http://economix.fr

Model economic phenomena with CART and Random Forest algorithms

Document de Travail
Working Paper
2017-46

Benjamin David

EconomiX - UMR7235
Université Paris Nanterre
Bâtiment G, 200, Avenue de la République
92001 Nanterre cedex

Fax : 33 (0) 1 40 97 41 98
Email : secretariat@economix.fr

Model economic phenomena with CART and Random Forest algorithms

Benjamin David[a]

Abstract

The aim of this paper is to highlight the advantages of algorithmic methods for economic research with quantitative orientation. We describe four typical problems involved in econometric modeling, namely the choice of explanatory variables, a functional form, a probability distribution and the inclusion of interactions in a model. We detail how those problems can be solved by using "CART" and "Random Forest" algorithms in a context of massive increasing data availability. We base our analysis on two examples, the identification of growth drivers and the prediction of growth cycles. More generally, we also discuss the application fields of these methods that come from a machine-learning framework by underlining their potential for economic applications.

Keywords: decision trees, CART, Random Forest

JEL Classification: C4, C18, C38, C55

[a]EconomiX-CNRS, Université Paris Nanterre, France. email: benjamin.david11@hotmail.fr

1 Introduction: some obstacles to econometric modeling

Over the past few years and decades, economics has placed an increasing emphasis on econometric applications. Thus, almost 70% of the articles published in the three main economics journals in 2011 were empirical works (Hamermesh, 2013). This orientation toward quantitative methods is driven by digitalization, enhancements of econometric tools and computer capacities growth, which has deeply reoriented economic research. The more common approach in econometric applications is the consideration of a variable Y explained with J predictors X_j linked to Y by a specific functional form determining the type of relationship. A simple example is the use of a linear function such as $Y = f(X_j) = \alpha + \sum_{j=1}^{J} \beta_j X_j$ with α and β_j as parameters whose values describe the relationship. The statistical work of an economist is to estimate α and β_j on the basis of a limited number of realizations (the observations) of the variables Y et X_j which could be considered as random variables[1] [2]. This type of approach is consistent with the seminal view of Haavelmo (1944) in which economic phenomena must be studied in probabilistic framework.
The estimation of the parameters of an econometric model could serve three purposes:

- identification of an effect: does the variable X_j affect variable Y?
- quantification of an effect: what is the importance of a measured effect?
- prediction of the Y values on the basis of the estimated model: for some fixed values of X_j which might be the probable values of Y?

This probabilistic approach called "stochastic data modelling" by Breiman (2001b)is the standard for the most part of econometrics works and manuals. It provides interesting quantitative results by using a large collection of tools built on solid mathematical foundations. Their effectiveness explains why this approach is present in other fields such as biology, medical research, engineering or political sciences. However, from an economist point of view, the use of this approach could be hindered by several obstacles linked to the complexity and the diversity of economic phenomena. Among these difficulties, four seem to be prominent.

(i) Choice of explanatory variables

The first important question in a modeling attempt is the choice of variables to be included in a model. Which are the best suited variables for explaining the variable of interest? A natural solution is to refer to the economic theory that is in line with the proposition of Frisch (1933). Indeed, this founder of econometrics considers that one of the dimensions of this field is economic theory, in addition to mathematics and statistics. Estimation, that is, a statistical task, relies primarily on quantifiable relationships identified on a theoretical basis. Nevertheless, this interesting strategy

[1] The probabilistic nature of the model is made visible by the addition of an error term (ϵ). Thus, the relationship takes this following form: $Y = f(X_j) = \alpha + \sum_{j=1}^{J} \beta_j X_j + \epsilon$

[2] In most cases, the stochastic nature of X_j variables is called into question for reasons of simplification. The perspective is to consider Y conditional to fixed values of X_j.

faces the diversity of theoretical proposals. From this perspective, a simple example is the study of growth determinants, because recognized models underline the role of many factors behind this dynamic. The standard models constructed by Ramsey (1928), Solow (1956) and Swan (1956) consider two factors, which are capital and labor. Other authors identify several additional factors such as "learning by doing" (Romer, 1986), education (Lucas, 1988), government spending (Barro, 1999), research and development (Aghion and Howitt, 1992) and energy (Kümmel et al., 2010), to name just a few. Furthermore, Sala-I-Martin (1997b) notes that many other predictors were introduced in empirical papers. He founds some 60 variables (growth drivers) with an associated coefficient significantly different from 0 at least once in a regression. It should be noted that since this publication, other variables have been used and the number of predictors can be higher if we consider their lagged values. The choice of variables contains a two-fold objective, which is the identification of the "good" specification in order to avoid the omitted variables bias and the reach of sufficient degrees of freedom to ensure feasibility and precision of estimation. This obstacle could be significant if data used are macroeconomic with annual frequency and the number of possible predictors is large.

(ii) Choice of a functional form

The choice of a functional form refers to the selection of a function of predictor variables able to approximate the observations of an output variable. This is a very important step in econometric modeling because it determines the kind of relationship to estimate; only parameters values will be estimated. For historical reasons, simplicity and conformity to data, a large number of works address the economic process in a linear framework. Indeed, a method such as Ordinary Least Squares (OLS) has been available to economists for a long time; it is accessible for non-specialists and it might be suitable for a certain number of phenomena with a linear pattern. This is also explained by the fact that it is possible to transform a nonlinear model in a linear model by transforming the variables used. This enables a consideration of different functional forms by conserving the linearity in the parameters but not in the variables. From this perspective, we can cite the example of the standard Cobb-Douglas function (Cobb and Douglas, 1928) or many power laws observed in economy, finance or other fields (Mandelbrot, 1963; Gabaix, 2016), which can be transformed with a logarithm function.

However, all economic relationships are not linear or linear after transformation, which complicates the identification of an adapted functional form. Indeed, nonlinearity takes on a wide diversity of forms because the relationships between economic variables could include many specific features such as threshold, structural breaks, deceleration or reinforcing. Nonlinearity seems to be at work in many economic topics such as in the link between wage, education and experience (Mincer, 1974), in Ricardian equivalence (OCDE, 2015) or in the link between oil prices and the value of the dollar (Coudert and Mignon, 2016). In order to avoid potential specification errors, it is possible to make a closer inspection of data by using graphical representations, to refer to previous studies on the same topic or to practice specification tests allowing provision of insights on the form of the relationship studied. In despite the precautions, the risks of misspecifications are substantial and can produce some

imprecisions in the estimations or even some spurious results if the functional form used is significantly unsuitable.

(iii) Choice of a probability distribution

Another important step in econometric modeling is the selection of a probability distribution $P(X, Y)$ to model the joint distribution of variables.[3] This choice can be the basis of estimation because some approaches are usable only if a specific probability distribution is defined before estimation. This is the case of all models estimated by maximum likelihood. In addition, the choice of distribution can be necessary to apply statistical tests or to compute confidence and prediction intervals. For instance, a student test on slopes from a linear regression is relevant under the normality hypothesis.[4] Therefore, this step is important because it conditions the estimated parameters' values and the level of confidence in these results. In some cases, the choice could be not too difficult due to the statistical issue. For a logistic regression, it seems logical to consider a Bernouilli distribution when Y is a binary variable, and a multinomial distribution when the number of categories is higher than two. The form of the problem leads naturally to these choices.

It is also possible to remove the risks associated with the choice of unsuitable distribution by studying the residuals of a model through specific statistical tests, which can establish the adequacy of a particular distribution for a given confidence level. On the other hand, economic literature could guide the modeler since specific probability distributions have been identified in particular contexts. For example, it seems that financial data have singular characteristics such as heavy tails of returns distribution (Cont, 2001). It means that the number of extreme events (large deviations around the mean) is higher than if returns were normally distributed. For this reason, estimation of (G)ARCH models (Engle, 1982; Bollerslev, 1986) by quasi maximum likelihood is possible by using a student distribution or a "General Error Distribution" (GED), which have heavy tails.

However, the use of specific probability distribution could be viewed as a strong hypothesis because all economic phenomena are not precisely specified and the power of statistical tests could be weak with samples with few observations (see for example Jarque and Bera (1980) test). In addition, the diagnostics could be dependent on the threshold chosen for a type 1 error or different according the test used. Thus, despite precautions, there can be uncertainty about the probability distribution or it can be impossible to specify it.

(iv) Inclusion of interactions[5]

A particular type of nonlinearity occurs when the impact of a variable on the variable to be modeled depends on the levels or the variations of other explanatory variables. Numerous economic questions suppose this kind of configuration. Will a

[3] The choice of a functional form already mentioned is a part of this step because the functional form can be viewed as the conditional expectation function of Y given X.

[4] This is true with finite samples but under additional hypothesis, $\hat{\beta}$ converges asymptotically towards the standard normal distribution.

