A comparison of three learning methods to predict N_2O fluxes and N leaching

Nathalie Villa-Vialaneix[1,2], Marco Follador[3] and Adrian Leip[3]

1- Université de Perpignan - IUT de Carcassonne
Domaine Universitaire d'Auriac, Avenue du Dr Suzanne Noël, 11000 Carcassonne - France

2- Institut de Mathématiques de Toulouse
Université Paul Sabatier, 118 route de Narbonne, 31068 Toulouse cedex 9 - France

3- European Commission, Joint Research Center - Climate Change Unit
Via Enrico Fermi, 2749, 21027 Ispra - Italy

Abstract. The environmental costs of intensive farming activities are often under-estimated or not included into rural development plans, even though they play an important role in addressing future society's needs. This paper focuses on the use of statistical learning methods to predict N_2O emissions and N leaching under several conservative scenarios, in order to provide an alternative approach to deterministic models on a macro-scale. To that aim, three learning methods, namely neural networks (multilayer perceptrons), SVM and random forests, are compared and provide accurate solutions.

1 Introduction

1.1 Context and purpose

Agriculture is a multi-goal system since it has to meet the needs of current and future generations while continuing to preserve natural resources. The last Food Agricultural Organization summit in Rome warned that more than 1 billion people are chronically undernourished worldwide [10] and that reversing this worrisome hunger trend could require more-intensive farming practices. The environmental costs of these activities are often unmeasured [27] even though they cause a degradation of other ecosystem services essential for human well-being. In 2003, the Common Agricultural Policy (CAP) revision introduced new measures to improve the compliance with more sustainable environmental and agricultural standards as prerequisite to receiving direct payments. To quantify the effectiveness of these measures several indicators have been selected to describe the nitrogen (N) and carbon (C) cycles across farmlands. In this paper, we focus on N_2O emissions and N leaching; the first indicator is a powerful greenhouse gas and the second one is an important source of fresh water pollution.

The estimation of N dynamics demands detailed simulation based methods and their integrated use to correctly represent complex and nonlinear interactions into cropping systems. To calculate the N_2O flux and N leaching from European arable lands, a modeling framework has been developed by combining

the agro-economical model CAPRI [8] with the biogeochemical model DNDC-EUROPE [15].

Despite the great power of modern computers, the use of deterministic models at macro-scale is often prohibited, because of computational needs and parametrization constraints [23]. Metamodeling is known to be a soaring application in many disciplines to approximate the expensive code of detailed models. [26] and [7] resume the benefits of metamodeling as follows:

- A better understanding of the relationship between input and output
- Easer integration into other processes
- Faster execution and responding scenario analysis for optimization and exploration of studied system
- Easier applicability across different scales and site-specific calibrations

In this paper we compare the performances of 3 different statistical learning methods to approximate the long and complex CAPRI/DNDC-EUROPE computer analysis codes; the metamodels have subsequently been integrated into the Cross Compliance Assessment Tool (CCAT), a large simulation platform which aims to provide an exhaustive assessment of the impact of CAP measures [11].

1.2 Description of the data

The Homogeneous Spatial Mapping Unit (HSMU) is the minimal geographical unit used for our simulations [15]. The main environmental informations are European soil [13] and daily meteorological data [20]; farm management and land use information have been obtained from CAPRI at regional level and subsequently dis-aggregated at HSMU resolution. Its processing is described in details in [15]. The original DNDC-EUROPE input dataset is very large and many parameters are required to feed the detailed model. To perform the simulations, we decided to screen out the less important variables from the whole input dataset; at the end the list of predictors comprised of the following information: N_FR (N input through fertilization; kg/ha y), N_MR (N input through manure spreading; kg/ha y), Nfix (N input from biological fixation; kg/ha y), Nres (N input from root residue; kg/ha y), BD (Bulk Density; g/cm^3), SOC (Soil organic carbon in topsoil; mass fraction), PH (Soil PH), Clay (Ratio of soil clay content), Rain (Annual precipitation; mm/y), Tmean (Annual mean temperature; °C), Nr (Concentration of N in rain; ppm).

