Russian Journal of Earth Sciences

Special Issue: "Advances in Environmental Studies, from the VIII International Scientific and Practical Conference 'Fundamental and Applied Aspects of Geology, Geophysics and Geoecology Using Modern Information Technologies' "

Comprehensive Analysis of Mudflow Data Using Machine Learning Methods

M. M. Gedueva[1]*, E. V. Kyul[1], L. A. Lyutikova[2], E. A. Korchagina[1], and Z. S. Nirova[1]

[1] Center for Geographical Research, Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences, Nalchik, Russia

[2]Institute of Applied Mathematics and Automation, Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences, Nalchik, Russia

* **Correspondence to:** Maryana Gedueva, m.gyaurgieva@mail.ru

The paper presents a comprehensive analysis of mudflow basin parameters, conducted using machine learning methods. For the northern slope of the Greater Caucasus, data on the main parameters of mudflow basins were analyzed to build models that allow forecasting mudflows with certain characteristics. A set of machine learning methods was used (clustering, search for association rules, logistic regression, etc.). Key factors of mudflows were identified, models were developed for classifying mudflow types and predicting the volume of one-time removal of material, and a number of association rules with high reliability were identified that describe the relationships between factors influencing mudflow processes. The obtained results show great potential in the application of learning in the tasks of analysis and forecasting of mudflow processes. Ultimately, this will allow, based on the addition of mudflow data and updating of existing mudflow maps, to develop more effective measures to reduce the impact of mudflows on the environment to a minimum.

Keywords: Mudflow, mudflow basin, mudflow activity, mudflow formation factors, machine learning methods, analysis models, clustering, multiparameter regression, association discretization rules.

Citation: Gedueva M. M., Kyul E. V., Lyutikova L. A., Korchagina E. A., and Nirova Z. S. (2025), Comprehensive Analysis of Mudflow Data Using Machine Learning Methods, *Russian Journal of Earth Sciences*, 25, ES6003, EDN: RTNMVL, https://doi.org/10.2205/2025ES001072

Introduction

Studying the factors influencing the formation and characteristics of mudflows is a pressing issue in the context of climate change and the increasing frequency of extreme weather events [*Khvorostov*, 2004; *Kondratieva et al.*, 2015].

This paper presents the results of a study aimed at studying the characteristics of mudflows using modern machine learning methods.

The purpose of the study is to identify key factors influencing the occurrence of mudflows, as well as to develop a model for classifying mudflow types and predicting the volume of one-time material removal. Currently, due to the active development of the mountainous part of the Caucasus, research into hazardous natural processes, including mudflows, is becoming a priority.

The object of the study is the northern slope of the Greater Caucasus – the territories of the republics of the Russian Federation: Dagestan (RD), Chechen Republic (CHR), Ingushetia (RI), North Ossetia–Alania (NO–Alania), Kabardino-Balkaria (KBR), Karachay-Cherkessia (KCHR), Adygea (RA).

The subject of the study is the mudflow activity of the mountainous territories of Russia on the northern slope of the Greater Caucasus.

Research Article

Received: 2025-09-10
Accepted: 2025-11-10
Published: 2025-12-10

Practical Significance

Conducting a comprehensive analysis of data on the characteristics of mudflows using machine learning methods can be useful, since it will allow for a better understanding of the nature and mechanisms of mudflow formation, identifying key factors influencing the characteristics of mudflows, building predictive models for assessing the risks and consequences of mudflows, and grouping mudflows by similar characteristics for further study [*Lombardo and Mai*, 2018]. The results of such a study can be applied in engineering practice, in planning and managing mudflow-hazardous areas, as well as for scientific purposes to deepen knowledge about mudflow processes [*Rahmati et al.*, 2019].

Materials and methods of research. The main methods in the work were machine learning.

To build various models for analyzing the characteristics of mudflows, the Mudflow Hazard Cadastre of the South of the European Part of Russia 2015 edited by N. V. Kondratieva was used, which is a system for collecting, processing, storing and analyzing data on mudflow processes in a certain territory [*Kondratieva et al.*, 2015]. In our case, this is the territory of the northern slope of the Greater Caucasus [*Korchagina et al.*, 2021; *Kyul et al.*, 2019; *Nirova et al.*, 2024].

Discussion of Results

Let us analyze some parameters of mudflow basins in the study area, based on data from the Mudflow Hazard Cadastre of the South of the European Part of Russia (Tables 1–3, Figures 1–4).