[5] This problem could be included in the choice of a functional form. However, we devote a specific paragraph to this topic due to its importance.

fiscal stimulus have the same effect depending on the structure of consumption and propensity to save? Will the effect of a development aid policy have the same effect according to the state of the main macroeconomic variables?
(Morgan and Sonquist, 1963) explain that these "interactions" are common in social sciences and that certain economic variables are sums of interactions by construction. For example, interest in industrialization of countries or regions returns to study an interactive process between capital accumulation, labor productivity growth, firms and territory reorganizations, evolution of human capital, etc. Thus, to study the impact of one of these factors without taking into account the other is probably an unproductive approach. It is even possible to wonder if the majority of economic variables interact and if separable effects do not fall within the exception. Indeed, economic systems can be viewed as "complex systems" (Arthur, 1999) that suppose that the elements in an economic system are heterogeneous, in interaction and linked by nonlinear retroactions. Thus, it appears difficult to describe economy with only "mechanical", unidirectional and independent relationships.
From a technical point of view, this situation of interaction makes it hard to simply consider an additive function of predictors to approximate the variable of interest. It becomes necessary to model more complex links with specific patterns such as threshold and symbiotic relationship. In economic research, interactions are taken into account in a variety of ways. One popular method consists of adding interaction terms in regression models. If the effect of a variable X_1 on Y depends on the level of X_2, one can create an additional variable $X_1 \times X_2$ whose associated coefficient captures the interaction. The marginal effect of X_1 would be a function of the X_2 values. Nonlinear econometric models such as switching regression models can also integrate this kind of dependence (see Teräsvirta et al. (2010)). In this way, the output variable is approximated by different regimes determined by the value of a transition variable. For specific range of values of the transition variable, Y is described by one regime and for other values by other regimes. Thus, the estimated coefficients are a function of the regime considered. These approaches are very useful but have a cost in terms of degrees of freedom and in the capacity to interpret the results. It is therefore difficult to consider a large number of interactions with these methods.

"Big data"

Obstacles described in the previous lines are inherent to econometric modeling but they can be reinforced by the evolution of the informational context of economic research. Indeed, econometric works depend on the amount of available data and the ability to process them, which have both been significantly increased. On data availability, Einav and Levin (2014a) summarize the situation: "data is now available faster, has greater coverage and scope, and includes new types of observations and measurements that previously were not available". These authors also identify the end of "rectangular data" (Einav and Levin, 2014b) that correspond to data where the number of variables is less than the number of observations. This new order of magnitude of the applied statistical problems is technically based on the increase in the number of devices capable of digitizing, storing and transforming information (personal computers, servers, smartphones, etc.). The main providers of these new data are mainly governments and firms that compile huge clients, vendors, house-

holds or business file bases (Einav and Levin, 2014b).
It should also be noted that national and international institutions are involved in the creation of very large databases. For instance, the International Monetary Fund (IMF) provides very detailed data on international trade with the "DOTS" database. For a country like Canada, we can find all past information on its imports and exports. Data are available for some trading partners over a period of 65 years at a monthly frequency. This is not really "big data" because variables can be observed a few thousand times; however this is an order of magnitude unprecedented for certain thematics.
This huge amount of these newly available data reinforces the obstacles described above, transforming a problem of lack of data in a problem of plenty of data. Indeed, they give the possibility of using new variables and asking new economic questions for which there are no prior expertise. Uncertainty about the relevant variables, the functional form, the underlying probability distribution and the interactions are therefore amplified.

"Blind spots" and algorithmic solutions

The four obstacles previously described could therefore pose significant problems in econometric modeling. In some cases, they are marginal issues or they can be circumvented. However, if they are important or if they are combined, there may be some "blind spots" for empirical analysis because standard econometric tools cannot intrinsically model the considered phenomena. It is possible that some applications cannot be achieved or that certain results cannot be obtained by construction. In this paper, we argue that one possible access to these blind spots relies on the use of decision trees [6] built with the "Classification and Regression Tree" (CART) (Breiman et al., 1984) and "Random Forest" (RF) (Breiman, 2001a) algorithms.
Before presenting these approaches, which come from a machine-learning framework, we can first remove the problem of their legitimacy in economic research. Indeed, the first formulation of a tree-based method[7] was done by an economist, James Morgan (Morgan and Sonquist, 1963), who proposed with his coauthor the "Automatic Interaction Detector" (AID). Furthermore, we can recall that econometrics, even if it is a part of economics, is clearly linked to other fields such as mathematics or computer science. Historically, it has used many tools from other fields. To give just one example, we refer to the statistical concepts and methods created by Ronald A. Fisher for biology or genetics applications, which have become standard for economists (maximum likelihood, etc.). Inversely, some econometric methods have been successfully applied in other fields such as cointegration in climatology (Schmith et al., 2012) or GARCH modeling in hydrology (Wang, 2006). Therefore, it seems appropriate to take an interest in nonstandard approaches if they offer good performance and are able to shed light on economic issues. Among them, decision trees have been highly recognized to the point of being considered as a part of "Top 10 algorithms in data mining" (Wu et al., 2008), and Random Forest, which is an extension of decision

[6] Decision trees are either regression trees or classification trees. Other algorithms than CART can be used to estimate these models.

[7] For a presentation of the history and the different types of regression and classification trees, reader could refers to (Loh, 2014).

trees, has been viewed as "one of the most accurate general-purpose learning techniques available" (Biau, 2012). Logically, these methods have been applied in a wide variety of fields, such as particle physics (Collaboration, 2012), genetics (Goldstein et al., 2010) or computer vision (Shotton et al., 2011).

Despite this multidisciplinary recognition, these methods have curiously only been minimally presented in manuals and barely used in economic papers. (Einav and Levin, 2014a) note, "The common techniques in this sort of data mining—classification and regression trees, lasso and methods to estimate sparsemodels, boosting, model averaging, and cross-validation—have not seen much use in economics". Some research on the *econpapers* database shows, for instance, that only 14 papers using CART were published in an economic peer-review journal over the period 1984-2016 [8] [9] while this account is based on a large definition of economics, including financial purposes. A recent counter-example is the Varian's article (Varian, 2014) within which the author tried to promote a set of tools from computer science (including decision trees).
Most of the time, the use of these approaches is justified by the achievement of good performance in prediction (in or out sample). This interesting argument (see section 4) must not hide that they are also able to solve technical problems in econometric modeling. It is on this point that we insist in this paper by showing that CART and Random Forest algorithms could overcome the obstacles described in the introduction.
Before starting the presentation of these methods, it appears necessary to provide some clarifications. First, this work is not a criticism of standard econometric tools, but instead, it aims to highlight the advantages of other complementary approaches. Moreover, this work is not a technical[10] and exhaustive presentation. We only take the major features and we leave aside many adjoining and interesting contributions. Thus, this analysis is solely focused on the base approaches, namely CART (Breiman et al., 1984) and Random Forest (Breiman, 2001a). We don't consider numerous extensions and enhancements such as the "Boosting" approach (Freund and Schapire, 1996), the "Conditional Inference Forest" (Hothorn et al., 2006), the oblique trees (Heath et al., 1993), the time series trees and forests (Sela and Simonoff, 2011 ; Deng et al, 2013), some concurrent algorithms such as ID3, C4-5 (Quinlan (1979) ;Quinlan (1993)) or adaptations to specific topics such as quantile regression (Meinhausen, 2006), survival analysis (Ishwaran et al., 2008), ranking analysis (Clémençon et al., 2013), clustering (Yan et al., 2013) or online data (Denil et al., 2013).
The rest of this paper is organized as follows. Section 2 presents an example of a regression tree built with CART, while section 3 proposes an example of a Random Forest classifier dedicated to a prediction issue. Each of these sections includes a description of how these two methods address the modeling barriers discussed in the first section. Section 4 proposes a discussion on the application fields in economics of the methods and the last section concludes this work.

[8] "CART" keyword was used in econpapers (http://econpapers.repec.org/) for this research.
[9] This count includes decision trees built with CART but not RF's use.
[10] For more details on theoretical aspects of these two approaches, see for CART (Breiman et al., 1984), Biau and Scornet (2016) for RF.

2 Modeling with CART

Identification of growth drivers

In order to underline the benefits of CART and Random Forest algorithms, we take the example of the Sala-I-Martin (1997b) work on the identification of growth determinants.[11] As previously stated in the introduction, this topic is an excellent illustration of model uncertainty because many variables can be considered as growth drivers. These variables are numerous and difficult to prioritize.

The aim of this analysis is to explain the cross-country differences in terms of production growth between 119 countries[12] with 62 explanatory variables. Data are in cross-section and the variable of interest is the average growth between 1960 and 1992.[13] To address this issue, Sala-I-Martin (1997b) studies the complete distribution of each $\hat{\beta}_j$ from many models based on different combinations of variables. He ranks the variables according to the sizes of the density functions intervals that do not include zero. This topic was also investigated by Fernández et al. (2001), who use "Bayesian model averaging" (BMA) methodology. This method also relies on the estimation of many models but in a Bayesian framework, the variables' importance is established on the base of the posterior probability of each model.

