environments

MDPI

Article

Enhancing Cover Management Factor Classification Through Imbalanced Data Resolution

Kieu Anh Nguyen and Walter Chen *

Department of Civil Engineering, National Taipei University of Technology, Taipei 10608, Taiwan; rosenguyen@ntut.edu.tw

* Correspondence: waltchen@ntut.edu.tw; Tel.: +886-(2)-27712171 (ext. 2628)

Abstract: This study addresses the persistent challenge of class imbalance in land use and land cover (LULC) classification within the Shihmen Reservoir watershed in Taiwan, where LULC is used to map the Cover Management factor (C-factor). The dominance of forests in the LULC categories leads to an imbalanced dataset, resulting in poor prediction performance for minority classes when using machine learning techniques. To overcome this limitation, we applied the Synthetic Minority Over-sampling Technique (SMOTE) and the 90-model SMOTE-variants package in Python to balance the dataset. Due to the multi-class nature of the data and memory constraints, 42 models were successfully used to create a balanced dataset, which was then integrated with a Random Forest algorithm for C-factor classification. The results show a marked improvement in model accuracy across most SMOTE variants, with the Selected Synthetic Minority Over-sampling Technique (Selected_SMOTE) emerging as the best-performing method, achieving an overall accuracy of 0.9524 and a sensitivity of 0.6892. Importantly, the previously observed issue of poor minority class prediction was resolved using the balanced dataset. This study provides a robust solution to the class imbalance issue in C-factor classification, demonstrating the effectiveness of SMOTE variants and the Random Forest algorithm in improving model performance and addressing imbalanced class distributions. The success of Selected_SMOTE underscores the potential of balanced datasets in enhancing machine learning outcomes, particularly in datasets dominated by a majority class. Additionally, by addressing imbalance in LULC classification, this research contributes to Sustainable Development Goal 15, which focuses on the protection, restoration, and sustainable use of terrestrial ecosystems.

Keywords: SMOTE; soil erosion; cover management factor; Random Forest; imbalanced data

check for updates

Citation: Nguyen, K.A.; Chen, W. Enhancing Cover Management Factor Classification Through Imbalanced Data Resolution. *Environments* **2024**, *11*, 250. https://doi.org/10.3390/environments11110250

Academic Editor: Brian D. Fath

Received: 10 October 2024
Revised: 8 November 2024
Accepted: 9 November 2024
Published: 12 November 2024

1. Introduction

Soil erosion, the process by which soil is displaced by natural forces, presents a significant challenge to agricultural productivity [1]. It contributes to land degradation, diminished soil fertility, and the pollution of air and water, with water erosion of particular concern [2]. Human activities such as deforestation and inadequate land management exacerbate this issue, making it an urgent environmental concern. The relationship between soil erosion and land use and land cover (LULC) is crucial, as changes in land cover—such as urbanization and agricultural expansion—can accelerate soil erosion. Forests and natural vegetation serve as protective barriers, stabilizing soil with their root systems, while alterations to these covers increase the risk of erosion. Understanding this interaction is essential for designing effective land management strategies that mitigate environmental degradation and promote sustainable land use practices.

Numerous studies have explored the link between soil erosion and LULC. Liu et al. [3] demonstrated that the grid cell method was more accurate in predicting soil erosion in Taiwan's Shihmen Reservoir watershed, an area subject to high rainfall erosivity. Similarly, Chen et al. [4] emphasized the role of appropriate land cover in controlling soil erosion, noting that cropland and grassland resulted in the lowest runoff and soil loss in southern

China's red soil hilly regions. Zhang et al. [5] highlighted the impact of LULC changes on soil erosion in the Jiuyuangou watershed, showing that vegetation restoration helped reduce erosion until extreme rainfall events became more frequent. Wen and Deng [6] called for further research into the combined effects of LULC and climate change on soil erosion, stressing the importance of large-scale soil erosion modeling.

A significant challenge in data analysis, particularly in LULC studies, is the issue of imbalanced data, where certain classes are underrepresented. This imbalance can lead to biased classifiers and poor model performance. In LULC analysis, an imbalanced dataset may result in inaccurate predictions, especially for minority classes. Therefore, addressing class imbalance is essential for generating reliable insights. Various techniques have been developed to address this issue, with resampling methods like over-sampling and under-sampling being commonly used. While under-sampling reduces instances of the majority class, potentially leading to information loss, over-sampling generates synthetic data to enhance the minority class representation. This study prefers over-sampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) [7] and its variant, Adaptive Synthetic Sampling (ADASYN) [8], which improve class balance by creating synthetic data points.

SMOTE has been widely applied in various fields to address imbalanced datasets, often improving model performance. For example, SMOTE combined with Random Forest has been used to manage skewed particle datasets in particle physics, improving the accuracy of particle state analysis [9]. In hyperspectral imaging, SMOTE has been used to classify imbalanced hyperspectral data, significantly enhancing accuracy across models such as Convolutional Neural Networks (CNN) [10]. Similarly, SMOTE-based CNN models optimized with sparrow search algorithms have improved flight delay classification [11]. In space weather forecasting, a SMOTE-based Super Learner ensemble improved the classification of ionospheric scintillation events, achieving high accuracy even in adverse conditions [12]. In healthcare, SMOTE has enhanced the classification of patient safety event reports by combining neural natural language processing techniques with machine learning models, significantly improving accuracy [13].

In LULC classification, various SMOTE variants have been implemented to address imbalanced data. For instance, G_SMOTE [14] has been applied by Douzas et al. [15] and Ebrahimy et al. [16], while kmeans_SMOTE [17] has been explored by Fonseca et al. [12]. Standard SMOTE and its variant, ADASYN, have also been utilized [18]. However, most studies limit their scope to one or two oversampling methods. This study seeks to address this gap by applying a comprehensive range of oversampling techniques to classify the Cover Management factor (C-factor) in Taiwan's Shihmen Reservoir watershed.

