S&DS 425/625 Report

Elder Veliz & Emmanuelle Brindamour

2024-12-16

Abstract

An overview of your report, including one or so sentences on each of these:

- a non-technical description of the problem you are trying to solve or the question you are trying to answer, and why you are trying to answer that question
- a non-technical description of the data, where it came from, and what it contains, including possibly the predictors, the outcome, and the observations
- a non-technical description of what kind of analysis you did, including high-level description of what the predictors were, what the outcome was, and how to interpret the results of the model
- a brief summary of the models that are used
- a non-technical description of the results of the model and main takeaways.

An abstract is one paragraph with text only and is aimed at a technical audience. This appears at the beginning of the report.

Executive Summary

An executive summary is typically longer than the abstract, up to a page, could possibly contain key visualizations, tables, or other figures that help communicate either the raw data or the results of the model, and is intended for someone outside of the data science/analytics team of an organization. It is important to be as concise as possible, and describe each of those points above without using language that is overly technical and not part of commonly used English. The executive summary is a separate document.

Note that in the abstract, executive summary, and throughout the report you should avoid using first-person singular pronouns like "I" and "me", even if you are the only author. Use "we" or use passive voice.

Introduction

Understanding and predicting changes in land cover and land use over time is critical for addressing pressing environmental challenges, such as climate change, deforestation, urbanization, and agricultural productivity. Remote sensing data, particularly from satellite imagery, provides a powerful tool for monitoring these changes over large geographic areas and long time periods. By leveraging satellite-derived spectral bands and vegetation indices, such as the Normalized Difference Vegetation Index (NDVI), researchers can classify land cover types, assess their spatial-temporal dynamics, and derive actionable insights for decision-making in environmental management.

The motivation for this project arises from the need to integrate spatial and temporal dimensions into land cover modeling. While static models offer a snapshot of land cover at a *single* point in time, they fail to capture seasonal oscillations, inter-annual variability, and spatial dependencies inherent in natural systems. For example, vegetation exhibits predictable seasonal cycles that can be modeled using sinusoidal functions, revealing patterns in NDVI amplitude and frequency that correspond to land cover types. Additionally, nearby pixels often share similar characteristics due to ecological and geographic continuity, underscoring the importance of spatial relationships in classification models. These dynamics are not only scientifically interesting but also essential for improving the accuracy and interpretability of predictive models.

The data used in this study includes satellite-derived spectral bands (e.g., from the Landsat collection) and associated geospatial features, such as NDVI, land cover labels, and pixel locations. Temporal attributes such as month and year provide the foundation for modeling seasonal and long-term trends, while spatial attributes enable the incorporation of neighborhood-level statistics. Together, these features provide excellent data for exploring spatial-temporal patterns and improving land cover classification.

This paper is organized as follows: Section 2 provides an overview of the data, including exploratory analysis and visualizations that highlight key relationships and patterns. Section 3 describes the modeling approaches, including both regression and classification methods, with a focus on integrating temporal and spatial features into the predictive framework. Section 4 discusses the visualization and interpretation of the results, comparing model performance and interpretability across approaches. Finally, Section 5 presents conclusions, recommendations, and ideas for future research directions, emphasizing the potential of spatial-temporal models for advancing land cover studies. By systematically exploring these dimensions, this project aims to contribute to a deeper understanding of land cover dynamics and their applications in environmental monitoring and management.

Data Exploration and Visualization

To develop an effective predictive framework for land cover classification, a detailed exploration of the dataset is essential. This section investigates key characteristics of the data, examines relationships between predictors and the outcome variable, and provides visual evidence to support modeling assumptions. Through descriptive statistics and visualizations, we aim to justify the choice of features, highlight relevant patterns, and assess the validity of the modeling approach.

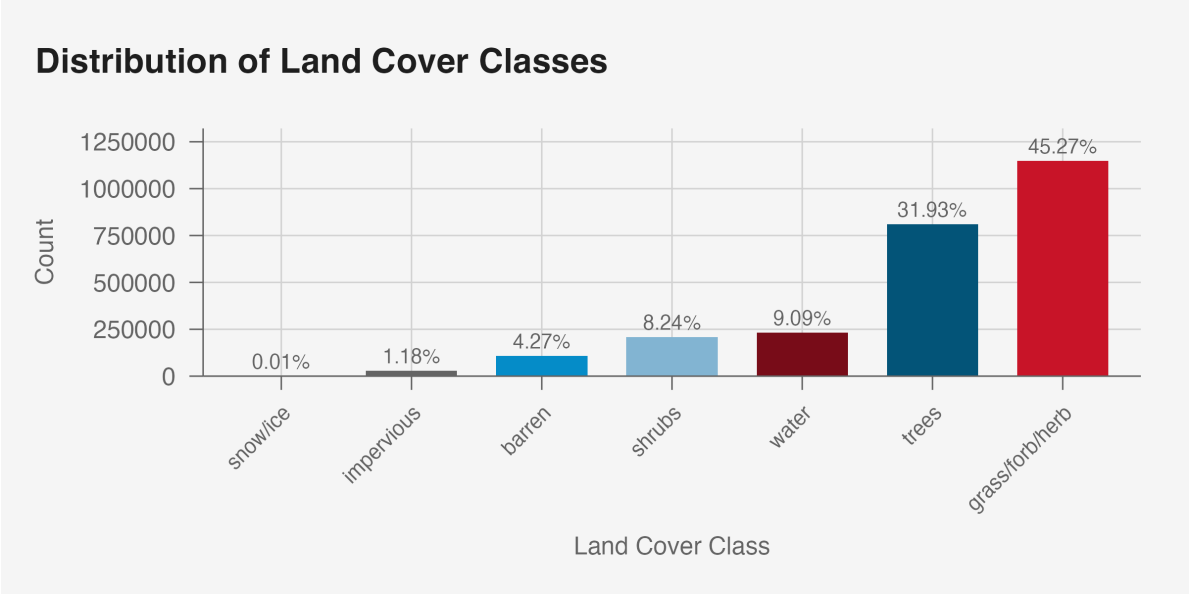
Overview of the Dataset

The dataset comprises satellite-derived spectral bands (e.g., B1 through B7), the NDVI (Normalized Difference Vegetation Index), and metadata such as plot IDs, geographic coordinates (`lat` and `lon`), and temporal information (`month` and `year`). The primary outcome of interest is `dominant_landcover`, a categorical variable representing land cover classes. A review of the dataset variables follows:

- `PlotID`: A unique identifier for each observation.
- `Month` and `Year`: Categorical, temporal features indicating the month and year of the observation.
- `SR_B1`: A continuous measurement of the "blue" spectral band, which captures visible blue light and is sensitive to water bodies and atmospheric aerosols.
- `SR_B2`: A continuous measurement of the "green" spectral band, which captures visible green light and is sensitive to vegetation health and possibly land-water boundaries.
- `SR_B3`: A continuous measurement of the "red" spectral band, which captures visible red light and is usually used to capture chlorophyll absorption in vegetation.
- `SR_B4`: A continuous measurement of the "near-infrared" spectral band, which captures near-infrared light and is sensitive to vegetation density.
- `SR_B5`: A continuous measurement of the "shortwave infrared 1" spectral band, which captures shortwave infrared light and is sensitive to moisture content in soil and vegetation.
- `SR_B7`: A continuous measurement of the "shortwave infrared 2" spectral band, which captures shortwave infrared light and differentiates vegetation stress, soil properties, and geology.
- `NDVI`: A continuous measurement of the Normalized Difference Vegetation Index, which quantifies vegetation density and health based on the contrast between red and near-infrared light (effectively, a non-linear function of `SR_B3` and `SR_B4`).
- `Lat` and `Lon`: The latitude and longitude coordinates of the pixel.
- `Dominant_Landcover`: A categorical outcome variable representing the dominant land cover class at the pixel level.
- `Dominant_LandUse`: A categorical outcome variable representing the dominant land use class at the pixel level.

- `Season`: A categorical variable for the season of observation.
- `Veg`: A binary variable indicating whether the pixel's dominant landcover is vegetated (1 for "Grass/Forb/Herb", "Shrubs", and "Trees") or non-vegetated (0 for "Snow/Ice", "Impervious", "Barren", and "Water").
- `EastWest`: A binary variable indicating whether the pixel is located east (1) or west (0) of the median longitude.
- `NorthSouth`: A binary indicating whether the pixel is located north (1) or south (0) of the median latitude.
- `Elevation_meters`: A continuous measurement of the elevation in meters above sea level at the pixel location.

Exploration of Outcome



The distribution of land cover classes reveals substantial class imbalances, which will have important implications for the modeling process. The "Grass/Forb/Herb" class constitutes the majority, accounting for 45.27% of the observations, followed by the "Trees" class at 31.93%. Together, just these two classes dominate the dataset, representing over three-quarters of all observations. Their prevalence, nevertheless, is consistent with the ecological characteristics of United States.