Several scenarios related to agricultural choices are studied: the definition of scenarios is an important policy decision tool to compare conservative vs. conventional intensive farming activities to carry out an environmental cost-benefit analysis; a good model should be able to approximate the outputs of interest for various scenarios. A comparative study of several approaches is then useful to provide guidelines on the choice of a learning method as well as to evaluate the accuracy of each algorithm for a given task. We designed 5 scenarios according to the CAP measures [11]:

- S1: Baseline scenario (conventional corn cropping system)
- S2: similar to S1 without tillage
- S3: similar to S1 with a limit in N_MR at 170 kg/ha y
- S4: rotation between corn (2y) and catch crop (3y)
- S5: similar to S1 with an additional application of fertilizer in winter

The Europe of 25 Member States is covered by more than 200 000 HSMUs but to reduce the running time we decided to select a representative sample subset for corn crops by applying a minimum threshold criteria in land use. The number of simulation units has been so decreased to around 20 000; only the S4 scenario have been based on a larger sample set (about 40 000 HSMU) because it simulated the rotation of corn with alfalfa and we had to include all the HSMUs which contain both crops. Finally, the final data set results in 5 scenarios, each with two variables to predict (N_2O flux and N leaching) and 11 variables to make the prediction. The scenarios contain, respectively, 18 794 observations (scenario 1), 18 830 observations (scenario 2), 18 800 (scenario 3), 40 536 (scenario 4) and 18 658 (scenario 5). Each HSMU is described by the 11 predictors presented below and by the 2 target variables (N_2O discharge and N leaching) that have been extracted from the geobiological simulator DNDC-EUROPE.

2 Description of the experiments

2.1 A short review about learning methods used

Three of the most popular learning methods developed during the past years have been compared to address this question of which kind of approach should be implemented by the Climate Change Unit to obtain fast and accurate estimations of greenhouse gases discharge as well as N leaching. These 3 methods are briefly described below:

- Multilayer perceptrons (MLPs) come from the original model called perceptron that was introduced on the end of the 50's by Rosenblatt and became very popular after the wide increase in the computational capacities of computers. MLPs have been continuously improved and studied and the work of [24] and [3] provide a general presentation of these methods and of their properties. To avoid overfitting, a weight decay is frequently used (see [14]): the mean squared error optimized to learn the weights of the perceptron is penalized by the norm of the weights to avoid large and instable weights. In the experiments described below, a weight decay has been used for a one-hidden-layer perceptron with a sigmoid activation function on the hidden layer and a linear activation function on the output layer.

  Multilayer perceptrons are already well-known in the geostatistics community: several publications have already emphasized the usefulness of that

tool in remote sensing [2], ecological modeling [16] and land use modeling [17, 29], among others. Moreover, MLP are implemented in several GIS software that model land use evolution as in, e.g., the widely used commercial software Idrisi© (`http://www.clarklabs.org/products/index.cfm`).

- Support Vector Machines (SVMs) are unusual in geostatistics. SVMs were introduced by [4] and were originally designed to address classification problems. [28] presents an extension to the regression case by the way of an ϵ-insensitive loss function. Since that, other variants of kernel methods to regression problems have been developed such as the kernel ridge regression [25]. A SVM with ϵ-insensitive loss function and a Gaussian kernel was used for the modeling of N_2O flux and N leaching.

- Random forests are the most recent of the three studied methods since they were first introduced by [5]. This paper combines the ideas of bagging developed by [5] and that of feature selection by [1, 12] for improving the regression tree method [6]. Basically, the method consists in computing a large number of randomly under-efficient regression trees and then to average them. Despite its recency, random forests have gaven rise to an increasing interest in the past year in geostatistics. Among others, we refer to the article of [21], in the field of remote sensing, and the paper of [22], in the field of ecology.

2.2 Methodology

All experiments were performed using the free software R [19] to enable their implementation and diffusion in the Climate Change Unit. Existing R packages have been used: the package `nnet` for multi-layer perceptrons, the package `e1071` for SVM (see [9]), the package `randomForest` for random forests.