2. Materials and Methods

In previous work, Tsai et al. [19] applied machine learning to classify the Cover Management factor (C-factor) for the Shihmen Reservoir watershed in Taiwan. The C-factor, a crucial component of the Revised Universal Soil Loss Equation (RUSLE) model, evaluates the impact of land cover and management practices on soil erosion. It measures how well vegetation and management strategies protect soil from erosion, with changes in land cover or management affecting the C-factor and influencing erosion susceptibility. Including the C-factor in soil erosion models is critical for assessing the effectiveness of land management in mitigating erosion.

One significant challenge in this analysis is the class imbalance in the dataset, where the majority class (C = 0.01) comprises over 92.5% of the data. This imbalance can skew model training, leading to the underrepresentation of the minority classes and less accurate predictions. Addressing this imbalance is essential for producing reliable soil erosion assessments. In the study watershed, the predominance of forest areas results in low C values, which may bias the model towards underestimating overall C values and, consequently, soil erosion risk. This bias has practical implications for land management, as it could lead to the under-prioritization of areas with higher erosion potential, thereby affecting resource allocation and conservation efforts. Balancing the dataset allows for a more comprehensive erosion assessment that better supports informed land management strategies.

This study seeks to classify the C-factor using a Random Forest model, incorporating various techniques to address data imbalance. Specifically, 42 oversampling methods were applied to balance the dataset before model training. These methods were designed to improve the representation of the minority classes, enhancing model performance and contributing to more accurate soil erosion assessments. This approach holds potential for improving land management strategies through better erosion risk predictions.

2.1. Data Collection

Machine learning input data is structured into two key components: predictor variables and the target variable. The target variable, which represents the outcome the algorithms are designed to predict, is the C-factor in this study. The C-factor was derived from a look-up table [20,21] and the 2004 LULC map of the Shihmen Reservoir watershed (Figure 1). As shown in Figure 1, 12 distinct C-value classes were assigned based on 23 LULC types, with forest being the predominant land cover class. This classification differs from typical LULC problems, as various LULC types can correspond to the same C-factor class for the purpose of soil erosion calculations.

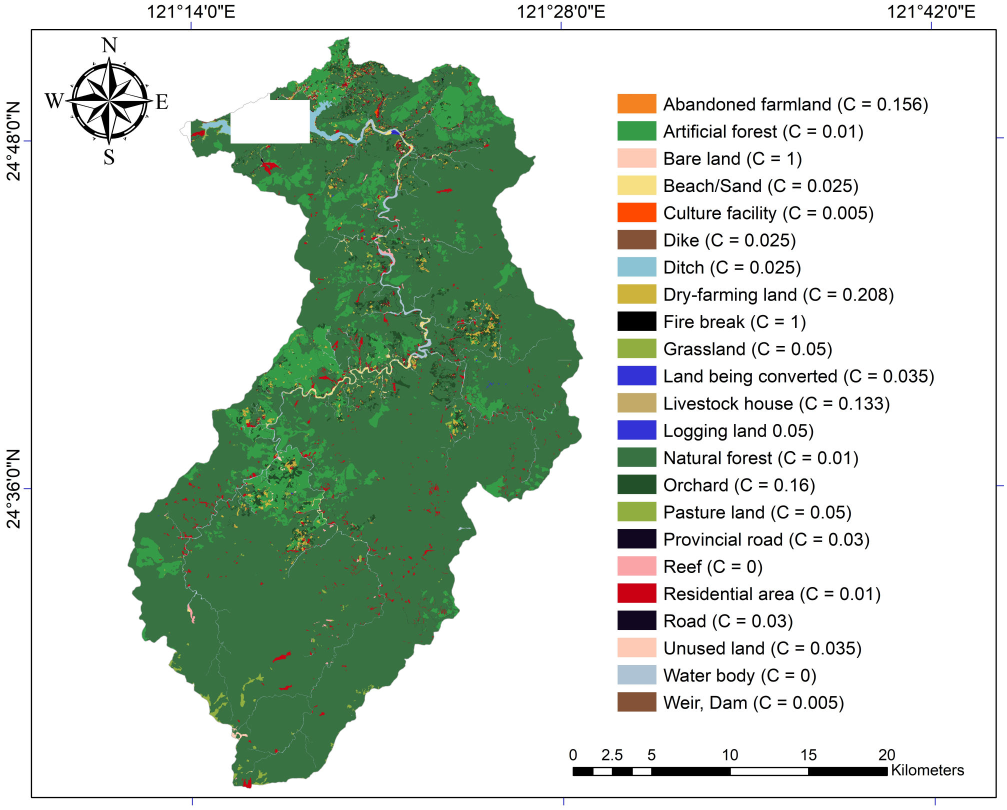

Figure 1. Map of LULC distribution and corresponding C-factors.

The predictor variables, which are used to forecast the C-factor, include eight key factors: elevation and slope, derived from a 10 m resolution Digital Elevation Model (DEM), the Normalized Difference Vegetation Index (NDVI) and the Soil Adjusted Vegetation Index (SAVI), obtained from SPOT 5 satellite imagery, and distance to road and distance to river, calculated using ArcGIS's Near tool, which measures proximity from point data to the nearest road or river. Additionally, geological and soil data were incorporated into the analysis.

A significant challenge in this study was the pronounced imbalance in the LULC classes of the Shihmen Reservoir watershed, with forest (C = 0.01) covering 92.5% of the area. This dominance of a single class can cause machine learning models to skew predictions toward the majority class, reducing the accuracy of predictions for minority classes. Addressing this imbalance is crucial for reliable classification outcomes.

2.2. Data Preparation and Model Development

To address the issue of class imbalance, a range of oversampling techniques from the smote-variants package [22] in Python (version 3.10.12, Python Software Foundation, Wilmington, DE, USA) were utilized to enhance the representation of minority classes. Initially, the dataset was divided into 70% for training and 30% for testing. The SMOTE algorithm was applied to the training dataset to balance the class distribution. However, due to the large size of the resulting dataset, which exceeded the processing capacity of Python, the training data was downsampled to 4% for subsequent analysis. Finally, the performance of models trained on the oversampled data was compared with those trained on the original, imbalanced data.

