

# CONSTRUÇÃO DE UM DATASET QUE UNIFICA FATORES DA PRODUÇÃO DE LEITE REGIONAL DO BRASIL

Dados da produção  
de leite brasileira

Aluno: Max Felipe Silva Dos Santos

Professor: Sergio Serra e Jorge Zavaleta





# Introdução e Contexto Econômico

- A relevância do setor agropecuário no Brasil é considerável, atualmente tal setor contribuiu com 23,5% do PIB brasileiro<sup>1</sup>
- Dentro do setor agropecuário, mesmo não sendo uma das principais forças, o setor lácteo gerou cerca de 77,1 bilhões de reais em 2020 (4% do PIB do agronegócio)<sup>2</sup>
- Nesse contexto As regiões Sul e Sudeste (ex: Minas Gerais e Paraná) lideram com 68% da produção nacional.
- Porém regiões como o nordeste vem em crescente aumento em sua produção leiteira
- Dado tais fatores, e as disponibilidades de informações disponibilizadas sobre a produção do leite surge a oportunidade de explorar e agregar informações para entender tendências e correlações regionais na produção de leite.

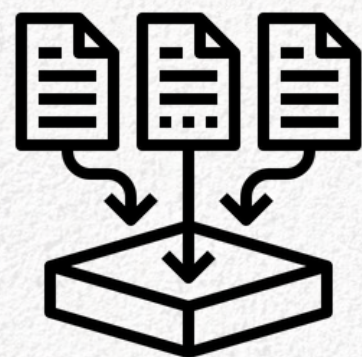
1. PIB do agronegócio registra crescimento de 6,49% no primeiro trimestre de 2025. Disponível em: <https://www.cepea.org.br/br/pib-do-agronegocio-brasileiro.aspx>

2. Grigol., Natália. **Por que monitorar os preços do leite e dos lácteos?** Cepea. Disponível em: <https://www.cepea.org.br/br/por-que-monitorar-os-precos-do-leite-e-dos-lacteos.aspx>



# Problema e Justificativa

- **Dispersão de Dados:** Informações econômicas, climáticas e de produção estão distribuídas pela WEB e não agregadas .
- **Necessidade de Unificação:** É necessário unificar dados para entender como fatores externos influenciam a produção regional.<sup>3</sup>
- **Integridade e Análise:** O crescente volume de dados exige processos de armazenamento e unificação que garantam integridade e facilitem análises estatísticas e de IA.
- **Proveniência:** É crucial garantir a rastreabilidade do pré-processamento e captura dos dados para assegurar confiabilidade<sup>4</sup>.



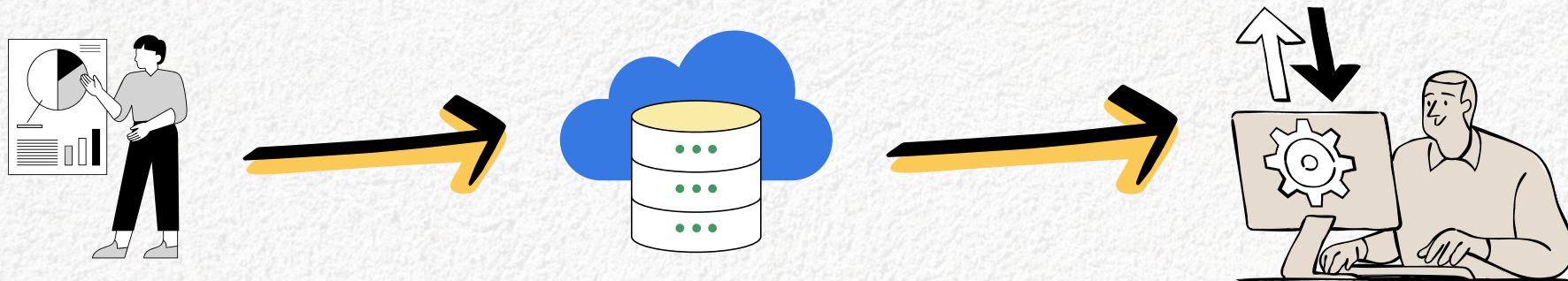
3. CABRERA, Victor E. et al. Data Integration and Analytics in the Dairy Industry: Challenges and Pathways Forward. *Animals*, v. 15, n. 3, p. 329, jan. 2025.

