Análise de Dados Climáticos e Históricos da Produção de Algodão no Brasil

Raphael Mauricio Sanches de Jesus[1]

[1]Programa de Pós-Graduação em Informática – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brazil

raphael.mauricio@gmail.com

Abstract. *This paper presents an analysis of climatic and historical data related to cotton production in Brazil, integrating advanced techniques for data preprocessing, visualization, and predictive modeling. The PROV methodology was applied to ensure reproducibility and process transparency. Results reveal significant correlations between climatic variables and productivity, providing actionable insights for agricultural decision-making.*

Resumo. *Este trabalho apresenta uma análise de dados climáticos e históricos relacionados à produção de algodão no Brasil, integrando técnicas avançadas de pré-processamento, visualização e modelagem preditiva. A metodologia PROV foi empregada para garantir a reprodutibilidade e a transparência do processo. Os resultados revelam correlações significativas entre variáveis climáticas e produtividade, oferecendo insights aplicáveis para a tomada de decisão agrícola.*

1. Introdução

A agricultura desempenha um papel central na economia brasileira, sendo responsável por uma significativa parcela do Produto Interno Bruto (PIB) e pela geração de empregos no país. O Brasil é reconhecido como um dos maiores produtores mundiais de grãos, fibras e outros produtos agrícolas, com destaque para culturas como soja, milho, arroz, feijão e algodão. Esse protagonismo está diretamente ligado à capacidade do setor de responder a desafios climáticos e de mercado, aliados a avanços tecnológicos e práticas agrícolas cada vez mais eficientes.

De acordo com o levantamento mais recente da Companhia Nacional de Abastecimento (Conab), a produção total de grãos da safra 2023/2024 foi estimada em 298,41 milhões de toneladas, representando uma redução de 21,4 milhões de toneladas em relação à safra anterior. As adversidades climáticas, como irregularidade nas chuvas e excesso de precipitação em algumas regiões, foram os principais fatores que contribuíram para a queda na produtividade de culturas importantes como soja e milho. Por outro lado, culturas como algodão e arroz apresentaram aumentos significativos na produção, impulsionados pelo crescimento da área cultivada [de Abastecimento (Conab) 2024b].

A Conab realiza o acompanhamento constante da safra de grãos no Brasil, monitorando as condições de desenvolvimento das principais culturas agrícolas. Esse acompanhamento inclui a produção mensal do Boletim de Acompanhamento da Safra Brasileira de Grãos e do Boletim de Monitoramento Agrícola, documentos que fornecem

informações cruciais para agentes envolvidos nos desafios da agricultura, segurança alimentar e abastecimento nacional. Além de permitir uma análise detalhada das condições do setor, essas publicações servem como base para o monitoramento e formulação de políticas agrícolas e de abastecimento [de Abastecimento (Conab) 2024a].

Neste cenário, a análise e o processamento de dados agrícolas e climáticos tornam-se essenciais para compreender as tendências da produção e mitigar os impactos de fatores externos. Este trabalho apresenta uma abordagem baseada em ciência de dados, utilizando modelos de rastreabilidade e proveniência, com o objetivo de aprimorar o entendimento sobre o desempenho das culturas agrícolas e fornecer subsídios para a tomada de decisão no setor. A metodologia PROV foi empregada para garantir a rastreabilidade dos dados e a transparência no processo analítico, alinhando-se às boas práticas científicas e promovendo maior confiança nos resultados apresentados.

Com a integração de dados históricos e climáticos e a aplicação de técnicas avançadas de análise, esperamos contribuir para o desenvolvimento de estratégias mais resilientes e sustentáveis, fortalecendo o papel do Brasil como um líder global na produção agrícola.

2. Descrição dos Dados

Foram utilizados dois *datasets* principais: o **Dataset Histórico de Algodão** (1976-2024) e o **Dataset de Variáveis Climáticas** (2000-2024). Os dados foram integrados para explorar relações entre produtividade agrícola e variáveis ambientais.

2.1. Dataset Histórico de Algodão

O *Dataset Histórico de Algodão* abrange informações detalhadas sobre a produção dessa cultura no Brasil, com registros que vão do período de 1976/1977 até 2023/2024. Este conjunto de dados foi obtido por meio de registros disponibilizados pela Companhia Nacional de Abastecimento (Conab). A seguir, descrevemos as principais colunas e informações contidas no *dataset*:

- **Área plantada (*hectares*)**: Representa o total de terras destinadas ao cultivo de algodão em diferentes estados e regiões.
- **Produtividade do Algodão em Caroço (*kg/ha*)**: Indica a eficiência da produção antes da separação da pluma e do caroço.
- **Produtividade da Pluma (*kg/ha*)**: Refere-se à quantidade de pluma de algodão produzida por hectare.
- **Produtividade do Caroço de Algodão (*kg/ha*)**: Indica a eficiência na produção do caroço, subproduto do processamento do algodão.
- **Rendimento da Pluma (*%*)**: Mede a proporção de pluma em relação ao total produzido, indicando a qualidade da fibra.
- **Produção Total em Caroço (*toneladas*)**: Quantidade total de algodão em caroço produzida por estado, região ou nacionalmente.
- **Produção de Pluma (*toneladas*)**: Volume de pluma de algodão obtido após o beneficiamento.
- **Produção de Caroço de Algodão (*toneladas*)**: Quantidade de caroço resultante do processamento da produção total.

