International Journal of Geo-Information

MDPI

Article

Determining Cover Management Factor with Remote Sensing and Spatial Analysis for Improving Long-Term Soil Loss Estimation in Watersheds

Fuan Tsai [1,2], Jhe-Syuan Lai [3], Kieu Anh Nguyen [4] and Walter Chen [4,*]

[1] Center for Space and Remote Sensing Research, National Central University, Zhongli City, Taoyuan 320, Taiwan; ftsai@csrsr.ncu.edu.tw
[2] Department of Civil Engineering, National Central University, Zhongli City, Taoyuan 320, Taiwan
[3] Department of Civil Engineering, Feng Chia University, Taichung 40724, Taiwan; jslai@fcu.edu.tw
[4] Department of Civil Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; t106429401@ntut.edu.tw
* Correspondence: waltchen@ntut.edu.tw; Tel.: +886-2-27712171 (ext. 2628)

Abstract: The universal soil loss equation (USLE) is a widely used empirical model for estimating soil loss. Among the USLE model factors, the cover management factor (C-factor) is a critical factor that substantially impacts the estimation result. Assigning C-factor values according to a land-use/land-cover (LULC) map from field surveys is a typical traditional approach. However, this approach may have limitations caused by the difficulty and cost in conducting field surveys and updating the LULC map regularly, thus significantly affecting the feasibility of multi-temporal analysis of soil erosion. To address this issue, this study uses data mining to build a random forest (RF) model between eight geospatial factors and the C-factor for the Shihmen Reservoir watershed in northern Taiwan for multi-temporal estimation of soil loss. The eight geospatial factors were collected or derived from remotely sensed images taken in 2004, a digital elevation model, and related digital maps. Due to the memory size limitation of the R software, only 4% of the total data points (population dataset) in each C-factor class were selected as the sample dataset (input dataset) for analysis using the stratified random sampling method. Seventy percent of the input dataset was used to train the RF model, and the other 30% was used to test the model. The results show that the RF model could capture the trend of vegetation recovery and soil loss reduction after the destructive event of Typhoon Aere in 2004 for multi-temporal analysis. Although the RF model was biased by the majority class's large sample size (C = 0.01 class), the estimated soil erosion rate was close to the measurement obtained by the erosion pins installed in the watershed (90.6 t/ha-year). After the model's completion, we furthered our aim to address the input dataset's imbalanced data problem to improve the model's classification performance. An ad-hoc down-sampling of the majority class technique was used to reduce the majority class's sampling rate to 2%, 1%, and 0.5% while keeping the other minority classes at a 4% sample rate. The results show an improvement of the Kappa coefficient from 0.574 to 0.732, the AUC from 0.780 to 0.891, and the true positive rate of all minority classes combined from 0.43 to 0.70. However, the overall accuracy decreases from 0.952 to 0.846, and the true positive rate of the majority class declines from 0.99 to 0.94. The best average C-factor was achieved when the sampling rate of the majority class was 1%. On the other hand, the best soil erosion estimate was obtained when the sampling rate was 2%.

Keywords: universal soil loss equation; USLE; NDVI; SAVI; multi-temporal remote sensing; spatial data mining; C-factor; random forest

Citation: Tsai, F.; Lai, J.-S.; Nguyen, K.A.; Chen, W. Determining Cover Management Factor with Remote Sensing and Spatial Analysis for Improving Long-Term Soil Loss Estimation in Watersheds. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 19. https://doi.org/10.3390/ijgi10010019

Received: 20 November 2020
Accepted: 2 January 2021
Published: 6 January 2021

1. Introduction

Severe soil erosion will increase soil sedimentation and severely reduce the water storage and supply capabilities of reservoirs. A previous analysis revealed that 30% of US croplands had excessive soil erosion rates [1]. Soil erosion has been one of the core topics in agriculture, natural resources conservation, and other related fields since the end of the 1920s [2]. A significant trend in soil erosion study was developing various measurements and prediction models for different locations and applications, such as the AGNPS (agricultural non-point source pollution model [3]), CREAMS (chemicals, runoff and erosion from agricultural management systems [4]), EPIC (erosion-productivity impact calculator [5]), SWRRBWQ (simulator for water resources in rural basins-water quality [6]), WEPP (water erosion prediction project [7]), and USLE (universal soil loss equation [8,9]). Despite being the earliest developed model, USLE is still a widely used empirical model and has been used at the national and international levels to estimate soil loss throughout the world [10].

There are six crucial factors in the USLE model, including the rainfall erosivity factor (R_m-factor), soil erodibility factor (K_m-factor), slope length factor (L-factor), slope steepness factor (S-factor), cover management factor (C-factor), and support practice factor (P-factor). The outcome of the USLE model represents the average annual soil loss (A_m). However, the USLE calculation excludes landslides, gully erosions, riverbed or bank erosions, and sediment depositions [11–13]. Among the USLE factors, the C-factor (ranging from 0 to 1) is related to the land-use/land-cover (LULC). Thus, it may cause a thousand-fold difference in soil erosion estimation (0.001 vs. 1).

A summary of past studies of soil erosion in a Taiwanese watershed showed that the calculated soil erosion rate varies from 1 to 3310 t/ha-year [14]. Another study also demonstrated that the non-uniform distribution of the USLE factors might cause a substantial discrepancy in soil erosion estimation [15]. Therefore, it is necessary to develop a strategy to derive more consistent and reliable factors when applying the USLE model for soil loss evaluation.

This study focused on the assessment of the C-factor. Traditionally, the C-factor was determined based on plot experiments or in-situ investigations [9]. However, such works were too time-consuming and uneconomical to apply everywhere [16]. Renard et al. [17] derived a function to simulate the long-term field experiments by using the prior-land-use sub-factor, canopy-cover sub-factor, surface-cover sub-factor, surface-roughness sub-factor, and soil-moisture sub-factor to evaluate the C-factor. Although the derived function improved the efficiency of the long-term field experiments, significant field works were still unavoidable. Nowadays, connecting the C-factor with the LULC map using a look-up table is an effective strategy [18]. However, this strategy is limited by the update period of the map, and therefore evaluating the multi-temporal soil erosion is difficult. Similarly, instead of the LULC map, a time-series result can be obtained using satellite remote sensing images and supervised classification techniques [19,20]. The limitations here are that ground truth points are needed to assess the accuracy of the image classification, and the results are inferior to the LULC map produced from fieldwork. Performing regression analyses to connect features derived from remotely sensed images (e.g., normalized difference vegetation index, NDVI) with the C-factor is another common approach [21,22]. Again, the developed relationships may have a large margin of error, fail to provide any physical meanings, and may be sensitive to vegetation phenology and soil conditions [23]. According to the comparison above, Table 1 summarizes the typical approaches for estimating the C-factor.

Table 1. Comparison of typical approaches for estimating the C-factor.

	Approach	Method	Advantage	Challenge
1	LULC map by field survey	1. Conduct field surveys to generate a LULC map 2. Convert the LULC classes to C-factors by a look-up table	Most accurate LULC data	Time-consuming with a lengthy update period
2	LULC map by remote sensing	1. Use remote sensing images to generate a LULC map 2. Convert the LULC classes to C-factors by a look-up table	Multi-temporal evaluation	Ground truthing needed to ensure accuracy
3	Empirical NDVI equation	1. Establish an empirical equation between NDVI values and C-factors 2. Use the equation to convert NDVI values to C-factors	Simple and multi-temporal evaluation	Sensitive to vegetation phenology and soil conditions
4	This study	1. Use the official LULC map and a look-up table to convert the LULC classes to C-factor classes 2. Use geospatial factors (including NDVI and SAVI) to train an RF model to predict C-factors	More accurate model and multi-temporal evaluation	More complicated

To improve upon previous research on this topic, the goal of this study was to construct a model between the C-factor and related geospatial data (including but not limited to those derived from remotely sensed images) using a data mining approach. If a credible model can be constructed from official data collected in earlier years, then the model can be applied to predict the C-factors for later years by merely updating the remote sensing images, thus achieving multi-temporal analysis without having to update the LULC map every year. Furthermore, using this novel framework to assess and predict the C-factors based on geospatial factors, we will be able to better estimate the rate of soil erosion at a large scale, such as the watershed scale. As shown in Table 1, we used the official LULC map and a look-up table to determine the C-factors in this study. An RF model was then built to predict the C-factor values from the geospatial factors for use in the USLE model.

The rest of the paper is organized as follows. Section 2 introduces the study area, the geospatial data, and the data mining algorithm. Section 3 presents the C-factor modeling and soil erosion estimation results and a discussion focusing on the class imbalance (imbalanced data) problem. Finally, Section 4 concludes this paper and provides possible future research directions.

