


**PROCESSAMENTO E LIMPEZA DE DADOS DE PRODUTIVIDADE AGRÍCOLA:  
APLICAÇÃO DE UM SCRIPT EM PYTHON**

**PROCESSING AND CLEANING OF AGRICULTURAL PRODUCTIVITY DATA:  
APPLICATION OF A PYTHON SCRIPT**

**PROCESAMIENTO Y LIMPIEZA DE DATOS DE PRODUCTIVIDAD AGRÍCOLA:  
APLICACIÓN DE UN SCRIPT DE PYTHON**

 <https://doi.org/10.56238/arev7n7-037>

**Data de submissão:** 03/06/2025

**Data de publicação:** 03/07/2025

**Raphael Prazeres da Silva**

Graduando em Engenharia Agrícola — Universidade Federal de Sergipe  
E-mail: rprazeress@gmail.com

**Welington Gonzaga do Vale**

Doutor em Produção Vegetal — Universidade Federal de Sergipe  
E-mail: valewg@gmail.com

**Janyelle do Nascimento Silva**

Mestranda em Ciência da Propriedade Intelectual — Universidade Federal de Sergipe  
E-mail: janyelle.engagricola@gmail.com

**Valfran José Santos Andrade**

Mestre em Recursos Hídricos — Universidade Federal de Sergipe  
E-mail: valfranjose40@gmail.com

**Patricia de Azevedo Castelo Branco do Vale**

Doutora em Ciência Animal — Universidade Federal de Sergipe  
E-mail: patriciavale78@gmail.com

**Adilson Machado Enes**

Doutor em Engenharia Agrícola — Universidade Federal de Sergipe  
E-mail: adilsonenes@gmail.com

**Diego Andrade Pereira**

Engenheiro Mecânico — Universidade Federal de Sergipe  
E-mail: diegoandrade\_senai@yahoo.com.br

---

**RESUMO**

Desenvolveu-se um *script* em Python para processar e limpar dados de produtividade agrícola de colhedoras em uma fazenda localizada em Brasnorte (MT), visando melhorar a confiabilidade em Agricultura de Precisão. O código, utilizando a biblioteca Pandas, aplicou três etapas: 1) filtragem por estado operacional (apenas registros "Efetivo"); 2) remoção de *outliers* (valores <500 kg/ha ou >dobro da média) e 3) ajuste iterativo das médias por equipamento. Os dados tratados foram interpolados no QGIS utilizando o método IDW. Os resultados mostraram que 58,8% dos dados brutos foram descartados no Talhão 1 e 66,9% no Talhão 2, principalmente devido a falhas ou sensores zerados. As

médias de produtividade aumentaram de 2,67 t/ha para 3,67 t/ha (Talhão 1) e 2,52 t/ha para 3,82 t/ha (Talhão 2), com a eliminação de valores extremos. Os mapas gerados revelaram áreas críticas nas bordas e regiões com falhas. Conclui-se que a ferramenta é eficaz na automação da limpeza dos dados de produtividade, no entanto estudos futuros devem considerar a inclusão de validação cruzada para reforçar a confiabilidade dos resultados.

**Palavras-chave:** Agricultura de precisão. Análise de dados. Mapa de produtividade.

## ABSTRACT

A script written in Python was developed to process and clean agricultural yield data from grain harvesters on a farm located in Brasnorte (MT), aiming to improve data reliability in Precision Agriculture. The code, using the Pandas library, followed three main steps: (1) filtering by machine operation status (retaining only “Effective” records), (2) removal of outliers (values <500 kg/ha or >twice the average), and (3) iterative adjustment of machine-specific yield values to match the field average. The cleaned data were interpolated in QGIS using the IDW method. The results showed that 58.8% of the raw data were discarded in Field 1 and 66.9% in Field 2, mainly due to failures or zeroed sensors. Yield averages increased from 2.67 t/ha to 3.67 t/ha (Field 1) and from 2.52 t/ha to 3.82 t/ha (Field 2), with the elimination of extreme values. The generated maps highlighted critical zones near field edges and data gaps. The results suggest that the tool efficiently automates data cleaning, though future studies should consider including cross-validation to reinforce the reliability of the results.

**Keywords:** Precision agriculture. Data analysis. Yield map.

## RESUMEN

Se desarrolló un script en Python para procesar y depurar datos de productividad agrícola de las cosechadoras en una finca ubicada en Brasnorte (MT), con el objetivo de mejorar la confiabilidad en la agricultura de precisión. El código, utilizando la biblioteca Pandas, aplicó tres pasos: 1) filtrado por estado operativo (solo registros "Efectivos"); 2) eliminación de valores atípicos (valores <500 kg/ha o >2 veces el promedio); y 3) ajuste iterativo de promedios por equipo. Los datos procesados se interpolaron en QGIS mediante el método IDW. Los resultados mostraron que el 58,8% de los datos brutos se descartaron en la Parcela 1 y el 66,9% en la Parcela 2, principalmente debido a fallas o sensores puestos a cero. Los promedios de productividad aumentaron de 2,67 t/ha a 3,67 t/ha (Parcela 1) y de 2,52 t/ha a 3,82 t/ha (Parcela 2), con la eliminación de valores extremos. Los mapas generados revelaron áreas críticas en los bordes y regiones con fallas. Se concluye que la herramienta es eficaz para automatizar la limpieza de datos de productividad; sin embargo, estudios futuros deberían considerar la inclusión de validación cruzada para reforzar la fiabilidad de los resultados.

**Palabras clave:** Agricultura de precisión. Análisis de datos. Mapa de productividad.

## 1 INTRODUÇÃO

A agricultura de precisão (AP) tem se consolidado como uma abordagem essencial na modernização do agronegócio, integrando tecnologias avançadas para monitorar e otimizar as operações agrícolas. Sensores instalados em máquinas agrícolas, como colhedoras, são capazes de registrar uma vasta quantidade de informações por segundo, abrangendo desde dados operacionais até condições ambientais (Costa et al., 2015). Essa capacidade de coleta massiva de dados permite um monitoramento detalhado da lavoura, possibilitando a identificação de variabilidades intra e entre talhões, o que é fundamental para a tomada de decisões mais assertivas (Silva; Silva-Mann, 2020).

As colhedoras modernas estão equipadas com sistemas automáticos de medição de produtividade, que integram sensores de fluxo de massa, umidade e posicionamento geográfico. Esses sensores coletam dados em tempo real durante a operação de colheita, permitindo o cálculo instantâneo da produtividade com base no volume de grãos colhido por unidade de área (Pereira; Molin, 2003; Li et al., 2005). O sensor de fluxo, geralmente localizado no elevador de grãos limpos, estima o volume colhido por tempo, enquanto sensores de umidade garantem o ajuste para valores padrão. Combinados ao sinal de GPS, esses dados são registrados e armazenados automaticamente, formando a base dos mapas de produtividade.

