Evaluating NDVI as a Proxy for LiDAR-Derived Understory Metrics: A Tile-Based Approach in Taal, Philippines

J.D. Casisirano[1,2*]

[1]Department of Geodetic Engineering, University of the Philippines, Diliman, Quezon City, Philippines (jarencecasisirano@gmail.com)
[2]Training Center for Applied Geodesy and Photogrammetry, University of the Philippines, Diliman, Quezon City, Philippines

Keywords: LiDAR, Understory structure, NDVI, Voxel metrics, Leaf area density, Tropical forest, Principal component analysis

Abstract

This study assesses the ability of NDVI to reflect forest structural complexity derived from airborne LiDAR data in a tropical forest context, specifically over a portion of Taal Volcano, Philippines. Using 100 m × 100 m tiles, five LiDAR metrics were computed—canopy cover ratio, voxel occupancy ratio, fractional cover, normalized cover, and leaf area density (LAD)—to represent both canopy and understory structure. NDVI was derived from Sentinel-2 TOA imagery and summarized per tile. Correlation analysis showed strong alignment between NDVI and canopy cover (r = 0.836), moderate to strong correlation with LAD and fractional cover, weak correlation with voxel occupancy (r = 0.224), and a negative relationship with normalized cover (r = –0.381). A multiple regression model combining all five metrics explained 82.2% of NDVI variation, with voxel and canopy metrics having the strongest predictive weight. Principal component analysis (PCA) revealed that LAD dominated the first principal component, which alone explained 95.3% of variance among LiDAR metrics; regression using the first two PCs still achieved an R^2 of 0.771. These results confirm that NDVI cannot directly measure understory structure, but its variation can be approximated by combining LiDAR-derived vertical metrics. The findings highlight the value of LiDAR for detecting vegetation complexity in multilayered tropical forests and underscore NDVI's limitations in resolving vertical structural information.

1. Introduction

1.1 Background of the study

Forests are vertically structured ecosystems composed of multiple vegetation layers, each contributing to biodiversity, carbon storage, and ecological resilience. Among these, the understory layer—comprising shrubs, saplings, regenerating trees, and herbaceous cover—plays a pivotal role in supporting wildlife habitats, modulating forest microclimates, and facilitating successional processes (cite: understory role in forest ecosystems). Despite its ecological importance, the understory remains among the most challenging forest layers to monitor remotely, due to its position beneath a typically dense and complex canopy.

Traditional satellite-based vegetation monitoring tools, particularly the Normalized Difference Vegetation Index (NDVI), have become ubiquitous in large-scale forest assessments. NDVI is widely used due to its simplicity, cost-efficiency, and proven ability to track canopy-level photosynthetic activity (cite: NDVI overview or use in forest monitoring). However, NDVI and similar 2D vegetation indices are inherently limited in capturing vertical structure, especially in multi-strata forest systems (cite: NDVI limitations in vertical structure or understory detection). In such environments, particularly tropical forests, NDVI may saturate at high biomass levels and fail to distinguish structurally complex layers such as the understory (cite: NDVI saturation and structural blindness).

In contrast, LiDAR (Light Detection and Ranging) offers the unique ability to capture the three-dimensional distribution of vegetation, including both canopy and understory layers. By emitting laser pulses and measuring their return time and intensity, airborne LiDAR systems can characterize forest height, density, and vertical complexity at fine spatial resolutions. Numerous studies have demonstrated LiDAR's capacity to quantify metrics such as canopy height models, vertical foliage profiles, and leaf area index (cite: general LiDAR applications in forestry). More advanced metrics, such as voxel occupancy, fractional cover, and leaf area density (LAD), have shown promising results in modeling understory vegetation, particularly when validated against field data (cite: LiDAR-derived understory metrics; possibly Venier et al. 2019 or similar).

A notable example is the work of Venier et al. (2019), which evaluated several LiDAR-derived metrics across boreal forests in Canada and demonstrated strong relationships between voxel-based LiDAR measures and field-observed understory cover. Their findings support the use of voxel-based and stratified LiDAR metrics in ecological applications, especially when field sampling is infeasible. However, studies like this are often limited to temperate or boreal forests, where canopy layering and vegetative density are relatively moderate compared to tropical regions.

Tropical forests, including those in the Philippines, exhibit exceptional vertical and horizontal heterogeneity, characterized by dense canopies, frequent layering, and spatially variable light environments (cite: tropical forest structure). In such contexts, the ability of NDVI to detect or infer structural variation is even more questionable. Although national initiatives such as the DREAM and Phil-LiDAR programs have produced extensive airborne LiDAR datasets for the Philippines, relatively few studies have explored the application of LiDAR for understory structure mapping in tropical ecosystems (cite: Phil-LiDAR coverage or gaps in tropical LiDAR use).

This study addresses that gap by adapting the methods of Venier et al. (2019) to a tropical forest setting, using high-resolution airborne LiDAR data collected over the Taal Volcano region in Batangas, Philippines. The research seeks to evaluate whether NDVI can serve as a meaningful proxy for structural vegetation attributes, particularly those derived from LiDAR that target the understory layer. Five structural metrics were computed: canopy cover ratio (based on DBH threshold), voxel occupancy (0.5–3.5 m), fractional cover, normalized cover, and leaf area density. These metrics represent both canopy-level and understory conditions. NDVI was extracted from Sentinel-2 Top-of-Atmosphere reflectance imagery for the same area and time period (2017), and tile-level NDVI means were computed.

By statistically comparing NDVI with each LiDAR-derived metric, and further using multiple linear regression and principal component analysis (PCA), the study investigates not just how well NDVI aligns with structural complexity, but also what types of structure NDVI is most sensitive to. The use of PCA, in particular, helps to reveal whether NDVI reflects a general structural axis (e.g., canopy density), or a more layered signature, such as that contributed by leaf area density in the understory.

Ultimately, this research contributes to the growing body of literature that examines the strengths and limitations of remote sensing tools in vertically complex tropical forest systems, and highlights the continued relevance of LiDAR for detailed structural analysis, especially where field validation is not feasible.

