Enhanced Water Body Extraction Using PCA-Enhanced SVM Classifiers: A Case Study in the Three Gorges Region with GEE

Wenlei Yang – 259732

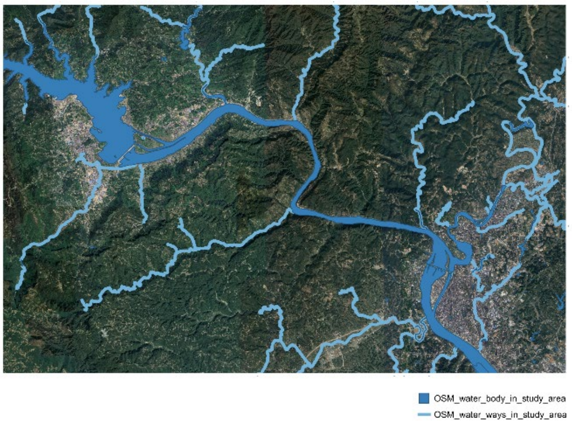
ABSTRACT

This report evaluates the performance of SVM and PCA-SVM classifiers for extracting water bodies using Google Earth Engine (GEE). The study focuses on the Three Gorges region, specifically the Xiling Gorge from Zigui to Yichang, China, with a water area of about 73 square kilometres. The study used Sentinel-2 data from two different periods (2020-08 and 2020-09) to evaluate the classification performance under different hydrological and atmospheric conditions, as severe flooding occurred in August 2020 in southern China and September 2020 after the flooding. The study compares textural features derived from NDWI and PCA and GLCM to improve classification accuracy. The results highlight the potential of PCA-SVM to differentiate water bodies under complex environmental conditions and cloud interference.

1. INTRODUCTION

Water body Segmentation is an important task in environmental monitoring, flood management and hydrological studies. Satellite-based Earth observation data (e.g., Sentinel-2 imagery) provide a lot of multispectral information that can be used for water body mapping. Machine learning techniques including SVM are also increasingly applied to land cover classification. PCA-SVM is firstly performed by Principal Component Analysis (PCA) with reduced features and texture features are extracted by GLCM and then combined with the original bands and classified by SVM to improve the accuracy.

This study was conducted on the GEE platform, focusing on the Three Gorges region, specifically the Xiling Gorge between Zigui and Yichang. This region is characterized by long river systems and complex surrounding terrain, making it a challenging testbed for water body classification. In order to compare the accuracy of water body classification, I selected Sentinel-2 data from two time periods (August and September 2020).



🟦 OSM_water_body_in_study_area
— OSM_water_ways_in_study_area

Figure 1. Overview of the Xiling Gorge

2. MATERIALS AND METHODS

2.1 DATA PREPARATION

2.1.1 Data Selection

Sentinel-2 and Landsat-8 multispectral images of August and September 2020 were selected. The hydrological and atmospheric conditions in these two months were very different: in August 2020, severe flooding occurred in southern China, resulting in elevated water levels and heavy cloud cover. In contrast, September 2020 had less disruptive cloud cover, immediately following the flooding period. Due to the nature of remotely sensed data, cloud cover is an unavoidable challenge. Therefore, a rigorous cloud mask was first performed and the performance of PCA-SVM was compared and investigated under both cloudy and cloud-free conditions.

Sentinel-2 and Landsat-8 images were chosen because they are spectrally and temporally compatible with each other. The Sentinel-2 data with higher spatial resolution is used as the relative truth to compare the classification performance of SVM on low precision imagery. To minimize seasonal differences, the closest time points within the same month were first chosen. Although there are inevitably small differences between the datasets, this approach ensures reasonable consistency in the water column range under similar hydrologic conditions.

