Linear Vs non-linear learning methods

A comparative study for forest above ground biomass, estimation from texture analysis of satellite images

Hippolyte TAPAMO [1,2] — Adamou Mfopou [1,2] — Blaise Ngonmang[1,2] — Pierre Couteron [4] — Olivier Monga[1,3]

[1] IRD UMI 209 UMMISCO,
Université de Yaoundé I,
B.P. 337 Yaoundé, Cameroun

[2] LIRIMA, Equipe IDASCO,
Faculté des Sciences, Département d'Informatique,
B.P. 812 Yaoundé, Cameroun

[3] UPMC, Paris 6,
EDITE,
Département d'Informatique,
B.P. 812 Yaoundé, Cameroun

[4] IRD UMR AMAP
S/C Cirad, TA A-51/PS2
Boulevard de la Lironde
34 398 Montpellier Cedex 5, France

ABSTRACT. The aboveground biomass estimation is an important question in the scope of Reducing Emission from Deforestation and Forest Degradation (REDD framework of the UNCCC). It is particularly challenging for tropical countries because of the scarcity of accurate ground forest inventory data and of the complexity of the forests. Satellite-borne remote sensing can help solve this problem considering the increasing availability of optical very high spatial resolution images that provide information on the forest structure via texture analysis of the canopy grain. For example, the FOTO (FOurier Texture Ordination) proved relevant for forest biomass prediction in several tropical regions. It uses PCA and linear regression and, in this paper, we suggest applying classification methods such as k-NN (k-nearest neighbors), SVM (support vector machines) and Random Forests to texture descriptors extracted from images via Fourier spectra. Experiments have been carried out on simulated images produced by the software DART (Discrete Anisotropic Radiative Transfer) in reference to information (3D stand mockups) from forests of DRC (Democratic Republic of Congo), CAR (Central

Received, October 07, 2012
Revised, October 22, 2013
Accepted, November 13, 2014

African Republic) and Congo. On this basis, we show that some classification techniques may yield a gain in prediction accuracy of 18 to 20%.

RÉSUMÉ. L'estimation de la biomasse aérienne reste une question ouverte dans le cadre de la Réduction des Emissions dues à la Déforestation et la Dégradation des forêts (cadre REDD de CNUCC). Cette estimation est particulièrement difficile pour les pays tropicaux en raison de l'absence de données sur les inventaires forestiers et de la complexité des forêts. Dans ce contexte, la télédétection peut contribuer à la résolution de ce problème compte tenu de la disponibilité croissante d'images à très haute résolution spatiale. Les méthodes basées sur l'analyse de la texture du grain de la canopée de ces images permettent d'obtenir des informations sur la structure de la forêt et de prédire les valeurs de la biomasse. Par exemple la méthode FOTO (Fourier Texture Ordination) s'est révélée pertinente pour la prédiction de la biomasse forestière dans plusieurs régions tropicales. Elle utilise l'analyse en composantes principales et la régression linéaire pour estimer les valeurs de biomasse. Dans ce papier, nous proposons l'application de méthodes non linéaires de régression tels que k-NN (k plus proches voisins), SVM (Séparateur à Vaste Marge) et les Forêts Aléatoires sur des descripteurs de texture extraits à partir d'images au travers des spectres de Fourier. Nous appliquons et comparons les résultats obtenus par ces méthodes non linéaires sur des images simulées de scènes forestières produites par le logiciel DART (Discrète Anisotrope Radiative Transfert). Les simulations ont été faites en référence à des maquettes 3D basées sur des informations de terrain provenant des forêts de la RDC (République Démocratique du Congo), de la RCA (République CentrAfricaine) et du Congo. Les résultats obtenus montrent que l'utilisation des techniques de régression non linéaires permettent d'obtenir un gain de précision de 18 à 20% sur la prédiction de la biomasse.

KEYWORDS : aboveground biomass, estimation, supervised learning, regression, support vector machines, random forests, k-nearest neighbor.

MOTS-CLÉS : Biomasse forestière, apprentissage supervisé, regression, machines à vecteurs support, forêts aléatoires, k plus proches voisins

1. Introduction

Tropical forests play an important role in the carbon storage process and the understanding and modeling of this process has become a key ecological question [16, 25]. Tropical deforestation and forest degradation account for a large share of anthropogenic carbon emissions [13] and incentives and policies are debated as to reduce this share as part of the Reducing Emissions due to Deforestation and forest Degradation (REDD+) framework. Achieving a successful implementation of carbon's emission reduction policies requires the development and testing of accurate and robust methods for estimation and monitoring of forest above-ground biomass (AGB) over extensive areas in the tropical regions [16].

