**Explorando o “web scraping” na aquisição de dados para a previsão de colheitas de grãos e aquisição de insumos na agricultura de precisão**.

Ana Paula Hemsing Brum¹;Pedro Alexander Velasquez-Vasconez²

¹ Graduando do curso de MBA em Data Science e Analytics. Graduada em Letras Língua Espanhola e Pós-graduada em Finanças Empresariais, Rua Pelotas 345, Glória, 98785-145 Santa Rosa Rio Grande do Sul, Brasil

² Orientador, Pedro Alexander Velasquez-Vasconez. Escola de Ciencias Básicas, Tecnología e Ingeniería. Universidad Nacional Abierta y a Distancia. Código postal 520001 Pasto (Nariño), Colômbia.

**Explorando o “web scraping” na aquisição de dados para a previsão de colheitas de grãos e aquisição de insumos na agricultura de precisão**.

**Resumo**

Neste trabalho, explorou-se o uso do “web scraping” para desenvolver uma robusta infraestrutura de dados, com o intuito de auxiliar analistas e cientistas de dados na coleta de informações cruciais para a agricultura de precisão. O objetivo central foi criar um conjunto de dados abrangente que atuasse como uma fonte primordial de informações para pesquisadores e profissionais do setor. Esse conjunto de dados foram direcionados principalmente para a projeção de colheitas e otimização na aquisição de insumos. A metodologia empregada focou na coleta de dados via “web scraping”, integrando-os com outras tecnologias agrícolas avançadas. Os resultados demonstraram que essa abordagem não só agiliza o processo de coleta de dados, mas também enriquece as análises, permitindo decisões mais informadas e estratégicas. Conclui-se que a implementação do “web scraping” na agricultura é uma ferramenta poderosa, essencial para futuros estudos e para a promoção de uma agricultura mais sustentável e eficiente.

**Palavras-chave:** Agricultura de precisão; análise de dados agrícolas; planejamento de cultivos; tomada de decisão agrícola; gestão agrícola sustentável.

**Introdução**

Dados são a moeda do século XXI, e seu valor é inestimável. Essa frase de impacto é amplamente reconhecida e frequentemente citada por influentes na área de tecnologia, como Andrew Grove, co-fundador da Intel, e Peter Sondergaard, ex-diretor de Pesquisa do Gartner Group. Ela ilustra o consenso sobre o foco central das informações digitais, que desempenham um papel essencial na transformação de setores inteiros, inclusive na agricultura.

O uso de tecnologia na agricultura não é uma novidade recente. Ao longo de muitos anos, já incorporamos sistemas de posicionamento global (GPS) e sensores climáticos para aprimorar nossas práticas agrícolas. No entanto, a partir de 2010, testemunhamos uma crescente popularização de tecnologias avançadas, como a Internet das Coisas (IoT) e veículos aéreos não tripulados, conhecidos como drones, no contexto agrícola.

A crescente demanda por tecnologias mais avançadas continua a impulsionar a criação de novos produtos e serviços de forma constante. Nos últimos anos, startups especializadas nesse setor, popularmente chamadas de agrotechs ou agtechs, ganharam destaque. Esse fenômeno não se restringe apenas ao Brasil; é uma tendência global. De acordo com os dados mais recentes do Senso AgTech Brasil, atualmente, temos aproximadamente 300 empresas ativas nesse segmento. Comparativamente, em 2011, apenas nove startups desse tipo operavam no Brasil (Mondin, 2016). Essa evolução notável em um período de apenas 12 anos evidencia a rápida transformação e o crescimento exponencial desse setor.

A ascensão das tecnologias na agricultura não só revolucionou a forma como gerenciamos nossas operações agrícolas, mas também desempenhou um papel fundamental na coleta e análise de dados agrícolas. Com o advento de sistemas de monitoramento, como sensores climáticos e drones equipados com câmeras e sensores, a quantidade de informações disponíveis para agricultores e pesquisadores aumentou exponencialmente. Esses avanços tecnológicos não apenas permitem uma coleta de dados mais precisa e em tempo real, mas também capacitam a agricultura de precisão, possibilitando decisões mais informadas e estratégicas em todas as etapas do ciclo agrícola.

A coleta de dados na agricultura moderna é alimentada por uma variedade de fontes de informações valiosas. Além dos dispositivos de monitoramento e sensores, fontes de dados externas desempenham um papel crucial na compreensão do ambiente agrícola. Isso inclui dados climáticos de estações meteorológicas, informações sobre safras passadas e atuais de agências governamentais, dados de mercado de commodities agrícolas, dados colhidos pelos drones, e equipamentos agrícolas e muito mais.

Este trabalho tem como objetivo central desenvolver uma sólida infraestrutura de dados a partir do "“web scraping”". A base de dados ou "conjunto de dados" resultante servirá como uma valiosa fonte de informações para pesquisadores e profissionais da agricultura. A ênfase recai principalmente sobre o uso desses dados em pesquisas agrícolas, particularmente na projeção de colheitas e na otimização da aquisição de insumos. Nossa expectativa é que este trabalho não apenas forneça recursos essenciais para estudos futuros, mas também promova a aplicação eficaz do “web scraping” na agricultura de precisão, contribuindo assim para a tomada de decisões fundamentadas no setor agrícola.