In contrast, other land cover classes are significantly underrepresented. For instance, "Water" and "Shrubs" account for 9.09% and 8.24% of observations, respectively, while "Barren" and "Impervious" cover only 4.27% and 1.18%. The "Snow/Ice" class constitutes a mere 0.01% of observations, reflecting minimal spatial extent. This extreme imbalance suggests challenges

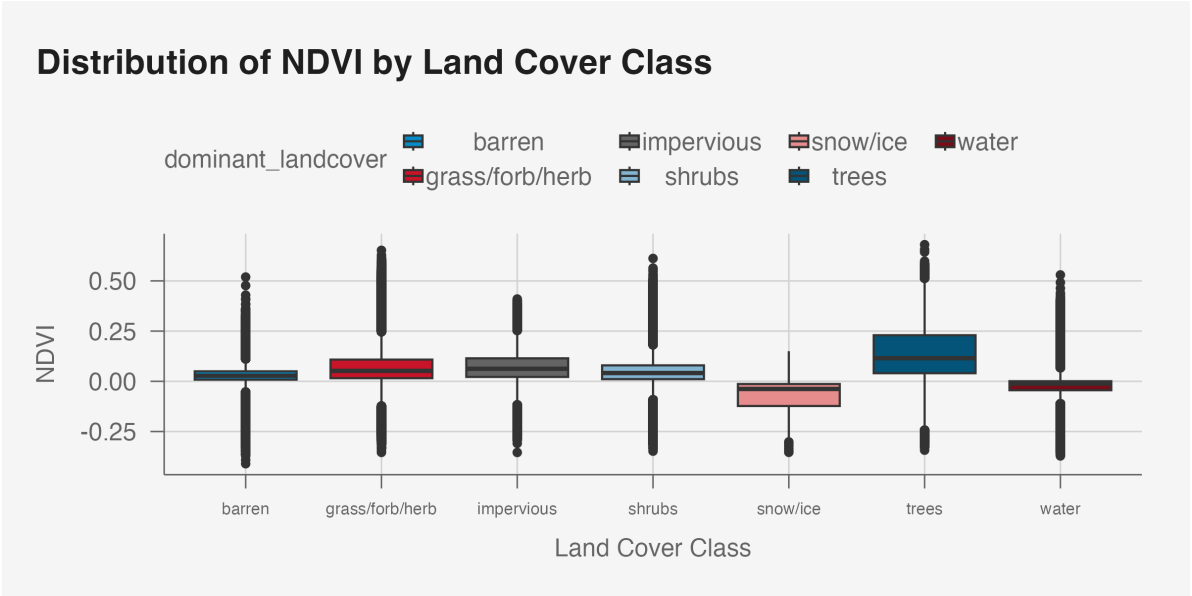
for classification models, as underrepresented classes may be overshadowed during training, leading to biased predictions.

From a modeling perspective, class imbalance necessitates careful consideration of strategies to mitigate its effects. Techniques such as class weighting, oversampling of minority classes, or undersampling of dominant classes could be employed to ensure balanced representation during training. Furthermore, performance metrics beyond overall accuracy, such as the F1-score, precision-recall curves, or area under the ROC curve (AUC), will be considered to evaluate model performance more effectively across all classes.

Exploration of Predictors

Spectral Bands and NDVI:

Visualizing the spectral band values across different land cover types reveals a potential relationship between the outcome and predictor. Below, a boxplot of NDVI across land cover classes illustrates this relationship:



The distribution of NDVI values across land cover classes reveals both distinct patterns and notable overlaps that highlight the complexities of land cover classification. As shown in the boxplot visualization, classes such as "Trees" exhibit consistently higher NDVI values (though, with a relatively wide range), indicative of dense vegetation and strong photosynthetic activity. This aligns with ecological expectations, as tree canopies reflect substantial near-infrared light and absorb red light, resulting in high NDVI values. Conversely, classes like "Water" show

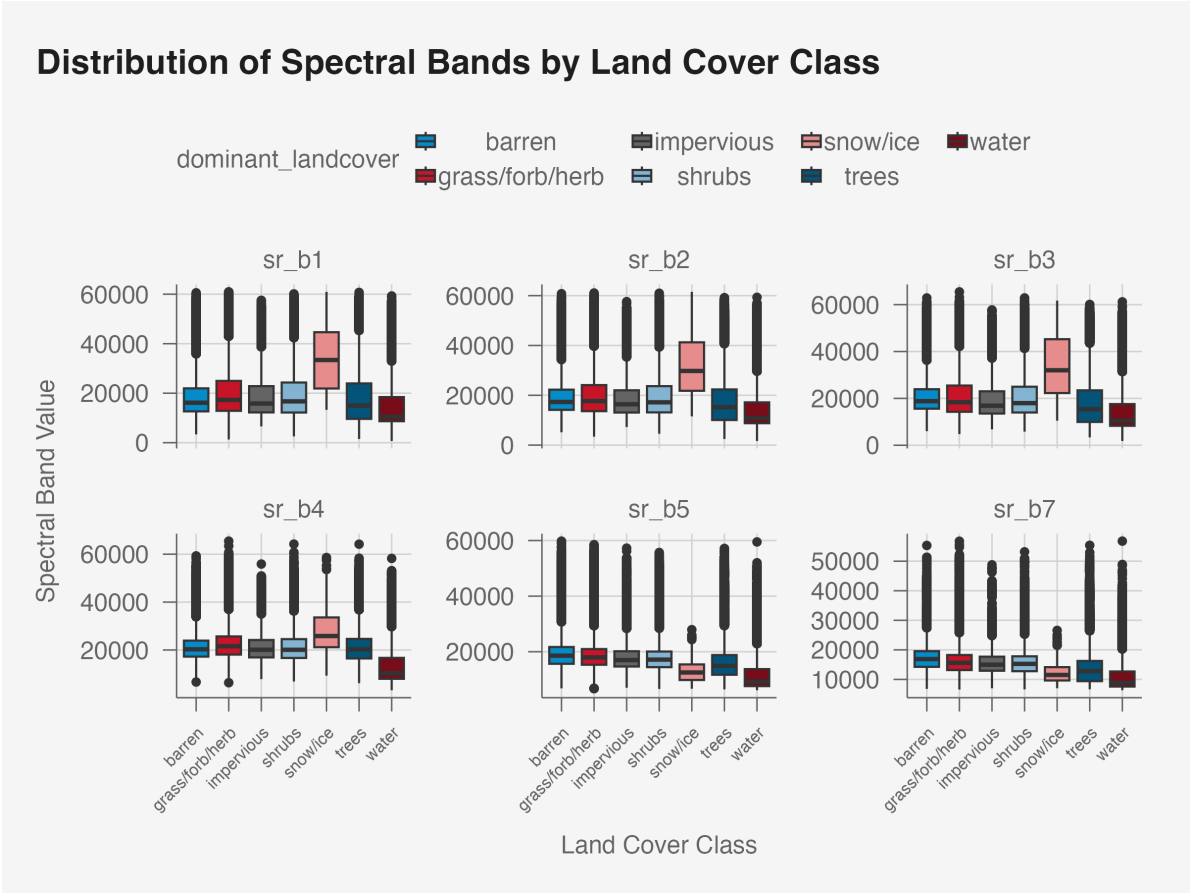
consistently negative NDVI values, reflecting the spectral signature of water, which absorbs near-infrared light and reflects visible wavelengths.

However, certain classes, such as "Grass/Forb/Herb" and "Impervious," demonstrate significant overlap in their NDVI distributions. This overlap may stem from seasonal variations in vegetation density or the heterogeneity of impervious surfaces, which can include both bare soil and constructed materials. Similarly, "Shrubs" and "Barren" exhibit NDVI ranges that overlap with both vegetative and non-vegetative classes, likely due to transitional or mixed land cover types.

This overlap underscores the need for advanced classification models that can handle complex, potentially *non*-linear relationships between predictors and outcomes. Random Forests and other ensemble methods, which are well-suited to datasets with overlapping class boundaries, are particularly promising. Additionally, incorporating supplementary predictors, such as spectral bands (e.g., B1-B7), temporal features (e.g., month and year), and spatial aggregates (e.g., mean NDVI of neighboring pixels), may further improve model performance by providing additional context for distinguishing between classes.

The variability within classes, as depicted by the interquartile ranges and outliers, also suggests the presence of heterogeneity within each land cover type. For example, the broader NDVI range for "Trees" may reflect differences in vegetation density, health, or canopy structure across regions. Similarly, the variability in "Grass/Forb/Herb" could be attributed to seasonal growth cycles or mixed vegetation types.

Ultimately, while NDVI provides meaningful separability for certain land cover classes, its limitations in distinguishing overlapping classes highlight the importance of *multi*-predictor models. This analysis justifies the use of advanced classifiers that integrate additional spectral, temporal, and spatial features to address the inherent complexity of land cover classification.



Similarly, these boxplots faceted by spectral bands across land cover classes reveal key patterns that inform the potential predictors' utility for classification. For Bands 1, 2, and 3, which correspond to the visible blue, green, and red portions of the spectrum, the distribution of values remains relatively constant across the bands. This limited variability suggests strong collinearity among these bands—suggesting the need for dimensionality reduction techniques, such as Principal Component Analysis (PCA), to consolidate their information into fewer predictors without sacrificing interpretability.

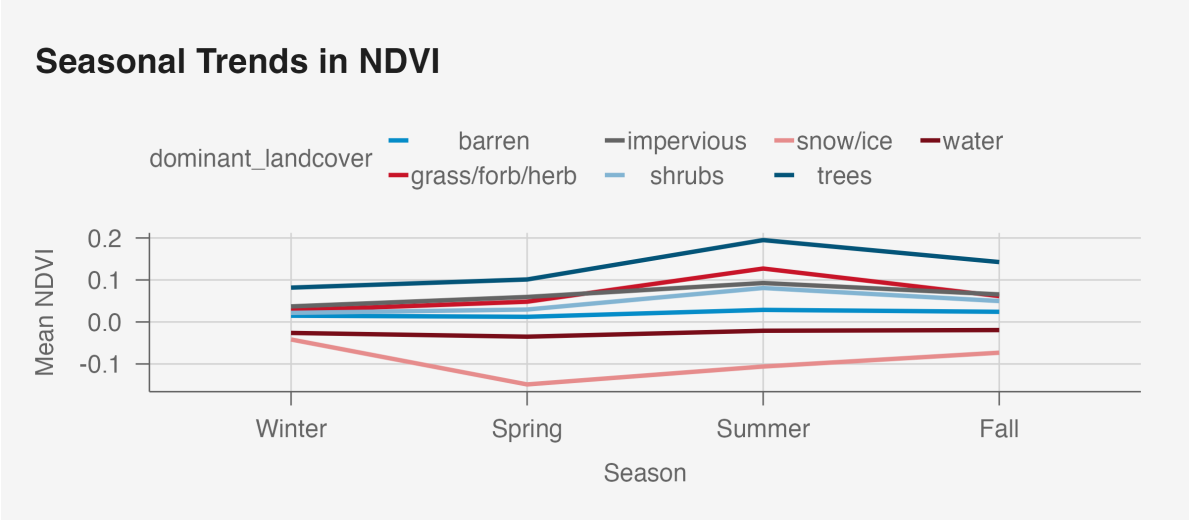
For Bands 5 and 7, which capture reflectance in the shortwave infrared (SWIR) region, the distributions show a similar pattern of consistency across bands, though with slightly more variability compared to the visible bands. These bands are known to be sensitive to moisture content and soil properties, which may provide complementary information to indices like NDVI. However, the moderate to strong inter-correlation observed between these bands also highlights the potential redundancy in their spectral information. As a result, selecting the most informative band or combining them into derived indices may enhance the model.