Entretanto, a grande quantidade de dados gerados apresenta desafios significativos. Além do volume, a qualidade dos dados é uma preocupação central, uma vez que informações coletadas por sensores podem conter erros, ruídos e inconsistências, decorrentes de falhas nos dispositivos, interferências ambientais ou problemas operacionais (Fizza et al., 2022). Essas imperfeições comprometem a análise da produtividade e podem levar a decisões equivocadas no manejo agrícola. Além disso, registros duplicados ou informações coletadas quando a colhedora não está em operação efetiva de colheita podem distorcer os cálculos de rendimento da lavoura. Para garantir a confiabilidade das informações, é essencial realizar um processamento e limpeza dos dados brutos, filtrando apenas aqueles que realmente representam a produtividade real do campo.

Diante desse contexto, este estudo tem como objetivo desenvolver e aplicar um script em Python para o tratamento de dados da produtividade agrícola, aplicando-o a um estudo de caso realizado em uma propriedade rural localizada no município de Brasnorte no estado do Mato Grosso.

## 2 OBJETIVOS

### 2.1 OBJETIVO GERAL

Desenvolver e aplicar um script em Python para limpeza de dados de produtividade agrícola.

## 2.2 OBJETIVOS ESPECÍFICOS

- Identificar possíveis problemas nos dados brutos;
- Aplicar técnicas de limpeza e filtragem;
- Comparar os dados antes e depois do tratamento.

## 3 REVISÃO DE LITERATURA

A agricultura de precisão (AP) tem se destacado no uso de tecnologias voltada para a coleta e processamento de dados agrícolas, possibilitando que os produtores adotem estratégias de manejo mais eficientes (Tschiedel; Ferreira, 2002). Essa metodologia se fundamenta na aplicação otimizada de insumos agrícolas, ajustando seu uso de acordo com a variabilidade espacial e temporal das lavouras (Silva; Silva-Mann, 2020). Para isso, a AP depende da captação, armazenamento e análise de grandes volumes de dados, proporcionando tomadas de decisão mais assertivas e ganhos significativos na produtividade (Basso et al., 2020).

A obtenção e o processamento dessas informações são viabilizados por um conjunto de tecnologias, incluindo sensores embarcados em máquinas agrícolas, softwares de análise e equipamentos automatizados. Esses sistemas operam em diferentes níveis de automação, variando desde processos parcialmente manuais até soluções totalmente autônomas (Basso et al., 2020). Nesse contexto, sensores instalados em máquinas agrícolas desempenham um papel essencial ao transmitir informações ambientais em tempo real, fornecendo dados valiosos para o produtor rural. Essa tecnologia tem se consolidado como uma tendência na otimização dos processos agrícolas, permitindo um controle mais preciso das operações no campo (Costa et al., 2015).

Além dos sensores, os sistemas de monitoramento de produtividade instalados em colhedoras têm papel fundamental na coleta automática de dados durante a colheita. Esses sistemas são compostos por sensores de fluxo de massa, sensores de umidade e receptores GNSS, os quais permitem estimar a produtividade em tempo real, associando o volume colhido à posição geográfica da máquina (Chandel et al., 2013). Quando devidamente calibrados, esses monitores oferecem informações confiáveis que subsidiam a geração de mapas de produtividade detalhados (Grisso et al., 2002).

Apesar dos avanços tecnológicos, a qualidade dos dados agrícolas ainda enfrenta desafios significativos. Sensores podem apresentar problemas como ruídos, inconsistências e erros na coleta de dados, afetando a precisão das análises e, conseqüentemente, a tomada de decisões (Tschiedel; Ferreira, 2002). Segundo Menegatti e Molin (2004) os arquivos brutos podem conter erros de posicionamento; rendimento nulo, ausente ou discrepantes; largura da plataforma irregular; do intervalo de enchimento; umidade nula ou ausente e distância nula entre pontos. Além desses, Sudduth e Drummond (2007)

também citaram o tempo de esvaziamento do granelheiro; atraso de tempo do grão na colhedora e mudanças rápidas de velocidade - como erros comuns em dados brutos.

Para mitigar essas limitações, o pré-processamento de dados surge como uma etapa essencial, envolvendo técnicas como limpeza, normalização e transformação para garantir a qualidade das informações utilizadas (Maldaner; Molin; Spekken, 2022). No contexto da AP, são amplamente empregadas estratégias como a remoção de outliers, filtragem e ajustes de calibração para corrigir inconsistências nos dados coletados. A automação desse processo, por meio de scripts, tem demonstrado ser uma alternativa eficiente para melhorar a precisão das análises, além de reduzir significativamente o tempo necessário para a preparação dos dados (Damico, 2025). Entre as abordagens mais utilizadas na limpeza de dados agrícolas, destacam-se os filtros de máximo e mínimo rendimento da colheita, que, em algumas situações, são os únicos métodos aplicados (Sudduth; Drummond, 2007). Esses filtros são amplamente empregados em estudos voltados para a remoção de valores extremos e o aprimoramento da qualidade dos dados (Sudduth; Drummond, 2007; Gimenez; Molin, 2004; Vega et al., 2019).

Gimenez e Molin (2004) aplicaram um filtro, com base na média, para remoção de dados discrepantes (outliers) dos sensores de produtividade e largura de corte da plataforma,  $n \cdot me'dia$  para valores maiores e  $n/me'dia$  para valores menores. Sudduth e Drummond (2007) sugerem um valor próximo de 0 (zero) para remover dados de produtividade baixa e o valor de máximo potencial da cultura para a remoção dos valores de alta produtividade.

Após o tratamento dos dados, autores como Santi et al. (2013) aplicam técnicas de interpolação para a geração de mapas contínuos de produtividade. Entre os algoritmos mais utilizados estão a Distância Inversa Ponderada (IDW) em que a saída é uma camada raster contendo valores em toda a extensão dos dados (Chandel et al., 2013) e a Krigagem destacada por considerar a dependência espacial entre os pontos amostrados (Bottega et al., 2012). A escolha do algoritmo impacta diretamente na precisão do mapa gerado, que servirá como base para decisões de manejo localizado.

A dependência espacial, nesse contexto, refere-se à correlação existente entre valores amostrados em posições geográficas próximas, e pode ser quantificada por meio de semivariogramas ou índice de Moran. Essas características são fundamentais para o uso da Krigagem e reforça a necessidade de entender o comportamento espacial da variável estudada (Igaz et al., 2021).

Para avaliar a qualidade dos modelos gerados, a validação cruzada é frequentemente empregada. Essa técnica consiste em dividir o conjunto de dados em partes, utilizando algumas para calibrar o modelo e outras para testá-lo, permitindo mensurar o erro de predição e garantir maior

confiabilidade na interpretação dos mapas gerados. Igaz et al. (2021) utilizaram a validação cruzada para definir o método mais confiável em sua pesquisa.