**Introdução ao “web scraping”**

1. **Origem do termo**

O termo “web scraping” refere-se ao conjunto de técnicas utilizadas para extrair automaticamente informações de sites, ao invés de copiá-las manualmente, podendo ser traduzido como raspagem de dados. Essa prática evoluiu com o crescimento da internet e se tornou uma ferramenta essencial em diversas áreas, incluindo publicidade na web, ciências sociais, saúde pública e pesquisa acadêmica.

A origem do termo remonta ao início dos anos 2000, quando a internet começou a se consolidar como uma fonte rica em dados. Com a proliferação de websites e a necessidade crescente de acessar informações de forma rápida e eficiente, surgiram as primeiras técnicas, inicialmente rudimentares, muitas vezes dependendo de processos manuais ou semi-automatizados.

Um dos primeiros registros acadêmicos do uso do termo "web scraping" pode ser encontrado no artigo "Exploiting web scraping in a collaborative filtering-based approach to web advertising" (Vargiu & Urru, 2012). Este estudo discute o uso da técnicas na área de publicidade online, ilustrando como ela começou a ser aplicadas em contextos comerciais.

Com o passar dos anos, o “web scraping” evoluiu de uma prática para extrair informações de páginas da web para um campo de estudo interdisciplinar. Em um artigo de 2021, Luscombe, Dick e Walby discutem o “web scraping” como uma técnica cada vez mais importante na produção de dados nas ciências sociais, destacando sua relevância e os desafios técnicos, legais e éticos associados [(Luscombe, Dick & Walby, 2021)](https://consensus.app/papers/algorithmic-thinking-interest-navigating-hurdles-luscombe/e4acd90922745f29884560e7c418c192/?utm_source=chatgpt)

Além disso, o artigo "A Pipeline-oriented Processing Approach to Continuous and Long-term web scraping” (Huber, Knoll & Döller, 2022) ilustra a evolução contínua das técnicas utilizadas, demonstrando como a abordagem tem se adaptado para enfrentar desafios de longo prazo e contínuos na extração de dados da web [(Huber, Knoll & Döller, 2022)](https://consensus.app/papers/pipelineoriented-processing-approach-continuous-huber/80ce08b8b08a5bd2b092d207288d35e3/?utm_source=chatgpt).

“Web Crawling”, “web scraping” e “data mining” são termos frequentemente usados na coleta de dados, mas cada um possui suas características distintas. O “web crawling” refere-se ao processo de navegação automatizada pela internet para indexar e buscar informações em websites, sendo a base para os motores de busca. O “web scraping”, por outro lado, é a técnica de extrair dados específicos de páginas da web de maneira estruturada, geralmente para uso específico em projetos ou análises. Já o “data mining”, ou mineração de dados, é um processo mais abrangente que envolve a análise de grandes conjuntos de dados para descobrir padrões, tendências e correlações. Enquanto o “web scraping” foca na coleta de dados específicos de sites, o “data mining” lida com a análise e interpretação desses dados, muitas vezes incluindo informações obtidas através de “web scraping”, para obter insights valiosos e tomar decisões baseadas em dados.

1. **Importância do “web scraping” para a agricultura**

O “web scraping” desempenha um papel crucial na agricultura de precisão, especialmente quando combinado com tecnologias da Internet das Coisas (IoT) e computação em nuvem. Um exemplo notável é a utilização do “web scraping” para coletar dados não estruturados de sites relacionados à agricultura. Esses dados, quando integrados com informações provenientes de dispositivos IoT, oferecem insights valiosos para a gestão de culturas. Isso inclui, por exemplo, o cálculo da evapotranspiração das culturas, um indicador essencial para o manejo inteligente da irrigação (Martínez et al., 2020).

Além disso, o “web scraping” é fundamental para a aplicação de técnicas de aprendizado de máquina na agricultura de precisão. A coleta automatizada de dados de diversas fontes online possibilita a criação de modelos preditivos para variáveis agrícolas, como umidade do solo, previsão de rendimento de culturas, e detecção de doenças e pragas. Este processo melhora significativamente a produtividade e a sustentabilidade na agricultura. Os dados coletados através do “web scraping” podem ser integrados com sensores IoT e outros dispositivos, proporcionando uma visão abrangente das condições agrícolas e promovendo uma agricultura mais inteligente e eficiente (Sharma et al., 2021).

A coleta de dados a longo prazo sobre preços e disponibilidade de insumos agrícolas é outra aplicação vital do “web scraping”. Essas informações permitem identificar tendências de mercado, eficiência de fornecedores e a disponibilidade de certos insumos. Com isso, os agricultores podem planejar melhor suas compras futuras, armazenamento de insumos, seleção de fornecedores e antecipar mudanças de mercado.

Estes exemplos destacam a relevância crítica do “web scraping” na agricultura de precisão. Ao disponibilizar dados atualizados e pertinentes, o “web scraping” não apenas capacita os agricultores e pesquisadores a tomarem decisões mais informadas e precisas, mas também agiliza e enriquece o processo de análise de dados. Com a aquisição automática e contínua de dados, é possível realizar análises mais rápidas e abrangentes, identificando padrões e insights que seriam difíceis de detectar manualmente. Esta capacidade de análise rápida e completa é fundamental para a implementação de práticas agrícolas mais sustentáveis e eficientes, permitindo reações imediatas a mudanças nas condições agrícolas e otimizando o uso de recursos.