Table 1. Generalized parameters of mudflow basins by study regions

Name of the subject of the Russian Federation	Total area of basins, km²	Total length of channels, km	Average slope of channels, ‰, max
Eastern Caucasus			
Republic of Dagestan	12488,3	4267,6	658
Chechen Republic	2432,7	506,3	335
Republic of Ingushetia	690,6	179,9	333
Total indicators	15611,6	4953,8	1326
Central Caucasus			
North Ossetia–Alania	2165	792,7	862
Kabardino-Balkaria	4972,7	1738,3	1150
Total indicators	7137,7	2531	2012
Western Caucasus			
Karachay-Cherkessia	2143	1169,8	1223
Republic of Adygea	362,3	313	375
Total indicators	2505,3	1169,8	1598

Building Analysis Models

The key elements of the mudflow cadastre are the mudflow genesis (categorical), mudflow type (categorical), basin area, S, km² (numerical), average channel slope, α (numerical), river length, L, km (numerical), source height (numerical), maximum one-time removal volume, W, m³ (numerical), maximum volume of solid mudflow deposits, m³ (numerical). The machine learning models considered in this paper will analyze data based on these characteristics [*Kondratieva et al.*, 2015].

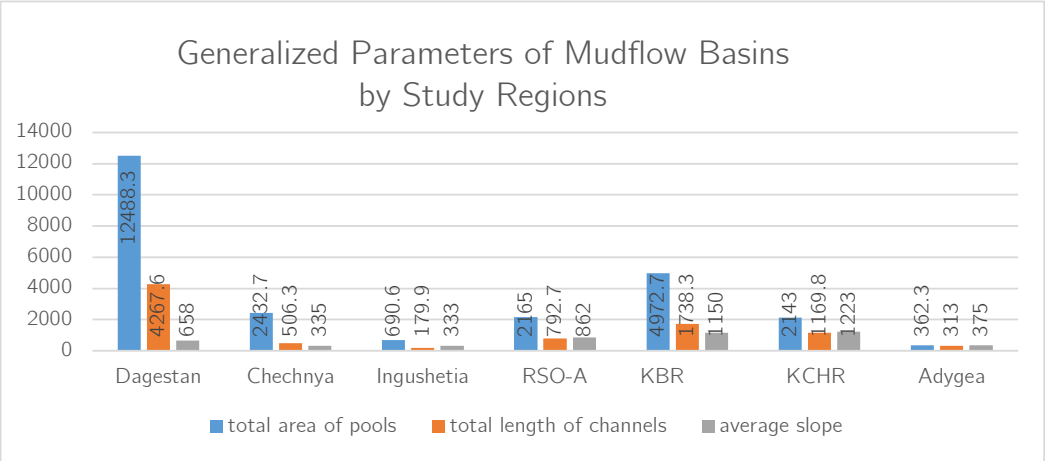

Figure 1. Summary indicators of mudflow basin characteristics in the republics of the northern slope of the Greater Caucasus.

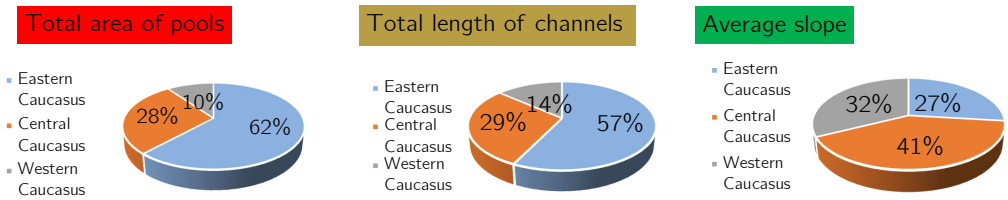

Figure 2. Summary indicators of mudflow basin characteristics: Eastern, Central, Western Caucasus.

Data Clustering Model

Clustering of mudflow data can be used to identify high-risk zones, as it allows dividing the territory into zones with different levels of mudflow risk. And also to analyze the relationships between mudflow processes and other factors. This allows for a better understanding of the causes of mudflows and the development of more effective measures to prevent them. To solve the problem of mudflow clustering with known basin characteristics (area, channel slope, channel length, volume of solids removed), a clustering algorithm, such as *K*-Means, can be used.

As a result of data processing, the Elbow method was used to obtain the optimal number of clusters. As a result, three clusters were obtained with the following characteristics:

Table 2. Number of mudflow basins by mudflow genesis (by study regions)

Name of the subject of the Russian Federation	Genesis of the mudflow					
	D	SD	L	LD	No data	Total Basins
Republic of Dagestan	486	26	2	20	0	487
Chechen Republic	37	4	1	9	2	44
Republic of Ingushetia	23	2		6	0	23
Republic of North Ossetia–Alania	102	–	22	32	0	119
Kabardino-Balkarian Republic	199	12	36	58	0	232
Karachay-Cherkess Republic	269	1	11	40	0	281
Republic of Adygea	29	–	2	8	0	29

Note: Genesis of the water component: D – rain, SD – snow-rain, L – glacial, LD – glacial rain.