2.2.1. Data Pre-Processing

The initial phase of this study involved comprehensive data pre-processing to convert raw raster data into a format compatible with machine learning analysis. This process required transforming the raster data into point data, which allows for more efficient application of machine learning algorithms. Once the data was converted, stratified random sampling was applied to ensure the balanced representation of all C-factor classes within the dataset. Specifically, the point dataset was divided, with 70% for training and 30% for testing. This stratified approach ensured that the distribution of C-factor classes in both the training and testing datasets reflected the overall class distribution, thereby improving the reliability of subsequent model training and evaluation.

2.2.2. Handling Imbalanced Data

To address the significant class imbalance—where certain land cover categories overwhelmingly dominate—the smote-variants package [22] was employed. This package offers a variety of synthetic oversampling techniques specifically designed for imbalanced datasets. The basic SMOTE method generates synthetic samples by interpolating between existing minority class data points, thereby enhancing representation without simply duplicating data. However, traditional SMOTE can face limitations when applied to multi-class datasets, leading to the development of numerous SMOTE variants that address these specific challenges in different contexts.

The smote-variants package (V 0.7.3) now includes 90 models, 65 of which are suited for multi-class classification tasks [22,23]. By applying these models, this study conducted an extensive exploration of synthetic oversampling techniques, aiming to resolve the complexities associated with imbalanced C-factor classes. This approach not only enriched the model development process but also provided valuable insights into how different SMOTE variants interact with the unique characteristics of the land cover data.

In the original dataset, the class with $C = 0.01$ had more than 7 million points. To manage this imbalance, a stratified random sampling method was used to divide the data, with 70% for training and 30% for testing. However, the $C = 0.01$ class in the training dataset still contained over 4.9 million points, far exceeding the other classes. To address this, we applied SMOTE to upsample all other classes to match the size of the second-largest class, which had 151,879 points. Following this, we downsampled the augmented training dataset to 4% of its size to enable efficient analysis using Python and to train the Random Forest model. As a result, the $C = 0.01$ class contained 196,604 points, while all other classes had 6075 points each. This approach effectively balanced the dataset while preserving the original dataset characteristics and optimizing computational resources for model development (see Table 1).

Table 1. Dataset sizes used in the analysis to compare various SMOTE variants across different C-values.

C-Value	Original (100%)	70% Train	30% Test	4% of 70% Train	4% of 30% Test	70% Train + SMOTE	70% Train + SMOTE, 4%
0	216,970	151,879	65,091	6075	2604	151,879	6075
0.005	1110	777	333	31	13	151,879	6075
0.01	7,021,560	4,915,092	2,106,468	196,604	84,259	4,915,092	196,604
0.025	47,629	33,340	14,289	1334	572	151,879	6075
0.03	37,714	26,400	11,314	1056	453	151,879	6075
0.035	4235	2965	1271	119	51	151,879	6075
0.05	46,598	32,619	13,979	1305	559	151,879	6075
0.133	73	51	22	2	1	151,879	6075
0.156	342	239	103	10	4	151,879	6075
0.16	141,672	99,170	42,502	3967	1700	151,879	6075
0.208	60,060	42,042	18,018	1682	721	151,879	6075
1	14,099	9869	4230	395	169	151,879	6075

2.2.3. Random Forest Model

The Random Forest algorithm [24] was selected for C-factor classification in this study because of its well-established accuracy and flexibility in handling both continuous and categorical variables. As an ensemble learning method, Random Forest works by constructing multiple decision trees during training and then aggregating their predictions to produce a final output. This process increases the model's robustness and reduces the risk of overfitting, as the ensemble approach mitigates the errors that may arise from individual trees. In this study, the model was configured with 1000 decision trees, a number chosen to ensure stability in the results and enhance predictive performance.

One of the key strengths of the Random Forest algorithm is its capacity to process diverse data types, making it particularly well-suited for complex datasets like the one used in this analysis. By accommodating both numeric and categorical input variables, the algorithm can efficiently handle the varied nature of the predictor variables, such as elevation, slope, vegetation indices, and distance measures. This versatility, combined with its high predictive accuracy and ability to manage large datasets, made Random Forest an ideal choice for classifying the C-factor in the Shihmen Reservoir watershed.

2.3. Accuracy Indices

This study used a comprehensive set of evaluation metrics to assess the performance of the classification model, focusing on different dimensions of accuracy. The evaluation indices included Precision, Sensitivity, Specificity, G-mean, F1-Score, Overall Accuracy, and the Kappa coefficient. These metrics are commonly calculated from the confusion matrix, which provides a detailed summary of a model's classification performance. As noted by [25], the formulas for these accuracy metrics can differ depending on whether the classification task is binary or multi-class, underscoring the importance of selecting the appropriate calculation method for each context.

In binary classification, the confusion matrix consists of four key components: True Positives (TP), which represent instances where the model correctly predicts the positive class; False Positives (FP), where the model incorrectly predicts the positive class when it is actually negative; True Negatives (TN), which occur when the model correctly predicts the negative class; and False Negatives (FN), which arise when the model incorrectly predicts the negative class when it is actually positive.

For multi-class classification, the confusion matrix extends to a square matrix of dimensions $k \times k$, where k denotes the number of classes. Each element in the matrix at position (i, j) corresponds to the number of instances of class i that were predicted as class j. The diagonal elements represent correctly classified instances for each class, while off-diagonal elements reflect misclassified instances.

Precision, also referred to as User's Accuracy, measures the ratio of correctly predicted positive instances to the total predicted positives [26]. In essence, it evaluates the model's ability to accurately identify positive cases from the pool of predicted positives. A higher Precision value indicates a lower false positive rate, signifying the model's ability to minimize the misclassification of negative instances as positive. Precision for each class was computed using Equation (1), while the mean Precision across all classes in the multi-class setting was calculated using Equation (2).