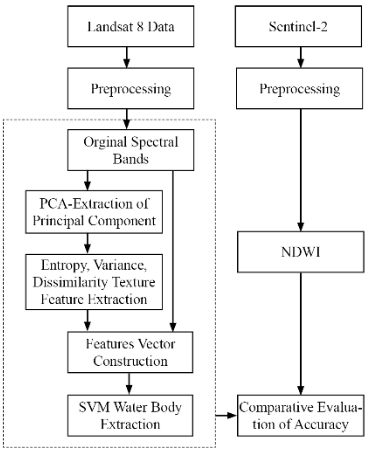
2.1.2 Preprocessing Steps

a) Cloud Masking: Cloud masking was performed using the Sentinel-2 Scene Classification Layer (SCL). The SCL categories for cloud shadow (value 3), high clouds (value 9), and cirrus clouds (value 10) were identified and combined to form a cloud mask. To mitigate edge artifacts, the mask was dilated by 30 meters using a focal minimum filter.

b) Orthorectification: Ensured geometric alignment of images.

2.2 METHODS

The workflow first performs water body segmentation with high-resolution by Sentinel-2 to create relative true value. PCA is then applied to Landsat8 construct feature vectors, which are combined with texture features for input into the SVM water body classification algorithm. Finally, the process concludes with accuracy evaluation to assess classification performance.



Landsat 8 Data
Sentinel-2
Preprocessing
Preprocessing
Orginal Spectral Bands
PCA-Extraction of Principal Component
Entropy, Variance, Dissimilarity Texture Feature Extraction
NDWI
Features Vector Construction
SVM Water Body Extraction
Comparative Evalua-tion of Accuracy

Figure 2. Flowchart of Working Process

2.2.1 Sentinel-2 and NDWI for Water Segmentation

This process leverages Sentinel-2 data at a 10m resolution to extract water body boundaries, with NDWI playing a pivotal role. NDWI is a widely used spectral index for water body identification. It enhances water features by utilizing the strong absorption of near-infrared (NIR) radiation by water and its high reflectance in the green band, providing high accuracy and reliability for water body delineation.

$$NDWI = \frac{Green - NIR}{Green + NIR}$$

A positive NDWI value typically indicates water, while negative values suggest non-water surfaces like soil or vegetation. NDWI is effective but may struggle in distinguishing shallow water or areas with mixed land cover.

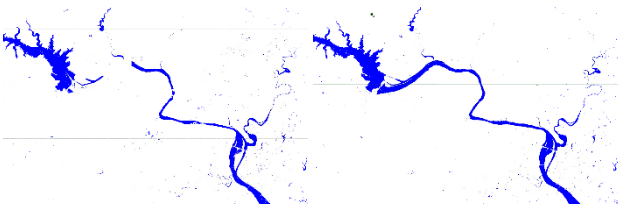


Figure 3. NDWI Images for Two Time Epochs Computed From Sentinel-2, spatial resolution:10m

2.2.2 Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique used to transform correlated features into a set of uncorrelated components. These components capture the maximum variance in the data. Mathematically, PCA transforms data as:

$$\mathbf{C} = \mathbf{R \Lambda R}^{\mathrm{T}}$$
$$\mathbf{y} = \mathbf{R}^{\mathrm{T}}\mathbf{x}$$

where **x** = original data matrix
C = covariance matrix of the input data
Λ = diagonal matrix containing the eigenvalues of C
R = the matrix whose columns are the eigenvector
y = the transformed data

In this study, PCA was applied to spectral bands, and the first principal component (PC_1) was used for texture analysis, as it represents the most significant variance in the dataset.

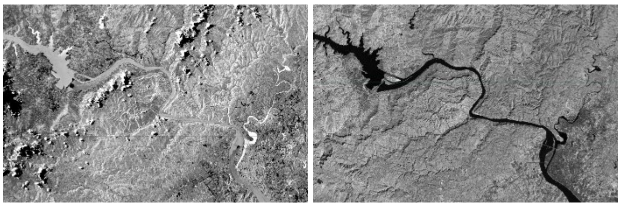


Figure 4. PC1 Band Derived From Landsat 8, spatial resolution:30m

2.2.3 Gray Level Co-occurrence Matrix (GLCM) [1]

GLCM is a texture analysis technique that quantifies the spatial relationship between pixel intensity values. It derives texture features by computing statistics over the co-occurrence matrix, which reflect spatial patterns in the image.

Key features extracted from GLCM include:

a) **Entropy**: Quantifies the randomness in texture. Higher entropy corresponds to more unpredictable patterns
b) **Variance:** Measures the spread of intensity values. Higher variance represents diverse and complex textures.
c) **Dissimilarity**: Highlights local intensity differences, capturing sharp variations in pixel intensities.