The following methodology has been applied to compare the three methods:

1. For each scenario, the dataset was randomly separated into a training set and a test set on the basis on 80% of the observations for the training set and 20% for the test set;

2. The training step was then performed for each of the three methods from the training set of each scenario and for both outputs to predict (N_2O flux and N leaching). During this step, several parameters had to be tuned:

   - MLPs required the tuning of the number of neurons on the hidden layer and of the penalization parameter associated with the weight decay. These tunings were made using a simple validation strategy on the training set. This approach was preferred to cross-validation to avoid a high computational cost that would have resulted from the number of observations. Moreover, the validation set was built by randomly selecting half of the whole training dataset and thus has a size always larger than 7 000, which should lead to a good robustness.

- SVMs required the tuning of three parameters: the parameter of the Gaussian kernel, the value of ϵ associated with the ϵ-insensitive loss function and the regularization parameter. The tuning of ϵ was avoiding by setting it equal to 1 which corresponds approximately to the second decile of the target variable for every scenario and every output: this choice fitted the standard proposed by [18] who suggest to have a number of Support Vectors smaller than 50% of the training set. Two other parameters were tuned in the same way as the parameters of MLP.
- Several parameters could have been tuned for random forests such as the number of trees in the final forest or the number of variables randomly selected to build a given split. But it is known that this method is less sensitive than the two others to parameter tuning so the default values implemented in the `randomForest` package, based on useful heuristics, were directly used. Moreover, the full learning process always led to a stabilized out-of-bag error.

In addition, to avoid problems due to the convergence to a local minimum, of the optimization algorithm involved in MLP, the whole learning process (including tuning) was repeated 5 times for this method, with different random initializations.

3. Finally, two quality measures were computed on the test set to compare results in each scenario and for each output to predict: the usual mean squared error and R^2 (Pseudo-R^2 was used: $1 - \frac{\text{SS}_{\text{model}}}{\text{SS}_{\text{default}}} = 1 - \frac{\sum_{i \in \text{test set}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{test set}} (y_i - \bar{y}_{\text{train}})^2}$ where y_i are the true output values, $\hat{y}_i$ are the predicted output values and $\bar{y}_{\text{train}}$ is the default prediction based on the mean of the outputs in the training set).

   Additionally, to give an indication of which variables are important in the prediction, an "importance" measure was calculated. For random forests, the importance is quite common: for a given predictor, the values of out-of-sample observations are randomly permuted; the mean squared error is then calculated based on all out-of-sample sets for all trees in the forest. The increase in the mean squared error compared to the out-of-sample mean squared error calculated with the true values of the predictor is called the importance of the predictor.

   Unfortunately, MLPs and SVMs are not based on bootstrapping so out-of-sample observations do not exist for these methods. Hence, importance cannot be defined or directly compared to the one given for random forests. Nervelessness, a close definition can be introduced by using the validation set selected for the tuning process and by comparing the mean squared error of permuted inputs to the true squared error on this validation set.

2.3 Preprocessing of variables

The input variables were rescaled to have 0 mean and a standard deviation equal to 1, in the training and test sets separately. Moreover, correlations between the predictors were studied. Additionally, a previous study of the input variables shows a great asymmetry in the distribution of N_2O flux and N leaching, as shown in Figure 1. Using the same experimental protocol (described in

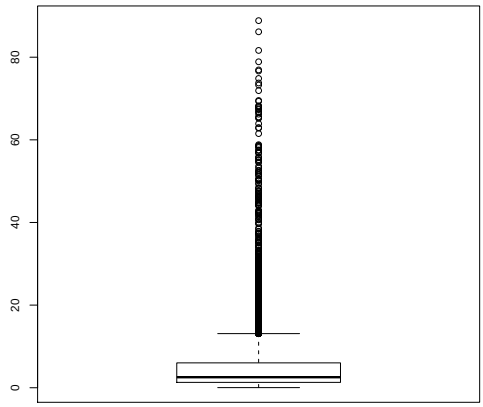

Fig. 1: N_2O flux in scenario 1

section 2.2), two kinds of variables were learned:

- the original variables corresponding to the original values of N_2O flux and N leaching;
- preprocessed variables corresponding to the log-values of N_2O flux and N leaching.