4. GLAVIC, Boris *et al.* Provenance for data mining. *In*: : TaPP'13.USA: USENIX Association, 2 abr. 2013. . Acesso em: 18 nov. 2025



# Objetivos do Trabalho

- Objetivo Geral: Coletar dados dispersos na rede, unificá-los em formato de acessível processamento e criar visualizações intuitivas destes dados.
- Objetivos Específicos:
  - Buscar informações tabulares de produção, economia e clima.
  - Mapear instituições de pesquisa e inovação.
  - Unificar em um dataset único e estabelecer proveniência.
  - Criar apresentações visuais (mapas).





# Trabalhos Relacionados

Foi realizado uma revisão narrativa, buscando encontrar em português e em inglês trabalhos dos quais tratassem do tema do presente trabalho de maneira similar, tentando entender como o trabalho pode se diferenciar.

Essa busca foi realizada através da utilização da seguinte string de busca e suas respectivas versões inglês e português:

**Inglês:** ("data science" OR "machine learning" OR "artificial intelligence" OR "statistical analysis" OR "predictive modeling" OR "big data" OR "data mining" OR "data analytics" OR "quantitative analysis" OR "computational model") AND ("milk")

**português:** ("ciência de dados" OR "aprendizado de máquina" OR "análise de dados" OR "modelagem estatística" OR "análise preditiva" OR "big data" OR "mineração de dados" OR "análise quantitativa") AND ("leite" OR "pecuaria")

**Scielo**



**SOL**



**Google scholar**



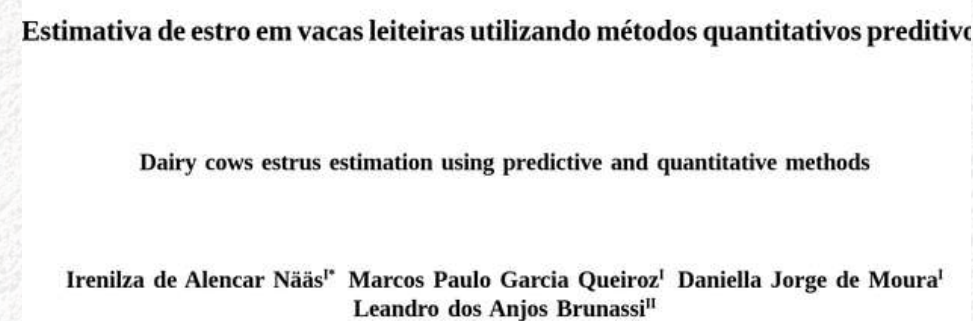


# Resultados de pesquisas Brasileiras

- Foi encontrado um número pequeno de pesquisas brasileiras que diretamente analisam a produção de leite, tais pesquisas focam na utilização de algoritmos e modelos computacionais



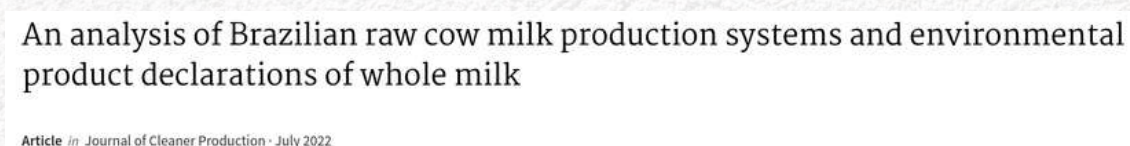
Analisa o consumo, produção e **tendências** na produção do leite porém não aplica uma análise estatística com os dados



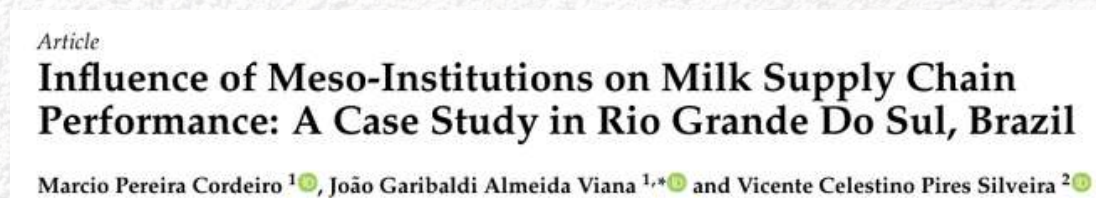
Utiliza modelos preditivos e algoritmos baseados em Inteligência artificial, porém foca a análise em **prever o estro**



