Impervious land classification using bootstrap principal component analysis

IOP Conf. Series: Materials Science and Engineering **508** (2019) 012116 doi:10.1088/1757-899X/508/1/012116

Impervious land classification using bootstrap principal component analysis

Janson Hendryli*, Dyah Erny Herwindiati and Junita Merdi

Informatics Department, Faculty of Information Technology
Universitas Tarumanagara, Jakarta 11440, Indonesia

*jansonh@fti.untar.ac.id,

Abstract. This paper proposes the bootstrap principal component analysis for the classification of an impervious land area. The impervious surface is closely related to the urban development area, and therefore, crucial to the urban planning of the city. In this paper, we use the remote sensing technology to map each pixel of the images as impervious land, water area, or empty lands. The case study is the Jabodetabek area in Indonesia, which is a megapolitan area consisting of megacities of Jakarta, Bogor, Depok, Tangerang, and Bekasi. The area is the political and economic center of the country. We evaluate the model using Cohen's kappa coefficient, which shows that the model has excellent performance in classifying those three land-cover classes.

1. Introduction

The development of the urban area is closely related to the land-cover and land-use change of the area over time [1]. The decision process of urban development involves many parties, and the decision can affect many aspects in society. Economically, the expansion of city areas can bring prosperity and sources of income to the region. Though unplanned development of urban areas can lead to many environmental issues, such as floods and soil degradation. Therefore, knowledge of land-use and land-cover within an area becomes increasingly essential for the stakeholders [2]. By mapping and analyzing the regional changes in land-use, it is hoped that the development of the area can avoid the deterioration of the environmental quality and the destruction of the ecosystem [3].

One of the key indicators of economic development in urban areas is the impervious area [4, 5], e.g., buildings, rooftops, paved roads, and parking lots, in the region. With the rapid urbanization, the manual process of surveying and analyzing the land-cover by hand can be tedious, inefficient, and ineffective. Therefore, a system which can map a large area automatically can be helpful for the urban development planning. By utilizing the remote sensing technology and satellite imagery, the system can process the image of a particular region and classify each pixel as one of the land-covers. This paper proposes the impervious land classification system using the bootstrap principal component analysis and regression model. Additionally, the system can also detect the water area and empty lands, as those land-covers are closely related to the impervious area.

2. Data and Methods

This section discusses the data and methods used in this research. The impervious land classification system employs the remote sensing data retrieved from the Landsat 7 satellite. The digital values of each

Landsat 7 bands are analyzed using the bootstrap resampling and the principal component analysis to classify each pixel as either impervious surface, water bodies, or empty lands.

2.1. Data

This research uses the Landsat 7 satellite images that are freely provided by the United States Geological Survey and can be downloaded from the GloVis website[1]. The satellite is the seventh iteration of the Landsat satellite program that was launched on April 15, 1999, and currently being managed by the NASA. The Landsat 7 is equipped with the enhanced thematic mapper plus (ETM+) sensor [6], which consists of seven thermal bands and one panchromatic band. On May 31, 2003, the scan line corrector (SLC) instrument experienced permanent failure resulting in data gaps on every image collected since July 2003 [7]. Despite the failure, the Landsat 7 still operates normally, even though some preprocessing methods to fill the data gaps are needed to compensate the instrument failure. Moreover, the Landsat 7 satellite has 16 days repeat cycle.

