# Remote Sensing for Forest Recovery: Proposal

Benjamin Frizzell Zanan Pech Mavis Wong Hui Tang

2025-05-05

## Table of contents

## 1 Abstract

Monitoring afforestation progress across hundreds of sites is a significant challenge. This project explores using site-level data and satellite-derived vegetation indices from Canada's 2 Billion Trees program to build machine learning models that predict tree survival over seven years. We will train models ranging from logistic regression to more advanced techniques like random forests, gradient boosting, and deep learning (RNNs, LSTMs) to capture temporal patterns. The goal is to evaluate predictive features and modeling strategies that enable scalable, cost-effective monitoring of afforestation efforts.

## 2 Introduction

Afforestation is crucial for combating climate change and supporting biodiversity. Trees help purify the atmosphere and provide food, nutrients, and habitats for countless species (Government of Canada 2023).

To advance these goals, the Canadian government launched the **2 Billion Trees** program, aiming to plant two billion trees nationwide by 2031 with support for provinces and territories (Natural Resources Canada 2021).

Monitoring survival at scale is a major challenge. While remote sensing enables broad environmental monitoring, detecting small or newly planted trees remains difficult due to limited canopy coverage and weak spectral signals (University of British Columbia Master of Data Science Program 2025).

This study aim to address two key questions: - Can remote sensing reduce the need for physical site visits? - Which modelling approach is most effective, and how soon after planting can survival be reliably predicted?

# 3 Data Description

# 4 Data Description

The dataset used in this study includes field-measured survival rates of afforested sites collected by Forest Ontario, as well as satellite data products from the Harmonized Landsat Sentinel-2 (HLS) project.

## 4.1 Site Features

| | Description |
|---|---|
| ID | Site ID |
| PixelID | Pixel ID |
| Area_ha | Area of Site |
| Season | Planting Year |
| PlantDt | Planting Date |
| prevUse | Previous Land Use of Site |
| Planted | Number of Trees Planted |
| SpcsCmp | Species Composition of Site |
| Type | Species Type of Site |
| SrvvR_1, ..., SrvvR_7 | Field Measured Survival Rate at Year 1-7 (Target) |
| AsssD_1, ..., AssD_7 | Date of Field Survival Rate Measurement |
| NmbrPlO | Number of Trees Originally Planted |
| NmbrPlR | Number of Trees Replanted |
| NmbrPlT | Total Number of Trees Planted |
| ImgDate | Image Date of the Remote Sensing Data |
| Year | Image Year of Remote Sensing Data |
| DOY | Day of Year of the Remote Sensing Data |

## 4.2 Spectral Indicies

| Index | Description |
|---|---|
| NDVI | Normalized Difference Vegetation Index |
| SAVI | Soil-Adjusted Vegetation Index |
| MSAVI | Modified Soil-Adjusted Vegetation Index |
| EVI | Enhanced Vegetation Index |
| EVI2 | Two-band Enhanced Vegetation Index |
| NDWI | Normalized Difference Water Index |

| Index | Description |
|---|---|
| NBR | Normalized Burn Ratio |
| TCB | Tasseled Cap Brightness |
| TCG | Tasseled Cap Greenness |
| TCW | Tasseled Cap Wetness |

### 4.3 Out-of-range Values

During EDA, we noticed that there are out-of-range values in the vegetation indicies and survival rates. With the exception of TCB, TCW and TCG, the vegetation indicies should range between -1 to 1(2018, n.d.a, 2025, 2023, n.d.b; Sinergise, n.d.; Mondal 2011). The survival rates should not exceed 100. We will be removing these out-of-range records from the dataset.

### 4.4 Previous Land Use and Species Composition

Class imbalance was observed in 'SpcsCmp' column. Given that there are over 300 categories in 'SpcsCmp', it would not be practical for us to use this column as a predictor for our model.

Severe class imbalance are also observed in the 'prevUse' column, as such, we would also not be using 'prevUse' as a predictor for our model.

### 4.5 Species Type

From 'SpcsCmp', we are able to impute the species type. Sites are classified as Conifer(Deciduous) if the proportion of softwood(hardwood) species >= 80%, else, they would be classified as mixed.