1.2 Significance of the study

Monitoring forest structure, particularly the understory layer, remains a critical challenge in tropical forest management, biodiversity conservation, and ecological research. While satellite-derived vegetation indices such as NDVI are widely used in remote sensing applications, their ability to reflect understory conditions is limited. This is especially problematic in tropical forests, where dense, multi-layered vegetation often conceals understory structure from traditional optical sensors. As a result, there is a need for more advanced and structure-sensitive methods that can complement or validate NDVI-based observations.

This study is significant for several reasons. First, it applies a tile-based LiDAR processing workflow to derive structural vegetation metrics—specifically voxel cover, leaf area density, fractional cover, normalized cover, and canopy cover—using airborne LiDAR data from the Taal Volcano region. By spatially summarizing these metrics at a 100m resolution, the study provides an interpretable and scalable method for analyzing vertical vegetation structure, including understory components, in a format compatible with satellite imagery.

Second, the study critically evaluates the relationship between NDVI and LiDAR-derived metrics, using both correlation and regression techniques. This includes a principal component analysis (PCA) that identifies the dominant structural axes captured by LiDAR and their explanatory power over NDVI. The finding that NDVI may be more closely aligned with leaf area density (LAD)—an understory-sensitive metric—than with canopy cover alone challenges common assumptions about NDVI's limitations and suggests it may carry more structural information in vertically complex forests than previously recognized.

Finally, the study contributes to the relatively limited body of literature that applies LiDAR for understory characterization in tropical forest ecosystems. While similar methodologies have been developed and tested in boreal and temperate zones, few have been implemented in Southeast Asia, despite the availability of high-quality LiDAR data from national programs such as DREAM and Phil-LiDAR. By adapting and extending these methods to the tropical Philippine context, this research underscores the continued value of airborne LiDAR for vegetation monitoring and demonstrates a pathway for remote understory assessment in regions where field validation is constrained by logistics, cost, or access.

1.3 Research Objectives

This study aims to assess the capacity of NDVI to reflect LiDAR-derived forest structural metrics in a tropical forest setting. Specifically, the objectives are:

1. To extract and compute LiDAR-derived structural metrics—including canopy cover, voxel occupancy, fractional cover, normalized cover, and leaf area density—using a tile-based processing approach.
2. To acquire Sentinel-2 NDVI imagery and compute tile-level NDVI statistics corresponding to the LiDAR coverage area.
3. To evaluate the relationship between NDVI and LiDAR-derived metrics, with emphasis on understory representation, using correlation analysis and regression modeling.
4. To apply principal component analysis (PCA) to identify dominant structural dimensions within the LiDAR data and examine their influence on NDVI.
5. To evaluate the predictive performance of the regression models using a 70/30 train-test split, and compare predicted NDVI values to the observed NDVI layer

2. Methodology

2.1 Study Area

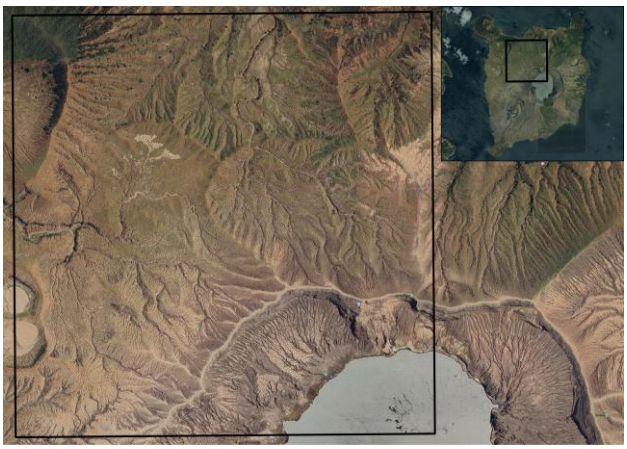

Figure 1. Study area map

The study was conducted on a forested portion of Taal Volcano Island, situated in Batangas, Philippines. Taal Volcano is located within a large freshwater caldera and is among the most active volcanoes in the country. The specific area analyzed lies

in the northeastern quadrant of the island (Figure 1), characterized by sloping terrain and dense vegetation cover. This area was selected based on the presence of continuous forest canopy and the availability of high-resolution airborne LiDAR data.

The LiDAR dataset was sourced from the Phil-LiDAR 1 Program, which captured topographic data across major Philippine watersheds between 2014 and 2017. Although the exact date of acquisition is not explicitly documented, the dataset used in this study is assumed to reflect conditions around 2017. The area was divided into 100 m × 100 m analysis tiles, and only those falling within dense forest cover were included in the structural and spectral analysis.

2.2 Methods

2.2.1 Overview

This study followed a tile-based approach to analyze the relationship between NDVI and LiDAR-derived structural metrics, with emphasis on characterizing forest understory structure. The methodology consisted of four major phases: (1) LiDAR data preprocessing, (2) computation of structural metrics, (3) NDVI acquisition and integration, and (4) statistical analysis and modeling. The complete methodology is summarized in Figure 2.

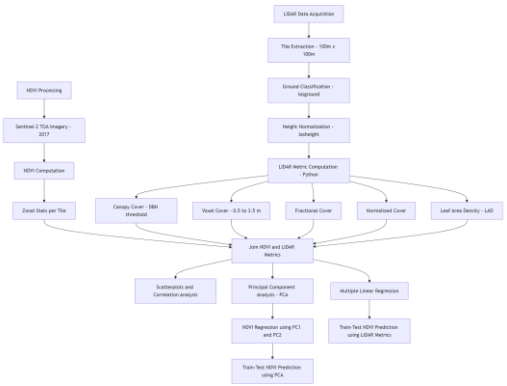

Figure 2. Flowchart of the research methodology

The diagram illustrates how airborne LiDAR data was preprocessed and used to compute five structural metrics, which were then spatially aligned with Sentinel-2 NDVI and used as predictors in statistical models (regression and PCA) to evaluate NDVI's relationship with forest structure.

2.2.2 Tile Framework and Spatial Resolution Context

To spatially align LiDAR-derived structural metrics with Sentinel-2 NDVI, the study area was divided into 100 m × 100 m grid tiles (Figure 3). This resolution was selected to balance three priorities: (1) compatibility with Sentinel-2's 10–20 m pixel scale, (2) adequate point density within each tile for robust LiDAR metric computation, and (3) computational efficiency for multi-metric modeling.