A qualidade da informação é essencial quando os mapas representam a base de dados para o processo de tomada de decisão. Erros decorrentes da automatização podem levar a interpretações equivocadas, apresentando áreas com produtividades irreais (Menegatti; Molin, 2004). Nesse cenário, a linguagem de programação Python tem se destacado como uma ferramenta poderosa na ciência de dados agrícolas devido à sua flexibilidade e robustez. Bibliotecas como Pandas e NumPy permitem a manipulação e análise de grandes volumes de dados, facilitando a automação de processos complexos (Sapre; Vartak, 2020). Estudos demonstram que a aplicação do Python, combinada a outras ferramentas, pode aprimorar significativamente a eficiência no processamento de dados agrícolas (Damico, 2025).

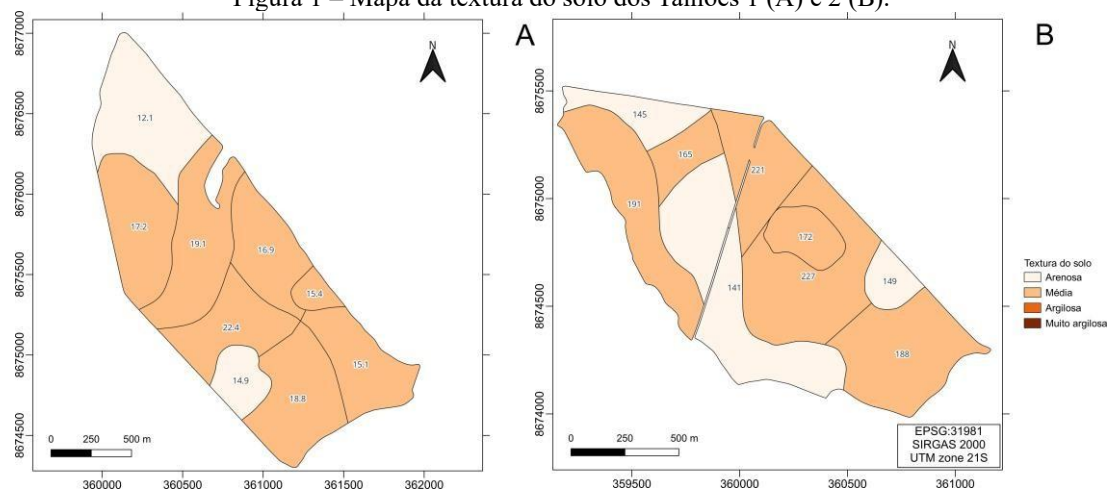
## **4 MATERIAL E MÉTODOS**

### **4.1 ESTUDO DE CASO**

Os dados utilizados neste estudo foram coletados durante a colheita na safra da soja 2024/25, utilizando seis colhedoras da mesma marca, todas com potência nominal de 378 cv e capacidade de armazenamento de 11.600 L. As máquinas estavam equipadas com balança por sensor de carga e plataforma Draper de 35 pés (10,66 m). A coleta foi realizada em dois talhões distintos: o Talhão 1 com uma área de 234,86 ha, e o Talhão 2 com 154,15 ha. As lavouras estão localizadas em uma fazenda no município de Brasnorte, Mato Grosso, cuja classificação climática, segundo Köppen (1936), é Aw (tropical com estação seca no inverno).

O mapa de textura do solo dos talhões, referente à camada de 0 a 10 cm (Figura 1), revela uma predominância de textura média, com teores de argila variando entre 15% e 35%. Em menor proporção, também são identificadas áreas com textura arenosa, caracterizadas por teores de argila inferiores a 15% (Santos et al., 2018).

Figura 1 – Mapa da textura do solo dos Talhões 1 (A) e 2 (B).



As colhedoras são equipadas com um sistema de telemetria responsável pela aquisição dos dados brutos. As informações foram baixadas no formato Comma separated values (CSV). Os atributos utilizados do arquivo CSV podem ser visto na Tabela 1.

Tabela 1 – Informações que devem conter no arquivo CSV.

| Atributo               | Descrição                         |
|------------------------|-----------------------------------|
| cd_equipamento         | Código do equipamento             |
| cd_estado              | Código de estado dos equipamentos |
| vl_rendimento_colheita | Valor de rendimento da colheita   |

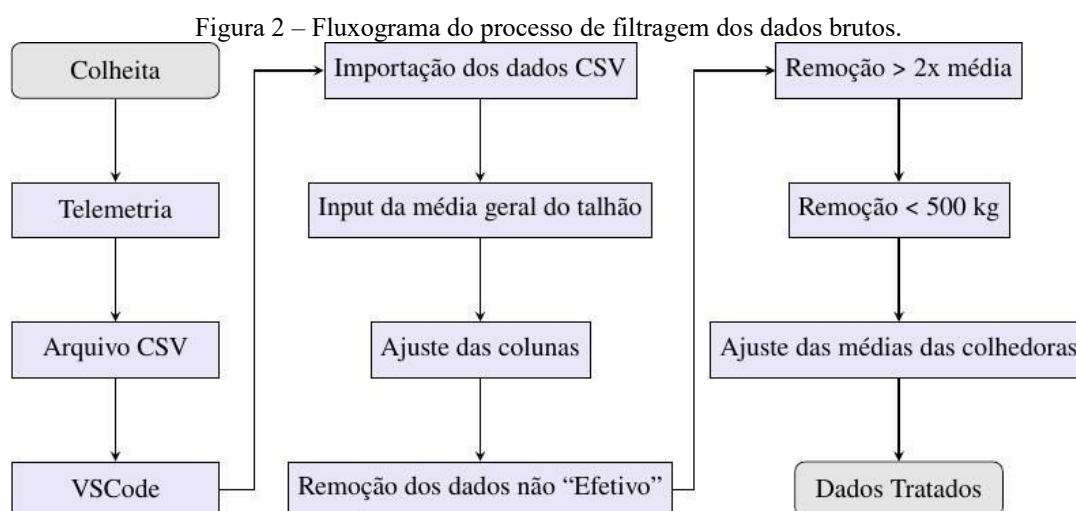
## 4.2 O SCRIPT

Para a realização do processamento dos dados, desenvolveu-se um script em Python (versão 3.9.16) utilizando o software Visual Studio Code (VS Code) (versão 1.98.2), um editor de código-fonte gratuito e de código aberto desenvolvido pela Microsoft. No VS Code, o usuário precisa apenas informar o caminho do arquivo contendo os dados brutos do talhão no formato CSV e a média da produtividade do mesmo, calculada com base no volume de grãos armazenado após a colheita.

