Using machine learning regarding suspended sediment concentration in a proglacial river

Keyvan Diba

Abstract

This study assesses the potential of applying machine learning techniques and algorithms to suspended sediment concentration in a proglacial river to see if it is possible to predict sediment concentration as a function of discharge and to observe whether these two variables are correlated.

1 | Introduction

Glaciers are active erosional agents, continuously crushing and abrading the ground over which they move. Therefore, glacierized basins are characterized by rapid meltwater transfer with high and variable suspended sediment concentrations (SSCs) (Hubbard et al., 2005). It is known that rivers draining glacierized areas generally transport significantly more solid matter in suspension than the ones draining non-glacierized areas (Hubbard et al., 2005). Thus, suspended sediment concentration reflects the availability of material for transportation at the glacier bed and the ability of the flow to transport it (Perolo et al., 2019). However, suspended sediment concentration is far from constant and vary systematically at a number of different time-scales (Perolo et al., 2019). Instream measures showed that suspended sediments and bedload respond differently to diurnal flow variability (Perolo et al., 2019).

Most suspended sediment concentration depends on the availability of fine material while bedload depends mainly on the competence of the flow (Perolo et al., 2019). Suspended sediment concentrations are therefore generally positively correlated with discharge at the diurnal and seasonal time-scales (Hubbard et al., 2005). Some studies were made on this topic and showed that sub-seasonal changes in relationships between suspended sediment transport and discharge demonstrate that the structure and hydraulics of the subglacial drainage system critically influenced how basal sediment was accessed and transported (Swift et al., 2005). Such studies have shown that sediment evacuation is largely dependent on the increased availability of meltwater during the melt season but is poorly related to discharge at annual scale (Swift et al., 2005).

The aim of this study is therefore to apply a machine learning algorithm to a dataset composed of continuous values of discharge and suspended sediment concentration measured during the 2021 melting season, and to see if it is possible to predict the expected suspended sediment concentrations based on discharge values and to assess if they are correlated or not.

2 | Data

2.1 Study area

The Otemma glacier is located in Switzerland in the canton of Wallis, in the southwest of the Valais Alps along the Italian border. Runoff from melting ice and snow correspond to the source of the proglacial flow called the Dranse de Bagnes. This stream is then captured in the artificial lake of Mauvoisin.

Figure 1 - Proglacial margin of the Otemma glacier
Source : Davide Mancini

2.2 Input data

The data used in this study were given by the professor Stuart Lane. They represent a continuous data collection of the suspended sediment concentration as well as the discharge during the 2021 summer melt season.

Figure 2 – Aerial view of the alluvial plain of the Otemma glacier. Source : Stuart Lane

Two datasets from two measuring stations were provided: (1) the "GS1" dataset from the measuring station downstream of the alluvial plain (Figure 3) and (2) the "GS2" dataset from the measuring station upstream of the alluvial plain at the outlet of the glacier (Figure 4).

	Time	C	SD_C	Q	SD_Q	QC	SD_QC	NoUse	NoUse2
0	244.003	0.9668	0.0068	2.6796	0.2546	2.5907	0.4839	3.0745	2.1068
1	244.007	0.9507	0.007	2.6995	0.2548	2.5664	0.4763	3.0427	2.0901
2	244.01	0.9451	0.007	2.6278	0.2541	2.4836	0.4722	2.9558	2.0114
3	244.014	0.9445	0.0071	2.7611	0.2554	2.6078	0.4743	3.0821	2.1335
4	244.017	0.9382	0.0071	2.7123	0.2549	2.5447	0.4703	3.0151	2.0744
...	...	...	...	...	...	...	...	...	...
5284	271.417	0.7465	0.0097	1.3525	0.2523	1.0096	0.37	1.3796	0.6396
5285	271.424	0	0	1.4786	0.2496	0	NaN	NaN	NaN
5286	271.431	0	0	1.4404	0.2503	0	NaN	NaN	NaN
5287	271.438	0	0	1.3914	0.2513	0	NaN	NaN	NaN
5288	271.444	0	0	1.3613	0.252	0	NaN	NaN	NaN

Figure 3 - Summary of the GS1 dataset (5289 rows and 9 columns)