The comparison between the two approaches shows that using the original variables is the best choice. The following results then restrict to that case, avoiding the preprocessed case.

3 Results and comments

3.1 Numerical results

The numerical results (mean squared errors and corresponding R^2 on test sets) are summarized in Table 1. Several comments follow from these results:

- The best predictions are accurate for all scenarios and for both variables to predict with R^2 always greater than 0.8 (and often greater than 0.9). In

		MLP	random forest	SVM
N_2O	Scenario 1	3.278 (90.7%)	**2.705 (92.3%)**	3.141 (91.0%)
	Scenario 2	6.741 (80.3%)	**5.139 (85.0%)**	6.081 (82.3%)
	Scenario 3	6.450 (85.1%)	**5.194 (88.0%)**	5.757 (86.7%)
	Scenario 4	3.597 (88.6%)	**3.011 (90.5%)**	4.350 (86.3%)
	Scenario 5	4.559 (80.6%)	**3.540 (84.9%)**	4.064 (82.7%)
N leaching	Scenario 1	555.7 (89.7%)	351.2 (93.5%)	**179.9 (96.6%)**
	Scenario 2	413.5 (91.4%)	331.8 (93.1%)	**147.1 (97.0%)**
	Scenario 3	447.4 (90.6%)	453.8 (90.4%)	**187.6 (96.0%)**
	Scenario 4	474.2 (80.5%)	317.8 (86.9%)	**229.1 (90.6%)**
	Scenario 5	759.1 (86.5%)	401.9 (92.9%)	**308.0 (94.5%)**

Table 1: Comparison of the prediction performances of the 3 methods on the test sets: mean squared errors and corresponding R^2 (in parentheses). For a given scenario and a given variable to predict, the best method is in bold.

term of R^2 values, the N leaching is often better predicted than the N_2O flux, except for scenario 4.

- MLPs has a bad prediction accuracy and is almost always the worse method. On the contrary, SVMs and random forests both yield interesting results: random forests are always the best method for the prediction of N_2O and SVMs are always the best method for the prediction of N leaching. In addition, for all scenarios and for the two variables to predict, the best method always obtains significantly better results than the second best one, according to a Wilcoxon paired test with level 1% on the residues; the only exception is the prediction of N_2O flux in scenario 4 where random forests does not obtain significantly better results than MLP.

3.2 Most important variables

Figure 2 gives the importance of the predictors by decreasing order for the prediction of N_2O flux by random forest. Only a few variables appear to be important, depending on the scenario: the variables SOC and PH are the most important but, into a much lesser extent, Nr, N_FR and N_MR are also important. The emission of N_2O seems to depend on soil attributes, soil organic matter and PH; the other important factors are the main N inputs represented by the fertilization and manure amendment, and the N concentration in rain which indirectly indicates the amount of N input through depositions.

In the same way, for the prediction of N leaching, a larger number of variables appears to be important for the accuracy of the prediction. Except for scenario 4, the important variables are PH, Nres, SOC, N_MR, N_FR, day and rain. Additionally, in scenario 4, Nfix is also important. The N leaching seems to be influenced mainly by soil attributes, SOC, PH and texture (clay), and by the