No single band provides clear separability between all land cover classes, as evidenced by the substantial overlap in their distributions. For example, vegetative classes such as "Shrubs,"

"Grass/Forb/Herb," and "Trees" share overlapping ranges across multiple bands, reflecting the shared spectral characteristics of these land cover types. This overlap suggests that relying on any *single* spectral band as a predictor is insufficient for robust land cover classification.

Temporal Features:

The dataset includes monthly and yearly observations, enabling an analysis of seasonal oscillations and inter-annual trends. Below, we visualize temporal variations using time-series plots:



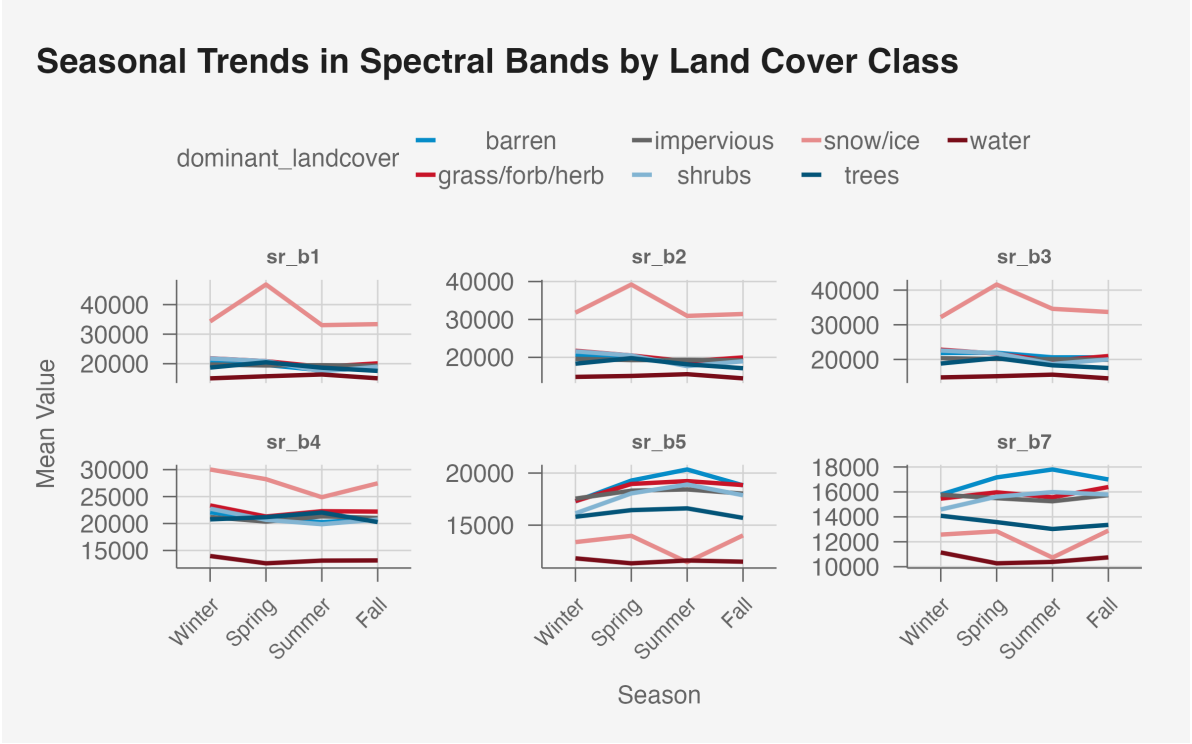
The visualization of seasonal NDVI trends across land cover classes reveals distinct temporal patterns that align with ecological expectations. As shown in the time-series plot, land cover types such as "Trees", "Grass/Forb/Herb", and "Shrubs" exhibit clear seasonal peaks during the summer months (June to August), corresponding to periods of maximum photosynthetic activity and vegetation growth. The NDVI for these classes declines during the winter months (October to February), reflecting the dormancy or reduced vegetation cover typical of temperate climates.

"Snow/Ice", on the other hand, shows a very prominent decline in NDVI in the late-Winter and Spring months (February to June), likely due to the melting of snow and ice, which reduces the reflectance in near-infrared bands. Conversely, "Water" maintains relatively stable NDVI values throughout the year, consistent with its spectral signature that is less influenced by seasonal changes. Classes like "Barren" and, albeit to a limited extent, "Impervious" also show minimal seasonal variation, as these land cover types are largely non-vegetative and therefore less responsive to seasonal shifts.

The overlap in seasonal NDVI patterns for certain classes (such as "Grass/Forb/Herb", "Impervious", and "Shrubs") underscores the importance of incorporating temporal features into the classification framework. These temporal dynamics can be modeled using sinusoidal functions, where NDVI is represented as a combination of sine and cosine terms to capture periodic oscillations. For example: $\text{NDVI}(t) = A \cdot sin(2\pi \cdot f \cdot t + \phi) + B \cdot cos(2\pi \cdot f \cdot t + \phi)$, where A and B are the amplitudes, f is the frequency (e.g., one cycle per year), t is time (month or a numerical mapping of season), and ϕ is the phase shift. These derived features, such as amplitude and phase, can serve as additional predictors in classification models.

The seasonal trends also highlight the need for models that account for temporal dependencies. Incorporating features like month or Fourier-derived seasonal components into machine learning classifiers can improve their ability to distinguish between classes with overlapping NDVI ranges. For example, while "Grass/Forb/Herb" and "Shrubs" may have similar NDVI distributions at specific times of the year, their seasonal trajectories clearly differ, providing an additional dimension for separation.

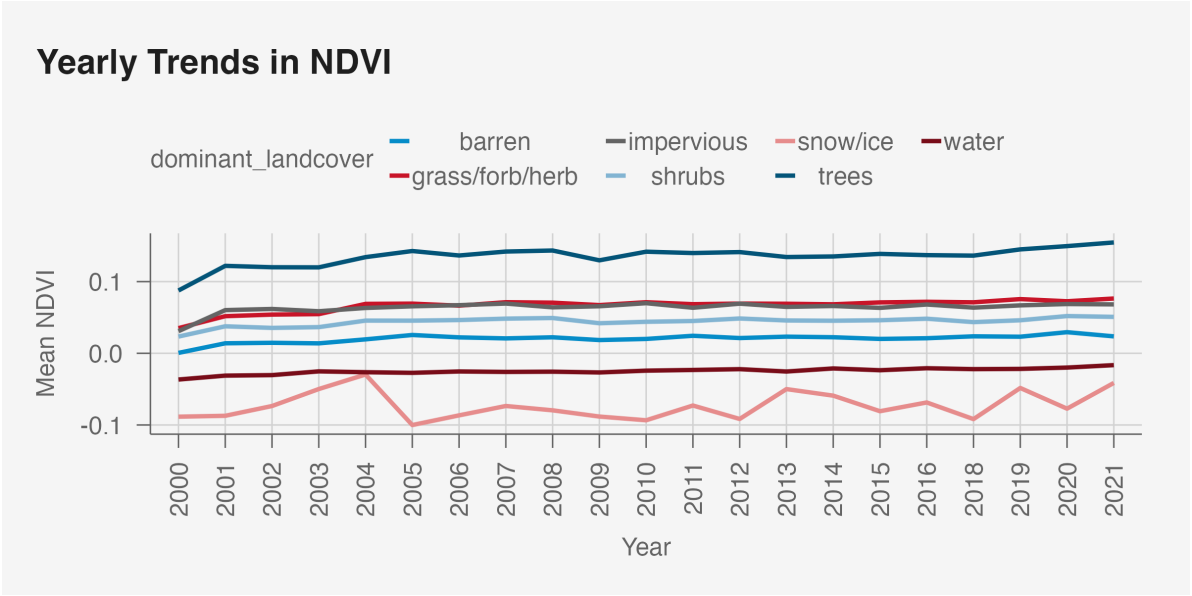
Overall, the observed periodic patterns in NDVI justify the inclusion of temporal features in the modeling framework. These features not only enhance the interpretability of the models but also provide a biologically meaningful basis for improving classification accuracy.