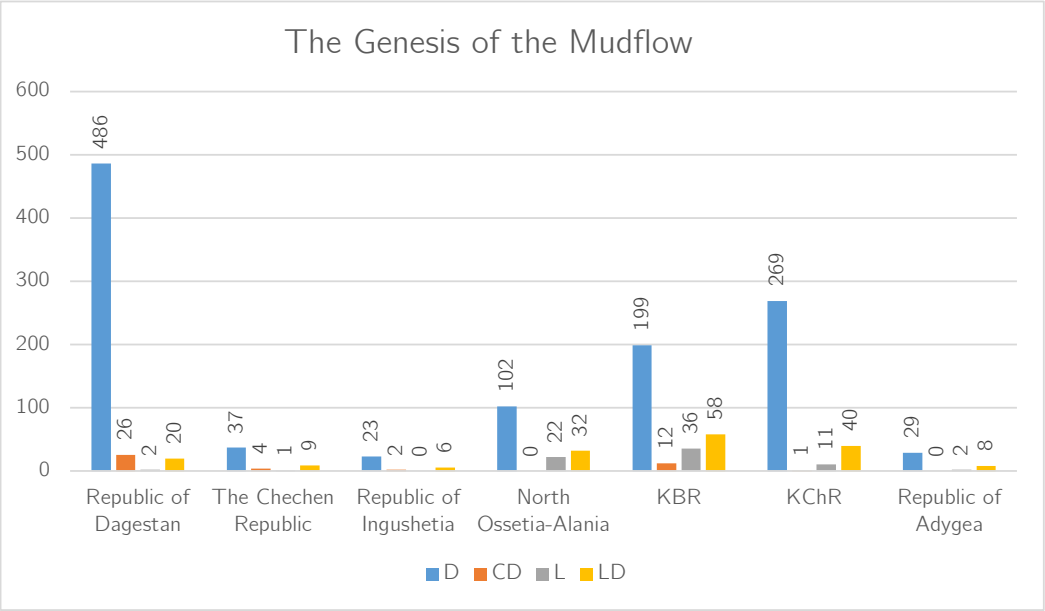

Figure 3. Number of mudflow basins by mudflow genesis (by study regions).

Cluster 0:

- Basin area: 21.20 km² – the average basin size is relatively small.
- Mudflow volume (M1): 274,713 m² – the average mudflow volume is quite significant.
- Slope: 6.86 – the average channel slope is quite steep.
- Source height: 3090.75 m – mudflows in this cluster start from a fairly high height.
- Genesis of mudflow: Rainfall (D) – this type of mudflow is formed as a result of intense rainfall.
- Mudflow type: mud-rock (GK) – consists of a mixture of mud, sand, gravel and stones of various sizes. The ratio of mud to stones can vary depending on the conditions of mudflow formation.

Cluster 1:

- Basin area: 589 km² – this cluster is characterized by significantly larger basins.
- Mudflow volume (M1): 50,000 m² – the average volume of mudflows is relatively small.
- Slope: 13 – Low slope, which is not typical for mudflows.
- Source height: 3100 m – mudflows in this cluster start from a high altitude.

Table 3. Number of mudflow basins by mudflow type (by study region)

Name of the subject of the Russian Federation	Mudflow type				
	GK	VK	NV	LK	No data
Republic of Dagestan	87	26	–	–	378
Chechen Republic	23	5	–	–	23
Republic of Ingushetia	8	–	–	–	2
Republic of North Ossetia–Alania	96	39	–	1	19
Kabardino-Balkarian Republic	152	123	–	–	1
Karachay-Cherkess Republic	255	2	88	–	–
Republic of Adygea	11	19	–	–	2

Note: Mudflow type by granulometric composition: GK – mud-rock, VK – water-rock, NV – alluvial, LK – ice-rock, "–" – no data.