Each tile serves as a consistent spatial unit for aggregating both 3D LiDAR information and surface NDVI values. The gridded framework also allows for straightforward visual comparison and statistical integration. Only tiles within the northeastern quadrant of Taal Volcano Island that exhibited continuous forest cover and acceptable point density were included in the analysis.

This tiling approach forms the foundation for all subsequent correlation and regression modeling, ensuring that structural metrics and spectral data are directly comparable across space.

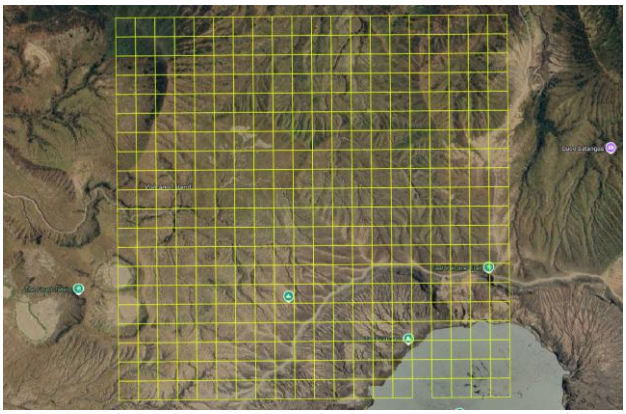

Figure 3. Gridded overlay of the study area showing the 100 m × 100 m tiles used for metric aggregation.

2.2.3 LiDAR Data Preprocessing

Airborne LiDAR data provided in .laz format were obtained from the Phil-LiDAR 1 program. The tiles were merged and then retiled into 100 m × 100 m grid units to serve as the basis for analysis. Preprocessing was performed using LAStools and included the following steps:

- Ground classification was done using the lasground tool with the -wilderness setting to detect subtle elevation changes in vegetated terrain.
- Height normalization was performed using lasheight, which calculated the height of all points above the classified ground surface.
- The resulting normalized point cloud tiles were saved as .las files for further processing in Python.

Only tiles with complete forest cover and adequate point density were retained for analysis.

2.2.4 Computation of LiDAR-Derived Metrics

Five structural metrics were computed from the height-normalized point clouds using custom Python scripts. These metrics were chosen to reflect both canopy and understory vegetation structure:

- Canopy Cover Ratio (DBH threshold): Computed as the proportion of all first returns that fall above 1.37 meters in height (approximating breast height), relative to the total number of first returns in the tile.
- Voxel Cover: Computed as the number of 1 m^3 voxels within the 0.5 to 3.5 meter vertical range that contain at least one LiDAR return, divided by the total number of voxels in that range, representing understory presence.
- Fractional Cover: Computed as the number of LiDAR returns within the 0.5 to 3.5 meter range (understory) divided by the sum of returns in that range plus all ground-classified returns.
- Normalized Cover: Computed as the number of returns in the 0.5 to 3.5 meter understory range divided by the total number of first returns in the tile, measuring signal penetration below the canopy.

- Leaf Area Density (LAD): Computed by taking the negative natural logarithm of the ratio of returns within a vertical height bin to the cumulative number of returns in and below that bin, divided by an extinction coefficient constant. This metric estimates foliage density per height bin.

Table 1 summarizes the structural focus and rationale behind each selected LiDAR metric:

Table 1. Structural Focus and Rationale Behind Selected LiDAR-Derived Metrics

Metric	Structural Focus	Rationale for Inclusion
Canopy Cover	Overstory (top layer)	Captures dense canopy reflectance directly tied to NDVI.
Voxel Occupancy	Understory (0.5-3.5 m)	Detects presence/absence of vegetation in 3D-critical for mapping structure where NDVI fails.
Leaf Area Density	Understory (0.5-3.5 m)	Measures density, not just presence, of foliage-used to refine voxel insights.
Fractional Cover	Understory vs ground	Highlights dominance of vegetation in lower strata-potentially correlated with NDVI in canopy gaps.
Normalized Cover	All heights (first returns)	Shows signal penetration-acts as an inverse proxy for canopy density.

Each metric was computed at the tile level and stored as a separate CSV file before being merged into a unified dataset.

2.2.5 NDVI Acquisition and Zonal Statistics

While the LiDAR metrics were computed to capture 3D vegetation structure across vertical layers, NDVI served as the dependent variable in all subsequent statistical analyses. Its computation was aligned to the LiDAR tile framework to allow direct spatial modelling.

NDVI was derived from Sentinel-2 Level-1C Top-of-Atmosphere (TOA) reflectance imagery for the period March to May 2017, corresponding to the presumed LiDAR acquisition window. NDVI was computed using the normalized difference of the near-infrared (B8) and red (B4) bands. A cloud-filtered composite was created using Google Earth Engine (GEE).

The resulting NDVI raster was exported and clipped to the LiDAR study area. Using the tiles_100m.shp grid, zonal statistics were calculated in QGIS to obtain the mean NDVI value per tile. These values were exported and joined with the LiDAR metrics, resulting in the final dataset.

2.2.6 Statistical Analysis and Modeling

This phase aimed to evaluate the extent to which NDVI variation can be explained by LiDAR-derived structural metrics—particularly those associated with understory complexity. NDVI served as the dependent variable, while all LiDAR metrics were treated as predictors.

Correlation and Scatterplots

Pairwise Pearson correlations were computed between NDVI and each LiDAR-derived metric to assess individual linear relationships. Scatterplots were generated to visually interpret the direction and strength of these associations.

Multiple Linear Regression

A multiple linear regression model was constructed with NDVI as the response variable and the five LiDAR metrics as predictors. This model was used to test whether NDVI can be approximated from a combination of structural variables, and to identify which metrics contribute unique explanatory power beyond their individual correlations.

Principal Component Analysis (PCA)

To examine structural co-variation and reduce dimensionality, PCA was applied to the five LiDAR metrics. PCA was used not only for data simplification, but also to determine whether NDVI aligns with a dominant latent axis of vegetation structure—such as overall foliage density in the understory—rather than being driven by isolated features. A second regression model using the first two principal components was also evaluated to compare predictive performance with the original metric-based model.

