Relationship between Cover-Management Factor and Spatial Data

Ting-Chun Lin[1], Jhe-Syuan Lai[2], Fuan Tsai[3] and Walter W. Chen[4]

[1,2,3] Center for Space and Remote Sensing Research, National Central University,
No.300, Jhongda Rd., Jhongli City, Taoyuan County 32001, Taiwan
[4] Department of Civil Engineering, National Taipei University of Technology,
No.1, Sec. 3, Zhongxiao E. Rd., Taipei 10608, Taiwan
suaygiho@hotmail.com

KEY WORDS: Cover-Management Factor, Vegetation Cover Factor, Universal Soil Loss Equation (USLE), Data Mining, Spatial Analysis

ABSTRACT: The cover-management factor (C-factor) is an important factor in Universal Soil Loss Equation (USLE). A commonly used method for C-factors estimation is to assign values to land use classes obtained from USLE guide tables. However, land use maps area seldom updated in most area. This study constructs a relationship between C-factor and spatial data with a data mining algorithm. The study area is Shimen reservoir watershed in Taiwan. The SPOT images, 10m DEM, a land use map and some spatial data were used in this research. The SPOT images were processed with a topographic correction with Minnaert constant. After that, Gray Level Co-occurrence Matrix (GLCM) was used to build some statistical feature indicators for data mining. Decision tree and random forest classifier were used to build the data mining model. In this step, the species simplification and separation were considered to optimize the result. The overall accuracy of model construction is about 75% with a kappa value of 0.75. The overall agreement of C-factor estimating base on the data mining model is about 60%.

1. INTRODUCTION

Universal Soil Loss Equation (USLE) is a solution to predict the soil loss amount caused by erosion processes in Taiwan Soil and Water Conservation Specification (Lin and Lin, 2008). Among the soil erosion risk factors of USLE, C-factor is related to the land use type with its commonly used values 0.001 to1. It will cause a thousandfold difference due to this widely value range. Thus, C-factor estimation is an important issue for soil erosion study, and environmental protection. C-factor look up table with land use map is a common and easy way to estimate the C-factor (Moore and Wilson, 1992). With the advantage of remote sensing, people tried to use the regression between vegetation indices and C-factor to calculate the soil erosion amount (Alexandridis et al., 2015). However, these methods have some limitations in data acquisitions or updating; they may also result in unreasonable estimation. Therefore, this study tries to construct a relationship between C-factor and spatial data with a data mining algorithm. Decision tree is one of the most powerful and popular approaches in knowledge discovery and data mining (Rokach and Maimon, 2014). Random forest develops an ensemble of decision trees and the final classification is obtained by combining results from the trees by voting (Breiman and Cutler, 2001). Therefore, this study tries to explore large and complex bodies of spatial data in order to discover useful patterns, constructing a relationship between C-factor and spatial data with a data mining algorithm.

2. MATERIALS AND METHODOLOGY

This study estimates the C-factor from 2004 to 2008 in a watershed. The study procedure base on a data mining algorithm is shown as following:

(a) A C-factor look-up table and land use information were used to calculate a C-factor map as decision attribute.
(b) Satellite images after topographic correction were used to estimate the vegetation indices and textures information. Both of them and some spatial data were the conditional attributes of the data mining.
(c) Data mining model construction was processed by decision attributes and conditional attributes described above.
(d) After constructing the data mining model using training data, spatial information were inserted into the model to get the integral C-factor estimation.

2.1 Study Area and Materials

The study area is shown in Figure 1, which is the Shimen reservoir watershed in Taiwan. SPOT 4 and SPOT 5 images, 10m DEM and other spatial data were used in this research. SPOT images were used to compute the vegetation indices and texture information. A 10 meter resolution DEM was used to produce the height and slope attributes. Geology, road, steam soil and land use information were also considered into this study.

