E3S Web of Conferences **583**, 01009 (2024) https://doi.org/10.1051/e3sconf/202458301009
ITESE-2024

Machine learning estimation of rock masses displacement

V.V. Kukartsev[1,2], *I.I.* Kleshko[2*], *N.A.* Dalisova[3], and *V.V.* Khramkov[2]

[1]Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia
[2]Artificial Intelligence Technology Scientific and Education Center, Bauman Moscow State Technical University, Moscow, Russia
[3]Krasnoyarsk State Agrarian University, Krasnoyarsk state agrarian university, 660049, Krasnoyarsk, Russia

Abstract. This paper presents a comprehensive analysis of the factors affecting landslide occurrence in Iran based on a dataset containing information on more than 4000 landslide cases. Both natural (slope, height, rainfall, distance to rivers and faults) and anthropogenic (type of land use) factors were studied. A random forest model was used to predict landslide risk and assess the significance of various factors. The results show that the most significant factors are terrain slope, elevation and distance to water bodies and tectonic faults. These findings can be used to develop preventive measures and improve landslide risk management strategies in the region.

1 Introduction

Landslides are one of the most common and destructive natural disasters in the world. In Iran, as in other countries, they often cause significant infrastructure damage, economic losses and loss of life [1-39]. To reduce risks and improve landslide prediction methods, it is important to understand what factors influence their occurrence. In this paper, we analyze a dataset that includes various natural and anthropogenic factors contributing to landslides in Iran.

This dataset includes information and factors affecting landslides in Iran. The dataset contains more than 4000 landslide hazards in Iran, each characterized by multiple factors that can contribute to landslides.

These factors include natural factors such as slope, climate, tectonic activity, etc. as well as human factors such as land use.

2 Description of the data set

These factors can have a significant impact on the likelihood of a landslide occurring. For example, slope is one of the most important natural factors affecting landslides. Landslides are more likely to occur on steep slopes.

This dataset contains 11 columns as follows:

IDENTIFICATOR

* Corresponding author: ilya-kleshko@mail.ru

LONG: longitude of the landslide point.
LAT: Latitude of the landslide point.
SUB_Basin: The name of the watershed of the landslide point.
Elevation: The elevation of the landslide point above sea level (m).
AAP(mm): Average annual precipitation at the landslide point.
RiverDIST(m): Distance of the landslide point from the river.
FaultDIST(m): Distance of the landslide point from a fault.
Landuse_Type: the type of land use at the landslide point.
Slope (percent): Slope at the landslide point (range of values from 0 to 100%).
Slope (degrees): Slope at the landslide point (range of values from 0 to 90).
GEO_UNIT: Geologic unit of the landslide point.
DES_GEOUNI: Description of the geologic unit of the landslide point.
Climate_Type: Climate type of the landslide point.
DES_ClimateType: Description of the climate type of the landslide point.
Data loading and preliminary data analysis

In the first step, we will load the data and perform an initial analysis of the data to understand the structure of the data and identify possible problems such as missing values.

Before proceeding with the analysis, it is necessary to perform data processing:

Missing values: check for missing values and, if necessary, fill them in or delete the corresponding records.

Coding of categorical variables: translate categorical variables into numerical format for further analysis and modeling.

Data normalization: let's normalize the numerical data to improve the performance of machine learning algorithms.

First, let's load the data and do some preliminary analysis.

```
Первых несколько строк данных:
   ID    LONG     LAT SUB_Basin  Elevation  AAP(mm)  RiverDIST(m)  \
0   1  52.326  27.763    Mehran      617.0      137   1448.705292
1   2  52.333  27.772    Mehran      944.0      137    344.299484
2   3  52.326  27.763    Mehran      617.0      137   1448.705292
3   4  52.333  27.694    Mehran       55.0      137   1889.828623
4   5  52.324  27.682    Mehran       20.0      137    874.201691

   FaultDIST(m)      Landuse_Type  Slop(Percent)  Slop(Degrees) GEO_UNIT  \
0   40639.57890         poorrange      42.240669      22.899523  EOas-ja
1   40135.02913  mix(woodland_x)      68.219116      34.301464  KEpd-gu
2   40639.57890         poorrange      42.240669      22.899523  EOas-ja
3   42189.54442              rock      12.141766       6.922833    Mlmmi
4   43010.08400         poorrange       2.216230       1.269598   MuPlaj

                                       DES_GEOUNI Climate_Type  \
0  Undivided Asmari and Jahrum Formation , regard...       A-M-VW
1                     Keewatin Epedotic quartz diorite       A-M-VW
2  Undivided Asmari and Jahrum Formation , regard...       A-M-VW
3   Low weathering grey marls alternating with ba...       A-M-VW
4   Brown to grey , calcareous , feature - formin...       A-M-VW

                                    DES_ClimateType
0  Warm and humid, with a humid period longer tha...
1  Warm and humid, with a humid period longer tha...
2  Warm and humid, with a humid period longer tha...
3  Warm and humid, with a humid period longer tha...
4  Warm and humid, with a humid period longer tha...

Список столбцов в данных:
Index(['ID', 'LONG', 'LAT', 'SUB_Basin', 'Elevation', 'AAP(mm)',
       'RiverDIST(m)', 'FaultDIST(m)', 'Landuse_Type', 'Slop(Percent)',
       'Slop(Degrees)', 'GEO_UNIT', 'DES_GEOUNI', 'Climate_Type',
       'DES_ClimateType'],
      dtype='object')
```