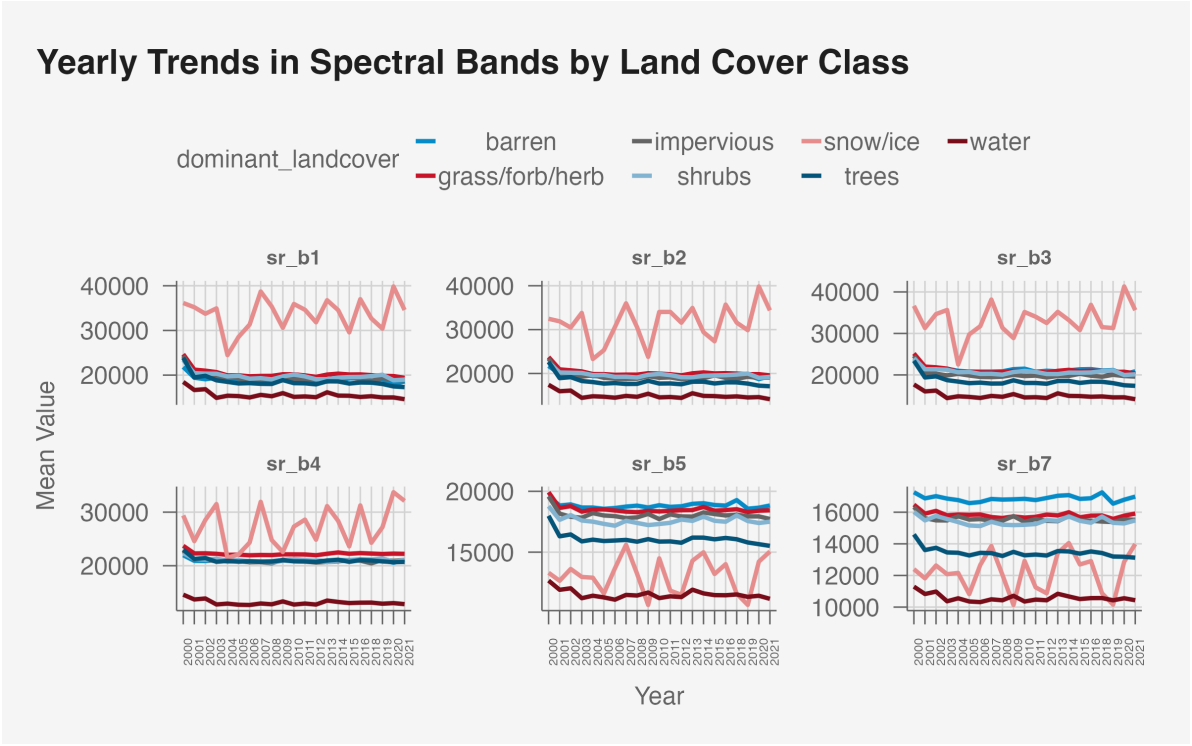


We extend this analysis to the spectral bands. The seasonal trends in spectral bands across land cover classes reveal similar periodicity, particularly in the "Snow/Ice" class, which exhibits

distinct seasonal variations across all bands. This behavior is most pronounced in bands 1, 2, and 3 (visible spectrum), where reflectance values peak in the Spring, likely due to snow and ice melt influencing surface properties. The similarity in trends across these bands reinforces their shared sensitivity to visible light, suggesting they capture *redundant* information.

Shortwave infrared bands 5 and 7 display more nuanced seasonal trends, particularly for "Trees" and "Barren." These classes show clear increases in reflectance during summer, for "Trees" aligning with the seasonal growth of vegetation and moisture-related changes that SWIR bands are sensitive to. Notably, the "Impervious" class remains relatively stable across all seasons in these bands, reflecting the consistency of its surface properties. The overall variability in Band 4 is less pronounced compared to Bands 5 and 7, which suggests SWIR bands may provide additional discriminative power for land cover classes with more prominent moisture or structural changes.





The analysis of yearly NDVI trends provides valuable insights into the stability and variability of vegetation dynamics across different land cover classes from 2000 to 2021 (the time range we focus on). The "Trees" class demonstrates consistently high NDVI values throughout the period, with a slight upward trend in recent years. This stability reflects the resilience of forested areas within the dataset and suggests minimal deforestation or even reforestation in some regions. Similarly, the "Grass/Forb/Herb" and "Shrubs" classes exhibit relatively stable NDVI values with limited inter-annual variability, indicative of consistent vegetation cover. These trends highlight the long-term stability of these vegetative land cover types, which is critical for assessing ecosystem health and predicting land cover changes.

In contrast, the "Snow/Ice" class exhibits considerable variability in NDVI, particularly between 2002 and 2005. This fluctuation may be attributable to climatic factors such as variations in snow and ice coverage due to changing weather patterns or broader impacts of global warming. The unstable NDVI trend observed for this class aligns with the hypothesis of reduced snow and ice cover over time, signaling potential shifts in cryospheric dynamics that merit further investigation. Non-vegetative class "Barren" maintains consistently low NDVI values, reflecting their lack of vegetation and spectral characteristics. This stability over time indicates minimal changes in land use or surface conditions in areas dominated by barren land. The "Water" class similarly shows stable negative NDVI values near zero, confirming its spectral separability and consistency over the study period. Concerning is the nearly identical trends

among "Impervious" and "Grass/Forb/Herb" which suggest distinguishing these classes will prove difficult.

The observed trends suggest that temporal features, such as year, may provide *limited* additional predictive power for classes like "Impervious" and "Grass/Forb/Herb," given their relative stability. However, for classes like "Snow/Ice," where inter-annual variability is more pronounced, temporal features may enhance the model's ability to capture dynamic changes. Overall, this analysis underscores the potential of temporal trends in NDVI to provide meaningful insights into land cover changes while emphasizing the need for tailored modeling approaches that account for the *unique* characteristics of each land cover class.

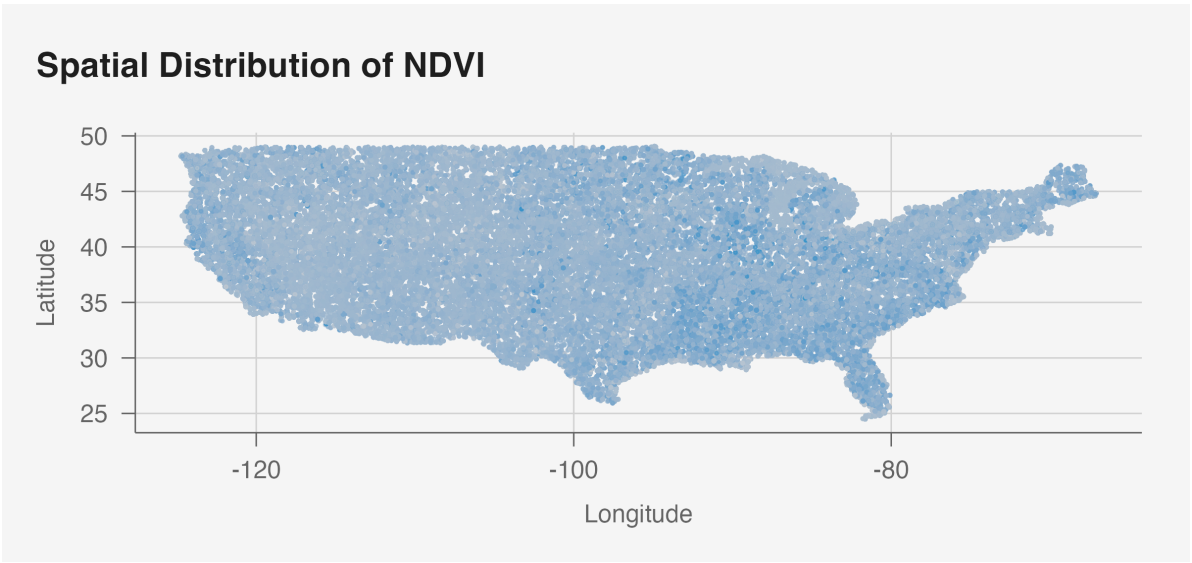
The analysis of yearly spectral band trends aligns closely with the NDVI observations, particularly in the clear separability of the "Snow/Ice" and "Water" classes. "Snow/Ice" demonstrates pronounced periodicity across visible bands (B1, B2, B3), with reflectance peaks likely driven by seasonal variations in snow cover. The Water class, in contrast, maintains consistently low reflectance values across all bands, highlighting its distinct spectral signature.

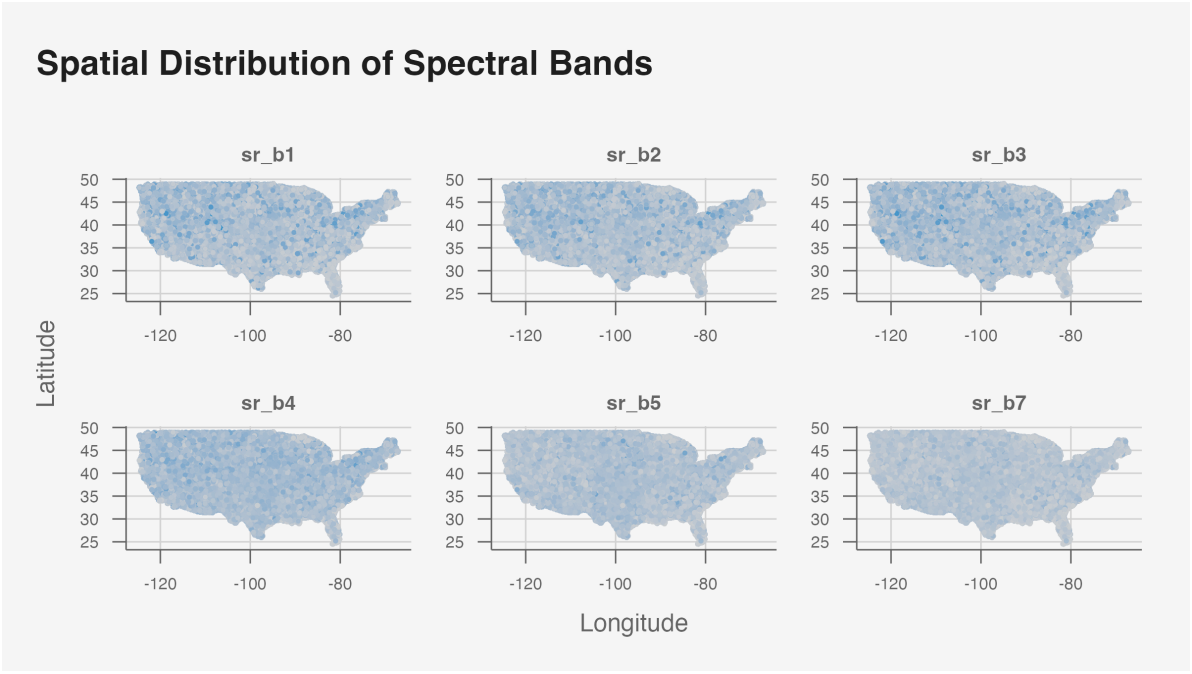
For the remaining land cover classes, visible bands (B1, B2, B3) show limited variability, reaffirming their redundancy in distinguishing between vegetative and non-vegetative classes. However, Bands 5 (SWIR1) and 7 (SWIR2) provide more nuanced differences, particularly for classes like "Trees" and "Barren," where seasonal and inter-annual changes are more evident.