Esses dados foram organizados em planilhas por ano-safra e permitem análises em diferentes níveis de agregação: estadual, regional e nacional. O arquivo está no formato Excel e contém múltiplas abas que correspondem a diferentes períodos e métricas. Essa riqueza de informações possibilita o estudo detalhado de tendências de produção ao longo do tempo.

2.2. Dataset de Variáveis Climáticas

O *Dataset de Variáveis Climáticas* foi extraído do repositório disponível no Kaggle, intitulado **"Brazil Weather Information by INMET"** [de Meteorologia (INMET) 2024]. Esse conjunto de dados consolidou informações coletadas pelo Instituto Nacional de Meteorologia (INMET), por meio do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP), abrangendo o período de 2000 até 2024.

Os dados são derivados de estações meteorológicas automáticas distribuídas em diversas localidades do Brasil e foram organizados em diferentes níveis de granularidade, incluindo séries temporais anuais e resumos agregados. A seguir, destacamos os principais elementos do *dataset*:

Colunas do *Dataset* Agregado (weather_sum_all)

O *dataset* utilizado neste trabalho é a versão resumida (*summarized data*), que apresenta os seguintes atributos:

- **ESTACAO (*Station*)**: Identificador da estação meteorológica onde os dados foram coletados.
- **DATA (*Date*)**: Data do registro no formato *YYYY-MM-DD*.
- **rain_max (*mm*)**: Precipitação máxima registrada no dia (mm).
- **rad_max (*KJ/m²*)**: Radiação global máxima registrada no dia (KJ/m²).
- **temp_avg (°*C*)**: Temperatura média diária do ar (°C).
- **temp_max (°*C*)**: Temperatura máxima registrada no dia (°C).
- **temp_min (°*C*)**: Temperatura mínima registrada no dia (°C).
- **hum_max (*%*)**: Umidade relativa máxima registrada no dia (
- **hum_min (*%*)**: Umidade relativa mínima registrada no dia (
- **wind_max (*m/s*)**: Velocidade máxima do vento no dia (m/s).
- **wind_avg (*m/s*)**: Velocidade média do vento no dia (m/s).

Origem e Processamento dos Dados

Os dados brutos foram coletados diretamente do INMET e estão disponíveis no BDMEP, enquanto o autor do *dataset* no Kaggle unificou as informações em quatro formatos:

- **weather_YYYY**: Dados meteorológicos brutos, organizados por ano.
- **weather_sum_YYYY**: Dados resumidos, organizados por ano, com métricas calculadas (máximas, mínimas e médias).
- **weather_sum_all**: Dados resumidos consolidados, contendo todas as informações disponíveis em um único arquivo.
- **Station Information**: Informações sobre as estações meteorológicas utilizadas e seus últimos registros.

A versão utilizada neste trabalho, **weather_sum_all**, permite uma análise agregada de longo prazo, ideal para identificar padrões climáticos históricos e anomalias.

Relevância dos Dados Climáticos para o Estudo

Os dados climáticos fornecem informações essenciais sobre os fatores ambientais que influenciam a produtividade agrícola, especialmente para a cultura do algodão. As principais variáveis destacadas, como precipitação, radiação, temperaturas e umidade, estão diretamente relacionadas ao desenvolvimento das plantas.

Além disso, a granularidade diária e o período histórico de 24 anos (2000–2024) permitem explorar relações complexas, como:

- Correlações entre eventos climáticos extremos (secas, chuvas intensas) e reduções de produtividade.
- Impacto da radiação solar e das temperaturas extremas sobre o rendimento das lavouras.
- Identificação de tendências de longo prazo, como o aumento da temperatura média ou mudanças nos padrões de precipitação.

O cruzamento dos dados climáticos com os históricos de produção do algodão é fundamental para compreender as dinâmicas de produção agrícola no Brasil e criar modelos preditivos robustos para futuras safras.

Disponibilidade e Referência

O conjunto de dados utilizado neste estudo pode ser acessado no Kaggle no seguinte endereço: `https://www.kaggle.com/datasets/gregoryoliveira/brazil-weather-information-by-inmet`. Mais detalhes sobre as estações meteorológicas e o catálogo completo estão disponíveis no portal do INMET.

A integração desses dados é fundamental para compreender a relação entre as condições climáticas e a produção agrícola no Brasil, especialmente para a cultura do algodão.

2.3. Importância dos Dados

Esses *datasets* fornecem uma base sólida para a análise dos impactos das condições climáticas na produção agrícola. Enquanto o *Dataset Histórico de Algodão* captura o desempenho produtivo ao longo do tempo, o *Dataset de Variáveis Climáticas* oferece uma visão detalhada dos fatores ambientais que influenciam diretamente essa produção. Juntos, esses dados permitem explorar tendências históricas e prever cenários futuros para o setor agrícola no Brasil.

3. Métodos de Pré-Processamento

Para garantir a qualidade dos dados e a confiabilidade das análises, aplicamos uma série de técnicas de pré-processamento nos dois *datasets* principais: o de produção de algodão e o de variáveis climáticas. Esses métodos são essenciais para eliminar inconsistências, normalizar os dados e preparar os *datasets* para a análise exploratória e modelagem subsequentes. Abaixo, descrevemos os principais passos realizados:

3.1. Pré-Processamento dos Dados de Produção de Algodão

Técnicas aplicadas incluem:

1. Remoção de duplicatas e normalização dos dados.
2. Conversão de formatos inconsistentes (e.g., datas e unidades).
3. Classificação das estações do ano com base em períodos climáticos.

Esses passos garantiram qualidade e consistência nos *datasets*, viabilizando análises precisas.

