S&DS 425/625 Report

Elder Veliz & Emmanuelle Brindamour

2024-12-14

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. Predictors include spectral bands, NDVI, temporal features, and potential spatial aggregates derived from neighboring pixels.

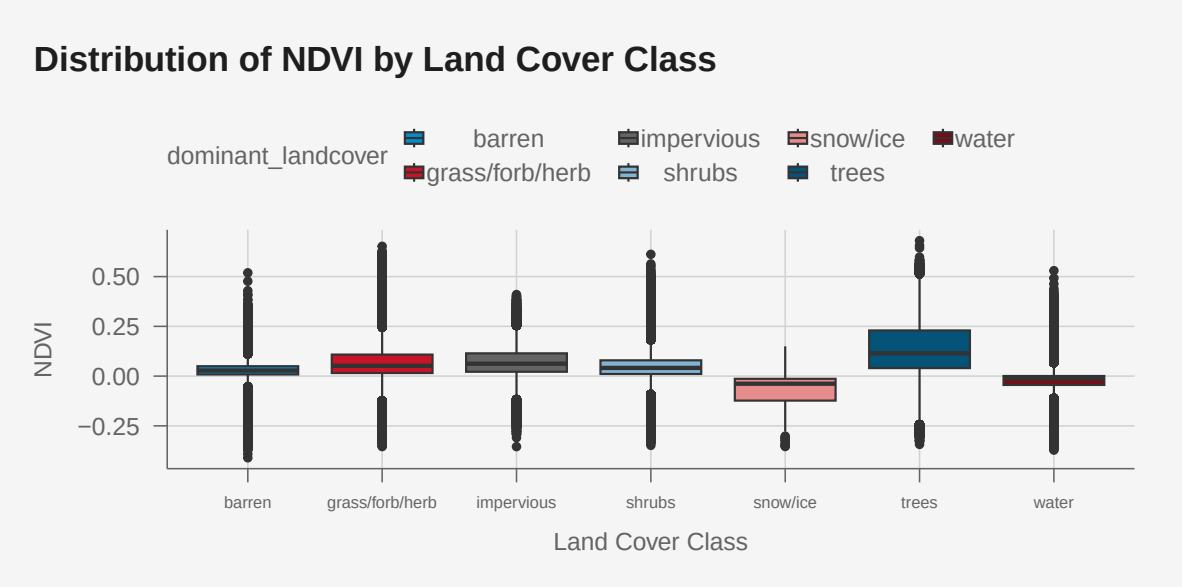
- `PlotID`: A unique identifier for each observation.
- `Month` and `Year`: Temporal features indicating the month and year of the observation.
- `SR_B1`: The "blue" spectral band captures visible blue light and is sensitive to water bodies and atmospheric aerosols.
- `SR_B2`: The "green" spectral band captures visible green light and is sensitive to vegetation health and possibly land-water boundaries.
- `SR_B3`: The "red" spectral band captures visible red light and is usually used to capture chlorophyll absorption in vegetation.
- `SR_B4`: The "near-infrared" spectral band captures near-infrared light and is sensitive to vegetation density.
- `SR_B5`: The "shortwave infrared 1" spectral band captures shortwave infrared light and is sensitive to moisture content in soil and vegetation.
- `SR_B7`: The "shortwave infrared 2" spectral band captures shortwave infrared light and differentiates vegetation stress, soil properties, and geology.
- `NDVI`: The Normalized Difference Vegetation Index 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`: The categorical outcome variable representing the dominant land cover class at the pixel level.
- `Dominant_LandUse`: The categorical outcome variable representing the dominant land use class at the pixel level.
- `Season`: The season (Winter, Spring, Summer, Fall) of observation.

- `Veg`: A binary variable indicating whether the pixel is vegetated (1) or non-vegetated (0).
- `EastWest`: A binary variable indicating whether the pixel is located east (1) or west (0) of the median longitude.
- `NorthSouth`: A binary variable indicating whether the pixel is located north (1) or south (0) of the median latitude.
- `Elevation_meters`: The elevation in meters above sea level at the pixel location.
- `Elevation_Bins`: A categorical variable representing the elevation range of the pixel.