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$\text{Precision} = \frac{1}{k}\sum_{i=1}^{k} Precision_i \tag{2}$$

where

$$k = \text{number of classes in the confusion matrix}$$
$$i = \text{a specific class index } (1, 2, \ldots, \text{k})$$

Sensitivity (Recall), also known as Producer's accuracy, measures the model's ability to correctly identify relevant instances within the dataset [26]. It is calculated as the ratio of correctly predicted positives to the total actual positives, offering insight into the model's effectiveness in detecting true positives. A higher recall indicates fewer false negatives, reflecting better detection accuracy. Sensitivity for each class was computed using Equation (3), while the mean Sensitivity across all classes was calculated using Equation (4), ensuring a balanced evaluation across the multi-class setting.

$$Sensitivity_i = \frac{TP_i}{TP_i + FN_i} \tag{3}$$

$$\text{Sensitivity} = \frac{1}{k}\sum_{i=1}^{k} Sensitivity_i \tag{4}$$

The F1-Score (Equation (5)), which is the harmonic mean of Precision and Recall, provides a comprehensive assessment of a model's performance, particularly in scenarios where class imbalance is prevalent.

$$\text{F1-Score}_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{5}$$

The Kappa Coefficient (Equation (6)) [27], or Cohen's Kappa, quantifies the agreement between predicted and actual classes while accounting for the agreement that might occur by chance.

$$\kappa = \frac{N\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k}(n_{i+} \times n_{+i})}{N^2 - \sum_{i=1}^{k}(n_{i+} \times n_{+i})} \tag{6}$$

where

$$n_{i+} = \text{total number of times class } i \text{ was predicted}$$
$$n_{+i} = \text{total number of times class } i \text{ is the true label}$$
$$n_{ii} = \text{number of samples correctly classified for class } i$$
$$N = \text{total number of samples}$$

Overall accuracy (Equation (7)) calculates the proportion of correctly classified instances (n_{ii}) among the total number of samples (N), providing a comprehensive assessment of the model's correctness. As a fundamental evaluation measure, overall accuracy serves as a useful benchmark for comparing the performance of different models and assessing their effectiveness in real-world applications.

$$\text{Overall Accuracy} = \frac{1}{N}\sum_{i=1}^{k} n_{ii} \tag{7}$$

Specificity (Equation (8)), also known as the True Negative Rate, is the ratio of true negatives to the sum of true negatives and false positives. It measures the ability of the model to identify negative instances.

$$\text{Specificity}_i = \frac{TN_i}{TN_i + FP_i} \tag{8}$$

G-mean (Equation (9)) is a metric that balances both sensitivity and specificity, providing a single value that reflects the geometric mean of the true positive rate and true negative rate. It is particularly useful for imbalanced datasets.

$$\text{G-mean}_i = \sqrt{\text{Sensitivity}_i \times \text{Specificity}_i} \tag{9}$$

3. Results and Discussion

In this section, we present the results of our analysis evaluating the effectiveness of various imbalanced data handling methods when applied in combination with the Random Forest model. The performance of these methods was systematically compared to determine the most effective approach. Afterward, the best-performing SMOTE method was compared to the baseline model, which used 4% of the total data points from each C-factor class of the Shihmen Reservoir watershed to split the training and test datasets. This baseline model, referred to as the baseline reduced dataset model (imbalanced dataset), was created using stratified random sampling without any oversampling.

This comparison highlights the improvements achieved through the application of oversampling techniques. It is important to note that in this study both the baseline reduced dataset model and the models augmented by SMOTE were tested against the full, unreduced test dataset (i.e., not reduced to 4%). Consequently, the accuracy indices reported here differ slightly from those presented in a previous study [19].

3.1. Performance of Different Imbalanced Data Methods

Table 2 summarizes the accuracy of various SMOTE variants used to predict the C-values within the study area using machine learning techniques. Due to the large dataset size and the computational demands of certain SMOTE techniques, memory limitations arose during analysis, and only 42 out of the 65 models could be retained for this study. Although steps were taken to reduce the dataset size, as detailed in Section 2.2.2, some SMOTE variants remained computationally infeasible, even after these reductions. This highlights the practical challenges associated with the dataset's size, imbalanced class distribution, and the high resource requirements of certain SMOTE methods.

The ranking column was created based on sensitivity, which measures the proportion of actual positives correctly identified. The Selected Synthetic Minority Over-sampling Technique (Selected_SMOTE) model [28] achieved the highest sensitivity at 0.6892. Although its overall accuracy ranked third at 0.9524, behind the Random Walk Over-Sampling (RWO_sampling) (0.9533) [29] and Combined Cleaning and Resampling (CCR) (0.9529) [30] models, Selected_SMOTE excelled in both sensitivity and the Kappa coefficient (0.6395), which is why it was ranked first.

In contrast, the lowest-ranked model was the Polynomial Curved-Bus Topology (polynom_fit_SMOTE_poly) [23], with an overall accuracy of 0.9513 and a sensitivity of 0.3377, reflecting its relatively poor performance. Similarly, the Denoising Autoencoder-based Generative Oversampling (DEAGO) model [31] had an accuracy of 0.9518 and a sensitivity of 0.3378, also struggling with the dataset's imbalance. Both models exhibited lower sensitivity than the original baseline model, indicating their difficulty in correctly identifying minority class instances, which ultimately affected their ranking, despite reasonable overall accuracy.

Table 2. Performance of different imbalanced data methods in classifying the C-factor.