Figure 1 shows that survival rates and spectral indices differs by species type, suggesting that 'Type' would be a viable feature to use as predictor for our model.

### 4.6 Trends and Seasonality

Clear seasonality was observed in the spectral indices, where the signals peaked during the summer and dropped during winter months (Figure 2). This would be something that we need to address in our model.

Figure 3 also shows a positive relationship between vegetation indicies and tree age. Minimal change in vegetation indicies was observed between age 1 to age 4. Indicating potential difficulties in survival rate prediction for earlier years.

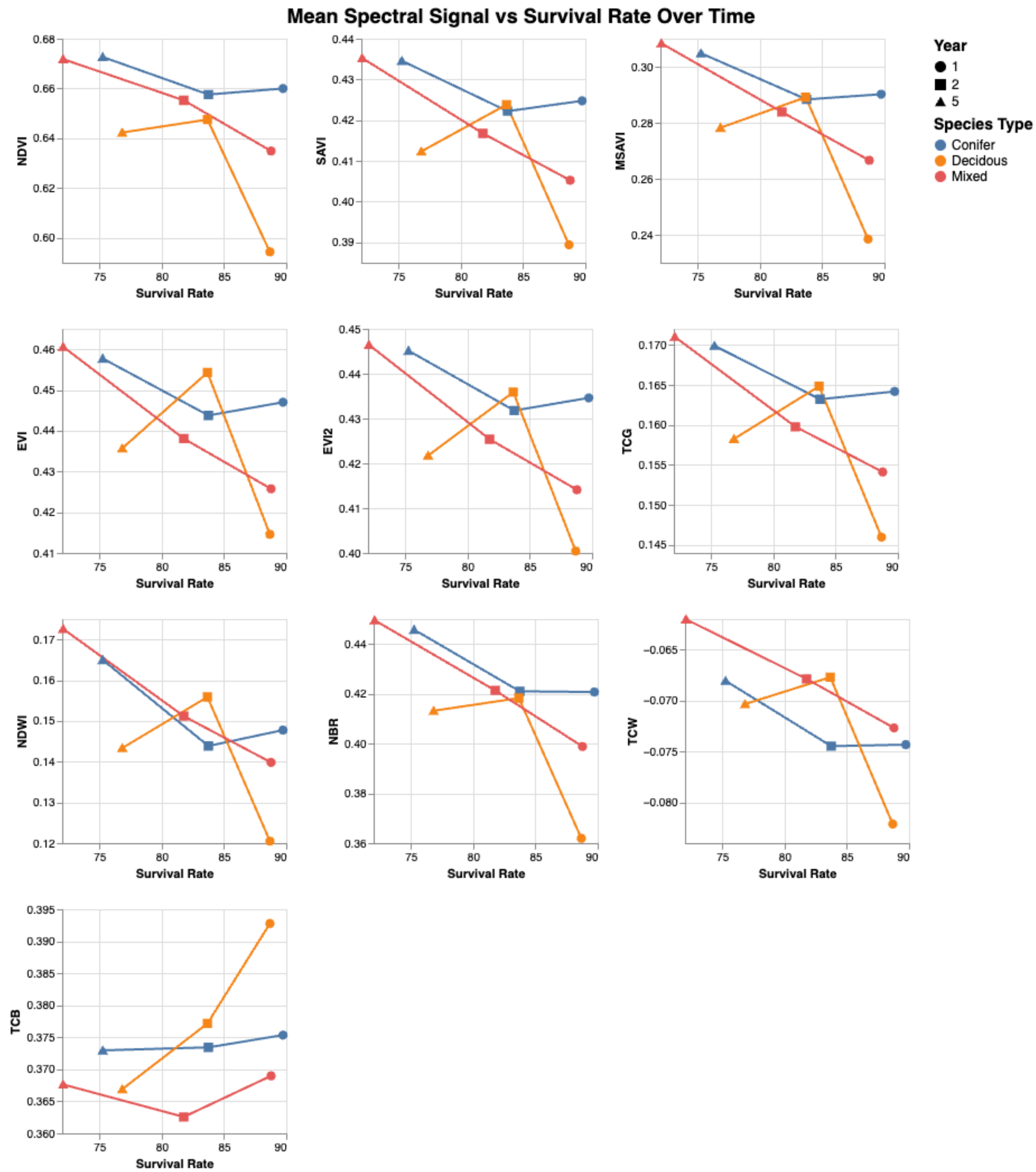


Figure 1: Plot showing mean survival rate and vegetation index signals for different species type in Year 1, 2 and 5. There is significant difference in the relationship between survival rate and VI signals for different species type. Conifers has a smaller signal response to change in survival rate. Deciduous shows the strongest response in the first two years. Mixed type shows a linear relationship between survival rate the spectral signal.

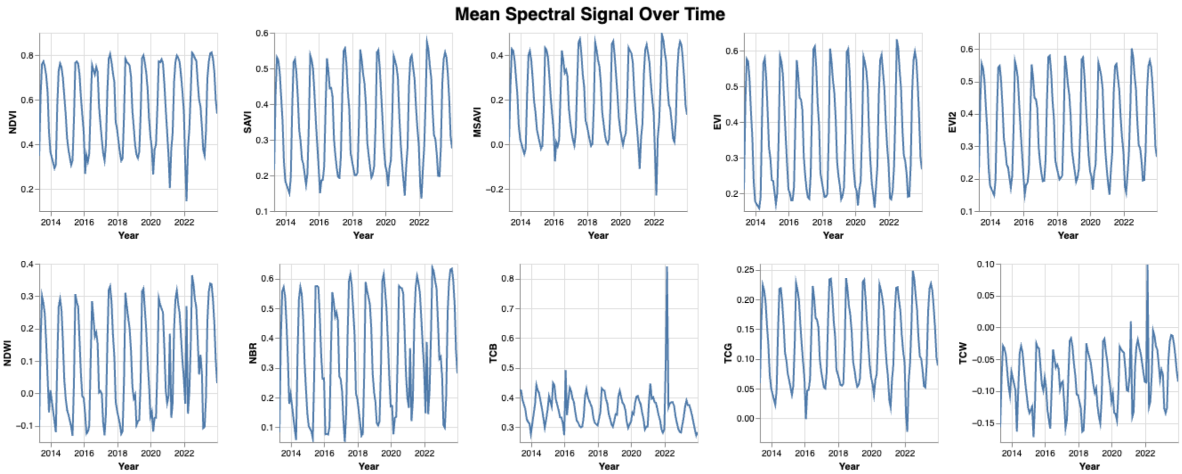


Figure 2: Plot showing seasonality in vegetation indices.