O *dataset* de produção de algodão foi carregado de um arquivo no formato Excel. O processo de tratamento incluiu as seguintes etapas:

1. **Carregamento e Organização:** Os dados foram carregados utilizando a biblioteca `pandas`, configurando as colunas de acordo com os anos da série histórica (1976 a 2024).
2. **Renomeação de Colunas:** Criamos nomes dinâmicos para as colunas, representando os anos da série histórica.
3. **Remoção de Valores Agregados:** Linhas contendo totais nacionais ou regionais (`BRASIL`, `NORTE/NORDESTE`) foram excluídas, garantindo que os dados permanecessem desagregados por estado ou região.
4. **Transformação de Dados:** O *dataset* foi transformado de um formato largo para o formato longo (*wide-to-long*), onde cada linha representa uma combinação de ano, região/estado e área plantada.
5. **Conversão e Tratamento de Dados:** Colunas relevantes foram convertidas para tipos numéricos, enquanto valores ausentes ou inválidos foram removidos.
6. **Verificação de Consistência:** Foram realizadas verificações para garantir que apenas registros válidos fossem utilizados, eliminando inconsistências e duplicatas.

O resultado final foi um *dataset* em formato longo, ideal para análises temporais e regionais, contendo as colunas `Região/UF`, `Ano` e `Área Plantada`.

3.2. Pré-Processamento dos Dados Climáticos

O *dataset* de variáveis climáticas foi carregado de um arquivo no formato CSV. O tratamento realizado incluiu as etapas abaixo:

1. **Carregamento Inicial:** Os dados foram carregados com a biblioteca `pandas`, e a coluna de datas (`DATA (YYYY-MM-DD)`) foi convertida para o tipo `datetime`, permitindo operações temporais.
2. **Extração de Informações Temporais:** Foram extraídos os anos e meses de cada registro, criando as colunas `Ano` e `Mes`.
3. **Definição de Estações do Ano:** Com base nos meses, as estações do ano foram definidas utilizando `bins` para classificação. As categorias criadas foram `Verão`, `Outono`, `Inverno` e `Primavera`.
4. **Verificação de Dados Faltantes:** Checamos a existência de valores nulos nas colunas criadas para estações do ano, garantindo que todos os meses estivessem corretamente mapeados.
5. **Visualização Inicial:** Uma pré-visualização dos dados foi realizada para verificar a consistência e identificar possíveis problemas antes da análise.

O resultado foi um *dataset* que inclui informações temporais detalhadas e categorizadas, contendo as seguintes colunas principais:

- **DATA:** Data no formato `YYYY-MM-DD`.
- **Ano e Mes:** Informações extraídas da coluna `DATA`.
- **Estacao:** Estação do ano (Verão, Outono, Inverno ou Primavera).
- **Variáveis Climáticas:** `rain_max`, `rad_max`, `temp_avg`, `temp_max`, `temp_min`, `hum_max`, `hum_min`, `wind_max` e `wind_avg`.

3.3. Normalização e Tratamento Geral

Além das etapas específicas para cada *dataset*, foram aplicadas as seguintes transformações gerais para ambos:

1. **Remoção de Duplicatas:** Linhas redundantes foram eliminadas para evitar viés nas análises.
2. **Normalização de Caracteres:** Caracteres especiais em nomes de colunas ou valores categóricos foram tratados usando a biblioteca `unidecode`, garantindo consistência entre as entradas.
3. **Conversão de Tipos:** Todas as colunas numéricas foram convertidas para `float` ou `int`, eliminando problemas com formatação inconsistente.

3.4. Importância do Pré-Processamento

O pré-processamento aplicado assegurou a integridade e a qualidade dos dados, permitindo a integração entre os dois *datasets*. A preparação cuidadosa dos dados reduz o risco de erros nas análises e facilita a aplicação de técnicas estatísticas e modelos preditivos confiáveis.

4. Metodologia de Proveniência (PROV)

A proveniência de dados desempenha um papel fundamental no projeto, garantindo transparência, rastreabilidade e reprodutibilidade dos resultados. O modelo de Proveniência (PROV) desenvolvido utiliza o padrão proposto pelo W3C, que permite documentar de maneira formal as relações entre agentes, entidades e atividades envolvidas.

4.1. Descrição do Modelo

Conforme ilustrado na Figura 1, o modelo PROV do projeto é composto pelos seguintes elementos:

- **Agentes:** Representam os responsáveis pelo desenvolvimento e execução do projeto. Os agentes incluem:
  - *UFRJ:* Representa a instituição principal.
  - *PPGI:* Programa de Pós-Graduação em Informática, ao qual a disciplina MAI712 está vinculada.
  - *MAI712:* Disciplina responsável pelo projeto.
  - *Raphael:* Desenvolvedor principal do projeto.
  - *getProv.py:* Script utilizado para gerar o modelo PROV.
- **Entidades:** Os objetos de dados utilizados ou produzidos durante o projeto, incluindo:
  - *dados-algodao:* Dataset histórico de algodão.

  - *dados-clima:* Dataset de variáveis climáticas.
  - *seasonal_trends, regional_potential, climatic_influences, historical_trends, predicted_areas:* Resultados gerados pela análise.
  - *visualizacao-output:* Resultado da etapa de visualização, como gráficos e mapas.
- **Atividades:** As ações realizadas no pipeline analítico, incluindo:
  - *processamento-dados:* Atividade de processamento inicial dos *datasets*.
  - *visualizacao:* Atividade de criação de gráficos e representações visuais.