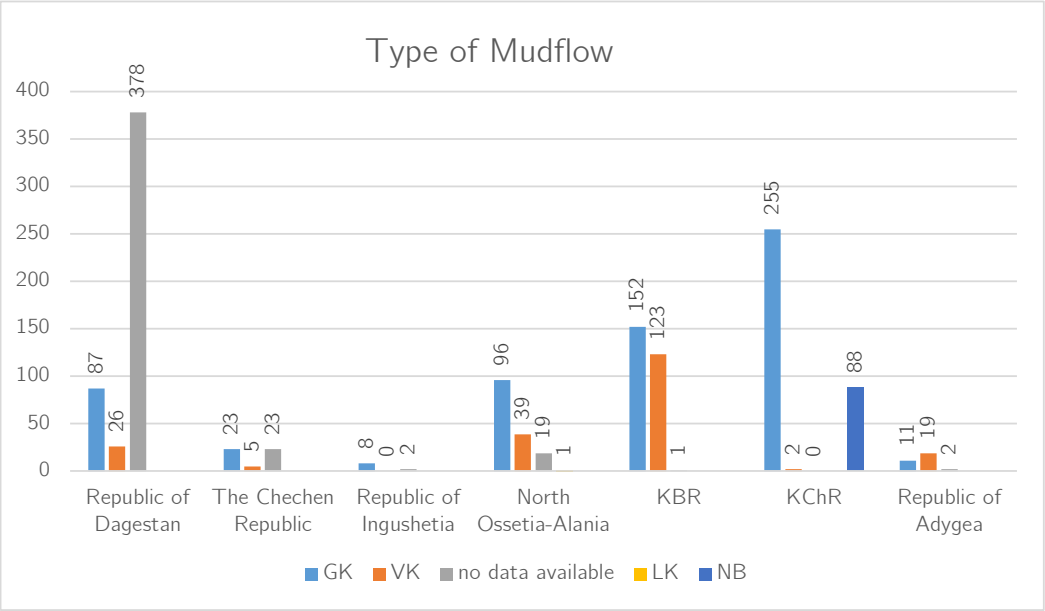

Figure 4. Number of mudflow basins by mudflow type (by study region).

- Genesis of mudflow. Glacial-rain, rain (L–D; D) – this cluster has mixed types. Glacial-rain, rain mudflows are a special type of mudflows that occur in mountainous areas where there are glaciers. They combine the characteristics of rain mudflows and mudflows formed by melting glaciers.
- Mudflow type: water-rock and mud-rock (GKVK) – this may mean that the mudflows in this cluster are mixed – water-rock and mud-rock (GK and VK), containing both large debris and finely dispersed material.

Cluster 2:

- Basin area: 15.78 km^2 – the average basin size is relatively small.
- Mudflow volume (M1): 39,922.5 m^2 – the average volume of mudflows is significant.
- Slope: 6.5 – the average slope of the riverbed is quite steep.
- Source height: 1774.03 m – mudflows in this cluster start at a lower height compared to cluster 0.
- Genesis of mudflow: Rain (D).
- Mudflow type: Water-rock (VK).

It can be concluded that cluster 1 is distinguished by large basins and a small slope, which is not typical for mudflows, clusters 0 and 2 differ in source height and mudflow volume, but have similar genesis and mudflow type.

Figure 5 shows clustering by features.

Figure 6 shows a visual representation of the characteristics of each cluster obtained by clustering the data. It helps to understand how the clusters differ from each other and which features are most important for separating them.

Visualization of the results of data clustering, where each cluster is represented by a set of points on a graph, is shown in Figure 7.

Overall, the clustering results indicate a relatively weak structure in the data, with three clusters that differ in physical characteristics and flow types. Further analysis may be needed to better understand the structure of the data and determine the optimal number of clusters [*Radeev*, 2021].

Building a Multivariable Regression Model

Multivariable regression is built to predict a dependent variable (target) using multiple independent variables (predictors).

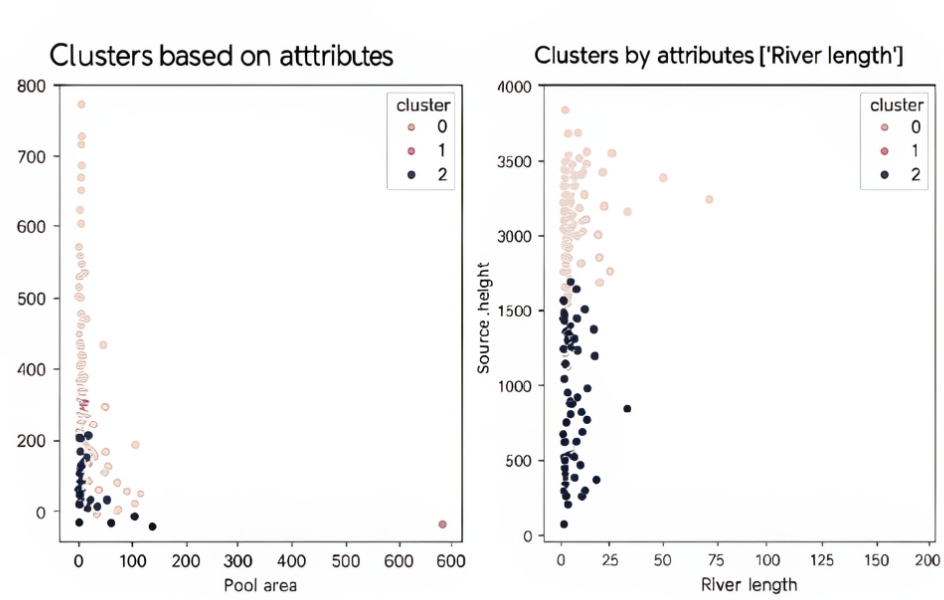

Figure 5. Clustering by features.

