

Supported by:



Federal Ministry  
for the Environment, Nature Conservation  
and Nuclear Safety



Federal Agency for  
Nature Conservation



# EOCap4Africa

## 9 Raster Analysis

### d) Land Cover Classification Practical



**INES Ruhengeri**  
Institute of Applied Sciences



# Learning Objectives



Understand how to create high-quality training data for classification

Learn best practices for selecting representative training samples

Apply training data to classify land cover using a Random Forest model

# Training Data



Training data is the foundation of supervised classification

- The quality of training data directly impacts classification accuracy
- Poor training data can lead to misclassification and unreliable results
- Training data should be well-distributed, balanced, and spectrally distinct

# Training Data



## Characteristics of good training data

- **Representative** – Covers all land cover classes in the study area
- **Balanced** – Avoid class imbalances by ensuring roughly equal sample sizes
- **Spatially Distributed** – Spread across different locations to account for variability
- **Spectrally Pure** – Use homogeneous areas to avoid mixed pixels
- **Independent Validation Set** – Keep separate data for accuracy assessment



# R for Random Forest Classification

- **Better Model Control** – Allows fine-tuning of hyperparameters (e.g., number of trees, depth, feature selection)
- **Faster Processing** – More efficient for large datasets compared to QGIS
- **Advanced Accuracy Metrics** – Generates confusion matrices, feature importance scores, and cross-validation
- **Better Performance Tracking** – Can visualize classification accuracy and analyze errors
- **Seamless Integration with GIS** – Results can be exported back into QGIS for visualization and further spatial analysis

QGIS can run Random Forest, but it has **limited customization** and may struggle with **large datasets**

# Task



## **Create your own Land Cover Classification!**

As an example, we are investigating Wetlands in Rwanda

1. Create your own training data in QGIS
2. Run a Random Forest Model in RStudio
3. Visualise your results in RStudio or QGIS

# Interpretation of Models



Output of the Random Forest model -> but what does this mean?

```

Confusion matrix:
      agriculture forest urban water wetlands class.error
agriculture      79      2      6      5      0 0.14130435
forest           1     148      0      5      0 0.03896104
urban            17      2     11     20      1 0.78431373
water             6      1      2    648      2 0.01669196
wetlands          3      4      1     28      8 0.81818182
    
```

# Interpretation of Models



Actual → Predicted	agriculture	forest	urban	water	wetlands	class.error
agriculture	79	2	6	5	0	14.1% misclassified
forest	1	148	0	5	0	3.9% misclassified
urban	17	2	11	20	1	78.4% misclassified
water	6	1	2	648	2	1.7% misclassified
wetlands	3	4	1	28	8	81.8% misclassified



# Interpretation of Models



## Good Performances

- Water (97.8% accuracy) – The model is classifying water very well, with only 1.7% error
- Forest (96.1% accuracy) – Also good, with only 3.9% misclassified cases

## Bad Performances

- Urban (only 11/51 correct, 78.4% error) – The model struggles to distinguish urban areas, misclassifying them as agriculture and water
- Wetlands (only 8/44 correct, 81.8% error) – The worst class! Wetlands are being confused with water (28 cases)



# Interpretation of Models

## 1. Not Enough Training Data for Certain Classes

- Urban and wetlands have very high misclassification rates
- They likely have too few training samples or are too similar to other classes (e.g., wetlands vs. water)
- Increase the number of training points for urban and wetlands

## 2. Overlapping Spectral Signatures

- Wetlands and water are confused because they likely have similar spectral reflectance
- Urban areas are confused with agriculture and water, which may indicate that urban pixels include mixed land cover types
- Try adding more spectral bands to improve separability

# Interpretation of Models



## 3. Class Imbalance

- Water (648 cases) dominates the dataset, while urban (11 cases) and wetlands (8 cases) are underrepresented
- The Random Forest model will naturally be biased toward classes with more data
- Balance the dataset by using equal numbers of training samples per class

