maior precisão e torna as consultas espaciais mais ágeis, permitindo filtros específicos e visualizações geoespaciais sem a necessidade de processamento adicional.

Assim, este documento encontra-se organizado da seguinte maneira: a Seção 2 discute os trabalhos relacionados encontrados na literatura; a Seção 3 descreve a metodologia empregada na coleta e padronização dos dados; a Seção 4 discute as etapas para a construção do *dataset*; a Seção 5 apresenta algumas aplicações do uso do *dataset* e, finalmente, a Seção 6 traz as considerações finais e propostas para trabalhos futuros.

2. Trabalhos Relacionados

Na literatura, são encontrados diversos trabalhos que abordam dados relacionados à área da educação no Brasil. Esse interesse está ligado a um movimento de incentivo ao uso dos microdados abertos do INEP, que ampliou as possibilidades de pesquisa e análise para apoiar políticas públicas. [de Castro Soares et al. 2021] realizaram uma revisão sistemática da literatura a fim de identificar estudos que visassem extrair conhecimento a partir de bases de dados compartilhadas pelo INEP tais como SAEB, IDEB, Censo Escolar e Indicadores Educacionais. Os autores encontraram 407 trabalhos, revelando um grande interesse da comunidade na análise de dados da educação no Brasil. A seguir, discutimos alguns trabalhos que utilizam dados disponibilizados pelo INEP.

[da Fonseca and Namen 2016] exploram bases do INEP utilizando métodos de mineração de dados com o intuito de analisar informações da Prova Brasil a fim de compreender como as características do corpo docente podem influenciar o desempenho em Matemática de alunos do Ensino Fundamental. O trabalho abrange desde a preparação dos dados até a aplicação das técnicas de classificação, apresentando como principal resultado a relação entre fatores como formação, tempo de experiência dos professores e o desempenho dos alunos. A preparação dos dados exigiu etapas de limpeza, seleção de atributos e tratamento de inconsistências, as quais poderiam ser facilitadas caso tivessem um conjunto de dados previamente padronizado e organizado como o proposto em nosso

trabalho.

[Saraiva et al. 2023] apresentaram uma análise descritiva dos cursos de Tecnologia da Informação e Comunicação (TIC) no Brasil, utilizando microdados do Censo da Educação Superior de 2015 a 2021. No estudo, os autores examinaram a evolução no número de cursos, investigaram taxas de evasão e conclusão e traçaram o perfil dos estudantes considerando modalidade de ensino, sexo e situação de matrícula. Para realizar a análise foram implementados *scripts* em R e construído um *dataset*. Algumas etapas realizadas no trabalho seriam facilitadas caso existisse um conjunto de dados integrado, padronizado e enriquecido como o proposto em nosso trabalho. A inclusão de dados de localização também facilitaria a visualização dos resultados em mapas.

Recentemente, [Yamanaka et al. 2024] propuseram uma metodologia estatística para validar o alinhamento de colunas na evolução do esquema de um banco de dados integrado a partir de 12 anos do Censo Escolar no Brasil. O estudo combina testes de aderência (Kolmogorov–Smirnov, Anderson–Darling, Welch's t-test e F-test) para identificar e mapear atributos equivalentes entre os diferentes esquemas dos microdados publicados anualmente, atingindo taxas de acerto superiores a 85%. Essa abordagem confirma a necessidade de padronização das edições de microdados do INEP, o que realizamos em nosso trabalho considerando o intervalo de 2009 a 2023 para dados da Educação Superior.

[Barros et al. 2022] propõem a construção de um *dataset* que reúne dados unificados do Censo da Educação Básica referentes aos anos de 2010 a 2021. Os autores aplicaram limpeza e padronização dos campos e também dividiram em tabelas temáticas (como escola, turma, matrícula e professor), criando um novo dicionário de dados. Apesar de apresentar um objetivo bastante similar à nossa proposta, os *datasets* produzidos são distintos visto que nosso trabalho propõe a integração de microdados do Censo da Educação Superior, considerando o período de 2009 a 2023. Além disso, nossa proposta provê o enriquecimento dos dados do INEP mediante a inclusão de coordenadas geográficas, para facilitar analises espaciais.

