-Concept Note-

Advanced Urban-Rural Classification: Integrating Building Type Recognition in Remote Sensing with Machine Learning and GIS

Sila Tonboot

*Centre For Health Equity Monitoring Foundation

Release: 15th March 2022

1. Introduction

The delineation of urban and rural areas is crucial for urban planning, socioeconomic research, and environmental management. Traditional classification methods often lack the nuance to capture the complexity of modern landscapes. This paper proposes an advanced approach to urban-rural classification by integrating building type recognition into remote sensing techniques using machine learning and Geographic Information Systems (GIS), specifically QGIS.

The rapid urbanization witnessed globally in recent decades has blurred the once-clear boundaries between urban and rural areas. Peri-urban zones, suburban sprawl, and rural towns with urban characteristics challenge conventional classification methods. These traditional approaches, often based on population density, administrative boundaries, or simplistic land use categories, fail to capture the intricate mosaic of the built environment that characterizes contemporary landscapes.

Remote sensing has emerged as a powerful tool for analyzing land use and land cover at various scales. However, most current applications in urban-rural classification focus on basic land cover types or simple built-up area detection. They often fail to distinguish between different types of urban fabric, such as high-density city centers, suburban residential areas, or industrial zones. This limitation hinders our ability to understand the complex spatial patterns of human settlements and their implications for planning and policymaking.

Recent advancements in machine learning, particularly in the field of computer vision, offer new opportunities to extract more detailed information from remote sensing data. Convolutional Neural Networks (CNNs) have shown remarkable success in image recognition tasks, including the detection and classification of objects in satellite imagery. By leveraging these techniques, we can move beyond simple built-up area detection to identify specific building types and urban morphologies.

The integration of GIS, specifically open-source software like QGIS, provides a robust framework for spatial analysis and visualization of the results. GIS allows for the incorporation of multiple data sources, spatial statistics calculation, and the creation of detailed maps that can inform decision-making processes. The combination of machine learning for image analysis and GIS for spatial processing and visualization creates a powerful toolset for advanced urban-rural classification.

This paper proposes a methodology that harnesses these technological advancements to develop a more nuanced and accurate approach to urban-rural classification. By recognizing and analyzing different building types - such as single-family homes, apartment blocks, commercial structures, and industrial facilities - we can create a more sophisticated understanding of the urban-rural continuum. This approach not only improves the accuracy of classification but also provides valuable insights into the functional characteristics of different areas.

The proposed methodology has wide-ranging applications. In urban planning, it can help identify areas of urban sprawl, inform zoning decisions, and guide infrastructure development. For socioeconomic research, it provides a more accurate spatial framework for analyzing patterns of economic activity, social inequality, and quality of life. In environmental management, it can assist in assessing the impacts of urbanization on ecosystems and natural resources.

By combining advanced machine learning techniques with the analytical power of GIS, this approach represents a significant step forward in our ability to understand and manage the complex landscapes of the 21st century. The following sections will detail the methodology, discuss potential applications, and address challenges and considerations for implementation.

2. Methodology

2.1 Data Acquisition and Preprocessing

- Data Sources: High-resolution satellite imagery and LiDAR (Light Detection and Ranging) data will be obtained for the study areas.
- Preprocessing in QGIS:
    1. Georeferencing: Assigning geographic coordinates to the imagery to align it with a map projection.
    2. Atmospheric correction: Removing atmospheric effects from satellite imagery to improve accuracy.
    3. Image mosaicking: Combining multiple images into a single seamless image covering the entire study area.

2.2 Building Detection and Segmentation

- Method: Convolutional Neural Network (CNN) using U-Net architecture
- Process: The CNN analyzes the preprocessed imagery to identify and outline individual buildings.
- Equation: $S(x) = \text{argmax}_c P(c|x)$
    - $S(x)$ is the final segmentation map
    - c represents the class (building or non-building)
    - $P(c|x)$ is the probability that pixel x belongs to class c
- Output: A binary map showing building footprints

2.3 Building Type Classification

- Method: Multi-class CNN

- Process: The CNN analyzes each detected building and classifies it into categories (e.g., residential, commercial, industrial, institutional)
- Equation: $y = \text{softmax}(Wx + b)$
    - y is the probability distribution over building types
    - W is the weight matrix (learned by the network)
    - x is the input feature vector (derived from the building image)
    - b is the bias term
- Output: A map with each building colored according to its classified type

2.4 Feature Extraction

- Method: Spatial analysis in QGIS
- Process: For each 1 km² grid cell, calculate:
    1. Density of each building type: $\rho i = Ni / A$
        - ρi is the density of building type i
        - Ni is the number of buildings of type i
        - A is the area of the grid cell (1 km²)
    2. Diversity index: $D = -\Sigma(pi * \ln(pi))$
        - D is the diversity index
        - pi is the proportion of buildings of type i
- Output: A table with each grid cell and its calculated features

2.5 Urban-Rural Classification

- Method: Random Forest classifier
- Process:
    1. Train the classifier using the extracted features
    2. Apply the trained classifier to categorize each grid cell along an urban-rural continuum
- Feature importance calculation: $Ij = \Sigma(\text{decrease in node impurity}) / (\text{number of trees})$
    - Ij is the importance of feature j
    - This helps identify which features are most crucial for classification
- Visualization: Use QGIS to create maps showing the final urban-rural classification

This methodology combines advanced machine learning techniques (CNNs for building detection and classification) with traditional GIS analysis (feature extraction and visualization in QGIS). The process moves from raw imagery to a detailed urban-rural classification, providing a nuanced view of the landscape based on the types and distributions of buildings present.