4.2. Fluxo de Dados e Dependências

O fluxo de dados é iniciado com os *datasets* brutos de algodão e variáveis climáticas, que são utilizados na atividade de processamento. Essa atividade gera entidades de saída representando os principais resultados analíticos, que por sua vez servem como insumos para a atividade de visualização. O produto final da etapa de visualização é uma coleção de gráficos e mapas interativos, representados pela entidade *visualizacao-output*.

4.3. Justificativa do Uso de Proveniência

O uso do modelo PROV no projeto é motivado por várias razões:

1. **Transparência:** O modelo permite documentar todas as etapas do processo analítico, facilitando a compreensão pelos stakeholders.
2. **Reprodutibilidade:** Pesquisadores e desenvolvedores podem reproduzir as análises, garantindo consistência nos resultados.
3. **Auditoria:** Em caso de dúvidas ou inconsistências, o modelo permite rastrear cada etapa e identificar possíveis fontes de erro.

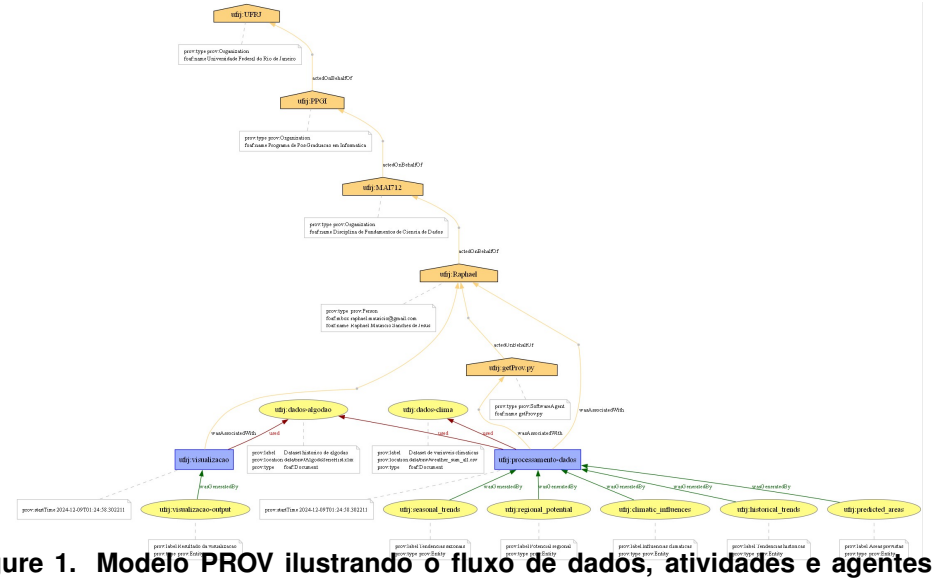

Figure 1. Modelo PROV ilustrando o fluxo de dados, atividades e agentes no projeto.

5. Visualização e Análise

As visualizações geradas neste estudo têm como objetivo revelar tendências históricas, padrões climáticos e a influência de variáveis climáticas na área plantada de algodão no Brasil. A seguir, detalhamos cada figura e suas respectivas análises.

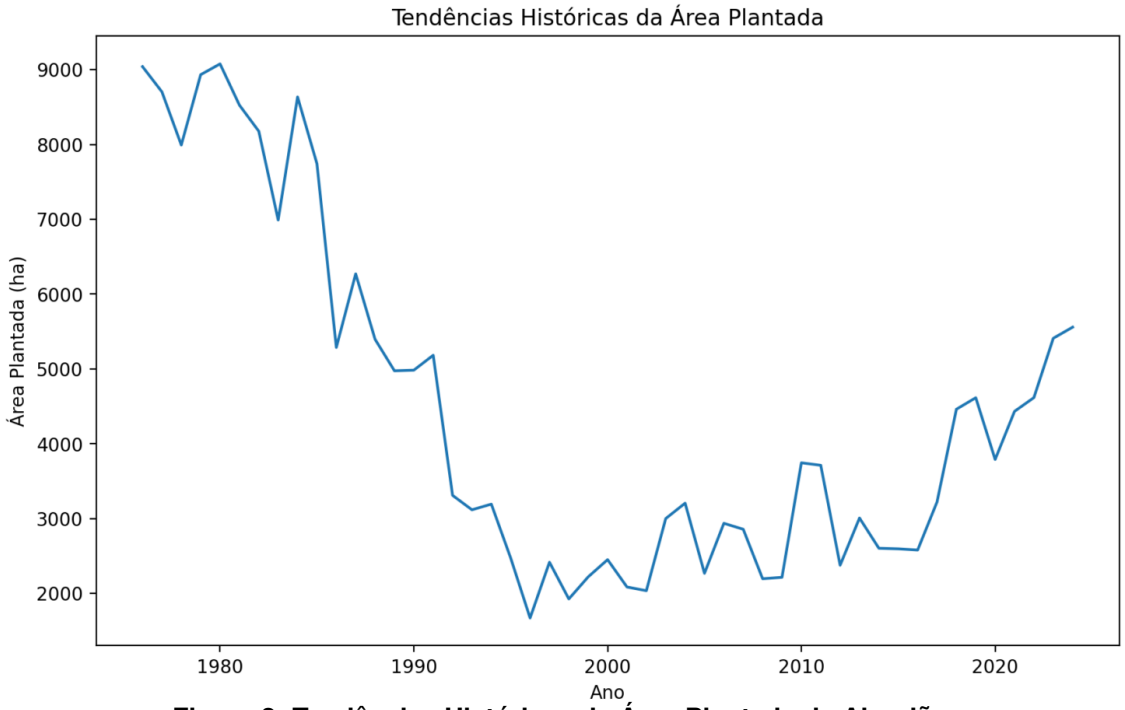

Figure 2. Tendências Históricas da Área Plantada de Algodão.