Exploration of Predictors

Spectral Bands and NDVI:

Visualizing the spectral band values across different land cover types reveals distinct patterns. For example, vegetative areas are expected to exhibit higher NDVI values, while urban or barren areas may show lower values. 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

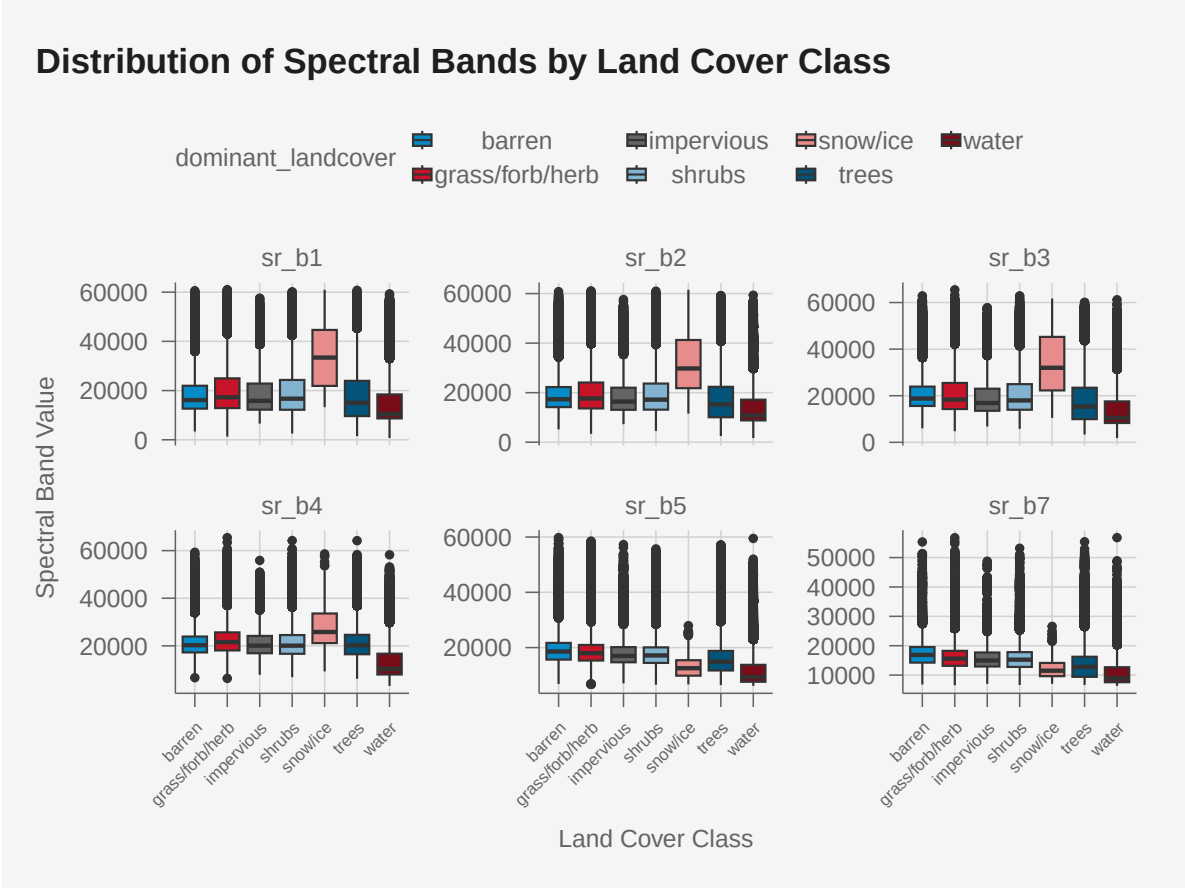
consistently negative NDVI values, reflecting the spectral signature of water, which absorbs near-infrared light and reflects visible wavelengths.

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. This foundation will guide the development of robust predictive models in subsequent sections.



