Andrew Engellant
Master of Science in Business Analytics
University of Montana
March 2025

Capstone Project: Vegetation Health Monitoring

Executive Summary

I designed and implemented a comprehensive data engineering solution that transforms raw satellite imagery into actionable vegetation health insights for Missoula County, Montana. This Vegetation Health Monitoring Tool addresses the growing need for timely, accurate environmental monitoring in agricultural management and forestry sectors.

By leveraging NASA's Sentinel-2 satellite data and implementing a robust data pipeline, I created a user-friendly web application that allows stakeholders to visualize vegetation health patterns across Missoula County throughout the growing season. The system processes complex satellite imagery into standardized vegetation indices, providing valuable insights into plant health, biomass distribution, and seasonal changes.

This report details the technical development process, the system's capabilities, and its potential applications for land management professionals, agricultural consultants, and environmental researchers. The solution demonstrates how modern data engineering practices can transform complex remote sensing data into accessible, valuable information for decision-makers.

Introduction

Project Background

Climate change, resource management challenges, and growing demands for sustainable land use have created an urgent need for sophisticated monitoring tools. While satellite technology provides frequent remote monitoring of our environment, the technical complexity of processing this data has traditionally limited its accessibility to specialized scientists.

For my capstone project, I sought to bridge this gap by creating a data pipeline and visualization tool that transforms complex satellite data into intuitive, actionable insights about vegetation health. This project represents the culmination of my data analytics education, combining skills in data engineering, web application development, and data visualization.

Problem Statement

Land managers, agricultural consultants, and environmental researchers need accessible, up-to-date information about vegetation health to make informed decisions. However, several challenges have impeded this access:

1. Traditional vegetation monitoring relies on infrequent and labor-intensive field assessments
2. Raw satellite imagery requires significant preprocessing before it becomes useful
3. Technical barriers prevent non-specialists from leveraging satellite data effectively

I addressed these challenges by creating an end-to-end solution that automates the acquisition, processing, and visualization of vegetation health data derived from satellite imagery, making powerful environmental insights available to non-technical users.

Project Significance

This project's significance extends beyond technical innovation. By democratizing access to vegetation health data, this tool enables:

- Agricultural professionals to monitor crop health across large areas with minimal field visits
- Forestry managers to detect early signs of stress or disease in forest stands
- Environmental researchers to track seasonal patterns and long-term trends
- Land conservation organizations to monitor the health of protected areas

System Architecture

I designed the system architecture with three primary components that work together to deliver a seamless user experience:

1. Data Processing Pipeline: Acquires, processes, and transforms satellite imagery into vegetation health metrics
2. Web Server: Serves processed geospatial data for visualization
3. Interactive Web Application: Provides an intuitive interface for exploring and analyzing vegetation health data

Each component performs specialized functions while maintaining clear integration points with other parts of the system. The architecture follows modern data engineering principles of modularity, automation, and scalability.

Data Processing Pipeline

The data processing pipeline forms the foundation of the system, handling the complex transformation of raw satellite data into standardized, analysis-ready formats. The pipeline performs several critical functions:

1. Data Acquisition: Uses a given geographic boundary (Missoula County in this project) and automatically searches and retrieves Sentinel-2 satellite imagery that are within or intersect with the geographic boundaries.
2. Geospatial Processing: Extracts the red, green, blue and near-infrared imagery layers, as well as the SCL image. Data outside of the geographic boundaries are then removed and all images are transformed to a consistent resolution and coordinate system.
3. Cloud Masking: Uses the SCL to identify and remove cloud-covered areas that would otherwise distort the analysis
4. Vegetation Index Calculation: Computes the Normalized Difference Vegetation Index (NDVI), a proven metric for vegetation health assessment
5. Data Standardization: Converts processed data into Cloud Optimized GeoTIFF (COG) format with optimized structure for web delivery
6. Statistical Summarization: Calculates key metrics including median NDVI values, variance, and vegetation abundance percentages

This pipeline runs automatically and can be implemented to run on a scheduled basis to fetch up-to-date vegetation health information for continual monitoring, or it can be used to retrieve and process all satellite imagery over a given historic date range to build a database of vegetation health information.

Web Server

The web server component acts as the bridge between the processed data and the user interface. I implemented this using FastAPI and TiTiler, which provide:

1. Tile-Based Data Serving: Large geographic boundaries such as Missoula Country contain tens of millions of pixels, which is too large to effectively load on a web application. This server converts the large geospatial datasets into small chunks called map tiles. As the user pans across the map, the server only delivers the tiles that are in view. This architecture ensures fast performance even when dealing with large geospatial datasets that would otherwise be too cumbersome for web-based visualization.
2. Dynamic Assessment of Available Days: Since the Sentinel-2's orbit only passes a given geography every few days, processed imagery is only available for certain given days. This server assesses the folder structure of the directory containing these images and provide information about the available days to the web application, which dictates what days appear in the user day selection feature.