A Figura 2 apresenta a evolução da área plantada de algodão ao longo das décadas. Observa-se uma redução significativa entre os anos 1980 e 2000, seguida por uma recuperação gradual a partir de 2010. Esse comportamento está associado a fatores econômicos, mudanças climáticas e avanços tecnológicos, como o uso de sementes geneticamente modificadas.

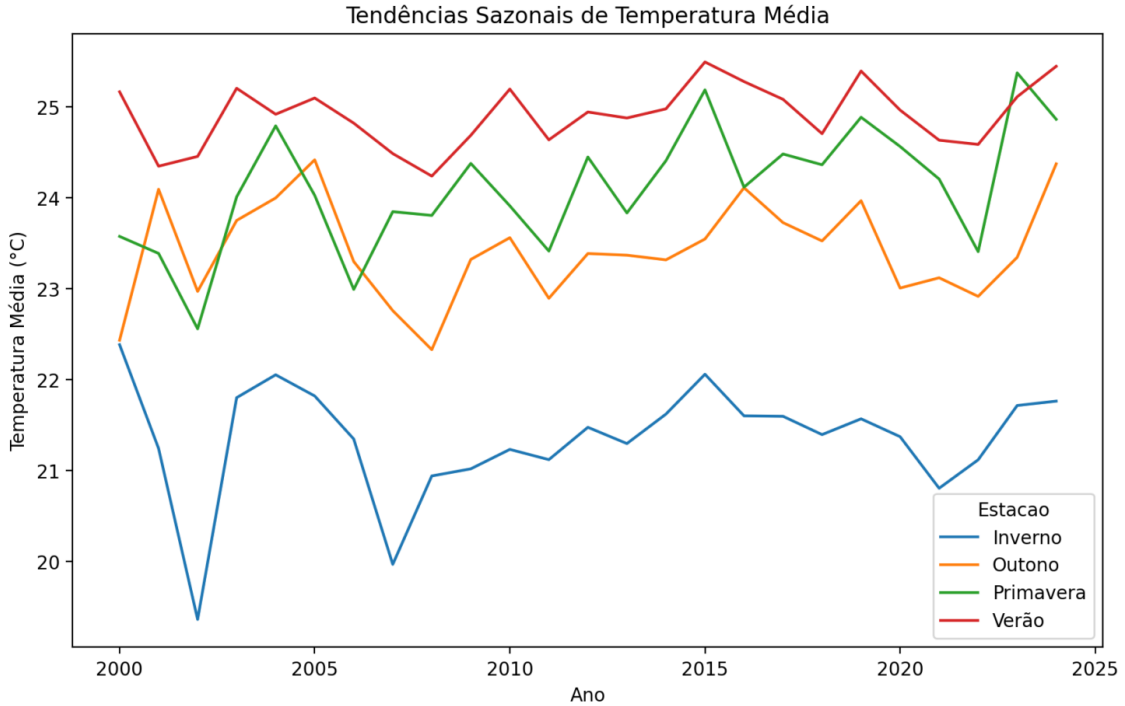

Figure 3. Tendências Sazonais de Temperatura Média.

A Figura 3 destaca as tendências sazonais de temperatura média ao longo dos anos. Verifica-se um aumento na temperatura média em todas as estações, especialmente no verão, o que pode impactar diretamente no ciclo de crescimento do algodão. Estas mudanças climáticas ressaltam a importância de selecionar períodos adequados para o plantio.