No.	Model	Overall Accuracy	Kappa	Sensitivity	Rank
1	ADASYN	0.9478	0.5942	0.6028	23
2	AND_SMOTE	0.9517	0.6295	0.6228	19
3	ASMOBD	0.9487	0.6117	0.6521	10
4	Borderline_SMOTE1	0.9484	0.5941	0.5447	29
5	Borderline_SMOTE2	0.9481	0.5881	0.5260	30
6	CCR	0.9529	0.5973	0.5041	31
7	CE_SMOTE	0.9509	0.6242	0.6331	16
8	cluster_SMOTE	0.9516	0.6264	0.6465	12
9	DEAGO	0.9518	0.5734	0.3378	41
10	distance_SMOTE	0.9517	0.6248	0.6276	18
11	Edge_Det_SMOTE	0.9511	0.6252	0.6492	11
12	G_SMOTE	0.9511	0.6267	0.6637	5
13	Gaussian_SMOTE	0.9521	0.5799	0.3503	39
14	kmeans_SMOTE	0.9524	0.5984	0.3611	38
15	Lee	0.9517	0.6301	0.6186	21
16	LN_SMOTE	0.9517	0.6286	0.6187	20
17	LVQ_SMOTE	0.9521	0.5797	0.4258	34
18	MCT	0.9510	0.6281	0.6577	8
19	MSMOTE	0.9515	0.6272	0.6278	17
20	NDO_sampling	0.9510	0.6281	0.6779	3
21	NRAS	0.9520	0.6311	0.5510	28
22	NT_SMOTE	0.9514	0.6266	0.6539	9
23	PDFOS	0.9520	0.5827	0.3804	36
24	polynom_fit_SMOTE_bus	0.9491	0.5869	0.5571	26
25	polynom_fit_SMOTE_mesh	0.9486	0.5821	0.5542	27
26	polynom_fit_SMOTE_poly	0.9513	0.5712	0.3377	42
27	polynom_fit_SMOTE_star	0.9518	0.5773	0.3668	37
28	Random_SMOTE	0.9512	0.6252	0.6686	4
29	ROSE	0.9522	0.5893	0.4351	33
30	RWO_sampling	0.9533	0.6042	0.4911	32
31	Safe_Level_SMOTE	0.9508	0.6234	0.6433	14
32	Selected_SMOTE	0.9524	0.6395	0.6892	1
33	SL_graph_SMOTE	0.9508	0.6233	0.6032	22
34	SMMO	0.9499	0.5992	0.5614	25
35	SMOBD	0.9520	0.5809	0.3485	40
36	SMOTE	0.9510	0.6268	0.6809	2
37	SMOTE_D	0.9507	0.6254	0.6599	7
38	SMOTE_OUT	0.9518	0.6151	0.6351	15
39	SMOTEWB	0.9511	0.6262	0.6439	13
40	SN_SMOTE	0.9510	0.6254	0.6615	6
41	SSO	0.9524	0.5851	0.4114	35
42	TRIM_SMOTE	0.9514	0.6172	0.5904	24

The results highlight the variability in performance across different oversampling techniques. The Selected_SMOTE model demonstrated strong capabilities in improving classification accuracy, particularly under conditions of significant data imbalance. This finding provides a preliminary basis for informing the selection of imbalance-handling methods in machine learning applications, with the understanding that dataset characteristics play a substantial role in determining the best approach. Unlike traditional SMOTE, which treats all features equally during synthetic data generation, Selected_SMOTE focuses on synthesizing specific features based on feature selection. This approach assigns more weight to important attributes that contribute more to classification decisions, creating synthetic samples that are both meaningful and relevant.

Figure 2 presents a scatter plot comparing the Kappa Coefficient (y-axis) and Sensitivity (x-axis) for both the SMOTE variants and the baseline reduced dataset model. The baseline model, which uses 4% of the total data points in each C-factor class of the Shihmen Reservoir watershed without SMOTE, is represented by a single red cross. This model exhibits a relatively low Overall Accuracy of approximately 0.9521 and a Kappa Coefficient of around 0.5805. The majority of the points, shown in blue, represent the 42 SMOTE variants. Notably, most of these variants achieve higher Sensitivity values than the baseline model, with the exception of the polynom_fit_SMOTE_poly and DEAGO models. Additionally, most SMOTE variants also surpass the baseline model in terms of the Kappa Coefficient, further underscoring the benefits of these techniques.

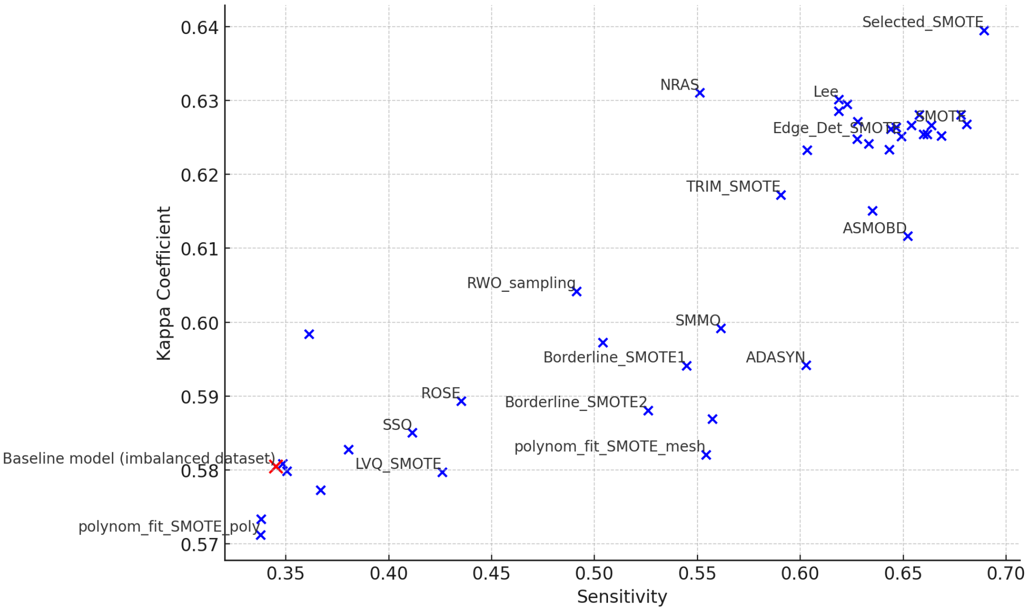

Figure 2. Comparison of sensitivity and Kappa Coefficient between SMOTE variants (blue crosses) and the baseline reduced dataset model (red cross) using a scatter plot.