Forest AGB assessment stems from field work, which combines estimations of biomass at individual tree level (generally from allometric equations using input variables easy to measure such as trunk diameter at breast height (dbh)), as well as sampling large territories to enumerate trees and measure such variables [18]. Such forest inventories which are costly and labor intensive can neither cover large sampling areas nor achieve high temporal frequency of observations as requested for monitoring purposes. Remote sensing techniques have to be used to overcome this limitation and help interpolate scarce field information in space and time [7]. Among them, satellite-borne optical remote-sensing is cheap compared to airborne techniques and its relevance with respect to forest biomass assessment has increased thanks to the increasing availability of high to very high spatial resolution (VHRS) images featuring pixels of sizes of 1 m or less. VHRS imagery allows linking inter-pixel variations (i.e. texture) to forest canopy patterns, notably crown sizes [19]. Texture information on two-dimensional textural images has been shown to correlate with some variables depicting the 3D structure of forest stands as to indirectly predict stand biomass. Among the methods to quantify texture on images based on optical data, those grounded on image texture analysis, like the FOTO Method [6, 22], showed promising results for applications to biomass predictions in wet tropical forests, notably because the texture indices showed no evidence of saturation at high biomass levels [20, 24].

The FOTO Method is a 3-step process: (1) 2-D Fast Fourier Transform and r-spectra computation applied on very high resolutions images, 2) ordination of r-spectra using Principal Component Analysis (PCA), and (3) Linear Regression (LR) model for estimating biomass values or stand parameters from reference datasets (field data or data-derived simulations). PCA and LR are the two mathematical models used for inferring the output of the method. Most of the models used in the published applications found in the FOTO literature are linear [5, 6, 22, 2, 20, 21, 3] and cannot capture all aspects of problems that are inherently nonlinear. In recent decades, several machine learning methods based on nonlinear models emerged to address regression, classification or estimation problems. In this study, we investigate nonlinear methods that can be useful in enhancing steps (2) and (3) to better address biomass estimation in tropical forests.

2. Foto Method

The basic idea of the FOTO method is to estimate the aboveground biomass using canopy grain. This canopy grain depends on the spatial distribution of trees and on the shapes and dimensions of the crowns. Crown dimensions display stable correlations with trunk dimensions (notably dbh) which are the basic predictors of biomass at tree level [1]. Since repetitiveness is the most important characteristic of the texture, the measurement of the degree of repetitiveness in the canopy grain is an important component of the FOTO method. This is done by a 2-dimensional Fourier Transform that shifts canopy grain from spatial to frequency domain [21]. The main steps of the FOTO method are presented in figure 1.

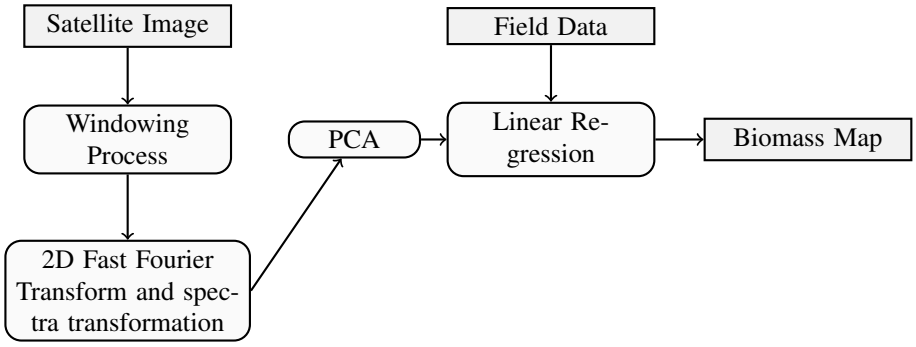

Figure 1. *The FOTO Method*

2.1. 2-D Fast Fourier Transform and r-spectra computation

Before applying the Fast Fourier Transform, it is necessary to define the square window size in which the 2D Fourier Transform is computed. The image is then partitioned into square windows in which Fourier radial spectra are computed. The Fourier coefficients are obtained by a convolution between image values and waveforms of varying directions, and spatial frequency are used to compute r-spectra [21]. The r-spectra neglect orientation information and have proved useful in order to summarize textural properties related to coarseness/fineness, which are of great importance in dealing with canopies of natural forests [6]. Windows with a coarse texture tend to yield r-spectra that are skewed towards small spatial frequencies, while fine-texture lead to spectra that are more balanced.