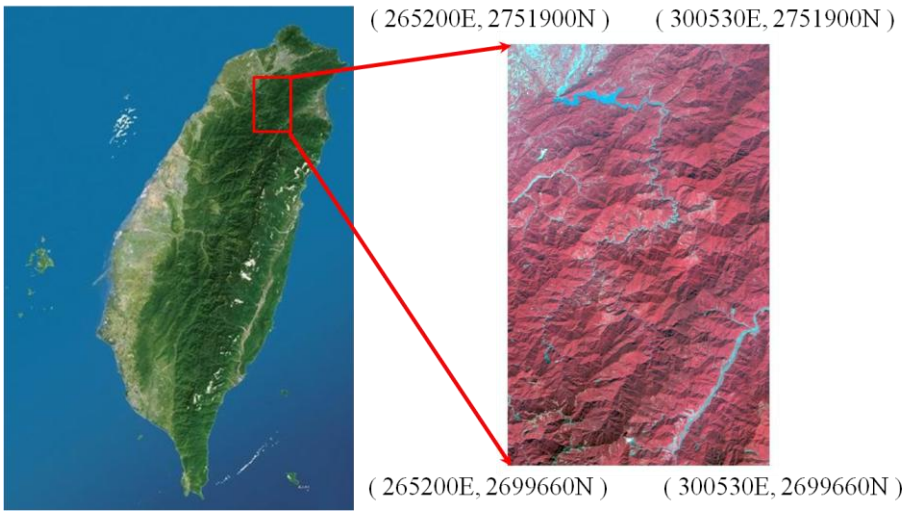

Figure 1. study area

2.2 C-factor Estimation

After adjusting the topographic effect with Minnaert correction, this study inserts spatial information, consisting of elevation, slope, road, steam, geology, soil, vegetation indices and texture information, into data mining algorithm as conditional attributes. NDVI (Normalized Difference Vegetation Index) and SAVI (Soil Adjusted Vegetation Index) are used in this study. These two commonly used indices could be calculated as Eq. (1) and Eq. (2). Here *NIR* means the reflectance value of near infrared band; *RED* means the reflectance value of red band; *L* is the soil correction factor, 0.5 is its commonly used value. GLCM (Gray Level Co-occurrence Matrix) was also employed to produce the texture information with its statistical indicators. Some commonly used indicators like angular second moment, contrast, homogeneity, entropy and mean were considered in this step. On the other hand, C-factor, the decision attribute of data mining in this study, is estimated according to a conversion from land use map with USLE look up table publish by Jhan (2014). The data mining factors and its land use type are shown as Table 3. In this table, *M* is the size of co-occurrence matrix; *P* is the marginal-probability matrix.

$$NDVI = \frac{NIR-RED}{NIR+RED} \quad (1)$$

$$SAVI = \frac{NIR-RED}{NIR+RED+L}(1+L) \quad (2)$$

Decision Tree was used in the preliminary data mining phase to rank the influential conditional attributes for the following processing. Besides waterbody (C=0), there are 11 data mining factors, shown in the Table 1, considered in the decision tree algorithm. After evaluating the precision and recall of the data mining results, data mining factors will be adjusted for the optimized calculation. Only influential conditional attributes will be included in an improved data mining with Random Forest algorithm.

Table 1. preliminary data mining factors

data mining factors		main land use types
1	C=0.005	cultural facility, weir, dam and reservoir
2	C=0.01	forest and residential area
3	C=0.025	dike, ditch, aqueduct, shoal, beach and wetlands
4	C=0.03	road
5	C=0.035	unused land
6	C=0.05	cutting woods, grassland and bush
7	C=0.133	animal farms
8	C=0.156	wasteland
9	C=0.16	orchard
10	C=0.208	dryland
11	C=1	fire breaking, reef and bare space

3. RESULTS AND DISCUSSIONS

In the preliminary data mining phase, the overall accuracy is about 56.181% with a kappa value of 0.518. The average of precision and recall are just about 0.5. Especially the factor C=0.03, C=0.16 and C=0.208 were not so good. According to the preliminary results, the confusing situation between C=0.16 and C=0.208 would lead a bad accuracy. To optimize the data mining algorithm, this study combines these two data mining factors together and defines a new factor by averaging these two C-factors. On the other hand, the land uses type of factor C=0.03 only contains road related usage. Therefore, this study takes this factor as an extra consideration. After the data mining optimization described above, there are only 9 factors considered in the following process. Table 2 shows the results of Random Forest in 2004 as an example. The overall accuracy is increased to 78.520% with a kappa value of 0.758. Both precision and recall are better than the preliminary results. According to the same analysis process, Table 3 shows the overall agreement between the look-up table approach and the data mining generated C-factor during 2004 to 2008, which is about 60.513% to 62.392%.

Table 2. data mining results of Random Forest (2004)

data mining factors		precision	recall
1	C=0.005	1	0.897
2	C=0.01	0.657	0.697
3	C=0.025	0.821	0.920
4	C=0.035	0.786	0.815
5	C=0.05	0.808	0.600
6	C=0.133	0.833	1
7	C=0.156	1	1
8	C=0.184	0.605	0.639
9	C=1	0.679	0.633
average		0.788	0.785

Table 3. the overall agreements of C-factor estimation

	2004	**2005**	**2006**	**2007**	**2008**
overall agreements	61.096%	62.392%	60.513%	61.299%	60.552%

4. CONCLUSIONS

This study employed data mining to analyze the C-factor of USLE in the Shimen reservoir watershed in Taiwan. Minnaert Correction was used to correct the topographic effect for the vegetation indices estimation and texture indicators construction. Beside these information, elevation, slope, geology, soil, road and river data were also considered as conditional attributes. According to the analysis results, the overall agreement between the look-up table of USLE and the data mining generated C-factor is better than 60%. This suggests that about 40% of the C-factor values should be changed from the land-use map look-up table. In the future work, this study will follow the same processing to estimate the C-factor maps. After that, erosion soil loss amount can be calculated for Shimen reservoir watershed sediment and dredging study.

5. REFERENCES

Alexandridis, T. K., Sotiropoulou, A. M., Bilas, G., Karapetsas, N., and Silleos, N. G., 2015. The Effects of Seasonality in Estimating the C-Factor of Soil Erosion Studies, Land Degradation & Development, Vol. 26, No. 6, pp. 596-603.

Breiman, L., Cutler, A., Liaw, A., and Wiener, M., 2001. Breiman and Cutler's Random Forest for Classification and Regression. R package version 4.5–16. Available: http://CRAN.R-project.org/package=randomForest. Accessed 2010 September 18.

Jhan, Y. K., 2014. Analysis of Soil Erosion of Shihmen Reservoir Watershed, Department of Civil Engineering at the National Taipei University of Technology (Master's thesis)

Lin, W. Y. and Lin, L. L., 2008. Application and Discussion of Soil Erosion Prediction Model. Journal of Soil and Water Conservation, Volume 40, Issue 3, pp. 357-368.

Moore, I. D. and Wilson, J. P., 1992. Length-Slope Factors for the Revised Universal Soil Loss Equation: Simplified Method of Estimation, Journal of Soil and Water Conservation, Vol. 47, No. 5, pp. 423-428.

Rokach, L., Maimon, O., 2008. Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing, Singapore.