Exploration of Spatial Patterns

Spatial Relationships:

The geographic proximity of pixels merits exploration of potential correlations between neighboring observations. A heatmap of NDVI values can highlight spatial continuity:





The visualization of NDVI and spectral band values across the United States highlights the spatial continuity of the predictors, where adjacent pixels exhibit similar values due to shared environmental and land cover characteristics.

High NDVI values, indicative of dense vegetation and robust photosynthetic activity, are concentrated in the eastern U.S. and parts of the Pacific Northwest which are regions charac-

terized by forested landscapes and favorable climatic conditions that support high vegetation density.

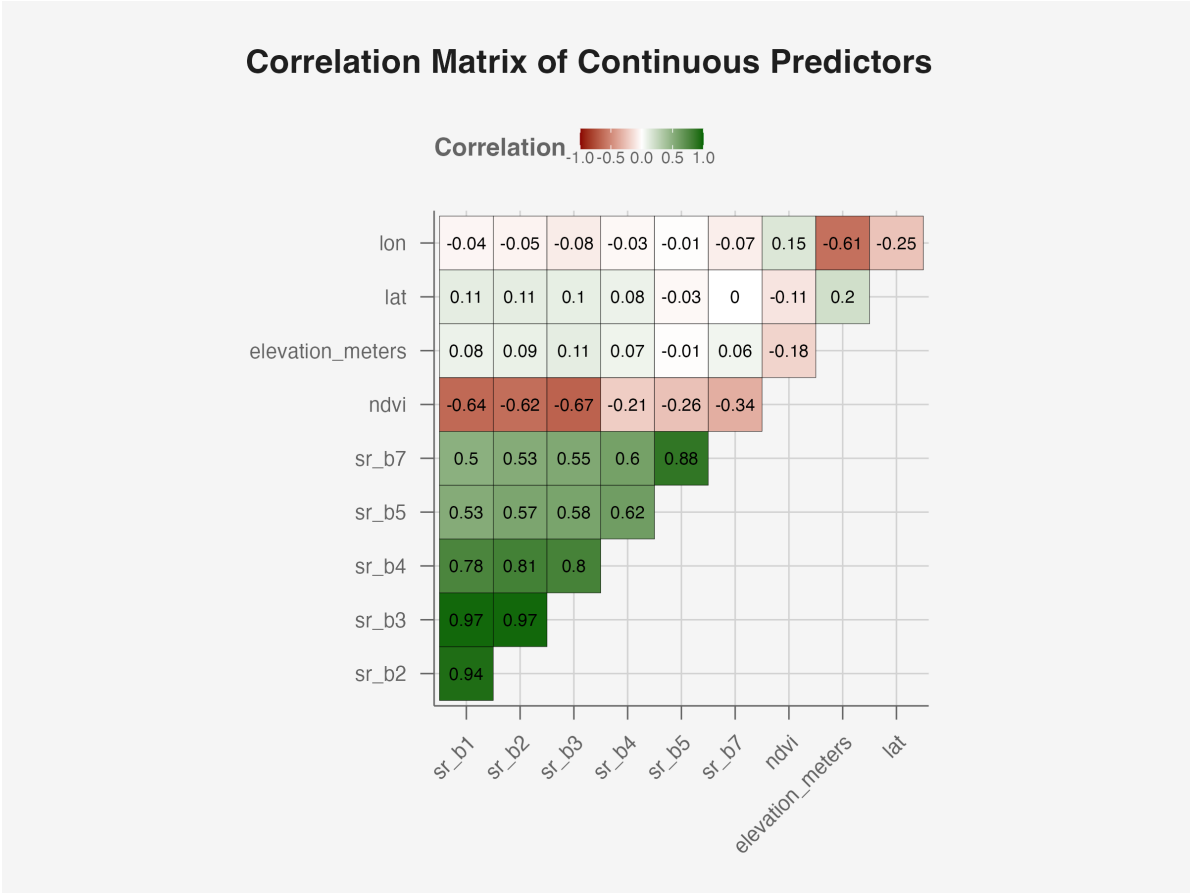
In contrast, areas with low NDVI values are predominantly located in the arid western and southwestern United States. These regions, dominated by deserts, sparse vegetation, and barren land, naturally exhibit reduced photosynthetic activity. The stability and clustering of NDVI in these regions emphasize the potential for spatial correlations to enhance predictive modeling. For example, neighboring pixels are likely to belong to the same land cover class, providing an opportunity to use spatial aggregates such as the mean or variance of NDVI values within a defined neighborhood as additional predictors in classification models.

Furthermore, the spatial clustering observed in the map suggests the importance of incorporating spatial relationships into the modeling framework. By considering the influence of nearby pixels, the model may better capture the continuity and heterogeneity inherent in natural landscapes. For instance, spatial smoothing or feature engineering techniques that aggregate NDVI values across a local neighborhood can provide additional context for distinguishing between land cover types with similar spectral signatures.

This analysis supports the hypothesis that spatial dependencies play a critical role in determining NDVI variability and, by extension, land cover classification. Incorporating spatial features may not only enhance model performance but also align with the underlying ecological processes that govern land cover patterns. We will leverage these spatial relationships by integrating *neighborhood-level* statistics.

Preliminary Correlation Analysis

A correlation matrix of predictors (e.g., spectral bands and NDVI) helps identify multicollinearity, which may impact model selection:



The correlation matrix provides insights into the relationships among the continuous variables, highlighting both redundancies and potential predictors for land cover classification.

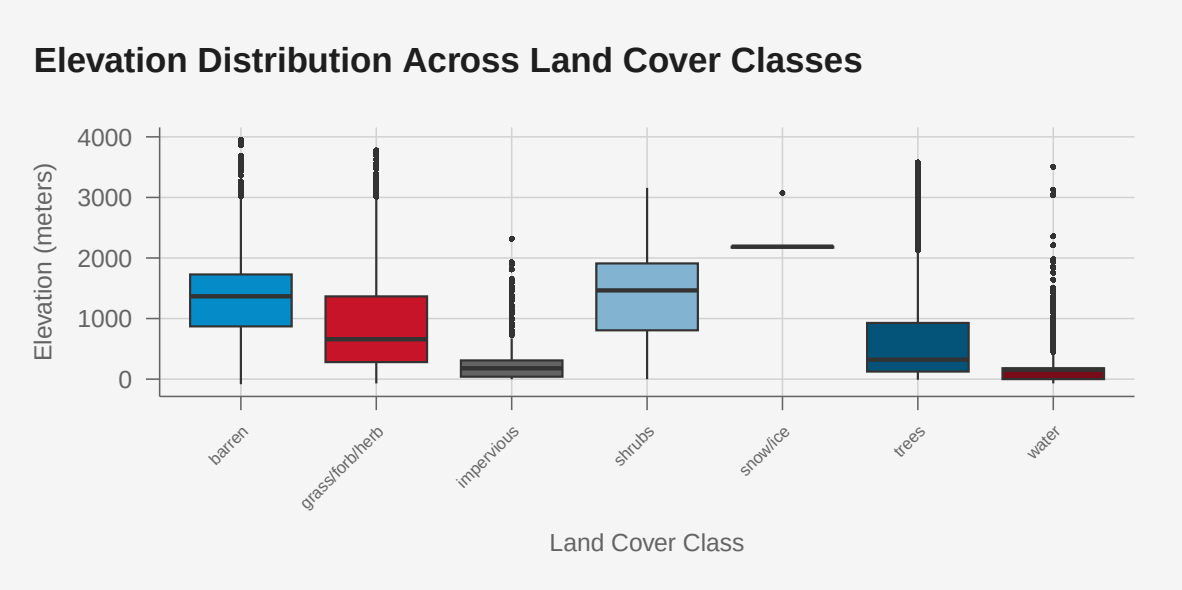
While NDVI exhibits weak to moderate negative correlations with several spectral bands, these relationships highlight the inherent complexity of vegetation dynamics and the limitations of *linear* correlation measures. Specifically, NDVI demonstrates weak negative correlations with the near-infrared band `SR_B4` (-0.21) and the shortwave infrared band `SR_B5` (-0.26), aligning with the fact that NDVI is a *non*-linear function of red and near-infrared reflectance. The weak linear correlation underscores that the predictive power of these bands may lie not in their independent contributions but in their interactions, which may be more effectively captured through non-linear models.

The visible bands `SR_B1`, `SR_B2`, and `SR_B3` exhibit near-perfect collinearity, with correlation coefficients exceeding 0.94. This strong interdependence indicates that these bands measure nearly identical surface properties in the visible spectrum, such as reflectance from bare soil, water, and vegetation—aligning with previously identified redundancies across the predictor categories. While these bands are critical for understanding vegetation absorption and reflection in the blue, green, and red ranges, their redundancy poses a risk of multicollinearity in

predictive models. Without dimensionality reduction, their inclusion as independent predictors may lead to instability and overfitting.

Beyond the visible spectrum, the infrared bands `SR_B4`, `SR_B5`, and `SR_B7` show moderate to strong inter-correlations, with values above 0.6 in some cases. These bands capture distinct but related properties, including vegetation structure, moisture content, and thermal properties. While their interdependence is not *as* pronounced as the visible bands, it suggests complementary information that could enhance classification models when used judiciously. In contrast, spatial variables such as latitude and longitude display negligible correlations with NDVI and other spectral predictors, indicating that geographic location alone may not strongly predict vegetation health. Elevation, similarly, shows weak correlations across all variables, suggesting it may have a limited direct impact on spectral properties in this dataset. However, spatial variables could still prove valuable in capturing non-linear regional trends or serving as inputs for spatial-temporal models.

The observed correlations provide critical guidance for feature engineering and model selection. The weak correlations between NDVI and individual spectral bands further reinforce the need to explore non-linear models, such as Random Forests, that can capture complex interactions. At the same time, the high collinearity among certain bands necessitates dimensionality reduction to mitigate redundancy and improve computational efficiency. Incorporating indices such as NDVI or enhanced vegetation indices may allow for a more nuanced representation of vegetation health, leveraging the interplay of spectral bands more effectively. While spatial and elevation variables appear weakly correlated with NDVI, they may still play an important role in explaining regional variations when combined with other predictors.



