MaLDIveS : Machine Learning Diagnostic Soil

Jibril Mammeri[1,2], Xavier Brochet[1,2], Thierry Heger[2,3], Sven Bacher[4], Magdalena Steiner[4], Carlos Peña[1,2]

[1]*School of Business and Engineering Vaud (HEIG-VD), University of Applied Sciences Western Switzerland (HES-SO), Switzerland*

[2]*Swiss Institute of Bioinformatics (SIB), Switzerland*

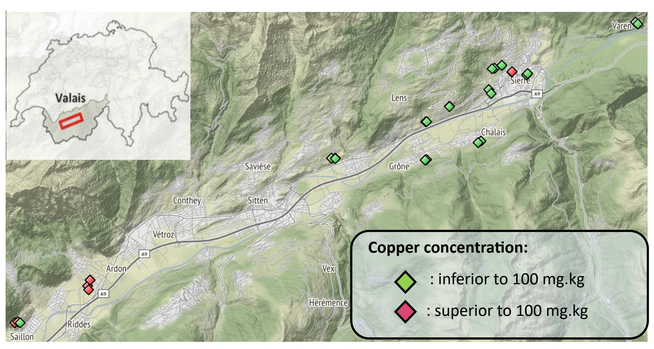
[3]*School of Viticulture and Enology (Changins), University of Applied Sciences Western Switzerland (HES-SO), Switzerland*

[4]*Department of Biology, Unit of Ecology and Evolution, University of Fribourg, Switzerland*

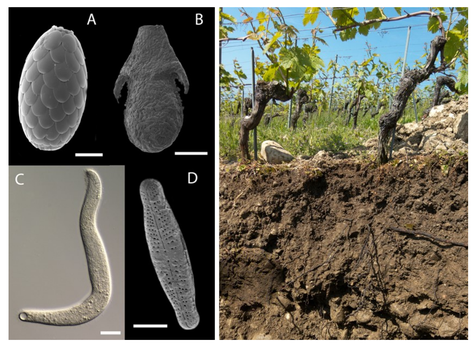
MaLDiveS is an ongoing multidisciplinary project, which aims to develop a new method of biomonitoring based on next-generation sequencing data and their treatment by machine learning methods. It will allow to assess the impact of treatment on the health and quality of the vines soil. This work will improve the understanding of pesticides and other environmental factors impact on protists communities. Our main objective is to identify bioindicators associated with environmental stress and to caracterize the behavior of their relative abundance, which will lead to the construction of diagnostic models.

Overview

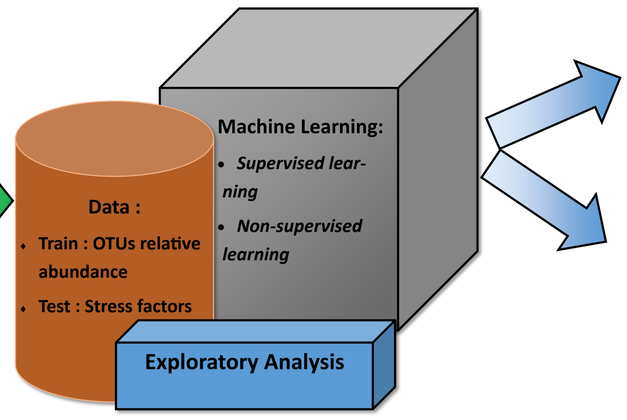
There is a growing concern among consumers and farmers about the impact that agricultural practices and, more specifically, pesticide usage have on the health and the quality of the soil. To assess this impact, and measure soil quality, soil organisms can be used as bioindicators. Among the different groups of soil organisms, protists are ideal candidates to be used as bioindicators of soil health. Protists are abundant, diversified and also very sensitive to natural and anthropogenic disturbances.



Map of the 33 Valais vineyards studied by Prof. Bacher for the european project PromESSing (Promoting Eco-System Services in Grapes) . In total, 178 samples were collected during two succesive seasons (2015-2016).



