Suspended bedload in a proglacial river

Alessandro Giovanardi

Abstract

Suspended bedload in rivers is a complex metric that influences the whole river ecosystem. Therefore, being able to predict it can provide a lot of information about the ecosystems, and hence about the habitat. Machine learning is then useful and a powerful method to determine a parameter with and thanks to other already known variables and to obtain good results. The main result that was found is that predicting an upstream segment with data collected downstream is easier and better.

1. Research question

Can suspended bedload be predicted by training the model with and thanks to the concentration of suspended bedload and the flow of an alpine river?

2. Introduction

As we all know, the amount of sediment transported in a river depends on river properties like flow rate or particle size, to name a few. This is crucial for knowing and predicting the supply of sediment by the river. These properties are extremely important to the river ecosystem and will determine the diversity and quality of the habitat of the species living there. For example, suspended sediment carry a lot of thing as nutrients and organic compound crucial for life in rivers (2). It is important to know that although this parameter is essential for the river ecosystem, if its concentration is too high in the water, light penetration will be affected, which will have a negative effect on the river (1). The fauna and flora will then partly depend on its amount in the water. In our case, given the region where these data were collected, most of the suspended sediment comes directly from glacier erosion (4). Therefore, this amount is season dependent. The majority will be released in spring and summer when the glacier melts (3). This project will then, try to establish what control the amount of suspended bedload downstream according to two different properties and find out which one determines the most this quantity. The two parameters studied are: the concentration of the suspended bedload and the discharge of the river. Then, as we have also the amount of suspended bedload of the river for the two stations, we will try to predict this amount going from downstream to upstream.

3. Data

The two datasets used for this project were provided by my thesis professor and collected in Arolla. Each of them has the following parameters: T, C, Q, QC. T stands for the time, C for the concentration of suspended bedload, Q for the discharge and last, QC for the amount of suspended bedload in the river. As we will see later too, QC is defined and depend on C and Q. For that, two measuring stations are needed, see Figure 1. The first one (the upstream), is located downstream of the glacier alluvial plain. Regarding the second station (the downstream), this point is situated close to the outlet of the glacier river. These two locations will be used to answer our research question in a conceptual way.

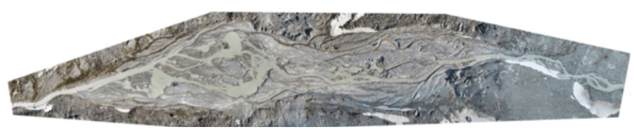

Figure 1. Aerial photo of station locations

Here the specific aspect of my dataset. First, my dataset is composed of 28759 instances for each parameter. Regarding the mean and the standard deviation for C and Q see Figure 2.

4. Methodology and methods

To answer the research question, this project will apply the following methodology. First, after loading the data, these are separated into a train, validation and test set following this 80%, 10%, 10% distribution. This by having previously normalized the input data. Then, a multi-linear regression followed by a random forest regression prediction is performed in order to highlight which parameters would explain the most the amount of suspended bedload found downstream. This will rank the different features together to see which is more important. Several parameters

```
Mean of C before normalisation: 1.54
Mean of Q before normalisation: 4.96
Mean of C2 before normalisation: 1.61
Mean of Q2 before normalisation: 6.35
Standard deviation of C : 0.91
Standard deviation of Q : 1.93
Standard deviation of C2 : 1.03
Standard deviation of Q2 : 2.15
```

Figure 2. Mean and standard deviation for C and Q

will then be calculated to analyse the performance of the model such as R^2, MSE or RMSE. This will allow us to assess the quality of our results. The coefficients for the multi-linear regression or the features importance for the random forest regression are calculated at the same time to estimate the parameter influencing the most the amount of suspended bedload in the river. The same methodology will be used for predicting the amount of suspended sediment in the water going from downstream to upstream.

5. Results

The following graph shows the variation of the concentration of the suspended bedload in the water as well as the variation of the flow over time. Some peaks can be observed for the concentration of the suspended bedload, although the concentration is the rest of the time rather constant. Concerning the flow, we observe that it is much more subject to daily variations and is quite different between the two stations.

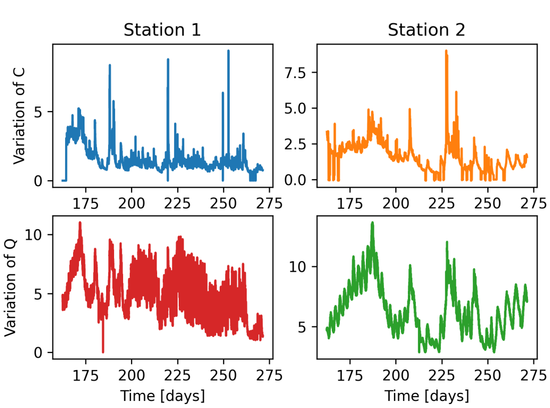

Figure 3. Variation of C and Q in time for the two stations

Now looking at the values in the table, when the random forest is used, the precision is better, and the errors are smaller. The errors decrease when we run the model on the test set. The accuracy ranges from 5% to about 90% for the random forest regression. Secondly, the prediction from downstream to upstream gives more accurate results and less errors than the other way. As for the coefficients, they vary a lot between the two methods.