Train-Test Regression Validation

To evaluate how well the models generalize, a 70/30 train-test split was applied. The subsets were determined using random sampling of the 100 m x 100 m tiles. The regression models were trained on 70% of the data and tested on the remaining 30% to predict NDVI. This approach was used for both the full metric model and the PCA-based model. The results were evaluated using R^2 and RMSE and compared to the original regression results trained on the full dataset.

3. Results and Discussion

3.1 NDVI and LiDAR Comparison

Before presenting quantitative results, it is useful to illustrate the fundamental difference in how NDVI and LiDAR represent forest structure. Figure 4 shows a side-by-side comparison of the NDVI composite and a rendered LiDAR point cloud for the same area in Taal Volcano Island. The NDVI image, derived from Sentinel-2 reflectance, displays surface-level greenness as a smooth gradient, emphasizing areas of high canopy photosynthetic activity. In contrast, the LiDAR rendering visualizes the 3D structure of the vegetation, revealing crown textures, height variations, and layering effects not captured in 2D indices.

This visual comparison underscores a core premise of this study: while NDVI is sensitive to top-of-canopy greenness, it lacks the vertical resolution needed to detect structural variation beneath the canopy. LiDAR, on the other hand, provides the three-dimensional information necessary to assess understory presence, density, and vertical complexity—metrics that are especially critical in multilayered tropical forests.

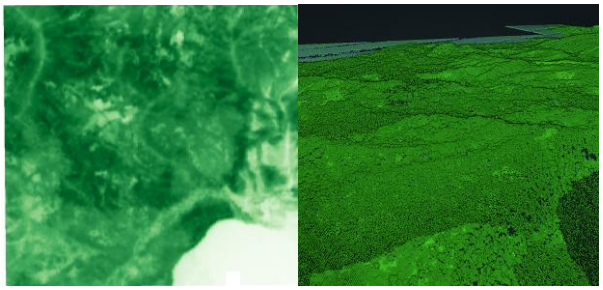

Figure 4. Side-by-side comparison of the NDVI composite (left) and the LiDAR point cloud rendered in 3D (right). LiDAR image taken from *Potree Viewer*.

3.2 NDVI and LiDAR Metrics Correlations

3.2.1 NDVI vs. Canopy Cover

A strong positive relationship was found between NDVI and the LiDAR-derived canopy cover ratio, with a Pearson correlation coefficient of r = 0.836 (Figure 5). The scatterplot exhibits a clear upward trend, with NDVI increasing as canopy cover becomes denser. At lower canopy cover values (below 0.2), NDVI is consistently low, while mid to high canopy cover values show a broader spread but still maintain a positive alignment. This trend confirms that NDVI is primarily sensitive to top-of-canopy greenness, as it responds to the cumulative reflectance of photosynthetically active vegetation, especially in the near-infrared spectrum (Hart et al., 2022).

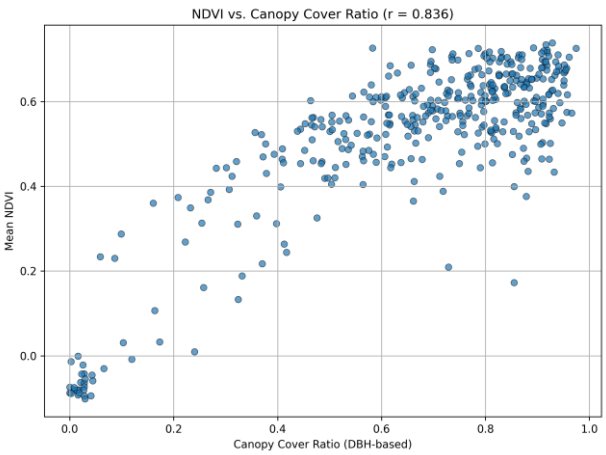

Figure 5. NDVI vs Canopy Cover scatter plot

This result shows that NDVI is a reliable indicator of canopy closure and overstory density, particularly in dense or mature forest stands. However, the observed widening in NDVI range at higher canopy cover levels may reflect the influence of additional structural factors, such as vertical layering, understory complexity, or saturation effects common in high-biomass environments. While NDVI clearly captures the overstory signal, these nuances underscore the importance of integrating 3D metrics—such as those derived from LiDAR—to detect understory variation not evident in 2D spectral indices alone.

3.2.2 NDVI vs. Voxel Cover Ratio

The relationship between NDVI and voxel occupancy ratio—a LiDAR-derived metric representing vegetation presence within the 0.5–3.5 m understory layer—was weak (r = 0.224) (Figure 6). While NDVI values tended to be low in areas with minimal voxel occupancy, there was considerable spread in NDVI across moderate to high voxel values, indicating that NDVI does not reliably track understory structure.

At low voxel occupancy values (i.e., <0.10), NDVI values were also low, often clustering below 0.2. However, once voxel values approached the 0.10–0.15 range, a distinct shift occurred: NDVI values rapidly increased, jumping above 0.4 and often clustering between 0.5 and 0.7. Beyond this threshold, voxel values continued to increase, but NDVI did not follow the same upward trend—instead, it plateaued or became more scattered. This stepwise response suggests that NDVI may only begin to respond to understory presence when structural density crosses a certain visibility or biomass threshold, after which additional structural variation in the understory is masked or saturated by the overstory reflectance.

This pattern confirms the limitations of NDVI in capturing fine-scale vertical structure, especially in tropical forests where overstory layers are dense and persistent year-round. NDVI, being a surface reflectance index, is inherently biased toward top-of-canopy greenness, and even where the understory is structurally significant, it remains optically obscured. The voxel metric, in contrast, is derived from LiDAR returns in stratified 3D space, offering a more direct measurement of vegetation presence within the understory range. Thus, the poor correlation between NDVI and voxel cover illustrates that NDVI is not a reliable indicator of understory complexity in vertically layered environments.