Fig. 1. Preliminary analysis

After that, data preprocessing can be done. For example:

1. Checking and processing missing values.

2. Coding of categorical variables (if necessary).
3. Normalization or standardization of numerical data.

Next, an analysis of correlations between variables can be performed:

Based on this analysis, it will be possible to understand which factors most strongly influence landslide risk.

Different algorithms can be used to build machine learning models to predict landslide risk, for example:

Data loading and preliminary analysis: Data were loaded and initial analysis was performed to understand the structure of the data and identify possible problems such as missing values.

Processed missing values, coded categorical variables and normalized the numerical data to make it suitable for machine learning.

Building and visualizing the correlation matrix helped to identify strong and weak relationships between different variables, which can be useful for building models and interpreting results.

Histograms and boxplots showed the distribution and possible outliers in the numerical data, which helps to better understand the structure of the data.

3 Data analysis

A random forest model was trained to predict landslide risk. The evaluation of the model showed the accuracy and importance of different attributes, which helps to understand which factors are most significant in predicting landslides.

Several graphs were obtained as a result.

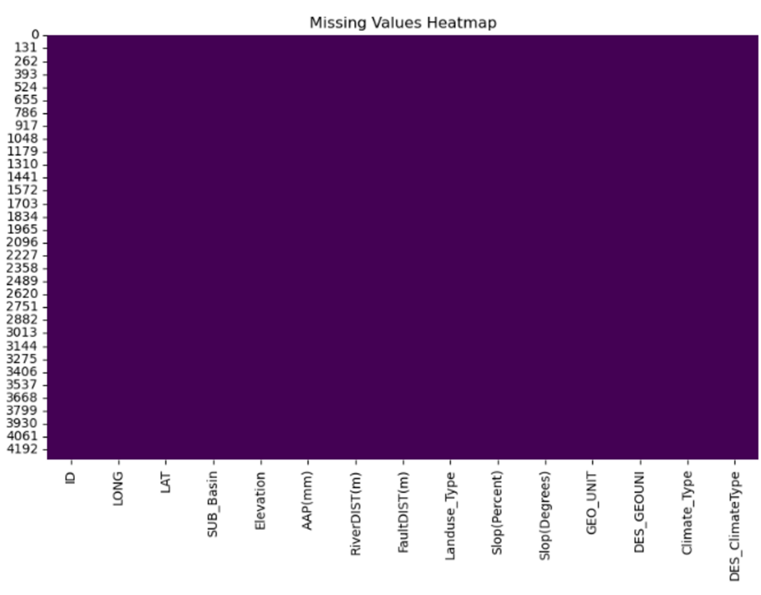

Fig. 2. Heat map of missing values

This heatmap visualizes the presence of missing values in the dataset.

A dark purple color indicates the presence of data (i.e., no missing values).

The absence of other colors suggests that there are no missing values in any columns in your dataset.

To identify relationships between variables, we will construct and visualize a correlation matrix. The correlation coefficients will allow us to understand which factors are most strongly associated with each other and with landslide risk.