Table 1. Summary of the accuracy and errors of the two methods predicting from upstream to downstream. LR stands for Linear Regression and RF for Random Forest

	LR		RF	
	Y VALID	Y TEST	Y VALID	Y TEST
ACCURACY [%]	5.19	5.19	89.72	89.72
R^2	0.046	0.058	0.24	0.3
MSE	117.86	107.68	93.54	79.44
RMSE	10.86	10.38	9.67	8.91

Table 2. Summary of the accuracy and errors of the two methods predicting from downstream to upstream

	LR		RF	
	Y VALID	Y TEST	Y VALID	Y TEST
ACCURACY [%]	5.14	5.14	93.76	93.76
R^2	-0.044	0.065	0.49	0.52
MSE	73.32	60.82	36.17	31.12
RMSE	8.56	7.80	6.01	5.58

6. Discussion

Figure 3 allows us to understand that the variability is significant in time and space. Even though the two stations are not very far from each other, the differences between them are not negligible. This reinforces the point that only two metrics to describe and predict the amount of suspended bedload in the water may not be sufficient. The uncertainty will therefore remain high as we can observe through the various error values calculated after the multi-linear regression and the random forest see tables 1 and 2. The first thing that can be observed in the results is that the random forest regression is more accurate and, as expected, gives better results. However, the errors found for the multi-linear regression, or the random forest are quite bad. They are all far from the ideal values, in our case 0, but tend to be smaller when the random forest is used and when the prediction goes from downstream to upstream. This means that the model predicts the y value without taking the input data properly in consideration. Moreover, this tells us that the results obtained cannot be taken very seriously because their

Table 3. Summary of the coefficient finds for the two methods predicting from upstream to downstream

	LR		RF	
	C	Q	C	Q
COEFFICIENT	0.5480	1.1056	0.5086	0.4913

Table 4. Summary of the coefficient finds for the two methods predicting from downstream to upstream

	LR		RF	
	C	Q	C	Q
COEFFICIENT	2.0043	-0.1081	0.5797	0.4203

uncertainty is high even if the precision is good. For the second scenario, moving from downstream to upstream, the results are interesting. Besides being better, it means that it is easier to predict the amount of suspended sediment in the river by knowing its concentration and the downstream discharge. It is true that in the reverse direction, the uncertainty would tend to be greater because many different factors can influence these variables before arriving downstream. Doing it the other way around, even though the factors are assumed to always be present, gives more information about the river and in this case, it is simpler to predict other variables. The coefficients found with linear regression are suspicious and vary significantly between the two scenarios. However, the results obtained with the random forest seem to be decent. These values suggest that there is a slight difference between the two inputs, i.e., C and Q, and that C explains the amount of suspended bed in the rivers a little more than flow. As mentioned earlier, implementing complexity by adding variables that explain the amount of suspended sediment in the river would help this research obtain significantly less error and better results. The uncertainty with just two variables is huge otherwise.

7. Conclusion

The prediction of variables such as the amount of suspended bedload in rivers is important to be able to assess the overall quality of habitats in rivers and, if necessary, take measures to restore good quality. It is a crucial parameter that affects life in rivers and plays a determining role for the whole river ecosystem. This study tends to show that predicting these parameters is easier when done with variables collected not as usual upstream, but downstream of the segment, of the wanted section. Unfortunately, the uncertainty of the model is large because only two variables are taken into account, even if the accuracy of the model is good.

8. Other resources

The complete code and the dataset used for this work can be found on GitHub at this link: The dataset is called LGS1_2021 and LGS2_2021.

`https://github.com/aless820/2022_ML_Earth_Env_Sci`

References

[1] Pia Bessell-Browne, Andrew P Negri, Rebecca Fisher, Peta L Clode, Alan Duckworth, and Ross Jones. Impacts of turbidity on corals: The relative importance of light limitation and suspended sediments. *Marine Pollution Bulletin*, 117(1-2):161–170, 2017.

[2] Peng Gao. Understanding watershed suspended sediment transport. *Progress in Physical Geography*, 32(3):243–263, 2008.

[3] P Perolo, M Bakker, C Gabbud, G Moradi, C Rennie, and SN Lane. 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, 2019.

[4] Darrel A Swift, Peter W Nienow, and Trevor B Hoey. Basal sediment evacuation by subglacial meltwater: suspended sediment transport from haut glacier d'arolla, switzerland. *Earth Surface Processes and Landforms*, 30(7):867–883, 2005.