Realizou uma análise na produção de leite na região sul do Brasil, com foco em **clusterização e quociente de localização** (analisar especializações de microregiões)



Faz uma **Avaliação do Ciclo de Vida (ACV)** dos sistemas de produção de leite cru de vaca entre o estado do paran  e de Minas gerais utilizando dados obtidos de entrevistas



Faz uma pesquisa exploratoria e descritiva de dois munic pios do RS utilizando dados do IBGE, ap s essa an lise foi realizada uma an lise do **processo de produ  o e industrializa  o do leite**

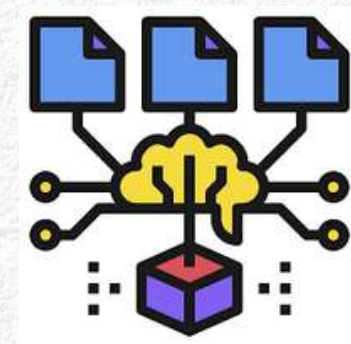




# Discussão sobre os artigos brasileiros

Percebe-se que nos artigos Nacionais a utilização de algoritmos de IA e modelos estatísticos estão associados a definir e identificar características de qualidade na produção do leite apenas um artigo que analise diferenças entre diferentes regiões na produção foi identificado

Não encontrado na pesquisa realizada um artigo que trabalhe a produção de leite pensando por cada região brasileira, e ou artigo que trabalhe utilizando dados variados disponíveis sobre as características de cada região quanto a sua capacidade de produção ou fatores relacionados.





# Resultados de pesquisas Estrangeiras

- Foi encontrado um número considerável de pesquisas estrangeiras que realizam análise preditiva de dados levando em conta fatores de produção do leite



Big data analytics for empowering milk yield prediction in dairy supply chains

Publisher: IEEE [Cite This](#) [PDF](#)

Proponhe a criação de uma ferramenta de produção de leite que facilita ao **produtor analisar e poder prever a produção de leite** em uma vaca ou um grupo de vacas utilizando ML

Original papers

Comparison of deep learning models for milk production forecasting at national scale

Luigi Cesarini <sup>a</sup>, Rui Gonçalves <sup>b</sup>, Mario Martina <sup>a</sup>, Xavier Romão <sup>c</sup>, Beatrice Monteleone <sup>a</sup>, Fernando Lobo Pereira <sup>b</sup>, Rui Figueiredo <sup>d</sup>

Analise a validade da **aplicação de diferentes modelos de machine learning**, utilizando também deep learning para previsão da produção de leite no contexto de dados escassos

Home > Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 1 > Conference paper

**Machine Learning Approaches for Dairy(Milk) Quality Assurance**

Testando diversos modelos de ML para **prever a qualidade do leite** dado informações de diversos parâmetros do leite

PAPER • OPEN ACCESS

Smart modelling of dairy milk production with machine learning

To cite this article: T K Samson *et al* 2023 *IOP Conf. Ser.: Earth Environ. Sci.* 1219 012008

Testam diferentes algoritmos de inteligência artificial para **prever a produção de leite** utilizando dados obtidos localmente pela universidade de Bowen

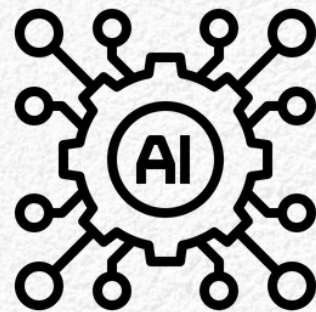




# Discussão sobre os artigos Estrangeiros

Apesar de diferentes artigos buscaram utilizar modelos ML para construção de modelos de predição com o foco de analisar a produção do leite, não foi encontrado um artigo que realiza comparativamente a análise entre diferentes regiões, nem levantam dados das características de produção destas regiões e como influenciam na produção do leite.