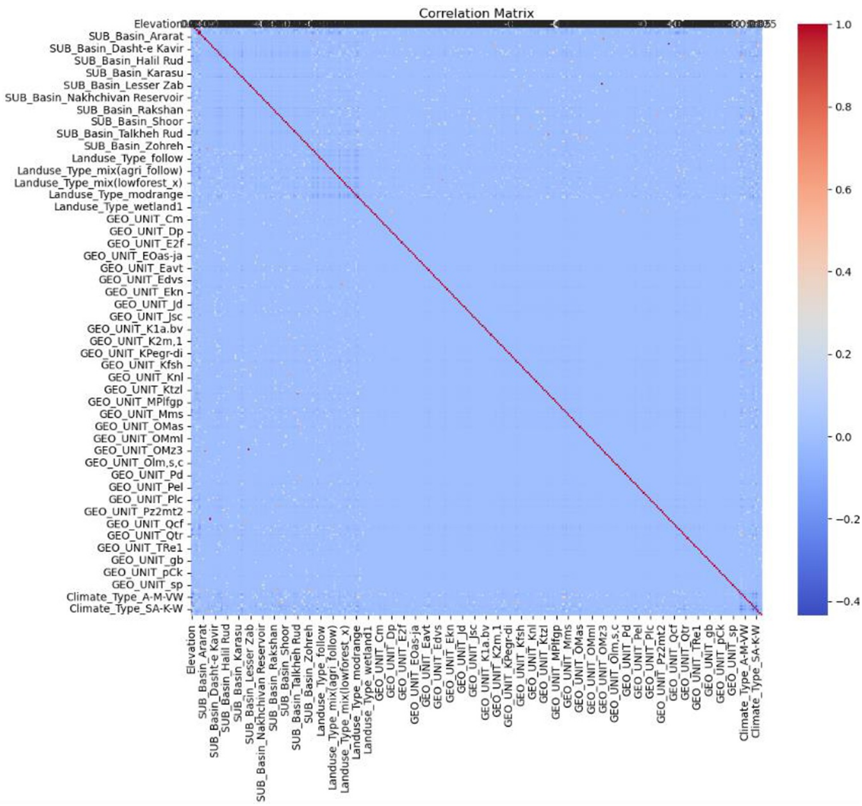

Fig. 3. Correlation matrix

This matrix shows the correlation coefficients between pairs of variables in the dataset.

The color scale on the right indicates the strength and direction of the correlations, with values ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation).

Diagonal elements are 1 (each variable is perfectly correlated with itself).

Blue colors indicate positive correlations and red colors indicate negative correlations. The lighter the blue, the stronger the positive correlation; the lighter the red, the stronger the negative correlation.

The graph shows that most of the variables have low correlation with each other because the matrix is mostly blue with little variation.

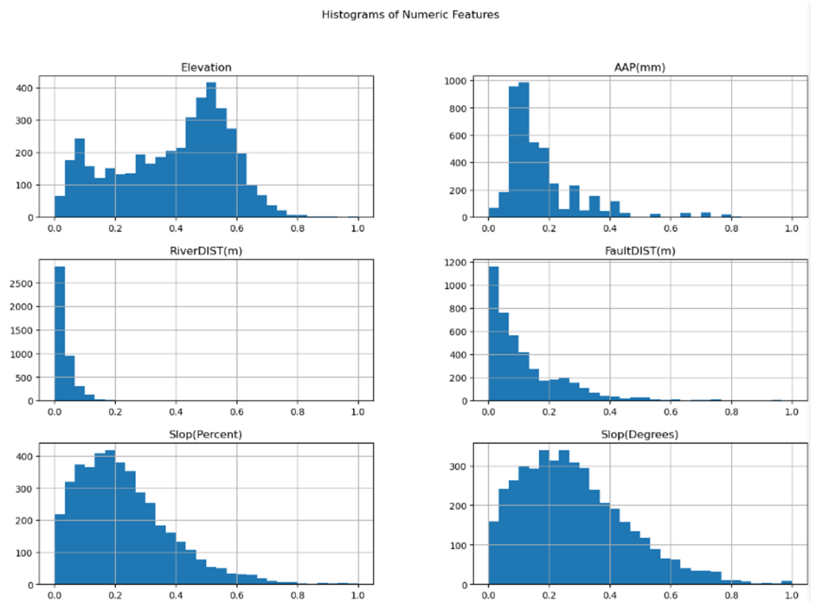

Fig. 4. Histogram of numeric features

These histograms provide a visual representation of the distribution of each numeric feature after normalization.

Elevation: The data is distributed over the entire range, with a peak in the middle.

AAP(mm): Shows a right-sided distribution with most values at the bottom.

RiverDIST(m): Strongly right-sided distribution, indicating that most values are centered at the bottom with a long tail.

FaultDIST(m): Also right-sided distribution, similar to RiverDIST(m).

Slop(Percent) and Slop(Degrees): Both have a somewhat right-handed distribution, but are more evenly distributed than the previous features.