3. Descrição e Padronização dos Dados

Esta seção detalha a origem, estrutura e aprimoramentos realizados nos dados utilizados na pesquisa. Inicialmente, apresenta-se a composição dos microdados brutos do Censo da Educação Superior obtidos junto ao INEP, seguida da motivação para integração e enriquecimento do *dataset* visando ampliar suas possibilidades analíticas.

3.1. Dados do INEP

Os dados utilizados neste trabalho foram obtidos a partir dos microdados anuais do Censo da Educação Superior, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Para cada ano, os microdados incluem milhares de registros que abrangem informações sobre diferentes aspectos de alunos, docentes, técnicos educacionais, instituições de ensino superior e cursos de graduação. O INEP[1] disponibiliza esses dados desde o ano de 1995 até o ano de 2023, sua mais recente publicação.

[1]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior

Os microdados do INEP reúnem centenas de campos no seu esquema original, o que dificulta sua apresentação completa neste artigo. Em virtude disso, as Tabelas 1 e 2 apresentam uma visão resumida e agrupada por temas dos principais elementos presentes nos esquemas que descrevem, respectivamente, cursos e IES. Os campos dos cursos estão organizados em seis categorias-chave: dados regionais, características institucionais e de curso, vagas oferecidas, além de perfis detalhados de inscritos, matriculados e concluintes. Esses perfis incluem desagregações por sexo, faixa etária, cor/raça, pessoas com deficiência (PCDs), tipo de financiamento e origem escolar. Os campos das IES englobam dados geográficos, atributos administrativos das IES, informações cadastrais de endereço e métricas sobre corpo docente e técnico-administrativo, incluindo contagens por titulação, regime de trabalho e características demográficas.

Tabela 1. Visão resumida em grupos de campos do esquema de cursos

Grupos de campos	Descrição	Qtd.
Dados regionais	Informações regionais do curso, como região, UF, município, localiza-se na capital.	10
Dados do curso	Características da IES (rede, categoria administrativa) e do curso (nome, código, grau acadêmico, gratuidade, modalidade e nível).	19
Dados sobre vagas	Quantidade de vagas oferecidas, detalhadas por turno e tipo.	9
Dados de inscritos e ingressantes	Contagem de inscritos e ingressantes por sexo, turno, seleção e demais categorias.	53
Dados de matrícula	Contagem de matrículas por sexo, turno, faixa etária, cor/etnia, deficiência, entre outras categorias.	39
Dados de concluintes	Contagem de concluintes por categorias.	39

Para cada ano, os dados são compartilhados em arquivos compactados que incluem dicionário de dados, um arquivo para informações dos cursos e instituições de ensino superior, ambos em formato CSV. Conforme mencionado na seção 2, pesquisadores possuem grande interesse na produção de análises históricas de dados da educação. Em se tratando dos dados da educação superior, a análise histórica possivelmente demandaria as seguintes atividades: (i) baixar cada arquivo compactado correspondente ao ano compreendido no intervalo a ser analisado; (ii) descompactar cada arquivo baixado; (iii) verificar de forma manual ou automatizada as diferenças entre os esquemas de anos distintos; (iv) implementar uma padronização dos esquemas e (v) prover uma visão integrada dos dados para o intervalo de tempo considerado. Essas etapas, embora não apresentem alta complexidade, consomem recursos e tempo que poderiam ser melhor dedicados diretamente à análise dos dados. Portanto, demonstra-se assim a necessidade de um *dataset* integrado para facilitar a realização de análises históricas de dados do INEP.