1. **Desafios técnicos e éticos**

Um dos principais desafios técnicos do “web scraping” é garantir a validade dos dados coletados. É crucial que os dados sejam precisos e representativos para serem úteis em aplicações agrícolas. A complexidade dos websites e a dinâmica dos dados podem afetar a precisão e a confiabilidade dos dados coletados [(Boegershausen et al., 2022)](https://consensus.app/papers/fields-gold-scraping-data-marketing-insights-boegershausen/291d157b837859c5b92eb38515001376/?utm_source=chatgpt).

Além disso, o “web scraping” na agricultura apresenta desafios específicos, como a necessidade de integrar dados de fontes variáveis e superar dificuldades técnicas como interfaces de usuário complexas ou medidas anti-raspagem dos websites [(Hillen, 2019)](https://consensus.app/papers/scraping-food-price-research-hillen/13e4679b6ac25fbfaaf953bd6b013026/?utm_source=chatgpt). Esses desafios técnicos se entrelaçam com considerações éticas importantes. Em particular, a privacidade dos dados coletados deve ser priorizada, exigindo consentimento quando necessário e conformidade com regulamentações sobre privacidade de dados [(Rennie et al., 2020)](https://consensus.app/papers/scraping-public-health-gains-ethical-considerations-data-rennie/13bd38bcf97e556b9252584ed0a4308c/?utm_source=chatgpt). A legislação brasileira sobre a coleta de dados é amplamente abordada pela Lei Geral de Proteção de Dados Pessoais (LGPD), Lei nº 13.709/2018. Esta lei define regras sobre o tratamento de dados pessoais, incluindo dados coletados na internet, porém como no caso desse estudo não coletaremos dados pessoais a mesma não necessariamente se aplicam.

Além dessas considerações legais e éticas, práticas gerais de “web scraping” também são fundamentais para evitar problemas técnico ou éticos, como a sobrecarga dos servidores dos sites. É importante limitar a frequência das solicitações, implementando atrasos entre elas, e definir limites claros para o escopo da raspagem, evitando assim a coleta excessiva de dados e a criação de "lixo digital". A inclusão de cabeçalhos apropriados nas solicitações, como um User-Agent que identifica claramente o scraper, ajuda a tornar as atividades de raspagem mais transparentes, contribuindo para uma abordagem mais responsável e sustentável.

1. **Principais bibliotecas**

Este capítulo explora as bibliotecas utilizadas para “web scraping”, com foco especial na linguagem de programação Python, amplamente reconhecida por sua versatilidade e facilidade de uso na área de ciência de dados, oferecendo uma ampla gama de bibliotecas de raspagem de dados que facilitam a coleta de dados de diferentes tipos de websites.

**4.1 Beautiful Soup**

A biblioteca “beautiful soup” é utilizada para a extração de dados de arquivos HTML e XML. Ela cria uma árvore de análise a partir da página analisada, permitindo o acesso aos dados de maneira fácil e intuitiva (Richardson, 2007).

Essa biblioteca apresenta diversas características que a tornam popular para "web scraping". Facilita a navegação na árvore DOM (“Document Object Model”) de arquivos HTML ou XML, possibilitando a manipulação dos elementos da página, o que é especialmente útil para extrair dados de sites com estruturas complexas e aninhadas.

Além disso, “beautiful soup” permite a extração de dados por meio de seletores CSS ou expressões XPath, o que garante flexibilidade e precisão na escolha de elementos específicos, como textos, imagens e links, permitindo que o usuário colete exatamente o que precisa de uma página web.

Outra vantagem significativa é a sua fácil integração com a biblioteca “requests”, amplamente usada em Python para acessar páginas web. Combinando essas duas ferramentas, é possível enviar requisições HTTP e, em seguida, utilizar “beautiful soup” para analisar e extrair os dados desejados do HTML retornado.

No que diz respeito aos casos de uso, a biblioteca é comumente utilizada para a extração de dados de tabelas e listas em páginas web, sendo essencial para coletar grandes volumes de informações de forma estruturada. Além disso, é amplamente empregada para coletar textos ou links de artigos, blogs ou relatórios, facilitando a criação de bases de dados a partir de múltiplas fontes online.

Neste trabalho, utilizamos essa biblioteca para coletar dados históricos do fenômeno “El Niño”, disponibilizados no site da NOAA (Administração Oceânica e Atmosférica Nacional dos EUA), em formato de tabela.

**4.2 Scrapy**

"Scrapy" é um framework de "scraping" rápido e de alto nível para Python, utilizado para criar robôs de "scraping" que podem navegar em sites e coletar dados de forma eficiente e escalável.

Uma de suas principais vantagens é o suporte à raspagem assíncrona, o que permite uma coleta de dados rápida e eficiente. Essa funcionalidade é particularmente útil ao lidar com grandes volumes de dados de vários sites simultaneamente, pois permite a execução de múltiplas "requisições" de maneira não bloqueante, otimizando o tempo de execução do processo de "scraping".

Além disso, "Scrapy" oferece uma estrutura robusta para escrever "spiders", que são scripts personalizados projetados para seguir links e explorar sites automaticamente. Essa capacidade de navegação permite que os usuários raspem dados de sites complexos que contêm múltiplas camadas de links, extraindo informações de maneira sistemática e abrangente.