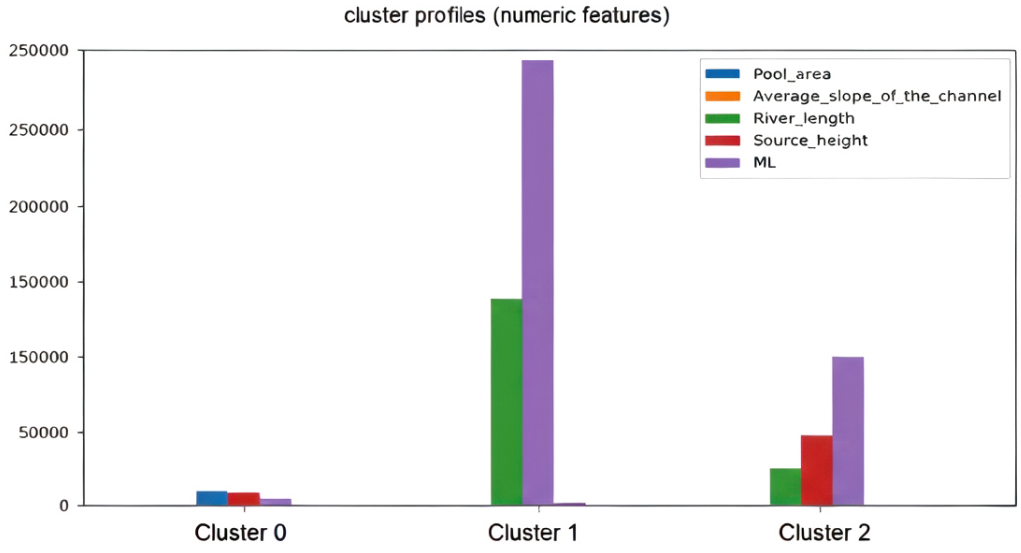

Figure 6. Cluster profile.

This is important when studying the relationship between variables, as it helps to understand how multiple variables work together to influence the target variable. Multivariable regression is used to predict future values of the target variable based on the values of the independent variables, allowing us to determine which independent variables are most important in predicting the target variable. In our model, the target variable corresponds to the value of the maximum one-time removal volume ('M1').

The results of building a multivariable regression model were generally predictable. The built multivariable regression model does not adequately describe the data. MSE = 92,477,727,488.7331: This is a very large MSE value, indicating significant errors in the model's predictions.

R-squared = 0.1235: This is a low *R*-squared value, which means that only about 12% of the variation in the target variable is explained by the model. The remaining 88% of the variation remains unexplained.

The inadequacy of the model is clear from the graphs in Figures 8–10.

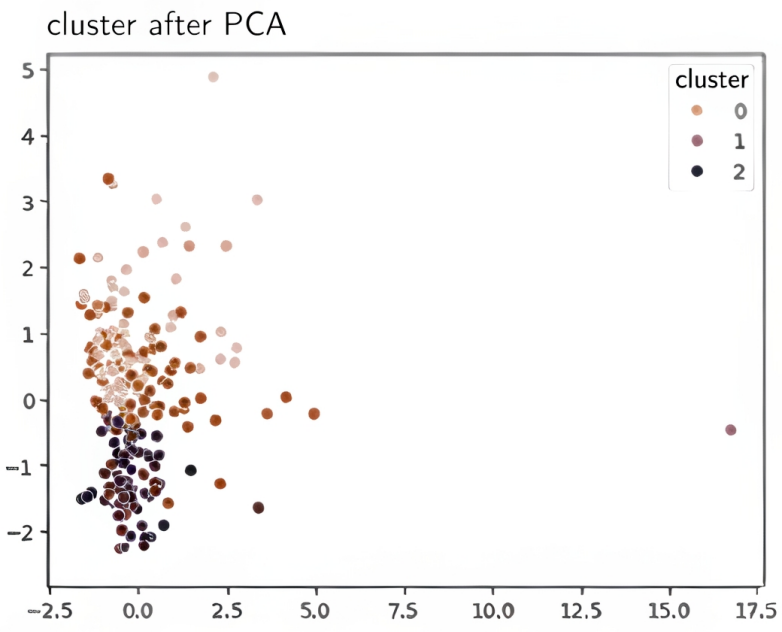

Figure 7. Mudflow clusters.