2. Methods

To address the research goal, it is necessary to develop a strategy to derive more consistent and reliable factors when applying the USLE model to evaluate soil loss. With the rapid development of geospatial technologies and data availability, there is a great potential to improve the USLE modeling based on geospatial data, such as remotely sensed images and geographic information system (GIS) data layers, to obtain a more accurate soil loss estimation at the regional scale, especially for long-term and multi-temporal studies. The data types, data processing, and data mining analysis are described in the following sub-sections.

2.1. Study Area

A mountainous area of 760 km² of the Shihmen Reservoir watershed in northern Taiwan was selected as the study area (Figure 1). The elevation of the study area ranges approximately from 250 to 3500 m above sea level, measured from a 10-m digital elevation model (DEM). The terrain increases steeply from north to south, and steep slopes are ubiquitous in the watershed. The average annual precipitation is approximately 2500 mm. There are 13 geological formations and six soil types in the study area [24]. The major land-cover is forest (both natural and artificial) and there is little agricultural activity.

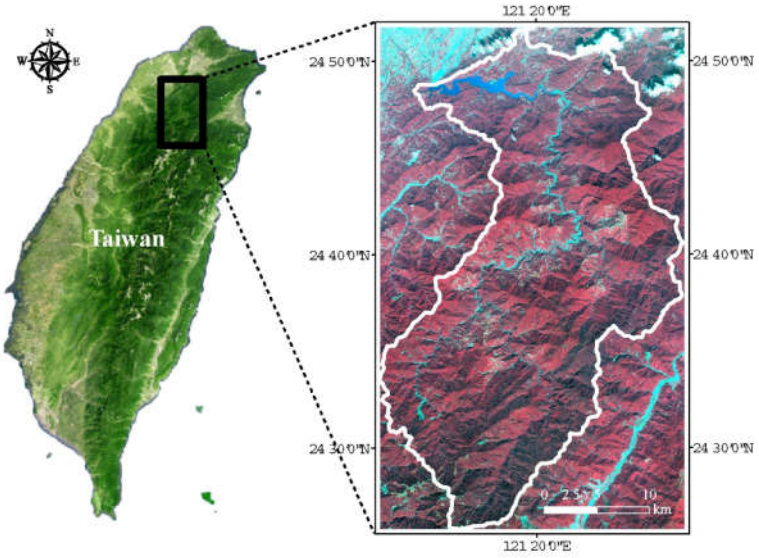


Figure 1. The location of the study area and a corresponding SPOT false-color satellite image.

The Shihmen Reservoir watershed is one of the major reservoirs in Taiwan, providing drinking water to more than three million people in three northern cities. However, heavy rainfall induced by typhoons, such as Typhoon Aere in August 2004, may generate a large amount of debris and driftwood, resulting in a water supply shortage and causing various water resource management problems. Hence, a long-term land cover monitoring project from 2004 to 2009 was implemented using remote sensing technologies to support water resource and water supply management [25]. Based on the collected geospatial data, this study further explored the effectiveness of the USLE model by modeling the C-factor in the Shihmen Reservoir watershed to estimate the multi-temporal soil erosion rates.

2.2. Materials and Data Preprocessing

To connect the geospatial data to the C-factor, a total of eight attributes (as listed in Table 2) were considered in the modeling. They are elevation, slope, NDVI, soil adjusted vegetation index (SAVI), shortest distance to roads, shortest distance to rivers, geology, and soil type. In contrast to studies using many attributes, we used only eight attributes because research has shown that attribute reduction could improve the model performance [26]. We also used fewer attributes because of the need to maximize the number of data points that can be processed by the R software (see Section 2.3). The purpose of preprocessing is three-fold, including data collection, labeling, and feature derivation. Some of the derived features could be obtained by spatial analysis from the original data. For example, the elevation and slope information could be derived from the DEM. As for the multi-temporal satellite images listed in Table 3, this study generated the NDVI and SAVI indices year by year for modeling purposes. Equations (1) and (2) show the NDVI and SAVI equations, where *NIR* and *RED* represent the radiance or reflectance of near-infrared and red bands, respectively, and *L* indicates the soil correction factor that is commonly set

to 0.5 [27]. However, the mountainous topography might distort the spectral responses so that the same land cover types located on phototropic and apheliotropic areas might have significant variations. To reduce the topographic effect, this study applied a typical Minnaert correction [28] to rectify the spectral responses based on a non-Lambertian assumption before deriving the vegetative features. Similar to the use of DEM and SPOT images, the rest of the data, such as the distance information to roads and rivers, were generated by spatial analysis. Additionally, the vector data were rasterized to 10 × 10 m to be compatible with one another.

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

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

where *NIR* and *RED* are the radiance or reflectance of near-infrared and red bands, and L is the soil correction factor (0.5).

Table 2. Types of geospatial data that were used in this study.

Original Data	Derived Data	Note
DEM	Elevation (numeric) Slope (numeric)	Cell size: 10 m
SPOT satellite images	NDVI (numeric) SAVI (numeric)	See Table 3 for details
Road map	Distance to road (numeric)	Measurement unit: meter
Stream map	Distance to river (numeric)	
Geology map	Geological formation (categorical)	1. Alluvium 2. Hsitsun Formation 3. Kueichulin Formation 4. Mushan Formation 5. Nanchuang Formation 6. Nankang Formation 7. Paling Formation 8. Peiliao Formation 9. Shihti Formation 10. Szeleng Sandstone 11. Taliao Formation 12. Talu Shale 13. Terrace Deposits
Soil map	Soil type (categorical)	1. Silt/silty loam 2. Loam 3. Loamy fine sand/coarse sandy loam/sandy loam 4. Very fine sand/loamy very fine sand 5. Clay 6. Fine sand/loamy sand/loamy coarse sand 7. No data

Table 3. SPOT satellite images that were used in this study.

Sensor	Date	Spatial Resolution (m)	Spectral Bandwidth (um)	Radiometric Resolution (bits)
SPOT 5	2004/10/12 2006/07/19 2008/09/21	10	Green: 0.50–0.59 Red: 0.61–0.68 NIR: 0.79–0.89	8
SPOT 4	2005/07/25 2007/07/19	20		

After feature derivation and assembly, a look-up table was used to assign C-factor values to the official 2004 LULC map (the only map available during the study period). The C-factor's point values were based on the research of Jhan [29] and Lin [30], which in turn were based on the design manual of the Soil and Water Conservation Bureau of Taiwan. Without conducting numerous experiments in the field to determine the C-factors at various locations, the look-up table provides the next best option to assign credible C-factor values to different LULC classes in the study area. As can be seen from Figure 2, there were 23 land use classes assigned to 12 different C-factor classes. The higher the C-factor, the less the land cover and the higher the soil erosion. The correspondence between the LULC classes and the C-factor classes are summarized in Table 4. The geospatial data and the corresponding C-factor values of each grid cell in the study area were extracted and assembled as an analytic dataset (herein referred to as the population dataset) used by the data mining algorithm (after sampling) to create a C-factor model. The model is used for multi-temporal analysis of C-factor change and soil erosion.

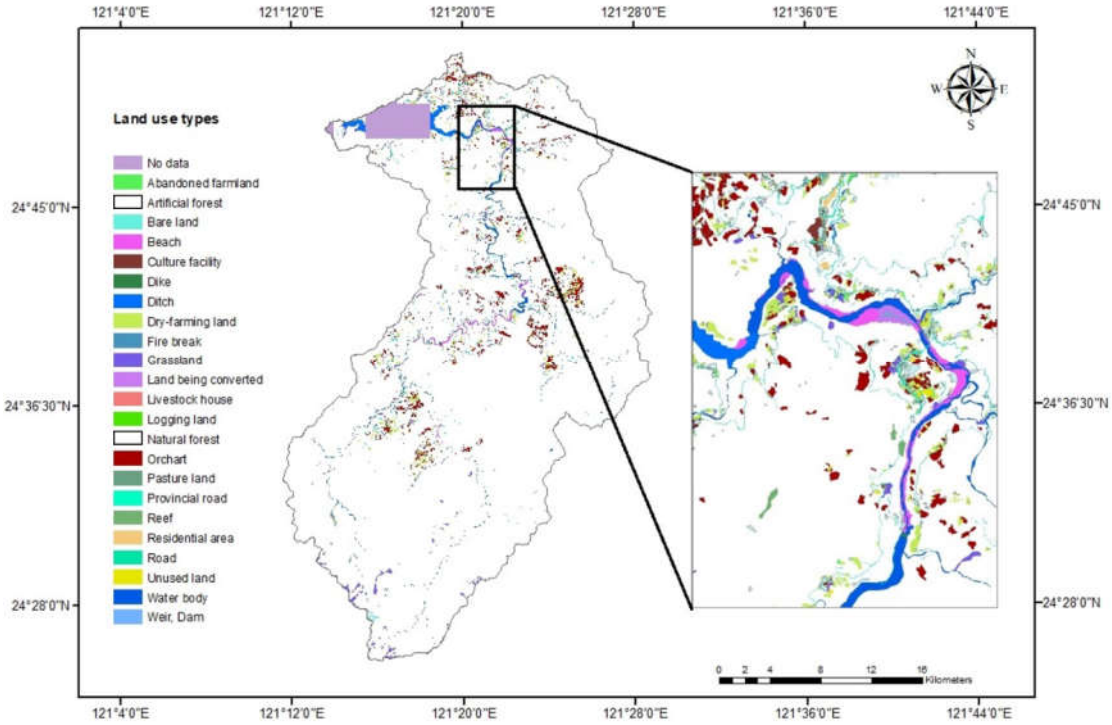


Figure 2. LULC classes of the study area (forests are labeled in white for contrasting other LULC classes).