***Left** : Optical and electron microscopy images illustrating four different groups of soil protists: A) Testate amoebae from the order of Euglyphida ; B) amoebae from the order of Arcellinida ; C) Ciliate ; D) Diatom, Source : T. Heger and W. Bourland. **Right**: Vine and soil profile photo: M. Motta (Changins).*



Sample collection

178 samples were collected from 33 vineyards in Valais describing 31 environmental factors

DNA extraction and quantification

From the samples, the relative abundance of 1364 OTUs *(Operational Taxonomic Unit)* corresponding to different species protists was computed

Data analysis

Statistical and machine learning methods will be used to interpret our datasets, build soil quality diagnostic models and identify bioindicators.

Methodology

Initial data
- 178 samples
- 1364 OTUs
- 31 environmental factors

Exploratory analysis
- Univariate distribution
- Correlation matrix
- Principal Component Analysis (PCA)
- T-distributed Stochastic Neighbor Embedding (T-SNE)
- Self-Organizing Maps (SOM)

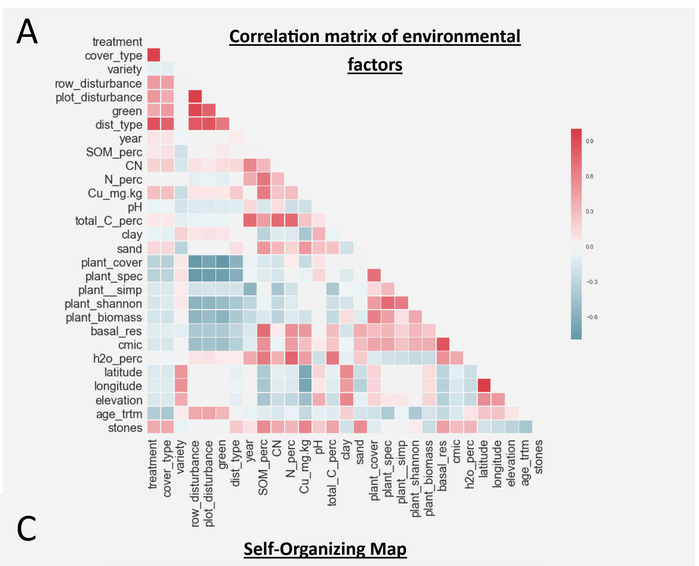
Feature Selection
- **Correlation** : 336 OTUs were highly correlated (>0,95)
- **Variance** : The relative abundance of 574 OTUs didn't vary (<0.05)
- We made four datasets :
  - **Full (1364 OTUs)**
  - **Non-correlated (1028 OTUs)**
  - **Low-variance (790 OTUs)**
  - **Non-correlated and low-variance (688 OTUs)**

Machine Learning
- Predictive model building with 5-fold cross-validation, based on three approaches :
  - Random Forests (RF)
  - Support Vector Machines (SVM)
  - Artificial Neural Networks (ANN)

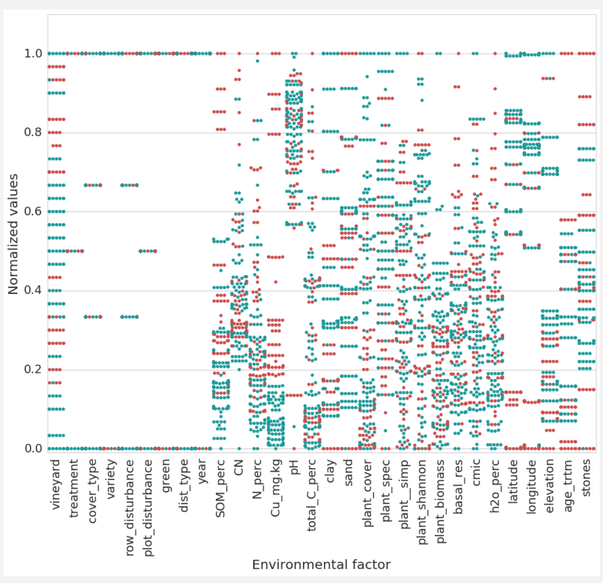
Exploratory analysis

- Some stress factors were correlated between each other (A,B) but none was correlated with an OTU.
- However we can distinguish a cluster of copper overcharged samples on the SOM drew from OTUs relative abundance (C).
- We hypothesize that soil copper pollution can be predicted from the relative abundance of protists.