2.2. Textural ordination

For each particular window, r-spectra are computed and saved in a general table. Each row of the table is the r-spectra of a given window, whereas each column contains the portions of the variance of image radiance explained by a given spatial frequency or wavenumber. This table of r-spectra is submitted to standardized Principal Component Analysis (PCA) which is an eigenvector analysis of the correlation matrix between spatial frequencies. The most prominent components (generally three) are used as texture indices. This enables the ordination of windows along texture gradients [21]. The next step in the process is the application of linear regression to the result produced by the PCA.

2.3. Linear Regression

Linear regression is an attempt to model the relationship between a dependent variable and one or more explanatory variables by fitting a linear equation to observed data. The case of one explanatory variable is called simple linear regression and for more than one explanatory variable, the process is called multiple linear regression [9, 26]. In linear regression, data are modeled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models. The biomass estimation model in FOTO is a multiple linear regression model that uses the first three factorial axis of the PCA.

3. Alternative Regression Methods

3.1. Support Vector for Regression

3.1.1. Principle

As for all the other regression techniques, the basic idea of SVR is to find a function that approximates the training points well by minimizing the prediction error. [27] The main difference between SVR and LR is that all deviations up to a user-specified parameter ϵ are simply discarded. Moreover, when minimizing the error, the risk of over-fitting is reduced by maximizing simultaneously the flatness of the function. Another difference is that what is minimized is normally the predictions' absolute error instead of the squared error used in linear regression. (There are, however, versions of the algorithm that use the squared error instead.)

The value of ϵ defines a tube around the regression function and controls how closely the function will fit the training data. Too large a value will produce a meaningless predictor: in the extreme case, when 2ϵ exceeds the range of class values in the training data,

the regression line is horizontal and the algorithm just predicts the mean class value. On the other hand, for small values of ϵ there may be no tube that encloses all the data. In that case some training points will have non-zero error, and there will be a trade-off between the prediction error and the tube's flatness. For the linear case, the support vector regression function can be written:

$$\hat{y}(X) = \hat{w_0} + \sum_i \alpha(i) X_i^T X \quad [1]$$

where w_0 and $\alpha(i) \geq 0$ are numeric parameters that have to be determined by the learning algorithm. The X_i corresponding to $\alpha(i) > 0$ are support vectors.

To map non linear functions, the product $X_i^T X$ can be replaced by a kernel $K(X_i, X)$. A kernel is a function that maps the data into a higher dimension where the linear mapping is possible. Generally linear mapping in the enlarged space achieve better performances [15]. Examples of kernels are:

– the linear kernel: $K = X_i^T X$, this corresponds to the non kernel SVR;

– the polynomial kernel: $K(x, x') = (1 + < x, x' >)^d$, with d the degree of the polynomial;

– the radius basis function (RBF): $K(X, X') = exp(-\frac{||X-X'||^2}{2\sigma^2})$

3.1.2. Advantages

Due to the possibility to use kernels, one of the main advantages of SVR is that they can represent very complex functions.

3.1.3. Limitations

When compared to LR, the main drawback of SVR is the training and prediction times that are more greater.

3.2. K-Nearest Neighbors

3.2.1. Principle

K-Nearest Neighbors (KNN) methods are memory-based, and require no model to be fit [15]. Given a query point x_0 , we find the k training points $x(r), r = 1, ..., k$ closest in distance to x_0 , and then predict the mean value among the k neighbors. The distance function used depends on the problem and on the data at hand. Examples of usual distance functions are Euclidean and Manhattan distances.

3.2.2. Advantages

The main advantage of KNN is that it does not require learning. Indeed the training set need only to be stored. Another advantage is that it can represent complex functions and captures local variations because its decision depends only on the local neighborhood.

3.2.3. Limitations

The main drawback is the storage need. Indeed all the training instances need to be stored. Some techniques such as help to reduce the number of training examples that must be stored[14, 10].

One other drawback is the prediction time which is more expensive than that of LR. Indeed, to compute the prediction for an instance x, one need to compute the similarity between x and all the other instances of the training set. The complexity can be reduced by clustering the data points [17].

3.3. Artificial Neural Networks

3.3.1. Definition and Principles

An Artificial Neural Network is a two-stage regression or classification model represented by a diagram network. Neural networks encompass a large class of models and learning methods and are nonlinear statistical models[15]. Such networks are organized in layers made of a number of interconnected *nodes* which contain an *activation function*. Data are provided to the network via the *input layer* and with one or more *hidden layers* where the processing is done using a system of weighted *connections*. The last hidden layer is linked to the output layer where the result is given as a vector(resp. scalar) if it is used for classification (resp. regression). When the network has only one hidden layer, it is also called *single hidden layer network* or *single layer perceptron*. Figure 2 shows an example of an artificial neural network with one hidden layer.