In this study, GLCM features were calculated using the first principal component (PC_1), which retains the most critical spatial information. These features enhanced the classification performance by revealing underlying spatial patterns.

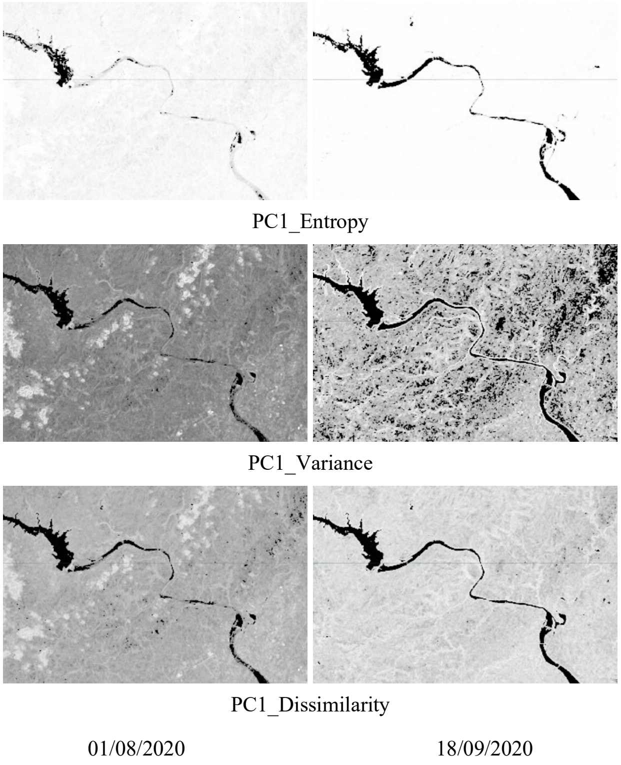


Figure 5. NDWI Images for Two Time Epochs Computed From Sentinel-2, spatial resolution:10m

2.2.4 Build Feature Vectors

Two distinct feature vectors were constructed for SVM classification:

Spectral Bands Only: This vector consisted of the original spectral bands (Blue, Green, Red, NIR).

Spectral Bands + Texture Features: This extended vector included the spectral bands, and PCA-GLCM derived texture features (entropy, variance, dissimilarity), with 7 dimensions in total.

To evaluate the classification performance, 1,600 sample points were generated in total, evenly divided between water and non-water classes (800 points each). These samples were derived using NDWI-based thresholds for water body delineation. The dataset was split into training (70%) and validation (30%) subsets to ensure robust accuracy assessment.

The comparison of these feature sets and sampling points allowed for evaluating the added value of texture features in enhancing classification accuracy.

2.2.5 Support Vector Machine (SVM) - Classification

SVM is a supervised machine learning algorithm used for classification tasks. It identifies an optimal hyperplane that separates data into distinct classes. For nonlinear datasets, the radial basis function (RBF) kernel is commonly used to map data into a higher-dimensional feature space where it becomes linearly separable. The decision function can be expressed as:

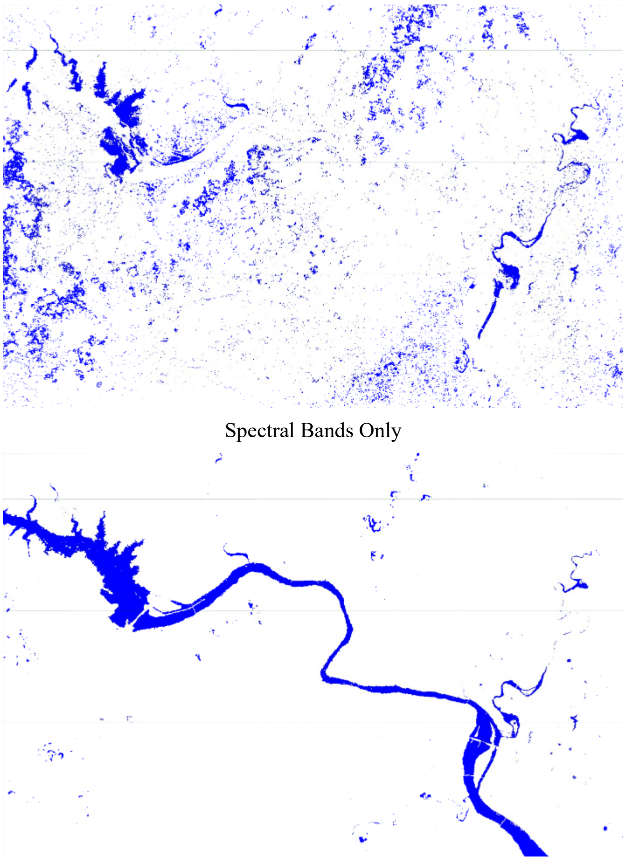
$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b$$

