Executive Summary

Introduction and Context

Accurate land cover classification is essential for effective environmental management and decision-making. The reliable differentiation of vegetation types, water bodies, barren lands, and other categories supports strategies to address climate change, deforestation, habitat loss, and urban expansion. Advances in remote sensing have greatly expanded the potential of satellite imagery for this purpose, yet challenges remain. Common issues include significant class imbalances, overlapping class boundaries, and complex, non-linear relationships among various predictors. To address these issues, this study systematically evaluated four modeling approaches, each designed to integrate spatial and temporal dimensions more effectively.

Data and Predictors

The dataset employed in this study integrates information from the U.S. Geological Survey's LCMAP program and Landsat-7 imagery, resulting in a rich source of spectral and geospatial data. Predictors include multiple spectral bands (B1-B7), the Normalized Difference Vegetation Index (NDVI), elevation measurements, geographic coordinates, and temporal attributes such as month, year, and season. This comprehensive dataset captures the heterogeneity of landscapes across the United States and provides a foundation for modeling complex ecological interactions. However, pronounced class imbalances—where dominant categories (e.g., trees and grass/forb/herbaceous cover) vastly outnumber rare classes (such as snow/ice)—present a persistent challenge for classification accuracy.

Modeling Approaches

Four approaches were considered. Multinomial logistic regression served as the baseline, offering a linear and interpretable model but struggling to accurately classify minority and spectrally similar classes. Classification and Regression Trees (CART) introduced non-linearity, improving upon simple linear models but still faltering against class imbalance and subtle spectral overlaps. Incorporating sinusoidal transformations into logistic regression to model seasonal variations yielded only negligible improvements, suggesting that simple parametric solutions are insufficient in complex contexts.

Key Results and Performance

Random forests, an ensemble-based approach, emerged as the most robust solution. By aggregating multiple decision trees and exploring non-linear patterns, random forests achieved a macro-averaged F1-score of 0.812 and a balanced accuracy of 0.877—significantly surpassing the performance of earlier models. This approach managed class imbalances more effectively, improving prediction for both dominant categories and rare classes such as snow/ice. Variable importance measures indicated that NDVI, elevation, and vegetation indicators were the most influential predictors. Principal components derived from correlated spectral bands further enhanced performance by reducing the noise introduced by multicollinearity.

Implications for Practice

While the random forest model requires greater computational resources and does not offer the transparent coefficient-based interpretability of simpler models, its accuracy and adaptability provide compelling advantages for large-scale, real-world applications. Environmental managers, policymakers, and researchers can rely on random forests to glean more reliable land cover insights, ultimately supporting better-informed decisions about resource allocation, conservation measures, and urban planning.

Recommendations and Future Directions

Future efforts should target improved handling of class imbalances, potentially through advanced sampling strategies or cost-sensitive learning, and the integration of additional data sources—such as climate or soil information—to further refine predictions. Hybrid modeling approaches that combine the interpretability of linear methods with the flexibility of ensemble methods may also prove fruitful. Ongoing exploration of these avenues stands to enhance the precision, scalability, and policy relevance of land cover classification methodologies.