Scalable Analysis of Vegetation Anomalies Preceding Plant Disease Outbreaks

THRISHA AMBAREESHARAJE URS URS, University of California, Riverside, USA
SOHUM DAMANI, University of California, Riverside, USA
YASHASWINI DIGGAVI, University of California, Riverside, USA
VEDANT BORKUTE, University of California, Riverside, USA
SHREYANGSHU BERA, University of California, Riverside, USA

Early detection of plant diseases is an important challenge in both agricultural and forested ecosystems, where delayed identification can cause significant ecological and economic damage. This project investigates whether large-scale vegetation anomalies observable in historical remote sensing data precede documented plant disease outbreaks. By analyzing long-term vegetation trends prior to known disease detection events, the study aims to identify deviations in vegetation behavior that may indicate early stress. The project follows a data-driven and exploratory approach and does not assume the existence of strong predictive signals. Although initial experimentation may be conducted on smaller datasets due to resource constraints, the project is structured so that it can be extended to significantly larger datasets in the future.

ACM Reference Format:
Thrisha Ambareesharaje Urs Urs, Sohum Damani, Yashaswini Diggavi, Vedant Borkute, and Shreyangshu Bera. 2026. Scalable Analysis of Vegetation Anomalies Preceding Plant Disease Outbreaks. 1, 1 (January 2026), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Group Name: Lake it Easy+Pulse Crew
Student IDs: Thrisha Ambareesharaje Urs Urs (862638215), Sohum Damani (862621529), Yashaswini Diggavi (862620058), Vedant Borkute (862552981), Shreyangshu Bera (862485337)
Emails: turs001@ucr.edu, sdama009@ucr.edu, ydigg001@ucr.edu, vbork001@ucr.edu, sbera004@ucr.edu

1 Background

Plant diseases have historically caused significant ecological and economic damage. In agricultural regions, disease detection often relies on human observation and localized monitoring. In contrast, forested and remote regions lack continuous surveillance. Advances in satellite imagery and environmental data collection enable long-term observation of vegetation behavior across large geographic regions. Analyzing such data effectively requires approaches that go beyond traditional single-machine scripts and instead follow scalable data processing paradigms.

Authors' Contact Information: Thrisha Ambareesharaje Urs Urs, turs001@ucr.edu, University of California, Riverside, Riverside, California, USA; Sohum Damani, sdama009@ucr.edu, University of California, Riverside, Riverside, California, USA; Yashaswini Diggavi, ydigg001@ucr.edu, University of California, Riverside, Riverside, California, USA; Vedant Borkute, vbork001@ucr.edu, University of California, Riverside, Riverside, California, USA; Shreyangshu Bera, sbera004@ucr.edu, University of California, Riverside, Riverside, California, USA.

2 Motivation

The motivation for this project is mainly practical. Plant diseases can cause long-term ecological damage and economic loss if they are not identified in a timely manner. In agricultural settings, farmers and local communities can often observe visible signs of disease and take corrective action. However, in forested and remote regions, such direct monitoring is limited or nonexistent, allowing diseases to spread undetected for extended periods. Analyzing vegetation behavior over time using data-driven methods can provide valuable insight into early stress patterns that may otherwise go unnoticed.

In addition, this project offers a secondary technical motivation by enabling the application of big-data concepts to a real-world environmental problem. However, the emphasis remains on understanding and interpreting vegetation changes rather than guaranteeing predictive accuracy.

3 Project Description

The primary objective of this project is to analyze historical vegetation trends in regions where plant diseases have been reported. For selected disease events, vegetation behavior will be studied over multiple years before the documented detection date. The analysis seeks to identify deviations or anomalies in vegetation patterns that may precede disease outbreaks.

4 Datasets

These are some of the satellite optical multispectral datasets we identified as useful for this study, and we will begin our analysis with these sources.

Geographic scope: Will pertain to a specific disease outbreak, e.g. Western United States for the Bark Beetle Outbreak (see Reference 2).

Temporal range: Sentinel-2 (10 m resolution) from its operational start in 2015/2016 and Landsat Collection 2 Level-2 (30 m resolution) surface reflectance data from ~1982 to present (see Reference 3).

Approximate volume: Petabytes of imagery archived publicly (see Reference 5).

Access/processing options: Available programmatically via platforms such as Google Earth Engine, Copernicus Open Access Hub, USGS EarthExplorer, AWS Open Data, Microsoft Planetary Computer, or direct download (see Reference 3).

