Cover Letter

Combining class-weighted algorithm and machine learning models in landslide susceptibility mapping: a case study of Wanzhou section of the Three Gorges Reservoir, China

Huijuan Zhang,Yingxu Song,Shiluo Xu,Yueshun He,Zhiwen Li,Xianyu Yu,Ye Liang,Weicheng Wu,Yue Wang

Dear Editors-in-Chief,

please find the enclosed manuscript "Combining class-weighted algorithm and machine learning models in landslide susceptibility mapping: a case study of Wanzhou section of the Three Gorges Reservoir, China" which we are submitting for exclusive consideration for publication in Computers & Geosciences. We confirm that the submission follows all the requirements and includes all the items of the submission checklist.

In this contribution, to solve the imbalanced landslide samples (landslides, non-landslides) in the landslide susceptibility evaluation, the application of the class-weighted algorithm combined with traditional machine learning (logistic regression) and ensemble machine learning models (LightGBM and random forest) have been investigated. Wanzhou section of the Three Gorges Reservoir area, China, where the number of landslide samples is 19 times more than non-landslide samples, is chosen as an example. The landslide inventory database was produced using field investigation and remote sensing images provided by Google Earth. Of the 233 landslides in the inventory, 40% were used for validation, and the remaining 60% were used for training purposes. Twelve environmental parameters (elevation, slope, aspect, curvature, distance to river, NDVI, NDWI, rainfall, seismic intensity, land use, TRI, lithology) were used as inputs of the models to produce landslide susceptibility map (LSM). The AUC value, Balanced accuracy, and Geometric mean score were used to estimate the quality of models. Research has found that the weighted models (weighted logistic regression, weighted LightGBM, weighted random forest) are better than unweighted methods and the weighted random forest method has the best performance. The class-weighted algorithm turned the susceptibility evaluation problem into a cost-sensitive problem by setting unequal weights for different classes, which is probably to be applied to the landslide susceptibility evaluation in other areas.

We provide the source codes in a public repository with details listed in the section "Code availability".

Thanks for your consideration.
Sincerely,

Yingxu Song
Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, School of Information and Engineering, East China University of Technology, Nanchang, 330013, Jiangxi, China; yxsong@ecut.edu.cn

Highlights

Combining class-weighted algorithm and machine learning models in landslide susceptibility mapping: a case study of Wanzhou section of the Three Gorges Reservoir, China

Huijuan Zhang,Yingxu Song,Shiluo Xu,Yueshun He,Zhiwen Li,Xianyu Yu,Ye Liang,Weicheng Wu,Yue Wang

- The imbalanced landslide samples (landslides, non-landslides) in the landslide susceptibility evaluation is emphasized.
- The class-weighted algorithm combined with machine learning (Logistic regression) and ensemble machine learning models (LightGBM and random forest) were applied to the landslide susceptibility evaluation.
- The weighted models are applicable for solving the problem of imbalanced landslide samples and have improved the landslide susceptibility mapping well.

Combining class-weighted algorithm and machine learning models in landslide susceptibility mapping: a case study of Wanzhou section of the Three Gorges Reservoir, China

Huijuan Zhang[a,b], Yingxu Song[c], Shiluo Xu[d], Yueshun He[e], Zhiwen Li[f], Xianyu Yu[g], Ye Liang[h], Weicheng Wu[a] and Yue Wang[b]

[a] *Key Lab of Digital Land and Resources and Faculty of Earth Sciences, East China University of Technology, Nanchang, 330013, Jiangxi, China*
[b] *School of Earth Sciences, East China University of Technology, Nanchang, Jiangxi Province 330013, China*
[c] *Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, School of Information and Engineering, East China University of Technology, Nanchang, 330013, Jiangxi, China; yxsong@ecut.edu.cn*
[d] *School of Information Engineering, Huzhou University, Huzhou 313000, China; xushiluo@163.com*
[e] *East China University of Technology, Nanchang, 330013, Jiangxi, China; heys@ecut.edu.cn*
[f] *School of Environmental and Chemical Engineering, Foshan University, Foshan, 528000, China; lizw1982@163.com*
[g] *School of Civil Engineering, Architecture and Environment, Hubei University of Technology, Wuhan, Hubei Province 430074, China; yuxianyu@hbut.edu.cn*
[h] *Jiangxi Engineering Technology Research Center of Nuclear Geoscience Data Science and System, East China University of Technology, Nanchang, 330013, Jiangxi, China; liangye@ecut.edu.cn*
[a] *Key Lab of Digital Land and Resources and Faculty of Earth Sciences, East China University of Technology, Nanchang, 330013, Jiangxi, China; wuwch@ecut.edu.cn/wuwc030903@sina.com*
[b] *School of Earth Sciences, East China University of Technology, Nanchang, Jiangxi Province 330013, China; 2020210058@ecut.edu.cn*