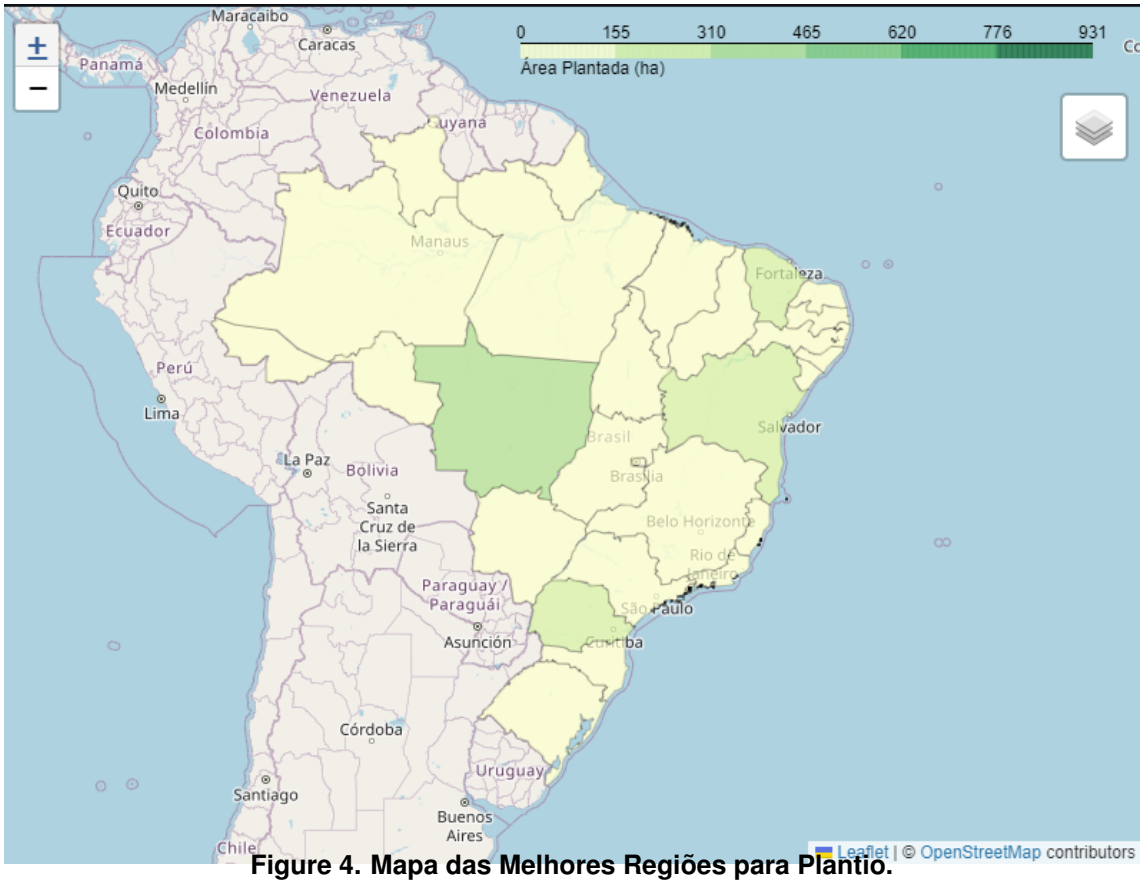

Figure 4. Mapa das Melhores Regiões para Plantio.

A Figura 4 apresenta um mapa interativo das áreas plantadas de algodão no Brasil, destacando as regiões com maior potencial para cultivo. Estados como Mato Grosso e Bahia apresentam áreas significativamente maiores e mais consistentes, indicando sua importância estratégica para a produção nacional.

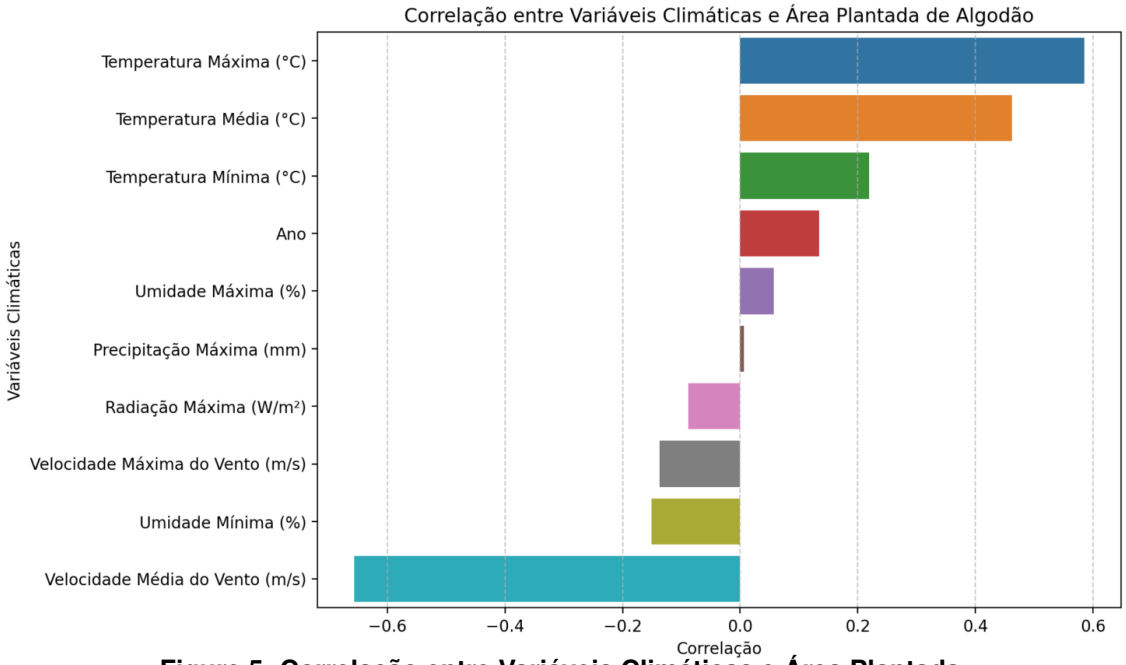

Figure 5. Correlação entre Variáveis Climáticas e Área Plantada.

A Figura 5 ilustra a correlação entre variáveis climáticas, como temperatura máxima e precipitação, e a área plantada. Observa-se que a temperatura máxima tem a maior influência positiva, enquanto a umidade relativa mínima apresenta correlação negativa. Essas informações são cruciais para prever e otimizar a produção de algodão.