3. QGIS Integration

QGIS, an open-source Geographic Information System (GIS) software, is integral to this methodology, providing powerful tools for data management, analysis, and visualization. Here's a detailed explanation of how QGIS will be used:

3.1 Data Preparation

a) Importing and managing satellite imagery and LiDAR data:

- Use QGIS's "Add Raster Layer" function to import satellite imagery.
- For LiDAR data, use the "LAStools" plugin to import and process point cloud data.

b) Creating grid cells for analysis:

- Utilize the "Create Grid" tool in QGIS to generate a 1 km² grid over the study area.
- This grid will serve as the basis for feature extraction and classification.

c) Preprocessing imagery:

- Use the Semi-Automatic Classification Plugin (SCP) for:
  - Atmospheric correction: Applying DOS1 (Dark Object Subtraction) method.
  - Pansharpening: Improving spatial resolution of multispectral imagery.
  - Mosaicking: Combining multiple image tiles into a seamless dataset.

3.2 Feature Calculation

a) Building density calculation:

- Use QGIS's "Count Points in Polygon" tool to count buildings in each grid cell.
- Apply field calculator to compute density: $\rho i = Ni / A$

b) Spatial statistics:

- Develop custom Python scripts using PyQGIS to calculate advanced spatial statistics.

3.3 Result Visualization and Analysis

a) Creating thematic maps:

- Use QGIS's symbology options to create color-coded maps based on building types and urban-rural classifications.
- Utilize the "Print Layout" tool to design publication-quality maps with legends, scale bars, and north arrows.

b) Spatial analysis:

- Apply QGIS's built-in spatial analysis tools like:

- Cluster analysis to identify urban centers
- Buffer analysis to examine proximity relationships
- Overlay analysis to compare classification results with other spatial data (e.g., population density)

3.4 Integration with Machine Learning

QGIS can be seamlessly integrated with Python-based machine learning libraries:

a) Setting up the environment:

- Install required Python libraries (e.g., TensorFlow, scikit-learn) in QGIS's Python environment.

b) Developing a plugin:

- Create a custom QGIS plugin that implements the CNN and Random Forest models.
- Use PyQGIS to access and manipulate spatial data within the machine learning workflow.

4. Potential Applications in Health Workforce Distribution in Thailand

4.1 Health Facility Accessibility Analysis

Using QGIS, researchers can create detailed maps of health facility distributions in relation to urban-rural classifications:

- Map health facilities (hospitals, clinics, community health centers) across Thailand.
- Overlay this with the urban-rural classification derived from building type analysis.
- Use QGIS's network analysis tools to calculate accessibility metrics:
  - Travel time isochrones from population centers to nearest health facilities.
  - Service area analysis to identify underserved regions.

This analysis can help identify disparities in healthcare access between urban and rural areas, informing decisions on where to establish new health facilities.

4.2 Time-Series Analysis of Health Workforce Distribution

QGIS's temporal capabilities can be leveraged to analyze changes in health workforce distribution over time:

- Create time-series maps of health worker density (e.g., doctors per 1000 population) for different urban-rural categories.
- Use QGIS's time manager plugin to visualize how health workforce distribution has changed alongside urbanization patterns.
- Perform trend analysis to forecast future needs based on urban growth patterns.

This temporal analysis can help policymakers understand how urbanization trends are affecting health workforce distribution and predict future needs.

4.3 Policy Support for Equitable Health Workforce Distribution

The detailed urban-rural classification can inform policies aimed at achieving more equitable health workforce distribution:

- Use QGIS to create suitability maps for new healthcare facilities or incentive zones for health workers:
  - Combine urban-rural classification with other factors like population density, existing healthcare coverage, and socioeconomic indicators.
  - Use multi-criteria decision analysis (MCDA) tools in QGIS to identify optimal locations for interventions.
- Develop a QGIS plugin for policy scenario testing:
  - Allow policymakers to input different resource allocation scenarios.
  - Visualize the potential impact on health workforce distribution and accessibility.

5. Challenges and Considerations

5.1 Computational Resources Processing high-resolution imagery and running complex machine learning models require significant computational power.

5.2 Model Generalization Ensuring the building type classification model generalizes well across different geographic contexts is crucial.

5.3 Data Privacy Handling high-resolution building data raises privacy concerns that must be addressed.

6. Conclusion

This concept paper presents an advanced approach to urban-rural classification by integrating building type recognition into remote sensing analysis through machine learning and GIS techniques. The use of QGIS as a central tool in this methodology offers powerful capabilities for data processing, analysis, and visualization.

The proposed approach leverages cutting-edge machine learning techniques and the robust spatial analysis capabilities of QGIS to extract detailed information from remote sensing data, offering a sophisticated understanding of urban and rural landscapes. While challenges exist in computational requirements and model generalization, the potential applications of this approach are significant for urban studies, planning, and policymaking.

Future research should focus on optimizing the integration of machine learning models with QGIS, developing plugins for streamlined workflow, and validating the approach across diverse geographic contexts. As GIS and machine learning technologies continue to evolve, methodologies like this will play a crucial role in advancing our understanding and management of urban and rural environments.