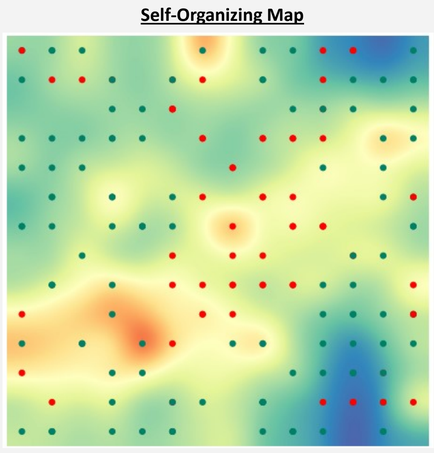
A



B

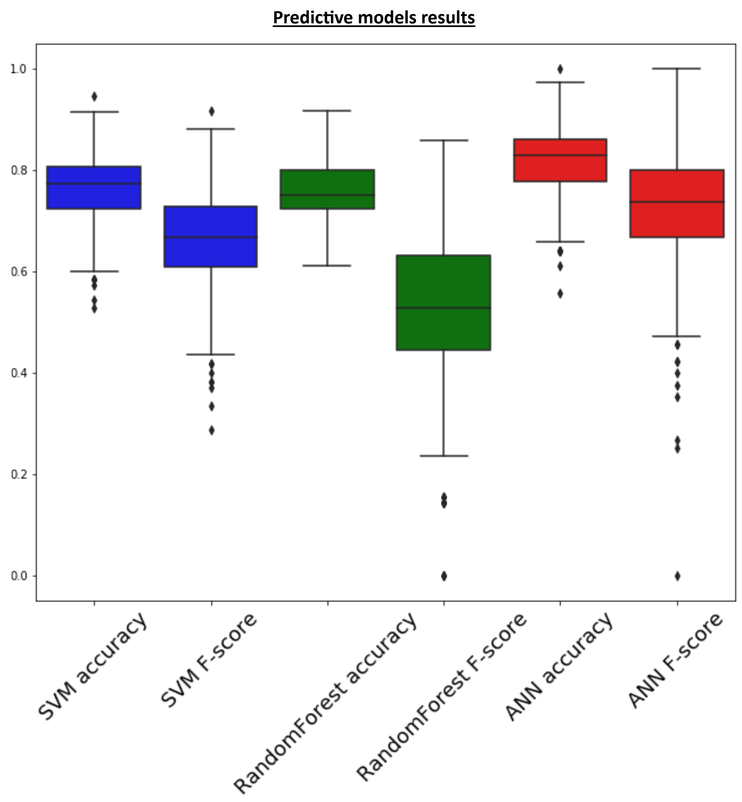


C



***A**: Correlation matrix of the environmental factors; **B**: Distribution of the normalized values of environmental factors colored by copper concentration of each sample (**red** : copper concentration > 100mg.Kg, **blue** : copper concentration < 100mg.Kg); **C** : Self-organizing map realized from the relative abundance of each OTU by sample and then colored according to copper concentration (**red** : copper concentration > 100mg.Kg, **blue** : copper concentration < 100mg.Kg).*

Machine Learning



Boxplot of the accuracy and F-score distribution for 100 ANN, SVM and RandomForest models built to predict copper contamination with a 5-fold cross-validation.

We were able to predict with around 80% accuracy whether a sample was copper contaminated with each model, but the F-score reveals that there is a major discrepancy between our sensitivity and specificity for some models.

Conclusion

- We obtained better specificity than sensitivity due to the lower number of copper overcharged samples in our dataset.
- By weigthing our training set and adjusting our models we will aim to improve our predictive power.

	ANN	SVM	RandomForest
Sensitivity	71.01%	67.39%	36.11%
Specificity	89.09%	81.54%	96.07%
Accuracy	82.69%	76.65%	74.85%
F1-score	74.39%	67.03%	50.41%

Table of the sensitivity, specificity, accuracy and F-score mean for 100 ANN, SVM and RandomForest models built to predict copper pollution with a 5-fold cross-validation.

Future work:

-Optimize and enhance our predictive models

-Build models to predict other stress factors

-Identify the most relevant protists to diagnostic soil quality

-Study new datasets realized from *in vitro* experiments

-Test other machine learning methods (KNN, Fuzzy logic, Ensemble Learning, ...)