Table 4. LULC classes merged by C-factor classes (revised from Jhan [29] and Lin [30]).

C-Factor Class	Land Use/Land Cover Class
0	Waterbody, reef
0.005	Railway related facility, weir, dam, cultural facility
0.01	Natural forest, artificial forest, residential area
0.025	Dike, ditch, aqueduct, shoal, beach, wetland
0.03	Provincial road, road
0.035	Unused land, land being converted
0.05	Pastureland, logging land, grassland, shrubland
0.133	Livestock house
0.156	Abandoned farmland
0.16	Orchard
0.208	Dry-farming land
1	Fire break, bare land

2.3. Data Mining Analysis

To extract useful, unknown, and potential information from the vast dataset, data mining is an efficient approach [31,32]. Random Forests (RF), one of the popular data mining algorithms proposed by Breiman [33], has excellent performance in analyzing many complicated remote sensing issues [34–37]. The procedure for applying the data mining algorithm to construct a C-factor model for soil erosion estimation is illustrated in Figure 3. We used the geospatial data derived from GIS and SPOT images and the official 2004 LULC map (from field surveys) to construct the C-factor model using the randomForest() package of the R software. Among the data mining procedures, the RF algorithm is a supervised approach that adopts multiple decision trees (DT), bootstrap aggregation (bagging), and internal cross-validation techniques. It integrates all tree-based results into the best model for analysis [33]. Recent years have seen increased attention being given to the RF algorithm in the geo-informatics domain. Belgiu and Dragut [34] reviewed its application and future direction in remote sensing. Nevertheless, constructing the C-factor model based on the RF algorithm and geospatial data has rarely been conducted until now. The primary benefit of the RF algorithm is that it can avoid the over-fitting issue to improve prediction accuracy [38]. The RF algorithm employs measures such as the Gini index, information gain (IG), or entropy to evaluate the degree of impurity of discrete or numeric input data. The smaller the Gini index of an attribute, the higher the priority should be selected to construct a conditional node and ignore the other attributes. The RF algorithm performs numerous iterations and randomly divides the training dataset (in terms of the number of data and the number of attributes) into many subsets to build many trees and generate better results than the DT method. The detailed steps can be seen in Guo et al. [39].

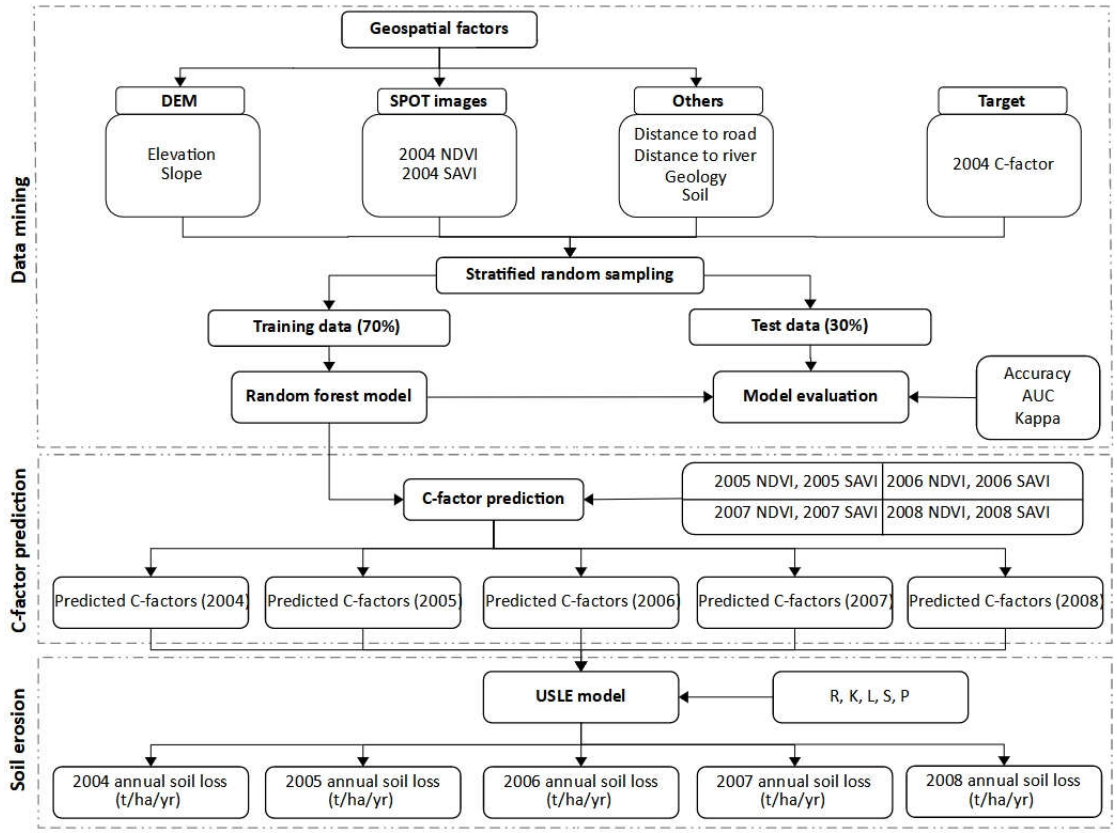


Figure 3. Research steps of this study.

Owing to the memory size limitation of R, only about 4% of the population dataset can be imported and used in the modeling (303,682 points out of 7,592,062 points). Therefore, we used the stratified random sampling method to select the same percentage of data points from each of the 12 C-factor classes (herein referred to as the sample dataset or the input dataset), as shown in Table 5. Note that the input dataset is highly unbalanced, with 92.5% of the data from the C = 0.01 class (forests). The percentage of the rest of the C-factor classes ranges from close to 0% to 2.9%.

Table 5. The percentage composition of C-factor classes in the input datasets under different majority class sampling rates (C = 0.01 class).

No.	C-factor Class	Total no. of Points (Population)	4% of the Majority Class	2% of the Majority Class	1% of the Majority Class	0.5% of the Majority Class
			% of the Total Sample (Input Dataset)			
1	0	216,970	8679 (2.9%)	8679 (5.3%)	8679 (9.3%)	8679 (15.0%)
2	0.005	1110	44 (0.0%)	44 (0.0%)	44 (0.0%)	44 (0.1%)
3	0.01	7,021,560	280,862 (92.5%)	140,431 (86.0%)	70,216 (75.5%)	35,108 (60.6%)
4	0.025	47,629	1905 (0.6%)	1905 (1.2%)	1905 (2.0%)	1905 (3.3%)
5	0.03	37,714	1509 (0.5%)	1509 (0.9%)	1509 (1.6%)	1509 (2.6%)
6	0.035	4235	169 (0.1%)	169 (0.1%)	169 (0.2%)	169 (0.3%)
7	0.05	46,598	1864 (0.6%)	1864 (1.1%)	1864 (2.0%)	1864 (3.2%)
8	0.133	73	3 (0.0%)	3 (0.0%)	3 (0.0%)	3 (0.0%)
9	0.156	342	14 (0.0%)	14 (0.0%)	14 (0.0%)	14 (0.0%)
10	0.16	141,672	5667 (1.9%)	5667 (3.5%)	5667 (6.1%)	5667 (9.8%)
11	0.208	60,060	2402 (0.8%)	2402 (1.5%)	2402 (2.6%)	2402 (4.1%)

12	1	14,099	564 (0.2%)	564 (0.3%)	564 (0.6%)	564 (1.0%)
	Total	7,592,062	100.0%	100.0%	100.0%	100.0%

After sampling, 70% of the input dataset was used as the training data and inputted into the RF algorithm to create the C-factor model. The remaining 30% was the test data and used to assess the performance of the model. Once the model was validated, it was applied to the population dataset to generate the C-factor maps from 2004 to 2008. Finally, the C-factor and the other USLE factors were combined in the USLE model to calculate the soil erosion rates from 2004 to 2008 (see Section 2.4 for details).