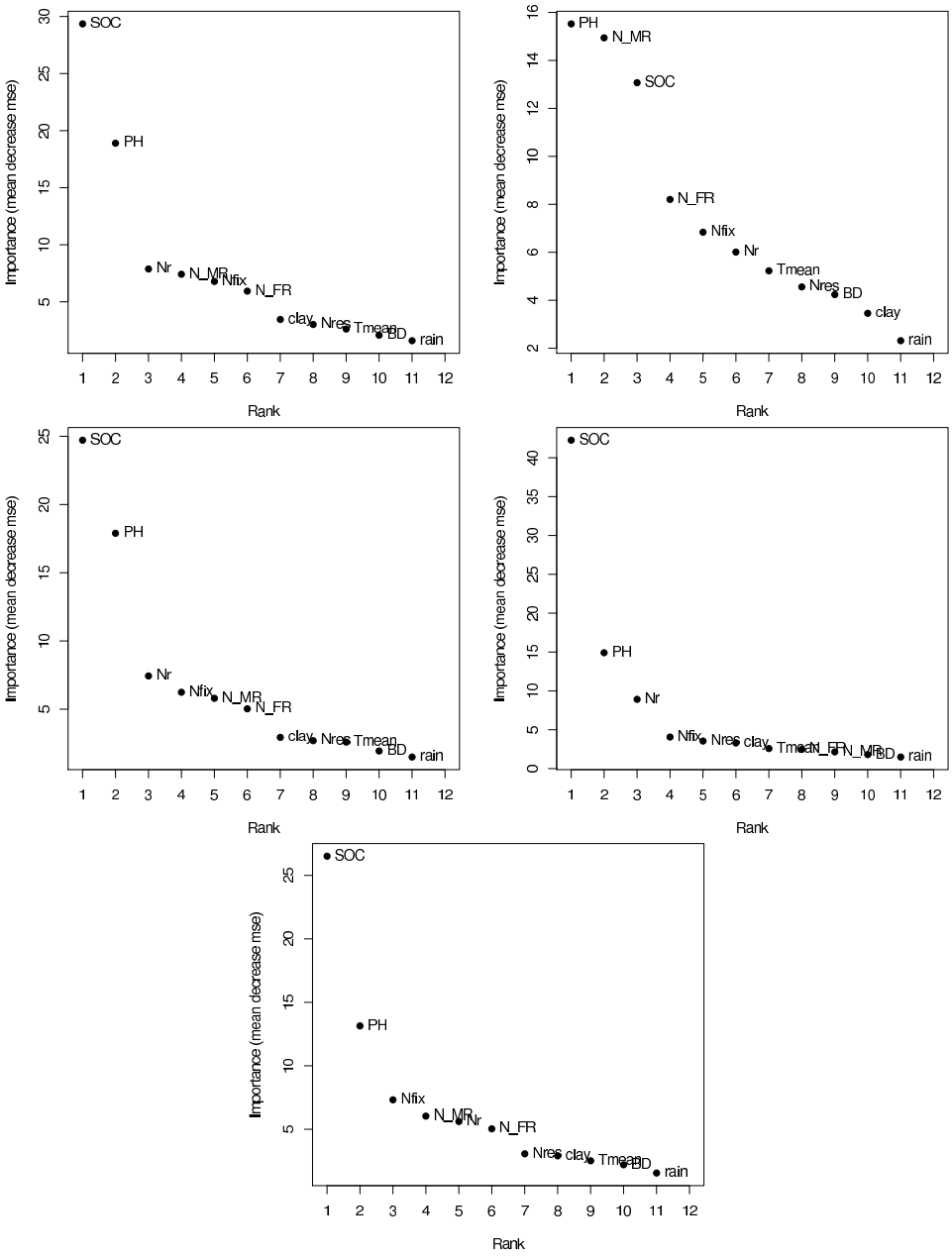

Fig. 2: Importance of the variables in function of their rank (by decreasing order of the importance) for the prediction of N_2O flux by random forest in scenario 1 (top left), 2 (top right), 3 (middle left), 4 (middle right) and 5 (bottom)

N inputs through fertilization, manure application and root residue; the annual rainfall events obviously play an active role in leaching.

4 Conclusion

Three methods have been compared to predict N_2O flux and N leaching in various scenarios. SVMs and random forests achieve the best performances, the first one to predict N leaching and the second one to predict N_2O flux. They are an interesting alternative solution at a macro scale as they can provide fast and accurate estimates for a large number of new inputs. In particular, random forests have a very low computational cost to provide new predictions whereas SVMs can be more demanding (they require the calculation of a kernel matrix with size $N_{\text{new}} \times N$ where N is the number of observations in the training set and N_{new} is the number of new observations to predict).

Moreover, the models also give strong indications about important variables needed to obtain accurate results: the first simulations show that the variables emphasized by the "importance" index can have bio-geochemical interpretations. A deeper analysis should be conducted to confirm this conclusion.