3.3.2. Advantages

An Artificial Neural Network can capture many kinds of relationships and therefore allows the user to quickly and relatively easily model phenomena which otherwise may have been very difficult or impossible to represent correctly. Fourier spectra falls well in their category work well when the relationships between variables are not well understood, and when the volume of data is very large.

3.3.3. Limitations

Artificial Neural Network tend to be slower to train than other types of networks. Regarding its structure, it is a parallel computer system and the slowness of the training step is due to the fact that individual artificial neurons are usually processed sequentially.

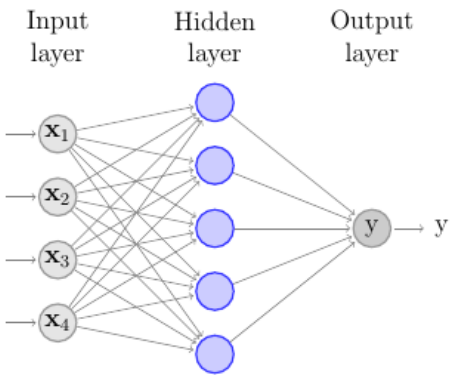

Figure 2. *Single layer perceptron*

3.4. Random Forests

There has been a lot of interest in ensemble classifiers, i.e. methods that generate several classifiers and aggregate their predictions [8]. More precisely, an ensemble classifier constructed from a given training data set, predicts the class of a previously unseen object by combining the predictions obtained from these basic classifiers. This combination aims at improving the accuracy of the basic classifiers.

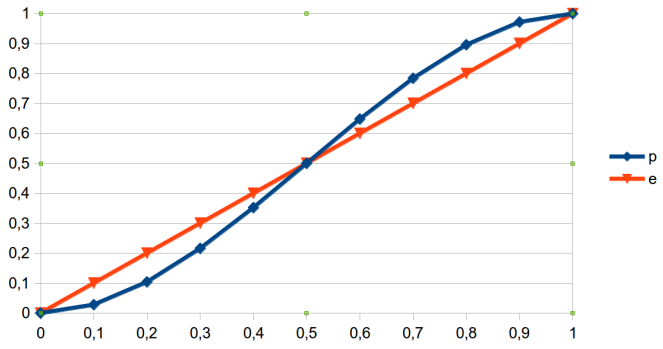

Figure 3. *Error probability for an ensemble of 2 independent classifiers with error rate* e.

For instance, let's consider an ensemble of three classifiers. If these classifiers are identical then the ensemble classifier will produce the same result as the basic classifiers and we will have no gain in the combination. On the other hand, assume that the three basic classifiers have the same error rate e and are independent. Since the ensemble clas-

sifier takes a majority vote on the results produced by the basic classifiers, it will produce an incorrect decision when two of the basic classifiers are wrong, i.e. with probability $p = 3e^2(1-e) + e^3$. As it can be seen in figure 3 , $p < e$ (i.e. the ensemble is worse than the individual classifiers) if $e < 1/2$, and $p > e$ (i.e. the ensemble classifier performs better than the individual ones) if $e > 1/2$. On the other hand, $p = e$ if $e \in \{0, 1/2, 1\}$. More generally, it can be shown that for an ensemble classifier to be more accurate than any of its individual components, there are two conditions : the basic classifiers must be accurate (i.e. they must perform better than the random guessing whose error rate is $1/2$), and they must be diverse, i.e. they must make uncorrelated errors.

Two well-known ensemble methods that are constructed by manipulating the training set are boosting [23] and bagging [4]. In bagging, several training sets are created by resampling (with replacement) the original training set according to a uniform probability distribution. These samples have the same size as the original data set. It can be shown that on average a sample training set contains approximately 63% of the original training data. This method is particularly efficient when the basic classifier is very sensitive to fluctuations in the training data because its variance is reduced as compared to the variance of the basic classifier. Boosting is an iterative technique that forces the classifier to focus on examples that are hard to classify. More precisely, all examples have initially the same weight. The classifier obtained after each step is used to classify all the elements of the training set. The weights of examples whose classes are not predicted correctly are increased, whereas the weights of examples that are classified correctly are decreased. Various versions of boosting are obtained by varying the way the weights are updated and by considering various techniques for the aggregation of the predictions made by basic classifiers.