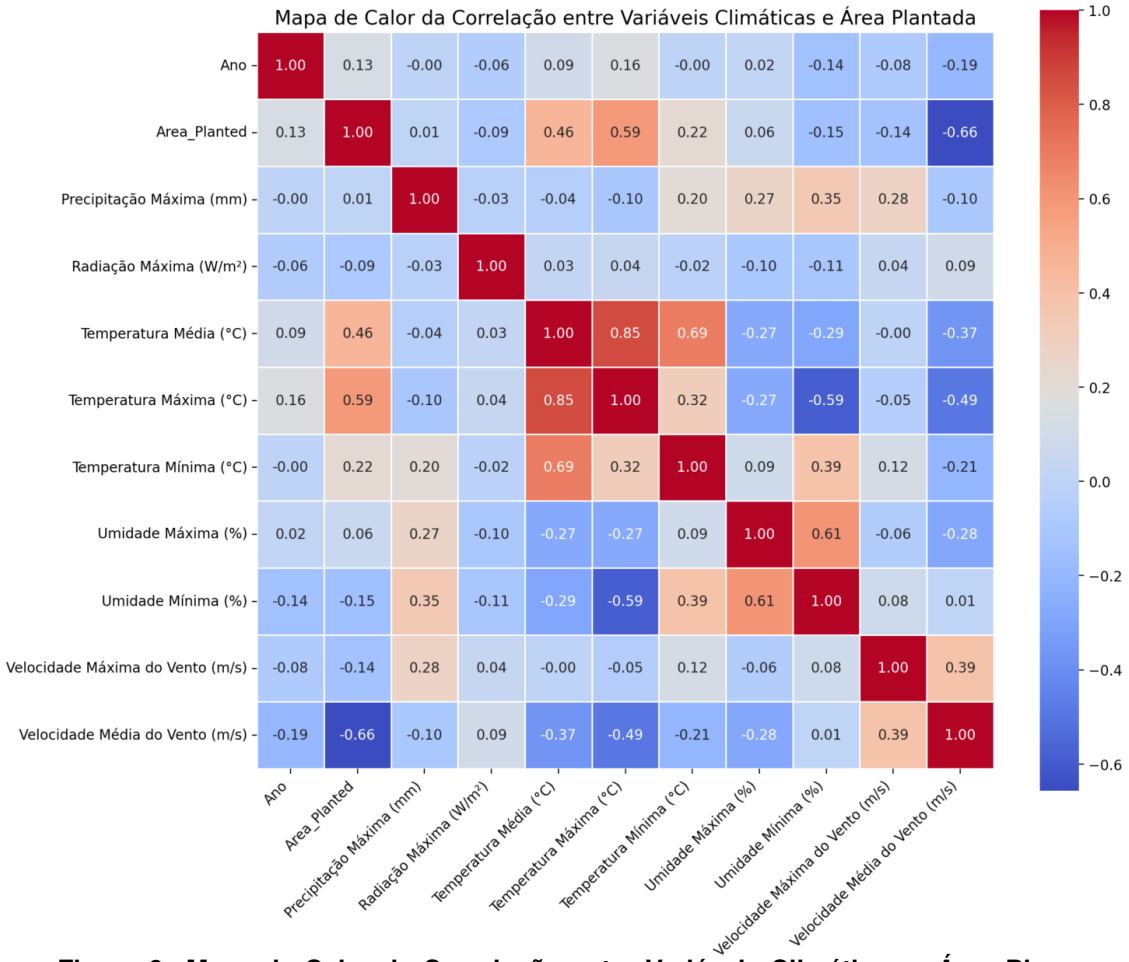

Figure 6. Mapa de Calor da Correlação entre Variáveis Climáticas e Área Plantada.

A Figura 6 complementa a análise da Figura 5, exibindo um mapa de calor que reforça as relações estatísticas entre as variáveis climáticas e a área plantada. Este tipo de visualização facilita a identificação de padrões significativos.

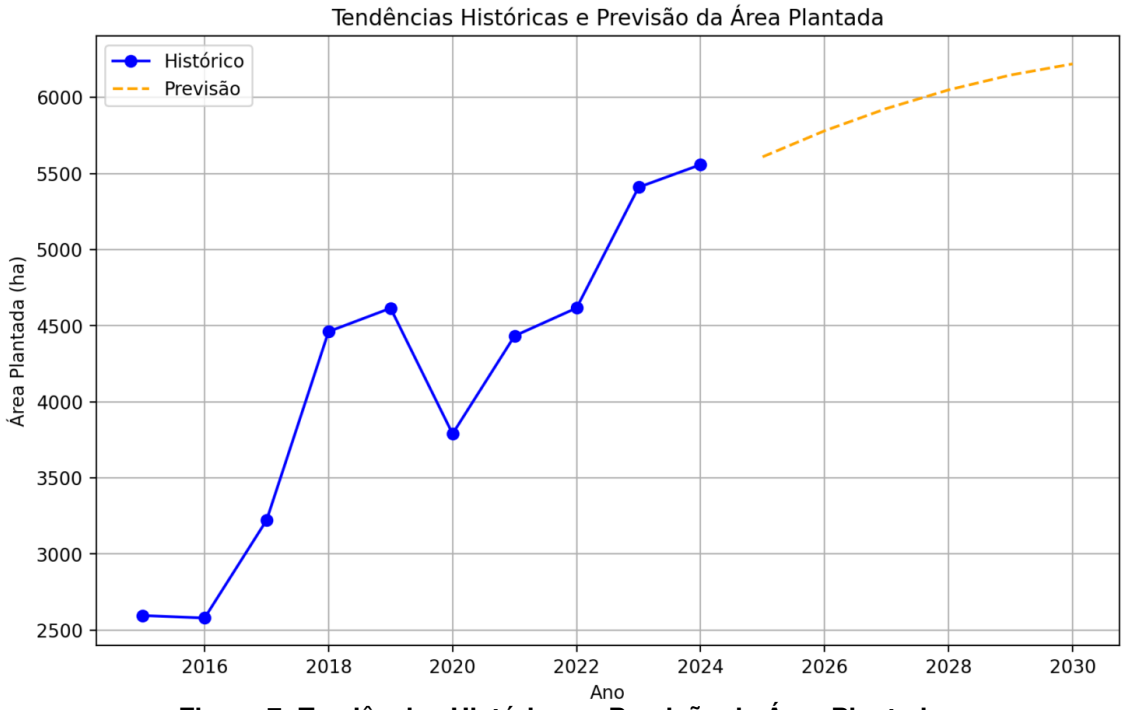

Figure 7. Tendências Históricas e Previsão da Área Plantada.