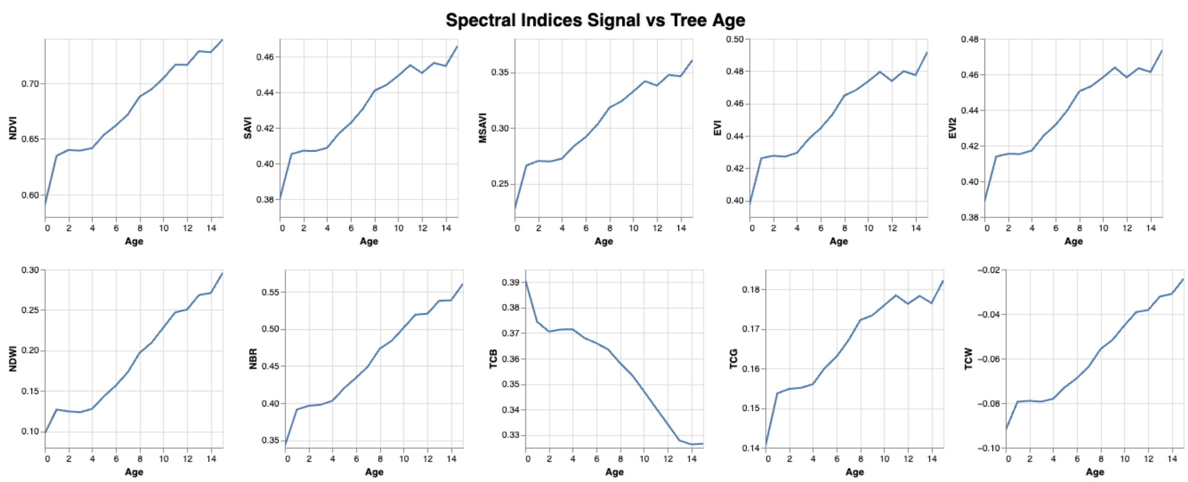


Figure 3: Plot showing mean spectral signal by tree age. With the exception of TCB, spectral signal increases as tree matures. A negative relationship was observed for TCB due to decreasing surface brightness as canopy cover increases.

### 4.7 Collinearity

From Figure 4, we observed strong collinearity between vegetation indices except for TCB. The strongest correlation between survival rate and vegetation indicies is observed in Year 7.

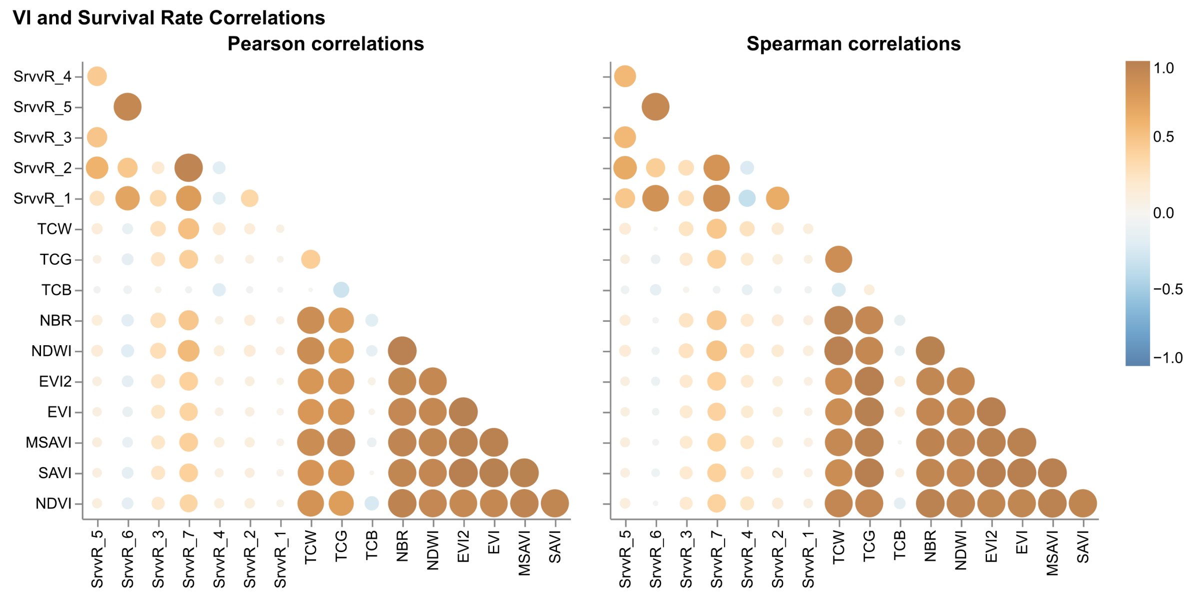


Figure 4: Correlation plot showing strong collinearity between vegetation indicies.

## 5 Data Engineering

### 5.1 Train-Test Splitting

Splitting the dataset into training and testing subsets is necessary to prevent data leakage, but requires nuance due to it's heirarchical structure. We perform the train-test split by **unique sites** to ensure pixels and time step records for a particular site appear in only one of the two subsets.