Breiman [4] has proposed a method called random forests, that is specifically suited for decision tree classifiers. In this technique, the basic classifiers are decision trees obtained by manipulating the input features. More precisely, a random vector is incorporated into the tree construction process by selecting randomly at each node F input features for splitting. This means that the decision to split a node is taken by examining not all the attributes, but rather by considering the subset consisting of F. Randomness can also be increased by considering bagging techniques to construct m training sets. Decisions for such ensemble classifiers are taken using a majority rule for classification and weighted sum for regression.

These methods can be improved quantitatively by using only a subset of the most relevant attributes.

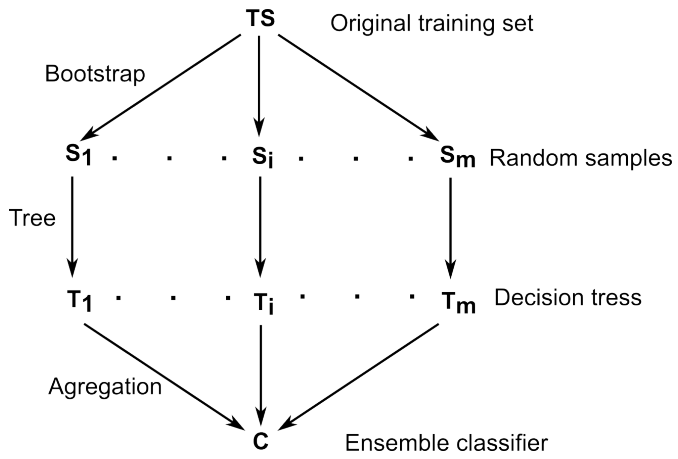

Figure 4. *Random Forests principle*

4. Feature Selection

In machine learning and statistics, attributes (or features or variables) selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that data contain many redundant or irrelevant features. Note that different features are said to be redundant if they provide the same information. Irrelevant features provide no useful information. Feature selection techniques provide three main benefits when constructing predictive models:

– improved model interpretability,

– shorter training time,

– enhanced generalisation by reducing overfitting.

Feature selection is also useful as part of a data analysis process, as it shows which features are important for prediction, and how these features are related.

In the experimentations presented in section 4, the Correlation Feature Selection (CFS) measure is used. It evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classes of objects and not correlated to each other" [12].

Let f and f' be two features, let denote $r_{f,f'}$ the correlation coefficient between f and f' with respect to the training set. For a subset of features S and a feature f, we denote

the average of $\overline{r_{f,f'}}$ for $f' \in S$. Hall has suggested to evaluate the merit of feature subset S of size k with respect to the class c by the following formula:

$$Merit(S) = \frac{k\overline{r_{c,f}}}{\sqrt{k + k(k-1)\overline{r_{f,f}}}}$$

In this formula, the numerator represents the correlation between class c and the features f of S. The denominator is a normalizing factor. Let us consider the situation where all the features of S have the same correlation coefficient r with c, clearly $\overline{r_{c,S}} = r$. If the elements of S are very strongly correlated, i.e. $r_{f,f'} \approx 1$ for each $f, f' \in S$, then $r_{S,S} = 1$ and $Merit(S) = r$. On the contrary, if the features in S are independent from each other i.e $r_{f,f'} = 0$ for $f \neq f'$, then $r_{S,S} = \frac{k}{k \times k} = \frac{1}{k}$ because there are k diagonal elements equal to 1 and a total of k^2 elements. As a consequence, $Merit(S) = \frac{k \times r}{\sqrt{2k-1}}$. This illustrates the fact that Merit(S) is high when features in S are highly correlated to class c and not correlated to each other. In this paper, the feature selection process aims at identifying a feature subset S with Merit as high as possible and superior to a certain threshold. This process can be done according to algorithm 4 described by Hall. In this algorithm, MAX represents the maximum number of levels to test when the best Merit does not change. We took $MAX = 5$ in this article. In fact, when MAX is greater than 5, there is no improvement in the results and the number of subsets tested increases, which leads to increased computation time.

5. Experiments

5.1. Dataset description

Reference data have been provided by N. Barbier using a simulator of 3D forest mock-ups "Allostand" as in [2] in order to roughly mimic in terms of tree densities and dbh distributions forest types that have been locally observed in three countries of Central Africa. Canopy images were simulated from mock-ups using a radiative transfer model (DART, [11]). DART is a software tool based on an approach that combines a ray-tracing model and discrete ordinate methods to simulate, simultaneously in several wavelengths of the optical domain, remotely sensed images of heterogeneous natural and urban landscapes. DART software is made at CESBIO and can be downloaded for free at their website *http://www.cesbio.ups-tlse.fr/fr/dart.html*.