Outro aspecto importante quando tratamos de análises de dados educacionais no Brasil, país de dimensão continental, é a capacidade de compreender as características distintas de cada região. Ou seja, a necessidade de prover análises espaciais sobre aces-

Tabela 2. Visão resumida em grupos de campos do esquema de instituições

Grupo de campos	Descrição	Qtd.
Dados regionais	Informações geográficas e regionais da IES (UF, município, mesorregião etc.)	12
Dados institucionais	Características organizacionais e administrativas da IES (rede, mantenedora, código INEP etc.)	10
Dados relativos a docentes	Contagem de docentes por regime, titulação, faixa etária, sexo, cor/etnia e outras categorias	31
Dados de endereço	Informações de localização postal da IES (logradouro, bairro, CEP etc.)	5
Dados relativos a técnicos	Contagem de técnicos-administrativos por nível de escolaridade, sexo e titulação	15

sibilidade, distribuição de vagas, evasão, perfil do docente, etc. No entanto, os dados brutos do INEP não disponibilizam a geolocalização de cursos e IES, limitando-se a fornecer o endereço completo (para IES) e o município e a UF (para cursos). Tal fato impõe uma restrição significativa para o uso imediato de softwares de geoprocessamento, tornando essencial o enriquecimento desses dados com coordenadas geográficas (latitude e longitude). Essa adição possibilita operações geoespaciais como cálculo de distâncias, identificação de áreas com baixa cobertura de serviços de Educação Superior e a integração com indicadores demográficos e socioeconômicos, ampliando o potencial analítico da base de dados.

3.2. Descrição do *Dataset*

O *Dataset* construído neste trabalho compreende dados da série histórica de 2009 a 2023, abrangendo 15 edições do Censo da Educação Superior, o que permite analisar tendências ao longo desse período. Os dados de anos anteriores não foram incluídos devido às diferenças significativas entre os esquemas. A partir 2009, os dados do INEP sofreram poucas mudanças em seu esquema, considerando o intervalo de 2009 até 2022 o esquema permaneceu inalterado. O ano de 2023 foi o único da série histórica que teve mudanças no esquema, sendo adicionados três novos campos descritos a seguir:

- `IN_CONFESSIONAL` (indica se a IES é confessional)
- `IN_COMUNITARIA` (indica se a IES é comunitária/sem fins lucrativos)
- `TP_REDE` (classifica a IES como pública ou privada).

tendo os dois primeiros sido adicionados em ambos os esquemas e o último apenas no de IES.

A fim de unificar a série histórica, o esquema de dados de 2023 foi adotado como padrão para o período de 2009 a 2022. Dessa forma, os campos `IN_CONFESSIONAL` e `IN_COMUNITARIA` foram adicionados aos esquemas de cursos e IES, enquanto

TP_REDE foi incluído apenas no esquema de IES, pois já era um dado presente no esquema de cursos desde 2009. Como resultado, o esquema de cursos passou de 200 campos para 202, e o de IES, de 81 para 84 campos. A fim de preservar a integridade analítica e eliminar a ambiguidade entre um valor booleano e um valor inexistente, os registros de presentes em anos anteriores, mas ausentes na edição de 2023, tiveram tais campos tratados como dados nulos.

Complementando os dados originais do INEP, o *dataset* foi enriquecido com coordenadas geográficas (latitude e longitude) das Instituições de Educação Superior. Essas informações foram obtidas a partir do georreferenciamento dos endereços cadastrais combinado com outros atributos da instituição, permitindo análises espaciais refinadas que vão além da granularidade municipal disponível nos microdados. Essa camada geoespacial amplia significativamente o potencial investigativo da base, especialmente para pesquisas sobre equidade no acesso ao ensino superior por região, facilitando o uso do *dataset* em softwares de geolocalização. A descrição completa dos campos está no dicionário de dados presente no repositório público do *Dataset*, disponível em: https://doi.org/10.5281/zenodo.16782940.

4. Construção do Dataset

A construção do *dataset* pode ser descrita em três etapas principais: (i) Extração dos dados brutos (ii) Junção das bases anuais e (iii) Enriquecimento com informações geoespaciais. A Figura 1 apresenta uma visão geral do *workflow* de construção do *dataset*, onde cada módulo corresponde a uma das etapas mencionadas anteriormente, as quais são descritas nas subseções seguintes.