O funcionamento do script depende da importação de bibliotecas essenciais para o processamento dos dados. A biblioteca Pandas foi utilizada para diversas operações, incluindo a importação de arquivos no formato CSV, organização, limpeza e análise estatística dos dados. As bibliotecas Matplotlib e Seaborn foram utilizadas para geração dos gráficos. Além disso, as bibliotecas Tkinter e OS, empregadas para desenvolver uma interface gráfica intuitiva, facilitou a interação do usuário e otimizou o processo de seleção dos arquivos diretamente no sistema operacional.

O script desenvolvido segue uma sequência estruturada de etapas (Figura 2). Os dados gerados durante a colheita são enviados ao sistema de telemetria, onde podem ser exportados no formato CSV.

Implementou-se o código no editor de texto VS Code, ambiente o qual desenvolveu-se todo o processo de tratamento dos dados brutos.



No interior do script, inicialmente são importadas as bibliotecas necessárias para o processamento dos dados. Em seguida, realiza-se a importação do arquivo CSV contendo os dados do talhão, bem como a definição de uma variável destinada a armazenar a média geral de produtividade do talhão, a qual foi calculada com base nos dados obtidos no momento do armazenamento dos grãos. Na sequência, são realizados ajustes na formatação de colunas, incluindo a padronização do separador decimal, a reorganização da ordem das colunas e a exclusão daquelas consideradas irrelevantes para a análise.

Após essa etapa de preparação, o script prossegue para a fase de remoção de dados com base nas informações fornecidas pela telemetria. A filtragem inicial considera o campo "cd\_estado", que indica o estado operacional da máquina no momento da aquisição de cada ponto. Durante a coleta, o sistema registra, dentre outros parâmetros, o tipo de atividade sendo realizada pela colhedora, como parada, manobra, deslocamento ou colheita propriamente dita. A Tabela 2 apresenta os códigos possíveis registrados durante esse processo.

Tabela 2 – Descrição dos códigos de estado dos equipamentos.

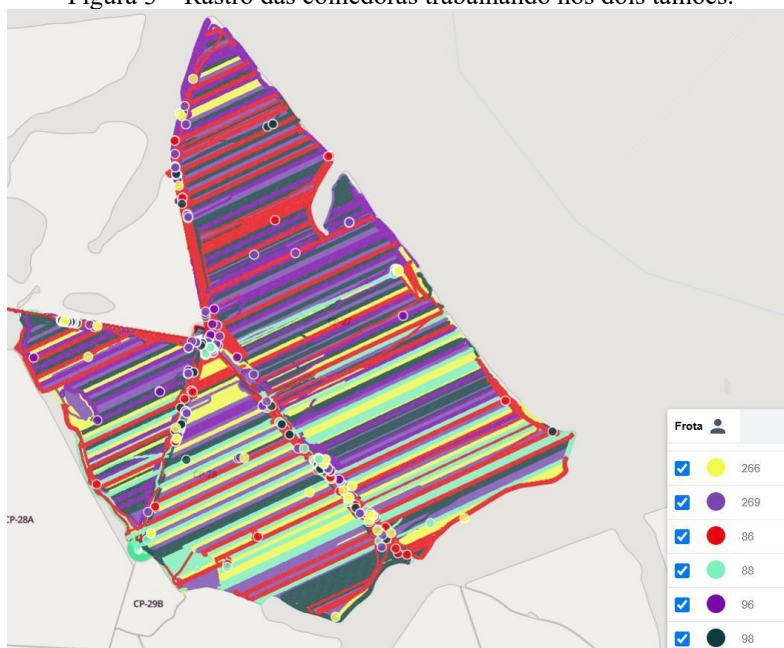
| cd_estado | Descrição                |
|-----------|--------------------------|
| B         | Descarregamento          |
| C         | Desloc p/Descarregamento |
| D         | Deslocamento             |
| E         | Efetivo                  |
| F         | Parada                   |
| M         | Manobra                  |

Neste trabalho, foram mantidas apenas as linhas cujo "cd\_estado" era igual a "Efetivo", pois representam o momento em que a máquina está efetivamente realizando a colheita. Os demais registros foram excluídos, uma vez que não correspondem à operação de colheita e, portanto, podem introduzir ruídos e distorções nos dados de produtividade.

O passo seguinte consistiu na filtragem dos dados relacionados ao "vl\_rendimento\_colheita", que representa a quantidade de grãos colhidos em toneladas por hectare, um procedimento comumente adotado nesse tipo de análise (Sudduth; Drummond, 2007). Para isso, foram excluídos os registros com valores superiores ao dobro da média da coluna "vl\_rendimento\_colheita", bem como aqueles inferiores a 500 kg, visando eliminar possíveis inconsistências nos dados. O valor de 500 kg foi adotado de acordo com os gestores da fazenda.

A última etapa do pré-processamento dos dados foi o ajuste das médias. Para padronização dos dados de todas as colhedoras, foi aplicado um ajuste no "vl\_rendimento\_colheita" de cada máquina com base na média geral do talhão. O código executa um ajuste iterativo dos valores de rendimento da colheita para cada equipamento, garantindo a homogeneização dos dados em relação à média, dentro de uma margem de erro predefinida. Esse ajuste tem como objetivo corrigir possíveis variações nos valores registrados pelos sensores, considerando que, durante a colheita de um mesmo talhão (Figura 3), independentemente do seu tamanho, as máquinas operam em proximidade.

Figura 3 – Rastro das colhedoras trabalhando nos dois talhões.



O detalhamento do código utilizado para ajuste das médias está descrito abaixo:

- Iteração sobre equipamentos: O código primeiro identifica todos os equipamentos em dados "cd\_equipamento".
- Ajuste iterativo: Para cada equipamento, o código entra em um loop while que continua até que a média de rendimento do equipamento esteja dentro da tolerância em relação à média geral.
- Cálculo do fator de ajuste: Dentro do loop, o script calcula um fator de ajuste com base na comparação entre a média de rendimento do equipamento e a média geral. Se a média do equipamento for menor que a média geral, o fator aumenta os valores de rendimento. Se for maior, o fator diminui os valores. Se forem iguais, o fator é 1 (nenhuma mudança).
- Aplicação do fator: O código multiplica os valores de "vl\_rendimento\_colheita" do equipamento pelo fator calculado, ajustando assim o rendimento.
- Verificação da tolerância: Após o ajuste, o código verifica se a diferença absoluta entre a média de rendimento ajustada e a média geral está dentro da tolerância. Se estiver, o loop termina. Caso contrário, o processo de ajuste continua.

Por fim, o script salva os dados gerando um novo arquivo CSV que pode ser aberto em um Sistema de Informação Geográfica (SIG), utilizado para visualizar, editar e analisar dados geográficos.

### 4.3 VISUALIZAÇÃO DOS DADOS

Para visualização/geração dos mapas de produtividade, foi utilizado o QGIS, um SIG gratuito e de código aberto. A interpolação dos dados foi feita pelo método IDW técnica usada para estimar valores em locais não amostrados com base em pontos de dados conhecidos, acessando a "Caixa de Ferramentas de Processamento" no QGIS e digitando no campo de busca "Interpolação IDW".