ARTICLE INFO

Keywords:
landslide susceptibility mapping
class-weighted algorithm
imbalanced landslide data
machine learning model
Three Gorges Reservoir area

ABSTRACT

This study aims to investigate the application of the class-weighted algorithm combined with traditional machine learning (logistic regression) and ensemble machine learning models (LightGBM and random forest) to the landslide susceptibility evaluation. Wanzhou section of the Three Gorges Reservoir area, China, which have numerous landslides and the number of landslide samples is 19 times more than non-landslide samples, is chosen as an example. The class-weighted algorithm focuses on the class-imbalanced problem of landslide and non-landslide samples in the assessment of landslide susceptibility and can turn the class-imbalanced issue into a cost-sensitive problem by setting unequal weights for different classes, which contribute to improving landslide susceptibility evaluation accuracy. The landslide inventory database was produced by field investigation and remote sensing images derived from Google Earth. Of the 233 landslides in the inventory, 40% were used for validation, and the remaining 60% were used for training purposes. Twelve environmental parameters (elevation, slope, aspect, curvature, distance to river, NDVI, NDWI, rainfall, seismic intensity, land use, TRI, lithology) were treated as inputs of the models to produce landslide susceptibility map (LSM). The AUC value, Balanced accuracy, and Geometric mean score were utilized to estimate the quality of models. The results showed that the weighted models (weighted logistic regression, weighted LightGBM, weighted random forest) have higher AUC values, Balanced accuracy, and Geometric mean scores than those of unweighted methods, which demonstrated that the weighted models exhibit better than unweighted methods, with the weighted random forest method having the best performance. The landslide susceptibility map of the Wanzhou section display that the high and very high landslide susceptibility are mainly distributed on both sides of the river. The insights from this research will be useful for ameliorating the landslide susceptibility mapping and the development of prevention and mitigation Wanzhou section.

CRediT authorship contribution statement

Huijuan Zhang: Conceptualization, Validation, Writing-original draft preparation, Writing-review and editing. **Yingxu Song:** Conceptualization, Methodology, Software, Validation, Investigation, Resources, Funding acquisition.

ORCID(s): 0000-0002-9273-2019 (Y. Song)

Shiluo Xu: Software, Resources. **Yueshun He:** Project administration, Funding acquisition. **Zhiwen Li:** Conceptualization. **Xianyu Yu:** Resources, Funding acquisition. **Ye Liang:** Funding acquisition. **Weicheng Wu:** Writing-review and editing. **Yue Wang:** Software.

1. Introduction

Landslide refers to a natural phenomenon in which the soil or rock mass on the slope slides downwards along the soft surface under the action of gravity or other external forces. Landslide is a common geological disaster, causing many economic losses and unfor-tunate casualties, such as devastating soil, vegetation, and dwellings, as well as critically blocking transportation lines and waterways(Abuzied et al., 2016; Chen et al., 2017). The China Geological Survey reported that there were 6181 geological disasters in 2019, including landslides, collapses, mudrock flows, the ground collapses, ground fissures, and land subsidence, resulting in 211 deaths, 13 missings, 75 injured and direct economic losses of 2.77 billion Yuan. Among them, 4020 landslides occurred, mainly distributed in Southwestern China, and brought about a large number of missing persons and severe economic losses. Various factors, such as natural factors (e.g., heavy rainfall, earthquake, loose lithology, and low vegetation coverage, etc.) and human-made factors (e.g., infrastructures con-struction and road irrigation, etc.) can trigger landslides(Wilde et al., 2018). Especially in recent years, the rapid urbanization and industrialization have increased the likelihood of landslide occurrence (Kocaman et al., 2020), which led to higher number of human casualties and more enormous loss of property. It is therefore of significant necessity to develop landslide susceptibility map, which represents the probability of the spatial distribution of landslides in a specific region based on historical landslides and related factors (Yu et al., 2016; Song et al., 2018).Government agencies have attempted to take various measures to reduce the casualties and financial losses caused by landslides. This process generally involves carrying out LSM, representing the probability of the spatial distribution of landslides in a specific region based on historical landslides and related factors (Yu et al., 2016; Song et al., 2018). Landslide susceptibility map can help government agencies to take preventable measures for reducing the casualties and financial losses caused by landslides.