4.1. Extração dos Dados

Conforme descrito na subseção 3.1, os dados brutos foram obtidos a partir dos microdados do Censo da Educação Superior disponibilizados pelo INEP. Inicialmente, considerando o intervalo de 2009 a 2023, os arquivos compactados foram baixados de forma manual, os quais continham um arquivo com dados das IES e outro com dados dos cursos, ambos em formato CSV. Posteriormente, os arquivos foram extraídos manualmente e organizados em uma única pasta.

4.2. Junção das bases anuais

A etapa de agregação dos microdados cadastrais de cursos (2009–2023) foi realizada através de um *script* desenvolvido em Python utilizando a biblioteca `Pandas`[2]. O processamento de junção dos dados ocorreu de forma similar para as bases de cursos e instituições de ensino superior, incluindo todos os campos presentes nos microdados originais (202 campos para cursos e 84 para IES).

A implementação do esquema padronizado do *dataset* ocorreu da forma a seguir. Inicialmente, o *script* copia o cabeçalho dos arquivos de 2023 para ser incorporado aos demais anos. Para cada ano, os arquivos CSV foram processados em blocos de dados (*chunks*). Para os blocos referentes aos anos de 2009 a 2022, foi realizado o enriquecimento dos dados de campos ausentes (IN_CONFESSIONAL, IN_COMUNITARIA e TP_REDE), conforme está detalhado na Seção 4.3.