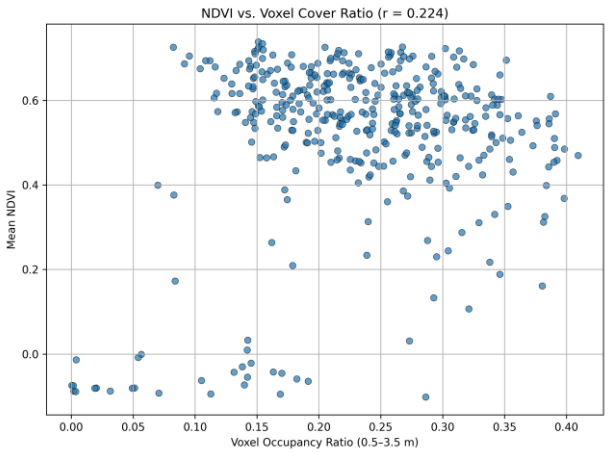

Figure 6. NDVI vs. Voxel Cover Ratio scatter plot

Importantly, this finding is in line with Venier et al. (2019), who demonstrated that voxel cover was among the strongest predictors of field-measured understory vegetation in boreal forests. By contrast, NDVI was not assessed as a predictor in their study, reinforcing the idea that NDVI and voxel cover represent different structural domains. This result strengthens their conclusion by showing that even in tropical forests—where vertical stratification is often more pronounced—NDVI does not align with voxel-derived understory structure. This reinforces the necessity of 3D structural data, such as LiDAR, for meaningful understory mapping, especially in ecosystems where spectral indices are constrained by saturation or canopy dominance.

3.2.3 NDVI vs. Leaf Area Density (LAD)

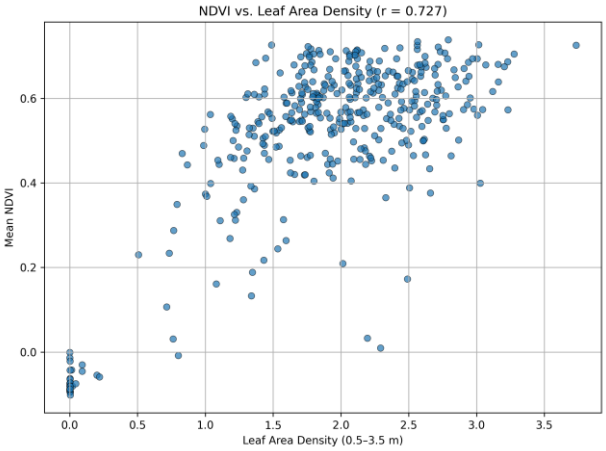

Figure 7. NDVI vs. Leaf Area Density (LAD) scatter plot

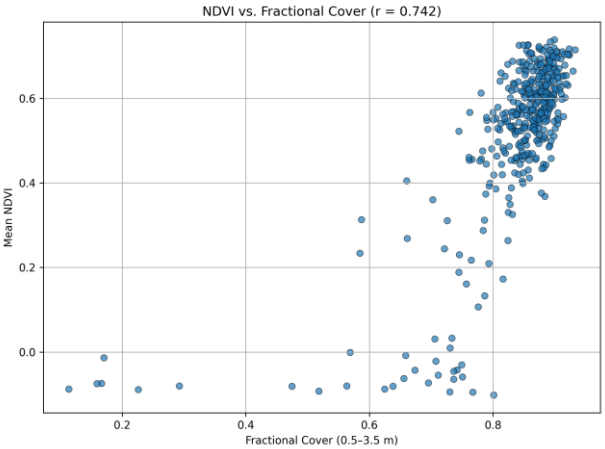

Figure 8. NDVI vs. Fractional Cover scatter plot

A strong positive relationship was also observed between NDVI and leaf area density (LAD) within the 0.5–3.5 m vertical stratum, with a correlation coefficient of r = 0.727 (Figure 7). LAD, unlike voxel occupancy which measures presence/absence, quantifies the density of vegetation returns in each height bin—offering a continuous estimate of foliage concentration. The scatterplot shows that as LAD increases, NDVI values also rise, particularly beyond LAD values of 1.0, where NDVI consistently exceeds 0.5. At very low LAD (<0.5), NDVI remains near zero, indicating the absence of both foliage and reflectance signal.

This result suggests that NDVI is somewhat sensitive to total vegetation density, even when it occurs in the understory range. However, this may be attributed to indirect effects, such as increased vertical layering, overall biomass, or canopy porosity that allows reflectance from lower strata to contribute to the surface signal. While NDVI is not directly targeting LAD, the correlation highlights the structural information embedded in vegetation indices under certain conditions. Nonetheless, it remains difficult to determine whether NDVI reflects understory LAD directly or as a byproduct of general stand density—again highlighting the value of LiDAR in distinguishing vegetation layers explicitly.

3.2.4 NDVI vs. Fractional Cover

The fractional cover metric, defined as the proportion of LiDAR returns within the 0.5–3.5 m understory stratum relative to the total number of understory and ground-classified returns, showed a strong correlation with NDVI (r = 0.742) (Figure 8). This result suggests that NDVI tends to increase in areas where a greater fraction of LiDAR returns originate from vegetated understory layers, as opposed to bare ground. This aligns with expectations that NDVI responds to increases in vegetative complexity, even in understory layers, when such complexity is structurally dominant.

However, the scatterplot revealed that fractional cover values were tightly clustered above 0.75 for most tiles, with only a small number of samples representing lower fractional cover. This limited variability may be due, in part, to the 100 m × 100 m tile resolution used in the analysis, which tends to average out small-scale spatial differences in vegetation structure. At this scale, most tiles likely encompass dense vegetation throughout, making it difficult to detect meaningful understory variation using fractional cover alone. While the statistical correlation is strong, the metric may therefore be less structurally informative than voxel occupancy or LAD in this context.