The SMOTE variants have proven highly effective in enhancing accuracy compared to models trained on imbalanced datasets. In cases where one class significantly outweighs others, traditional machine learning algorithms often favor the majority class, resulting in poor classification of the minority classes. SMOTE variants address this challenge by generating synthetic samples for the minority classes, effectively balancing the class distribution. By creating synthetic instances that closely mirror the characteristics of the minority classes, these techniques provide the algorithm with more representative data, allowing it to learn the underlying patterns more accurately.

This balanced approach leads to improved generalization and performance, enabling the algorithm to make more accurate predictions for both majority and minority classes. Moreover, by diversifying the dataset, models using SMOTE variants mitigate the risk of overfitting, further enhancing accuracy and robustness. The enhanced performance, as illustrated in the scatter plot, highlights the importance of employing SMOTE variants to effectively resolve class imbalance issues in machine learning.

3.2. Comparison of Minority Class Predictions

Table 3 presents performance metrics for each class, comparing models trained on the baseline reduced dataset model (imbalanced dataset) and the balanced dataset generated using the Selected_SMOTE method.

The model trained on the imbalanced dataset struggled to correctly identify minority classes, specifically for C = 0.133 and C = 0.156. For these classes, metrics such as precision, sensitivity, F1-score, and G-mean were either zero or undefined, indicating severe issues due to class imbalance.

In contrast, the Selected_SMOTE model, trained on the balanced dataset, showed substantial improvement, particularly for the minority classes. Metrics such as precision, sensitivity, and F1-score for C = 0.133 and C = 0.156 improved significantly, demonstrating the model's ability to better identify and classify instances of these minority classes.

This comparison highlights the critical importance of addressing class imbalance before training machine learning models. It underscores the necessity of implementing effective data preprocessing techniques, such as oversampling methods like SMOTE, to mitigate the effects of imbalance and to enhance overall model performance and reliability.

Furthermore, Figure 3 compares the sensitivity of the two models across each class, showing that the sensitivity of nearly all classes improved with the application of the Selected_SMOTE variant. This underscores the importance of mitigating dataset imbalance to achieve robust machine learning outcomes. The results emphasize the effectiveness of oversampling techniques, particularly the Selected_SMOTE approach, in addressing class imbalance for C-factors and improving the overall performance of machine learning models.

Table 3. Performance metrics for each class for the Selected_SMOTE and baseline reduced dataset (imbalanced dataset).

Model	Class	Precision	Sensitivity	F1-Score	Specificity	G-Mean
Selected_SMOTE Balanced dataset	0	0.8855	0.6927	0.7774	0.9974	0.8312
	0.005	0.3067	0.9099	0.4587	0.9997	0.9537
	0.01	0.9714	0.9845	0.9779	0.6432	0.7958
	0.025	0.6688	0.8413	0.7452	0.9974	0.9160
	0.03	0.5954	0.9438	0.7302	0.9968	0.9699
	0.035	0.1792	0.7677	0.2905	0.9980	0.8753
	0.05	0.6143	0.4507	0.5199	0.9983	0.6707
	0.133	0.0133	0.4545	0.0258	0.9997	0.6741
	0.156	0.2192	0.9320	0.3549	0.9998	0.9653
	0.16	0.5887	0.2333	0.3342	0.9969	0.4823
	0.208	0.4600	0.3936	0.4242	0.9963	0.6262
	1	0.4866	0.6660	0.5623	0.9987	0.8155
Baseline reduced dataset model - Imbalanced dataset	0	0.8345	0.6962	0.7591	0.9959	0.8327
	0.005	0.7677	0.2282	0.3519	1.0000	0.4777
	0.01	0.9604	0.9936	0.9767	0.4956	0.7017
	0.025	0.8061	0.6462	0.7174	0.9990	0.8035
	0.03	0.6710	0.7789	0.7209	0.9981	0.8817
	0.035	0.8290	0.2520	0.3865	1.0000	0.5020
	0.05	0.8126	0.1604	0.2679	0.9998	0.4004
	0.133	-	0.0000	-	1.0000	0.0000
	0.156	-	0.0000	-	1.0000	0.0000
	0.16	0.5794	0.1623	0.2535	0.9978	0.4024
	0.208	0.5794	0.1185	0.1967	0.9993	0.3441
	1	0.8820	0.1043	0.1865	1.0000	0.3229

Note: (-) undefined value.

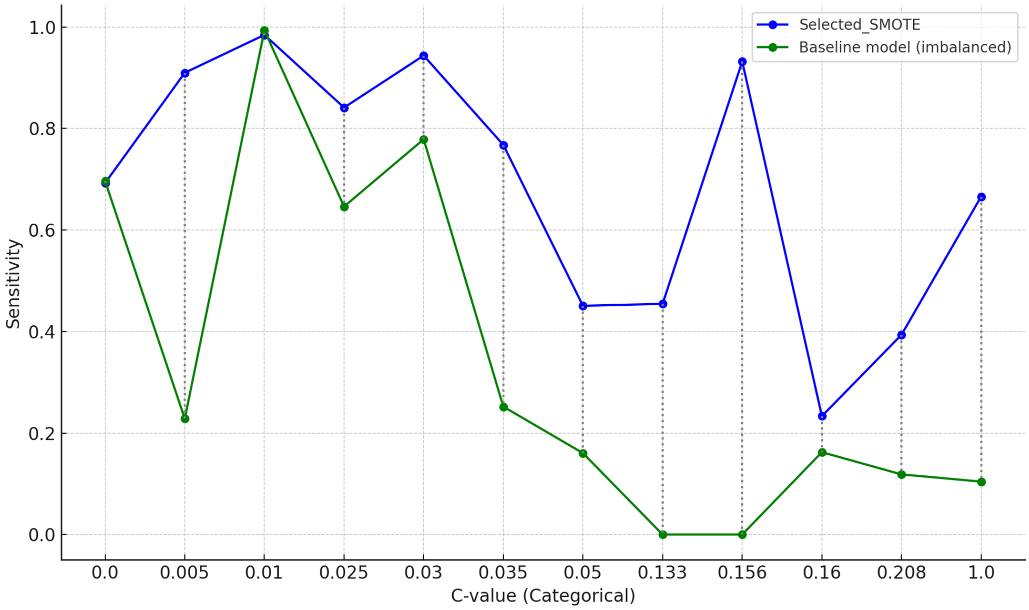

Figure 3. Producer's accuracy (sensitivity) between the baseline reduced dataset model (imbalanced dataset) and Selected_SMOTE model.