(1) **Sentinel-2 Level-2A and Landsat Collection 2 Level-2 surface reflectance imagery:** These datasets provide atmospherically corrected optical imagery with multispectral bands spanning the visible, near-infrared, red-edge, and shortwave infrared regions. Surface reflectance products allow the computation of vegetation indices such as NDVI (see Reference 1), which quantify vegetation health and support the detection of temporal changes associated with stress or disease. Sentinel-2's frequent revisit rate and high spatial resolution (10 m for key bands) make it particularly well suited to monitor forest dynamics and capture subtle variations in canopy condition.

(2) **ESA WorldCover global land cover map:** This dataset provides a global land cover classification at 10 m spatial resolution. We use WorldCover to identify and mask forested regions prior to vegetation index analysis. While NDVI measures greenness, it does not distinguish between land cover types; without land cover context, signals from grasslands, croplands, or other surfaces could be misinterpreted as forest health. Incorporating a land cover map ensures that analysis is restricted to true forest pixels and reduces contamination from non-forest land covers such as built-up areas, water bodies, or agricultural fields.

5 Main Outcome

The primary outcome of this project is retrospective analytics that examines historical vegetation behavior preceding known plant disease detection events. Rather than serving directly as an operational warning system, the framework evaluates whether anomalous vegetation patterns were present in the period leading to documented outbreaks. These anomalies are intended to guide analysis and interpretation rather than serve as definitive diagnoses, acknowledging the possibility of false positives. By allowing a closer examination of flagged regions, the project has the potential to support earlier intervention and reduce the spread of plant diseases, contributing to improved vegetation conservation. The framework is evaluated as an exploratory test mechanism, with the goal of assessing whether such signals could have been identified through historical data analysis and whether the approach is feasible for future implementation.

6 Relevance to Big Data

This project is relevant to big-data analytics due to the volume, variety, and temporal nature of the data involved. The analysis includes multi-year environmental datasets covering large geographic regions. Although initial experiments may use smaller subsets of data, the project is designed for scalability so that the same approach can be applied to larger datasets without fundamental changes. By attempting to identify vegetation stress months before it becomes observable, the system might turn raw data into a valuable tool for early disease prevention. This focus on extracting insights at scale directly aligns with the goals of big-data analytics.

7 Evaluation Plan

The project will be evaluated based on analytical rigor and adherence to big data design principles, with a primary focus on the system's computational performance. Evaluation criteria include measuring processing throughput and resource saturation, specifically peak memory usage and CPU utilization, as the dataset scales from a single year to a larger time series. Success is defined by the framework's ability to efficiently manage large scale multispectral data cubes on constrained local hardware while maintaining the quality of the spatiotemporal analysis.

8 Backup Project

If vegetation anomalies related to disease events are not clearly observable, the project will pivot to an alternative analysis examining long-term climate trends in a selected region. This backup project focuses on analyzing historical climate variables such as temperature and precipitation to identify long-term patterns and changes over time. This backup project follows the same scalable design philosophy and remains aligned with the course objectives.

9 Project Timeline

Table 1. Project Milestones and Proposed Timeline

Milestone	Timeline	Task Description
Data Acquisition	Week 1	Ingesting and preprocessing the dataset tiles for diseases affected forest regions, ensuring cloud-free pixel selection.
Baseline Creation	Week 2	Processing the historical Landsat time-series to calculate monthly NDVI (Normalized Difference Vegetation Index) means and standard deviations, establishing a "normal" forest health signature.
Anomaly Detection	Week 3	Developing the analytical engine to identify spatiotemporal patterns where current-year vegetation signals deviate significantly from the established historical baseline.
Evaluation	Week 4	Validating detected anomalies against historical disease records (e.g., USFS Bark Beetle maps) and finalizing the analytical framework and project report.

References

(1) https://www.earthdata.nasa.gov/topics/land-surface/normalized-difference-vegetation-index-ndvi
(2) https://tandf.figshare.com/articles/journal_contribution/Assessing_Combinations_of_Landsat_Sentinel-2_and_Sentinel-1_Time_series_for_Detecting_Bark_Beetle_Infestations/23566805?file=41345396
(3) https://deepwiki.com/microsoft/PlanetaryComputerExamples/4.1-satellite-imagery-datasets
(4) https://sanborn.com/blog/google-earth-engine-frequently-asked-questions/