Outro ponto forte do "scrapy" é sua integração com "pipelines" de processamento de dados. Essa funcionalidade facilita a limpeza, transformação e armazenamento dos dados coletados, proporcionando um fluxo contínuo desde a coleta até o armazenamento final. Essa integração é essencial para projetos de "scraping" que exigem processamento de dados em grande escala, garantindo que os dados sejam devidamente formatados e armazenados para uso posterior.

Quanto aos casos de uso, "scrapy" é amplamente utilizado para a coleta de dados de sites grandes e complexos que possuem várias páginas. Ele é ideal para projetos como o monitoramento de preços em lojas de "e-commerce", onde os preços de produtos podem mudar frequentemente e precisam ser rastreados em tempo real. Além disso, "scrapy" é eficaz na extração de dados estruturados para análises ou armazenamento em bancos de dados, tornando-o uma escolha preferida para analistas de dados e cientistas que necessitam de uma solução escalável e eficiente para "web scraping".

**4.3 Selenium**

"Selenium" é uma ferramenta para automatizar navegadores web. Embora seja amplamente usada para testes automatizados, também é útil para "scraping" de sites que utilizam conteúdo dinâmico carregado por "javascript". Neste trabalho, utilizamos muitas opções do seu pacote, principalmente para a coleta de dados do "INMET", onde foi necessário interagir com vários elementos da página.

Essa ferramenta popular para automação de navegadores web apresenta várias características que a tornam uma escolha ideal para a raspagem de dados de páginas dinâmicas. Uma de suas principais funcionalidades é a capacidade de interagir com elementos dinâmicos de uma página web. Isso inclui ações como clicar em botões, preencher formulários e navegar através de menus suspensos. Essas interações são cruciais quando se lida com sites que carregam conteúdo dinamicamente ou exigem ações do usuário para revelar certas informações.

Além disso, "selenium" oferece suporte a vários navegadores, incluindo "Chrome", "firefox" e "safari", o que o torna uma ferramenta versátil e compatível com diferentes ambientes de desenvolvimento e necessidades específicas de "scraping". Essa flexibilidade é importante para desenvolver scripts que funcionem de maneira consistente em diferentes plataformas e navegadores, garantindo que o processo de "scraping" seja robusto e adaptável.

Outra vantagem significativa dessa biblioteca é sua capacidade de ser usada em conjunto com a biblioteca "beautiful soup". Enquanto "selenium" pode carregar e interagir com páginas dinâmicas, "beautiful soup" pode ser utilizada para extrair dados após a renderização da página. Essa combinação é particularmente útil para sites que utilizam "javascript" ou que carregam conteúdo de forma assíncrona, permitindo que os desenvolvedores raspem dados que seriam inacessíveis usando apenas técnicas de "scraping" baseadas em HTML estático.

Quanto aos casos de uso, "selenium" é frequentemente utilizado para a extração de dados de sites que dependem de "javascript" intensivo ou que carregam conteúdo de forma assíncrona, como plataformas de redes sociais ou portais de notícias dinâmicos. Também é amplamente aplicado na simulação de interações de usuário para "scraping" de dados em sites que possuem autenticação ou proteção contra "bots". Nessas situações, "selenium" permite que os desenvolvedores automatizem o login, naveguem pelas páginas internas e coletem dados que, de outra forma, estariam protegidos contra o "scraping" convencional.

* 1. **Requests**

"Requests" é uma biblioteca "HTTP" para python, que facilita o envio de requisições "HTTP/1.1", com métodos como "get" e "post". Embora não seja uma ferramenta de "scraping" por si só, é frequentemente usada em conjunto com outras bibliotecas como "beautiful soup" e "scrapy" para obter o conteúdo "HTML" de uma página web.

A biblioteca "requests" é amplamente reconhecida por sua simplicidade e facilidade de uso ao enviar requisições "HTTP" em Python. Esta ferramenta permite que desenvolvedores realizem operações "HTTP", como "get" e "post", de maneira intuitiva e eficiente, tornando-a uma escolha popular para aqueles que precisam acessar e interagir com páginas web e "APIs". Sua interface amigável facilita o processo de requisição de dados de sites, sem a complexidade associada a muitas outras bibliotecas de rede. (Beazley, 2013)

Além disso, "requests" oferece suporte robusto para o gerenciamento de "cookies", autenticação e manipulação de cabeçalhos personalizados. Isso significa que os usuários podem facilmente manter sessões de login, lidar com autenticação de vários tipos (como "Basic", "Digest", "OAuth"), e ajustar cabeçalhos "HTTP" para imitar requisições de diferentes navegadores ou dispositivos, ou para atender a requisitos específicos de "API".

Nos casos de uso, é frequentemente utilizado para acessar páginas web e "apis" para coletar dados em formatos "HTML" e "JSON". Isso permite a extração de informações estruturadas ou semiestruturadas de uma ampla variedade de fontes online. Além disso, a biblioteca é essencial para enviar requisições "HTTP" para sites que exigem autenticação, permitindo que os desenvolvedores interajam com plataformas que protegem seus dados por trás de um sistema de login, ou que necessitam de manipulação específica de cabeçalhos para acessar determinados recursos. Essa funcionalidade torna "requests" uma ferramenta indispensável para desenvolvedores que trabalham com "scraping" de dados ou integração de " apis ".