### 5.2 Missing Data

We find excessive missingness in the columns `PlantDt`, `Type`, `NmbrPIO`, `NmbrPIR`, and `NmbrPIT`. `PlantDt`, `NmbrPIO`, `NmbrPIR`, and `NmbrPIT` relate to sites where replanting has occured, and can be removed as they are outside of the project scope, and `Type` can be fully imputed via string processing on the `SpcsCmp` column. There is a direct correspondence between missingness in

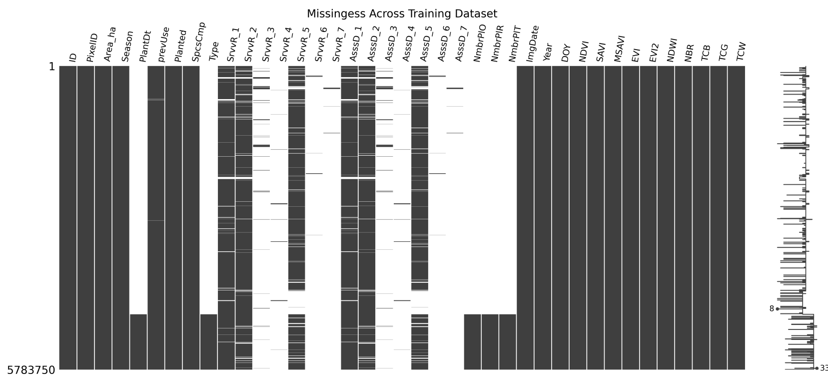


Figure 5: A plot visualizing missing record patterns across the training dataset. Grey-coloured records indicate rows that have recorded values.

survivial rate and assesment time, allowing for easier tracking of temporal dependence across survival records.

## 5.3 Feature Selection and Engineering

As mentioned in Section 3, strong collinearity between vegetation indices indicates a need for feature selection. We will begin with **Recursive Feature Elimination**, due to its compatbility with nonlinear ML models (Pedregosa et al. 2011). We also propose the use of **Bayesian Model Averaging**, as it is suitable for an analysis of multiple competing models (Hoeting et al. 1998). Domain knowlege of the vegetation indices as given by Zeng et al. (2022) will also be leveraged in this process. Very little feature engineering will be performed, but we aim experiment with tree density (number of trees per unit area) as a predictor.

## 5.4 Data Pivoting

Most machine learning models require the input data to contain just one target column. We will pivot the seven target columns into one, keeping track of temporalty using column names and assesment dates. We will then remove rows with missing survival rates, and with mismatching assesment and imaging dates.

## 5.5 Conversion to Classifier Problem

Since the survival rates are given as percentage proportions, they will be converted to binary classes to simplify the analysis and emphasize high-risk sites. Since most survival rates are within 70% - 100%, usefulness alongside class imbalance must be considered when deciding on a suitable threshold.

# 6 Modelling Techniques

## 6.1 Logistic Regression

Logistic regression will serve as a baseline to contrast performance against other, more complex models. It may also provide insight into feature importance through interpretable coefficients.

## 6.2 Aggregated Models: Random Forest and Gradient Boosting

Studies by Bergmüller and Vanderwel (2022) suggest that aggregated tree models may effectively utitlize vegetation indices to predict tree mortality. Random Forests and GBMs complement each other well, as the former produces many 'overfit' trees in tandem to produce more accurate responses, whereas the ladder iteratively produces 'underfit' trees which correct on mistakes made by the previous (Pedregosa et al. 2011).

## 6.3 Sequential Deep Learning Models: LSTMs, GRUs, etc.

If time permits, we may consider implementing sequential RNN-based deep learning models such as LSTMs and/or GRUs. These models have the benefit of capturing non-linear structure and may exploit changes in vegetation indices across time (Paszke et al. 2019).

# 7 Success Criteria

The success of this project will be evaluated based on the usefulness of the selected predictors and the performance of key evaluation metrics, including accuracy, log loss, F1 score, and ROC and PR curves. If our predictors are shown to contribute meaningfully to modeling tree survival rates, we aim to achieve at least 60% accuracy in correct predictions. Accuracy will serve as the primary metric for communicating results to general audiences—particularly government officials without data science expertise—as it provides a straightforward summary of model performance. However, given the imbalanced nature of our dataset, metrics such as log loss, F1 score, and the ROC and PR curves will be especially important for technical stakeholders and researchers within the federal government who are equipped to interpret more nuanced performance indicators.