A solution to avoid the problem of model uncertainty is the construction of a regression tree with CART. The basic idea is to split the predictor's spaces into different subspaces and to approximate the response of the output variable by its empirical mean in each subspace. On the question of growth determinants, CART produces the tree plotted in figure 1. The interpretation of this estimated model is as follows. If a country has an openness index ("YRSOPEN") of lower than 0.433, its growth rate depends on the value of its life expectancy ("LIFE060"). If it is under 43.25 years, the growth rate is 0.318 and otherwise the growth rate is equal to 1.532. If a country has an openness index higher than 0.433, predicted growth depends on the fraction of Buddhists in the population.[14] If it is superior to 8.5% the growth rate is equal to 5.571 and it is 2.610 otherwise.

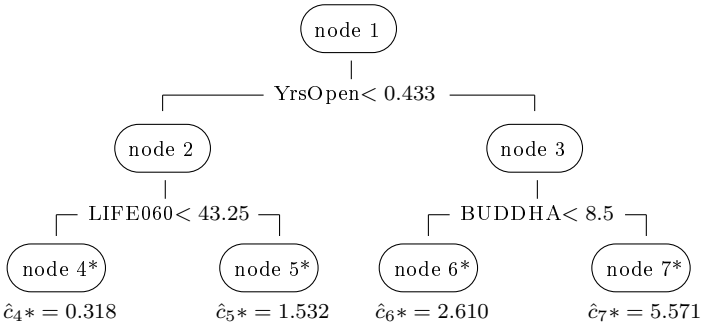
[11] We keep the same labels for the variables.

[12] Database contains more countries but there are missing values for the growth variable

[13] Data used are described in Table 3 in the appendix.

[14] The author has some reservations about the culturalist interpretation that can be made of this result. Note that cultural and religious variables are used here only for comparison with (Sala-I-Martin, 1997b) and (Fernández et al., 2001). These variables can also be interpreted in a very different way. The "BUDDHA" variable has a high value only for East Asian countries (Japan, South Korea, Taiwan, etc.), which have developed specific development strategies in line with the "flying goose model" described by Akamatsu (1962). The importance of this variable can be interpreted as the success of this strategy.

Figure 1 – regression tree (example 1)



The ease of interpretation is typical for trees built with CART. It doesn't require any statistical or mathematical skills and it is even familiar to economists who are used to tree representations (Varian, 2014). Initial space is the root node (node 1) and other spaces can be viewed as nodes (t) (or "leaves" or "regions") and we can define a part of a tree as a "branch" (T_t). It is also possible to have an analytical representation[15] of a tree which corresponds to a sum of constants, where $\hat{y}(x)$ is the predicted (fitted) value, I_{t*} a dummy variable taking 1 if the terminal node t_* is considered, $\hat{c}_{t*}$ is the predicted response in this node and N_t equal to the number of cases in node t:

$$\hat{y}(x) = \sum_{t*} I_{t*}\hat{c}_{t*} \tag{1}$$

$$I_{t*} = \begin{cases} 1, & \text{if } x \in t* \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \qquad \hat{c}_t = \bar{y}_t = \frac{1}{N_t}\sum_{x_n \in t} y_n$$

Estimation of regression tree with CART

Building a decision tree is a two-stage procedure in which the first is the succession of binary partitioning producing a tree structure (see figures 2 and 3). The initial space of predictors is split by maximizing the decrease of errors relative to an explanatory variable and a splitting point (s), which are considered the best splitting variable and point. Formally, we have:

$$\max_{j,s} \triangle R(s,t) = R(t) - R(t_l) - R(t_r) \tag{2}$$

$\triangle R(s,t)$ is the decrease of errors at node t, t_l and t_r indexes child nodes. In a regression case, the error of a node is defined as $R(t) = \frac{1}{N_t}\sum_{x_n \in t}(y_n - \hat{c}_t)^2$. The two child nodes are also partitioned with the same procedure and so on until there is a very large tree (T_{max}), which must be pruned.[16] Indeed, the risks include arbitrarily setting the size of the tree, constructing too simple tree or creating one that affords a very accurate (or even perfect) adjustment but at the cost of overfitting. A satisfying model must be complex enough to identify the structures in sample but it must have

[15] Mathematically, this is a "simple function".

[16] It is possible to set a hyperparameter controlling the number of observations in the terminal nodes before estimation in order to save computing resources.

a scope beyond a specific dataset. To prevent this risk of overfitting, CART includes a pruning procedure that penalizes (a posteriori) additional partitions.[17] Precisely, the idea is to minimize the cost-complexity criterion ($C_\alpha(T)$), which is the sum of errors in the terminal nodes and the number of splits carried. This criterion takes the following form:

$$C_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{3}$$

$R(T) = \sum_{t\in\tilde{T}} N_t R(t)$ corresponds to errors of T,$\tilde{T}$ is the set of terminal nodes of T, $|\tilde{T}|$ is the number of final leaves (nodes) and α is the cost of an additional split.[18]

Figure 2 – split 1

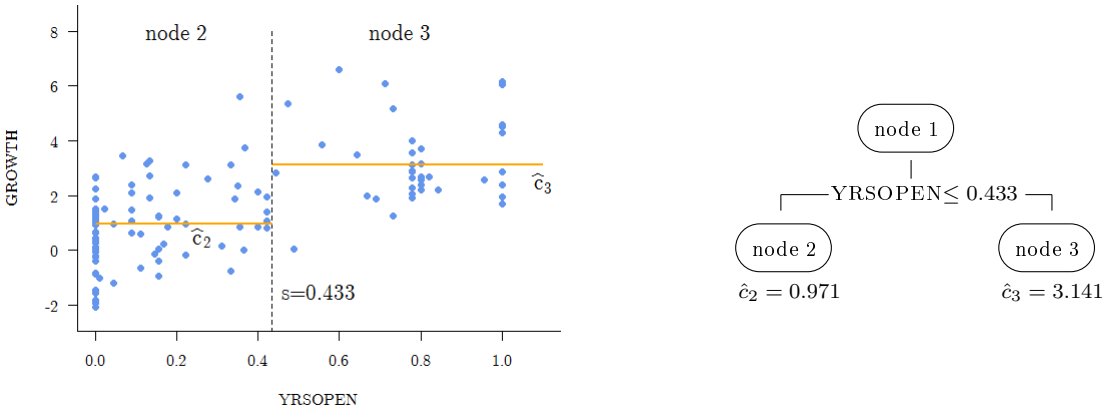
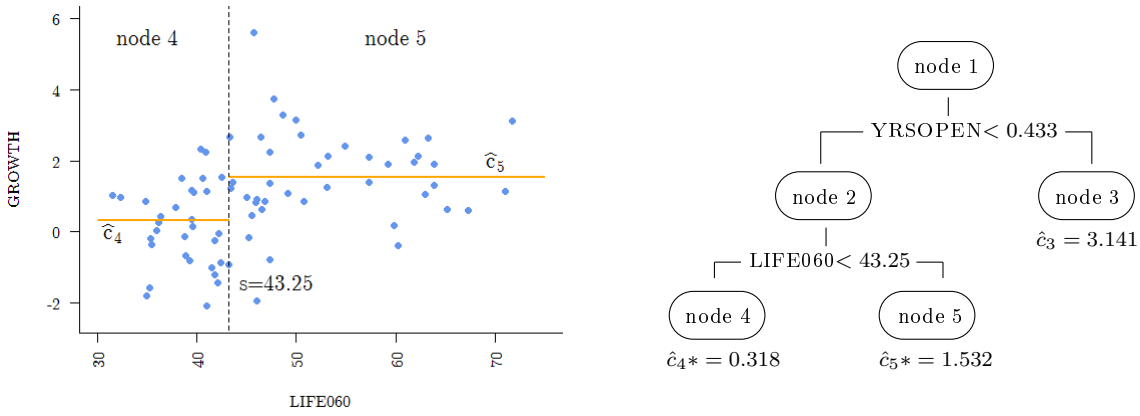


Figure 3 – split 2



[17] There are other pruning methods not presented in this paper.

[18] The procedure to identify the value of α is a cross-validation. For more details, readers may refer to Breiman et al. (1984) or Hastie et al. (2009).

Overcoming modeling problems

On the basis of the previous comments, several remarks can be made. The first is that the problem of the choice of the predictors (obstacle (i)) is solved by using CART. Indeed, the algorithm considers at each split all explanatory variables as a potential splitting variable. In the example, each partition was realized with the variable that minimizes the sum of quadratic errors. This explains why only three predictors are present in the final tree, while we initially considered 62 variables.

This automatic selection gives a first indication about the importance of each variable and the hierarchy between them. However, the need to select just one of them at each split could mask important predictors which are slightly less relevant than the selected variable. To overcome these possible "masked effects", Breiman et al. (1984) define a variable importance index including all variables, even those that were not selected in a final tree. Variable importance for X_j has the following form:

$$\text{Importance}\,(X_j) = \sum_{t \in T} \triangle R\left(\tilde{s}_t^j, t\right) \tag{4}$$

$\tilde{s}_t^j$ corresponds to the value of the splitting point of "surrogate" (substitute) split closest to the primary split. For each node t, algorithm searches the same partition with others variables than the splitting variable and $\triangle R(\tilde{s}_t^j, t)$ is the decrease of errors of each j variable.