Interactive Web Application

The web application provides the user-facing interface for exploring vegetation health data. I developed this using Shiny for Python, creating:

1. Interactive Map: Displays full color satellite imagery and NDVI layers with pan and zoom capabilities
2. Temporal Navigation: Allows users to select specific dates and observe changes over time
3. Data Summary Dashboards: Presents basic statistical information about vegetation health in easy-to-understand formats
4. Vegetation Health Indicators: Categorizes and displays the status of vegetation using intuitive visual cues

The interface design emphasizes accessibility for non-technical users while still providing the depth of information needed for professional analysis.

Data Processing Methodology

Satellite Data Source

For this project, I selected European Space Agency's Sentinel-2 satellite imagery as the primary data source for several compelling reasons:

1. Spatial Resolution: Sentinel-2 provides 10-meter resolution in key spectral bands, offering sufficient detail for county-level vegetation analysis
2. Temporal Frequency: The satellite orbital path captures new imagery approximately every 3 days, allowing for near-continuous monitoring during the growing season
3. Spectral Capabilities: Sentinel-2 records data in multiple spectral bands that are ideal for vegetation analysis, including red and near-infrared wavelengths
4. Open Data Policy: The European Space Agency provides Sentinel-2 data free of charge, making it an economical choice for operational use
5. Established Ecosystem: A mature ecosystem of tools and methodologies exists for working with Sentinel-2 data, including a Python integrated search API

I accessed this data through the Element84 Earth Search API, which provides a standardized interface to query and download Sentinel-2 imagery based on geographic location and date parameters.

Vegetation Health Metrics

To quantify vegetation health, I implemented several key metrics:

1. Normalized Difference Vegetation Index (NDVI): The primary metric, calculated as (NIR - Red) / (NIR + Red), where NIR is the near-infrared reflectance and Red is the visible red reflectance. This index leverages the fact that healthy vegetation has higher chlorophyl composition, which strongly absorbs red light and reflects near-infrared light. Values are standardized and can range from -1 to 1, however most vegetation will reflect values between 0.2 and 1, with higher values indicating healthier vegetation.
2. Vegetation Abundance Percentage: The proportion of the geography area covered by vegetation (NDVI > 0.2), indicating the extent of healthy plant cover.
3. Variability Coefficient: A measure of how heterogeneous the vegetation health is across the landscape, calculated as the statistical variance of NDVI values.
4. Cloud Cover Percentage: An important quality indicator that shows how much of the area was obscured by clouds during satellite data acquisition. This measure represents the percentage of the total geography that was obscured by clouds.

These metrics provide complementary perspectives on vegetation health, enabling more nuanced analysis than any single measure alone.

Technical Challenges and Solutions

The development process presented several technical challenges that required innovative solutions:

1. Cloud Masking: Clouds in satellite imagery can significantly distort vegetation indices and skew statistical measures. I implemented a classification-based approach using Sentinel-2's Scene Classification Layer (SCL) to identify and exclude cloud-covered pixels from analysis. Cloud pixels are included in the full color layer of the satellite image explorer tool but have been masked in the NDVI layer. This approach allows users visualize both cloud cover and which areas were removed from NDVI analysis. I chose to not impute missing NDVI values upon client request.
2. Geometric Alignment: As the Sentinel-2 satellite passes an area, it captures square images of the landscape as objects called scenes. When working with large geographies such as Missoula County, several scenes must be combined to cover the full geography. This process of combining multiple satellite images is called mosaicking and requires precise geometric alignment. I used the EPSG:32611 (UTM Zone 11N) projection for processing and EPSG:3857 (Web Mercator) for visualization to ensure consistency.
3. Data Volume: Processing high-resolution satellite imagery requires significant computational resources. I implemented chunked processing with the Dask framework to efficiently handle large datasets without exhausting system memory
4. Quality Control: Variations in cloud coverage and orbit paths can affect data quality. I established filtering criteria based on cloud coverage percentage and pixel count to ensure that only high-quality observations influence the analysis. For

the timeseries visualizations, only data from days where the Sentinel-2 satellite captured imagery of the full county landscape and cloud cover was less than 50% are included.