* 1. **Httpx**

A biblioteca "httpx" foi desenvolvida pensando em eficiência, alta performance e suporte a tecnologias modernas. Desenvolvida pela organização "Encode", que também criou o "starlette" e "django rest framework", o "httpx" se destaca por oferecer suporte a operações assíncronas e compatibilidade com o protocolo "HTTP/2". Lançado em 2019, o "httpx" veio suprir as novas necessidades do desenvolvimento moderno, que requerem tanto um fluxo mais eficiente quanto uma maior concorrência na execução de tarefas de rede.

Esta biblioteca possui uma série de características que o diferenciam de bibliotecas anteriores, como o "requests", que ainda é amplamente usada. Uma das maiores vantagens do “httpx" é o suporte a requisições assíncronas usando "asyncio". Isso significa que ele é capaz de realizar várias requisições simultâneas sem precisar bloquear o fluxo do programa. Essa abordagem assíncrona é extremamente útil ao lidar com grandes volumes de dados, como em "web scraping" ou integração com várias "APIs".

O "httpx" suporta o protocolo "HTTP/2", uma evolução do "HTTP/1.1" que introduz multiplexação, compressão de cabeçalhos e conexões persistentes mais eficientes. Isso resulta em requisições mais rápidas e melhor aproveitamento dos recursos de rede.

Para facilitar a transição, o "httpx" manteve uma "API" similar à biblioteca "requests", o que permite que desenvolvedores familiarizados com "requests" aprendam a usar o "httpx" de forma rápida e eficiente. A diferença está principalmente na possibilidade de usar "await" e lidar com tarefas assíncronas.

Embora essa biblioteca seja conhecida por seu suporte assíncrono, ele também permite fazer requisições síncronas, tornando-o flexível para vários cenários. Isso é útil em situações em que a simplicidade é mais importante que a eficiência.

Assim como o "requests", o " httpx" permite o uso de sessões e o gerenciamento de "cookies", facilitando o rastreamento do estado de uma conexão, algo essencial ao trabalhar com sites que exigem autenticação, e fornece suporte completo para requisições "https", com verificação de certificados "ssl".

**O papel fundamental dos conjuntos de dados na análise de dados**

A análise de dados é um componente crucial na agricultura de precisão, permitindo transformar grandes quantidades de dados brutos em insights valiosos. No coração dessa transformação estão os conjuntos de dados, estruturas de dados essenciais para manipulação e análise eficiente de dados. Este capítulo explora a importância dos conjuntos de dados, destacando suas funcionalidades e aplicações práticas na análise de dados.

Conjunto de dados são estruturas de dados colunares que permitem uma análise eficiente e organizada de grandes volumes de dados, essenciais na agricultura de precisão. O “web scraping” é uma das técnicas utilizadas para a coleta de dados para a construção de conjunto de dados que facilitam o processo de limpeza, manipulação e análise de dados, tornando-se uma ferramenta indispensável para agricultores e pesquisadores.

Exemplo de conjunto de dados referente aos dados coletados da produção de trigo que faz parte da etapa de resultados preliminares deste trabalho, figura 1:

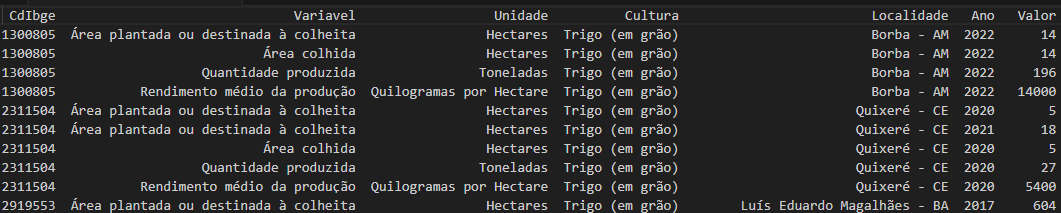


Figura 1. Representação do banco de dados criado a partir do script proposto neste trabalho. Fonte: IBGE

A principal vantagem dos conjuntos de dados é sua flexibilidade. Eles permitem a realização de várias operações de manipulação de dados, como seleção, filtragem, agrupamento e reordenamento, além de cálculos para previsões. Segundo Hadley Wickham, conhecido por suas contribuições para a linguagem R, conjuntos de dados tornam a manipulação de dados mais intuitiva, ao permitir que o usuário se concentre na manipulação dos dados em si, em vez de se preocupar com os detalhes de armazenamento (Wickham, 2014).

**Resultados**

Como resultados preliminares temos uma coletânea de dados contendo dados históricos de produção agrícola do Brasil, dados de identificação geográfica das unidades federativas e dados referente a situação climática. Abaixo teremos uma descrição mais abrangente sobre cada banco de dados e a granularidade das informações contidas.