The ranking presented[19] in Table 1 sheds light on the clear hierarchy between all variables because we observe the exclusion of 46 of them. Of the 16 remaining variables, the importance values are also clearly differentiated. For instance, one can say that the "YRSOPEN" is the most important predictor or that life expectancy ("LIFEE060") is four times more important than urbanization ("URB60"). Moreover, our application shows that CART is able to take into account the probable masked effects because the third variable in terms of importance is not in the final tree, which suggests that it had been masked in the splitting process by the selected variables.[20]

[19] Values are normalized in order to have a sum of importance equal to 100. For this example, the sum is equal to 99 due to rounding.

[20] The details of the execution of the algorithm confirm this point.

Table 1 – Variables' importance (example 1)

	CART		(Sala-I-Martin, 1997b)		(Fernández et al., 2001)	
Variable	Importance	Rank	Variable	Rank	Variable	Rank
YRSOPEN	19	1	EQINV	1	GDPSH60	1
LIFEE060	12	2	YRSOPEN	1	CONFUC	2
DPOP6090	10	3	CONFUC	1	LIFEE060	3
CIVLIBB	9	4	RULELAW	1	EQINV	4
ABSLATIT	9	4	MUSLIM	1	SAFRICA	5
PRIGHTSB	9	4	PRIGHTSB	6	MUSLIM	6
BUDDHA	7	7	LAAM	6	RULELAW	7
CONFUC	5	8	SAFRICA	8	YRSOPEN	8
URB60	4	9	CIVLIBB	8	ECORG	9
P60	3	10	REVCOUP	10	PROT	10
SAFRICA	3	10	MINING	11	MINING	11
S60	3	10	BMP1	12	NONEQINV	12
H60	3	10	PRIEXP70	13	LAAM	13
PI6089	1	14	ECORG	14	P60	14
FRAC	1	14	WARDUM	15	BUDDHIST	15
HINDU	1	14	NONEQINV	16	BMP1	16

Note 1: The underlined variables are those selected in the final tree.

Note 2: The variables having importance equal to 0 are not reported.

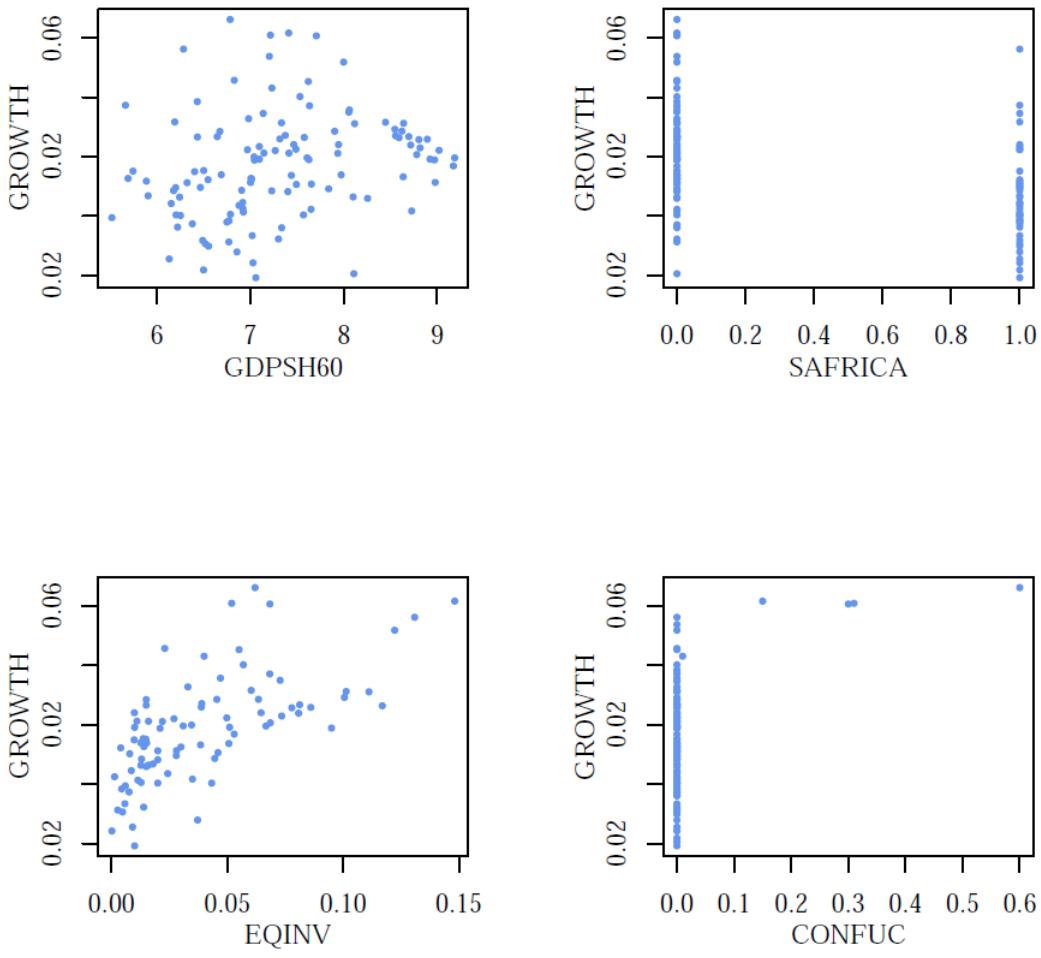
In view of the differences in terms of ranking (see Table 1), one can ask the question of how the results obtained with CART are not similar to those of Sala-I-Martin (1997b) and Fernández et al. (2001). A possibility is that the functional form used in these analyses is not adapted to the relationships considered (obstacle (ii)). Indeed, in these works, authors assume without statistical justification that the identification of the growth determinants can be carried in a linear framework. Fernández et al. (2001) simply argue that "Following the analyses in Levine and Renelt (1992) and Sala-I-Martin (1997b) tradition in the growth regression literature, we will consider linear regression models". Thus, the presence of nonlinear relationships is clearly a blind spot for these approaches because they cannot, by construction, grasp them. The problem is that a close look at the variables considered as the more important in these previous studies for explaining the differences in terms of average growth do not confirm without doubt the presence of linear patterns (see Figure 4). For instance, if we consider some of the most important variables according to Fernández et al. (2001)[21], the scatter plots do not display linear relationships. The joint distribution of "GDPSH60" (GDP per capita in logarithm in 1960) which is view as an "obvious variable" by Sala-I-Martin (1997b) and the growth is almost circular. The possibility of identifying a linear pattern in this case is only based on few observations, that is confirmed by the value of the correlation coefficient ($\hat{r} = 0.24$). Graphical inspection of the couples "GROWTH"/"CONFUC" and "GROWTH"/"SAFRICA" completely excludes a linear specification while the relationship between "GROWTH" and the variable "EQINV" seems to be slightly instable (approximately logarithmic). This quick evaluation is confirmed by the application of linearity tests which reject the presence of linear patterns.[22]

[21] These variables are also important according to the Sala-I-Martin (1997b) ranking.

[22] The results are not reported.

At least it is possible to recognize that there is uncertainty about the functional form adopted in these works that weakens the identification of key variables for explain average growth. Inversely, CART is not subject to this sort of problem because it is a nonparametric method which allows taking into account a large scope of nonlinearities. This algorithm realizes without a priori automatic detection structures by searching the most suitable functional form for a given dataset. Through multiple combinations, the successive splits can model very complex nonlinear relationships. In addition, it should be noted that CART can reach the conclusion that there is no link between variables. Indeed, the splits must be more informative than costly to be included in the final tree because the cost-complexity criterion must be minimized. Therefore, it is possible that for a given analysis, any split can compensate for the complexity of the model. This feature is very interesting because it gives the possibility of avoiding the selection of spurious links between the variables. This distinguishes CART from other nonparametric approaches, such as local regressions (LOESS, Cleveland (1979); LOWESS, Cleveland and Devlin (1988), which necessarily propose an estimated model.

Figure 4 – Scatter plots



This very flexible approach that combines a nonparametric view and automatic variable selection also has the advantage of simplicity in comparison with other methods such as BMA, which relies on Monte-Carlo Markov Chain (MCMC) to select relevant predictors.[23] Furthermore, estimation with CART doesn't need to specify a particular probability distribution. Indeed, from a theoretical point of view, the variable of interest Y and the predictors X_j are considered as random variables whose joint distribution are unknown, while their distribution is precisely defined in standard econometric approaches. Breiman (2001b) states that "The one assumption made in the theory is that the data is drawn *i.i.d.* from an unknown multivariate distribution." This fundamental difference means that some statistical tools are not defined for the case of tree-based method. Thus, there is no likelihood function or parameters to estimate. The ambition, common to algorithmic methods, is just to build, on the base of a learning sample $L(Y, X)$, a function of explanatory variables able to correctly approximate Y without searching to identify the generating process behind data.