Categorical and numerical data are two major geospatial data types. The randomForest() package of R uses the Gini index to split nodes in order to reduce impurity at each node [40]. We used 1000 trees in this study, and three variables were tried at each split. The Gini index of dataset D is defined in Equation (3), where m is the number of categories, n_i is the number of data points in the ith category, and N is the total number of data points. If a binary split was performed on attribute A, the Gini index given the split was defined in Equation (4), where D_1 and D_2 are the datasets after the split [41].

$$Gini(D) = 1 - \sum_{i=1}^{m} \left(\frac{n_i}{N}\right)^2 \quad (3)$$

$$Gini_A(D) = \frac{n_1}{N} Gini(D_1) + \frac{n_2}{N} Gini(D_2) \quad (4)$$

To verify the data mining-based C-factor model, this study calculated the confusion matrix (or error matrix), overall accuracy (*OA*), Kappa coefficient, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve for the quantitative evaluation. Overall accuracy refers to the percentage of correctly classified samples as shown in Equation (5), where M represents the element in the confusion matrix; M_{total} is the sum of M; M_{diag} is the sum of M on the diagonal line; Nc is the number of labels; and i and j are the row and column indices. The *Kappa* coefficient is shown in Equation (6), which reflects the reliability of the modeling results. When the *Kappa* coefficient is close to 1, it shows excellent agreement between prediction and observation. By contrast, the results are worse than random assignment when a negative Kappa value appears.

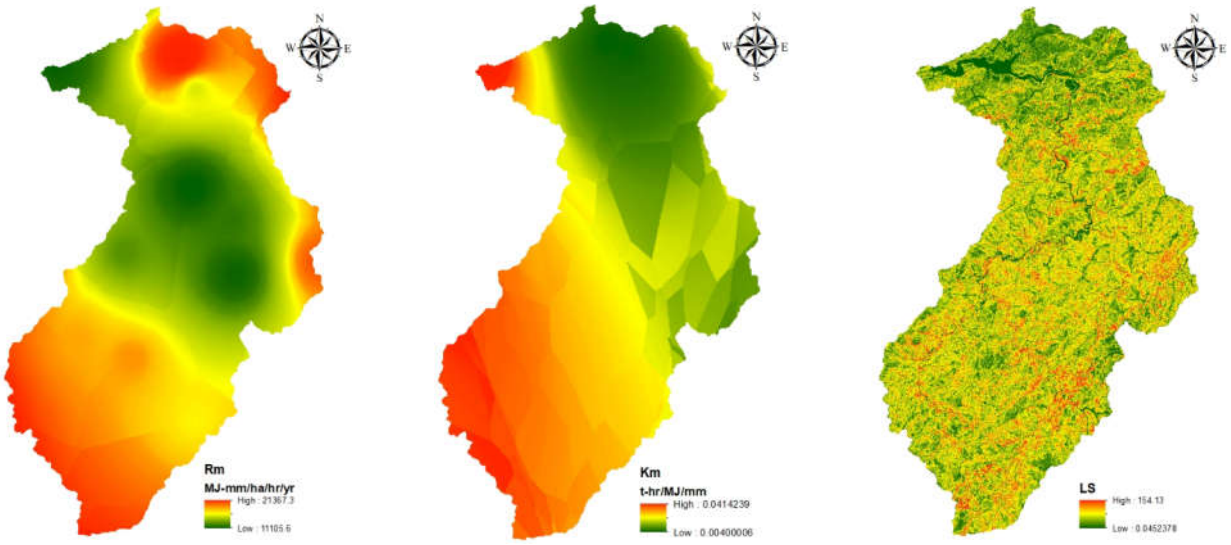
$$OA = \frac{\sum_{i=1}^{N_c} M_{ii}}{\sum_{i=1}^{N_c} \sum_{j=1}^{N_c} M_{ij}} = \frac{M_{diag}}{M_{total}} \quad (5)$$

$$Kappa = \frac{M_{total} M_{diag} - \sum_{i=1}^{N_c} (M_{+i} M_{i+})}{M_{total}^2 - \sum_{i=1}^{N_c} (M_{+i} M_{i+})} \quad (6)$$

2.4. USLE Computation

The USLE equation is shown in Equation (7), and the meanings of all USLE factors are listed in Table 6. Although the goal of the study is to evaluate the C-factor, other USLE factors are also needed to estimate the annual soil loss. For these factors, this study followed the investigations of Chen et al. [14] and Liu et al. [42] to produce the rainfall erosivity, soil erodibility, slope length, and slope steepness layers, respectively. We also assumed that the support practice factor is 1. The generated R_m- and K_m- factor distribution maps are shown in Figure 4a,b, and the L- and S-factors are combined as a topographic factor (LS-factor) shown in Figure 4c.

$$A_m = R_m \times K_m \times L \times S \times C \times P \quad (7)$$



(**a**) R_m-factor (**b**) K_m-factor (**c**) LS-factor

Figure 4. The distribution maps of R_m-, K_m-, and LS-factors.

Table 6. The USLE factors and their definitions and units.

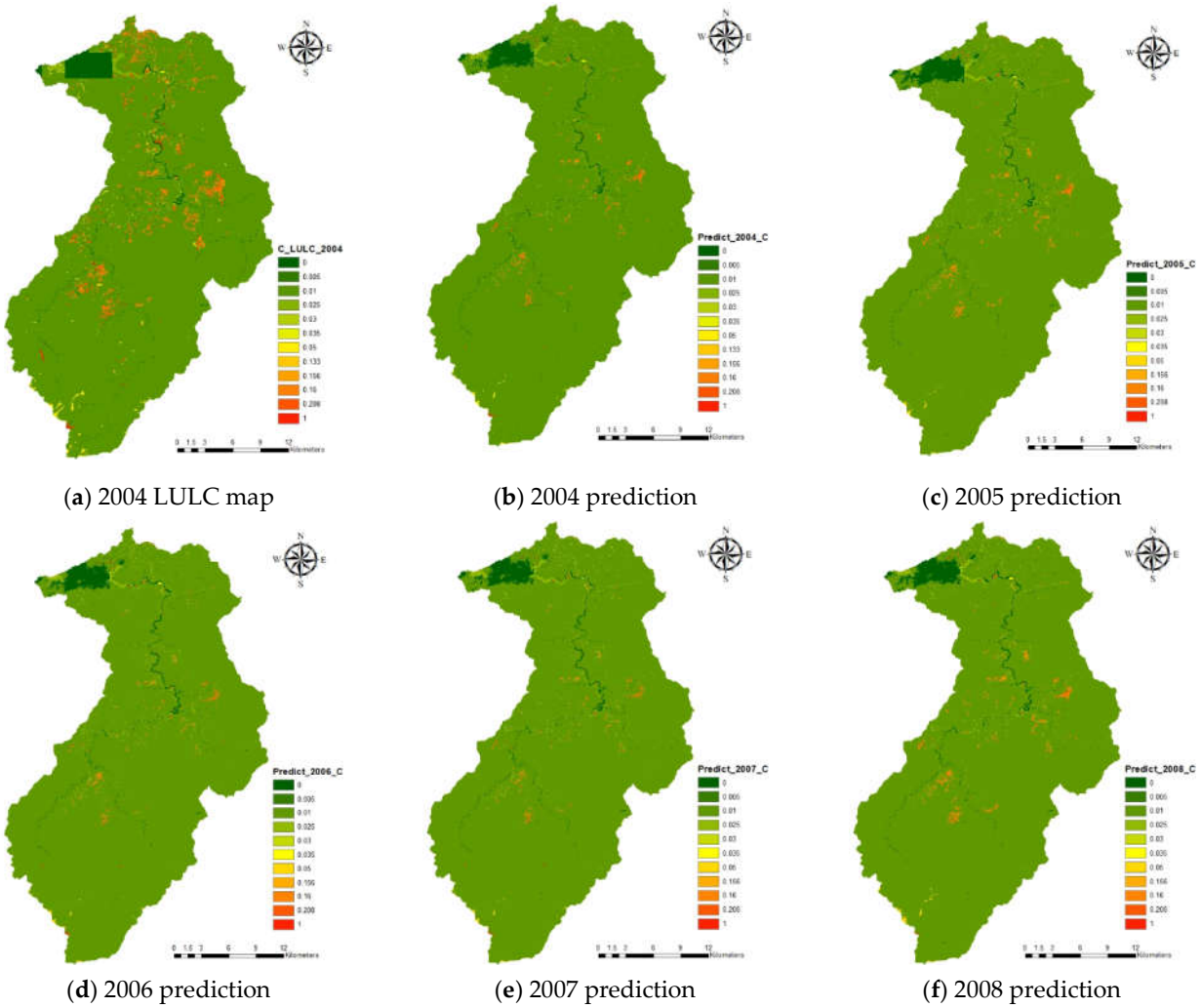
Symbol	Definition	Unit
A_m	Average annual soil loss	t/ha-year
R_m	Rainfall erosivity factor	MJ-mm/ha-hour-year
K_m	Soil erodibility factor	t-hour/MJ-mm
L	Slope length factor	--
S	Slope steepness factor	--
C	Cover management factor	--
P	Support practice factor	--

3. Results and Discussion

The results of this study are presented as tables, graphs, and statistical metrics in the following sub-sections.