**Produção agrícola de grãos**

Criado base de dados com o histórico de 2017 a 2023 com informações sobre a produção agrícola de grãos (soja, milho e trigo) no Brasil, usando como fonte de dados o IBGE, área plantada ou destinada à colheita, quantidade produzida, rendimento médio da produção, área colhida e rendimento médio da produção, nele temos:

* CdIbge: Código do Instituto Brasileiro de Geografia e Estatística para a localidade.
* Variável: dado referente ao tipo de valor retornado, podendo ser: área plantada ou destinada à colheita, quantidade produzida, rendimento médio da produção, área colhida e rendimento médio da produção.
* Unidade: unidade de medida que se aplica o valor, exemplo: hectares, toneladas, Quilogramas por Hectare.
* Cultura: tipo de cultura, neste caso, trigo (em grão), soja e milho.
* Localidade: Nome da localidade, incluindo a cidade e o estado.
* Ano: O ano a que os dados se referem.
* Valor: Valor correspondente à variável para o ano e localidade especificados.

A granularidade dos dados está a nível de município, podendo ser relacionada a tabela dim\_municipios pelo CdIbge.

Para essa coleta, utilizamos como fonte de dados a API do IBGE, que se destaca pela sua confiabilidade e abrangência. O Instituto Brasileiro de Geografia e Estatística (IBGE) é a principal referência em estatísticas oficiais no Brasil, reconhecido por fornecer dados de qualidade, atualizados e validados, abrangendo diversos aspectos sociais, econômicos e geográficos do país. Esses dados são amplamente utilizados por pesquisadores, governos e organizações internacionais, garantindo assim a credibilidade da informação coletada. O script de coleta “scraping\_safra.ipynb” utiliza essa API para assegurar que as informações obtidas sejam precisas e adequadas para o estudo em questão.

**Dados Climáticos**

Neste tópico, foram obtidas informações relacionadas ao balanço hídrico de 486 estações meteorológicas e de solo sob a gestão do INMET (Instituto Nacional de Meteorologia), abrangendo o período de 01/2017 a 08/2024. Os dados coletados incluem a data da coleta (data), o nome da estação, armazenamento hídrico (ARM), alteração de armazenamento (ALT), evapotranspiração real (ETR), déficit hídrico (DEF), excedente hídrico (EXC), evapotranspiração de referência (ETO), precipitação (P) e temperatura média (T). Além disso, foi registrada a característica do tipo de solo predominante na região de cada estação, sendo esses dados relevantes para a análise do comportamento hídrico e climático nas diversas regiões monitoradas, permitindo compreender padrões de variabilidade sazonal e regional, bem como embasar estudos de impacto ambiental e planejamento agrícola.

Para esta coleta utilizamos dados fornecidos pelo INMET disponibilizados para consumo do público em geral no site <https://sisdagro.inmet.gov.br/sisdagro/app/monitoramento/bhs>, o script usado para tal consumo está disponível no git sob o nome “scraping\_meteorologicos”.

O Instituto Nacional de Meteorologia (INMET) é um órgão governamental ligado ao Ministério da Agricultura, Pecuária e Abastecimento do Brasil, com mais de 100 anos de experiência na coleta e análise de dados meteorológicos. O INMET possui uma rede extensa de estações meteorológicas automáticas e convencionais distribuídas por todo o território nacional, o que garante uma ampla cobertura geográfica e uma coleta de dados precisa e consistente. Além disso, os processos de medição e validação de dados do INMET seguem padrões internacionais, assegurando que as informações sejam confiáveis e compatíveis com aqueles obtidos em outras partes do mundo. O INMET é amplamente reconhecido como uma fonte oficial e segura de dados meteorológicos, utilizados por pesquisadores, governos, agricultores e empresas para análises climáticas, previsões do tempo, planejamento agrícola e estudos ambientais. Dessa forma, a utilização dos dados do INMET neste estudo garante a qualidade e a precisão necessárias para uma análise robusta e fidedigna.

No contexto das informações climáticas, também foram coletados dados sobre os períodos propícios para o manejo do solo, colheita e semeadura, sendo que essas probabilidades foram calculadas pelo INMET e disponibiliza em seu site, permitindo o planejamento mais inteligente das práticas agrícolas.

O arquivo denominado “dias\_aptos\_manejo\_solo.xlsx” contém informações sobre a estimativa de dias considerados aptos para diferentes práticas agrícolas, como manejo do solo, colheita e semeadura, com base em dados meteorológicos. Esses dados são apresentados de forma organizada, permitindo uma leitura detalhada das condições climáticas para cada período do mês e cada estação meteorológica.

Cada linha do arquivo inclui as seguintes informações:

* Cod. IBGE: Código de identificação único da região ou município para o qual os dados foram registrados.
* Probabilidade: Tipo de probabilidade apresentada, como "Anual", indicando a frequência com que determinados dias são considerados aptos.
* Prática Agrícola: Tipo de prática em questão, como "Preparo do Solo", "Colheita" ou "Semeadura".
* Estação: Nome da estação meteorológica utilizada como base para os dados apresentados.
* Decêndio: Período de dez dias do mês. Cada mês é dividido em três decêndios: o primeiro (dias 1 a 10), o segundo (dias 11 a 20) e o terceiro (dias 21 até o final do mês).
* Mês: Mês ao qual os dados se referem.
* Dias Aptos: Número de dias considerados aptos para a prática agrícola em questão, com base nas condições climáticas.
* Porcentagem Dias Aptos: Probabilidade percentual de ocorrer o número especificado de dias aptos durante o decêndio.

Para facilitar a interpretação dos dados fornecerei um exemplo de leitura, considerando os dados gerados na tabela a seguir.