Section Conclusion

This exploratory analysis establishes the relationships between predictors and the outcome variable, supports the inclusion of temporal and spatial features, and provides insight into approaches for model selection. The next section will focus on modeling approaches, detailing the assumptions, methodology, and evaluation metrics used to predict land cover classes.

Modeling and Analysis

The objective of this section is to formalize and implement predictive models to classify land cover types using the predictors identified in the data exploration phase. The models we employ must address key challenges, including the complexity of spectral, temporal, and spatial relationships, overlapping class boundaries, and collinearity among predictors. This section details the regression and classification models tested, their assumptions, implementation, and evaluation.

Problem Setup

The task involves predicting the dominant land cover class (`dominant_landcover`), a categorical variable with multiple classes, using a set of continuous and categorical predictors. Specifically:

- Observation (X): Each row corresponds to a unique satellite-derived observation at a specific time (`year`, `month`) and geographic location (`lat`, `lon`).
- Predictors (columns of X): Continuous predictors include spectral bands (`SR_B1`, `SR_B2`, … `SR_B7`), `NDVI`, elevation (`elevation_meters`), and spatial-temporal variables like `month`, `year`, and `season`. Derived spatial aggregates and sinusoidal features (e.g., sine/cosine terms) will also be introduced during feature engineering.
- Outcome (y): The target variable is `dominant_landcover`, representing land cover classes such as "Trees," "Water," "Snow/Ice," etc.

Baseline Models: Multinomial Logistic Regression Model

Assumptions:

The multinomial logistic regression model assumes a *linear* relationship between the predictors and the log-odds of class membership. While this assumption simplifies interpretation, it

imposes constraints that may not align with the inherent complexity of the land cover classification problem. Specifically, land cover classes often exhibit non-linear relationships, overlapping boundaries, and significant multicollinearity among predictors (e.g., spectral bands), which may reduce the model's ability to generalize effectively. To mitigate multicollinearity, Principal Component Analysis (PCA) was employed to transform the highly correlated spectral bands into orthogonal components, thereby preserving the majority of variance while reducing redundancy and computational complexity.

Model Specification:

The multinomial logistic regression model predicts the probability of a pixel y belonging to a given land cover class C, where $c \in [1, K]$, as follows:

$$P(y = c|X) = \frac{\exp(\beta_{c_0} + \beta_c^\top X)}{\sum_{k=1}^{K} \exp(\beta_{k_0} + \beta_k^\top X)}$$

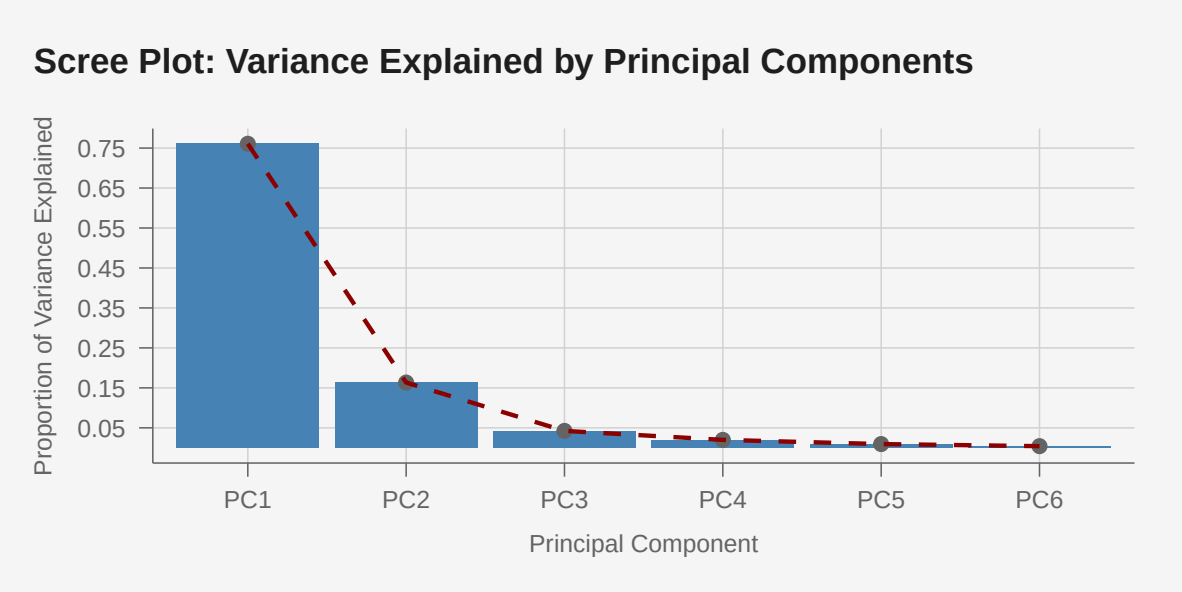
where X represents the predictors (spectral bands, NDVI, temporal features), β_c is the vector of coefficients for class c, and β_{c_0} is the intercept for class c.

The model estimates β using maximum likelihood estimation to minimize the discrepancy between the predicted and observed probabilities.

Implementation:

The predictors used in the model include spectral bands (`SR_B1` to `SR_B7`), `NDVI`, temporal features (`season`), elevation bins, vegetation indicators, and spatial features (`EastWest` and `NorthSouth`). However, due to the strong collinearity observed among the spectral bands (e.g., `SR_B1`, `SR_B2`, and `SR_B3`), dimensionality reduction is applied using Principal Component Analysis (PCA). PCA transforms the original spectral band features into orthogonal components, retaining the majority of the variance in the data while reducing redundancy and computational complexity.

Principal Component Analysis (PCA)



The scree plot of explained variance demonstrates that the first principal component (PC1) accounts for approximately 75% of the variance, while the second principal component (PC2) explains an additional 15%—together explaining over 90% of the total variance. Beyond PC2, the contribution of additional components (PC3 to PC6) is minimal, collectively explaining less than 10% of the variance, indicating that their inclusion offers limited additional information.

Given this pattern, we retain only PC1 and PC2 for downstream modeling, as they account for the vast majority of the variance in the spectral bands while significantly reducing dimensionality. Selecting only PC1 and PC2 simplifies the model and improves computational efficiency while maintaining the interpretability of the dominant trends in the spectral bands.

We proceed to fit the multinomial logistic regression model using PC1, PC2, `NDVI`, `season`, `elevation_meters`, `veg`, `east_west`, and `north_south` as predictors. We standardize the continuous predictors to have mean 0, standard deviation 1 to ensure that coefficients are on the same scale and interpretability is consistent across features.

Performance Evaluation:

Model performance is evaluated using accuracy, precision, recall, and F1-score.

Class	Precision	Recall	F1	Balanced Accuracy
Barren	0.8534	0.9158	0.8835	0.9544
Grass/Forb/Herb	0.6403	0.7816	0.7039	0.7091
Impervious	0.4747	0.2310	0.3108	0.6140
Shrubs	0.0561	0.0028	0.0052	0.4993
Snow/Ice	0.4375	0.0308	0.0576	0.5154
Trees	0.6425	0.5991	0.6200	0.7214
Water	0.9057	0.9357	0.9205	0.9630
Macro Averaged	**0.5729**	**0.4995**	**0.5002**	**0.7109**

The performance metrics for the multinomial logistic regression model demonstrate both strengths and limitations, reflecting variability in the model's ability to classify different land cover classes. The model performs well for dominant classes such as "Grass/Forb/Herb," "Trees," and "Water," achieving F1-scores of 0.704, 0.620, and 0.920, respectively. Notably, "Water" stands out as the best-performing class, with exceptionally high precision (0.906) and recall (0.936). This strong performance is likely attributable to the distinct spectral signature of water, which facilitates accurate classification. Similarly, the relatively high recall (0.782) and precision (0.640) for "Grass/Forb/Herb" indicate the model's capability to capture prevalent vegetative classes.

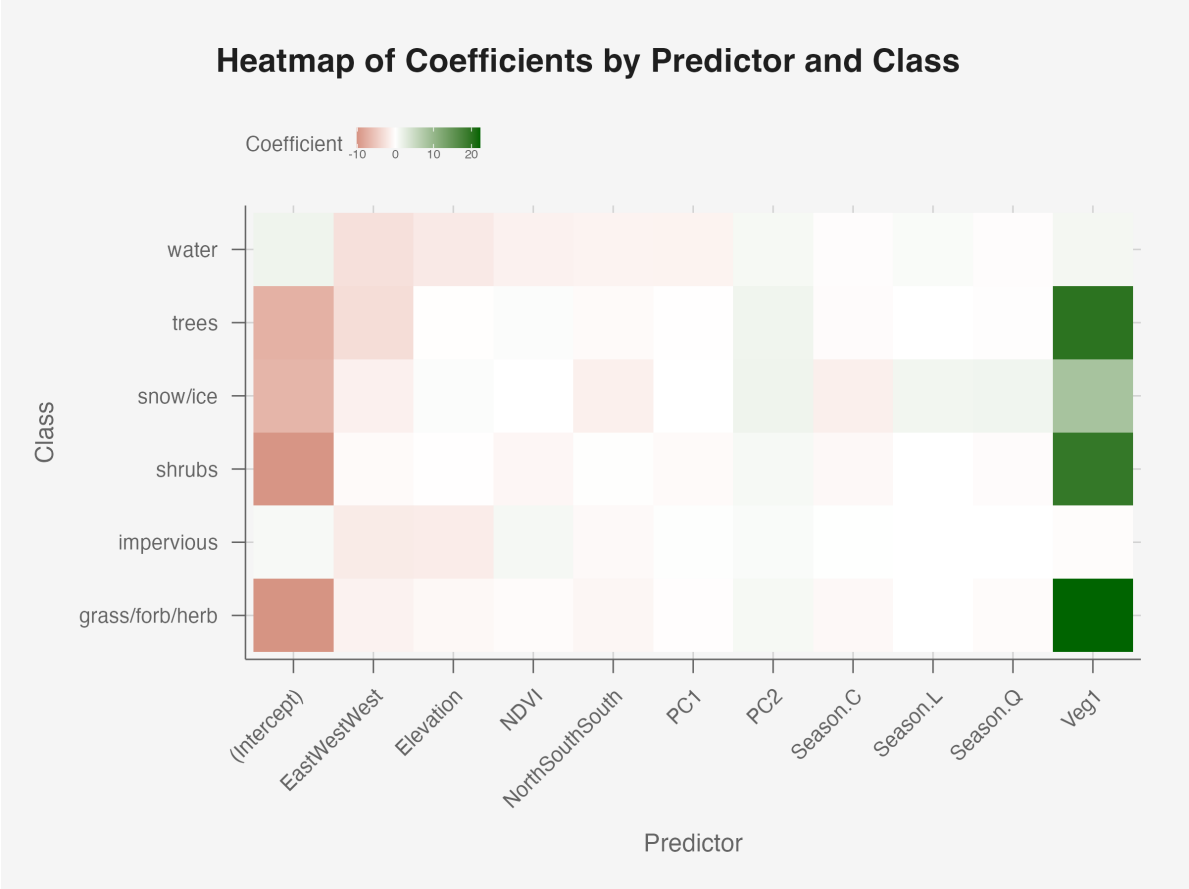
In contrast, the model struggles significantly with minority and spectrally overlapping classes. For instance, the "Shrubs" class exhibits an F1-score of only 0.005, driven by extremely low recall (0.003) and precision (0.056). Likewise, the "Snow/Ice" class achieves an F1-score of 0.058, with a recall of 0.031, indicating that these classes are rarely identified correctly. These results highlight the challenges of classifying underrepresented categories, where insufficient training examples and overlapping spectral features hinder the model's ability to generalize.