where $\mathbf{x}$ = input feature vector
$\mathbf{x_i}$ = support vectors
$\boldsymbol{K}(\mathbf{x},\mathbf{x_i})$ = kernel function
α_i = the Lagrange multipliers
b = the bias

SVM is robust in handling high-dimensional data. This study applied SVM to classify features in 30m resolution remote sensing imagery with and without Principal Component Analysis (PCA), to compare classification performance. SVM was particularly effective in binary classification tasks, such as identifying the presence of water bodies. While SVM is designed for binary classification, it can also be extended to multiclass classification using strategies like one-vs-one or one-vs-all.

3. RESULTS DISCUSSION

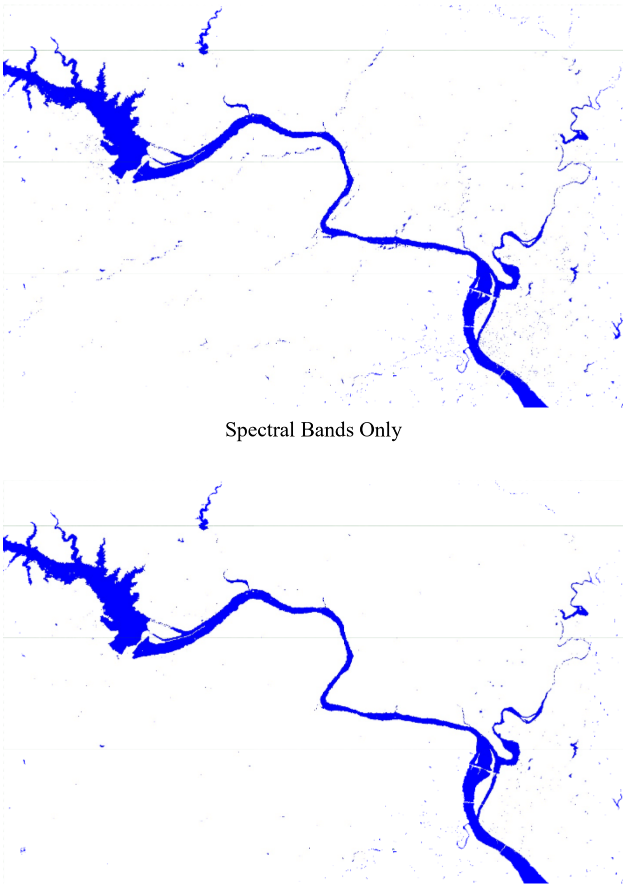
3.1 CLASSIFICATION RESULTS



Spectral Bands Only

Classification with Spectral Bands + Texture Feature

Figure 6. Comparison of Waterbody Classification Results for 01/08/2020



Spectral Bands Only

Classification with Spectral Bands + Texture Feature

Figure 7. Comparison of Waterbody Classification Results for 18/09/2020

3.2 Analysis of Classification Results

3.2.1 Confusion Matrix

August 2020

	Predicted Water	Predicted Non-Water
Water	206	7
Non-Water	6	4

Table 1. SVM - Spectral Bands Only

	Predicted Water	Predicted Non-Water
Water	212	1
Non-Water	0	10

Table 2. SVM - Spectral Bands Only + Texture Features

September 2020

	Predicted Water	Predicted Non-Water
Water	244	1
Non-Water	1	9

Table 3. SVM - Spectral Bands Only

	Predicted Water	Predicted Non-Water
Water	245	1
Non-Water	0	9

Table 4. SVM - Spectral Bands Only + Texture Features

Accuracy Metrics

Metric	**SVM (%)**	**PCA-SVM (%)**
Overall Accuracy (August)	94.17	99.55
Overall Accuracy (September)	99.22	99.61
Producer's Accuracy (Water)	91.15	99.50
User's Accuracy (Water)	97.16	99.58