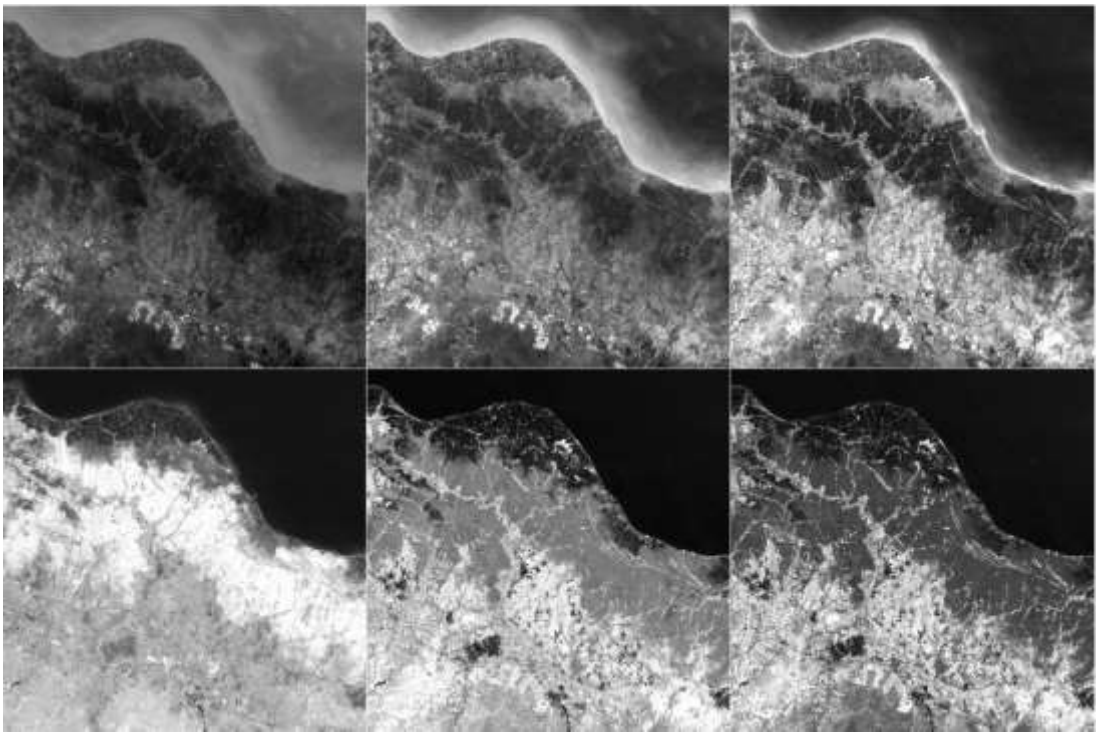

Figure 1. The Landsat 7 satellite imagery (band 1-5 and 7) showing the area of Bekasi, Karawang, and Cikarang.

Table 1 below shows the characteristic of each Landsat 7 bands. Satellite images from band 1, 2, 3, 4, 5, and 7 are used in this research. We experiment on 2006, 2008, 2010, 2011, and 2012 satellite images of Jakarta, Bogor, Depok, Tangerang, and Bekasi (Jabodetabek) area in Indonesia. The preprocessing of the data involves the cropping, gap filling, geometric correction, and overlays, which are all conducted using the ENVI software. We select 75 different locations in the Jabodetabek area as the training data.

2.2. Methods

The system starts with extracting 5 x 5 pixels of impervious, water, and land areas from the Landsat 7 satellite imagery. Afterward, the mean vectors of each band are computed and then standardized. The next step is the dimensionality reduction process by calculating the eigenvalue and eigenvector of the

[1] https://glovis.usgs.gov

matrix. After finding the decomposition matrix, we can train the principal component regression model. This trained model can then be used for the classification of impervious surface from the testing data.

Bootstrap resampling is a data sampling method where samples are taken arbitrarily with replacement from the population [8, 9]. We employ this sampling strategy in the experiments, specifically in the training phase.

Table 1. The characteristic of Landsat 7 ETM+ satellite bands [10]

BANDS	WAVELENGTH	RESOLUTION
Band 1 – Blue	0.45 – 0.52	30 m
Band 2 – Green	0.52 – 0.60	30 m
Band 3 – Red	0.63 – 0.69	30 m
Band 4 – Near Infrared (NIR)	0.77 – 0.90	30 m
Band 5 – Shortwave Infrared (SWIR) 1	1.55 – 1.75	30 m
Band 6 – Thermal	10.40 – 12.50	60 m
Band 7 – Shortwave Infrared (SWIR) 2	2.09 – 2.35	30 m
Band 8 – Panchromatic	0.52 – 0.90	15 m

The principal component analysis (PCA), which is also known as Karhunen-Loève transform, is a dimensionality reduction method which can be defined as the orthogonal projection of the data onto a lower dimensional linear space or as the linear projection that minimizes the average projection cost [11]. To calculate the PCA, we employ the singular value decomposition (SVD) method to compute the eigenvector which will be used by the PCA.