5. Performance Optimization: To enable smooth web visualization, I implemented image overviews (also referred to as pyramids). As the user zooms out, the image resolution decreases. I used averages to creates overviews at 20, 40, 80, 160, and 320-meter resolutions. I also optimized the data structure of the satellite images, storing them as Cloud Optimized GeoTIFFs, which uses appropriate compression and tiling schemes.

These technical decisions were crucial for creating a system that balances analytical clarity with operational efficiency.

System Capabilities and Features

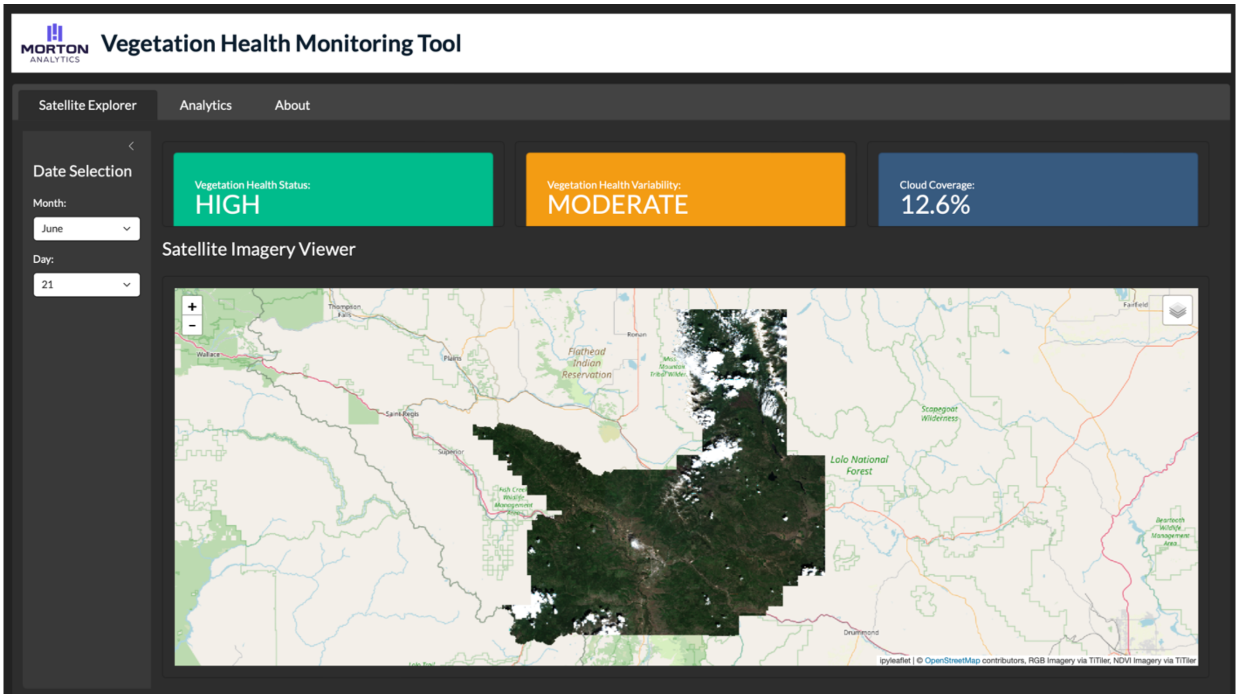
The Vegetation Health Monitoring Tool provides several key capabilities designed for practical use by land management professionals:

Interactive Satellite Exploration

The Satellite Explorer interface allows users to:

- View true-color satellite imagery showing the actual appearance of the landscape
- Overlay vegetation health (NDVI) visualization using a red-yellow-green color gradient
- Toggle between different map layers to compare visual and analytical perspectives
- Examine specific areas of interest through pan and zoom functionality
- Access imagery from different dates throughout the 2024 growing season

This visualization capability brings the power of satellite remote sensing to users without requiring specialized GIS software or technical expertise.



MORTON ANALYTICS Vegetation Health Monitoring Tool

Satellite Explorer
Analytics
About

Date Selection

Month: June

Day: 21

Vegetation Health Status: HIGH

Vegetation Health Variability: MODERATE

Cloud Coverage: 12.6%

Satellite Imagery Viewer

+
-
Thompson Falls
Wallace
Plains
Flathead Indian Reservation
Ronan
Saint Regis
Superior
Lolo National Forest
Scapegoat Wilderness
Fairfield
Drummond
Bearteeth Wildlife Management Area
ipyleaflet | © OpenStreetMap contributors, RGB Imagery via TiTiler, NDVI Imagery via TiTiler