The data is a set of gray scale square images of size 99×99 pixels, analogous to $1ha$ of forest in the field, with their corresponding AGB which is given in kilograms per hectare.

– Congo dataset: 176 images. The AGB values for this dataset range from $68184kg/ha$ to $480649kg/ha$ with a mean of $255805kg/ha$ and a standard deviation

Algorithm 1 Basic structure of working set algorithm

$S \leftarrow \emptyset$ ▷ S is the Current best subset
$M_S \leftarrow 0$ ▷ M_S is the Merit of the current subset S
$term \leftarrow MAX$ ▷ Number of levels evaluated after the current subset
for $k = 0$ to n **do** ▷ n is the number of items in the subset
  if $term > 0$ **then**
    $L_k \leftarrow$ all sets of k features containing S
    for all subset $S' \in L_k$ **do**
      $Merit_{S'} \leftarrow \frac{k \times r_{S',c}}{\sqrt{k+k(k-1) \times r_{S',S'}}}$
      $M_{S'} \leftarrow Merit_{S'}$
      if $M_{S'} > M_S$ **then**
        $S \leftarrow S'$
        $M_S \leftarrow M_{S'}$
      end if
    end for
    if $Size(S) = k$ **then** ▷ Testing if the current subset S is in the current level k
      $term \leftarrow Max$
    else
      $term \leftarrow term - 1$
    end if
  end if
end for
return S

of $93192kg/ha$.

– Democratic Republic of Congo dataset: 105 images. The AGB values for this dataset range from $55311kg/ha$ to $328223kg/ha$ with a mean of $167655kg/ha$ and a standard deviation of $55476kg/ha$.

– Central African Republic dataset 109 images. The AGB values for this dataset range from $3259kg/ha$ to $523208kg/ha$ with a mean of $287586kg/ha$ and its standard deviation is $103148kg/ha$.

These datasets are used as original training examples.

5.2. Attribute selection

The original dataset has 67 features labeled $F1, F2, \ldots, F67$ resulting from the r-spectra calculation step. The attribute selection process applied on the Central Africa Republic dataset produced 12 features. The same process applied on the Congo and the Democratic Republic of Congo datasets produced respectively 14 and 31 features. These features are used only for SVR and k-NN predictions. In the case of Random Forests,

Dataset	Number of selected attributes	Number of subset explored	Merit
Congo	14	854	0.791
DRC	31	1790	0.769
CAR	12	960	0.686

Table 1. *Table of data and the parameters used for subset selection*

prediction is performed on the entire dataset because the algorithm already includes a feature selection.

5.3. Performance evaluation

In order to compare two classifiers, we must be able to compare their performances. Ideally, the performance of a model is measured by the generalization error, i.e. the error made on unforeseen examples. The problem is that we don't have access to these examples. A way of overcoming this difficulty is to use the error made on a test set as estimate of the generalization error. In this approach, one can partition the original data set into two disjoint subsets : a training set and a test set.

Cross-validation is a nice way to insure that each example is used once for training and once for testing. For instance one can consider a partition of the data set into two subsets, use one for training and one for testing. The roles of these subsets are then swapped so that the training (resp. test) set becomes the test (resp. training) set. This 2-fold approach can be generalized to obtain a k-fold cross-validation : this method partitions the dataset into k equal-sized subsets. At the i^{th} iteration, the i^{th} subset is used as test set while the union of the other subsets serves as training set. The performance of the classifier is estimated by considering the average accuracy of the k models constructed during the k runs. The advantage of this approach is that each example is used exactly once for testing and $k-1$ times for training. This ensures that the estimate of predictive accuracy is less biased.

6. Results and Discussions

6.1. Results

For each classifier and each dataset, we used a ten-fold cross validation scheme to estimate the error. The evaluation criterion is the average mean absolute error (AMAE) i.e. the average of the absolute value of the difference between the real and the predicted AGB values for all the images of the dataset. The results are presented in Table 2, Table 3 and

Technique	Parameters	AMAE	Gain
Original FOTO		56590	
FOTO with SVR	RBF Kernel	46103.74	18.53 %
FOTO with k-NN	K=9	48344.67	14.57 %
FOTO with k-NN	K=11	48232.02	14.77 %
FOTO with RF	ntree=50	77554.91	-37.05 %
FOTO with **Attribute selection + SVR**	**RBF Kernel**	**45356.84**	**19.85 %**
Attribute selection + k-NN	K=5	45593.04	19.43 %
FOTO with **Attribute selection + k-NN**	**K=7**	**45280.02**	**19.99** %

