Digital mapping of soil waterlogging as a support to wetland delineation at regional scale: learning strategies and accuracy assessment

Christian Walter [A,B], **Marine Lacoste** [A,B], **Blandine Lemercier** [A,B]

[A] INRA, UMR 1069, Soil Agro and hydrosystem, Spatialization, Rennes, France, Christian.Walter@agrocampus-ouest.fr
[B] University Européenne de Bretagne, Agrocampus Ouest, 65 rue de St Brieuc, Rennes, France

Abstract

Environmental regulation at European level compels member states to a better protection of wetland areas. Systematic inventory and delineations of wetlands have therefore been recently undertaken in France and consider soil redoximorphic features using criteria fixed by decree. Digital soil mapping is tested as a support of wetland delineation at the regional scale of Brittany (France) and the study focuses on two objectives: (i) to compare the efficiency of learning strategies based on punctual observations or existing detailed soil maps to extrapolate spatial models of soil waterlogging; (ii) to assess the influence of prediction uncertainties of soil waterlogging on the final wetland delineation. A classification tree is used: MART (Multiple Additive and Regression Trees). MART allows the creation of predictive model based on point data and predictive variables: topography, landscape map obtain by remote sensing, airborne gamma-ray spectrometry, etc. Two kinds of models are inferred: P-Model is based on 5129 punctual soil descriptions; M-Model is based on existing soil maps at 1:25 000 scale covering 11% of the study area. Model validation by independent data set indicates an overall better performance of P-Model explained by a better spatial coverage. M-Model nevertheless performed slightly better in areas similar to its learning conditions. Probability estimates of soil waterlogging classes were finally combined to estimate the probability of occurrence of wetland conditions meeting the regulation criteria.

Key Words

Stochastic gradient boosting, classification trees, learning machine, soil waterlogging, wetland delineation.

Introduction

Enhanced access to attributes describing the physical environment and recent advances in digital soil mapping, GIS and statistics areas offer new perspectives for spatialization of soil properties. Detailed existing soil maps constitute an interesting information source on soil spatial distribution, but have a generally limited spatial extension. An alternative approach is to use punctual observations which may be more evenly distributed. The knowledge embedded into soil maps by soil surveyors or in punctual descriptions can be retrieved and explicitly formulated using environmental data (Moran and Bui 2002; McBratney *et al.* 2003).

This study considers the ability of Digital Soil mapping to assist wetland delineation at regional scale, with two objectives: (i) to compare the efficiency of learning strategies based on punctual observations or existing detailed soil maps to extrapolate spatial models of soil waterlogging; (ii) to assess the influence of prediction uncertainties of soil waterlogging on the final wetland delineation.

Material and methods

Study area

The study area is related to Brittany, a west French region of 27 020 900 ha. The geology of the region, influenced by several orogenies and transgressions, is complex. North and South Brittany mainly presents igneous and metamorphic rocks, whereas the centre of Brittany shows sedimentary rocks. Brittany is also covered by superficial deposits, particularly by Aeolian loam in the North of the area. Topography, parent material and climate gradients are considered to be the main factors of regional soil waterlogging variability.

Model creation and datasets

The MART (Multiple Adaptive Regression Tree) method was used to create predictive models. This method allows solving predictive learning problems building classification or regression trees and using “stochastic gradient boosting” (Fiedman 1999). This particular boosting method is known to significantly improve accuracy compared with simple regression and classification trees.

The model requires two kinds of input data: training data, which correspond to the response variable to predict, and environmental predictors. Two models were created:

- P-Model was based on 5 129 profiles (punctual observations of soil), and 4/5 of them were used as training data to build the models;
- M-Model was based on existing soil maps at 1:25 000 scale covering 11% of the study area. These maps were resampled and integrated as training dataset into the MART learning machine.

17 environmental predictors were used, compound by: (i) terrain attributes derived from a 50 m resolution DEM (elevation, slope.), (ii) emissions of K, Th and U derived from airborne gamma-ray spectrometry, (iii) geological variables at 1:250 000 scale (bedrock lithology and superficial deposits) and (iv) landscape map obtained by remote sensing.

Models outputs were extrapolated at regional scale with a 50 m x 50 m resolution, enabling the prediction of the occurrence probability of each soil waterlogging class.

Validation procedure

Models results were validated in four ways: (i) comparing the model predictions to all the profiles (internal validation); (ii) comparing the model predictions to 1/5 of the profiles, not used to create the models (cross-validation); (iii) comparing the model to existing soil maps at 1:25 000 scale covering 4% of the study area (external validation); (iv) estimation of the quality of the prediction by experts.