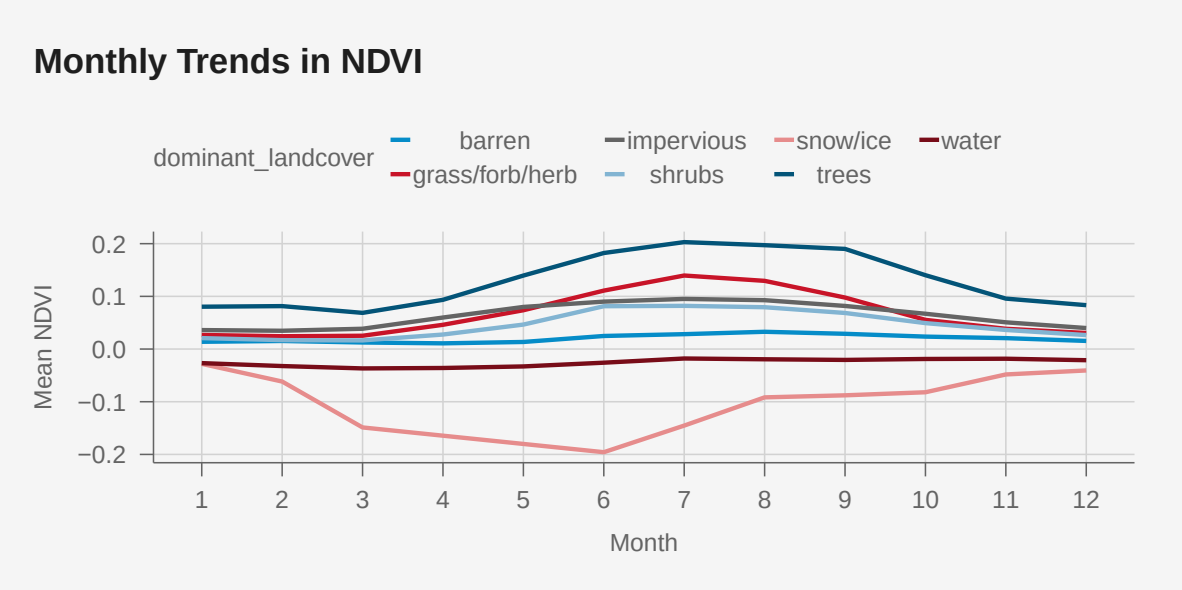
The analysis of spectral bands across land cover classes reveals key patterns that inform their utility for classification. For Bands 1, 2, and 3, which correspond to the visible blue, green, and red portions of the spectrum, the distribution of values remains relatively constant across most land cover classes. This limited variability suggests that these bands may lack the discriminatory power needed to differentiate between diverse land cover types. The strong collinearity among these bands further reinforces the need for dimensionality reduction techniques, such as Principal Component Analysis (PCA), to consolidate their information into fewer predictors without sacrificing interpretability.

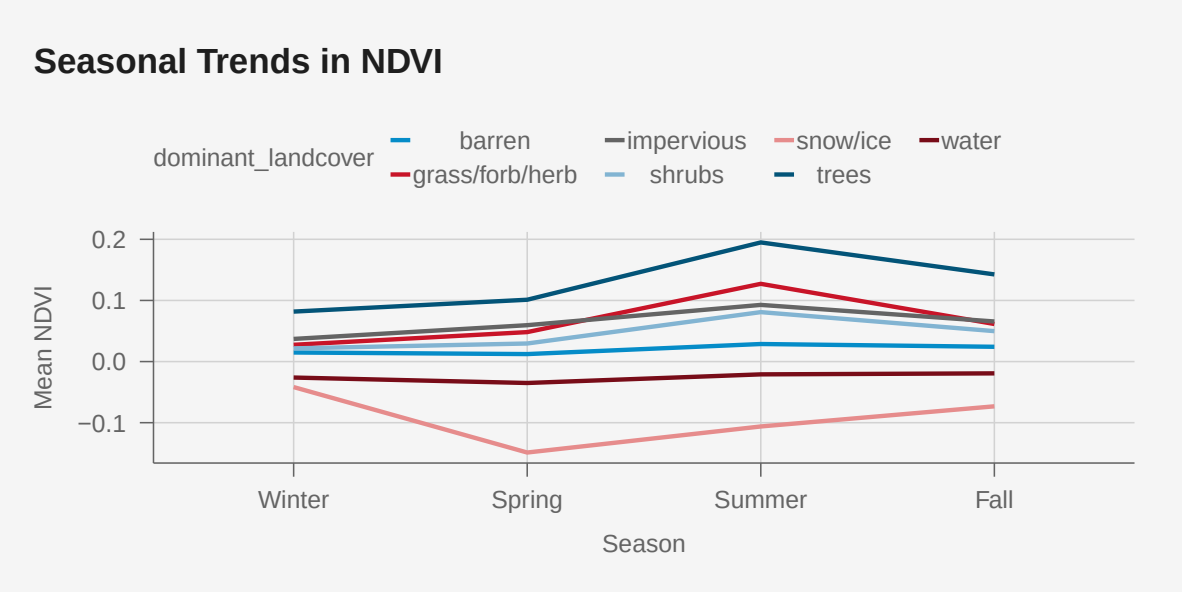
For Bands 5 and 7, which capture reflectance in the shortwave infrared (SWIR) region, the distributions show a similar pattern of consistency across land cover types, though with slightly more variability compared to the visible bands. These bands are known to be sensitive to moisture content and soil 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's efficiency and predictive accuracy.

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. Seasonal variations are expected, particularly in NDVI and vegetation-sensitive bands, which can be visualized 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.

