

## Summary

The current paper proposes a cost-effective and robust rural-urban classifier. Compared to existing Nielsen methodologies for rural-urban classification, the novelty of the current approach is that instead of spectral characteristics of the satellite images data, features extracted from road networks are the subjects of machine learning algorithms. The resultant rural-urban classifiers have displayed high performances in terms of classification accuracy and stability. More importantly, by utilizing road network characteristics as input features, the proposed classification framework is able to greatly reduce the cost of data acquisition and the amount of image processing prior to the construction of the rural-urban classifiers. If implemented, the proposed rural-urban classification methodology has the potential to facilitate regular re-evaluation of the rural-urban status for regions in rapidly developing countries such as Indonesia and India.

The Indonesian province of Bali is used as a proof-of-concept in the current study. Collaborating with Dr. Tung Whye Loon from the Nielsen Innovation Lab, I built an application for the automatic retrieval and vectorization of road network in the Python programming language (see Chapter 3). The resultant application was subsequently used to obtain road network information of 382 towns/villages in Bali to construct a labeled feature set with two-class rural-urban categorization. As described in Section 5.1, I applied four machine learning algorithms: decision tree, random forest, neural network and logistic regression to 1000 bootstrap samples generated from the labeled feature set. The results are presented in Section 5.1.1. With a detailed analysis of the empirical results as illustrated in Section 5.1.2, I found that a two-class Rural/Urban labeling scheme does not adequately represent the various degrees of urban development in Bali.

In Section 5.2, following k-means clustering and a review of literature on rural-to-urban land conversion in developing countries, it is shown that the dataset likely exhibits a five-cluster structure. After a detailed visual inspection of the regions in each identified cluster, I subsequently proposed a new labeling scheme consisting of five classes: ‘Core-Rural’, ‘Rural’, ‘Rural-Urban Fringe’, ‘Urban’ and ‘Core-Urban’. The performances of the benchmarking modeling techniques under the new five-class labeling scheme are presented in Section 5.3.1 and 5.3.2.