Por fim, a Figura 7 apresenta uma projeção da área plantada com base em tendências históricas. A análise prevê um aumento gradual na área plantada nos próximos anos, caso as condições climáticas e econômicas permaneçam estáveis. Essa previsão pode ser utilizada para planejamento estratégico e formulação de políticas públicas.

6. Resultados

Os principais resultados deste estudo incluem:

- Melhores períodos para plantio: As análises sugerem que as estações de Primavera e Verão são ideais para o plantio de algodão. Isso se deve às temperaturas médias elevadas e consistentes, radiação solar intensa, e níveis moderados de precipitação, proporcionando condições favoráveis ao desenvolvimento das plantas.
- Regiões de maior potencial produtivo: As regiões Nordeste e Centro-Oeste lideram em potencial para o cultivo de algodão. No Nordeste, destaca-se a Bahia, com infraestrutura e tecnologia avançadas, além de condições climáticas controladas. No Centro-Oeste, estados como Mato Grosso apresentam expansão significativa devido à ampla disponibilidade de terras e uso de tecnologias mecanizadas.
- Impactos climáticos: Identificamos variáveis climáticas com maior influência na área plantada, como: - Temperatura média (correlação de 0,46): Essencial para climas quentes e estáveis. - Temperatura máxima (0,59): Indicando a importância de dias quentes. - Velocidade média do vento (-0,66): Revelando o impacto negativo de ventos excessivos. - Precipitação máxima (0,35): Demonstrando a necessidade de equilíbrio na umidade.
- Tendências históricas: Observa-se um crescimento consistente da área plantada nos últimos anos, impulsionado pela adoção de tecnologias agrícolas modernas,

como sementes geneticamente modificadas e sistemas eficientes de irrigação, além do aumento no valor de mercado do algodão.
- Previsões futuras: Os modelos preditivos indicam uma expansão moderada da área plantada até 2030, influenciada por mudanças climáticas, disponibilidade de recursos e incentivos governamentais.

7. Discussão

Os resultados reforçam a importância de fatores climáticos no planejamento agrícola e apontam caminhos estratégicos para a maximização da produtividade do algodão no Brasil. Entre os destaques:

- Primavera e Verão como períodos críticos: A consistência de temperaturas e radiação solar nesses períodos oferece uma base sólida para práticas agrícolas planejadas.
- Regiões promissoras: O Nordeste e o Centro-Oeste não apenas lideram em potencial produtivo, mas também servem como referência para políticas públicas direcionadas ao desenvolvimento do agronegócio.
- Monitoramento climático: A correlação significativa entre variáveis climáticas e a área plantada destaca a necessidade de sistemas de monitoramento em tempo real e práticas adaptativas, como o uso de barreiras vegetativas contra ventos fortes.
- Tecnologias e mercado: O aumento da área plantada reflete o impacto direto de tecnologias avançadas e do crescente valor do algodão no mercado, sugerindo a relevância de incentivos à inovação agrícola.

8. Conclusão

Este estudo apresentou uma análise abrangente da produção de algodão no Brasil, destacando:

- A importância das condições climáticas ideais, com foco nas estações e nas regiões mais produtivas.
- O impacto de variáveis como temperaturas e ventos na área plantada.
- A evolução histórica impulsionada por inovações tecnológicas e crescimento do mercado.
- Previsões moderadas de expansão até 2030, sugerindo a necessidade de políticas públicas estratégicas.

Recomendações Finais:

1. Planejamento alinhado com as estações ideais: Incentivar o plantio nas épocas de Primavera e Verão, com base em monitoramento climático em tempo real.
2. Investimentos em regiões promissoras: Priorizar infraestrutura e tecnologias agrícolas no Nordeste e Centro-Oeste.
3. Adaptação às mudanças climáticas: Estimular a pesquisa em práticas agrícolas resilientes às mudanças climáticas e promover o uso sustentável dos recursos naturais.
4. Capacitação técnica: Implementar programas de capacitação para agricultores, com foco na aplicação de tecnologias modernas e na gestão climática.
5. Monitoramento contínuo: Expandir os sistemas de monitoramento climático e agrícola, para ajustar estratégias com base em tendências climáticas e de mercado.

Futuras pesquisas podem incorporar variáveis econômicas e sociais para oferecer uma visão ainda mais integrada do cultivo de algodão no Brasil.

Agradecimentos

Agradeço aos professore Sergio Serra e Jorge Zavaleta, pelas aulas incriveis e o entusiasmo pela pesquisa ao Programa de Pós-Graduação em Informática (PPGI/UFRJ) pelo suporte e oportunidade.

References

de Abastecimento (Conab), C. N. (2024a). Safra brasileira de grãos. https://www.conab.gov.br/info-agro/safras/graos. Acesso em: 26 nov. 2024.

de Abastecimento (Conab), C. N. (2024b). Último levantamento da safra 2023/2024 estima produção de grãos em 298,41 milhões de toneladas. https://www.conab.gov.br/ultimas-noticias/5728-ultimo-levantamento-da-safra-2023-2024-estima-producao-de-graos- Acesso em: 26 nov. 2024.

de Meteorologia (INMET), I. N. (2024). Brazil weather information by inmet. https://www.kaggle.com/datasets/gregoryoliveira/brazil-weather-information-by-inmet. Acesso em: 26 nov. 2024.