Project 0: Landslide Inventory Benchmarking

Mark Asch - ML-PREP

2025

landslide inventory

CONTEXT

Landslide Inventories

In many parts of the world today, landslide inventories are being, and have been produced. Recently, there is a lot of interest in the use of machine learning approaches to execute this task—see the recent review [ESR] and references therein. A benchmark is an essential tool for learning the LSM (landslide susceptibility mapping) methodology and then implementing new approaches, or applying it to new data.

In this preliminary project, we will:

- read and understand the benchmark paper [ESR];
- download the benchmark datasets from https://geomorphology.irpi.cnr.it/tools/slope-units that are to found in the section Details–Benchmark dataset and workflow;
- select and compare specific ML methods for generating susceptibility maps;

- compare with existing landslide inventories provided by PHIVOLCS and identify all similarities and differences.

Setup Instructions

1. Divide the ML-PREP team into 2 groups. Each group will work independently on the project and then we will present and compare results.

2. Select one of the 3 approaches and set up the necessary software environments:

   (a) Python with scikit-learn inside jupyter notebooks.
   (b) R with caret, or with mlr3 framework inside R notebooks.
   (c) ArcGIS...

3. Download the benchmark data files (gpkg, shp files).

Remark 1. I have a clear preference for the python-scikit approach, but it would be good if the 2 groups choose, each one, a different approach. I leave this up to you.

Project Steps

1. Read the paper and make a summary of the main approaches and findings. Eventually consult the references of the paper for further information.

2. Load data (see above). Write a very precise and complete description of the data contents, with detailed explanation of each one of the features. This will be important for later application to PHIVOLCS data.

3. Model selection: we will concentrate on 2 ML methods (Random Forests and XGBoost). Read about, and write a short presentation of each of these methods.

4. As in the paper, we will use cross-validation and an AUC-ROC metric to evaluate and compare the methods. Read up and ensure that you understand well these 2 concepts. Are there any alternatives? We will discuss, in particular, the concept of spatial cross-validation during the training sessions.

5. Variable/feature selection: the original dataset contains 26 feature variables and 2 target variables. What are the possible methods for performing variable selection? Please consult the machine learning lecture notes and [ISLP]. We would like to investigate (see point 7) what constitutes a minimal (or essential) set of features, and in particular in the context of what is available from PHIVOLCS.

6. Apply ML approaches and perform comparative evaluation based on carefully chosen metrics [APM]. Examine both the full dataset (26 variables) and the reduced dataset (19 variables, available in the codes directory). [Optional] Use RF and Shapley analysis to select a reduced set of variables.

7. Draw detailed conclusions. What do you think will need to be done to apply all the above to existing, available Philippines data?

Additional Steps

The paper [ISPRS] also performs a machine-learning based study of landslide susceptibility mapping, where the ML methods are trained on a carefully prepared inventory [ESDA], and then tested on unseen data (from the same inventory).

1. Read the paper and take careful note of the features, that include climatic data.

2. Study the workflow and explain the steps. Use the notebooks in the GitHub and identify all the steps.

3. All the cartographic data can be downloaded from Danish, public websites. Try and download this data, which is listed in the Data Availability Statement at the end of the paper.

4. The landslide inventory itself is in the Data folder of the GitHub,

5. Examine very closely the Notebook 8 and notice how cleanly and elegantly the actual machine learning can be performed with scikit-learn.

6. Even if you have not succeeded in downloading all the data, we can apply the 3 ML methods: random forest, logistic regression and support vector machine—see explanation on next slides. Note the performance evaluation metrics that are used.

7. Attempt to apply this methodology to the benchmark data from the previous exercise on Italian landslides. Note that it is not necessary to have all the feature variables of the Danish data—a subset of available features should work quite well.

8. Attempt to apply this methodology to a PHIVOLCS dataset. This supposes that all the preparation and pre-processing steps have been performed on this data. Note that once again, it is not necessary to have all the feature variables of the Danish data—a subset of available features should work quite well.

Recall: formulation of a machine learning problem

It is important to go back to the theoretical foundation of a machine learning problem, as already seen in the basic course on ML. Once we have a mathematical formulation, we no longer require the physical context since the problem has become a question of statistical estimation. We have "abstracted" away the physical context and our mathematical formulation is "universal" in some broad sense.

Recall: the mathematical framework

- Suppose we have :

  $\Rightarrow$ a response variable (to explain), Y,
  $\Rightarrow$ p explanatory[1] , variables, $X = (X_1, X_2, \ldots, X_p)$,
  $\Rightarrow$ n samples of data, giving an $(n \times p)$ matrix,

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

  $\Rightarrow$ a relationship between Y and X of the form

$$Y = f(X) + \epsilon$$

  where
  $\rightarrow$ f is an unknown function of $X_1, X_2, \ldots, X_p$
  $\rightarrow$ ϵ is a random error term, independent of X, and with zero mean

[1] Also called: features, attributes

- ML is then an ensemble of approaches for estimating f with the objectives of

  $\Rightarrow$ Prediction: $\hat{Y} = \hat{f}(X)$ where $\hat{f}$ is an estimation for f and $\hat{Y}$ is the resulting prediction

  $\Rightarrow$ Inference: to understand how Y varies as a function of X (correlations, importances, linearity, etc.)

References

References

[ESR] M. Alvioli, M. Loche, L. Jacobs, C. H. Grohmann, M. T. Abraham, K. Gupta, N. Satyam, G. Scaringi, T. Bornaetxea, M. Rossi, I. Marchesini, L. Lombardo, M. Moreno, S. Steger, C.A.S. Camera, G. Bajni, G. Samodra, E. E. Wahyudi, N. Susyanto, M. Sinčić, S. B. Gazibara, F. Sirbu, J. Torizin, N. Schüßler, B. B. Mirus, J. B. Woodard, H. Aguilera, J. Rivera-Rivera. A benchmark dataset and workflow for landslide susceptibility zonation. *Earth-Science Reviews*, Volume **258**, 2024, 104927. DOI or Science Direct

[ISLP] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor. *An Introduction to Statistical Learning*. Springer. 2023.

[APM] M. Kuhn, K. Johnson. *Applied Predictive Modeling*. Springer. 2013.

[ISPRS] A. Ageenko, L. C. Hansen, K. L. Lyng, L. Bodum and J. J. Arsanjani. Landslide Susceptibility Mapping Using Machine Learning: A Danish Case Study. *ISPRS Int. J. Geo-Inf.* 2022,11,324.

[ESDA] Luetzenburg, G., Svennevig, K., Bjork, A. A., Keiding, M., Kroon, A. A national landslide inventory for Denmark. *Earth System Science Data*, 14, 7 (2022), pp. 3157--3165, URL.