Tabela 1.Dias aptos para o preparo de solo na cidade de Santa Rosa-RS

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cod. IBGE | Probabilidade | Prática Agricola | Estação | Decêndio | Mês | Dias Aptos | Porcentagem Dias Aptos |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 2 | 47,2 |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 3 | 35,6 |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 4 | 26,7 |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 5 | 20,2 |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 6 | 12,1 |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 7 | 6,9 |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 8 | 5,2 |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 9 | 3,4 |
| 4317202 | Anual | Preparo do Solo | SANTA ROSA (A) - RS | 1 | 1 | 10 | 1,7 |

Fonte: INMET

Com base nessa tabela podemos fazer a seguinte leitura: para o primeiro decêndio de janeiro, a probabilidade de se ter 2 dias aptos ao preparo do solo é de 47,2%. Ou seja, há uma chance de 47,2% de ocorrerem 2 dias adequados para realizar o manejo do solo, considerando o nível de umidade e precipitação definidos como critério para essa prática agrícola.

Dados sobre a previsão de formação de geada e sua intensidade também são fornecidas pelo INMET, e foram coletas pelo script “scraping\_geada\_convencional.ipynb” a coleta abrange-o o período de 01/2017 a 08/2024 com as informações de :

* Cod. IBGE: Código de identificação único do município.
* UF: Unidade Federativa (estado) onde o município está localizado.
* Município: Nome do município.
* Dia de ocorrência: Data em que ocorreu a geada.
* Temperatura Mínima: Temperatura mínima registrada no dia da ocorrência.
* Intensidade: Intensidade da geada, classificada como "Fraca", "Moderada" ou "Forte".

Tabela:2 Exemplo de dados coletados sobre geada de estação convencional.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cod. IBGE | Uf | Município | Dia de ocorrência | Temperatura Mínima | Intensidade |
| 4216503 | SC | São Joaquim | 22/05/2017 | 3.0°C | Moderada |
| 4216503 | SC | São Joaquim | 16/05/2017 | 5.0°C | Fraca |
| 4301602 | RS | Bagé | 21/06/2017 | 4.8°C | Fraca |

Fonte:INMET

Na tabela de exemplo apresentada podemos realizar a leitura da seguinte maneira: para o município de **São Joaquim (SC),** no dia **22/05/2017**, foi registrada uma temperatura mínima de **3.0°C**, e a intensidade da geada foi classificada como **moderada**.

Para a analise da geada foram coletados dados de duas estações meteorológicas, automática e convencional, permitindo assim uma cobertura maior do território.

A coleta de dados meteorológicos também abrange a intensidade dos fenômenos "El Niño" e "La Niña", que, quando presentes e dependendo de sua intensidade, impactam significativamente a produção de grãos. "El Niño" está associado ao aquecimento anômalo das águas do Oceano Pacífico Equatorial, causando alterações climáticas que podem resultar em secas severas em algumas regiões produtoras, reduzindo o rendimento das lavouras devido à falta de umidade. Por outro lado, "La Niña" provoca o resfriamento dessas águas, o que pode levar ao excesso de chuvas em certas áreas, causando inundações, erosão do solo e comprometimento da qualidade e produtividade das culturas agrícolas. Essas variações climáticas afetam diretamente o planejamento agrícola, a previsão de colheitas e a alocação de recursos, tornando essencial o monitoramento desses fenômenos para mitigar seus impactos adversos na produção de grãos. Os dados utilizados para análise dos fenômenos "El Niño" e "La Niña" foram obtidos dos "National Centers for Environmental Prediction" (NCEP) dos EUA, que oferecem informações detalhadas e atualizadas sobre a intensidade e duração desses eventos climáticos.

Esses dados podem ser acessados no arquivo “[resultados\_oni.xlsx](https://github.com/ANA-BRUM/DadosAgricolas/blob/main/dados_meteorologicos/resultados_oni.xlsx)”.

Para apoiar a análise conjunta dos dados coletados incluímos um campo com o código do município (Cod. IBGE), sempre que o dados coletado estivesse ao nível de município e inclui na coleta uma tabela de municípios para realizar esse relacionamento. Abaixo o modelo de relacionamento esperado.

* Produção Agricola Nacional[Cod. IBGE] \*[<-]1 Dim Municipios[Cod. IBGE].
* Produção Agricola Nacional[Ano] \*[<-]\* Dim Calendario[Ano].
* Dias Aptos Manejo Solo[Cod. IBGE] \*[<-]1 Dim Municipios[Cod. IBGE].
* Dias Aptos Manejo Solo[Mês] \*[<-]\* Dim Calendario[MêsNúm].
* Geada Convenciona'[Cod. IBGE] \*[<-]1 Dim Municipios[Cod. IBGE].
* Geada Convenciona'[Dia de ocorrência] \*[<-]1 Dim Calendario[Date].
* Geada Automatico[Cod. IBG\*[<-]1 Dim Municipios[Cod. IBGE].
* Geada Automatico[Dia de ocorrência] \*[<-]1 Dim Calendario'[Date].
* Dados Metereologicos[Data] \*[<-]1 Dim Calendario[Date].
* Dados Metereologicos[Cod. IBGE] \*[<-]1 Dim Municipios[Cod. IBGE].
* Dim Calendario[MesAno] \*[<-]1 Oni[MesAno].