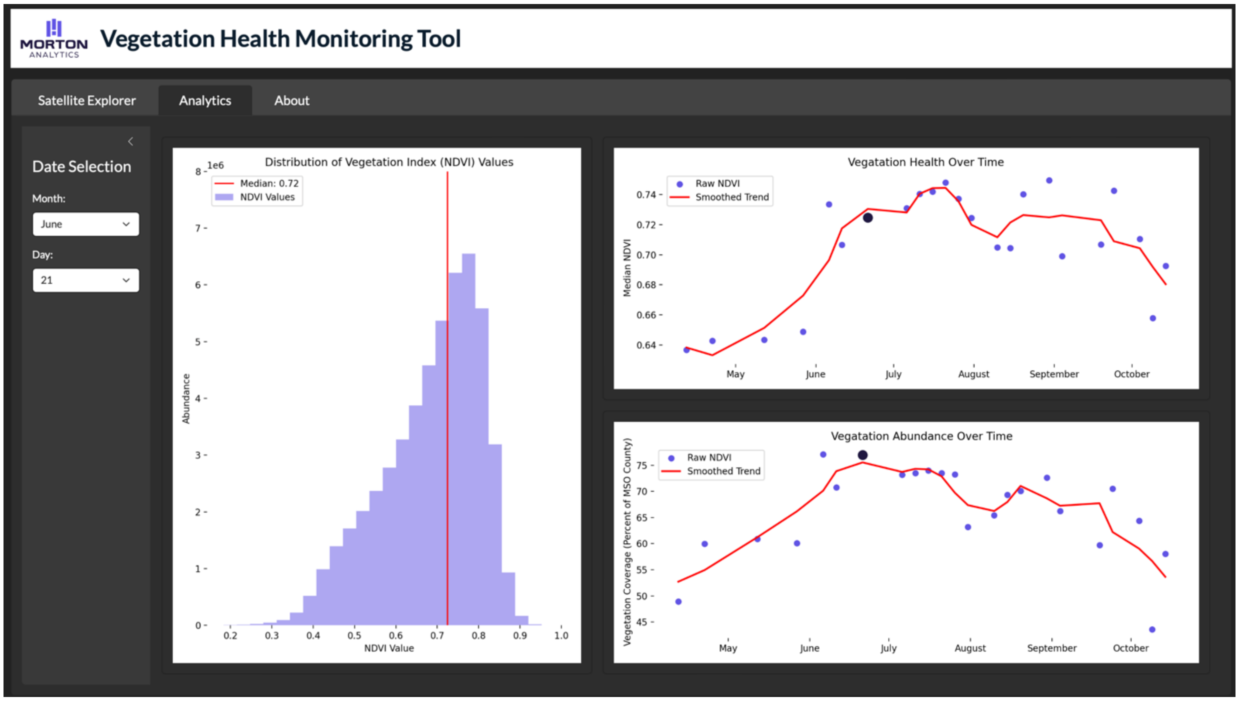
Figure 1: User interface of the interactive satellite explorer.

Vegetation Health Analytics

The Analytics dashboard provides deeper insights through:

- Histogram visualization showing the distribution of vegetation health values
- Time-series graphs displaying seasonal vegetation patterns
- Comparative analysis of current conditions versus historical trends
- Statistical indicators summarizing overall vegetation status
- Automated categorization of vegetation health into interpretable classes (Low, Moderate, High, Very High)

These analytics transform raw data into actionable information, supporting informed decision-making about land management practices.



Month: June

Day: 21

Figure 2: User interface of the analytics dashboard displaying NDVI histogram (left), vegetation health timeseries (right, upper), and vegetation abundance timeseries (right, lower).

Temporal Analysis

The system's temporal capabilities enable users to:

- Track vegetation changes throughout the growing season (April through October)
- Identify critical periods of growth, peak health, and senescence
- Detect unusual patterns that may indicate stress, disease, or management issues
- Understand the timing of seasonal transitions for different land cover types
- Establish baseline conditions for future comparison

This temporal dimension adds significant value by capturing the dynamic nature of vegetation systems that static assessments would miss.

Applications and Use Cases

The Vegetation Health Monitoring Tool supports numerous practical applications across different sectors:

Agricultural Consulting

Agricultural consultants can use this tool to:

- Monitor crop health and development across multiple client properties
- Identify areas of stress or underperformance that require targeted intervention
- Compare the effectiveness of different management practices based on vegetation response
- Provide data-backed recommendations to clients
- Document seasonal patterns to inform future planning

For example, a consultant could quickly identify irrigation issues in a client's field based on localized patterns of reduced NDVI values, potentially saving a crop before visible symptoms appear. Figure 3 demonstrates this capability. The left image shows satellite imagery of an anonymous crop. Applying the NDVI layer (right) reveals a significant discrepancy in health in the northern third of the crop.