A Figura 4 mostra os parâmetros utilizados para a composição do mapa de produtividade. A camada vetorial, são os dados tratados (arquivo CSV) gerados pelo script e transformados em um arquivo shapefile em coordenadas UTM (Universal Transverse Mercator) pelo próprio QGIS; o atributo de interpolação é a coluna que representa o rendimento da colheita; a distância para o coeficiente P foi mantido o padrão do interpolador; a extensão foi um arquivo shapefile do limite do talhão e o tamanho do raster de saída foi com base na resolução dos pixels X e Y de 10 m.

Figura 4 – Parâmetros da interpolação IDW.

**Interpolação IDW**

Parâmetros Log

Camada(s) de entrada

Camada vetorial: dados\_tratados\_vs\_cp27\_utm

Atributo de interpolação: 1.2 vl\_rendime

☐ Usar Coordenada Z para interpolação

| Camada vetorial            | Atributo   | Tipo   |
|----------------------------|------------|--------|
| dados_tratados_vs_cp27_utm | vl_rendime | Pontos |

Distância para coeficiente P: 2.000000

Extensão: 359934.8571,361973.8553,8674299.2982,8677004.9267 [EPSG:31981]

Tamanho do raster de saída

Linhas: 272 Colunas: 205

Tamanho do pixel X: 10.000000 Tamanho do pixel Y: 10.000000

Interpolado: [Salvar em arquivo temporário]

☒ Abrir arquivo de saída depois executar o algoritmo

0%

Avançado Executar processo em Lote... Executar Fechar Ajuda

**Interpolação IDW**

Gera a interpolação Ponderação pelo Inverso da Distância (IDW) de uma camada de pontos vetorial.

Pontos amostrais são ponderados durante a interpolação para que a influência de um ponto em relação a outro caia com a distância do ponto desconhecido criado.

## 5 RESULTADOS E DISCUSSÃO

A Tabela 3 apresenta as estatísticas descritivas dos dados de rendimento nos dois talhões, antes e depois do processo de limpeza e ajuste das médias. Observou-se um aumento na média em ambos os talhões após o tratamento, o que indicou que os dados brutos continham uma quantidade significativa de valores abaixo da média. Outros estudos também obtiveram aumento da média após a limpeza dos dados no software que eles desenvolveram (Sudduth; Drummond, 2007; Gimenez; Molin, 2004). A mediana zerada nos dois casos dos dados brutos aponta que, pelo menos, 50% dos registros de rendimento oriundos da telemetria apresentaram valor igual a zero. Além disso, houve uma redução no desvio padrão (St dev), o mesmo aconteceu nos estudos de Sudduth e Drummond (2007), Gimenez e Molin (2004), indicando maior uniformidade nos dados tratados. Os valores máximos caíram de 44,8 t/ha e 42,2 t/ha para 8,90 t/ha e 9,43 t/ha, respectivamente, evidenciando a remoção eficaz dos outliers e resultando em um conjunto de dados mais homogêneo e confiável para análises posteriores.

Tabela 3 – Estatísticas dos dados de rendimento da colheita.

| Estatística | Dados Brutos | Dados Tratados | Dados Brutos | Dados Tratados |
|-------------|--------------|----------------|--------------|----------------|
|             | Talhão 1     |                | Talhão 2     |                |
| Contagem    | 23291        | 9595           | 21055        | 6977           |
| Média       | 2,674        | 3,672          | 2,520        | 3,823          |
| Mediana     | 0,000        | 3,985          | 0,000        | 4,033          |
| St dev      | 3,469        | 1,373          | 3,810        | 1,484          |
| Mínimo      | 0,000        | 0,397          | 0,000        | 0,381          |
| Máximo      | 44,800       | 8,899          | 42,200       | 9,432          |
| Q1          | 0,000        | 2,989          | 0,000        | 3,068          |
| Q3          | 5,200        | 4,544          | 4,900        | 4,717          |
| IQR         | 5,200        | 1,556          | 4,900        | 1,649          |

Outra observação relevante refere-se à expressiva redução no número de dados após o processo de limpeza: aproximadamente 58,8% no Talhão 1 e 66,9% no Talhão 2 foram descartados. Para fins de comparação, no estudo conduzido por Gimenez e Molin (2004), as taxas de descarte foram de 41% e 21%, enquanto Sudduth e Drummond (2007) relataram a remoção de 13% a 27% dos dados em cinco talhões com áreas variando entre 11 ha e 48 ha, utilizando o software desenvolvido por eles.

A maior taxa de descarte observada neste trabalho pode estar relacionada à ocorrência de falhas em um dos sensores de uma das colhedoras, que resultou no registro de rendimento igual a zero em todos os pontos coletados por esse equipamento. Outro fato relevante foi a exclusão dos dados com o "c\_estado" diferentes de efetivo, foram removidos 35,4% dos dados do Talhão 1 e 39,9% do Talhão 2, em alguns sistemas de telemetria esses dados nem aparecem nos dados brutos.

As Tabelas 4 e 5 apresentam a quantidade dos dados excluídos, após o processo de limpeza, variando de 42,23% a 78,27% para o Talhão 1 e de 46,49% a 80,08% para o Talhão 2, excluindo o equipamento 266 que foi perdido 100% dos dados. Com o resultado foi possível verificar que alguns sensores podem estar com problemas ou descalibrados. Outra informação são as médias, de rendimento e umidade, por máquina em cada etapa do processamento.

Tabela 4 – Dados das colhedoras antes, durante e após tratamento (Talhão 1).

| cd_<br>equipa-<br>mento | Dados Brutos |                          |                    | Dados Tratados<br>sem ajuste da média |                          |                    | Dados Tratados<br>com ajuste<br>da média |
|-------------------------|--------------|--------------------------|--------------------|---------------------------------------|--------------------------|--------------------|--|
|                         | Contagem     | vl                       | vl                 | Contagem                              | vl                       | vl                 | vl                                       |
|                         |              | _rendimento<br>_colheita | _umidade<br>_graos |                                       | _rendimento<br>_colheita | _umidade<br>_graos |  |
| 86                      | 5288         | 3,152                    | 13,16              | 2366                                  | 6,649                    | 17,64              | 3,680                                    |
| 88                      | 2214         | 2,473                    | 10,82              | 1271                                  | 4,231                    | 14,79              | 3,682                                    |
| 96                      | 4284         | 2,647                    | 15,16              | 2475                                  | 4,450                    | 19,33              | 3,677                                    |
| 98                      | 4842         | 1,890                    | 5,54               | 2628                                  | 3,370                    | 6,89               | 3,657                                    |
| 269                     | 3935         | 4,991                    | 14,63              | 855                                   | 4,616                    | 18,99              | 3,669                                    |
| 266                     | 2728         | 0,000                    | 13,87              | -                                     | -                        | -                  | -  |