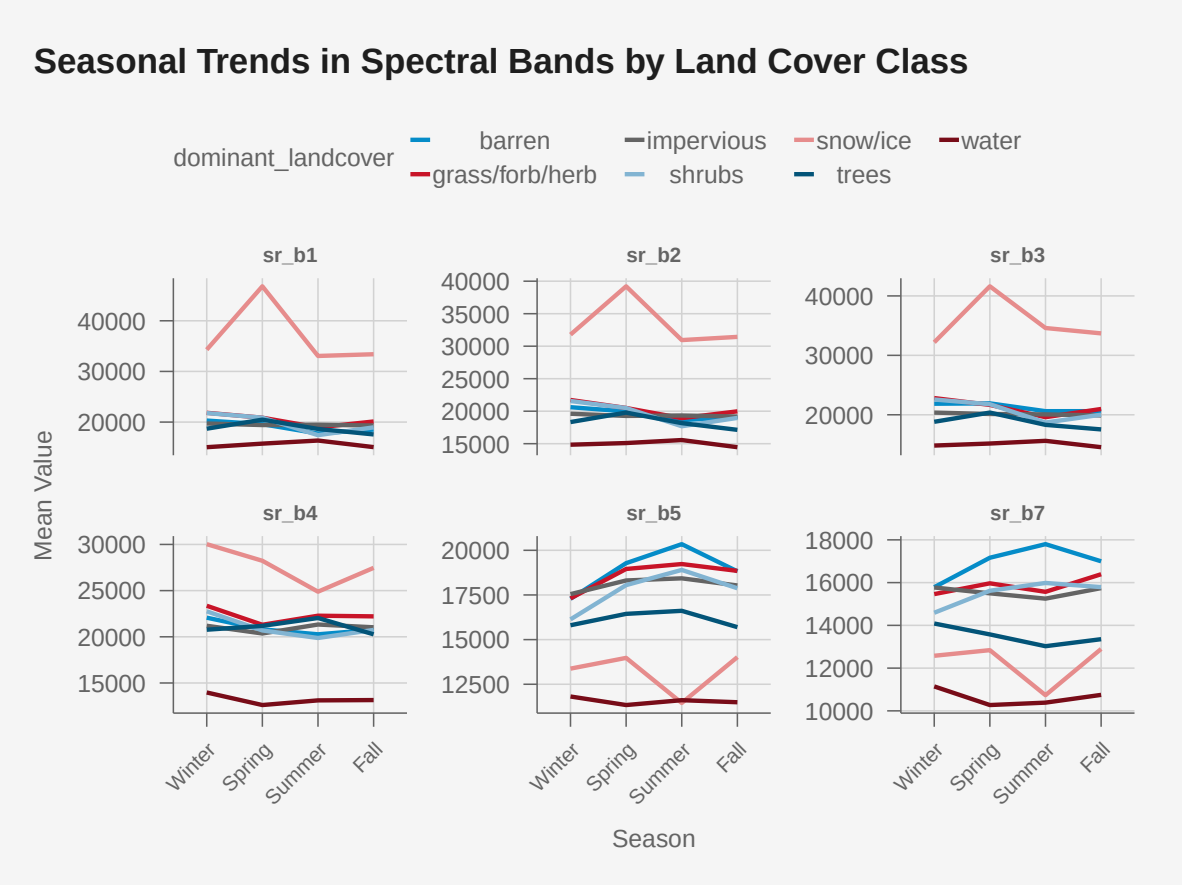
"Snow/Ice", on the other hand, shows a very prominent decline in NDVI in the late-Winter and Spring months (February to June), likely due to the melting of snow and ice, which reduces the reflectance in near-infrared bands. Conversely, "Water" maintains relatively stable NDVI values throughout the year, consistent with its spectral signature that is less influenced by seasonal changes. Classes like "Barren" and "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" 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: $NDVI(t) = A \cdot sin(2\pi \cdot f \cdot t + \phi) + B \cdot cos(2\pi \cdot f \cdot t + \phi)$, where A and B are the amplitudes, f is the frequency (e.g., one cycle per year), t is time (month), 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 learn-