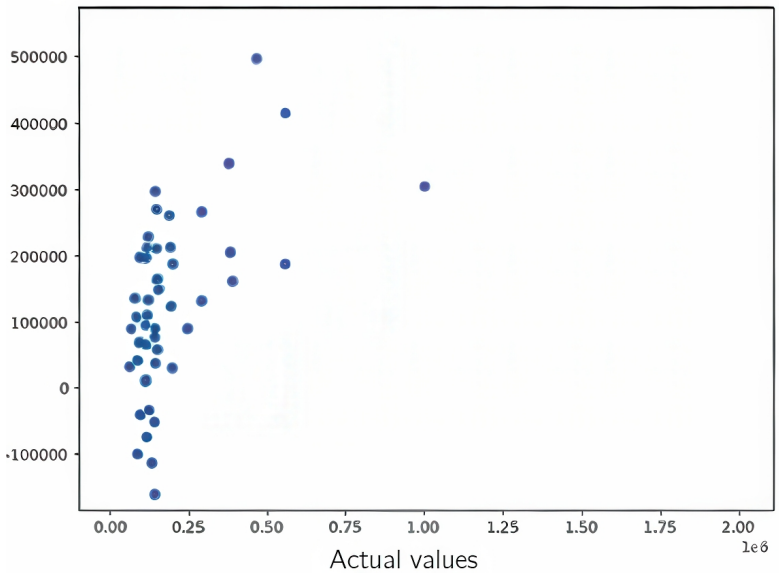

Figure 8. Scatter plot of predictions.

It is clear from the graphs that the reason for the poor model is the nonlinear relationships between the predictors and the target variable, which linear regression cannot capture.

To solve the problem of nonlinearity, it was decided to categorize the numerical data [*Zhuravlev*, 1978].

Building Models With Categorical Data

Dividing the numerical data into categories simplifies the analysis and interpretation of the results. Since instead of analyzing continuous values, you can get, for example, three discrete groups that are easier to compare and interpret. In our case, this is useful for analyzing the relationships with the categorical variables genesis and mudflow type ("Genesis_of_mudflow", "Mudflow_type").

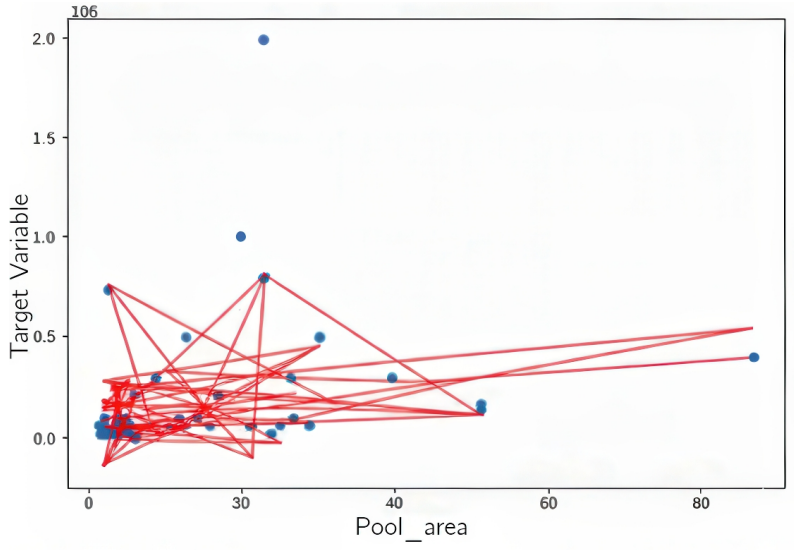

Figure 9. Dependence of the target variable.

Since our numerical data showed nonlinearity, moving to categories can help identify nonlinear relationships and reveal the influence of features. Correctly chosen categories can highlight important features.

In your case, you suggested dividing each field with numerical data into three categories: "0 – small", "1 – medium" and "2 – large". This is similar to quantile partitioning, where the range is divided into equal parts based on the number of observations. This may not be the most expressive way to categorize the data, and it certainly simplifies the data, but overall it follows the approach. We are turning a complex continuous value into a simpler categorical value.

Categorical variables are less sensitive to outliers and noise in the data, which makes the models more stable. Discretization allows us to account for nonlinearities by breaking the range into parts where the relationship can be approximated as linear.

Now the regression problem in the previous section is reduced to a classification problem, since the target variable is now categorical. Instead of predicting a continuous value, the maximum lump sum 'M1', we now predict which of three categories (0, 1, or 2) 'M1' belongs to.

After building a classification model using a decision tree, we obtained the following impressive results:

accuracy: 1.0

	precision	recall	f1-score	support
60.0	1.00	1.00	1.00	16
61.0	1.00	1.00	1.00	17
62.0	1.00	1.00	1.00	13
accuracy			1.00	46
macro avg	1.00	1.00	1.00	46
weighted avg	1.00	1.00	1.00	46

Figure 10. Result of classification of the maximum one-time removal volume.