Tabela 5 – Dados das colhedoras antes, durante e após tratamento (Talhão 2).

| cd_<br>equipa-<br>mento | Dados Brutos |                          |                    | Dados Tratados<br>sem ajuste da média |                          |                    | Dados Tratados<br>com ajuste<br>da média |
|-------------------------|--------------|--------------------------|--------------------|---------------------------------------|--------------------------|--------------------|--|
|                         | Contagem     | vl                       | vl                 | Contagem                              | vl                       | vl                 | vl                                       |
|                         |              | _rendimento<br>_colheita | _umidade<br>_graos |                                       | _rendimento<br>_colheita | _umidade<br>_graos |  |
| 86                      | 3515         | 2,970                    | 17,76              | 1410                                  | 6,766                    | 22,13              | 3,790                                    |
| 88                      | 3394         | 2,550                    | 11,04              | 1801                                  | 4,631                    | 15,59              | 3,832                                    |
| 96                      | 2947         | 1,569                    | 7,60               | 903                                   | 4,997                    | 19,46              | 3,808                                    |
| 98                      | 3835         | 1,899                    | 5,56               | 2052                                  | 3,378                    | 6,93               | 3,839                                    |
| 269                     | 4071         | 5,418                    | 16,23              | 811                                   | 4,367                    | 20,09              | 3,838                                    |
| 266                     | 3293         | 0,000                    | 16,02              | -                                     | -                        | -                  | -  |

Os mapas de produtividade gerados a partir dos dados brutos, baixados diretamente do sistema de telemetria da fazenda, estão apresentados nas Figuras 5 e 6.

Figura 5 – Mapa de produtividade em pontos do Talhão 1 (dados brutos).

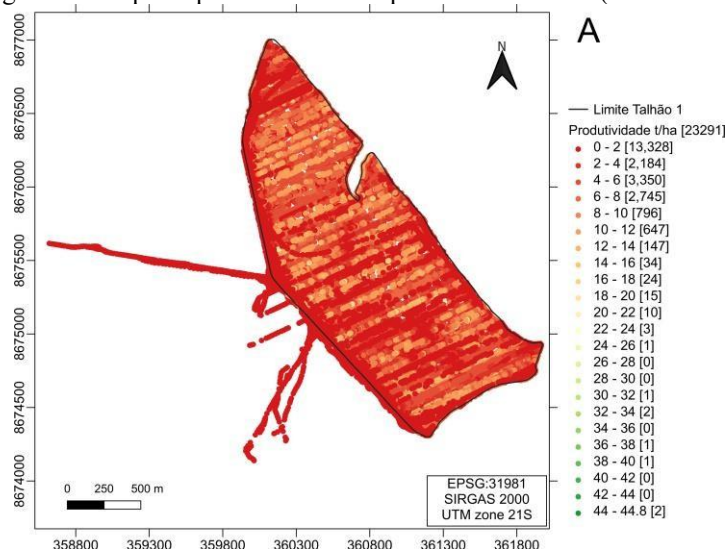
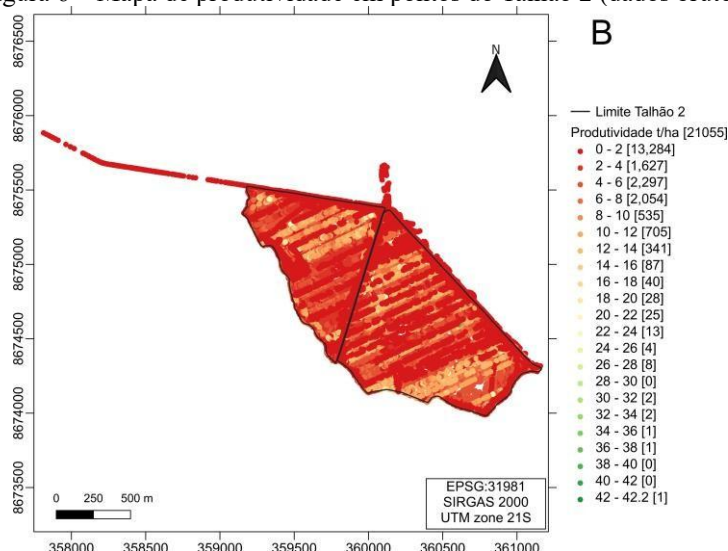


Figura 6 – Mapa de produtividade em pontos do Talhão 2 (dados brutos).



Observou-se, em ambos os talhões, a presença de pontos fora dos limites geográficos das áreas colhidas. No estudo conduzido por Menegatti e Molin (2004), erros de posicionamento representaram de 0,1% a 7,8% dos dados, evidenciando a recorrência desse tipo de inconsistência. Também há uma expressiva concentração de dados com valores inferiores a 2,0 t/ha, representando 57,22% dos pontos no Talhão 1 e 63,09% no Talhão 2. Como consequência dessas distorções, as médias de produtividade calculadas com base nos dados brutos foram de apenas 2,67 t/ha e 2,52 t/ha, respectivamente, valores significativamente inferiores às médias reais obtidas a partir da pesagem dos grãos no armazém, 3,68 t/ha e 3,84 t/ha, já ajustadas para 13% de umidade.

Já nas Figuras 7 e 8, são apresentados os mapas de produtividade gerados após o tratamento dos dados utilizando o script proposto.

Figura 7 – Mapa de produtividade em pontos do Talhão 1 (dados tratados).