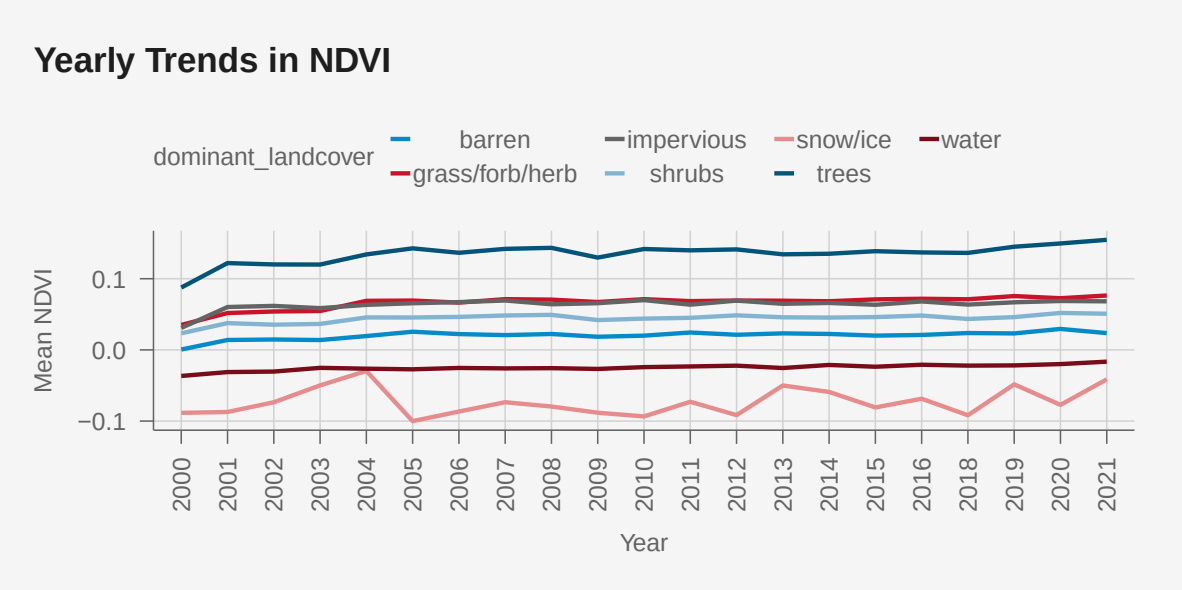
ing classifiers can improve their ability to distinguish between classes with overlapping NDVI ranges. For example, while "Grass/Forb/Herb" and "Shrubs" may have similar NDVI distributions at specific times of the year, their seasonal trajectories differ, providing an additional dimension for separation.

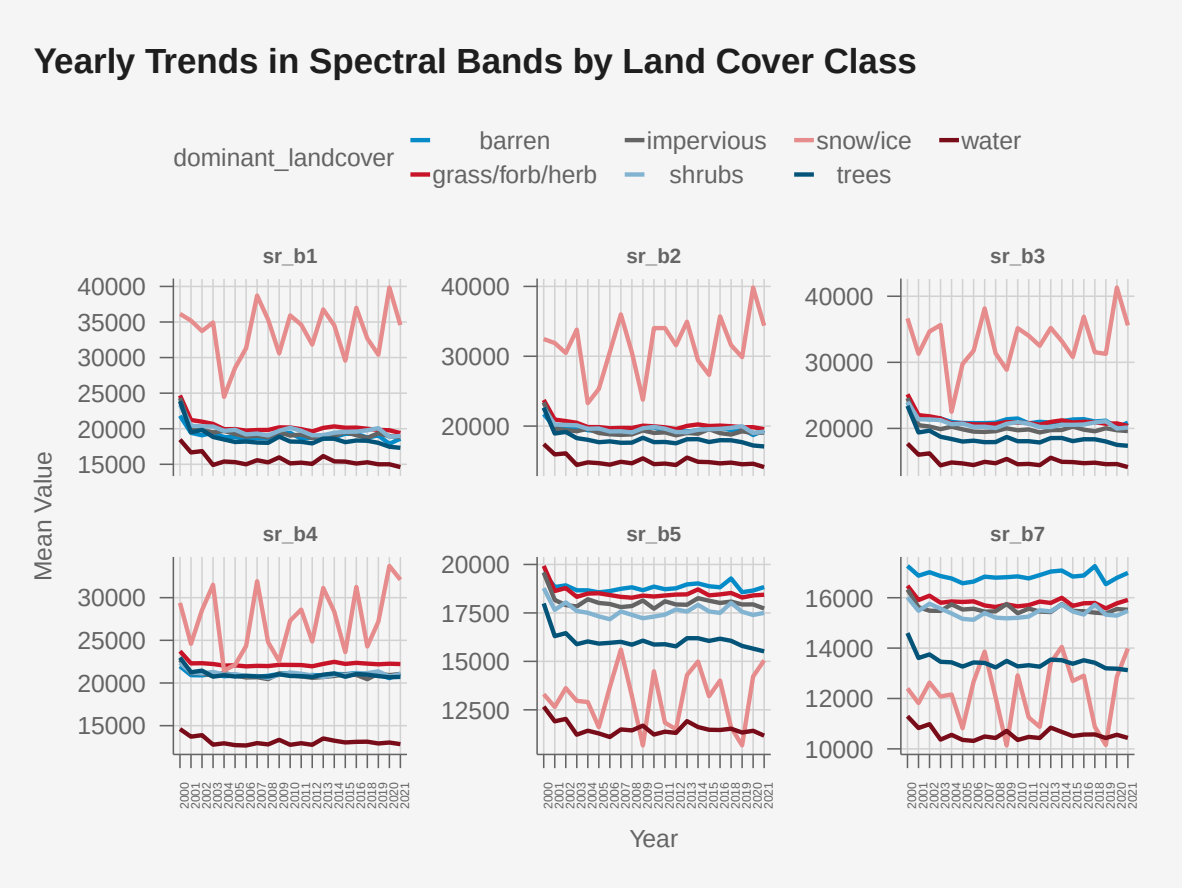
Overall, the observed periodic patterns in NDVI justify the inclusion of temporal features in the modeling framework. These features not only enhance the interpretability of the models but also provide a biologically meaningful basis for improving classification accuracy. Next steps will involve integrating these temporal dynamics with spatial and spectral features to construct a robust predictive framework.