Let X be the input matrix from the Landsat 7 satellite imagery of size $n \times p$ and the mean vector $\vec{\bar{X}} = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_p)$. The standardization process computes $B_{ij} = x_{ij} - \bar{x}_j$ for $i = 1, 2, 3, \dots, n$. We can then compute the eigenvalue λ of C and D, such that $|C - \lambda I| = 0$ and $|D - \lambda I| = 0$ where $C = B^T B$ and $D = BB^T$; and the eigenvector V such that $(C - \lambda I)V = 0$ and $(D - \lambda I)V = 0$. The decomposition matrix of B can then be calculated using SVD as in Eq. (1) where U denotes the decomposition matrix.

$$U = B\, V\, \lambda^{-1} \tag{1}$$

The principal component matrix can then be calculated as:

$$W = W_U^T \tag{2}$$

where $W_u = \lambda U^T$. The coefficient β of the regression model $\hat{Y} = W \cdot \beta$ can then be calculated using the normal equation as in Eq. (3).

$$\beta = (W^T W)^{-1} W^T Y \tag{3}$$

where Y denotes the actual class of impervious, water, or land area from the training data.

Initially, in order to evaluate the regression model at the testing phase, the principal components of the test data have to be calculated as in Eq. (4) where V is the eigenvector from the training phase.

$$W_{test} = X_{test}\, V \tag{4}$$

Afterward, the test data can be classified using the regression model as in Eq. (5) where β is the coefficient of the model. It should be noted that the output variable Y uses dummy dependent variables similar to [12].

$$Y_{test} = W_{test}\,\beta \tag{5}$$

3. Results and Discussions

The performance metric we employed for the model evaluation is the Cohen's kappa coefficient κ [13], which measures the inter-rater agreement of categorical items [14]. The coefficient can be computed as in Eq. (6) where p_o is the relative observed agreement or accuracy and p_e is the hypothetical probability of chance agreement. Complete agreement is shown by the value of $\kappa = 1$, while $\kappa = 0$ shows that there is no agreement. It should be noted that a negative value of κ is possible which implies that there is no effective agreement or the agreement is worse than random [14].

$$\kappa = 1 - \frac{1-p_o}{1-p_e} \tag{6}$$

The training data consist of 750 pixels retrieved from 75 different locations in Jabodetabek area, where we select 150 pixels from each region (50 pixels from each class). The number of testing data is 400 pixels, also from the Jabodetabek area, which consists of 150 impervious area, 100 water bodies, and 150 open lands. Moreover, the experiments are conducted 20 times. Figure 2 shows the result of each experiment.

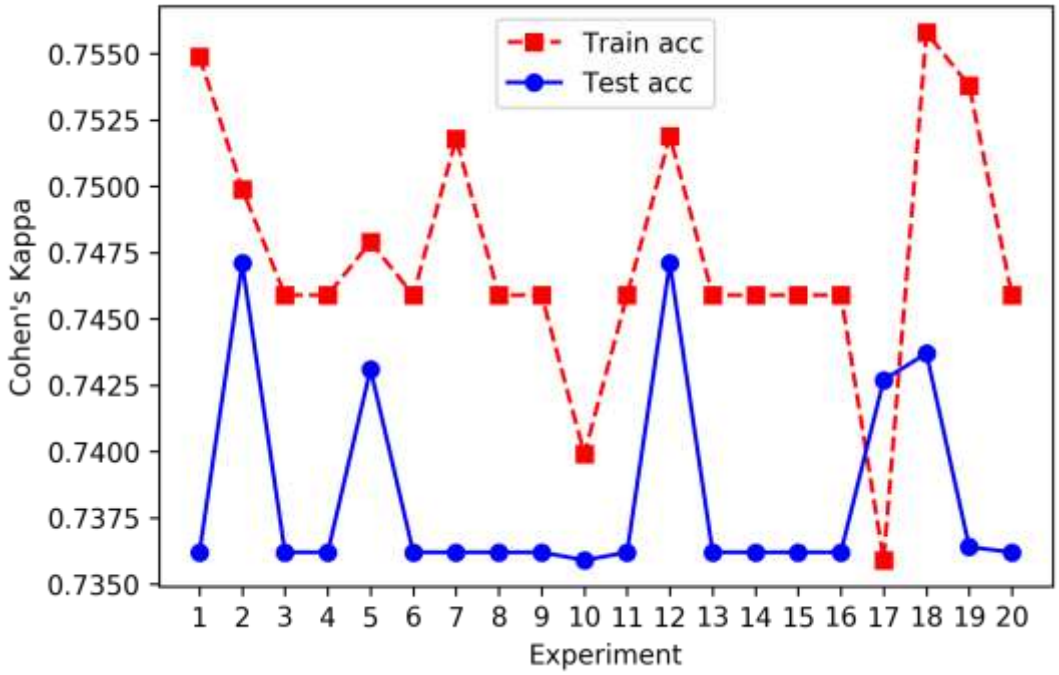

Figure 2. The Cohen's kappa of each experiment shows the average value of 0.747335 (training) and 0.73833 (testing).