On the other hand, the example of the growth drivers gives the opportunity to stress that CART considers all variables as potential interaction variables because each split is included in a sequence of successive partitioning. At each node, CART evaluates the ability of all variables to be the splitting variables and it tests all possible interactions for building a tree. Finally, the effect of variables will depend, in most of cases, on the values taken by other predictors, except if a terminal node is directly linked with the root node.
In our first application, the observation that the influence of "YRSOPEN" depends on the values of "LIFEE060" and "BUDDHA" is done by eliminating the less relevant other potential interaction variables. This constitutes a clear advantage compared to approach using interaction variables. Indeed, this latter strategy could lead to a problem in terms of degrees of freedom because the quantity of interaction variables hugely increases with the number of predictors[24] (number of interaction variables $= J!(2!(J-2)!))^{-1}$). Inversely, CART can produce a very precise model containing many interactions without any technical difficulty. On this basis, we can argue that this algorithm is able to completely solve the obstacle (iv) described in section 1.

3 From trees to forests

Classification and possible trees instability
Numerous contributions have been made to improve decision trees, the main one being Random Forest elaboration. This algorithm is close to the propositions of Ho (1998), Amit and Geman (1997), Dietterich (2000), and was finally defined by Breiman (2001a). The formulation of RF addresses several concerns, namely the enhancement of prediction accuracy and the solution of the problem of tree instability. Indeed, in some cases, trees built with CART (or other algorithms) can be

[23] Other variables selection methods are employed in economics such as "Bayesian Averaging of Classical Estimates" (BACE) (Sala-i Martin et al., 2004) or "GEneral TO Specific" (GETS) (Hendry and Krolzig, 2004). They also have the drawback of only supposing linear relationships.
[24] To consider all possible interactions, it should use 1891 interaction variables.

affected by small modifications of learning sample, which weakens their ability to predict and to be interpreted (Breiman, 1996).
To present RF and underline its qualities, we mobilize another example inspired by the analysis conducted by Osborn et al. (2005) on growth cycles prediction. Recall that "classical" cycles correspond to the alternation of periods of expansion and recession, while growth cycles involve the succession of accelerating and slowing production. We pay attention to this topic for the French economy over the period 1978 to 2014 (for more details on data see Table 4 in the Appendix). This example differs from the previous one because the variable "CYCLE" takes only two possible values (this is a binary variable). We consider the following coding: If growth is higher than growth observed in the previous month, "CYCLE" is equal to 1. Inversely, if growth is lower than growth observed in the previous month, "CYCLE" is equal to 0. For the identification of cycles, we use the chronology established by the Economic Cycle Research Institute[25] (ECRI).

As for the first example, the obstacle (i) is obvious because Osborn et al. (2005) identifies about ten variables able to influence growth cycle and consider, in addition, some lagged values of explanatory variables. This type of configuration can constitute a blind spot for standard econometric tools due to the number of predictors. For instance, in such cases, it is difficult to use a method such as logistic regression because the ratio of events to the number of observations is insufficient.[26] Osborn et al. (2005) use a variable selection algorithm ("n-search algorithm") for choosing the best model among all combinations of variables but with a maximum number of variables of 9. This constraint allows them to obtain interesting estimations, but at the expense of a loss of information due to the exclusion of predictors. A classification tree[27] built with CART is able to overcome this problem by constructing a model that predicts growth cycle by considering all predictors. The use of this algorithm shows its important flexibility because it can work with continuous or categorical as variables of interest. It should also be noted that explanatory variables can be of these two types. CART structure is the same as in the first example except that each split is not carried according to the decrease of quadratic errors. For classification purposes, there are three "impurity" criteria $i(t)$ that can be used at each node t: error classification rate $(1-max_k\left(p_{tk}\right))$, Gini index $(\sum_{k=1}^{K} p_{tk}\left(1-p_{tk}\right))$, cross entropy $(-\sum_{k=1}^{K} p_{tk}\ln\left(p_{tk}\right))$. Each node t is split by maximizing the decrease of impurity:

$$\max_{j,s} \triangle i(s,t) = i(t) - p_l i(t_l) - p_r i(t_r) \tag{5}$$

p_{tk} corresponds to the share of observations belonging to the class k in the node t while p_l and p_r are the share of cases falling in child nodes l and r. The predicted response in each terminal node depends on the most represented class. For example, for a two class problem, if a terminal node contains ten observations and that seven

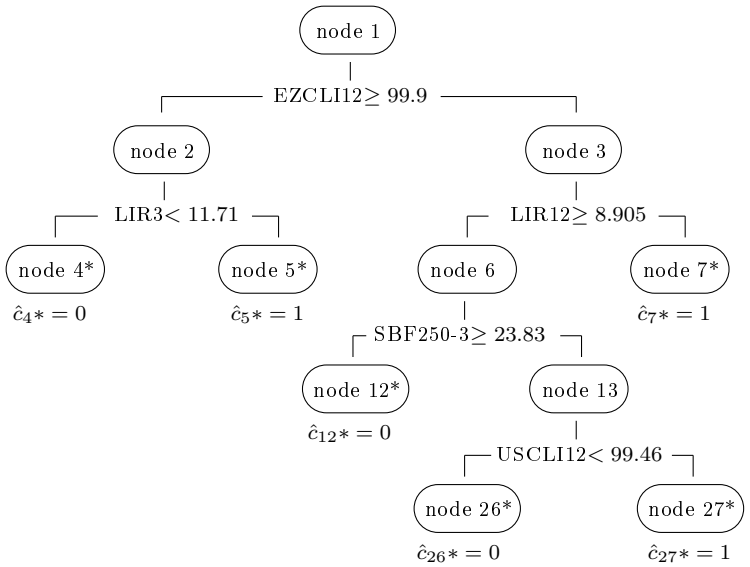
[25] https://www.businesscycle.com/

[26] Logistic regression is a very accurate classifier but Peduzzi et al. (1996) have shown that this method needs a learning sample containing at least 10 Events per Variable (EPV). EPV corresponds to the case where the variable of interest is equal to 1 (in binary classification). In the example used, EPV is equal to 97, making it possible to use 10 explanatory variables at most.

[27] This is a classification tree because the output variable is binary.

of them belong class 1, the model will predict class 1 with a probability of 70%. Model built with CART[28] is summarized in Figure 5. As in the first example, the interpretation of the tree is simple. For instance, if at a given month, "EZCLI12" is higher than 99.9 and "LIR3" is inferior to 11.71, one can predict that growth would be lower than in the previous month ($\hat{\text{CYCLE}} = 0$).

Figure 5 – Classification tree (example 2)



Aggregating trees as a solution

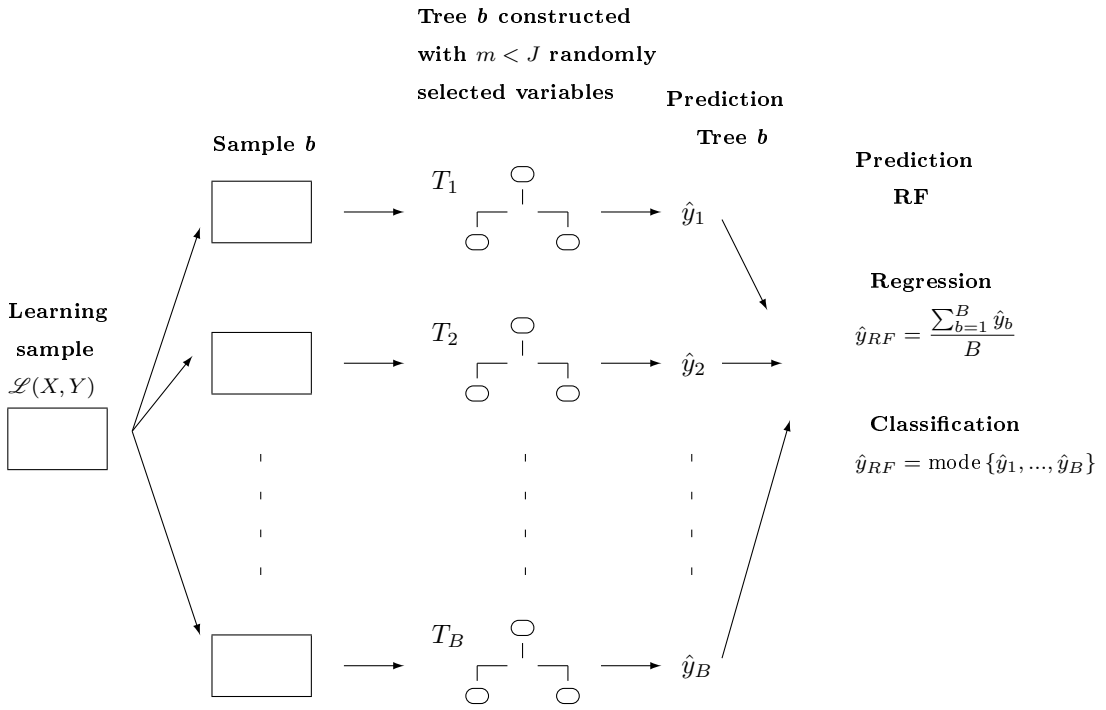
As noted before, a possible drawback of this kind of model is its potential instability because a perturbation of the learning sample could affect the tree structure. In these cases, prediction and interpretation are therefore questioned, but it should be noted that all trees are not sensible to dataset modifications. The instability problem can be solved and more accurate predictions can be obtained by using Random Forest. The central idea of this CART improvement is to use a collection of trees from B bootstrap samples created on the base of the original learning sample. Note that each bootstrap sample is created with only a fraction of the cases contained in the learning set and one can speak of "Out Of Bag"(OBB) observations to designate the data not used to generate each sample. The predicted values of each tree are aggregated to obtain the final prediction of the forest ("bagging", Breiman (1996)). A key point is that the trees are not correlated because each of them is estimated by randomly selecting only a portion of predictors.[29] It is possible to demonstrate that this way involves a significant reduction of the variance of the estimation.