The seasonal trends in spectral bands across land cover classes reveal notable 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 and Summer seasons, 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.

Bands 5 and 7 (representing shortwave infrared) display more nuanced seasonal trends, particularly for vegetative classes like "Trees" and "Grass/Forb/Herb." These classes show clear increases in reflectance during summer, aligning with the seasonal growth of vegetation and moisture-related changes that SWIR bands are sensitive to. Notably, the "Barren" and "Impervious" classes remain relatively stable across all seasons in these bands, reflecting the consistency of their 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 declining 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 "Trees" 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. These findings highlight the importance of incorporating nuanced temporal features into land cover classification models. 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.

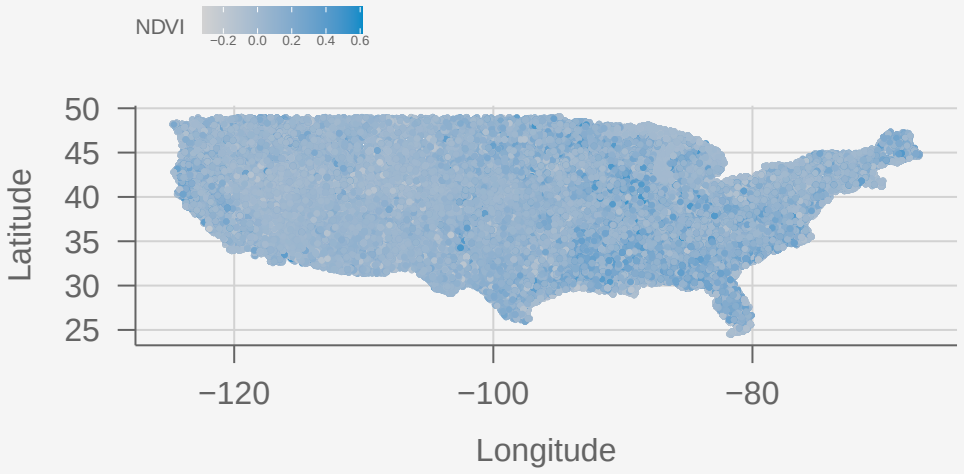
For the remaining land cover classes, visible bands (B1, B2, B3) show limited variability, reaffirming their redundancy in distinguishing between vegetative and non-vegetative classes. However, Bands 5 (SWIR1) and 7 (SWIR2) provide more nuanced differences, particularly for classes like "Trees" and "Grass/Forb/Herb,"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