References

[1] Y. Amit and D. Geman. Shape quantization and recognition with random trees. *Neural Computation*, 9:1545–1588, 1997.

[2] P. Atkinson and A. Tatnall. Neural networks in remote sensing. *International Journal of Remote Sensing*, 18(4):699–709, 1997.

[3] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995.

[4] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *5^{th} annual ACM Workshop on COLT*, pages 144–152. D. Haussler Editor, ACM Press, 1992.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] L. Breiman, J. Friedman, R. Olsen, and C. Stone. *Classification and Regression Trees*. Chapman and Hall, 1984.

[7] W. Britz and A. Leip. Development of marginal emission factors for N losses from agricultural soils with the DNDC-CAPRI metamodel. *Agriculture, Ecosystems and Environment*, 133(3-4):267–279, 2009.

[8] W. Britz and P. Witzke. Capri model documentation 2008. Technical report, CAPRI project, Bonn, Germany, 2008. Online at: http://www.capri-model.org/.

[9] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*. 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[10] FAO. More people than ever are victims of hunger. In *FAO Press conference on new hunger figures*, 19 June 2009.

[11] M. Follador and A. Leip. Derivation of DNDC metamodels to evaluate the impact of cross compliance measures. Technical report, EU-STREEP 44423-CCAT project, 2009.

[12] T. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, August 1998.

[13] R. Jones, R. Hiederer, E. Rusco, and L. Montnarella. Estimating organic carbon in the soil of europe for policy support. *European Journal of Soil Science*, 56:655–671, 2005.

[14] A. Krogh and J. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, volume 4, pages 950–957. Kaufmann, M., 1992.

[15] A. Leip, G. Marchi, R. Koeble, M. Kempen, W. Britz, and C. Li. Linking an economic model for european agriculture with a mechanistic model to estimate nitrogen and carbon losses from arable soils in europe. *Biogeosciences*, 5:73–94, 2008.

[16] S. Lek and J. Guégan. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, 120(2-3):65–73, August 2003.

[17] J. Mas, H. Puig, J. Palacio, and A. Sosa-López. Modelling deforestation using gis and artificial neural networks. *Environmental Modelling and Software*, 19:461–471, 2003.

[18] D. Mattera and S. Haykin. *Advances in Kernel Methods: Support Vector Learning*, chapter Support vector machines for dynamic reconstruction of a chaotic system, pages 209–241. MIT Press, 1998.

[19] R Development Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005.

[20] L. Orlandini and A. Leip. A high resolution dataset of european daily weather from 1901-2000 for applications with ecosystem models. In *Proceedings of NitroEurope IP Open Science Conference*, Gent, Belgium, 20-21 February 2008.

[21] M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, January 2005.

[22] J. Peters, B. De Baetsb, N. Verhoest, R. Samson, S. Degroeve, P. De Becker, and W. Huybrechts. Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, 207(2-4):304–318, October 2007.

[23] G. Pineros, A. Ordonez, J. Roosen, and V. M. Metamodeling: theory, concepts and application to nitrate leaching modeling. *Ecological Modeling*, 193(3-4):629–644, 2006.

[24] B. Ripley. Neural networks and related methods for classification. *Journal of the Royal Statistical Society, Series B*, 56(3):409–456, 1994.

[25] G. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98)*, pages 515–521, Madison, Wisconsin, USA, 1998.

[26] T. Simpson, J. Peplinski, P. Koch, and J. Allen. Metamodels for computer-based engineering design: survey and recommendations. *Engineering with Computers*, 17:129–150, 2001.

[27] D. Tilman, K. Cassman, P. Matson, R. Naylor, and S. Polasky. Agricultural sustainability and intensive production practices. *Nature*, 418:671–677, 2002.

[28] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[29] N. Villa, M. Paegelow, M. Camacho Olmedo, L. Cornez, F. Ferraty, L. Ferré, and P. Sarda. Various approaches to predicting land cover in mountain areas. *Communication in Statistics - Simulation and Computation*, 36(1):73–86, 2007.