3.3. Improving Minority Class Performance Through SMOTE Variants

Models trained on imbalanced datasets typically show reduced accuracy for minority classes due to two main factors. First, the limited representation of the minority classes in such datasets restricts the model's exposure to these instances, making it difficult to learn and distinguish the unique features of minority class samples. This lack of sufficient training data results in poor classification performance for the minority classes.

Second, traditional machine learning algorithms tend to focus on optimizing overall performance metrics, like overall accuracy, which often leads to bias toward the majority class. This bias results in the model prioritizing the correct classification of majority class instances at the expense of the minority classes, causing lower recall and F1-score for the underrepresented class.

SMOTE variants address these challenges by enhancing overall accuracy while specifically improving the classification of minority classes. Imbalanced datasets pose significant hurdles for traditional models, as they favor the majority class, leading to misclassification or omission of critical minority class instances. SMOTE variants mitigate this issue by oversampling the minority classes, ensuring that the model encounters enough minority instances during training. This exposure allows the model to better capture the nuances of the minority classes, leading to significant improvements in F1-score, recall, and overall performance.

By prioritizing the correct identification of minority instances, SMOTE variants not only improve model performance but also provide more reliable and actionable insights, particularly in scenarios where minority classes represent critical outcomes. The results of this study highlight that addressing dataset imbalance is crucial for achieving more accurate and reasonable performance across all classes, underscoring the importance of careful data balancing prior to model training.

3.4. Limitations of the Study

While this study demonstrates the effectiveness of SMOTE variants in enhancing classification accuracy for cover management factor estimation, several limitations should be acknowledged. First, our findings and conclusions may not be directly applicable to other application scenarios or datasets. The smote-variants package, which originally implemented 85 SMOTE techniques, provides a comprehensive comparison of these methods across 104 imbalanced datasets, with rankings based on classification accuracy and other performance metrics [32,33]. This comparison includes a "Ranking" section listing the top 10 SMOTE variants, yet only G_SMOTE from this top 10 aligns with our top 5 methods, while others from the smote-variants rankings fall between 21 and 42 in our study or do not appear in our results. This discrepancy suggests that the optimal choice of SMOTE variant is closely tied to dataset-specific characteristics, an insight that limits the generalizability of any single SMOTE variant's performance.

Second, our findings, together with prior studies, suggest that the effectiveness of SMOTE techniques is highly dependent on the unique characteristics of each dataset. While this study is the first to apply these SMOTE variants specifically for balancing data in cover management factor estimation, we found no established framework to categorize or universally recommend SMOTE techniques for similar applications. The variability in performance observed across datasets indicates that selecting an optimal SMOTE variant based solely on problem type or application field is impractical without considering the specific data characteristics involved.

Given these insights, it is premature to conclude that any particular application context alone can reliably guide the selection of a specific SMOTE variant to ensure optimal performance across diverse contexts. Instead, our findings emphasize the need for further research to systematically investigate the interaction between dataset properties and SMOTE performance. Such research could ultimately provide a clearer framework for selecting the most suitable SMOTE techniques across various applications, aided by empirical results from both our study and the comprehensive comparisons offered by the smote-variants package.

4. Conclusions

This study addresses the challenge of class imbalance in LULC classification, a key issue in accurately mapping the cover management factor within datasets dominated by majority classes. In the study area, the predominance of forested areas results in a heavily imbalanced dataset, which hinders machine learning models from effectively classifying minority classes within the C-factor. While previous models achieved reasonable overall accuracy, they struggled with identifying these minority classes accurately. To address this issue, our study aimed to balance the dataset prior to model training, thereby enhancing classification accuracy across all classes.

To address the class imbalance, the smote-variants package was used, applying various SMOTE techniques to create a more balanced dataset for C-factor classification. The results indicate substantial improvements in model performance across nearly all SMOTE variants. Selected_SMOTE excelled in both sensitivity (0.6892) and the kappa coefficient (0.6395), which contributed to its ranking as the top method overall, despite its overall accuracy ranking third at 0.9524. These results underscore the effectiveness of SMOTE variants in enhancing model performance on imbalanced datasets.

SMOTE techniques improve model performance by generating synthetic samples for minority classes, allowing the model to learn more effectively from the full range of class distributions. By addressing the imbalance, SMOTE variants provide a more comprehensive training set that captures the complexities of both majority and minority classes, resulting in improved classification accuracy across the board.

In summary, this study demonstrates that addressing class imbalance through SMOTE variants significantly enhances classification performance in C-factor modeling. The success of the Selected_SMOTE method illustrates the potential of balanced datasets in improving machine learning outcomes for imbalanced data, providing a valuable approach for future applications in LULC classification and other fields facing similar challenges.

Author Contributions: Conceptualization, W.C.; Data Curation, K.A.N.; Funding Acquisition, W.C.; Investigation, W.C. and K.A.N.; Methodology, W.C. and K.A.N.; Project Administration, W.C.; Resources, W.C.; Software, K.A.N.; Supervision, W.C.; Validation, W.C. and K.A.N.; Visualization, K.A.N.; Writing—Original Draft, W.C. and K.A.N.; Writing—Review & Editing, W.C. and K.A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially supported by the Ministry of Science and Technology (Taiwan) Research Project (Grant Number MOST 111-2121-M-027-001) and the National Science and Technology Council (Taiwan) Research Project (Grant Number NSTC 112-2121-M-027-001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because they are from a confidential third party. Requests to access the datasets should be directed to the corresponding author.