Table 2. *Results on Congo dataset*

Technique	Parameters	AMAE	Gain
Original FOTO		81890	
FOTO with SVR	RBF Kernel	66643.02	18.62 %
FOTO with k-NN	K=3	72857.69	11.03 %
FOTO with k-NN	K=5	70765.04	13.59 %
FOTO with **RF**	**ntree=50**	**67140.53**	**18.01 %**
FOTO with **Attribute selection + SVR**	**RBF Kernel**	**66633.88**	**18.63 %**
Attribute selection + k-NN	K=5	71077.77	13.20 %
Attribute selection + k-NN	K=3	70634.69	13.74 %

Table 3. *Results on CAR dataset*

Table 4 respectively for Congo, CAR (Central African Republic) and DRC (Democratic Republic of Congo) datasets.

We see from these results that the FOTO method is improved by using Support Vector for Regression with attribute selection. Depending on the dataset we can have an improvement of more than 19%. In fact the regression method must be selected according to the dataset in presence. For some datasets, the linear regression can be a good choice but for other datasets it can be a bad one.

For the Congo dataset, we obtained the major improvement with k-NN combined with attribute selection using $k = 7$ and a gain of 19.99% followed, in the quality of result by SVM with attribute selection with a gain of 19.85%

For the DRC dataset, the algorithms that yield the best improvements are SVM with attribute selection and Random Forests (RF) with respective gains of 23.24% and 21.95%.

Technique	Parameters	AMAE	Gain
Original FOTO		39760	
FOTO with SVR	RBF Kernel	31458.90	20.88 %
FOTO with k-NN	K=5	33608.78	15.47 %
FOTO with **RF**	**ntree=50**	**33079.94**	**21.95 %**
FOTO with **Attribute selection + SVR**	**RBF Kernel**	**30519.11**	**23.24 %**
Attribute selection + k-NN	K=2	33608.78	15.47 %

Table 4. *Results on DRC dataset*

For the CAR dataset, the algorithms that yield the best improvement are SVM with attribute selection and Random Forests with respective gains of 18.63% and 18.01%.

We notice from these results that the new algorithms presented here improve the original FOTO Method except Random Forests applied to Congo dataset, where the FOTO Method gives better results.

6.2. Discussions

In this work, we have explored some non-linear machine learning methods for biomass estimation in tropical forest areas based on canopy image texture. The dataset was composed of simulated images relating to three different forest types yielding diversified image textures and canopy aspects (open vs. closed canopy, fine vs. coarse texture). The relationship between tree biomass and crown dimensions is non-linear and the ensuing relationship between stand biomass and features of canopy texture is liable to be non-linear in many situations. We observed here that nonlinear methods applied on textural indices yield better results compared to the linear method based on PCA and linear regression. Average prediction errors as presented in tables 2, 3 and 4 confirm this assertion. In average, nonlinear methods have a gain of 19.50% compared to the linear method. Non-linearity is then an important factor to be taken into account in biomass estimation using canopy texture ordination.

PCA used by the FOTO method has the advantage to reduce the amount of information useful to understand the phenomenon being studied but may frequently fail to produce the most significant texture variables compared to what we have with attribute selection. Attribute selection provides more suitable variables to depict texture.

7. Conclusions

Assessing forest above-ground biomass (AGB) over extensive territories of poor accessibility is a challenging task. The FOTO method introduced by Couteron [6] and Proisy et al. [22] has proved powerful to provide reliable biomass predictions from optical canopy images in different regions in the tropics ([20, 24, 3]). In this paper, we have explored the possibility of improving the performance of this method by non-linear regression techniques. Experiments on simulated images of the canopy of the forests in reference to local stand characteristics in DRC (Democratic Republic of Congo), CAR (Central African Republic), and Congo demonstrate that our approach is robust and give results with less errors than in linear regression used in FOTO. Accuracy gains of about 20% have been obtained using the proposed approach.

8. Acknowledgement

We thank Nicolas Barbier (IRD, UMR AMAP) for providing the simulated images used in this paper as test dataset and Jean-Philippe Gastellu-Etchegorry (Univ. of Toulouse, UMR CESBIO) for granting the free use of the DART software to carry out such simulations.

We also thank Pierre Ploton (IRD, UMR AMAP and ENS Yaoundé, Department of Biological Sciences) for fruitful discussions.

We thank Professor Maurice Tchuente (UMMISCO, Yaoundé) for his valuable and constructive suggestions and inspiring comments. His willingness to give his time so generously has been very much appreciated.