# 8 Timeline

| Date & Time | Deliverable | Description |
|---|---|---|
| **May 2, 2025** | Proposal Presentations | Group oral presentations |
| **May 6, 2025 12:00** | Proposal Report – Draft to Mentor | Ungraded draft |
| **May 9, 2025 17:00** | Final Proposal Report | Final version to partner & mentor |
| **June 9, 2025 16:00** | Data Product – Runnable Draft | Draft pipeline & code to mentor |
| **June 12–13, 2025 9:00–16:30** | Final Presentation | Group presentation of modelling approach & key results |
| **June 25, 2025 12:00** | Final Data Product & Report | Final pipeline, data product, and technical report |

## 9 Conclusion

This project addresses the challenge of monitoring tree survival across hundreds of afforestation sites in Canada's 2 Billion Trees program. By leveraging site-level data and satellite-derived spectral indices, we evaluate machine learning and deep learning models to predict survival over time. Our approach spans from interpretable models like logistic regression to advanced methods such as random forests, gradient boosting, and recurrent neural networks. Identifying key predictors is essential for building effective, scalable tools to support national afforestation efforts.

## References


2018. *USGS*. https://www.usgs.gov/special-topics/remote-sensing-phenology/science/ndvi-foundation-remote-sensing-phenology#:~:text=Although%20there%20are%20several%20vegetation,example%2C%200.1%20or%20less).

———. 2023. *EOS Data Analytics*. https://eos.com/make-an-analysis/ndwi/.

———. 2025. *EOS Data Analytics*. https://eos.com/make-an-analysis/msavi/.

———. n.d.a. https://pro.arcgis.com/en/pro-app/latest/arcpy/spatial-analyst/savi.htm#:~:text=The%20Soil%2DAdjusted%20Vegetation%20Index,values%20between%20%2D1.0%20and%201.0.&text=L%E2%80%94The%20amount%20of%20green%20vegetation%20cover.

———. n.d.b. *UP42 Documentation*. https://docs.up42.com/help/spectral-indexes/nbr#:~:text=NBR%20ranges%20between%20%2D1%20and,have%20values%20close%20to%20zero.

Bergmüller, Kai O, and Mark C Vanderwel. 2022. "Predicting Tree Mortality Using Spectral Indices Derived from Multispectral UAV Imagery." *Remote Sensing* 14 (9): 2195.

Government of Canada. 2023. "2 Billion Trees Commitment." https://www.canada.ca/en/campaign/2-billion-trees.html.

Hoeting, Jennifer A, David Madigan, Adrian E Raftery, and Chris T Volinsky. 1998. "Bayesian Model Averaging." In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models*, 335:77–83. Citeseer.

Mondal, Pinki. 2011. "Quantifying Surface Gradients with a 2-Band Enhanced Vegetation Index (EVI2)." *Ecological Indicators* 11 (3): 918–24.

Natural Resources Canada. 2021. "2 Billion Trees Program." https://www.canada.ca/en/campaign/2-billion-trees/2-billion-trees-program.html.

Paszke, Adam, Sam Gross, Francesco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Sinergise, Sentinel-Hub By. n.d. "EVI (Enhanced Vegetation Index)." *Sentinel Hub Custom Scripts*. https://custom-scripts.sentinel-hub.com/custom-scripts/sentinel-2/evi/#:~:text=Values%20description%3A%20The%20range%20of,generally%20around%200.20%20to%200.80.

University of British Columbia Master of Data Science Program. 2025. "Remote Sensing for Forest Recovery." https://pages.github.ubc.ca/mds-2024-25/DSCI_591_capstone-proj_students/proposals/Remote_Sensing_for_Forest_Recovery.html.

Zeng, Yelu, Dalei Hao, Alfredo Huete, Benjamin Dechant, Joe Berry, Jing M Chen, Joanna Joiner, et al. 2022. "Optical Vegetation Indices for Monitoring Terrestrial Ecosystems Globally." *Nature Reviews Earth & Environment* 3 (7): 477–93.