3.1. Cover-Management Factor Modeling

According to the data preprocessing and study procedures described in previous sections, the C-factor RF model was constructed using the training data and tested with the test data. The C-factors were from the official 2004 LULC map (the only map available during the study period), and the results are shown in Table 7. Unlike a previous study [30], which only used at most 100 points from each C-factor class and substantially overestimated soil erosion, we tried to maximize the number of data points that could be processed to train the RF model given the memory size limitation of R. Consequently, 4% of the total data points (population dataset) were used (303,682 points selected out of 7,592,062 points). The training dataset result shows that OA = 1, Kappa = 1, and AUC = 1 (the confusion matrix is not shown here to avoid redundancy). This indicates that the RF model can correctly distinguish all 212,578 data points in the training dataset. By contrast, Table 7 shows that the result of the test dataset has less remarkable metrics. While OA is still very high (0.9516), Kappa is only 0.5741, and AUC is 0.7804. Using this RF model, we predicted the C-factor distribution maps from 2004 to 2008 using SPOT images. The resulting maps are shown in Figure 5b–f.



(**a**) 2004 LULC map (**b**) 2004 prediction (**c**) 2005 prediction

(**d**) 2006 prediction (**e**) 2007 prediction (**f**) 2008 prediction

Figure 5. The C-factor distribution maps of the (**a**) 2004 LULC map, (**b**) 2004 prediction, (**c**) 2005 prediction, (**d**) 2006 prediction, (**e**) 2007 prediction, and (**f**) 2008 prediction (4% sampling rate of the majority class).

Table 7. The confusion matrix of the test data (4% sampling rate of the majority class).

Actual	Predicted											
	0	**0.005**	**0.01**	**0.025**	**0.03**	**0.035**	**0.05**	**0.133**	**0.156**	**0.16**	**0.208**	**1**
0	1792	0	732	75	1	0	1	0	0	3	0	0
0.005	0	3	8	0	0	0	2	0	0	0	0	0
0.01	249	0	83,729	7	130	1	19	0	0	101	19	4
0.025	109	0	89	372	0	0	1	0	0	0	0	0
0.03	2	0	99	0	349	0	0	0	0	2	1	0
0.035	2	0	32	3	3	8	0	0	0	2	1	0
0.05	13	0	441	4	8	0	85	0	0	7	1	0
0.133	0	0	1	0	0	0	0	0	0	0	0	0
0.156	0	0	4	0	0	0	0	0	0	0	0	0
0.16	6	0	1400	0	24	0	0	0	0	251	19	0
0.208	5	0	559	0	17	1	0	0	0	58	81	0
1	1	0	133	7	2	0	0	0	0	1	0	25
Overall statistics	OA = 0.952				Kappa = 0.574				AUC = 0.780			

Figure 5a shows the true C-factor distribution from the official 2004 LULC map. The similarity between the prediction (Figure 5b–f) and the reference C-factor is evident. The red pixels (high C-factor values) of all maps cluster along the central river valley near the center and lower portions of the watershed. However, some of the C-factor cannot be distinguished reliably, and some omission and commission errors have occurred. As shown in column 4 (shaded) of Table 7, the most noticeable trend is a prediction bias towards the C = 0.01 class, representing natural and artificial forests. This result reflects the overwhelming majority of the C = 0.01 class in the sample (92.5%) when building the RF model (Table 5). Hence, the RF model tends to classify pixels into the C = 0.01 class. Because of this bias towards a low C-factor value (0.01), the average of 2004 C-factors predicted by the RF model is only 0.0115 (Table 8), which is lower than the true average of 0.0164 (official 2004 LULC map). Likewise, the predicted average C-factors are also lower in the subsequent years from 2005 to 2008.

Table 8. Comparison of C-factors between the official LULC map and the model predictions under different sampling rates of the majority class (C = 0.01 class).

	Official LULC Map (Mean)	**RF Model Prediction (Mean)**				
	2004	**2004**	**2005**	**2006**	**2007**	**2008**
4%	0.0164	0.0115	0.0114	0.0110	0.0110	0.0115
2%		0.0130	0.0132	0.0125	0.0124	0.0133
1%		0.0156	0.0164	0.0149	0.0146	0.0169
0.5%		0.0115	0.0114	0.0110	0.0109	0.0115

To reduce the classification error, we experimented with an ad-hoc down-sampling of the majority class technique that used only 2% data from the majority class (C = 0.01 class) while maintaining a 4% sample rate of the other minority C-factor classes. The percentages of data points from each of the 12 C-factor classes in the input dataset were previously shown in Table 5. Again, the result of the training dataset shows a perfect classification of OA = 1, Kappa = 1, and AUC = 1 (again, the confusion matrix is not shown here to avoid redundancy). The result of the test dataset (Table 9) shows that OA = 0.9230, Kappa = 0.6484, and AUC = 0.7807. Compared with Table 7, we can see that the OA decreases from 0.9516 to 0.9230, the Kappa increases from 0.5741 to 0.6484, and the AUC remains about the same. Using only 2% data from the C = 0.01 class, the predicted average C-factors from 2004 to 2008 range from 0.0124 to 0.0133 (Table 8), higher than the 4% case and closer to the true average C-factor value. In other words, the reduction of the sampling rate from 4% to 2% increases the Kappa coefficient at the expense of OA. Simultaneously, the reduction of the sampling rate also brings the predictions closer to the reference value (ground truth). Although we had adopted the stratified random sampling method to obtain a representative sample of all C-factor classes, it did not completely avoid the imbalanced data problem. The next section will present how the sampling rate of the majority class affects soil erosion estimates.

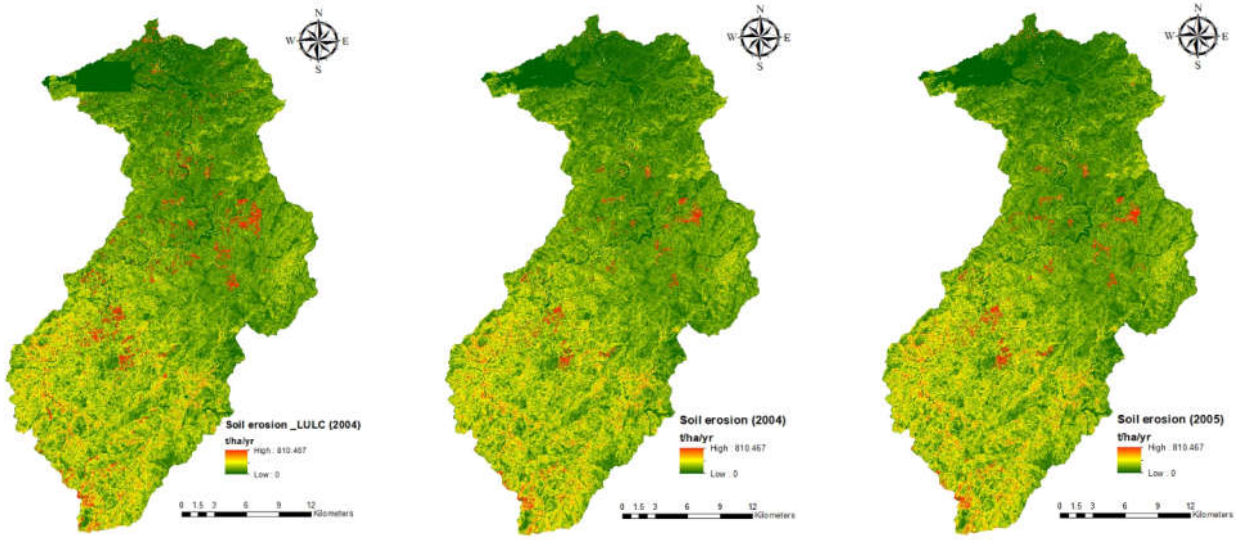
Table 9. The confusion matrix of the test data (2% sampling rate of the majority class).