[2] https://pandas.pydata.org/

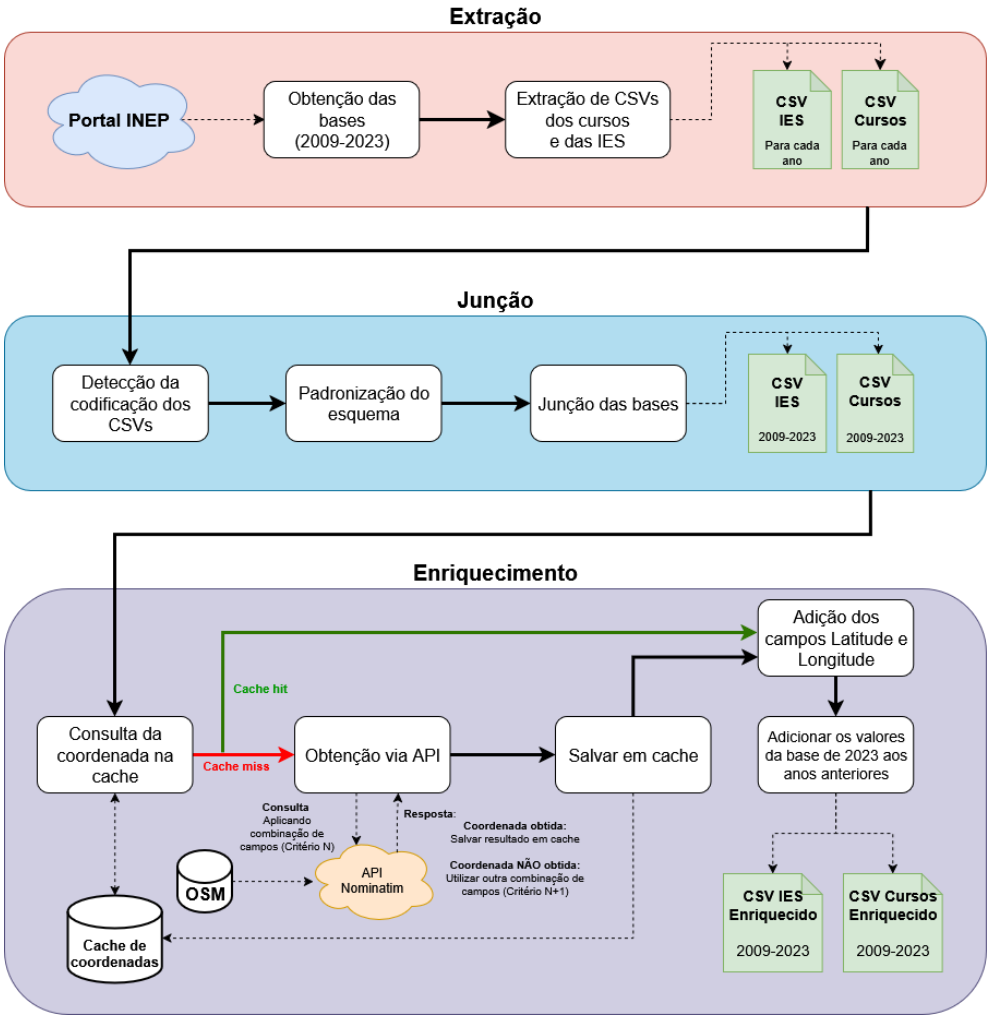

Figura 1. O *workflow* de construção do *dataset*

Logo após, o *script* realiza a validação do campo `NU_ANO_CENSO`, que indica o ano do censo ao qual a linha pertence, para assegurar a completude deste atributo em todas as linhas, realizando sua inserção caso esteja ausente. Após a validação, os *chunks* de dados foram integrados e atribuída a codificação UTF-8.

4.3. Enriquecimento com Dados Geoespaciais

Após a junção das bases, inicia-se a etapa de enriquecimento do *dataset* com coordenadas geográficas (latitude e longitude) das instituições e dos cursos. Este processo teve como objetivo possibilitar análises espaciais, observando aspectos regionais com maior precisão.

A geocodificação foi realizada por meio da API pública Nominatim[3], um serviço de geolocalização do projeto OpenStreetMap[4] (OSM), acessada por meio de *scripts* em Python. O processo de obtenção da geolocalização dos cursos e das IES compreende algumas tentativas de consulta aos dados do OSM, que podem atingir até 7 requisições.

[3]https://nominatim.org/
[4]https://www.openstreetmap.org/

Este processo encontra-se detalhado no workflow de construção, descrevendo a aplicação das possíveis combinações de campos utilizadas para a captura das coordenadas, as quais são apresentadas na Tabela 3.

Tabela 3. Combinações de campos utilizados para obtenção das coordenadas.

	CidadeUF_IES	*CidadeUF_Curso*	*Rua*	*Num*	*CEP*	*Nome_IES*	*Sigla*
C1	X		X	X	X		
C2	X				X	X	
C3	X					X	
C4	X				X		X
C5	X						X
C6	X		X		X		
C7	X				X		
C8		X				X	
C9		X					X
C10		X					

Os dados do INEP apresentam limitações quanto aos dados de localização dos cursos. Esses desafios ocorrem porque em uma IES com mais de um campus, normalmente, seu endereço é relacionado ao campus principal da instituição. Isso pode ocasionar inconsistência ao tentarmos determinar a localização do curso a partir de sua relação com a IES, no caso do curso estar localizado em um campus secundário.

Para superar este desafio, a obtenção das coordenadas de um curso inicia a partir da captura do código da IES a que está relacionado. Este código é utilizado para acessar o município da IES (CidadeUF_IES) e compará-lo com o município descrito nos dados do curso. Se apresentam o mesmo valor, assumimos que o curso está no campus principal, mesmo sabendo que não acontece para todos os casos. Em seguida, faz-se uma requisição à API Nominatim, que retorna a geolocalização mais correspondente à solicitação, utilizando a combinação de campos do critério C1, buscando uma coordenada para o endereço completo descrito nos dados da IES. Caso a combinação de campos não retorne uma coordenada válida, prossegue-se realizando requisições variando a combinação até o critério C7. As diferentes combinações garantem que, em algum momento, uma geolocalização válida seja retornada, mesmo que, em alguns casos, o nível de precisão da geolocalização possa diminuir de acordo com os critérios selecionados.

Se o município da IES e do curso apresentam diferentes valores, as requisições são feitas utilizando a combinação de campos do critério C8 ao C10, que usa o município do curso (CidadeUF_Curso) juntamente com dados das instituições (Nome_IES e Sigla).