Various methods and techniques, which can be defined as qualitative or quantitative, have been implemented in the landslide susceptibility assessment and have achieved notable progress (Fang et al., 2020; Guzzetti et al., 1999; Bui et al., 2020). Qualitative methods are based on expert knowledge to identify the main triggering factors, determine the weights of natural and human-made factors and acquire landslide susceptible zones (Aditian et al., 2018), such as analytic hierarchy process (AHP) (Barredo et al., 2000; Yalcin, 2008; Feizizadeh et al., 2014)(Barredo et al., 2000; Yalcin, 2008), interval pairwise comparison matrix (IPCM)(Ghorbanzadeh et al., 2019), and fuzzy logic models(Aksoy and Ercanoglu, 2012; Anbalagan et al., 2015; Shahabi et al., 2015; Roy and Saha, 2019). Whereas quantitative methods rely on mathematical models including the statistical and deterministic models(Abuzied et al., 2016; Reichenbach et al., 2018; Fang et al., 2020). With the rapid advancement of computer technology and the improvement of remote sensing (RS) and geographic information system (GIS) technology, the quantitative methods develop swiftly. Many studies have demonstrated that the quantitative approaches are more precise than qualitative methods because the qualitative methods have much subjectivity concerning the prediction of landslides(Aditian et al., 2018; Bui et al., 2020). Machine learning model which is one of the qualitative methods has the capability of handling non-linear data with different scales and from different type of sources(Bui et al., 2020). Different machine learning algorithms together with GIS and RS techniques have been widely applied to assess landslide susceptibility and perform well, such as LR (logistic regression), which were most widely used and often found successful in the landslide susceptibility evaluation (wenxian12/13) (Ayalew and Yamagishi, 2005; Eeckhaut et al., 2006; Bai et al., 2010; Akgun, 2012)(Sevgen et al., 2019). Additionally, the ensemble learning methods acting as an improvement of traditional machine learning models arise and show more robust performance in many real-world tasks, widely used in landslide susceptibility evaluation (Althuwaynee et al., 2014; Napoli et al., 2020; Hong et al., 2020; Saha et al., 2021). Random forest (RF) (Breiman, 2001), which is an extended variant of the bagging method, has a simple implementation and low computational overhead (Youssef et al., 2015; Kim et al., 2017). LigthGBM is a new member of the boosting ensemble models, having faster training efficiency, higher accuracy, and more robust ability to handle large-scale data (Song et al., 2018). However, landslide samples are often much less than non-landslide samples in almost every study region, leading to poor performance in landslide susceptibility evaluation whether the traditional machine learning or ensemble machine learning models are utilized. Some researchers have paid attention to the sample selection in the evaluation of landslide susceptibility, (Ada and San, 2017; San, 2014; Nefeslioglu et al., 2012).

The class-weighted algorithm treats the susceptibility assessment as a cost-sensitive issue and sets different misclassification weights for different categories (landslides, non-landslides). This method has been widely used to solve the unbalanced variety, but the application to landslide susceptibility assessment is still relatively few.

Wanzhou district of Chongqing is in the Three Gorges Reservoir area's hinterland, playing a significant role in the prevention and domination of geological disasters in the Three Gorges Reservoir area. In recent decades, because of the abundant precipitation and cyclical fluctuation of water level in the Yangtze River, landslides and other geological disasters in this area have increased significantly, seriously destroying the eco-logical environment and socially sustainable development. In this study, the Wanzhou section of Three Gorges Reservoir was selected as the research area, and the class-weighted algorithm combined with traditional machine learning model (Logistic regression) and ensemble machine learning models (LightGBM and random forest) were applied to the landslide susceptibility evaluation. The purpose of this research attempts to achieve the relatively optimal method in which the impact of unbalanced landslide samples can be minimized, and the accuracy of the landslide susceptibility map is improved, providing essential introductory information for mitigating the land-slide hazard by governmental subdivisions or decision-makers. Different from previous work, the novelty of this paper are 1) the class-weighted algorithm is firstly applied to landslide sus-

Table 1
Example of table.

	a	b	c
a	0.014	0.20	0.13
b	0.20	0.17	2.46
c	0.13	2.5	0.31

ceptibility mapping; 2) the advantages and disadvantages of traditional machine learning model (Logistic regression) and ensemble machine learning models (LightGBM and random forest) combined with class-weighted algorithm were com-pared in the Wanzhou section.