[28]For this example, we use the Gini index as impurity criterion.

[29]As suggested by the presentation of the algorithm, RF doesn't need a pruning procedure contrary to CART. For classification purpose, the default value of m is fixed at $\lfloor\sqrt{p}\rfloor$ while it is equal to $\lfloor\frac{p}{3}\rfloor$ or regression task. In practice, it is possible that another value of in the proximity of $\lfloor\sqrt{p}\rfloor$ or $\lfloor\frac{p}{3}\rfloor$ gives better results. In our application, the default value is $\lfloor\sqrt{17}\rfloor = 4$ but we set $m = 3$ for this reason. This choice is consistent with the first suggestion of Breiman who proposed to use the following formula: $m = \lfloor\log_2(p+1)\rfloor$.

The aggregation is different according to the type of tree used. In the case of the regression problem, the predicted values are the average responses of all trees and they correspond to the majority of votes (mode) for classification task.[30] In this case, the total number of votes over the number of trees is interpreted as the probability of the event studied. For instance, for a given observation, if three quarters of trees predict the value 0 and $P(0) = 0.75$ (see Figure 7 to see this distinction between prediction and probability). RF estimation is summarized in Figure 6.

Figure 6 – RF Algorithm



The use of a large number of trees removes the possibility of having a simple representation of an estimated model (it is just possible to plot the trees one by one). However this loss is compensated by accuracy gains and stability of the model. Indeed, beyond a certain number of trees used, errors committed by a forest reach an asymptotic limit that avoids the risk of overfitting. Adding new trees in the forest does not improve the quality of the estimation Breiman (2001a).

[30] It should be noted that confidence intervals can be computed by using "Infinitesimal Jackknife" (Wager et al., 2014).

Figure 7 – Out-of-sample predictions

Note 3: This figure shows the predictions of the RF model on out-of-sample data. The learning sample includes monthly observations over the period December 1978 - December 2002 ($n = 289$) and the data for prediction are available between January 2003 and December 2014 ($n = 144$). Blue areas represent different cycles. Above 0.5, these periods within growth rates are higher than rates observed in the previous month (CYCLE = 0) and under this threshold there are growth rates lower than rates observed in the previous month (CYCLE = 1).

RF algorithm shares with its predecessor CART many advantages such as its non-parametric nature. Any assumption is needed in order to produce a function of explanatory variables able to fit the values of the interest variable. A very large scope of functional forms can be considered without specifying any particular probability distribution that avoids obstacles (ii) and (iii). In the case of regression, a gain in precision is obtained in comparison with CART because the latter proposes vertical partitions and constant fitted values in each terminal node. Inversely, RF is based on an aggregation of trees which produces smoothed and individualized predicted values. There is a similar situation in classification because some strong constraints that are common to standard methods are relaxed. Indeed, the forms of link function of probit and logistic models force the predicted values in contrast to RF. For instance, a logistic regression will tend to favor values near to 0 or 1, which could be desirable but could also produce an underestimation or overestimation regarding values taken by explanatory variables. Moreover, it should be stressed that while probit and logistic models have nonlinear link functions, they produce linear separations for classifying statistical individuals. On the contrary, CART and RF afford nonlinear separations which allow them to achieve more accurate classification.

Number, selection and importance of variables

RF as CART has the ability to consider a very large number of predictors, even going so far as to have $J > N$ while having very efficient predictions. For example, Breiman (2002) presents a case with a learning sample containing 81 observations, a number of levels of the output variable of 3 and the number of explanatory variables equal to 4,682. This configuration, which is impossible to study with standard econometrics tools, is modeled with RF, which obtains an error rate of only 1.2%. Beyond this example, comparisons on datasets with very large dimensions (between 701 and 685,569 predictors) showed that RF is on average more accurate that other algorithms. Its relative performance even improves as the number of dimensions increases (Caruana et al., 2008).

The consideration of many variables is different (relative to CART) when RF is used. For each tree and at each node, the algorithm selects the most relevant splitting variable and splitting point in order to maximize the decrease of errors (or impurity) of the child nodes. As said previously, the aggregation of the predicted values by tree provides the predicted value by the forest. With CART, non-selected variables do not influence the predicted response, while RF uses almost all variables to produce predictions. Indeed, RF considers all trees that are constructed with different variables such that the probability that each variable would be selected at least one time in one tree is high. Thus, one can say that an RF model gives prediction on the basis of all explanatory variables.[31]

To improve understanding of the predictors' hierarchy, RF produces a ranking by importance but differently to CART. In fact, there are two important measures. The first is the Mean Decrease Accuracy (MDA) (Breiman, 2001a) that is obtained for a j variable by computing the difference between the error rate on OOB sample (e_{OBB}) and the error rate based on this OBB sample but with j values permuted with the j values of another OBB sample ($e_{OBBj'}$).

$$\text{MDA}(X_j) = \frac{1}{B}\sum_{1}^{B}(e_{OBB} - e_{OBBj'}) \tag{6}$$

The logic behind this calculus is that if variable is not important, the permutation of its values should not affect strongly the quality of the estimation.[32] The second measure is the Mean Decrease Impurity (MDI) (Breiman, 2002) which, for a given variable X_j, is the average value of the decrease in errors (or in impurity) on all nodes of all trees where X_j is used (j_t* is the index of variable used for splitting at node t.). Impurity is defined as Gini index for classification problems and by sum of squared errors for regression.

[31] A variable not used corresponds to a variable not selected at any node of any tree.

[32] MDA could be normalized by dividing by the standard deviation of differences.

$$\text{MDI}(X_j)_{\text{classification}} = \frac{1}{B}\sum_{1}^{B}\sum_{t\in T_b} I(j_t* = j)\left(\frac{N_t}{N}\triangle i(s,t)\right) \tag{7}$$

or

$$\text{MDI}(X_j)_{\text{regression}} = \frac{1}{B}\sum_{1}^{B}\sum_{t\in T_b} I(j_t* = j)\left(\frac{N_t}{N}\triangle R(s,t)\right) \tag{8}$$

In the example of growth determinants, computing variables' importance gives the following ranking:

Table 2 – Variables' importance (example 2)

Variables	MDA	MDA Rank	MDI	MDI Rank	Average rank
LIR12	17.893	1	10.320	1	1
OPTG3	16.234	3	9.924	2	2.5
OECDCLI3	17.091	2	9.155	4	3
LIR3	15.851	4	7.411	6	5
SIR12	15.828	5	8.387	5	5
EZCLI12	15.497	7	9.193	3	5
SIR3	15.548	6	7.018	7	6.5
USCLI3	14.538	8	6.454	9	8.5
OECDECLI3	14.243	9	6.496	8	8.5
SBF250-12	13.733	10	5.076	13	11.5
EZCLI3	13.526	11	5.897	12	11.5
USCLI12	13.393	12	5.963	11	11.5
OPTG12	13.327	13	6.108	10	11.5
RM1-12	7.625	14	2.693	14	14
EXR	5.193	15	2.377	15	15
SBF250-3	5.038	16	2.032	16	16
RM1-3	-0.272	17	0.517	17	17

These results mainly inform that the four variables situated in the last positions have a very low importance in contrast to other predictors. We remark that there is a gradual decline of variables' importance until the last four which are far from the others. One of these variables ("RM1-3") even has a negative value, demonstrating that its inclusion in the model reduces the accuracy of the adjustment. On the other hand, given that RF components are decision trees, this method can take into account interactions between explanatory variables. These interactions are even considered at a very fine level because the search of relevant interactions is done with multiple samples and with a limited number of predictors at each step. Thus, for each tree in the forest, the risk of a masked interaction is reduced due to the low probability of having two near variables in competition in the same tree. However, it is not possible to interpret these interactions as in simple tree because they are numerous and not specially attached to one tree.

4 Application fields for economic topics

On the basis of our previous analysis, it is possible to argue that CART and Random Forest algorithms are able to overcome the four obstacles described in the introduction one by one or if they are present jointly. These methods are therefore very useful in order to explore many economic and financial topics. In the next section, we discuss the application fields of these approaches by identifying the type of issues most suited for their use.