4 Conclusion

The conducted data analysis on factors affecting landslides in Iran identifies the key variables affecting the risk of landslides. The random forest model showed that the most significant factors are terrain slope, elevation and distance to rivers and faults. These results can be used to develop more effective measures for landslide prevention and mitigation.

References

1. Martyushev N.V. et al., Energies **16**, 2, 729 (2023)
2. Shutaleva A. et al., Sustainability **15**, 4, 3011 (2023)
3. Rezanov V.A. et al., Metal **12**, 121 2135 (2022)
4. Martyushev N.V. et al., Materials **16**, 9, 3490 (2023)
5. Kukartsev V.A. et al., Metals **13**, 2, 337 (2023)
6. Degtyareva K. et al., *Use of Computer Simulation Tools to Simulate Processes at the Foundry*. In 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE (2024)

7. Degtyareva K. et al., *Automated System for Accounting of Customers and Orders.* In 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-4). IEEE (2024)
8. Golik V.I. et al., MIAB. Mining Inf. Anal. Bull. **(11-1):** 175-189 (2023)
9. Panfilova T.A. et al., MIAB. Mining Inf. Anal. Bull. **(11-1):** 239-251 (2023)
10. Suprun E. et al., BIO Web of Conferences **84**. 01008 (2024)
11. Orlov V. et al. E3S Web of Conferences **460**. 07002 (2023)
12. Kravtsov K. et al. E3S Web of Conferences **458**. 09022 (2023)
13. Tynchenko V.S. et al. E3S Web of Conferences **458**. 01011 (2023)
14. Nelyub V.A. et al. (2023). *Correlation Analysis and Predictive Factors for Building a Mathematical Model.* In Proceedings of the Computational Methods in Systems and Software (pp. 14-25). Cham: Springer International Publishing.
15. Degtyareva K.V. et al. E3S Web of Conferences **458**. 02002 (2023)
16. Gantimurov A. et al. E3S Web of Conferences **431**. 03005 (2023)
17. Tynchenko V.S. et al. E3S Web of Conferences **458**. 07003 (2023)
18. Tynchenko V. et al. *Application of U-Net Architecture Neural Network for Segmentation of Brain Cell Images Stained with Trypan Blue*. In International Conference on High-Performance Computing Systems and Technologies in Scientific Research, Automation of Control and Production (pp. 170-181). Cham: Springer Nature Switzerland (2024)
19. Martyushev N.V. et al. Materials, **16(13),** 4529 (2023)
20. Orlov V. et al. E3S Web of Conferences **460**. 07002 (2023)
21. Kravtsov K. et al. E3S Web of Conferences **458**. 09022 (2023)
22. Tynchenko V.S. et al. E3S Web of Conferences **458**. 01011 (2023)
23. Semenova E. et al. *Using UML to Describe the Development of Software Products Using an Object Approach.* In 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-4). IEEE (2022)
24. Tynchenko V.S. et al. AIP Conference Proceedings **2700**, 1 (2023)
25. Chernykh N. et al. *Comparative Analysis of Existing Measures to Reduce Road Accidents in Western Europe.* In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE (2023)
26. Volneikina E. et al. *Simulation-Dynamic Modeling Of Supply Chains Based On Big Data.* In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE (2023)
27. Filina O.A. et al. Energies **17**. 1, 17 (2023)
28. Boychuk, I. P., Grinek, A. V., Martyushev, N. V., Klyuev, R. V., Malozyomov, B. V., Tynchenko, V. S., Kondratiev, S. I. (2023). Energies, **16(24),** 8101.
29. Golik V.I. et al. Materials **16**. 21. 7004 (2023)
30. Malozyomov B.V. et al. Energies **16**. 13. 5046 (2023)
31. Malashin I.P. et al. Polymers **16**. 1. 115 (2023)
32. Malozyomov B.V. et al. Energies **16**. 13. 4907 (2023)
33. Tynchenko V.S. et al. Journal of Physics: Conference Series **2373**, 6, 062015 (2022)
34. Martyushev, N. V. et al. Materials, **16(13),** 4529 (2023)
35. Tynchenko V. et al. Mathematics, **12(2),** 276 (2023)

36. Rogova D. et al. *Software System for Modeling Temperature Distribution During the Electron Beam Welding.* In 2022 IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS) (pp. 1-6). IEEE. (2022)
37. Kurashkin S. et al. Procedia Computer Science, 200, 83-90 (2022)
38. Tynchenko V. et al. Procedia Computer Science, **200**, 843-851 (2022)
39. Tynchenko V. et al. *Software for modeling brazing process of spacecraft elements from widely used alloys.* In 2022 21st International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE (2022)