These results reflect classification quality metrics, where accuracy is 1.0 or 100%.

Key Interpretation Points

Accuracy: 1.0. This means that all objects were correctly classified by the model, i.e., all model predictions matched the actual class labels.

Per-Class Metrics

For class 60.0: small MLP.

Precision = 1.00 (the proportion of correctly classified objects among all objects predicted to belong to this class).

Recall = 1.00 (the proportion of correctly classified objects among all objects that actually belong to this class).

F1-score = 1.00 (the harmonic mean of precision and recall).

Support = 16 (the number of objects in this class).

Similarly for classes 61.0 medium MLP and 62.0 large MLP.

General conclusion: the results indicate that the classification model shows ideal quality, correctly predicting the belonging of all objects to their respective classes. Such a high level of accuracy may indicate a good match between the features and classes used.

Building Association Rules

Association analysis is a data analysis method used to identify relationships between variables in data. It allows finding association rules that describe the joint occurrence of variable values in data [*Flach*, 2015]. This is done using the frequency of occurrence of an element or group of elements in the data, the proportion of transactions in which a certain set of elements appears, the probability that a certain element will appear in a transaction if another element is already present in it. We will use association analysis to find relationships between factors affecting mudflow processes, which can help in predicting risk and developing prevention measures.

The FP-Growth algorithm was used to find association rules.

After the algorithm ran, the most important rules were shown (Figure 11).

the most important associative rules

	antecedents	consequents	antecedent support	\
93735	(31, nan)	(62, 52, GK)	0.008621	
93733	(62, 52, GK, 42)	(31, 6, nan)	0.008621	
178007	(31, nan, 52, 42)	(62, GK)	0.008621	
27987	(31, 62)	(nan, GK)	0.008621	

	consequent support	support	confidence	lift	leverage	conviction	\
93735	0.008621	0.008621	1.0	116	0.008546	inf	
93733	0.008621	0.008621	1.0	116	0.008546	inf	
178012	0.008621	0.008621	1.0	116	0.008546	inf	
178007	0.008621	0.008621	1.0	116	0.008546	inf	
27987	0.008621	0.008621	1.0	116	0.008546	inf	

	zhangs_metric
93735	1.0
93733	1.0
178012	1.0
178007	1.0
27987	1.0

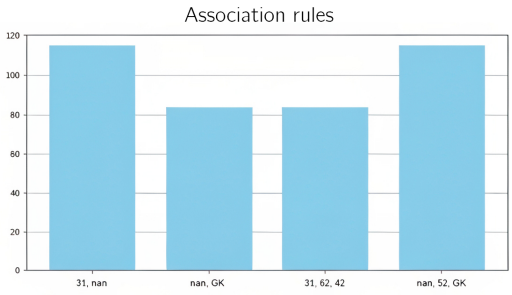

Figure 11. Basic association rules.

The quality of rules is assessed by the following characteristics.

Antecedents are a set of elements (in this case, numeric values) that together precede the occurrence of a certain set of elements in the Consequents. Consequents are a set of elements that are associated with the Antecedents.

Support for antecedents shows how often a set of elements occurs in the antecedents. Certainty shows how likely it is that if a set of elements occurs in the antecedents, then a set of elements in the consequents also occurs.

Lift shows how much more or less likely the occurrence of consequents is in the presence of antecedents, compared to the occurrence of consequents as a whole.

Leverage shows how strongly the antecedents and consequents are related to each other, compared to their individual occurrences. Conviction shows how likely it is that the rule is not random, that is, that the antecedents and consequents are truly related. Zhang's Metric is a comprehensive assessment of the quality of an association rule that takes into account support, confidence, and lift.

Analysis of the Quality of the Obtained Rules

All the rules identified as a result of the algorithm have very high confidence (1.0), that is, if a set of elements in the antecedents is found, then a set of elements in the consequents is guaranteed to be found. The lift for all rules is very high (116.0), which means that the probability of consequents occurring in the presence of antecedents is significantly higher than the probability of consequents occurring in general. Leverage and conviction are also very high, showing a strong connection between antecedents and consequents. Zhang's Metric is 1.0 for all rules, which indicates their high quality.

Thus, it can be argued that the presented association rules describe very strong and reliable connections between sets of elements in the antecedents and consequents.

The last rule states that with an average basin area and a large volume of solid mudflow deposits, a mudflow is highly likely to occur. Combining all five rules using Boolean algebra operations [*Lyutikova*, 2023], we come to the following conclusion: a mudflow, even with an average basin area, is characterized by a large volume of maximum one-time removal and a large maximum volume of solid deposits.

Large, small, medium are meant within the framework of our division.