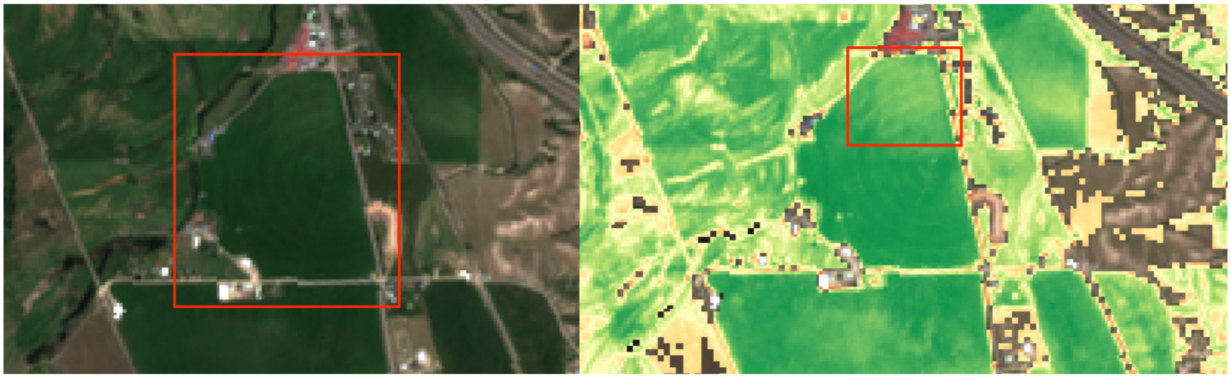


Figure 3: Example use of the Vegetation Health Monitoring Tool identifying a crop deficiency on an anonymous crop displayed in full-color base layer imagery (left) and NDVI layering (left).

Forestry Management

For forestry professionals, the system enables:

- Early detection of forest health issues including disease and pest outbreaks
- Assessment of recovery following wildfires or logging operations
- Monitoring of seasonal growth patterns in different forest stands
- Evaluation of the effectiveness of management interventions
- Documentation of long-term forest condition trends

A forestry manager could use historical NDVI patterns to determine optimal timing for silvicultural treatments based on when trees are most actively growing.

Environmental Monitoring

Environmental researchers and conservation organizations can leverage the tool for:

- Tracking the health of protected natural areas
- Documenting the impacts of climate variations on vegetation systems
- Assessing revegetation success in restoration projects
- Providing objective evidence for conservation funding and reporting

For instance, a watershed protection organization could monitor the effectiveness of restoration projects by tracking increases in vegetation health and abundance over time.

Limitations and Future Enhancements

While the current system provides significant value, I acknowledge several limitations and opportunities for future enhancement:

Current Limitations

1. Temporal Coverage: The current implementation includes data only from the 2024 growing season. Historical context from previous years would enhance trend analysis capabilities.

2. Geographic Scope: The system is currently optimized specifically for Missoula County, Montana. Expansion to larger regions would require additional data processing.

3. Spatial Resolution: The 10-meter resolution of Sentinel-2 imagery limits the detection of fine-scale vegetation patterns, particularly in heterogeneous urban environments or small agricultural plots.

4. Vegetation Type Discrimination: The current system tracks general vegetation health but does not distinguish between different vegetation types (e.g., crops, forests, grasslands).

Future Enhancement Opportunities

Several promising enhancement paths could address these limitations and extend the system's capabilities:

1. Multi-Year Historical Analysis: Incorporating historical satellite data from previous years would enable long-term trend analysis and anomaly detection.
2. Additional Layers: Implementing additional map layers such would provide more targeted insights depending on use cases. Some additional bands to incorporate could include the following:

    - Normalized Difference Moisture Index: Excellent indicators for drought monitoring

- Soil Adjusted Vegetation Index: More accurate vegetation health measurement in areas with sparse vegetation
- Normalized Burn Ratio: Useful for post-fire recovery monitoring in forested areas

3. Cloud Hosted Storage/Server: The application currently operates locally on a personal machine, which limits accessibility and scalability. Implementing cloud-based infrastructure for both data storage and server management would enable secure, remote access to the monitoring tool from any location.
4. Machine Learning Integration: Developing machine learning models to classify land cover and crop types, predict vegetation trends, or detect anomalies would add powerful analytical capabilities.
5. Imagery Augmentation: Supplementing Sentinel-2 data with higher resolution and more frequent imagery from commercial satellites or drones for priority areas would enable more detailed analysis.
6. Weather Data Integration: Correlating vegetation patterns with precipitation, temperature, and other climate variables would provide context for interpreting vegetation changes.