3.2.5 NDVI vs. Normalized Cover

The normalized cover metric—defined as the proportion of first returns originating from the 0.5–3.5 m understory range—was negatively correlated with NDVI (r = -0.381) (Figure 9). This indicates that tiles with a greater share of understory returns relative to all first returns tend to exhibit lower NDVI values. This result aligns with expectations, as an increase in normalized cover often reflects reduced canopy presence. Since NDVI primarily captures top-of-canopy greenness, it declines when first returns are dominated by understory signals, thereby reinforcing its limitations in detecting vertical vegetation structure in dense forests. Normalized cover, though structurally indirect, appears to function as a proxy for NDVI suppression in understory-dominant environments

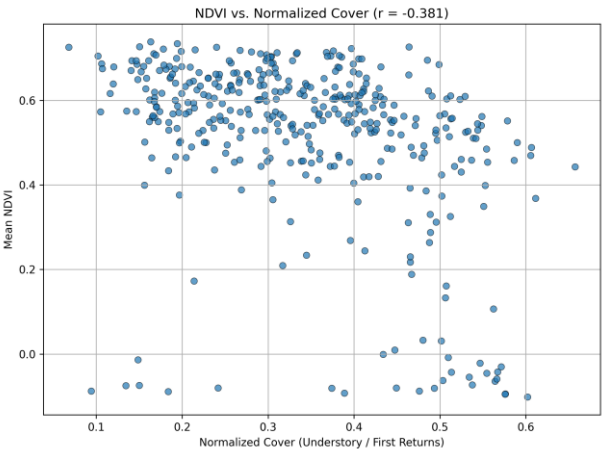

Figure 9. NDVI vs. Normalized Cover

3.3 Multiple Linear Regression Modeling

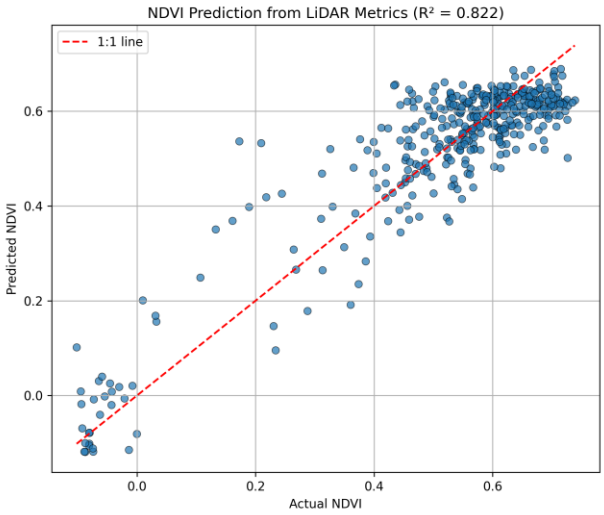

Figure 10. NDVI Regression from LiDAR Metrics

A multiple linear regression model was developed using all five LiDAR-derived structural metrics to predict mean NDVI across the study area. The model achieved an R^2 of 0.822 and an RMSE of 0.0819, indicating strong predictive performance and low residual error (Figure 10). The scatterplot shows that predicted NDVI values closely follow the 1:1 line across the full NDVI range, particularly in tiles with moderate to high vegetation greenness.

Interestingly, voxel occupancy ratio had the highest regression coefficient (0.7919), despite its relatively weak individual correlation with NDVI (r = 0.224). This suggests that voxel cover contributes unique structural information not captured by other variables. While NDVI alone may not align with voxel occupancy, when combined with canopy and density metrics, voxel cover helps explain areas where vertical complexity contributes indirectly to the surface reflectance signal. Its role may reflect structural heterogeneity or vertical layering, which affects NDVI in non-obvious ways. Canopy cover ratio had the second-highest coefficient (0.7365) and also showed the strongest direct correlation with NDVI (r = 0.836), making it the most intuitively interpretable predictor. As expected, NDVI strongly reflects top-of-canopy greenness, and canopy cover directly enhances NDVI through increased near-infrared reflectance. The model confirms canopy cover's dominant influence on optical vegetation indices in dense tropical forests.

Normalized cover had a moderate positive coefficient (0.2526), despite having a negative individual correlation with NDVI (r = –0.381). This reversal highlights the effect of multivariate control—once canopy cover and voxel occupancy are included, normalized cover adds residual explanatory power in specific conditions, possibly reflecting canopy gaps or understory exposure in a way that correlates with NDVI increase, not decrease. Fractional cover exhibited a moderate negative coefficient (–0.2244), which is somewhat surprising given its strong positive correlation with NDVI (r = 0.742). This inconsistency suggests that fractional cover may overlap structurally with other predictors, especially voxel occupancy and LAD. In the presence of more precise structural variables, its contribution appears partially redundant or inversely weighted. This is a classic case where collinearity among predictors reshapes their apparent influence in a multivariate context. Finally, leaf area density (LAD)—despite having a high individual correlation with NDVI (r = 0.727)—had the lowest regression coefficient (0.0355). This implies that the variation LAD explains in NDVI is largely shared with voxel and canopy cover. Once those stronger variables are included, LAD contributes very little additional information to the model. This doesn't invalidate its value, but rather suggests that its effect is not uniquely distinguishable in the presence of other strong predictors.

Taken together, these results highlight the strength of combining LiDAR-derived metrics to model NDVI, while also illustrating the complexities of multicollinearity and structural redundancy in ecological datasets. The model suggests that while NDVI may not directly reflect certain structural variables, such as voxel cover or LAD, these can still help predict NDVI when used in combination—underscoring the complementary role of 3D LiDAR structure in enhancing surface-level vegetation interpretation.

3.4 NDVI Prediction Using PCA

Principal component analysis (PCA) was used to reduce redundancy among the five LiDAR-derived structural metrics and to identify dominant structural gradients that may explain NDVI variation. The first principal component (PC1) alone explained 95.3% of the total variance across the input variables, while the second component (PC2) contributed an additional 2.9%. A regression model using only PC1 and PC2 achieved an R^2 of 0.771, only slightly lower than the full multiple regression model using all five original variables (R^2 = 0.822), indicating that nearly all of the NDVI variation could be captured by these two latent dimensions (Figure 11).

The PCA loadings revealed that PC1 was overwhelmingly dominated by leaf area density (LAD), which had a loading of 0.94, while other metrics contributed minimally—including canopy cover ratio (0.31), fractional cover (0.08), normalized cover (–0.13), and voxel occupancy ratio (–0.03). This indicates that the dominant structural axis in the LiDAR dataset was most strongly aligned with vertical foliage density in the 0.5–3.5 m understory layer. While other metrics capture complementary aspects of forest structure, they did not co-vary as strongly as LAD across the sampled area, making LAD the defining feature of PC1.