Para a obtenção das coordenadas das instituições, o processamento é similar ao realizado para os cursos, porém, utiliza-se apenas as combinações de campos do critério C1 ao C7.

Devido ao elevado número de instâncias na base agregada (cerca de 3 milhões de

cursos), foi criada uma cache de coordenadas como forma de otimizar o processamento. O objetivo da cache é armazenar as coordenadas retornadas após as requisições de cada combinação de município e instituição. Isso evita múltiplas requisições para cursos que tenham a mesma combinação, tornando o processo mais eficiente, já que necessita apenas de uma consulta simples na base.

Ao final do processo de obtenção de todas as coordenadas, as informações das geolocalizações dos cursos e das IES são adicionadas à base agregada. Cerca de 5.8% das coordenadas foram obtidas a partir das requisições feitas com os critérios C1 ao C7, 6.9% com os critérios C8 e C9 e 87.3% com o critério C10.

Além disso, os campos IN_CONFESSIONAL, IN_COMUNITARIA e TP_REDE, presentes apenas nos dados do ano de 2023, foram adicionados às bases de 2009 a 2022. Os valores foram obtidos a partir dos dados de 2023, através do atributo identificador da instituição (`CO_IES`). Para isso, foram utilizados *scripts* com a biblioteca `Pandas`. Nos casos onde alguma IES presente nos dados de 2009 a 2022 não fosse encontrada nos dados de 2023, foi atribuído o valor nulo aos campos, evitando ambiguidades.

Por fim, o *dataset* enriquecido com as coordenadas geográficas e com o preenchimento de campos para sua padronização, é disponibilizado em arquivos CSVs separados.

5. Aplicações

O *dataset* desenvolvido neste trabalho permite a realização de uma série de análises voltadas para proporcionar uma melhor compreensão das temáticas relacionadas à Educação Superior no Brasil, possibilitando revelar diferenças regionais e suas mudanças no decorrer dos anos.

Esta seção apresenta duas possíveis aplicações baseadas no *dataset*, as quais são descritas a seguir.

- **Distribuição espacial de cursos presenciais**: A aplicação representada na Figura 2 permite identificar a distribuição espacial dos cursos de medicina (destacados em vermelho) no Nordeste brasileiro nos anos de 2009 e 2023. Ao confrontarmos as diferenças entre os padrões de distribuição espacial da oferta de cursos de medicina nesses anos, percebe-se que houve um aumento do número de cursos de medicina em regiões interioranas para todos os estados. O estado da Bahia destaca-se pela adição de cursos e também pela boa distribuição ao longo de seu território. Em contraste, os estados do Piauí e Rio Grande do Norte apresentam uma menor ampliação de sua cobertura.
- **Visualização de microdados por categorias ao longo de um período**: A figura 3 exemplifica uma possível análise dos microdados do Censo da Educação Superior, considerando o número de ingressantes agrupados por sexo e região nos cursos de Tecnologias de Informação e Comunicação (TIC) em diferentes quadriênios durante o período de 2009 a 2020. Percebe-se que, embora a quantidade absoluta de mulheres nos cursos tem aumentado nos últimos anos, proporcionalmente, houve uma leve queda em todas regiões, exceto na região Sul onde a proporção entre ingressantes do sexo feminino e masculino manteve-se inalterada.

Cabe ressaltar que além dos exemplos apresentados nesta seção, o *dataset* pode ser utilizado para a análise de diversas outras questões importantes relacionadas à educação