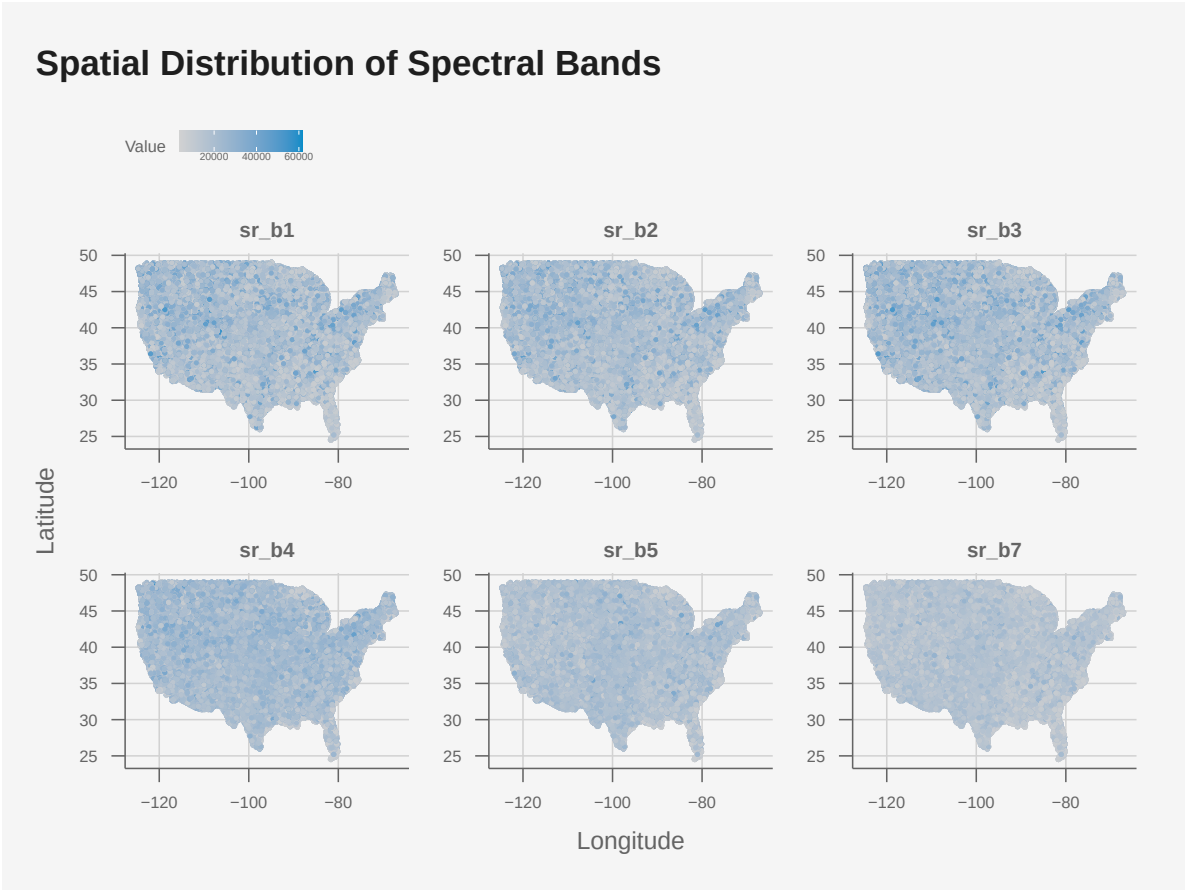
Exploration of Spatial Patterns

Spatial Relationships:

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

Spatial Distribution of NDVI





The spatial distribution of NDVI and spectral band values across the United States reveals significant patterns that align with ecological and geographic expectations. The visualization 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 characterized by forested landscapes and favorable climatic conditions that support high vegetation density.