The average Cohen's kappa from the experiments is 0.747335 and 0.73833 for the training and testing, respectively. It can be said that the result shows that there is a good agreement from the experiments and implies that the model shows a good performance.

4. Conclusions

This paper demonstrates the classification of impervious, water, and land area from Landsat 7 satellite images using the bootstrap principal component analysis and regression model. This research uses satellite imagery from the Jakarta, Bogor, Depok, Tangerang, and Bekasi (Jabodetabek) area in Indonesia. The principal component analysis uses the eigenvector calculated by the singular value decomposition method. The experiments show that the regression model can accurately classify the classes with good performance.

For the future works, the model can be further developed to classify other land covers, such as vegetation. A good automatic land cover classification may help the government in making urban planning decisions.

5. References

[1] J. R. Anderson, 1976. A Land Use and Land Cover Classification System for Use with Remote Sensor Data **964**. US Government Printing Office.

[2] D. E. Herwindiati, M. A. Djauhari, L. Jaupi, 2013. Robust Classification of Remote Sensing Data for Green Space Analysis. *Journal of Mathematics and System Science* **3**, pp. 180-186.

[3] F. Yuan, K. E. Sawaya, B. C. Loeffelholz, M. E. Bauer, 2005. Land Cover Classification and Change Analysis of the Twin Cities (Minnesota) Metropolitan Area by Multitemporal Landsat Remote Sensing. *Remote Sensing of Environment* **98**(2-3), pp. 317-328.

[4] C. L. Arnold, C. J. Gibbons, 1996. Impervious Surface Coverage: The Emergence of a Key Environmental Indicator. *Journal of the American Planning Association* **62**(2), pp. 243-258.

[5] J. Knorn, A. Rabe, V. C. Radeloff, T. Kuemmerle, J. Kozak, P. Hostert, 2009. Land Cover Mapping of Large Areas Using Chain Classification of Neighboring Landsat Satellite Images. *Remote Sensing of Environment* **113**, pp. 957-964.

[6] G. Chander, B. L. Markham, D. L. Helder, 2009. Summary of Current Radiometric Calibration Coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI Sensors. *Remote Sensing of Environment* **113**(5), pp. 893-903.

[7] D. E. Herwindiati, M. A. Djauhari, L. Jaupi, 2012. Robust Statistics for Classification of Remote Sensing Data. In Proceedings of International Conference on Computational Statistics, pp. 317-328.

[8] B. Efron, D. Rogosa, R. Tibshirani, 2004. Resampling Methods of Estimation. In N.J. Smelser, & P.B. Baltes (Eds.). *International Encyclopedia of the Social & Behavioral Sciences*, pp. 13216-13220. New York: Elsevier.

[9] B. Efron, R. Tibshirani, 1998. *An Introduction to the Bootstrap*. Chapman & Hall.

[10] Department of the Interior U.S. Geological Survey, 2018. *Landsat 7 Data Users Handbook*. Available online at https://landsat.usgs.gov/sites/default/files/documents/LSDS-1927_L7_Data_Users_Handbook.pdf.

[11] C. Bishop, 2006. *Pattern Recognition and Machine Learning*. Springer.

[12] D. E. Herwindiati, J. Hendryli, L. Hiryanto, 2017. Impervious Surface Mapping Using Robust Depth Minimum Vector Variance Regression. *European Journal of Sustainable Development* **6** (3), pp. 29-39.

[13] F. Henry, D. E. Herwindiati, S. Mulyono, J. Hendryli, 2017. Sugarcane Land Classification with Satellite Imagery using Logistic Regression Model. *Materials Science and Engineering* **185**.

[14] J. Cohen, 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **20** (1), pp. 37-46.