Examples of citations:

Gómez-Hernández and Srivastava (1990); Pebesma (2004); Hansen et al. (2018)

Examples of citations in parentheses:

(Gómez-Hernández and Srivastava, 1990; Pebesma, 2004; Hansen et al., 2018)

2. Methodology

This section includes an example of equation.

$$y = ax + b. \tag{1}$$

2.1. Subsection

This section contains another example of equation, different from Eq. 1.

$$y = ax^2 + bx + c \tag{2}$$

3. Algorithm and implementation

Example of algorithm:

4. Results

This section includes an example of figure (Figure 1), from de Figueiredo et al. (2021).

This section includes an example of table (Table 1).

Algorithm 1 Algorithm example

Input: ...

1. Step1
2. Step2;
3. Step3;

for i = 1,..., m **do**
    4. Step 4;
    for j = 2,..., n **do**
        5. Step 5;
        6. Step 6;
    end for
end for

Output: ...

4.1. Subsection

Text ...

5. Conclusions

Conclusions here...

6. Acknowledgments

The authors would like to acknowledge ...

Code availability section

Name of the code/library

Contact: e-mail and phone number

Hardware requirements: ...

Program language: ...

Software required: ...

Program size: ...

The source codes are available for downloading at the link: https://github.com/ . . . .

References

Abuzied, S., Ibrahim, S., Kaiser, M., Saleem, T., 2016. Geospatial susceptibility mapping of earthquake-induced landslides in nuweiba area, gulf of aqaba, egypt. Journal of Mountain Science 13, 1286–1303.

Ada, M., San, B.T., 2017. Comparison of machine-learning techniques for landslide susceptibility mapping using two-level random sampling (2lrs) in alakir catchment area, antalya, turkey. Natural Hazards 90, 237–263. doi:10.1007/s11069-017-3043-8.

Aditian, A., Kubota, T., Shinohara, Y., 2018. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of ambon, indonesia. Geomorphology 318, 101–111. doi:10.1016/j.geomorph.2018.06.006.

Aksoy, B., Ercanoglu, M., 2012. Landslide identification and classification by object-based image analysis and fuzzy logic: An example from the azdavay region (kastamonu, turkey). Computers & Geosciences 38, 87–98. doi:10.1016/j.cageo.2011.05.010.

Anbalagan, R., Kumar, R., Lakshmanan, K., Parida, S., Neethu, S., 2015. Landslide hazard zonation mapping using frequency ratio and fuzzy logic approach, a case study of lachung valley, sikkim. Geoenvironmental Disasters 2. doi:10.1186/s40677-014-0009-y.

Barredo, J., Benavides, A., Hervás, J., van Westen, C.J., 2000. Comparing heuristic landslide hazard assessment techniques using GIS in the tirajana basin, gran canaria island, spain. International Journal of Applied Earth Observation and Geoinformation 2, 9–23. doi:10.1016/s0303-2434(00)85022-9.

Bui, D.T., Tsangaratos, P., Nguyen, V.T., Liem, N.V., Trinh, P.T., 2020. Comparing the prediction performance of a deep learning neural network model with conventional machine learning models in landslide susceptibility assessment. CATENA 188, 104426. doi:10.1016/j.catena.2019.104426.

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z., Ma, J., 2017. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 151, 147–160.

Fang, Z., Wang, Y., Peng, L., Hong, H., 2020. A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. International Journal of Geographical Information Science 35, 321–347. doi:10.1080/13658816.2020.1808897.

de Figueiredo, L.P., Schmitz, T., Lunelli, R., Roisenberg, M., de Freitas, D.S., Grana, D., 2021. Direct multivariate simulation-a stepwise conditional transformation for multivariate geostatistical simulation. Computers & Geosciences 147, 104659.

Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S., Tiede, D., Aryal, J., 2019. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Remote Sensing 11, 196. doi:10.3390/rs11020196.

Gómez-Hernández, J.J., Srivastava, R.M., 1990. Isim3d: An ansi-c three-dimensional multiple indicator conditional simulation program. Computers & Geosciences 16, 395–440.

Guzzetti, F., Carrara, A., Cardinali, M., Reichenbach, P., 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, central italy. Geomorphology 31, 181–216.

Hansen, T.M., Mosegaard, K., Cordua, K.S., 2018. Multiple point statistical simulation using uncertain (soft) conditional data. Computers & Geosciences 114, 1–10.

Kocaman, S., Tavus, B., Nefeslioglu, H.A., Karakas, G., Gokceoglu, C., 2020. Evaluation of floods and landslides triggered by a meteorological catastrophe (ordu, turkey, august 2018) using optical and radar data. Geofluids 2020, 1–18. doi:10.1155/2020/8830661.