Actual	**Predicted**											
	0	**0.005**	**0.01**	**0.025**	**0.03**	**0.035**	**0.05**	**0.133**	**0.156**	**0.16**	**0.208**	**1**
0	2034	0	494	62	0	0	3	0	0	4	6	1
0.005	0	6	5	0	0	0	2	0	0	0	0	0
0.01	234	0	41,551	4	83	1	21	0	0	188	41	6
0.025	113	0	85	371	1	0	1	0	0	0	0	0
0.03	1	0	37	0	401	4	0	0	0	9	1	0
0.035	1	0	20	1	1	15	0	0	0	8	5	0

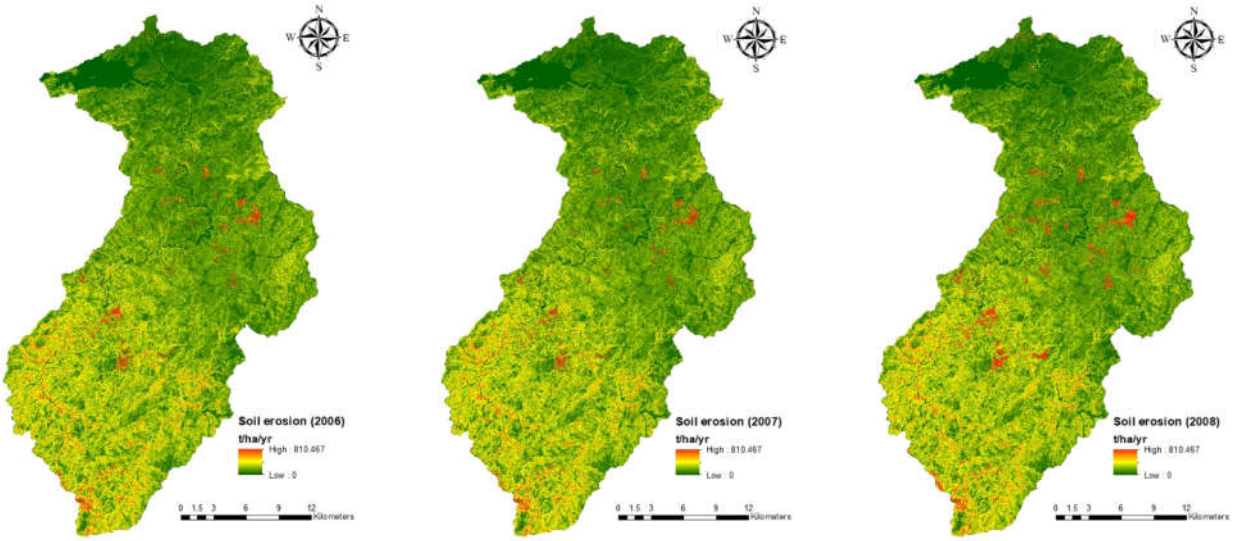
0.05	10	1	402	5	2	0	124	0	0	11	4	0
0.133	0	0	1	0	0	0	0	0	0	0	0	0
0.156	0	0	4	0	0	0	0	0	0	0	0	0
0.16	7	0	1095	0	24	0	0	0	0	540	34	0
0.208	9	0	440	0	19	0	0	0	0	123	130	0
1	2	0	130	2	0	0	0	0	0	2	1	32
Overall statistics	OA = 0.923				Kappa = 0.648				AUC = 0.781			

3.2. Soil Erosion Estimation

Up until 2017, the USLE model was the only method for estimating the amount of soil loss in the technical regulations for soil and water conservation in Taiwan [43]. Combining the R_m-, K_m-, and LS-factor layers (Figure 4) with the C-factor layer (Figure 5), the USLE model was used in this study to estimate the amount of soil erosion. The result based on the official 2004 LULC map is listed in the second column of Table 10 (116.3 t/ha-year). The multi-temporal evaluations from 2004 to 2008 based on the RF models (4% and 2%) are shown in columns 3–7 of the same table year by year. Comparing the results for 2004 (columns 2 and 3), the RF models generate lower than expected (true) soil erosion rates (88.2 and 95.1 vs. 116.3 t/ha-year). Similarly, the soil erosion rates are lower in the subsequent years from 2005 to 2008. Using the 2% RF model, we prepared the predicted soil erosion maps in Figure 6b–f for the years 2004 through to 2008, while the soil erosion map based on the official 2004 LULC map is shown in Figure 6a. As such, a high resemblance between the predictions (Figure 6b–f) and the reference value (Figure 6a) is evident. The red pixels (high soil erosion rates) of all maps cluster near the center and lower portions of the watershed, indicating good modeling results. The results of the 4% RF model are similar, but we did not include them here to avoid redundancy.



(**a**) 2004 LULC map (**b**) 2004 prediction (**c**) 2005 prediction



(**d**) 2006 prediction (**e**) 2007 prediction (**f**) 2008 prediction

Figure 6. The soil erosion maps of the (**a**) 2004 LULC map, (**b**) 2004 prediction, (**c**) 2005 prediction, (**d**) 2006 prediction, (**e**) 2007 prediction, and (**f**) 2008 prediction (2% sampling rate of the majority class).

Table 10. The estimation of annual soil loss based on C-factors of different years (4% and 2% sampling rates).

	Official LULC Map (2004)	RF Model Prediction						Erosion Pins (2008–2011)
		2004	2005	2006	2007	2008		
A_m (t/ha-year)	116.3	88.2	85.6	84.7	84.0	84.5	(4% of C = 0.01)	90.6
		95.1	94.4	92.4	91.6	93.2	(2% of C = 0.01)	

Chen et al. [14] compiled a table of the calculated amounts of soil erosion of the Shihmen Reservoir watershed from previous studies. The table shows that soil erosion ranges from 1 to 3310 t/ha-year. Our results are close to the lower end of the soil erosion range. By contrast, the only study using a similar methodology to this study [30], which related nine decision factors, including the gray level co-occurrence matrix (GLCM) to the C-factor, only used at most 100 points from each C-factor class. The study achieved a Kappa coefficient of 0.758, but estimated the soil erosion to be 359.4–629.9 t/ha-year.

Based on the erosion pins installed in the Shihmen Reservoir watershed [44] and the measurements collected from 8 September 2008 to 10 October 2011, the soil erosion depth ranged from 2.17 to 13.03 mm/year [37]. The average erosion depth is 6.5 mm/year, which is equivalent to 90.6 t/ha-year if the unit weight of soil is assumed to be 1.4 t/m^3 [42]. Thus, our results are more comparable with the erosion pin measurements. Specifically, it shows that the 4% sampling rate underestimates soil erosion, whereas the 2% sampling rate overestimates soil erosion.

It is worth noting that the long-term land cover and landslide monitoring project [25] from 2004 to 2009 indicated that only typhoon Aere in 2004 induced significant land degradation and mass movement in the period. Large amounts of debris and driftwood flowed into the Shihmen Reservoir. The high turbidity in the water caused the water distribution system to be shut down for an unprecedented 18 days. A similar situation has not happened since. The removal of land cover in the watershed is the reason why the calculated soil erosion based on the official 2004 LULC map is as high as 116.3 t/ha-year.

After Typhoon Aere, as the land stabilized and vegetation re-grew to provide new ground cover, soil erosion reduced substantially. This explains why the measured soil erosion is only 90.6 t/ha-year between 2008 and 2011 (Table 10).

According to Table 10, both of the 4% and 2% C-factor models predict a peak of soil erosion in 2004, followed by a gradual decrease until a small rebound in 2008. This is consistent with the vegetation recovery in the study area from 2005 to 2007 and the hike in rainfall brought by Typhoons Kalmaegi, Sinlaku, and Jangmi in 2008 [45]. However, the differences in soil erosion rates are not as marked as expected from year to year. Nevertheless, these results confirm that it is possible to develop a multi-temporal data mining model for the C-factor and the corresponding soil erosion.

3.3. Discussion

Although this study's results demonstrated the feasibility of constructing a data mining model for the USLE C-factor, the modeling results were affected by the majority class's sampling rate (C = 0.01 class). At the beginning of the study, we used the stratified random sampling method (instead of the simple random sampling method) to ensure that each class (strata) of the population dataset (total data points) were represented. It helped to avoid incorrect analysis, but it did not solve the class imbalance problem entirely. So, we experimented with the down-sampling of the majority class (C = 0.01 class) to 2% and kept the other minority classes at a 4% sample rate. A better result was achieved as indicated by the higher Kappa coefficient (but at the expense of lower OA). To investigate if the classification performance can be further improved, we applied an ad-hoc down-sampling of the majority class technique (similar to random under-sampling) to the population dataset with 1% and 0.5% sampling rates. The resulting percentage compositions of each of the 12 C-factor classes in the input dataset were previously shown in Table 5. After building corresponding RF models, the overall results are summarized in Table 11, which shows the OA, the Kappa coefficient, the AUC, the true positive rate of all minority classes combined, the true positive rate of the majority class, the average C-factor of the LULC map, the predicted average C-factor of 2004, the predicted soil erosion rate of 2004, the predicted soil erosion rate of 2008, and the measured soil erosion rate by the erosion pins.

Table 11. Comparison of typical metrics under different sampling rates of the majority class (C = 0.01 class).

Sampling Rate of the Majority Class	OA	Kappa	AUC	True Positive Rate of All Minority Classes Combined	True Positive Rate of the Majority Class	Average C-factor of LULC Map	Average C-factor (2004)	Soil Erosion of 2004 (t/ha-Year)	Soil Erosion of 2008 (t/ha-Year)	Erosion Pins (2008–2011)
4%	0.952	0.574	0.780	0.43	0.99	0.0164	0.0115	88.2	84.5	90.6
2%	0.923	0.648	0.781	0.53	0.99		0.0130	95.1	93.2	
1%	0.880	0.687	0.845	0.61	0.97		0.0156	104.3	108.0	
0.5%	0.846	0.732	0.891	0.70	0.94		0.0115	88.2	84.5	

It was found that the down-sampling strategy works well. With the decrease in the majority class sampling rate, the Kappa coefficient increases from 0.574 to 0.732, and the AUC increases from 0.780 to 0.891. Moreover, the true positive rate of all minority classes combined also increases from 0.43 to 0.70. However, the overall accuracy decreases with the down-sampling from 0.952 to 0.846, and the true positive rate of the majority class declines from 0.99 to 0.94.