Um dos possíveis motivos encontrados para que pesquisas estrangeiras foquem pouco na análise comparativa entre regiões pode estar relacionada a que tais fatores já são mais extensamente analisados pelo ministério da agricultura (EUA)<sup>5</sup> e por instituições privadas como a CLAL<sup>6</sup>



5. <https://www.ers.usda.gov/data-products/dairy-data>

6. [https://www.clal.it/en/index.php?section=produzioni\\_usa\\_latte\\_bovino](https://www.clal.it/en/index.php?section=produzioni_usa_latte_bovino)



# Conclusão da revisão



Existe lacunas para análise da produção de leite no Brasil a nível regional e análise de diferentes características regionais, tais como econômicas e climáticas, como elas influenciam na produção do leite e como elas se diferem, foi encontrado que diferentes algoritmos de ML já foram testados e validados, porém foram aplicados em parâmetros de qualidade do leite e ou em dados específicos de pequenas regiões.

É possível através da unificação de diversos elementos da produção de leite levantar uma base de dados sólida que permite com que sejam aplicadas técnicas e métodos para análise comparativa da produção de leite no Brasil. Algo semelhante já realizado extensivamente pelo setor privado no exterior.





# Metodologia – Fontes de Dados

- **Clima:** Dados de temperatura e precipitação (2000-2023) obtidos do *INMET/BDMET*.
- **Economia (Inflação):** Variação do IPCA obtida via IBGE.
- **Produção total e custo de produção:** Média de produção e preços ao produtor obtidos no *CILeite*.
- **preços de comercialização interna:** Preço do leite comercializado entre produtores *CEPEA*
- **Instituições de Pesquisa:** Localização geográfica de laboratórios da Embrapa e Rede Brasileira de Qualidade do Leite (RBQL).





# Metodologia – Processamento e Limpeza

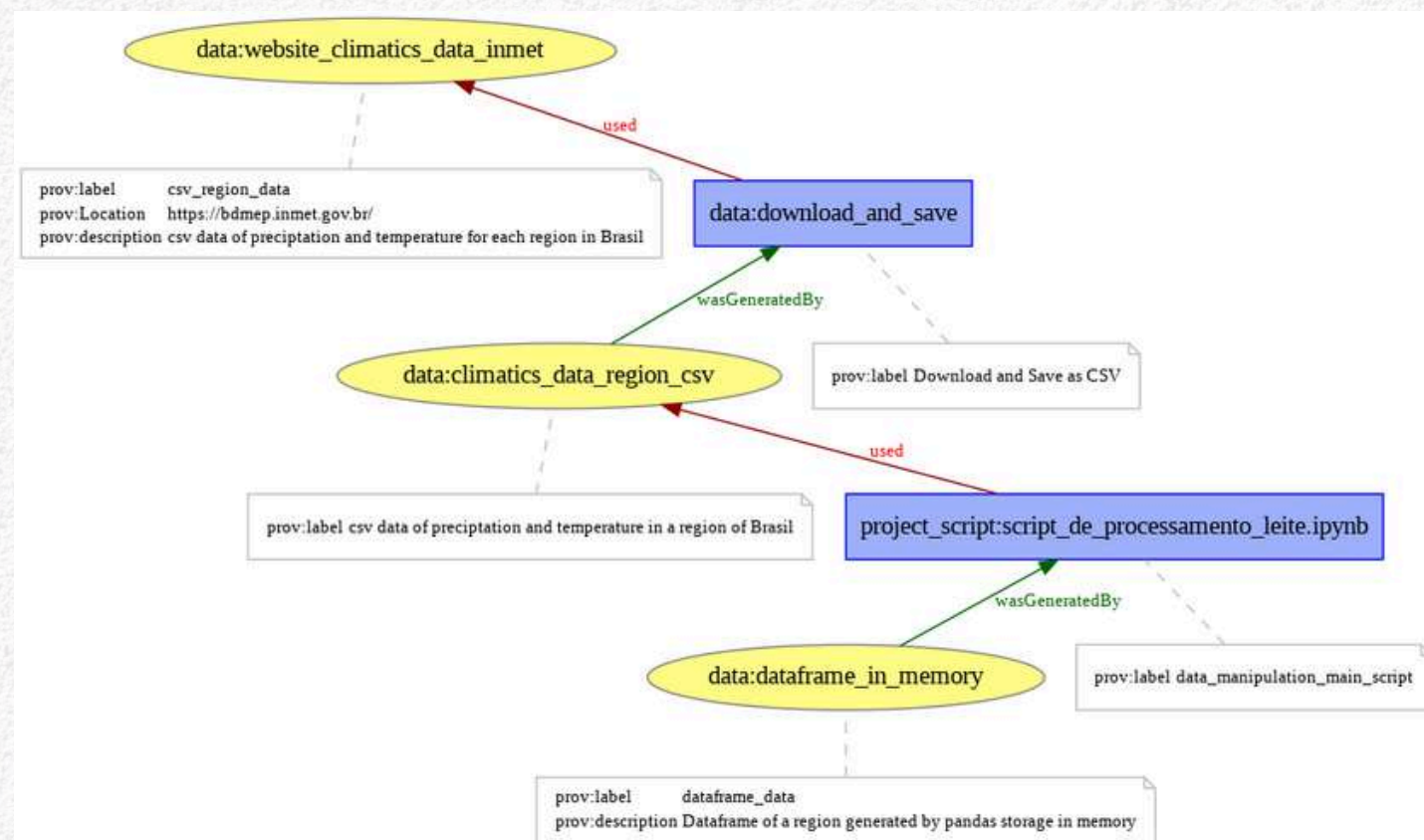
- **Ferramentas:** Uso de Python (versão 3.12.11) e bibliotecas como Pandas no ambiente Google Colab (com versionamento pré-definido).
- **Imputação de Dados:** Uso do algoritmo k-nearest neighbors (Scikit-learn) para preencher lacunas de dados<sup>7</sup>.
- **Transformação:** Agrupamento anual de dados mensais, anuais e trimestrais por região e suas respectivas formatações numéricas.
- **Unificação:** Junção dos dados separados por região em um único dataset.