These enhancements represent logical next steps for the evolution of this monitoring system based on user needs and technological opportunities.

Technical Implementation Details

The technical implementation of this project involved several specialized technologies and programming approaches:

Key Technologies

1. Python: The primary programming language used throughout the project, selected for its robust ecosystem of geospatial and data science libraries.
2. Rasterio, Rioxarray & Xarray: Core Python libraries used for reading, writing, and processing geospatial raster data with labeled dimensions, enabling analysis of multi-dimensional satellite data.
3. PySTAC Client: Python API client for searching and accessing Sentinel-2 satellite imagery through the SpatioTemporal Asset Catalog (STAC) protocol.
4. FastAPI & TiTiler: Python libraries used to establish the web framework and geospatial tile server that enable efficient delivery of large raster datasets over the web.
5. Shiny for Python: A web application framework that allows the creation of the interactive web application interface without requiring JavaScript expertise.
6. Matplotlib & Pandas: Python libraries visualization and data manipulation capabilities for statistical analysis of vegetation health metrics.

7. ipyleaflet: Interactive mapping library that powers the web-based satellite image viewer component.

Data Pipeline Workflow

The data pipeline follows a structured workflow for each satellite image acquisition:

1. Scene Selection: Query the STAC API to identify Sentinel-2 scenes covering Missoula County for a specific date, filtering out scenes with extreme cloud cover. If no scenes are found, the next day in the date range is queried.

2. Band Extraction: Extract the relevant spectral bands (Red, Green, Blue, Near-Infrared) and the SCL from the selected scenes.

3. Geometric Processing: Clip each band to the Missoula County boundary and reproject to a consistent coordinate system (EPSG:32611 for processing).

4. Cloud Masking: Apply masks to exclude pixels classified as clouds, cloud shadows, or snow in the SCL. Figure 4 below demonstrates this process in action. Pixels corresponding to clouds and shadows are marked and excluded in the creation of subsequent layers.

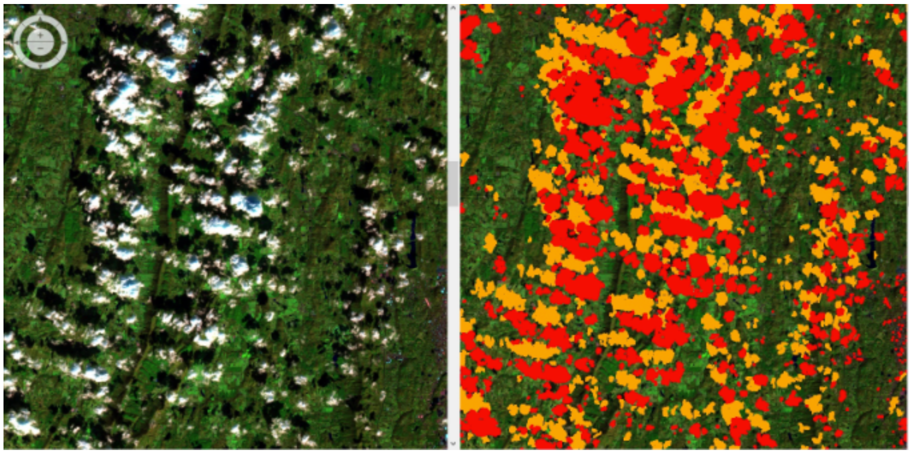


Figure 4: Full-color satellite image (left) and satellite image with cloud masking (right).

5. NDVI Calculation: Compute the Normalized Difference Vegetation Index layer using the formula (NIR - Red) / (NIR + Red) for all valid vegetation pixels.

6. RGB Composition: Create true-color composite layer by combining the Red, Green, and Blue bands (see Fig. 5) with appropriate scaling and color correction.

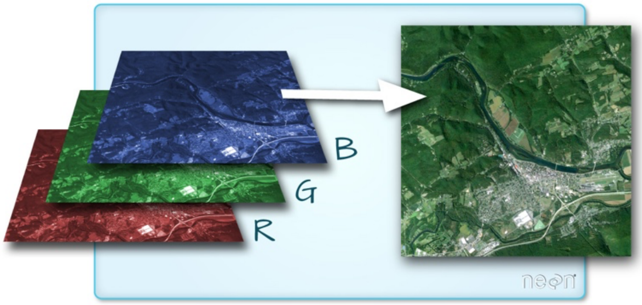


Figure 5: Visual demonstrating the formation of a true-color composite by combining the Red, Green and Blue reflectance bands.