At first, it appears that 0.5% is the best sampling rate in this study, but the average C-factor indicates otherwise. The C-factor value starts from 0.0115 at 4% and then becomes 0.0130 at 2% and 0.0156 at 1%, gradually approaching the reference value of 0.0164. However, the average C-factor suddenly dips back down to 0.0115 at 0.5%, deviating from the

reference value again. Therefore, judging from the average C-factor, 1% is the majority class's best sampling rate.

The unexpected result of the average C-factor is further investigated in Figure 7, in which the predicted percentage compositions of four different sampling rates were plotted for all minority classes. The red line indicates the "true" percentage composition from the 2004 official LULC map. As shown in the figure, the difference in the C = 0.16 class between the LULC map and the different sampling rates determines how accurate the final average C-factor is. As the sampling rate decreases from 4% to 2% and 1%, the respective percentage of the C = 0.16 class approaches and then surpasses that of the LULC map. When the sampling rate is further reduced to 0.5%, the percentage of the C = 0.16 class reverses course and drops back to the same level as that of the 4% sampling rate. This explains why the 0.5% sampling rate did not yield a better average C-factor, even though its other metrics (such as OA, Kappa, and AUC) were superior.

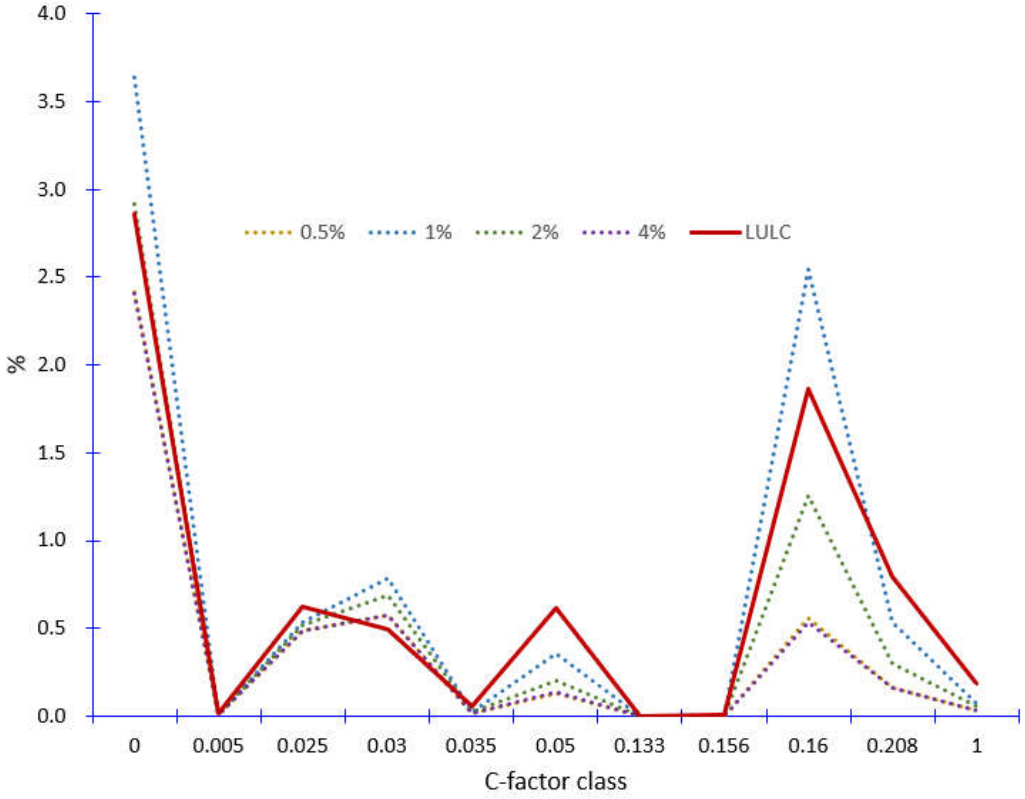


Figure 7. The model predicted percentage compositions of the minority classes in the 2004 population dataset under different sampling rates of the majority class. The red line indicates the true percentage composition from the 2004 official LULC map.

Finally, if we compare the predicted soil erosion rates with the rate measured by the erosion pins, we arrive at yet another different conclusion. The 2% sampling rate predicts a soil erosion rate of 93.2 t/ha-year, which is the closest to the measured rate of 90.6 t/ha-year. In this scenario, 2% is the best sampling rate. The best sampling rates under different criteria are shaded in Table 11 for easy comparison.

The results presented above suggest that, apart from the class imbalance problem, other factors are responsible for this study's modeling performance. Using the overall evaluation metrics (such as OA, Kappa, and AUC) in this study is not entirely appropriate and could be misleading. Since our goal was a two-step approach to model the C-factor

and eventually the soil erosion, the predicted soil erosion rate is the most important indicator. A more balanced dataset does not always yield a better modeling result in terms of soil loss, and we consider 2% to be the best sampling rate in this study.

4. Conclusions

Unlike previous studies, this research developed C-factor models based on data mining techniques to improve the soil erosion assessment in the Shihmen Reservoir watershed in northern Taiwan. Eight geospatial data were selected and used in the modeling. The multi-temporal vegetative indices (NDVI and SAVI) derived from multispectral satellite images were rectified by a topographic correction to reduce variations over time and better characterize the surface radiances of the same targets. The C-factor models built using the RF-based data mining algorithm were used with USLE to estimate the spatiotemporal soil losses from 2004 to 2008. The results were compared against past studies and the measurements of erosion pins. They showed promising classification performance.

It is found that the soil erosion rate in 2004 was the highest because of the unprecedented destruction of Typhoon Aere in 2004. As the vegetation in the watershed re-grew after the typhoon to provide new ground cover, the soil erosion rate decreased steadily until 2008, when a surge of rainfall occurred due to Typhoons Kalmaegi, Sinlaku, and Jangmi. This trend was successfully captured by the RF models, which demonstrates the feasibility of the multi-temporal analysis. Furthermore, using an ad-hoc down-sampling of the majority class technique (at 2% sampling rate), the soil erosion rate was predicted to be 93.2 t/ha-year, very close to the 90.6 t/ha-year measured by the erosion pins installed in the watershed.

In addition, this study provides a case of an imbalanced data problem that differs from other imbalanced data problems in that a more balanced dataset does not always yield a better modeling result. The best sampling rate of the majority class based on different metrics are summarized as follows:

1. Overall accuracy and true positive rate of the majority class: 4%
2. Kappa coefficient, AUC, and true positive rate of all minority classes combined: 0.5%
3. Average C-factor: 1%
4. Soil erosion rate: 2%

In summary, the results show that the proposed novel framework for assessing and predicting C-factors and soil erosion based on geospatial factors is both viable and practical. The method also has promising classification performance even when faced with an imbalanced data problem. An imbalanced data problem cannot be easily eradicated by removing records from the majority class. Therefore, future research is necessary to improve model performance and soil erosion estimates.