The "Impervious" class also performs poorly, with a recall of 0.231 and an F1-score of 0.311. This suggests difficulty in distinguishing impervious surfaces from similar classes, like "Grass/Forb/Herb" which we previously observed shared similar predictor distributions. The relatively low precision for this class (0.475) further underscores the model's inability to effectively separate it from other categories.

Macro-averaged metrics, including an F1-score of 0.500 and balanced accuracy of 0.711, provide an overall assessment of model performance. While the balanced accuracy indicates moderate effectiveness across all classes, the macro-averaged F1-score reflects the challenges posed by poor performance on minority classes. The disparity between these metrics suggests that while the model is effective for majority and spectrally distinct classes, it fails to adequately address class imbalance and overlapping class distributions.

These results highlight the model's strengths in classifying dominant, spectrally distinct classes and its weaknesses in handling overlapping class distributions and underrepresented categories.

Interpretation:



The heatmap of coefficients provides an empirical representation of the multinomial logistic regression model, offering insights into how each predictor influences the relative log-odds of classification into a specific land cover class. Each cell represents the coefficient for a predictor-class pair, where positive values indicate an increase in the log-odds of a pixel belonging to that class, while negative values denote a decrease relative to the reference category. Notably, the Vegetation Indicator (`Veg:1`) exhibits the largest positive coefficients across vegetated classes such as "Trees," "Shrubs," and "Grass/Forb/Herb", compared to the non-vegetated reference category (`Veg:0`). This result aligns with observations from the data exploration phase, where vegetated classes consistently demonstrated high NDVI values and distinct spectral patterns indicative of healthy vegetation. Conversely, non-vegetated classes, including "Water," "Snow/Ice," and "Impervious," display negligible or negative coefficients for `Veg:1`, reflecting their spectral dissimilarity from vegetated surfaces.

The intercept terms, particularly for "Grass/Forb/Herb" and "Trees," are non-negligible and negative, suggesting higher baseline probabilities for these majority classes when all predictors are at their reference values. This observation reinforces the class distribution imbalance

observed during the outcome exploration, where these vegetated classes accounted for over 75% of all observations.

The spatial predictors, `EastWest` and `NorthSouth`, reveal distinct trends. `EastWest:West` shows the largest negative coefficients for "Water" and "Trees", indicating a reduced likelihood of these classes occurring in the western regions. Conversely, "Grass/Forb/Herb" and "Shrubs" display smaller negative coefficients, suggesting a more balanced east-west distribution for these classes. This trend aligns with geographic observations, where dense vegetation (trees) and water bodies are more prominent in the eastern United States. `NorthSouth:South`, on the other hand, exhibits negligible coefficients across most classes, with a modest negative value for "Snow/Ice". This pattern is consistent with the climatological expectation that snow and ice are less likely to appear in southern regions closer to the equator, where temperatures are higher.

The `NDVI` coefficients highlight its varying relevance across land cover types. `NDVI` is slightly positive for "Impervious" surfaces, likely reflecting bare soil or sparse vegetation associated with impervious areas. In contrast, `NDVI` is negative (albeit, minimally) for "Water" and "Shrubs", consistent with their low or intermediate vegetation density, which was observed in the `NDVI` boxplots during data exploration. The negative relationship for "Water" reflects its distinct spectral signature, as water absorbs near-infrared light, resulting in consistently low NDVI values.

The principal components (`PC1` and `PC2`) derived from spectral bands demonstrate subtle yet interpretable patterns. PC1 exhibits negligible coefficients across most classes, except for small negative values for "Shrubs" and "Water". This result suggests that PC1 primarily captures spectral variations that do not contribute meaningfully to class separability. In contrast, PC2 displays small but positive coefficients across multiple classes, most prominently for "Trees" and "Snow/Ice". This positive relationship may reflect spectral nuances captured in PC2 that are more aligned with vegetated and cryospheric surfaces. Notably, "Impervious" surfaces show near-zero coefficients for PC2, suggesting minimal spectral distinction along this principal component axis.

The seasonal predictors (`Season:C`, `Season:L`, `Season:Q`) exhibit small but meaningful coefficients that align with their polynomial contrasts, given the ordered nature of the seasons. In multinomial regression, ordered factors are modeled using orthogonal polynomials to capture linear, quadratic, and cubic trends. `Season.C` (linear trend) is slightly negative for "Snow/Ice", reflecting a steady seasonal decline as winter transitions into warmer seasons. `Season.L` (quadratic trend) and `Season.Q` (cubic trend) coefficients are most noticeable and positive for "Snow/Ice", indicating that the relationship between seasonality and snow cover is non-linear. This aligns with the earlier seasonal NDVI analysis, which revealed pronounced temporal peaks and troughs for "Snow/Ice". For other classes, the seasonal predictors have minimal influence, reinforcing the earlier observation that the linear separability of temporal trends is limited.

The heatmap confirms and extends findings from the data exploration phase. The spatial predictors align with expected geographic distributions, particularly for "Water," "Trees," and "Snow/Ice". NDVI and principal components reveal varying relevance, with NDVI strongly differentiating "Water" and "Impervious" surfaces, while PC2 captures subtle spectral distinctions for "Trees" and "Snow/Ice". Seasonal predictors highlight the importance of non-linear trends, particularly for "Snow/Ice".

Appropriateness:

The multinomial logistic regression model serves as a simple and interpretable baseline for land cover classification. Its linear assumptions make it *less suitable* for the inherent non-linear relationships and overlapping boundaries observed in the data, particularly among spectrally similar classes such as "Shrubs" and "Grass/Forb/Herb." While PCA mitigated multicollinearity among spectral bands by reducing dimensionality, it abstracts raw spectral information, potentially limiting fine-grained class separability. Moreover, the model's performance is heavily influenced by class imbalance, as evidenced by strong results for majority classes like "Water" and "Trees" but poor performance for underrepresented classes such as "Shrubs" and "Snow/Ice." Thus, we may need a more flexible, non-linear models to handle the data's complexity.

Nevertheless, a key strength of the multinomial logistic regression model is its interpretability. Coefficients provide clear insights into predictor effects: positive values increase the log-odds of class membership, while negative values decrease it. For instance, the Vegetation Indicator (`Veg:1`) strongly differentiates vegetated classes ("Trees," "Grass/Forb/Herb") from non-vegetated ones, aligning with NDVI trends observed during data exploration. Spatial predictors (`EastWest` and `NorthSouth`) capture meaningful geographic patterns, such as the higher prevalence of "Water" and "Trees" in eastern regions.

NDVI coefficients confirm its ecological relevance, with negative values for "Water" and slight positives for "Impervious," consistent with their spectral signatures. Principal components (PC1, PC2) show subtle yet interpretable contributions, with PC2 capturing spectral variations for "Trees" and "Snow/Ice." Seasonal predictors, modeled through polynomial contrasts, reveal non-linear temporal patterns, notably for "Snow/Ice," which peaks in winter and declines thereafter.

While the results are interpretable and align with prior observations, the model's linear constraints and limited handling of class imbalance reduce its effectiveness for minority and complex classes. A more flexible, non-linear approach would better capture the data's intricacies while maintaining clarity of interpretation.

Temporal Models: Sinusoidal Regression

Seasonal periodicity observed in NDVI and spectral bands indicates that temporal dynamics are not adequately captured by standard linear models. To address this, sinusoidal regression introduces sine and cosine components to model periodic oscillations, which align with seasonal trends and provide a continuous representation of time-varying effects.

Model Specification:

Sinusoidal regression expresses temporal patterns as a combination of sine and cosine functions, mathematically represented as:

$$f(t) = A \cdot \sin(2\pi f t + \phi) + B \cdot \cos(2\pi f t + \phi)$$

Here:

- t represents the temporal variable (e.g., the numeric representation of `season`)
- A and B are the amplitude coefficients, which are estimated during model training. These coefficients determine the magnitude of the oscillation and capture the contribution of the sine and cosine terms to the overall periodic pattern.
- f denotes the frequency of the oscillation, which corresponds to the natural periodicity of the phenomenon under study. For seasonal data divided into four quarters (Winter, Spring, Summer, Fall), $f = \frac{1}{4}$, indicating one full cycle per year.
- ϕ represents the phase shift, accounting for the starting point of the oscillation. Since the data aligns naturally with the seasons (e.g., starts in the Winter), we assume ϕ to be 0 to simplify the model.