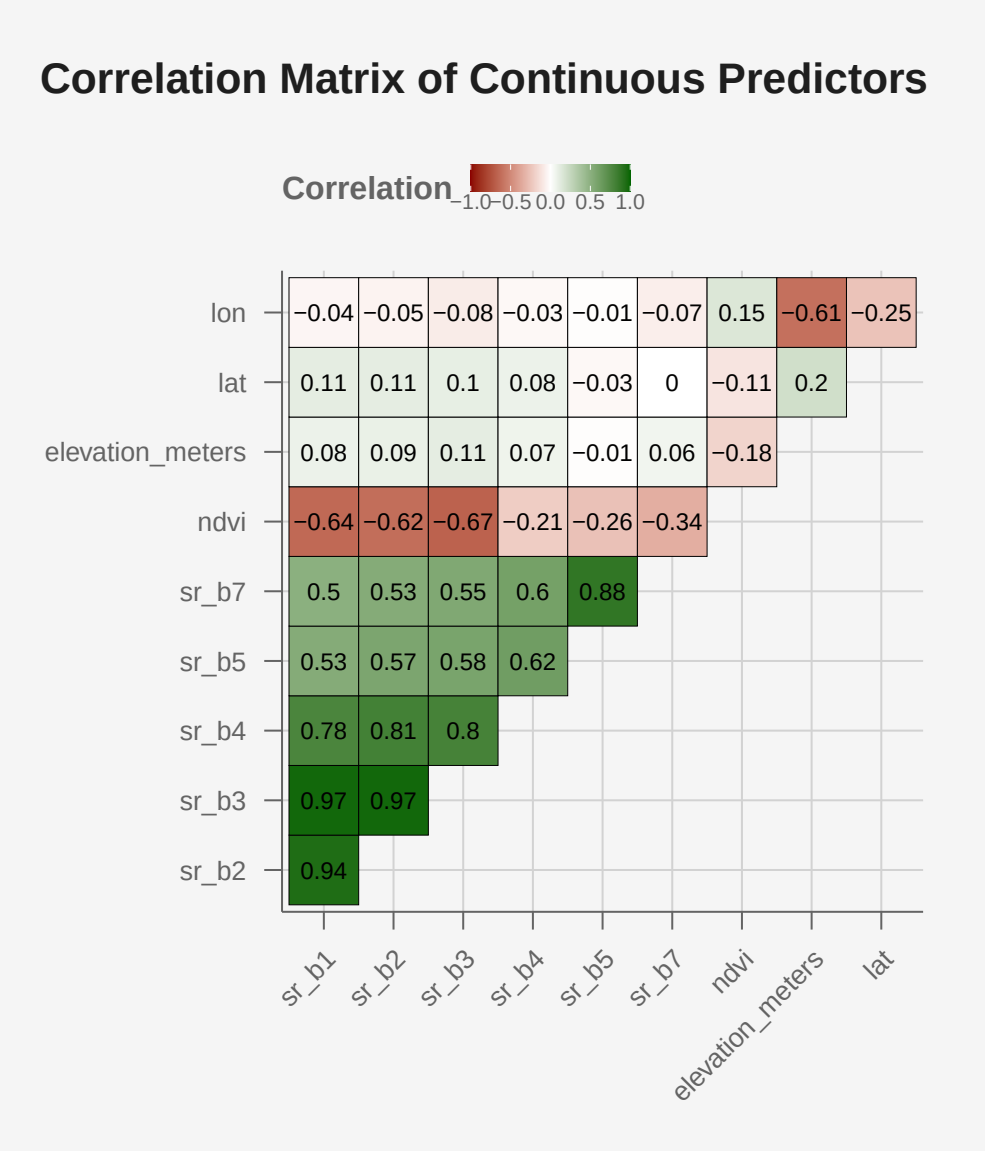
In contrast, areas with low NDVI values are predominantly located in the arid western and southwestern United States. These regions, dominated by deserts, sparse vegetation, and barren land, naturally exhibit reduced photosynthetic activity. The stability and clustering of NDVI in these regions emphasize 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 not only enhances model performance but also aligns with the underlying ecological processes that govern land cover patterns. Modeling efforts will leverage these spatial relationships by integrating neighborhood-level statistics and testing their impact on classification accuracy.

Preliminary Correlation Analysis

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



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. 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. Dimensionality reduction techniques, like PCA could help mitigate this high degree of interdependence.

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 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.

Assumptions for Modeling

From the data exploration, several assumptions emerge: 1. Predictor-Outcome Relationships: • Spectral bands and NDVI are assumed to have distinct distributions across land cover classes. • Temporal features (e.g., month) contribute to variations in spectral responses, particularly in vegetated areas. 2. Spatial Dependencies: • The inclusion of spatial aggregates accounts for the influence of neighboring pixels, improving classification accuracy. 3. Seasonal Patterns: • NDVI and other vegetation-sensitive bands exhibit periodic trends, justifying sinusoidal modeling.

Justifying Model Choices

1. Linear Models: NDVI trends suggest linear separability in some cases, making Ordinary Least Squares (OLS) a suitable baseline.
2. Random Forests: To handle potential overlaps in spectral responses between classes, Random Forests are well-suited for non-linear decision boundaries.
3. Temporal Models: Incorporating sinusoidal features into logistic regression or Random Forest classifiers accounts for seasonal oscillations, enhancing model performance.
4. Spatial-Temporal Extensions: Models that combine spatial aggregates and temporal features provide a comprehensive framework for land cover classification.

Section Conclusion

This exploratory analysis establishes the relationships between predictors and the outcome variable, supports the inclusion of temporal and spatial features, and provides justification 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/Analysis

Describe regression or classification model(s) used, or the analysis that was performed. For each regression or classification model, discuss

- any assumptions that are made
- the observation, the predictors, and the outcome (aka the rows of X, the columns of X, and y)
- what model you are using, and write out the model
- what the coefficients mean (when applicable) and how this is related to your problem
- appropriate measures of the performance of the model, such as measures of fit and predictive ability
- whether or not you think the model is appropriate for this kind of data, and why, and
- how easy/hard it is to interpret the results and explain them to either a technical or non-technical audience.

For other kinds of analysis, what you give is highly dependent on the type of analysis. But in general, talk about assumptions, if they are appropriate, how they might not be appropriate, and why you chose this type of analysis.

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.