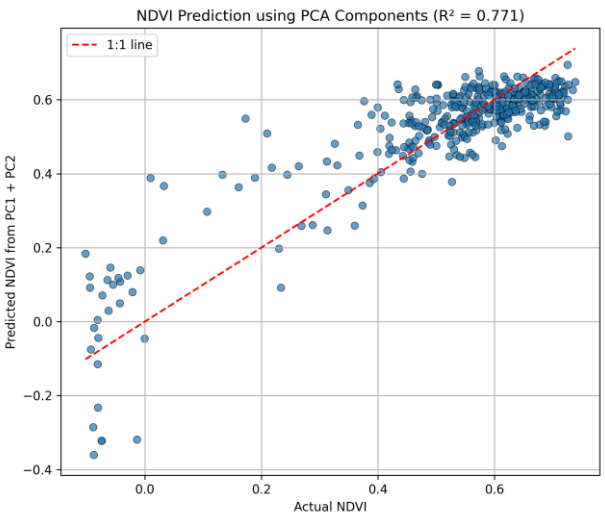

Figure 11. NDVI Prediction using PCA

Although LAD dominated PC1, it had the smallest coefficient in the multiple regression model. This reflects the different goals of PCA and regression: PCA identifies internal variation among the predictors, while regression focuses on how each metric contributes uniquely to explaining NDVI. Both voxel occupancy and LAD describe the same vertical stratum but measure it differently—voxel cover detects presence of structure, while LAD measures its density. Once voxel and canopy effects are accounted for in the regression, LAD's contribution becomes statistically redundant. This doesn't mean it lacks value—only that it doesn't explain additional NDVI variation that isn't already captured by other metrics.

These results demonstrate that NDVI variation in this study can be largely explained by a single dominant vegetation structure gradient, most closely tied to understory foliage density. While NDVI is not explicitly designed to measure vertical structure, its response appears to align with areas of high cumulative vegetative complexity. This further reinforces the value of combining LiDAR-based metrics for modeling NDVI patterns and shows how dimensionality reduction techniques like PCA can help isolate the most informative structural axes from a suite of 3D vegetation indicators.

3.5 Model Validation Using Train-Test Prediction

To assess the generalizability of the regression models beyond the dataset used for training, a 70/30 train-test split was implemented. This approach simulates a realistic scenario in which NDVI must be predicted in tiles where only LiDAR-derived structural data are available. Two predictive models were tested: one using the original five LiDAR metrics, and another using the first two components from Principal Component Analysis (PCA). Figures 12 and 13 illustrate the results of these models on the held-out test tiles.

3.5.1 NDVI Prediction from LiDAR Metrics

The first model employed multiple linear regression using five LiDAR-derived predictors: canopy cover, voxel occupancy, fractional cover, normalized cover, and leaf area density (LAD). These metrics were selected for their ability to capture both canopy and understory structure. The model was trained using 70% of the available tiles and tested on the remaining 30%.

As shown in Figure 12, the predicted NDVI values closely align with the observed NDVI values in the test set. The resulting R^2 value was 0.756, with an RMSE of 0.0889. The scatterplot shows a strong clustering of points along the 1:1 reference line (red dashed), indicating high predictive accuracy. Some dispersion exists at lower NDVI values, where structure is more heterogeneous and NDVI may be more influenced by external factors like soil reflectance or cloud shadowing. Nonetheless, the model demonstrates robust predictive capacity based on LiDAR structure alone.

The result confirms that NDVI variation across the landscape is strongly associated with LiDAR-observed canopy and midstory properties. Importantly, this performance was achieved without using the test set for training, lending confidence to the model's ability to generalize beyond the training tiles.

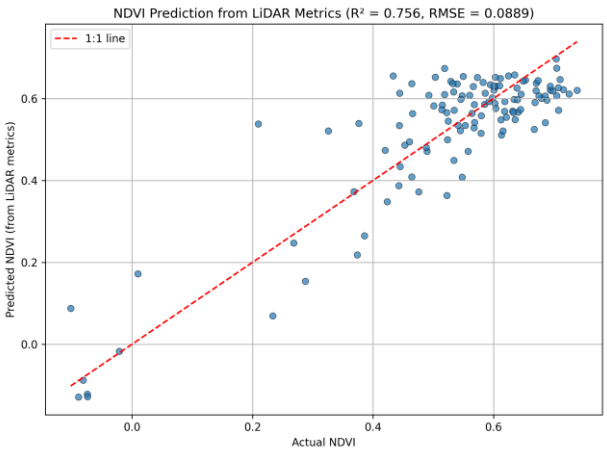

Figure 12. Predicted vs. actual NDVI values on the 30% test tiles using a regression model trained on all five LiDAR metrics

3.5.2 NDVI Prediction from PCA Components

The second model used PCA as a dimensionality reduction technique. The five original LiDAR metrics were transformed into orthogonal components, with the first two principal components (PC1 and PC2) used as predictors. PC1, which explained over 95% of the structural variance, was heavily dominated by LAD, followed by voxel occupancy and fractional cover.

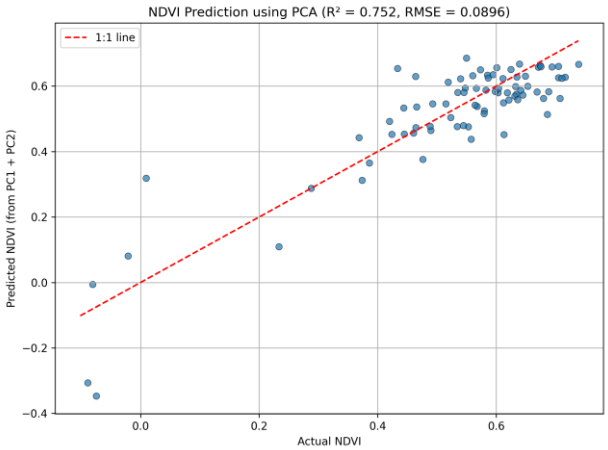

Figure 13. Predicted vs. actual NDVI values on the 30% test tiles using a regression model trained on PC1 and PC2 from PCA

Figure 13 presents the predicted vs. actual NDVI values from this PCA-based model. The model achieved an R^2 of 0.752 and RMSE of 0.0896, nearly identical to the performance of the full metric model. This shows that most of the structure needed to explain NDVI can be condensed into a latent gradient dominated by vertical foliage density.