7. Mosaic Creation: Combine multiple satellite scenes covering different parts of the county into seamless mosaics (see Fig. 6) for both NDVI and RGB products.

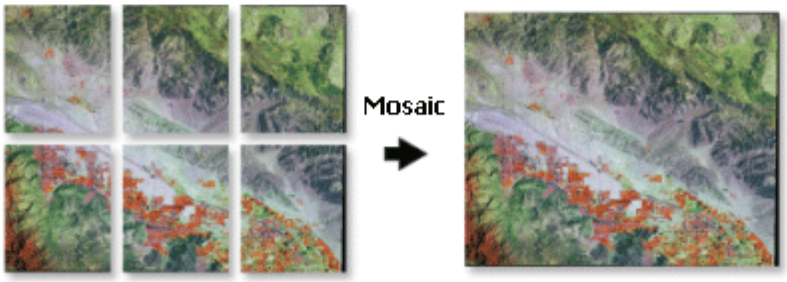


Figure 6: Creation of one seamless mosaic combining six individual scenes

8. Output Optimization: Convert final products to Cloud Optimized GeoTIFF format with appropriate compression, tiling, and overviews for efficient web delivery.

9. Statistical Analysis: Generate summary statistics and histograms of NDVI values for each processed date to support the analytics dashboard. Information for these visualizations are saved as Python pickle files and are used to quickly render visualizations in the dashboard.

This pipeline runs independently for each acquisition date. When implemented in a loop, users can adjust the data range parameter to build a temporal dataset that captures all imagery over the specified dates