Medium basin from 12.64 to 58.45 km^2, large volume of maximum one-time removal from 38,800.00 m^3, large maximum volume of solid deposits from 102,840.08 m^3.

Of course, it should be understood that the obtained rules are not a law of nature, they are rather strong associative rules, and not laws in the strict sense. Laws assume absolute truth, while these rules show statistical correlations that will be fulfilled with a high probability, but not necessarily in 100% of cases. Conclusions

A comprehensive analysis of mudflow data using machine learning methods allowed us to identify key factors in the presented data that determine their occurrence and characteristics. As part of the study, models were developed for classifying mudflow types and predicting the volume of one-time material removal.

The results of the study demonstrate a significant superiority of neural network-based models over traditional classification and forecasting methods. This indicates a great potential for using deep learning in the analysis and forecasting of mudflow processes.

However, to improve the accuracy and versatility of the models, it is necessary to continue research in the direction of improving data preprocessing. Particular attention should be paid to the development of hybrid models that combine the advantages of neural networks and other machine learning methods, which will improve the integration and interpretation of the results.

For a deeper understanding of the mechanisms of mudflow formation and dynamics, it is necessary to expand the range of input data for modeling, taking into account such parameters as the total amount of precipitation for a certain period, vegetation characteristics, anthropogenic factors, geomorphology and geological structure of the basin.

The results obtained during the study can be used to solve a number of priority tasks: 1) supplementing mudflow data; 2) updating cartographic material on their basis; 3) improving systems for monitoring and forecasting mudflow processes; 4) developing more effective measures to prevent and mitigate the consequences of mudflows.

Acknowledgments. The work was carried out within the framework of the state assignment of KBSC RAS.

References

Flach P. Machine learning: the science and art of building algorithms that extract knowledge from data. — M. : DMK Press, 2015. — 400 p. — (In Russian).

Khvorostov V. V. Extraordinary and ultra-mudflows in the Greater Caucasus // Proceedings of the V International Conference "Sustainable Development of Mountainous Areas: Problems and Prospects for the Integration of Science and Education". — 2004. — P. 248–286. — (In Russian).

Kondratieva N. V., Adzhiev A. Kh., Bekkiev M. Yu., et al. Mudflow hazard cadastre for the South of European Russia. — M.; Nalchik : Feoriya, 2015. — 148 p. — EDN: VHKUJH ; (in Russian).

Korchagina E. A., Gedueva M. M., Ataev Z. V., et al. Geoecological Research in the Territory of The Northern Slope of the Great Caucasus // News of the Kabardin-Balkar Scientific Center of RAS. — 2021. — Vol. 2, no. 100. — P. 126–138. — https://doi.org/10.35330/1991-6639-2021-2-100-126-138. — (In Russian).

Kyul E. V., Ezaov A. K. and Kankulova L. I. The Theoretical Basis of Geo-Environmental Monitoring of the Mountain Geosystems // Sustainable Development of Mountain Territories. — 2019. — Vol. 11, no. 1. — P. 36–43. — https://doi.org/10.21177/1998-4502-2019-11-1-36-43. — (In Russian).

Lombardo L. and Mai P. M. Presenting logistic regression-based landslide susceptibility results // Engineering Geology. — 2018. — Vol. 244. — P. 14–24. — https://doi.org/10.1016/j.enggeo.2018.07.019.

Lyutikova L. A. Methods for Improving the Efficiency of Neural Network Decision-Making // Advances in Automation IV. RusAutoCon 2022. Lecture Notes in Electrical Engineering, vol 986. — Springer International Publishing, 2023. — P. 294–303. — https://doi.org/10.1007/978-3-031-22311-2_29.

Nirova Z. S., Kyul E. V., Baidaeva Z. R., et al. Assessment of the River Basins State in Prielbrusye National Park // Dagestan State Pedagogical University. Journal. Natural and Exact Sciences. — 2024. — Vol. 18, no. 1. — P. 59–69. — https://doi.org/10.31161/1995-0675-2024-18-1-59-69. — EDN: LPPDGN ; (in Russian).

Radeev N. A. Avalanches Forecasting Using Machine Learning Methods // Vestnik NSU. Series: Information Technologies. — 2021. — Vol. 19, no. 2. — P. 92–101. — https://doi.org/10.25205/1818-7900-2021-19-2-92-101. — (In Russian).

Rahmati O., Kornejady A., Samadi M., et al. PMT: New analytical framework for automated evaluation of geo-environmental modelling approaches // Science of The Total Environment. — 2019. — Vol. 664. — P. 296–311. — https://doi.org/10.1016/j.scitotenv.2019.02.017.

Zhuravlev Yu. I. On an algebraic approach to solving recognition or classification problems // Problems of Cybernetics. Vol. 33. — Nauka, 1978. — P. 5–68. — (In Russian).