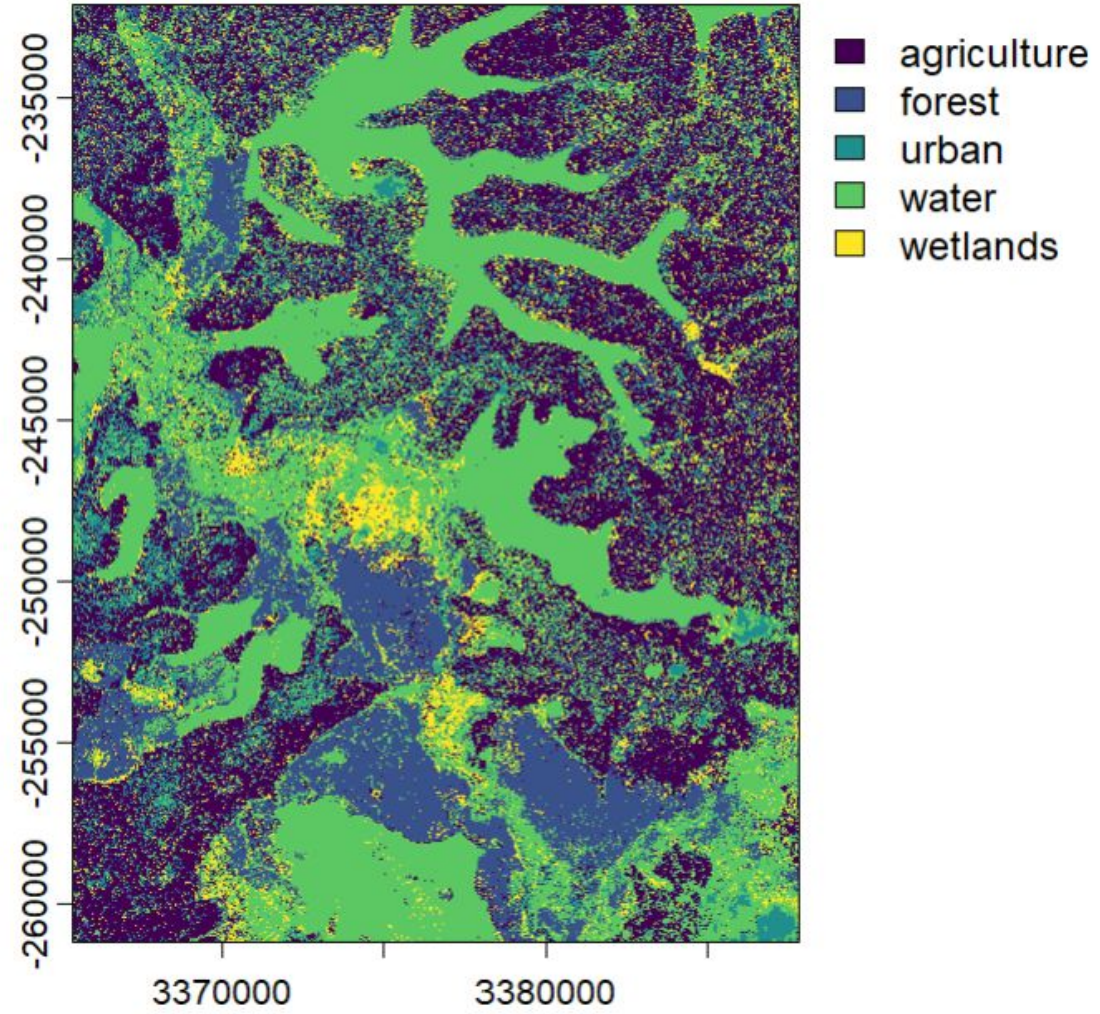
## 4. Feature Selection

- The features (raster bands) used for training might not be sufficiently different for urban, wetlands, and water
- Add additional data like:
  - Vegetation indices (NDVI, NDBI) to separate vegetation and built-up areas
  - Texture analysis to distinguish urban areas from natural ones

# Results



## Random Forest Land Cover Classification





# Summary & Key Takeaways

**Good training data** is essential for accurate land cover classification.

Training samples must be **well-distributed, balanced, and spectrally pure**.

QGIS can be used to **create training data, train the model and run the prediction**

Supported by:



Federal Ministry  
for the Environment, Nature Conservation  
and Nuclear Safety



Federal Agency for  
Nature Conservation



# Thank you for your attention!

Dr. Insa Otte, Hanna Schulten  
 (on behalf of the EOCap4Africa Team)  
 and colleagues

insa.otte@uni-wuerzburg.de



**INES Ruhengeri**  
 Institute of Applied Sciences