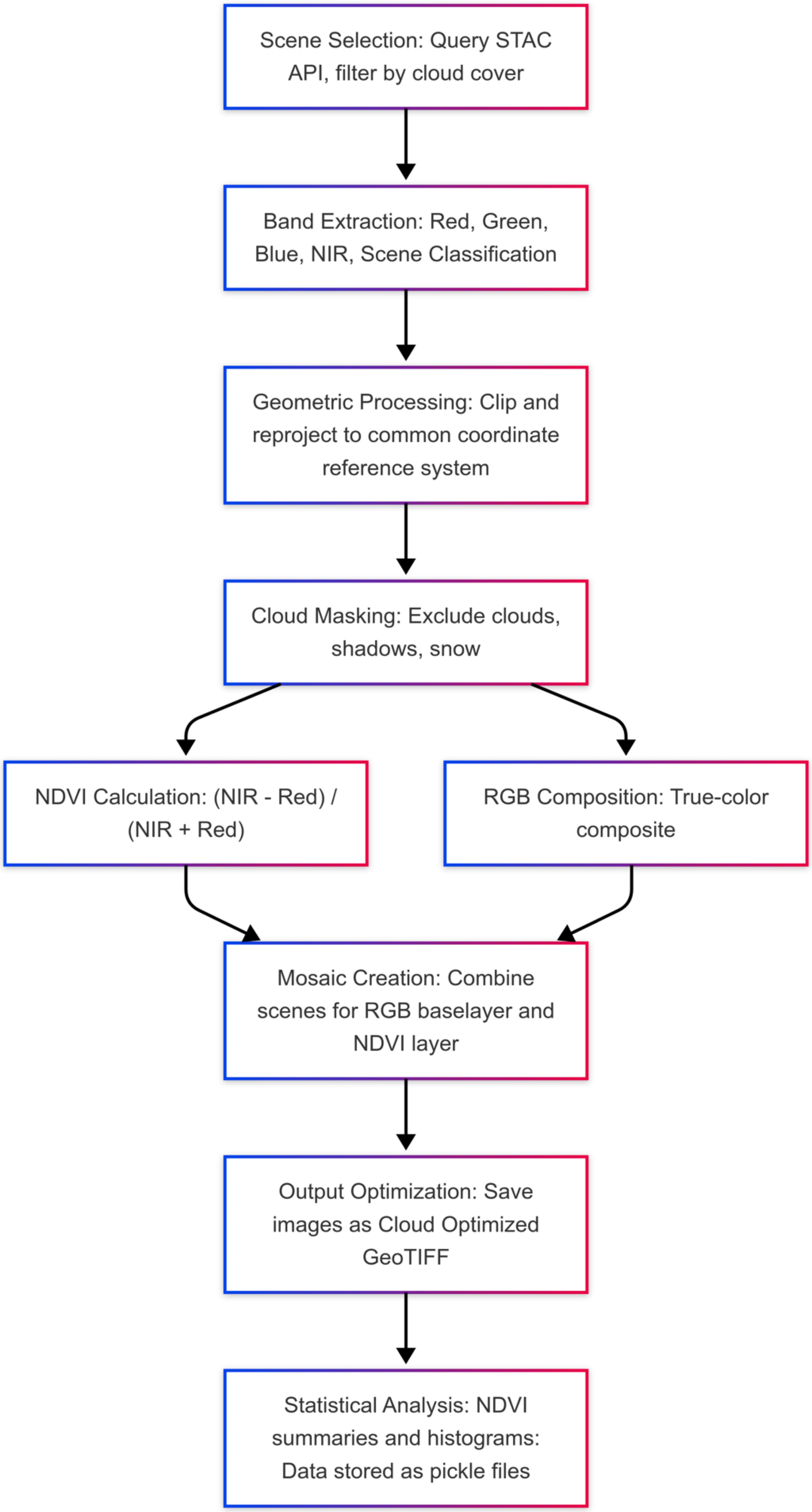

Scene Selection: Query STAC API, filter by cloud cover
Band Extraction: Red, Green, Blue, NIR, Scene Classification
Geometric Processing: Clip and reproject to common coordinate reference system
Cloud Masking: Exclude clouds, shadows, snow
NDVI Calculation: (NIR - Red) / (NIR + Red)
RGB Composition: True-color composite
Mosaic Creation: Combine scenes for RGB baselayer and NDVI layer
Output Optimization: Save images as Cloud Optimized GeoTIFF
Statistical Analysis: NDVI summaries and histograms: Data stored as pickle files

Figure 7: Data pipeline workflow flowchart.

Web Application Architecture

The web application follows a modern architecture that utilizes dynamic data serving practices to ensure smooth user interface functionality:

1. Backend Server: A FastAPI application with TiTiler integration that:
   - Serves map tiles from Cloud Optimized GeoTIFFs
   - Provides metadata about available dates and data quality
2. Frontend Application: A Shiny for Python application that:
   - Provides the user interface with interactive controls
   - Manages user selections
   - Communicates with the backend server to request appropriate data
   - Renders maps, charts, and statistical indicators
3. Reactive Framework: Uses Shiny's reactive programming model to automatically update visualizations when user selections change, creating a responsive experience without page reloads.

This separation of concerns enhances maintainability and allows each component to be optimized for its specific role in the system.

Conclusion

Project Outcomes

Through this capstone project, I successfully developed a comprehensive system that transforms complex satellite data into accessible vegetation health insights. The key outcomes include:

1. A fully automated data pipeline that processes satellite imagery into standardized vegetation health metrics
2. A responsive web server that efficiently delivers geospatial data for visualization
3. An intuitive web application that allows non-technical users to explore and analyze vegetation patterns
4. A demonstration of how modern data engineering practices can make remote sensing data accessible and valuable for practical decision-making

The Vegetation Health Monitoring Tool represents not just a technical achievement, but a practical solution to real-world challenges in environmental monitoring and land management.

Professional Growth

This project provided valuable professional growth opportunities that have prepared me for future data engineering challenges:

1. Deepening my expertise in geospatial data processing and analysis
2. Gaining experience in designing end-to-end data pipelines from acquisition to visualization
3. Developing skills in creating user-centered data products that bridge technical complexity and practical utility
4. Solving complex technical challenges through systematic problem decomposition and innovative approaches
5. Applying theoretical knowledge from my graduate program to a real-world implementation with practical constraints

These experiences have significantly enhanced my capabilities as a data professional and provided a foundation for continued growth in this field.

Final Reflections

This capstone project demonstrates how data engineering can transform raw data into valuable insights that support informed decision-making. By making sophisticated satellite analysis accessible to non-technical users, this system helps bridge the gap between advanced earth observation technologies and practical land management needs.

The Vegetation Health Monitoring Tool represents both the culmination of my graduate studies and a starting point for future innovations in environmental monitoring. I believe this work contributes meaningfully to the growing field of applied geospatial analytics and showcases the potential for data engineering to address important environmental challenges.

Appendix: Data Sources

Primary satellite data for this project was obtained from:

> **Sentinel-2 MSI Level-2A**: Atmospherically corrected surface reflectance products provided by the European Space Agency and accessed through the Element84 Earth Search STAC API (https://earth-search.aws.element84.com/v1)
>
> **Missoula County Boundary**: TIGER/Line shapefiles obtained from the U.S. Census Bureau (2022 release)