References

[1] C. Antin, R Pélissier, G. Vincent, and P. Couteron. Crown allometries are less responsive than stem allometry to tree size and habitat variations in an indian monsoon forest. *Trees*, 27:1485–1495, 2013.

[2] N. Barbier, P. Couteron, J.P Gastelly-Etchegorry, and C. Proisy. Linking canopy images to forest structural parameters: potential of a modeling framework. *Annals of Forest Science*, 69:305–311, 2012.

[3] J-F. Bastin, N. Barbier, P. Couteron, P Adams, A. Shapiro, J. Bogaert, and C. De Cannière. Aboveground biomass mapping of African forest mosaics using canopy texture analysis: towoard a regional approach. *Ecological Applications. In press.*

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

[5] P. Couteron. Quantifying change in patterned semi-arid vegetation by fourier analysis of digitized aerial photographs. *International Journal of Remote sensing*, 23(17):3407–3425, 2002.

[6] P. Couteron, R. Pelissier, E. Nicolini, and D. Paget. Predicting tropical forest stand structure parameters from fourier transform of very high-resolution remotely sensed canopy images. *Journal of Applied Ecology*, 42:1121–1128, 2005.

[7] R. DeFries, F. Achard, S. Brown, M. Herold, D. Murdiyarso, B. Schlamadinger, and C. de Souza. Earth observations for estimating greenhouse gas emissions from deforestation in developing countries. *Environmental Science and Policy*, 10:385–394, 2007.

[8] T. G. Dietterich. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 1–15, London, UK, UK, 2000. Springer-Verlag.

[9] N.R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley, 1998.

[10] Hatem A. Fayed and Amir F. Atiya. A novel template reduction approach for the k-nearest neighbor method. *Trans. Neur. Netw.*, 20(5):890–896, May 2009.

[11] J. P. Gastellu-Etchegorry. 3D modeling of satellite spectral images, radiation budget and energy budget of urban landscapes. *Meteorology and Atmospheric Physics*, 102:187–207, 2008.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, November 2009.

[13] N. L. Harris, and S. Brown, S. C. Hagen, S. S. Saatchi, S. Petrova, W. Salas, M. C. Hansen, P. V. Potapov, and A. Lotsch. Baseline map of carbon emissions from deforestation in tropical regions. *Science*, 336:1573–1576, 2012.

[14] P. E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:515–516, 1968.

[15] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[16] R. A. Houghton, F. Hall, and S. J. Goetz. Importance of biomass in the global carbon cycle. *Journal of Geophysical Research*, 114, 2009.

[17] Dalong Li, Steven Simske, Dalong Li, and Steven Simske. Training set compression by incremental clustering, 2011.

[18] D. Maniatis and D. Mollicone. Options for sampling and stratification for national forest inventories to implement redd+ under the unfccc. *Carbon Balance and Management*, 5(9), 2010.

[19] M. Palace, and M. Keller, G. P. Asner, S. Hagen, and B. Braswell. Amazon forest structure from ikonos satellite data and the automated characterization of forest canopy properties. *Biotropica*, 40:141–150, 2008.

[20] P. Ploton, R. Pélissier, C. Proisy, T. Flavenot, N. Barbier, S. N. Rai, and P. Couteron. Assessing aboveground tropical forest biomass using Google Earth canopy images. *Ecological Applications*, 22(3):993–1003, 2012.

[21] C. Proisy, N. Barbier, M. Guéroult, R. Pélissier, J. P. Gastellu-Etchegorry, E. Grau, and P. Couteron. Biomass prediction in tropical forests: The canopy grain approach. In *Remote Sensing of Biomass - Principles and Applications*. Temilola Fatoyinbo, 2012.

[22] C. Proisy, P. Couteron, and F. Fromard. Predicting and mapping mangrove biomass from canopy grain analysis using Fourier-based textural ordination of Ikonos images. *Remote Sensing of Environment*, 109(3):379–392, 2007.

[23] R. Shapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26(5):1651–1686, 1998.

[24] M. Singh, Y. Malhi, and S. Bhagwat. Biomass estimation of mixed forest landscape using a fourier transform texture-based approach on very-highresolution optical satellite imagery. *International Journal of Remote Sensing*, 35(9):3331–3349, 2014.

[25] Thomas F Stocker, Q Dahe, and Gian-Kasper Plattner. Climate change 2013: The physical science basis. *Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Summary for Policymakers (IPCC, 2013)*, 2013.

[26] Sanford Weisberg. *Applied linear regression*. John Wiley & Sons, 2014.

[27] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.