Author Contributions: Conceptualization, Fuan Tsai; Data curation, Fuan Tsai and Jhe-Syuan Lai; Formal analysis, Jhe-Syuan Lai, Kieu Anh Nguyen and Walter Chen; Funding acquisition, Fuan Tsai; Investigation, Jhe-Syuan Lai, Kieu Anh Nguyen and Walter Chen; Methodology, Fuan Tsai and Walter Chen; Project administration, Fuan Tsai; Resources, Fuan Tsai; Software, Jhe-Syuan Lai and Kieu Anh Nguyen; Supervision, Fuan Tsai and Walter Chen; Validation, Walter Chen; Visualization, Kieu Anh Nguyen; Writing—original draft, Fuan Tsai, Jhe-Syuan Lai and Walter Chen; Writing—review & editing, Fuan Tsai, Jhe-Syuan Lai, Kieu Anh Nguyen and Walter Chen. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported, in part, by the Ministry of Science and Technology of Taiwan (ROC) under project No. NSC 100-2628-E-008-014-MY3, MOST 108-2221-E-008-001-MY3, and MOST 109-2121-M-027-001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. National Research Council. *Soil Conservation: Assessing the National Resources Inventory*; National Academy Press: Washington, DC, USA, 1986.
2. Laflen, J.M.; Moldenhauer, W.C. *Pioneering Soil Erosion Prediction—The USLE Story*; World Association of Soil and Water Conservation: Beijing, China, 2003.
3. Young, R.A.; Onstad, C.A.; Bosch, D.D.; Anderson, W.P. *Agricultural Non-Point Source Pollution Model, Version 4.03. AGNPS User's Guide*; North Central Soil Conservation Research Laboratory: Morris, MN, USA, 1994.
4. Knisel, W.G. *CREAMS: A Field Scale Model for Chemicals, Runoff, and Erosion from Agricultural Management Systems*; U.S. Department of Agriculture, Science and Education Administration: Washington, DC, USA, 1980.
5. Sharpley, A.N.; Williams, J.R. *EPIC Erosion/Productivity Impact Calculator: 1. Model Documentation*; U.S. Department of Agriculture, Technical Bulletin: Washington, DC, USA, 1990; Volume 1768, p. 235.
6. Arnold, J.G.; Williams, J.R.; Nicks, A.D.; Sammons, N.B. *SWRRB: A Basin Scale Simulation Model for Soil and Water Resources Management*; Texas A&M University Press: College Station, TX, USA, 1990.
7. Nearing, M.A.; Foster, G.R.; Lane, L.J.; Finkner, S.C. A process-based soil erosion model for USDA-Water Erosion Prediction Project technology. *Trans. ASAE* **1989**, *32*, 1587–1593.
8. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall-Erosion Losses from Cropland East of the Rocky Mountains*; U.S. Department of Agriculture, Agriculture Handbook: Washington, DC, USA, 1965; Volume 282, p. 49.
9. Wischmeier, W.H.; Smith, D.D. *Predicting Rainfall Erosion Losses—A Guide to Conservation Planning*; U.S. Department of Agriculture, Agriculture Handbook: Washington, DC, USA, 1978; Volume 537, p. 67.
10. Sillanpää, M. *Micronutrient Assessment at the Country Level: An International Study*; FAO Soils Bulletin: Rome, Italy, 1990; Volume 63, p. 214.
11. Foster, G.R.; Wischmeier, W. Evaluating irregular slopes for soil loss prediction. *Trans. ASAE* **1974**, *17*, 305–309.
12. Griffin, M.L.; Beasley, D.B.; Fletcher, J.J.; Foster, G.R. Estimating soil loss on topographically non-uniform field and farm units. *J. Soil Water Conserv.* **1988**, *43*, 326–331.
13. Renard, K.G.; Foster, G.R.; Weesies, G.A.; Porter, J.P. RUSLE: Revised Universal Soil Loss Equation. *J. Soil Water Conserv.* **1991**, *46*, 30–33.
14. Chen, W.; Li, D.-H.; Yang, K.-J.; Tsai, F.; Seeboonruang, U. Identifying and comparing relatively high soil erosion sites with four DEMs. *Ecol. Eng.* **2018**, *120*, 449–463.
15. Li, J.-Y.; Yang, K.-J.; Chen, W. Approximation Equation of Erosion Calculation in GIS. In Proceedings of the 37th Asian Conference on Remote Sensing, Colombo, Sri Lanka, 17–21 October 2016; Volume 3, pp. 1837–1843.
16. Gabriels, D.; Ghekiere, G.; Schiettecatte, W.; Rottiers, I. Assessment of USLE cover-management C-factors for 40 crop rotation systems on arable farms in the Kemmelbeek watershed, Belgium. *Soil Tillage Res.* **2003**, *74*, 47–53.
17. Renard, K.G.; Foster, G.R.; Weesies, G.A.; McCool, D.K.; Yoder, D.C. *Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation (RUSLE)*; U.S. Department of Agriculture, Agriculture Handbook: Washington, DC, USA, 1997; Volume 703, p. 404.
18. Borrelli, P.; Märker, M.; Panagos, P.; Schütt, B. Modeling soil erosion and river sediment yield for an intermountain drainage basin of the Central Apennines, Italy. *Catena* **2014**, *114*, 45–58.
19. Dabral, P.P.; Baithuri, N.; Pandey, A. Soil erosion assessment in a hilly catchment of north eastern India using USLE, GIS and Remote Sensing. *Water Resour. Manag.* **2008**, *22*, 1783–1798.
20. Pandey, A.; Chowdary, V.M.; Mal, B.C. Identification of critical erosion prone areas in the small agricultural watershed using USLE, GIS and remote sensing. *Water Resour. Manag.* **2007**, *21*, 729–746.
21. Alexandridis, T.K.; Sotiropoulou, A.M.; Bilas, G.; Karapetsas, N.; Silleos, N.G. The effects of seasonality in estimating the C-factor of soil erosion studies. *Land Degrad. Dev.* **2015**, *26*, 596–603.
22. Van der Knijff, J.M.; Jones, R.J.A.; Montanarella, L. *Soil Erosion Risk Assessment in Italy*; European Soil Bureau: Brussels, Belgium, 1999; p. 58.
23. Ayalew, D.A.; Deumlich, D.; Šarapatka, B.; Doktor, D. Quantifying the Sensitivity of NDVI-Based C Factor Estimation and Potential Soil Erosion Prediction using Spaceborne Earth Observation Data. *Remote Sens.* **2020**, *12*, 1136.
24. Tsai, F.; Lai, J.-S.; Chen, W.W.; Lin, T.-H. Analysis of topographic and vegetative factors with data mining for landslide verification. *Ecol. Eng.* **2013**, *61*, 669–677.
25. Tsai, F.; Chen, L.C. Long-term landcover monitoring and disaster assessment in the Shihmen reservoir watershed using satellite images. In Proceedings of the 13th CeRES International Symposium on Remote Sensing, Chiba, Japan, 29 October 2007.
26. Lai, J.-S.; Tsai, F.; Lin, T.-Y.; Chen, W. Verification and susceptibility assessment for landslides using data mining techniques. *J. Photogramm. Remote Sens.* **2013**, *17*, 149–160. (In Chinese with English Abstract)
27. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309.
28. Minnaert, M. The reciprocity principle in lunar photometry. *Astrophys. J.* **1941**, *93*, 403–410.
29. Jhan, Y.K. Analysis of Soil Erosion of Shihmen Reservoir Watershed. Master's Thesis, National Taipei University of Technology, Taipei, Taiwan, 2014. (In Chinese with English Abstract)

30. Lin, T.-C. Establishment of Relationship between USLE Cover Management Factor and Spatial Data. Master's Thesis, National Central University, Zhongli City, Taoyuan, Taiwan, 2016. (In Chinese with English Abstract)
31. Ouyang, Y.; Luo, S.M.; Cui, L.H.; Wang, Q.; Zhang, J.E. Estimation of real-time N load in surface water using dynamic data-driven application system. *Ecol. Eng.* **2011**, *37*, 616–621.
32. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann: San Francisco, CA, USA, 2011.
33. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
34. Belgiu, M.; Dragut, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31.
35. Stumpf, A.; Kerle, N. Object-oriented mapping of landslides using Random Forests. *Remote Sens. Environ.* **2011**, *115*, 2564–2577.
36. Tsai, F.; Lai, J.-S.; Lu, Y.-H. Full-waveform LiDAR point cloud land cover classification with volumetric texture measures. *Terr. Atmos. Ocean. Sci.* **2016**, *27*, 549–563.
37. Nguyen, K.A.; Chen, W.; Lin, B.-S.; Seeboonruang, U.; Thomas, K. Predicting sheet and rill erosion of Shihmen reservoir watershed in Taiwan using machine learning. *Sustainability* **2019**, *11*, 3615.
38. Ismail, R.; Mutanga, O.; Kumar, L. Modeling the potential distribution of pine forests susceptible to Sirex Noctilio infestations in Mpumalang, South African. *Trans. GIS* **2010**, *14*, 709–726.
39. Guo, L.; Chehata, N.; Mallet, C.; Boukir, S. Relevance of airborne LiDAR and multispectral image data for urban scene classification using random forests. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 56–66.
40. Zhang, F.; Yang, X. Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection. *Remote Sens. Environ.* **2020**, *251*, 112105.
41. Han, J.; Kamber, M.; Pei, J. *Data Mining, Concepts and Techniques*, 3rd ed.; Morgan Kaufmann: Waltham, MA, USA, 2012.
42. Liu, Y.-H.; Li, D.-H.; Chen, W.; Lin, B.-S.; Seeboonruang, U.; Tsai, F. Soil erosion modeling and comparison using slope units and grid cells in Shihmen reservoir watershed in Northern Taiwan. *Water* **2018**, *10*, 1387.
43. Lin, W.Y.; Lin, L.L. Application and discussion of soil erosion prediction model. *J. Soil Water Conserv.* **2008**, *40*, 357–368. (In Chinese with English Abstract)
44. Lin, B.S.; Thomas, K.; Chen, C.K.; Ho, H.C. Evaluation of soil erosion risk for watershed management in Shenmu watershed, central Taiwan using USLE model parameters. *Paddy Water Environ.* **2008**, *14*, 19–43.
45. Li, D.-H. Analyzing Soil Erosion of Shihmen Reservoir Watershed Using Slope Units. Master's Thesis, National Taipei University of Technology, Taipei, Taiwan, 2017. (In Chinese with English Abstract)