Como em alguns situações o nome da estação meteorológica não representava um município, como por exemplo o arquipélago de Abrolhos na Bahia, inclui o código do município mais próximo ao local, para realizar o relacionamento.

Para construção do relacionamento realizou-se uma modelagem nos dados e a tabela Produção Agrícola Nacional representa a união da tabelas de produção agrícola de milho, trigo e soja.

Todas as coletas citadas nesse trabalho estão disponíveis no github https://github.com/ANA-BRUM/DadosAgricolas.

**Referências**

Balasundram, SK; Shamshiri, RR; Sridhara, S.; Rizan, N. 2023. O papel da agricultura digital na mitigação das mudanças climáticas e na garantia da segurança alimentar: uma visão geral. Sustentabilidade. Disponível em: <https://doi.org/10.3390/su15065325>. Acesso em: 10 jan. 2024.

Beazley, D., & Jones, B. 2013. Python Cookbook. 3ed. O'Reilly Media.

Boegershausen, J.;Datta, H.; Borah, A.; Stephen, A. T. Fields of Gold: Scraping Web Data for Marketing Insights. 2022. Disponível em:< https://journals.sagepub.com/doi/pdf/10.1177/00222429221100750>. Acesso em: 14 nov. 2023.

CRUZ, Felipe. Python: escreva seus primeiros programas. São Paulo, SP: Casa do Código, 2015. E-book. Disponível em: <https://plataforma.bvirtual.com.br>. Acesso em: 21 out. 2024.

GRUS, Joel. Data science do zero: noções fundamentais com Python. 2. ed. Rio de Janeiro: Alta Books, 2016. E-book. (1 recurso online). ISBN 9788550816463. Disponível em: <https://integrada.minhabiblioteca.com.br/books/9788550816463>. Acesso em: 30 set. 2023.

Hillen, Judith. “web scraping” for food price research. 2019. Disponível em:< https://www.emerald.com/insight/content/doi/10.1108/BFJ-02-2019-0081/full/html>. Acesso em: 14 nov. 2023.

Huber, S.; Knoll, F., Döller, M. 2022. A Pipeline-oriented Processing Approach to Continuous and Long-term “web scraping”. Disponível em: < https://www.scitepress.org/Link.aspx?doi=10.5220/0011275100003266>. Acesso em: 10 out. 2023.

Instituto Brasileiro de Geografia e Estatística [IBGE]. 2023. API serviço de dados. Disponível em: <https://servicodados.ibge.gov.br/api/docs/>. Acesso em: 5 set. 2023.

Kalenska, Svitlana. 2022. Food security and innovative solutions in crop production. Disponível em: <https://agriculturalscience.com.ua/en/journals/tom-13-2-2022/prodovolcha-bezpeka-ta-innovatsiyni-rishennya-v-roslinnitstvi>. Acesso em: 10 jan. 2024.

Luscombe, A., Dick, K.; Walby, K. 2022. Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to “web scraping” in the social sciences. Disponível em: < https://doi.org/10.1007/s11135-021-01164-0>. Acesso em: 02 out. 2023.

Martínez, J. V. Y.; Skarmeta, A. F.; Zamora-Izquierdo, M. A.; Ramallo-González, A. P. IoT-based data management for Smart Agricultur. 2020. Disponível em: < https://ieeexplore.ieee.org/document/9296443/authors#authors>. Acesso em: 02 out. 2023.

Mitchell,Ryan. 2019. Web Scraping com Python, coletando dados na Web moderna. [Novatec Editora](https://www.google.com.br/search?hl=pt-BR&sa=X&ved=2ahUKEwjtjo7xgKOJAxViHrkGHWqlF0oQre8FegQIAxAY&q=inpublisher:%22Novatec+Editora%22&tbm=bks&sxsrf=ADLYWIIbpcFCvJs6o4JvnC6B60lftBpaig:1729635112465). 2ed.Brasil

Mondin Mateus; Nonino Hermes. 2016. 1º censo agtech startups Brasil: confira resultados e análises de mapeamento inédito setor. Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, SP, Brasil. Disponível em: < https://www.startagro.agr.br/1o-censo-agtech-startups-brasil-confira-resultados-e-analises/>. Acesso em: 02 out. 2023.

Python Requests. Requests: HTTP for Humans. Disponível em: <https://docs.python-requests.org/>. Acesso em: 15 mar. 2024

Richardson, Leonard. Beautiful Soup Documentation. Disponível em: < https://www.crummy.com/software/BeautifulSoup/> Acesso em: 01 ago. 2024.

Sharma, A., Jain, A., Gupta, P., Chowdary, V. 2021. Machine Learning Applications for Precision Agriculture. Disponível em: <https://doi.org/10.1109/ACCESS.2020.3048415> Acesso em: 02 out. 2023.

Selenium. Selenium documentation. Disponível em: <https://www.selenium.dev/documentation/en/>. Acesso em: 15 mar. 2024.

Vargiu, Eloisa; Mirko Urru. 2012. Exploiting “web scraping” in a collaborative filtering- based approach to web advertising. Disponível em: <https://www.semanticscholar.org/paper/Exploiting-web-scraping-in-a-collaborative-based-to-Vargiu-Urru/584753b67ad5e40>. Acesso em: 02 out. 2023.