Table 5. Accuracy Metrics

3.2.2 Analysis

01/08/2020 - Results with Strong Cloud Interference

Spectral Bands Only: Affected by cloud interference, the classification results show obvious problems of fragmented and discontinuous water body morphology, leading to misclassification and loss of water body shape.

Spectral Bands + Texture Features: The texture features (entropy, variance, and discreteness) extracted by PCA significantly enhance the classification effect and maintain the integrity and coherence of the shape of the water body even in the presence of cloud interference.

Results: The texture feature effectively reduces the effect of data noise and improves the recognition of continuous water bodies.

18/09/2020 - Results Without Cloud Interference

Spectral Bands Only: With less cloud interference and higher sample quality, the classification results are highly accurate with clear and less noisy boundaries.

Spectral Bands + Texture Features: While still valid, the texture features introduce a certain amount of redundant information, resulting in slightly lower classification accuracy than models using only spectral bands, with overfitting or minor misclassification.

Results: When the spectral data quality is high, the addition of texture features does not significantly improve the classification results but may increase the model complexity.

August SVM and PCA-SVM classification correctness was significantly different (94.17% vs 99.55%). Under the cloud interference conditions, the spectral features of water bodies were disturbed, and the SVM model relying solely on spectral data could not accurately differentiate water bodies from non-water body areas. NDWI may not be able to accurately calculate the reflectance features of water bodies under strong cloud interference, resulting in fewer water body areas being identified. PCA-SVM makes up for the lack of spectral features by introducing texture features (e.g., entropy and heterogeneity in the GLCM) and can somewhat eliminate the effects of cloud interference. that can eliminate the effect of cloud interference to a certain extent.

4. CONCLUSIONS

PCA texture features have a significant advantage in the case of noisy or incomplete data, can enhance the classification effect by capturing spatial relationships under significant interference, providing additional spatial information; however, when the spectral information is sufficiently accurate and the interference is low, the gain of texture features is limited，the model using only spectral bands is simpler and more effective.

There is a resolution mismatch, the reference data is Sentinel-2 (10m resolution), while the classification data is Landsat-8 (30m resolution), the resolution mismatch may lead to the blurring of the boundary, affecting the generalization ability of the model.

The classification is limited by insufficient band information, especially the lack of key bands such as shortwave infrared (SWIR) to enhance the ability to distinguish between water and non-water bodies. At the same time, the digital elevation model (DEM) was not introduced to capture terrain information, which may limit the classification performance under complex terrain conditions. These issues can be used as the development direction of future research to further improve the accuracy and applicability of classification.

REFERENCES

[1] R. M. Haralick, K. Shanmugam and I. Dinstein, *"Textural Features for Image Classification,"* in IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610-621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[2] 周婷,汪炎,邹俊,等.基于 PCA 和 SVM 的遥感影像水体提取方法及验证[J].水资源保护,2023,39(2):180-189.(ZHOU Ting, WANG Yan, ZOU Jun, et al. *PCA and SVM-based algorithm of water area extraction from remote sensing images and its verification*[J]. Water Resources Protection, 2023, 39(2): 180-189. (in Chinese)

[3] Myint, S. W., Lam, N., & Tyler, J. (2002). *An evaluation of four different wavelet decomposition procedures for spatial feature discrimination in urban areas*. Transactions in GIS, 6(4), 403-429. https://doi.org/10.1111/1467-9671.00120

[4] 姜青香 刘慧平 利用纹理分析方法提取 TM 图像信息[J]. 遥感学报, 2004, 8 (5): 458-464. (JIANG Qingxiang, LIU Huiping. *Extracting TM image information using texture analysis* [J]. Journal of Remote Sensing, 2004,8(5):458-464. (in Chinese))

Code: https://github.com/wenleiyah/gee_svm