As the scatterplot in Figure 13 shows, predicted values are tightly clustered around the 1:1 line, again indicating strong predictive performance. While a few outliers exist, especially at the lower end of the NDVI spectrum, the overall distribution closely mirrors that of the metric-based model. This reinforces the idea that NDVI, although limited in vertical sensitivity, still reflects integrated structure when modeled from 3D inputs.

4. Conclusion

4.1 Conclusion

This study explored the relationship between NDVI and a suite of LiDAR-derived structural metrics in a tropical forest setting using tile-based analysis over Taal Volcano, Philippines. Building upon the framework of Venier et al. (2019), the research sought to evaluate whether NDVI, a widely used optical vegetation index, can capture understory structural complexity when compared to high-resolution airborne LiDAR data. Specifically, five LiDAR metrics—canopy cover ratio, voxel occupancy ratio, fractional cover, normalized cover, and leaf area density (LAD)—were extracted and analyzed across 100 m × 100 m tiles. These metrics were then individually and collectively assessed for their ability to explain spatial variation in Sentinel-2 NDVI.

Among the five metrics, canopy cover ratio showed the strongest individual correlation with NDVI ($r = 0.836$), which is consistent with NDVI's design as a canopy-sensitive index. Leaf area density and fractional cover also showed strong correlations ($r = 0.727$ and $r = 0.742$, respectively), while voxel occupancy showed a much weaker individual correlation ($r = 0.224$), and normalized cover was negatively correlated with NDVI ($r = –0.381$). These results confirm that NDVI is most sensitive to canopy-level greenness, and its ability to reflect understory structure is indirect at best.

A multiple linear regression model combining all five LiDAR metrics achieved a strong fit ($R^2 = 0.822$), demonstrating that NDVI variation can be effectively predicted from structural indicators when used in combination. Interestingly, voxel occupancy and canopy cover had the highest regression coefficients, even though voxel cover had only a weak individual correlation with NDVI. This suggests that voxel occupancy offers unique structural information—particularly about vertical complexity—that is not captured by other metrics but still influences NDVI when considered in a multivariate context. The inverse and unexpected coefficients of fractional cover and the minimal contribution of LAD also highlight the effects of multicollinearity and underscore the need for careful interpretation when using overlapping metrics.

Principal component analysis (PCA) was employed to reduce dimensionality and examine underlying structural gradients. The first principal component (PC1) explained 95.3% of total variance and was overwhelmingly driven by LAD, confirming that vertical foliage density in the 0.5–3.5 m range is the dominant structural signal in the LiDAR dataset. Regression using only PC1 and PC2 still explained 77.1% of NDVI variation, indicating that most of the LiDAR–NDVI relationship could be captured using a reduced structural representation. While LAD played a minor role in the multivariate regression, it dominated the internal variance structure of the predictor set, reflecting the nuanced interplay between statistical modeling frameworks.

To assess how well NDVI could be predicted in new, unsampled forest tiles, two additional regression models were developed using a 70/30 train-test split. One model used the full set of five LiDAR metrics, while the other used only the first two principal components from PCA. The metric-based model achieved an R^2 of 0.756 and RMSE of 0.0889 on the test set, while the PCA-based model performed similarly with an R^2 of 0.752 and RMSE of 0.0896. These results confirm that NDVI variation can be reliably predicted using LiDAR-derived structural information alone. The comparable performance of the reduced PCA model suggests that NDVI is most responsive to a dominant structural gradient, largely defined by voxel occupancy and LAD, rather than to specific input metrics individually.

In conclusion, while NDVI cannot directly detect forest understory structure, it can be partially explained—and even predicted—by a combination of LiDAR-derived metrics that capture both canopy and midstory complexity. NDVI is most sensitive to surface-level greenness but responds, indirectly, to deeper vegetative layers when modeled appropriately. These findings affirm the value of LiDAR in characterizing vertical forest structure and demonstrate the limitations of relying solely on NDVI in complex tropical systems. They also highlight the potential of LiDAR-based models to estimate NDVI in spatially or temporally data-poor regions, providing new opportunities for forest monitoring where optical observations are limited.

4.2 Recommendations

Future research building on this study should consider several refinements and extensions. First, the use of smaller spatial units (e.g., 10–30 m tiles) may better capture localized variation in understory structure, which is likely averaged out at the 100 m tile scale. This would also allow for more detailed spatial modeling, particularly in fragmented or transitional forest areas. Second, incorporating field-based validation data—such as ground-measured understory density or species composition—would provide a stronger basis for interpreting LiDAR-derived metrics and improving ecological inferences. This is especially relevant when distinguishing between structural density and spectral greenness as sources of NDVI variation.

Third, future studies may explore comparisons across different forest types, such as secondary growth, disturbed zones, or plantations, to test whether the same relationships hold in structurally simpler or more dynamic environments. Additionally, integration with other remote sensing platforms—such as UAV-based photogrammetry—could help expand the utility of this approach in regions where airborne LiDAR is limited or unavailable.

Finally, given that voxel occupancy and LAD were particularly useful for capturing understory structure, future work could investigate how these metrics behave in relation to biodiversity indicators, carbon estimates, or habitat quality, using NDVI as a supplementary, but not stand-alone, measure of vegetation structure.

Code and Data Availability

All scripts, metric outputs, and reproducible workflows related to this study are openly available in the GitHub repository:

https://github.com/jarencecasisirano/taal-lidar-understory.git

This includes Python scripts for LiDAR metric computation, statistical analysis, and the final processed CSV files and plots. The data used in this study is not included in the project repository due to data size.

References

Hart, J., Yeager, R., Riggs, D., Fleischer, D., Owolabi, U., Walker, K., Bhatnagar, A., & Keith, R. (2022). The Relationship between Perceptions and Objective Measures of Greenness. *International Journal of Environmental Research and Public Health*, 19. https://doi.org/10.3390/ijerph192316317.