7. AHN, Hyun; SUN, Kyunghee; KIM, Kwanghoon. Comparison of Missing Data Imputation Methods in Time Series Forecasting.



# Proveniência dos Dados

- **Ferramenta:** Utilização da biblioteca prov (W3C) em Python.
- **Rastreabilidade:** Mapeamento de todo o fluxo, desde a obtenção até a geração do dataset final.
- **Saída:** Geração de grafos e descrições textuais que documentam a linhagem dos dados.

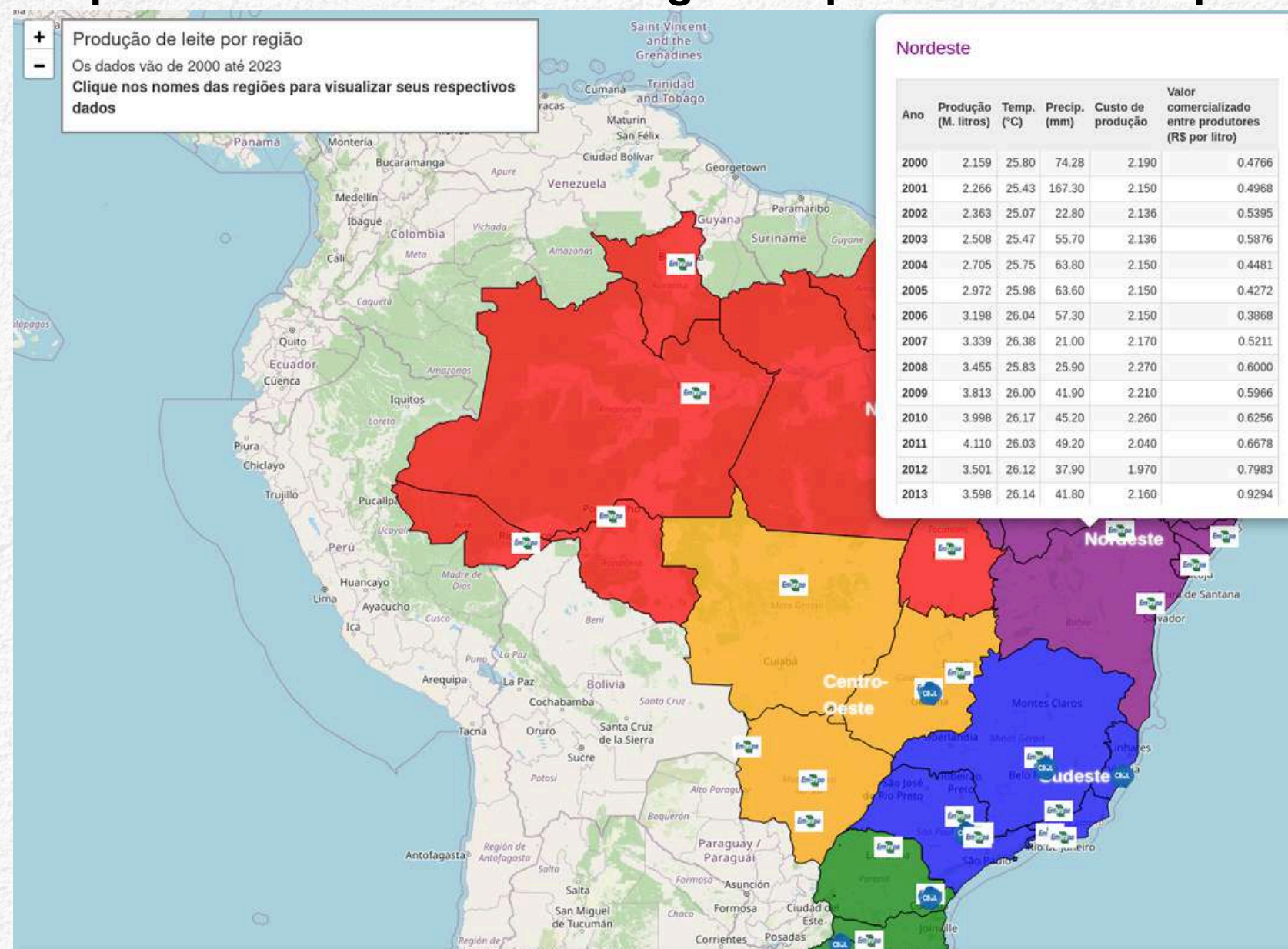


Exemplo de grafo de proveniência



# Resultados

- **Mapa Interativo:** Criação de uma apresentação visual com a biblioteca folium.
- **Funcionalidades:**
  - Tabelas dinâmicas com dados de cada região.
  - Uso de malhas geográficas do IBGE (API V3).
  - Marcadores geográficos das instituições de pesquisa (Sispel e Embrapa Gado de Leite).
- **Disponibilidade:** Dados e códigos disponibilizados na plataforma Zenodo





# Discussão e Conclusões

- **Centralização na Pesquisa em produção de leite no Brasil:** Concentração dos laboratórios e instituições de pesquisa nas regiões Sul e Sudeste.
- Poucos dados disponíveis sobre a região Norte e sua produção de leite, o que pode prejudicar visualização sobre os avanços ou retrações da região quanto a sua capacidade de produção
- **Crescimento do Nordeste:** Aumento de 191% na produção, sugerindo potencial para se tornar um novo epicentro mesmo com menos instituições de pesquisa.
- **Utilidade do Dataset:** O artefato permite que leitores avaliem o estado da produção por região de forma acessível.





# Limitações e Trabalhos Futuros

- **Limitações:** Falta de dados a nível municipal, dados referentes a região norte e ausência de variáveis de qualidade do leite (doenças, parâmetros físico-químicos).
- **Trabalhos Futuros:**
  - Expansão para incluir dados de qualidade e maior granularidade geográfica.
  - Aplicação de Inteligência Artificial para predição de tendências futuras.
  - Aplicar uma análise comparativa entre os dados encontrados nas diferentes regiões do Brasil com regiões de outros Países visualizar se existem fatores que variam de nação para nação e que determinam o aumento na produção de leite.





# OBRIGADO PELA ATENÇÃO

PERGUNTAS?