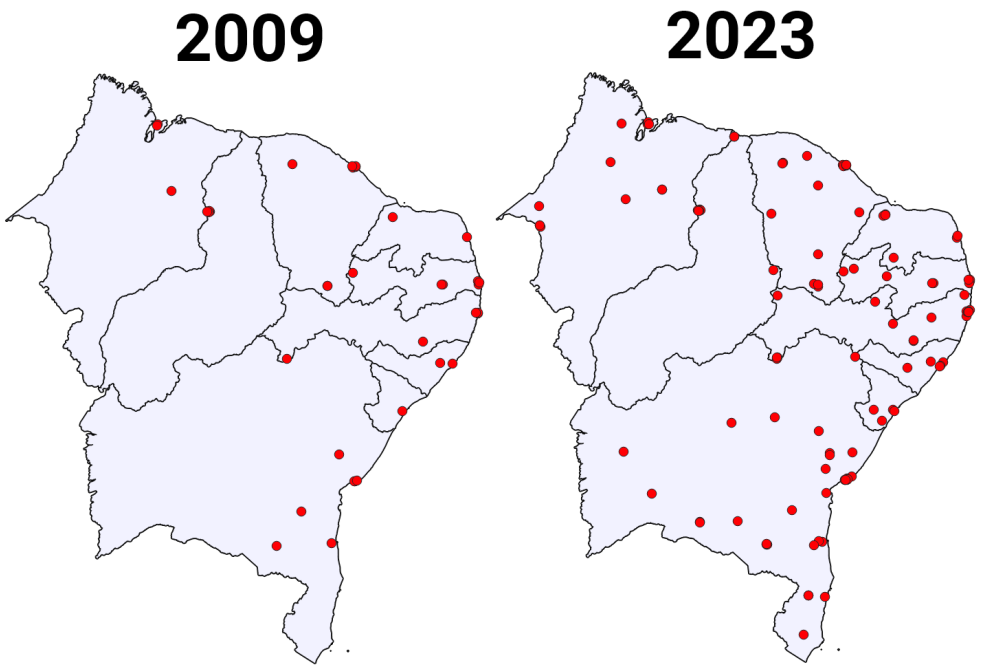

Figura 2. Distribuição espacial dos cursos de medicina na região Nordeste para os anos de 2009 e 2023.

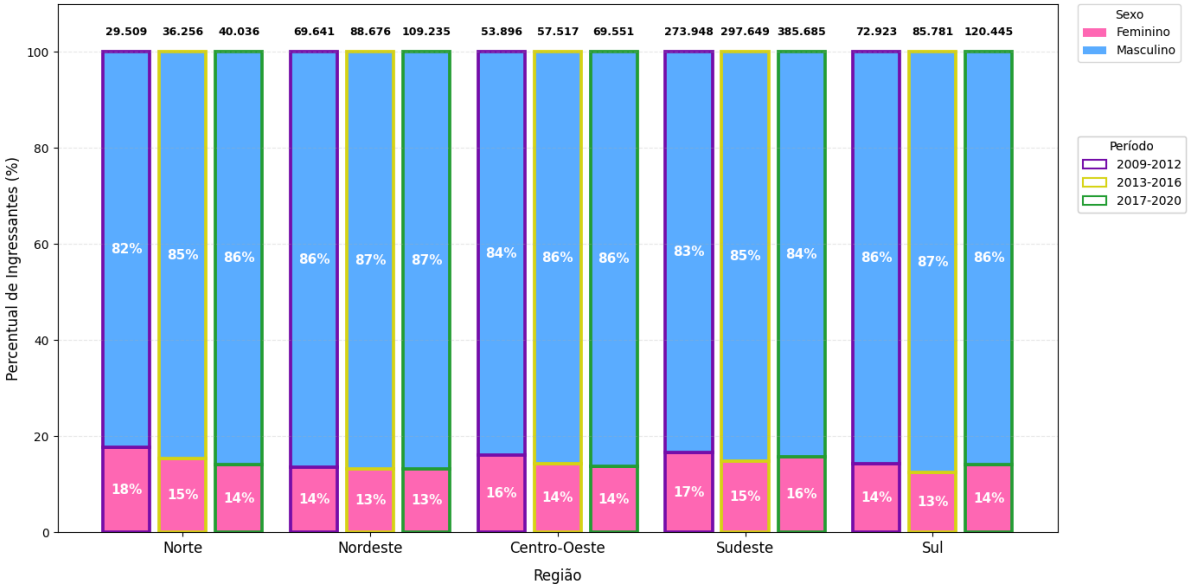

Figura 3. Distribuição de Ingressantes em Cursos de TIC por Sexo, Região e Período (2009-2020)

superior tais como evasão, tempo de permanência do aluno no curso, carência da oferta por região, dados de estudantes com deficiência, e outras. Desta forma, a publicação do *dataset* contribui para a realização de diagnósticos sobre a Educação Superior no Brasil.