Nefeslioglu, H., San, B., Gokceoglu, C., Duman, T., 2012. An assessment on the use of terra ASTER l3a data in landslide susceptibility mapping. International Journal of Applied Earth Observation and Geoinformation 14, 40–60. doi:10.1016/j.jag.2011.08.005.

Pebesma, E.J., 2004. Multivariable geostatistics in s: the gstat package. Computers & Geosciences 30, 683–691.

Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth-Science Reviews 180, 60–91. doi:10.1016/j.earscirev.2018.03.001.

Roy, J., Saha, D.S., 2019. GIS-based gully erosion susceptibility evaluation using frequency ratio, cosine amplitude and logistic regression ensembled with fuzzy logic in hinglo river basin, india. Remote Sensing Applications: Society and Environment 15, 100247. doi:10.1016/j.rsase.2019.100247.

San, B.T., 2014. An evaluation of SVM using polygon-based random sampling in landslide susceptibility mapping: The candir catchment area (western antalya, turkey). International Journal of Applied Earth Observation and Geoinformation 26, 399–412. doi:10.1016/j.jag.2013.09.010.

Sevgen, E., Kocaman, S., Nefeslioglu, H.A., Gokceoglu, C., 2019. A novel performance assessment approach using photogrammetric techniques for landslide susceptibility mapping with logistic regression, ann and random forest. Sensors 19, 3940.

Shahabi, H., Hashim, M., Ahmad, B.B., 2015. Remote sensing and gis-based landslide susceptibility mapping using frequency ratio, logistic regression, and fuzzy logic methods at the central zab basin, iran. Environmental Earth Sciences 73, 1–22.

Song, Y., Niu, R., Xu, S., Ye, R., Peng, L., Guo, T., Li, S., Chen, T., 2018. Landslide susceptibility mapping based on weighted gradient boosting decision tree in wanzhou section of the three gorges reservoir area (china). ISPRS International Journal of Geo-Information 8, 4. doi:10.3390/ijgi8010004.

Wilde, M., Günther, A., Reichenbach, P., Malet, J.P., Hervás, J., 2018. Pan-european landslide susceptibility mapping: ELSUS version 2. Journal of Maps 14, 97–104. doi:10.1080/17445647.2018.1432511.

Yalcin, A., 2008. Gis-based landslide susceptibility mapping using analytical hierarchy process and bivariate statistics in ardesen (turkey): Comparisons of results and confirmations. Catena 72, 1–12.

Yu, X., Wang, Y., Niu, R., Hu, Y., 2016. A combination of geographically weighted regression, particle swarm optimization and support vector machine for landslide susceptibility mapping: A case study at wanzhou in the three gorges area, china. Int J Environ Res Public Health 13, 487.

List of Figures

1 Caption here. Image from de Figueiredo et al. (2021). 9

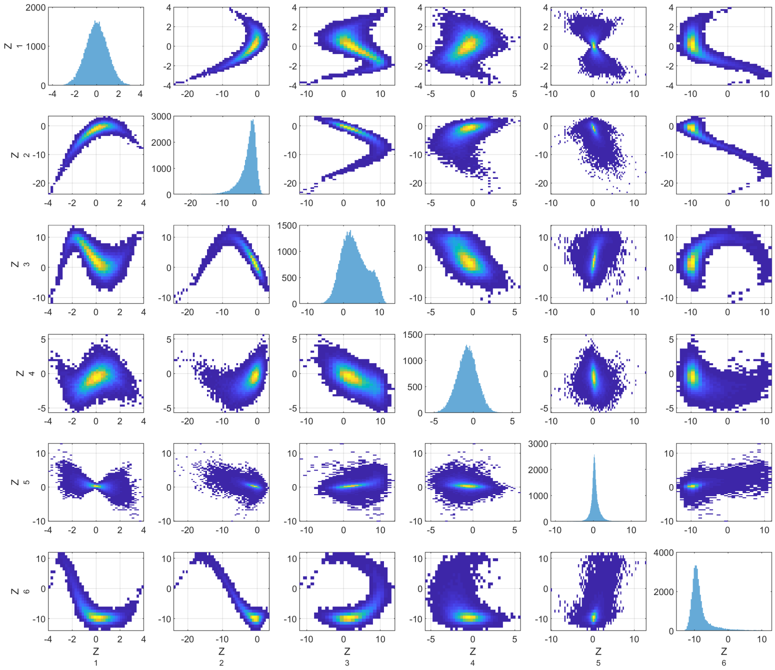

Figure 1: Caption here. Image from de Figueiredo et al. (2021).