Acknowledgments: The authors wish to extend their sincere gratitude to all those who offered support and feedback throughout the course of this research. We are also thankful to our institution for providing the resources and facilities that made this work possible. In preparing this manuscript, the authors used ChatGPT 4o (OpenAI, San Francisco, CA, USA) to enhance readability and language. Following its use, the authors thoroughly reviewed and edited the content as necessary and take full responsibility for the final published work. Additionally, the authors acknowledge that a native English speaker with a Ph.D. degree proofread the manuscript to improve the language quality.

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

References

1. Al-Kaisi, M. Soil erosion: An agricultural production challenge. *Integr. Crop Manag.* **2000**, *484*, 141–143.
2. McCool, D.K.; Williams, J.D. Soil erosion by water. *Encycl. Ecol.* **2008**, 3284–3290.
3. 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. [CrossRef]
4. Chen, J.; Li, Z.; Xiao, H.; Ning, K.; Tang, C. Effects of land use and land cover on soil erosion control in southern China: Implications from a systematic quantitative review. *J. Environ. Manag.* **2021**, *282*, 111924. [CrossRef]
5. Zhang, B.; Chen, Z.; Shi, X.; Wu, S.; Feng, H.; Gao, X.; Siddique, K.H. Temporal and spatial changes of soil erosion under land use and land cover change based on Chinese soil loss equation in the typical watershed on the Loess Plateau. *Soil Use Manag.* **2023**, *39*, 557–570. [CrossRef]
6. Wen, X.; Deng, X. Current soil erosion assessment in the Loess Plateau of China: A mini-review. *J. Clean. Prod.* **2020**, *276*, 123091. [CrossRef]
7. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
8. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–6 June 2008; pp. 1322–1328.
9. Yalcin Kuzu, S. Random forest based multiclass classification approach for highly skewed particle data. *J. Sci. Comput.* **2023**, *95*, 21. [CrossRef]
10. Özdemir, A.; Polat, K.; Alhudhaif, A. Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods. *Expert Syst. Appl.* **2021**, *178*, 114986. [CrossRef]
11. Deng, W.; He, Q.; Zhou, X.; Chen, H.; Zhao, H. A sparrow search algorithm-optimized convolutional neural network for imbalanced data classification using synthetic minority over-sampling technique. *Phys. Scr.* **2023**, *98*, 116001. [CrossRef]
12. Fonseca, J.; Douzas, G.; Bacao, F. Improving imbalanced land cover classification with K-Means SMOTE: Detecting and oversampling distinctive minority spectral signatures. *Information* **2021**, *12*, 266. [CrossRef]
13. Srivani, I.; Sridhar, M.; Swamy, K.C.T.; Ratnam, D.V. Multi-class classification of ionospheric scintillations using SMOTE-Super Learner ensemble technique. *Adv. Space Res.* **2024**, *73*, 3845–3854. [CrossRef]
14. Sandhan, T.; Choi, J.Y. Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 1449–1453.
15. Douzas, G.; Bacao, F.; Fonseca, J.; Khudinyan, M. Imbalanced learning in land cover classification: Improving minority classes' prediction accuracy using the geometric SMOTE algorithm. *Remote Sens.* **2019**, *11*, 3040. [CrossRef]
16. Ebrahimy, H.; Naboureh, A.; Feizizadeh, B.; Aryal, J.; Ghorbanzadeh, O. Integration of Sentinel-1 and Sentinel-2 data with the G-SMOTE technique for boosting land cover classification accuracy. *Appl. Sci.* **2021**, *11*, 10309. [CrossRef]
17. Douzas, G.; Bacao, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. [CrossRef]
18. Ebrahimy, H.; Mirbagheri, B.; Matkan, A.A.; Azadbakht, M. Effectiveness of the integration of data balancing techniques and tree-based ensemble machine learning algorithms for spatially-explicit land cover accuracy prediction. *Remote Sens. Appl. Soc. Environ.* **2022**, *27*, 100785. [CrossRef]

19. 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. [CrossRef]
20. 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)
21. 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)
22. Kovács, G. Smote-variants: A python implementation of 85 minority oversampling techniques. *Neurocomputing* **2019**, *366*, 352–354. [CrossRef]
23. Gazzah, S.; Amara, N.E.B. New oversampling approaches based on polynomial fitting for imbalanced data sets. In Proceedings of the 2008 Eighth IAPR International Workshop on Document Analysis Systems, Nara, Japan, 16–19 September 2008; pp. 677–684.
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. Chen, W. Revisiting "a hybrid UNet based approach for crop classification using Sentinel-1B synthetic aperture radar images": A comment aided by ChatGPT. *Multimed. Tools Appl.* **2024**. [CrossRef]
26. Bonnett, R.; Campbell, J.B. *Introduction to Remote Sensing*; CRC Press: Boca Raton, FL, USA, 2002.
27. Lillesand, T.; Kiefer, R.W.; Chipman, J. *Remote Sensing and Image Interpretation*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
28. Koto, F. SMOTE-Out, SMOTE-Cosine, and Selected-SMOTE: An enhancement strategy to handle imbalance in data level. In Proceedings of the 2014 International Conference on Advanced Computer Science and Information System, Jakarta, Indonesia, 13–14 October 2014; pp. 280–284.
29. Zhang, H.; Li, M. RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Inf. Fusion* **2014**, *20*, 99–116. [CrossRef]
30. Koziarski, M.; Woźniak, M. CCR: A combined cleaning and resampling algorithm for imbalanced data classification. *Int. J. Appl. Math. Comput. Sci.* **2017**, *27*, 635–645. [CrossRef]
31. Bellinger, C.; Japkowicz, N.; Drummond, C. Synthetic oversampling for advanced radioactive threat detection. In Proceedings of the 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 9–11 December 2015; pp. 948–953.
32. SMOTE-Variants Documentation. Ranking. Available online: https://smote-variants.readthedocs.io/en/latest/ranking.html (accessed on 5 November 2024).
33. Kovács, G. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Appl. Soft Comput.* **2019**, *83*, 105662. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.