Wetland delineation accuracy

Soil waterlogging classes were interpreted considering wetland delineation criteria fixed by decree. Probabilities of the different soil waterlogging classes available were thereafter combined to estimate the probability of occurrence of a wetland.

Results

P-Model prediction of soil waterlogging

Internal and cross-validation of the model showed an accuracy of 72% and 68%. The most important predictors were respectively parent material, land use, elevation from stream, and Compound Topographic Index. Overall accuracy for external validation was 57%. According to the experts, the prediction was globally satisfactory and appeared of homogeneous quality over the region.

M-Model prediction of soil waterlogging

Internal validation accuracy was 66 %. The most important predictors were the previously parent material, Compound Topographic Index, deviation from mean K emissions and bedrock lithology. Overall accuracy for external validation was 55%. According to experts, the prediction was of good quality in situations analog to the training areas, but of poor quality in other situations.

Wetland delineation

Probabilities of occurrence of soil waterlogging classes were combined to estimate the probability of each pixel to be considered as a wetland following regulations rules. This on-going work focuses on situations with high certainty to be included or excluded in wetland zones and also on situations of high uncertainty.

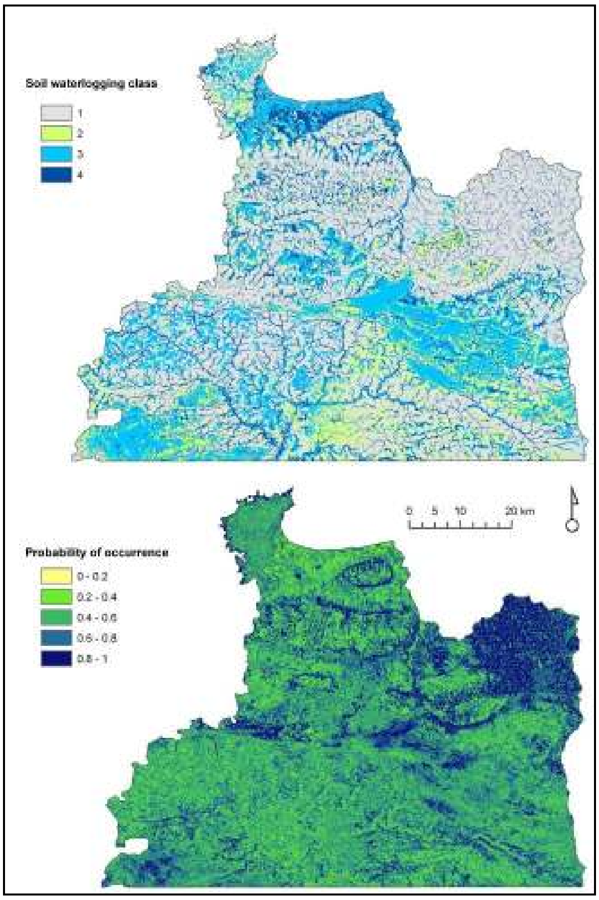

Figure 1. Prediction of soil waterlogging class and associated probability using the M-Model prediction on a subarea of Brittany

Discussion - Conclusion

The results of this work indicate an overall good performance of the MART predictions for soil waterlogging: 66 to 70% of pixels with existing soil information through profiles were correctly predicted. Soil waterlogging prediction was mostly influenced by the predicted soil parent material and topographic conditions for both models. The prediction of soil waterlogging appeared of good quality for the localization of the most redoximorphic areas, but less precise for intermediate classes. This means that digital soil mapping may be used to delineate with high reliability the most redoximorphic zones and that additional field work is needed to disentangle intermediate situations: we show that the probability of occurrence of a wetland zone as derived from a P-model is a relevant indicator to identify situations where additional information has to be gathered.

P-Model and M-Model predictions agreed at 66%. Model validation indicates an overall better performance of P-Model explained by a better spatial coverage. M-Model performed better in areas similar to their learning conditions. The external validation of the model using soil map training data showed an accuracy of 60%. So the two methods showed overall similarity, and the use of punctual information as training data in learning methods appears to be a promising approach.

References

Friedman JH (1999) Greedy function approximation: A gradient boosting machine. *Annals of Statistics* **29**(5)**,** 1189-1232.

Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. *Statistics in Medicin* **22**, *1365-1381.*

Grinand C, Arrouays D, Laroche B, Martin MP (2008) Extrapolating regional landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma* **143**, 180-190.

Jenny H (1941) 'Factors of soil formation.' (McGraw-Hill Publishing: New-York)

Moran CJ, Bui EN (2002) Spatial data mining for enhanced soil map modelling. *International Journal of Geographical Information Science* **16**(6), 533-549.