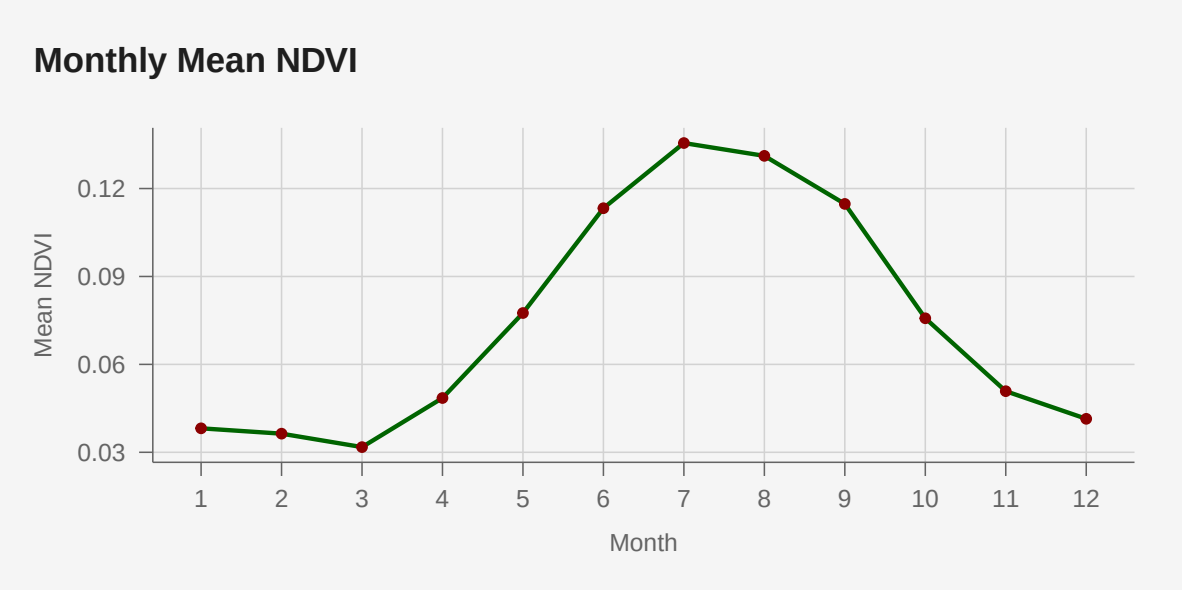
Thus, we extend the multinomial logistic regression model by incorporating sine and cosine transformations of the seasonal variable:

$$P(y = c|X, t) = \frac{\exp(\beta_{c_0} + \beta_c^\top X + \alpha_c \sin(2\pi f t) + \gamma_c \cos(2\pi f t))}{\sum_{k=1}^{K} \exp(\beta_{k_0} + \beta_k^\top X + \alpha_k \sin(2\pi f t) + \gamma_k \cos(2\pi f t))},$$

where α_c and γ_c represent the class-specific coefficients for the sine and cosine terms, respectively—quantifying how temporal periodicity influences the log-odds of classification into each land cover class.

Implementation:

Fourier Transformation



To account for the clear seasonal periodicity observed in `NDVI`, we performed a Fourier Transformation to identify the dominant frequency driving temporal oscillations. This step was motivated by the strong seasonal trends observed in the data, where `NDVI` peaks during the summer months and declines in winter, reflecting vegetation growth cycles. The Fourier Transformation decomposes the `NDVI` time series into its constituent frequencies, allowing us to quantify periodic patterns and determine the most significant frequency contributing to the observed variations.

The results of the Fourier Transformation indicate that the dominant frequency corresponds to $f = \frac{1}{12}$, representing a single annual cycle. This frequency accounts for the majority of the variance in `NDVI`, while higher-order frequencies have negligible contributions. This finding confirms that `NDVI` follows a regular, annual oscillatory pattern, which aligns with ecological expectations of seasonal vegetation changes.

To incorporate this seasonality into the multinomial regression model, we transformed the `month` variable into two sinusoidal features: $\sin(2\pi \cdot f \cdot \text{month})$ and $\cos(2\pi \cdot f \cdot \text{month})$. These sine and cosine terms capture the periodic nature of `NDVI` across months, providing a smooth representation of temporal trends. By including these derived features as predictors, the model gains the ability to account for seasonal variations in a flexible yet interpretable manner, enhancing its capacity to distinguish between land cover classes with strong temporal signals, such as "Grass/Forb/Herb" and "Snow/Ice."

Performance Evaluation:

Model performance is evaluated using accuracy, precision, recall, and F1-score.

Class	Precision	Recall	F1	Balanced Accuracy
Barren	0.8541	0.9160	0.8839	0.9545
Grass/Forb/Herb	0.6403	0.7789	0.7029	0.7085
Impervious	0.4729	0.2320	0.3112	0.6144
Shrubs	0.0555	0.0027	0.0052	0.4993
Snow/Ice	0.4706	0.0352	0.0656	0.5176
Trees	0.6396	0.6003	0.6193	0.7208
Water	0.9059	0.9357	0.9206	0.9630
Macro Averaged	**0.5770**	**0.5001**	**0.5012**	**0.7112**

Non-Linear Models: Random Forests

Given the limitations of linear models, Random Forests are introduced to capture non-linear relationships and feature interactions. Random Forests are ensemble models that construct multiple decision trees and combine their outputs to improve classification accuracy.

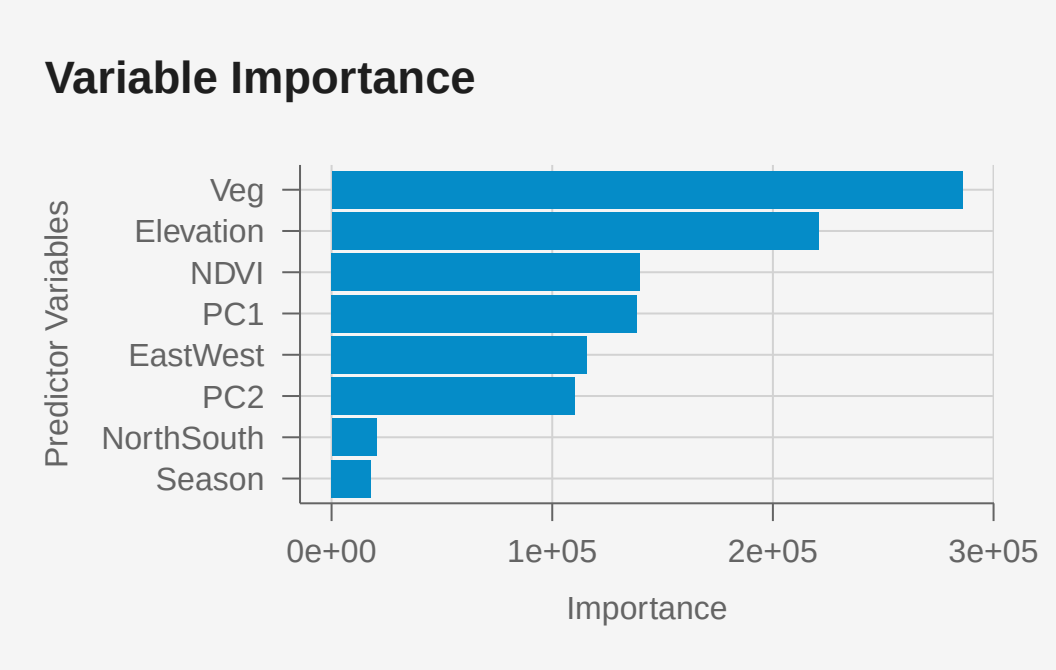
Assumptions:

Random Forests make no assumptions about the functional form of the predictors and outcomes, making them highly flexible for complex datasets. They can handle multicollinearity, non-linear relationships, and high-dimensional predictors effectively.

###Implementation: The Random Forest model will include all predictors used in the multinomial model. Spatial aggregates (e.g., neighborhood means of NDVI) will be introduced as additional features to capture local spatial dependencies.

Performance Evaluation:

The Random Forest model will be evaluated using out-of-bag (OOB) error, accuracy, precision, recall, and F1-score. Variable importance plots will help identify the most influential predictors.



Interpretation:

Random Forests provide insights into the relative importance of predictors, highlighting which spectral, temporal, and spatial features contribute most to land cover classification. #? For example, NDVI and PCA-derived features are expected to rank highly, given their ecological relevance.

Appropriateness:

Random Forests are well-suited for this problem, as they can handle overlapping class boundaries and complex interactions without requiring prior assumptions about the data. However, they lack interpretability compared to linear models.

Model Comparison and Validation

To compare the performance of the models (logistic regression, Random Forests, and sinusoidal extensions), we will use cross-validation. Model evaluation metrics, such as accuracy, precision, recall, and F1-score, will guide the selection of the best-performing model.

Section Conclusion

This section details the implementation of multiple classification models, including logistic regression (linear baseline), Random Forests (non-linear, ensemble), and sinusoidal regression (temporal modeling). Each model addresses specific challenges identified in the data exploration phase, such as overlapping class boundaries, temporal periodicity, and spatial dependencies. The next section will visualize and interpret the results, providing a comparative analysis of model performance and insights into the relationships between predictors and land cover classes.

Visualization and interpretation of the results

Create visualizations of the results when appropriate, focusing on visualizations that

- help describe aspects of the results that have real-world interpretation
- help the reader understand how the model addresses the problem you are studying.

Visualizations are one of the most powerful ways to communicate information to the reader, so it is important to spend time producing clear, descriptive, eye-catching visualizations.

Discuss the results of the model or models you chose, and describe how they are related to the problem statement or question that you were trying to answer in the project.

If you have built multiple models or types of analysis, compare the measures of performance and the ease of interpretability across models or types of analysis, stating which model or models performed best, and which model or models were most interpretable. Finally, decide which model or type of analysis is best for your particular problem based on some combination of performance and interpretability.

Conclusions and Recommendations

One or two paragraphs stating conclusions, recommendations, and ideas for future work and improvements.

References

List any references for your data source(s), other work or results, etc.