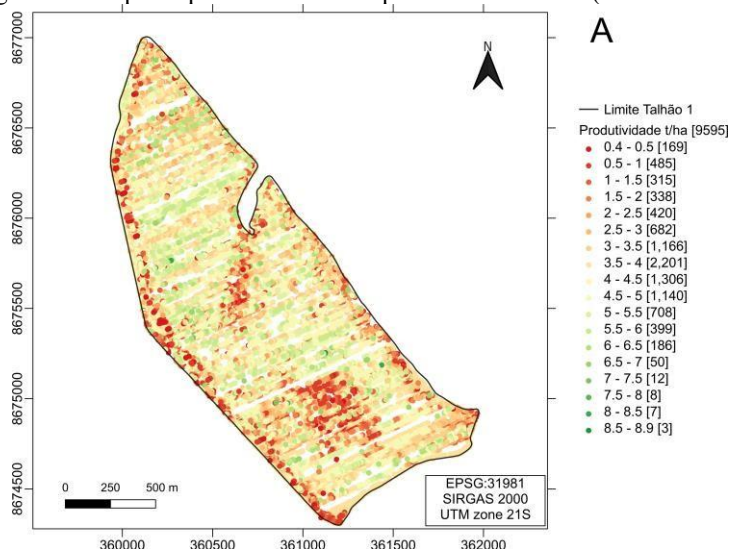
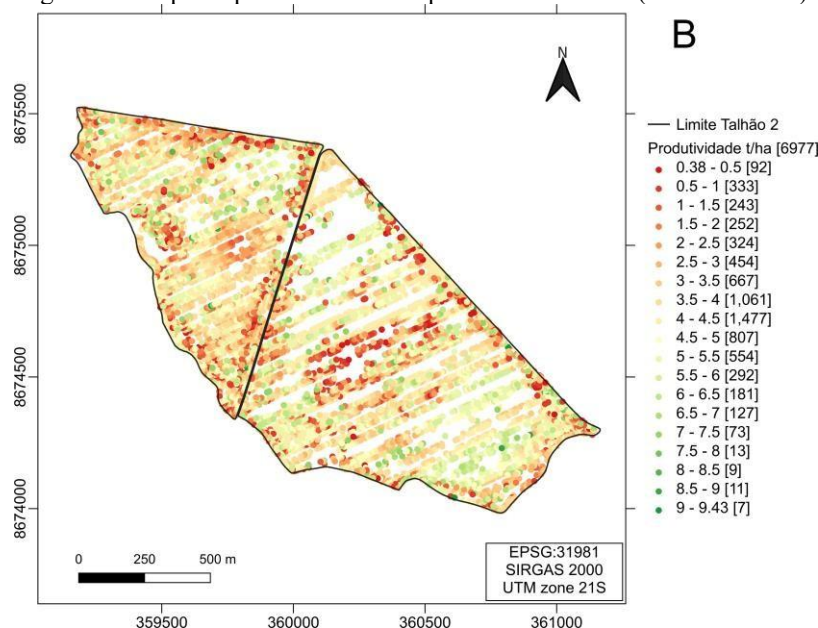


Figura 8 – Mapa de produtividade em pontos do Talhão 2 (dados tratados).



Os pontos localizados fora dos limites dos talhões foram excluídos, não por meio de um critério geográfico específico, mas por apresentarem valores abaixo do limite mínimo definido para rendimento.

Observou-se, especialmente nas bordaduras dos talhões, a presença de áreas com baixa produtividade. No Talhão 1, isso ocorre principalmente nas porções inferior e esquerda, enquanto no Talhão 2, observa-se na região superior — justamente na divisa entre os dois talhões, onde há uma estrada municipal. É provável que essa área funcione como zona de manobras e parada das máquinas ao final do expediente, o que pode ter provocado maior compactação do solo e, consequentemente, queda na produtividade.

Outro aspecto notável são as “falhas” ou lacunas visíveis em algumas regiões dos mapas. Isso se deve, além da natural redução de pontos ocasionada pela limpeza dos dados, ao fato de que uma das colhedoras registrou todos os valores de rendimento como zero, resultando em faixas sem dados válidos.

Mesmo antes da aplicação de técnicas de interpolação, os mapas já permitiam a identificação das zonas com menor produtividade, evidenciando o potencial do script para fornecer uma visualização inicial mais confiável da variabilidade espacial das lavouras.

Por fim foi gerado o mapa de produtividade após a realização da interpolação IDW (Figura 9 e 10), destacando as áreas de baixa em tons vermelhos e alta produtividade tons verdes. São esses mapas que são utilizados para a geração das zonas de manejo juntamente com outros.

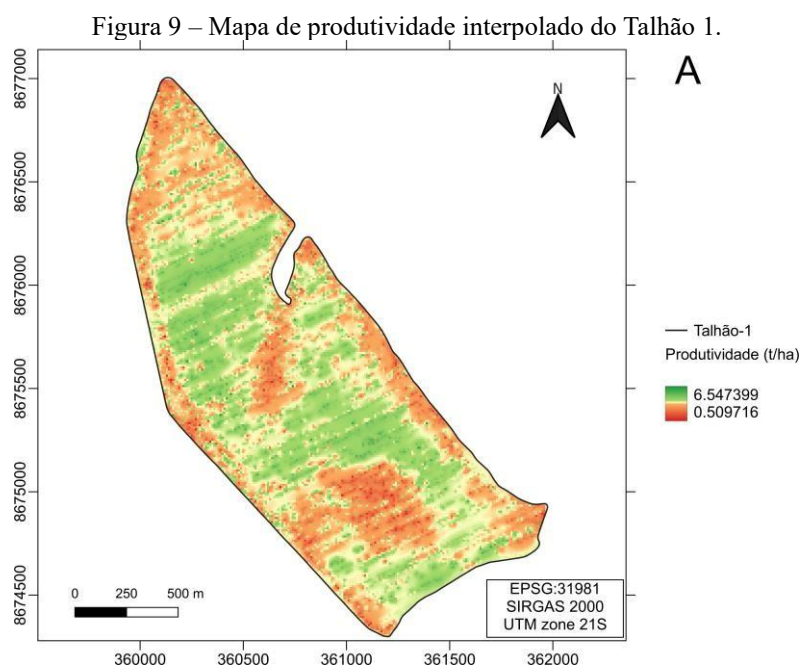
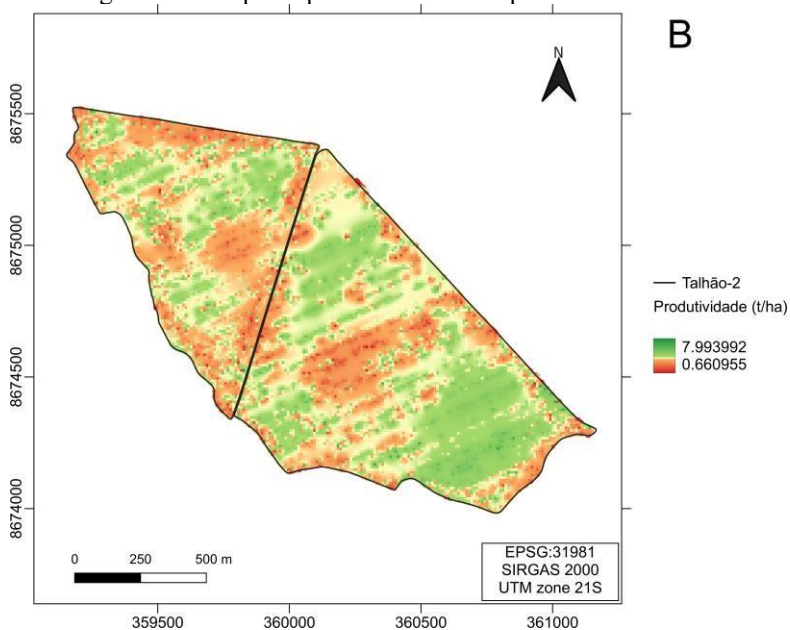


Figura 10 – Mapa de produtividade interpolado do Talhão 2.



