Comparative LULC Mapping of Boalkhali Upazila: Random Forest vs. K-Means

Introduction

Land Use/Land Cover (LULC) mapping is crucial for urban and environmental management in rapidly developing regions. This study evaluates the efficacy of two distinct classification methodologies—**Supervised (Random Forest, RF)** and **Unsupervised (K-Means)**—in mapping the highly heterogeneous floodplain landscape of Boalkhali Upazila using Sentinel-2 satellite imagery (Jan–Nov 2025) via Google Earth Engine (GEE). The primary challenge in this area, characterized by intermixed settlements and agriculture, is **spectral confusion**. The objective is to determine which approach yields a more accurate and reliable LULC product for local planning.

Methodology

Dataset and Pre-processing

Sentinel-2 Level-2A Surface Reflectance data from 2025 was used. A **median composite** was generated after rigorous cloud masking to create a stable, annual representation of the landscape, neutralizing seasonal effects. The final feature space included the 10 m bands (B_2, B_3, B_4, B_8). Five LULC classes were mapped: **Vegetation, Waterbody, Cropland, Settlement, and Bareland.**

A. Supervised Classification (Random Forest)

Ground-Truth (GT) Collection: 35 GT points were collected via visual interpretation.
Training and Testing: The total GT dataset was rigorously partitioned into two separate, non-overlapping sets:
Training Set: Used exclusively to train the Random Forest classifier.
Independent Testing Set: A total of **35 points** (6 Vegetation, 12 Waterbody, 6 Cropland, 6 Settlement, 5 Bareland) were held back and used **only** for the final accuracy assessment.
Classification: A Random Forest classifier (50 trees) was trained on the GT data to learn the complex spectral patterns corresponding to the five LULC classes.

B. Unsupervised Classification (K-Means)

Clustering: The K-Means algorithm was run to partition the feature space into (K=5) clusters, purely based on minimizing the spectral distance between pixels and cluster centroids.

Interpretation: The resultant clusters were then subjectively interpreted and assigned LULC labels by the analyst for comparative analysis. No formal accuracy assessment was performed due to inherent label ambiguity.

Results

A. Supervised Classification Accuracy

The Random Forest classification demonstrated high performance.

Metric	Value
Overall Accuracy (OA)	**85.71%**
Kappa Coefficient (kappa)	**0.8150**

Table 1: Supervised Confusion Matrix and Accuracy Assessment (N=35)

Classified	Vegetation	Waterbody	Settlement	Cropland	Bareland	Producer's Accuracy (%)	User's Accuracy (%)
Vegetation	**6**	0	0	0	0	100.00	100.00
Waterbody	0	**12**	0	0	0	100.00	100.00
Settlement	0	0	**5**	1	0	62.50	83.33
Cropland	0	0	2	**4**	0	66.67	66.67
Bareland	0	0	1	1	**3**	100.00	60.00

- **High Accuracy Classes:** Vegetation and Waterbody achieved 100\% accuracy (PA and UA), confirming their unique spectral separability.
- **Low Accuracy Classes:** The lowest accuracy values were for **Settlement** (PA 62.50%) and **Bareland** (UA 60.00%), indicating notable confusion with the Cropland class. For instance, 3 of the 8 true Settlement pixels were misclassified as Cropland or Bareland.

B. Area Statistics Comparison

Table 2: Area Statistics Comparison (km^2)

LULC Class	RF Supervised Area (km^2)	K-Means Unsupervised Area (km^2)	Area Difference (RF - K-Means, km2)
Vegetation	77.372	49.339	**+28.033**
Waterbody	7.575	6.931	+0.644
Cropland	31.579	38.223	-6.644
Settlement	6.195	11.649	**-5.454**
Bareland	8.415	24.993	**-16.578**
Total Area	131.136	131.136	0.000

Discussion

1. Performance and Reliability

The **Random Forest** classification significantly outperformed the unsupervised approach, providing a statistically reliable map with an OA of **85.71%** and a **Kappa** coefficient of **0.8150**. The K-Means map, with its inability to account for ground-truth reality, resulted in massive and unacceptable area misallocations.

2. Spectral Confusion and Area Misallocation

The area comparison is the most dramatic finding, directly illustrating the failure of the K-Means algorithm:

Vegetation/Bareland Error: K-Means severely underestimated Vegetation by 28.033 km^2 while simultaneously **overestimating Bareland** by 16.578 km^2.This highlights a fundamental spectral confusion where K-Means grouped healthy vegetation into non-vegetated classes.
Settlement Inflation: K-Means **overestimated Settlement** area by 5.454 km^2.This is consistent with the low User's and Producer's Accuracies for Settlement/Bareland in the supervised matrix, confirming that non-linear RF rules are necessary to separate urban/construction materials from fallow/bare land.

3. Low Accuracy of Urban-Rural Fringe Classes

The lower PA/UA for Settlement and Cropland in the RF result (62.5% and 66.67% respectively) points to the persistent issue of **mixed pixels**. The 10 m resolution is insufficient to cleanly separate small, intermixed patches of urban structure, homestead trees, and agricultural land in this highly complex floodplain environment.

4. Conclusion for Land Management

The superior accuracy and reliable area estimates of the Random Forest map make it the only trustworthy product for local governance. The RF output corrects the critical overestimation of urban sprawl and bare land seen in the K-Means result, providing necessary precision for **urban containment and agricultural zoning** in Boalkhali. Supervised classification is therefore the recommended methodology for LULC mapping in highly heterogeneous landscapes.