6. Considerações Finais

Este artigo apresentou o processo de construção de um *dataset* reunindo dados do Censo da Educação Superior no período de 2009 a 2023, para o qual foram adicionados dados de geolocalização dos cursos e IES. O *dataset* encontra-se disponível em formato aberto no repositório público do Zenodo: https://doi.org/10.5281/zenodo.16782940.

A agregação dos microdados anuais exigiu a padronização de campos entre versões de esquemas diferentes e sua disponibilização permite análises históricas importantes para a formulação de políticas públicas. Além disso, o enriquecimento com informações de geolocalização proporciona análises espaciais mais precisas, ampliando as possibilidades de pesquisas que abordem diferentes aspectos relacionados à desigualdade regional.

Como trabalhos futuros, pretende-se enriquecer o *dataset* com informações de outras fontes públicas, como dados demográficos e socioeconômicos do IBGE, facilitando a realização de análises ampliadas. Também está prevista a validação dos dados de geolocalização do *dataset* e a implementação de *dashboards* interativos para a visualização rápida dos principais indicadores educacionais, facilitando o processo de análise para pesquisadores que não dominam ferramentas avançadas. Além disso, pretende-se adicionar dados de anos anteriores a 2009, a partir da técnica proposta em [Yamanaka et al. 2024] que visa automatizar a integração de esquemas distintos, tornando o *dataset* ainda mais abrangente e útil para analises históricas de longo prazo.

7. Agradecimentos

Este trabalho foi realizado com apoio financeiro da UFAL, por meio do Programa Institucional de Bolsas de Iniciação Científica. Também agradecemos a colaboração dos membros do grupo de pesquisa, Ruan Tenório de Melo e José Matheus Santana Alves, pelas contribuições nas pesquisas que resultaram neste trabalho.

Referências

Alves, M. T. G. (2020). Caracterização das desigualdades educacionais com dados públicos: Desafios para conceituação e operacionalização empírica. *Lua Nova: Revista de Cultura e Política*, (110):189–214.

Barros, A., Alencar, A., Nascimento, A., Albuquerque, A., and Mello, R. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45, Porto Alegre, RS, Brasil. SBC.

da Fonseca, S. O. and Namen, A. A. (2016). Mineração em bases de dados do INEP: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. *Educação em Revista*, 32(1):133–157.

de Castro Soares, R., Neto, N. W., Coutinho, L. R., da Silva e Silva, F. J., dos Santos, D. V., and Teles, A. S. (2021). Mineração de dados da educação básica brasileira usando as bases do INEP: Uma revisão sistemática da literatura. *CINTED-UFRGS Revista Novas Tecnologias na Educação (RENOTE)*, 19(1):361–370.

Saraiva, R. L., Sousa, P. S. d., Araújo, A. A., and Souza, J. (2023). Análise descritiva dos microdados do censo da educação superior do INEP para cursos de tecnologia da informação e comunicação no Brasil – um panorama 2015–2021. In *Anais do XXXI*

Workshop sobre Educação em Computação (WEI), pages 443–453, João Pessoa, PB, Brasil. Sociedade Brasileira de Computação.

Vizzotto, P. A. (2020). Um panorama sobre as licenciaturas em física do Brasil: análise descritiva dos microdados do censo da educação superior do INEP. *Revista Brasileira de Ensino de Física*, 43(1):e20200112.

Yamanaka, M., de Almeida, D., de Almeida, P. R., Dominico, S., Peres, L., Sunye, M., and Almeida, E. (2024). Statistical validation of column matching in the database schema evolution of the brazilian public school census. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 498–509, Porto Alegre, RS, Brasil. SBC.