## 6 CONCLUSÃO

O desenvolvimento e aplicação do script em Python para limpeza dos dados brutos de produtividade agrícola demonstraram-se eficazes no tratamento de informações coletadas via telemetria. A proposta contribuiu para a automação do processo de pré-processamento, reduzindo interferências causadas por erros nos dados brutos e aumentando a confiabilidade das análises.

A ferramenta se mostrou uma alternativa acessível e eficiente, permitindo maior controle e padronização na preparação dos dados para geração de mapas de produtividade. A aplicação desse tipo de solução é fundamental para garantir a qualidade da informação em projetos de AP, apoiando decisões agrônômicas mais assertivas.

Comparado aos dados brutos, os dados tratados fornecem informações mais condizentes com a realidade dos talhões, permitindo uma melhor tomada de decisão para os gestores, possibilitando o uso do mapa interpolado à aplicações em taxa variável.

## REFERÊNCIAS

- BASSOI, L. H. et al. Agricultura de precisão e agricultura digital. TECCOGS: Revista Digital de Tecnologias Cognitivas, São Paulo, n. 20, maio 2020. Disponível em: <https://revistas.pucsp.br/teccogs/article/view/48542>. Acesso em: 3 jul. 2025.
- BOTTEGA, E. et al. Uso de diferentes interpoladores na geração de modelo digital de elevação. In: CONGRESSO BRASILEIRO DE AGRICULTURA DE PRECISÃO - CONBAP, 2012, Ribeirão Preto. Anais [...]. Ribeirão Preto: [s.n.], 2012.
- CHANDEL, N. et al. Idw interpolation of soybean yield data acquired by automated yield monitor. International Journal for Science and Emerging Technologies with Latest Trends, [S.l.], v. 13, p. 36-45, 2013.
- COSTA, F. et al. Uma visão geral da aplicação de sensores em máquinas agrícolas. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 10., 2015, Ponta Grossa. Anais [...]. Ponta Grossa: [s.n.], 2015.
- DAMICO, J. Methodology for cleaning, quality, and normalization of vegetation index data derived from Sentinel 2. [S.l.: s.n.], 2025. Disponível em: [https://www.academia.edu/126945828/Methodology\\_for\\_Cleaning\\_Quality\\_and\\_Normalization\\_of\\_Vegetation\\_Index\\_Data\\_Derived\\_from\\_Sentinel\\_2](https://www.academia.edu/126945828/Methodology_for_Cleaning_Quality_and_Normalization_of_Vegetation_Index_Data_Derived_from_Sentinel_2). Acesso em: 3 jul. 2025.
- FIZZA, K. et al. Evaluating sensor data quality in internet of things smart agriculture applications. IEEE Micro, [S.l.], v. 42, n. 1, p. 51-60, 2022. Disponível em: <http://dx.doi.org/10.1109/MM.2021.3137401>. Acesso em: 3 jul. 2025.
- GIMENEZ, L. M.; MOLIN, J. P. Algoritmo para redução de erros em mapas de produtividade para agricultura de precisão. Revista Brasileira de Agrocomputação, [S.l.], v. 2, n. 1, p. 5-10, 2004.
- GRISSE, R. D. et al. Yield monitor accuracy: Successful farming magazine case study. Applied Engineering in Agriculture, [S.l.], v. 18, n. 2, p. 147, 2002.
- IGAZ, D. et al. The evaluation of the accuracy of interpolation methods in crafting maps of physical and hydro-physical soil properties. Water, [S.l.], v. 13, p. 212, 2021.
- KÖPPEN, W. Das geographische System der Klimate. Berlin: Gebrüder Borntraeger, 1936. v. 1C, p. 1-44.
- LI, M. et al. Development of an intelligent yield monitor for grain combine harvester. In: LI, D.; WANG, B. (ed.). Artificial intelligence applications and innovations. New York: Springer-Verlag, 2005. p. 663-670. (IFIP — The International Federation for Information Processing). Disponível em: [http://link.springer.com/10.1007/0-387-29295-0\\_72](http://link.springer.com/10.1007/0-387-29295-0_72). Acesso em: 3 jul. 2025.
- MALDANER, L. F.; MOLIN, J. P.; SPEKKEN, M. Methodology to filter out outliers in high spatial density data to improve maps reliability. Scientia Agricola, [S.l.], v. 79, n. 1, 2022. Disponível em: <https://www.scielo.br/j/sa/a/>. Acesso em: 3 jul. 2025.

MENEGATTI, L. A. A.; MOLIN, J. P. Remoção de erros em mapas de produtividade via filtragem de dados brutos. *Revista Brasileira de Engenharia Agrícola e Ambiental*, [S.l.], v. 8, n. 1, p. 126-134, 2004. Disponível em: <https://www.scielo.br/j/rbeaa/a/>. Acesso em: 3 jul. 2025.

PEREIRA, F. J.; MOLIN, J. P. Bancada de ensaio para avaliação de monitores de produtividade de grãos. *Engenharia Agrícola*, [S.l.], v. 23, n. 3, p. 568-578, 2003.

SANTI, A. et al. Definição de zonas de produtividade em áreas manejadas com agricultura de precisão. *Revista Brasileira de Ciências Agrárias*, [S.l.], v. 8, n. 3, p. 510-515, 2013. Disponível em: <http://www.agraria.pro.br/ojs32/index.php/RBCA/article/view/v8i3a2489>. Acesso em: 3 jul. 2025.

SANTOS, H. G. et al. Sistema brasileiro de classificação de solos. Brasília: Embrapa, 2018.

SAPRE, A.; VARTAK, S. Scientific computing and data analysis using NumPy and Pandas. [S.l.], v. 7, n. 12, 2020.

SILVA, W. V. R.; SILVA-MANN, R. Precision agriculture in Brazil: current situation, challenges and perspectives. *Research, Society and Development*, [S.l.], v. 9, n. 11, p. e1979119603, 2020. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/9603>. Acesso em: 3 jul. 2025.

SUDDUTH, K. A.; DRUMMOND, S. T. Yield editor: software for removing errors from crop yield maps. *Agronomy Journal*, [S.l.], v. 99, n. 6, p. 1471-1482, 2007. Disponível em: <https://acsess.onlinelibrary.wiley.com/doi/abs/10.2134/agronj2006.0326>. Acesso em: 3 jul. 2025.

TSCHIEDEL, M.; FERREIRA, M. F. Introdução à agricultura de precisão: conceitos e vantagens. *Ciência Rural*, Santa Maria, v. 32, n. 1, p. 159-163, 2002. Disponível em: <https://doi.org/10.1590/S0103-84782002000100027>. Acesso em: 3 jul. 2025.

VEGA, A. et al. Protocol for automating error removal from yield maps. *Precision Agriculture*, [S.l.], v. 20, n. 5, p. 1030-1044, 2019. Disponível em: <https://doi.org/10.1007/s11119-018-09632-8>. Acesso em: 3 jul. 2025.