	Time	C	SD_C	Q	SD_Q	QC	SD_QC	NoUse	NoUse2
0	246.5347	1.3922	0.0253	4.2128	0.3126	5.865	0.8781	6.7431	4.987
1	246.5361	1.395	0.0254	4.2856	0.3119	5.9786	0.8791	6.8577	5.0995
2	246.5375	1.402	0.0258	4.3692	0.3111	6.1257	0.8829	7.0086	5.2428
3	246.5389	1.407	0.0261	4.4405	0.3104	6.2478	0.8854	7.1332	5.3624
4	246.5403	1.4105	0.0262	4.4719	0.31	6.3075	0.8874	7.1949	5.42
...	...	...	...	...	...	...	...	...	...
5204	271.3542	0	0	1.1966	0.219	0	NaN	NaN	NaN
5205	271.3611	0	0	1.1906	0.2191	0	NaN	NaN	NaN
5206	271.3681	0	0	1.1441	0.22	0	NaN	NaN	NaN
5207	271.375	0	0	1.1496	0.2199	0	NaN	NaN	NaN
5208	271.3819	0	0	1.3337	0.2176	0	NaN	NaN	NaN

Figure 4 - Summary of the GS2 dataset (5209 rows and 9 columns)

3 | Method

As a first step, it is necessary to process the data as soon as it is imported and loaded. The two datasets used in this study have two columns of data whose nature has not been specified, they have simply been considered "useless" by the person in charge of their acquisition. It is therefore imperative to remove them from the data to be processed in the algorithm. Moreover, there is a certain amount of "NaN" values that need to be removed from the datasets as well.
Following the data processing, there is the need to define the parameter or criterion that will be used to predict the suspended sediment concentration and to split the datasets into three separate sets: (1) train set, (2) test set, (3) validation set. Finally, it is necessary to perform a «RandomForest» classify to the training set. The resulting model can be applied to the validation set and a confusion matrix will be created in order to visualize the accuracy score of the dataset.
This methodology will have to be applied to the two datasets in order to be able to observe whether there are significant differences induced by the different positions of the two measuring stations.

4 | Results and discussion

The results obtained following the RandomForest classification using discharge as the main variable are not satisfactory enough to be able to effectively predict the suspended sediment concentration. Indeed, the final accuracy obtained is 62.9% for the dataset

coming from the measuring station located at the outlet of the glacier, which could be considered satisfactory. But, the final accuracy obtained for the second measuring station is 1.4%, allowing to doubt the effectiveness of this prediction.

Several discussion elements can be brought to try to explain why the use of the RandomForest algorithm did not produce satisfactory results. First, the dataset used in this study may not be suitable for this algorithm because it does not allow clear classified results. The implementation of a regression algorithm would have been a more coherent decision in relation to the nature of the data. It would also have been possible to formulate the research problem in a different way. Indeed, instead of trying to predict the amount of suspended sediment as a function of discharge and to see if these two variables are correlated, it would have been more appropriate to try to predict the behaviour of downstream suspended sediments as a function of the observed behaviour upstream. This research question would have been also more appropriate with the application of RandomForest algorithm.
In addition, several elements could have been added to the current code in order to improve it. In this study, any hyperparameters were used in the algorithm applied to the dataset. Adding hyperparameters could therefore directly improve it. Moreover, it would be possible to combine the algorithm with other methods in order to increase the overall performance of the code and refine the results obtained.

5 | Conclusion

The fact that the use of machine learning methods has not yielded conclusive results which make it possible to predict effectively the concentration of suspended sediment as a function of discharge does not question the potential of this field of study in the context of environmental sciences. Machine learning methods are powerful and suitable tools for dealing with complex environmental problems.

These are reliable tools and methods to put in place as long as there is a clear understanding of the problematic to be addressed and a clear idea of how to deal with it, which was not necessarily the case in this particular study.

6 | Link to the code

The code used to carry out this study can be found here.

7 | References

Bryn Hubbard, Neil F. Glasser—Field Techniques in Glaciology and Glacial Geomorphology-Wiley (2005).pdf. (s. d.).

Perolo, P., Bakker, M., Gabbud, C., Moradi, G., Rennie, C., & Lane, S. N. (2019). Subglacial sediment production and snout marginal ice uplift during the late ablation season of a temperate valley glacier. Earth Surface Processes and Landforms, 44(5), 1117-1136. https://doi.org/10.1002/esp.4562

Swift, D. A., Nienow, P. W., & Hoey, T. B. (2005). Basal sediment evacuation by subglacial meltwater : Suspended sediment transport from Haut Glacier d'Arolla, Switzerland. Earth Surface Processes and Landforms, 30(7), 867-883. https://doi.org/10.1002/esp.1197