4.1 Generality and flexibility

Access to blind spots and complex relationships

The ability of tree-based models gives them a feature of generality. Indeed, they are able to cover a large scope of functional forms (even involving highly nonlinear patterns) while neglecting the definition of a specific probability distribution. They can also take into account a large number of continuous or categorical predictors and interactions. As presented before, these characteristics provide solutions for econometric modeling but also suppose a more profound methodological approach. As outlined in section 1, economy can be viewed as one or several "complex system(s)" in the technical sense of the term Arthur (1999). This implies that many agents are in relation and react regarding the behavior of other agents, and that the economic variables change by retroactions and nonlinear relationships. All economy and some parts of it can be considered as a complex system (markets, industries, firms...). Arthur (1999) argues that because these objects are difficult to analyze, "conventional economic theory" simplifies the issues in order to make possible an analytical approach. It seems that many works using econometric tools proceed in the same way by assuming particular analytical forms in order to make applicable estimation and models available. It appears that decision trees built with CART, and especially Random Forest, can deviate from these technical constraints and thus better account for the "complexity" of economic phenomena in contrast to more conventional methods. A good example is the ability to include a very large number of interactions with strong nonlinearities that happen to be the counterpart of the many complex relationships between economic variables. RF is also able to take into account complexities in the non-theoretical sense of the term. Goldstein et al. (2010) underline that the methods from a machine-learning framework (including CART and RF) are very efficient for this kind of problem: "This means these algorithms may be more suited for identifying variants where the causal mechanism is unknown and complex".

Data mining

As stated in section 1 and shown in the two examples, there is often uncertainty on several structural characteristics of economic issues. This could come from diversity, incompatibility or even non-existence of theoretical frameworks. It also can be justified by the contradictory results of previous studies or a strong difficulty to specify the economic question. The increasing data availability ("big data") is also susceptible to reinforcing this problem.

A response in this context could be an exploratory analysis ("data mining") with

CART or RF in order to identify the main characteristics contained in data. On this issue, these algorithms are mostly presented as data mining methods in spite of the fact that they are able to do other kind of task. In order to stress this ability, we can recall that these methods can identify the most relevant predictors for regression and classification problems and to hierarchize them. Identification of the key variables, which is a very important question in economic research, can even be at the core of a statistical analysis with tree-based models instead of prediction (Goldstein et al., 2010; Verikas et al., 2011).
Tree-based models also give the possibility of grasping interactions and the presence of relatively homogeneous subgroups of statistical individuals. This latter capacity is very interesting for economists because data in cross-section and panel forms are very common in economics. This exploratory capacity is therefore particularly suitable for survey data (on employment, firms, etc.) that uses a large number of individuals and statistical variables. In the first example on the growth drivers, the regression tree identifies four subgroups depending on the values of the splitting variables.

Prediction
CART and RF are naturally oriented toward predictions and are very efficient in this kind of task. Their procedures of estimation and model validation are always based on the ability to predict by using new values not contained in the learning sample. For instance, building a tree with CART involves a step of cross-validation or the use of a test set for choosing the α value, which penalizes additional splits and identifies the final tree. The quality of a model, the variables' importance and the proximity matrix in RF framework are evaluated with OBB cases.
On a practical level, these methods are often used for prediction tasks for which very accurate performances are obtained (see examples in Breiman (2001b) or Caruana et al. (2008)). Generally, forests are superior to trees and are positioned very favorably compared to other approaches. To stress this ability, it should be noted that Fernández-Delgado et al. (2014) have evaluated the predictions of 179 classification approaches from 17 algorithm families on 121 datasets and that three of the five best models are RF models. These very convincing performances are added to other interesting characteristics evocated in this paper and are adapted to many economic issues. Examples are the estimation of default probabilities of banks, states, household or the prediction of key variables from a policy point of view such as inflation, unemployment or balance of trade.

4.2 Interpretation of tree-based models

The accuracy of in and out-sample predictions of an econometric model is an important quality. However, as said in section 1, economists are also interested in identifying and quantifying the effects of various explanatory variables on output variables. For this reason, they mostly resort to parametric models able to provide the sign and the magnitude of a given effect (by giving a marginal effect). For instance, in a log-log model, the slopes associated with predictors can be interpreted as elasticities and the coefficients obtained after a logistic regression can be viewed

as odd ratios. These statistical results give the possibility for formulating economic interpretation and policy recommendations.
A possible criticism of the use of tree-based model in economics is related to their supposed limit in terms of interpretation. Indeed, these algorithms that come from the machine-learning field can be view as "black box" (Breiman, 2001b) because they propose pertinent predictions but they may appear difficult to interpret. Their complexity can be linked with the large number of used variables, the large number of interactions (and trees for RF) or the absence of parameters and simple analytical form could reduce the attractiveness for economic works. For example, it may seem difficult to understand a forest based on hundreds of trees and characterized by even more interactions.

Firstly, it should be noted that a systematic use of tree-based models does not seem recommendable. The question of the level of prior information on variables and their relationships is fundamental. In the case where the problem is complex and has an unknown form, it seems very interesting to rely on CART or RF in order to perform an automatic structure identification instead of quantifying an uncertain relationship. Inversely, if the statistical issue is well defined, it seems more efficient to use a parametric method. The adjustment to the data would be more accurate and the results would be interpreted in statistical and economic terms. These two approaches can be complementary because it is possible to use a parametric approach after a preliminary analysis with CART or RF.
Secondly, forests and trees are not perfectly hermetic black boxes which just produce predicted values without giving information on the data structure. The algorithms evoked in this work are both able to carry a variable selection to build a model and to rank all variables considered. Even if they differ in the methodology, these rankings provide information on the capacity of the variables to reduce the prediction errors. It is quite natural to see the existence of causal effects behind these rankings. For this reason, Archer and Kimes (2008) conclude that "the RF methodology is attractive for use in classification problems when the goals of the study are to produce an accurate classifier and to provide insight regarding the discriminative ability of individual predictor variables." Furthermore, it is possible to add the sign of the effect to importance measures by inspecting the execution of algorithms. For single trees, the detail of their structure is summarized in the associated graphical representations which clearly present the relationships and interactions identified by CART. Thus, "single decision trees are highly interpretable" Hastie et al. (2009)). For example, in the case of the tree estimated in section 2, it is obvious that the relationship between average growth and openness index ("YRSOPEN") is positive. Regarding Random Forest, the situation is different because it is not possible to summarize the model in a single tree. It is only possible for a given variable to be studied on the basis of all trees how it determines the variable of interest.

However, a more convenient approach for studying the impact of explanatory variables on output variable consists in relying on the "partial dependence function" (PDF) defined by Friedman (2001). This kind of object is particularly useful for tree-based models but it can be used for many other type of model. For this reason,

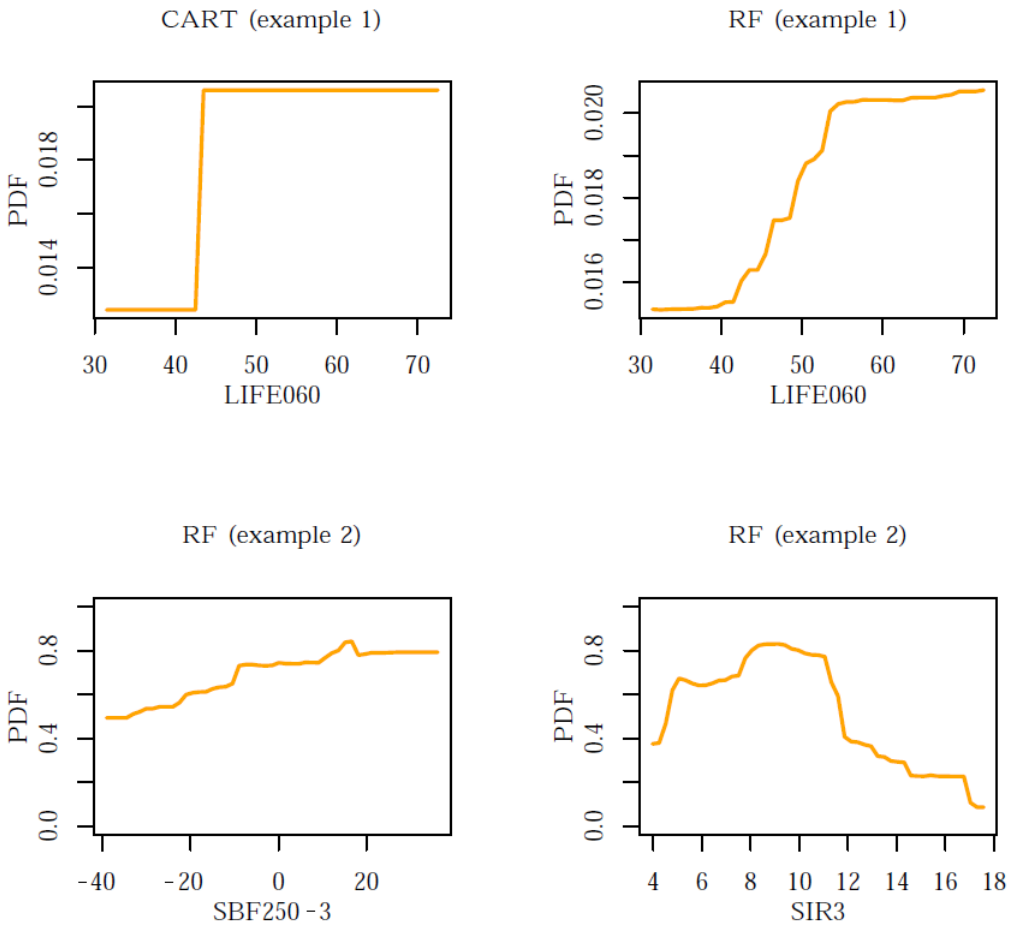
Hastie et al. (2009) stress that "partial dependence functions can be used to interpret the results of any 'black box' learning method." The general idea is to evaluate the mean predicted response of a model for each given value of one of several variables(s) X_S.[33] Formally, by considering a subset X_S of matrix X which contains predictors $S \subset 1, 2, \ldots, J$ and a complement set C. Partial dependence function can be estimated by using:

$$\hat{f}_S(X_S) = \frac{1}{N} \sum_{1}^{N} \hat{f}(X_S, X_{C_n}) \tag{9}$$

An important point is that the slope of $\hat{f}_S(X_S)$ may are not necessarily reflect marginal effects because it describes how the predicted response changes relative to X_S after taking into account the impact of the other variables. The slope of $\hat{f}(X_S)$ corresponds to marginal effects only when one variable in X_S is considered and that any interaction is present that is rare with tree-based models.
For instance, we plot four partial dependence functions from the applications of this work in Figure 8. The first plot on the left displays the partial dependence of "GROWTH" over life expectancy ("LIFE060"). It shows that for any values of "LIFE060", the predicted response is positive and that beyond a given threshold (43.25), the predicted value of growth is higher. This very simple conclusion reflects the simple structure of the underlying model, which has a limited size with only three explanatory variables and two interactions selected.
When a dependence function is estimated with a large tree or with a RF model, it is probable that very specific relationships will be obtained. If we take the same example but use an RF model (at the top right of the Figure 8), we observe that a growing and nonlinear relationship indicates that the increase or decrease of life expectancy has an effect concentrated in a specific interval (when "LIFE060" $\in (40, 55)$). The two other plots in Figure 6 represent two PDF from the second example. It is important to note that in a classification case, the mean response is computed by using the centered logit. These comments stress that the "black box" expression often associated with statistical models from the machine-learning framework does not seem adapted for tree-based models. Indeed, these approaches are able to identify relevant predictors, hierarchize them and to describe how the variable of interest evolves with values taken by explanatory variables.

[33] It is also possible to refine the analysis of partial dependence by decomposing the average response by studying the "Individual Conditional Expectation" (ICE) as proposed by Goldstein et al. (2014). It also gives the possibility to use median for aggregate individual responses.

Figure 8 – Partial Dependance Functions



5 Conclusion

The purpose of this work is to underline the benefits of the use of CART and RF algorithms for studying economic issues. After describing four typical problems occurring in econometric modeling, we have presented the two algorithms and shown how they can solve these obstacles through using two examples. The first is devoted to the identification of growth determinants and the second concerns the prediction of growth cycles. More generally, we have described the most suitable tasks for these approaches. It seems that CART and RF are particularly efficient for grasping complex patterns in data with a large number of variables, nonlinear relationships and interactions. They also appear adapted to data mining thanks to their ability to automatically detect structure and to make accurate predictions on the basis of a learning sample. Our analysis also stresses that these methods are able to produce models that can be interpreted, which is crucial from an economist's point of view. On this basis, it seems possible to argue that these tools are very useful for statistical works in economics in complement with other standard approaches. This claim naturally leads to considering the use in economic research of the numerous extensions of these ground algorithms and of other methods from the machine-learning framework, such as neural networks and Support Vector Machines (SVM).

Appendices

Table 3 – Data used in example 1

Variable Name	Variable definition	Countries
GROWTH	Average Growth 1960-1990	Algeria, Angola, Benin, Botswana,
GDPSH60	log(GDP per capita 1960)	Burkina Faso, Burundi, Cameroon,
LIFEE060	Life Expectancy in 1960	Cape verde, Cent'l Afr. Rep., Chad,
P60	Primary School Enrollment Rate in 1960	Comoros, Congo, Egypt, Ethiopia,
SAFRICA	Dummy Variable for Sub-Sahara African Countries	Gabon, Gambia, Ghana, Guinea-Bissau,
LAAM	Dummy Variable for Latin American Countries	Cote d'Ivoire, Kenya, Lesotho, Liberia,
BMP1	Black Market Premium	Madagascar, Malawi, Mali, Mauritania,
BMS6087	Standard Deviation Black Market Premium	Mauritius, Morocco, Mozambique, Niger,
GDC6089	Growth of Domestic Credit 1960-1990	Nigeria, Rwanda, Senegal, Seychelles,
STDC6089	Standard Deviation Domestic Credit	Sierra Leone, Somalia,South africa,
PI6089	Average Inflation Rate	Sudan, Swaziland, Tanzania, Togo,
STPI6089	Standard Deviation Inflation 1960-1990	Tunisia, Uganda, Zaire, Zambia
SCOUT	Outward Orientation	Zimbabwe, Barbados, Canada, Costa Rica,
AREA	Total Area of the Country	Dominican Rep., El Salvador, Guatemala,
FREEOP	Free Trade Openess	Haiti, Honduras, Jamaica, Mexico,
FREETAR	Tariff Restrictions	Nicaragua, Panama, Trinidad and Tobago,
DPOP6090	Average Growth Rate of Population 1960-1990	United States, Argentina, Bolivia, Brazil,
PYR60	Average Years of Primary School	Chile, Colombia, Ecuador, Guyana,
SYR60	Average Years of Secondary School	Paraguay, Peru, Suriname, Uruguay,
HYR60	Average Years of Higher School	Venezuela, Afghanistan, Bangladesh,
HUMAN60	Average Years of Schooling	Myanmar, Hong Kong, India, Indonesia,
S60	Secondary School Enrollment Rate in 1960	Iran, Iraq, Israel, Japan, Jordan
H60	Higher School Enrollment Rate in 1960	Korea, Malaysia, Nepal, Pakistan,
YRSOPEN	Number of Years Open Economy	Philippines, Singapore, Sri Lanka
GGCFD3	Public Investment Share	Syria, Taiwan, Thailand, Austria,
GVXDXE52	Public Consumption Share	Belgium, Cyprus, Denmark, Finland,
GEEREC1	Government Education Spending Share	France, Germany (West), Greece, Iceland,
GDE1	Defense Spending Share	Ireland, Italy, Luxembourg, Malta,
ASSASSP2	Political Assassinations	Netherlands, Norway, Portugal, Spain,
REVCOUP	Revolution and Coups	Sweden, Switzerland, Turkey,
PINSTAB2	Political instability	United Kingdom, Yugoslavia, Australia,
WARDUM	War dummy	Fiji, New Zealand, Papua New Guinea
PRIGHTSB	Political Rights	
CIVLIBB	Civil Liberties	
ABSLATIT	Absolute Lattitude	
FRAC	Ethnolinguistic Fractionalization	
DEMOC65	Index of Democraty 1965	
PRIEXP70	Primary Export in 1970	
RULELAW	Rule of Law	
URB60	Urbanization Rate	
RERD	Exchange Rate Distortions	
EQINV	Equipment Investment	
NONEQINV	Non-equipment Investment	
HUMANYL	Average Years of Schooling*log(GDP60)	
TOT1	Terms of Trade Growth	
WORK60L	Ratio Workers to Population	
LLY1	Liquid Liabilities to GDP	
BRIT	British Colony	
FRENCH	French Colony	
SPAIN	Spanish Colony	
BUDDHA	Fraction of Buddhist	
CATH	Fraction of Catholic	
CONFUC	Fraction of Confucius	
HINDU	Fraction of Hindus	
JEW	Fraction of Jewish	
MUSLIM	Fraction of Hindus	
PROT	Fraction of Protestant	
LFORCE60	Size Labor Force	
MINING	Fraction of GDP in Mining	
ECORG	Degree of Capitalism	
OTHFRAC	Fraction Population Speaking Foreign Language	
ENGFRAC	Fraction Population Speaking English	

Note 4: More details on data sources are in Sala-I-Martin (1996).

Table 4 – Data used in example 2

Variable	Definition	Source
CYCLE	Growth cycle If growth rate is higher than previous growth rate, CYCLE=1 otherwhise CYCLE=0	ECRI
OPTG3	Output gap lagged 3 months This serie is built by using HP filter on industrial production index	Datastream code: FROCIPRDG
OPTG12	Output gap lagged 12 months This serie is built by using HP filter on industrial production index	Datastream code: FROCIPRDG
RM1-3	M1 Money Supply ln difference on 3 months Original serie is divided by CPI	OECD
RM1-12	M1 Money Supply ln difference on 12 months Original serie is divided by CPI	OECD
SBF250-3	Stock market prices (SBF 250) ln difference on 3 months Original serie based on monthly mean	Datastream code: FRSHRPRCF
SBF250-12	Stock market prices (SBF 250) ln difference on 12 months Original serie based on monthly mean	Datastream code: FRSHRPRCF
SIR3	Short interest rate ln difference on 3 months Original serie based on monthly mean	Datastream code: FRINTER3
SIR12	Short interest rate ln difference on 12 months Original serie based on monthly mean	Datastream code: FRINTER4
LIR3	Long interest rate ln difference on 3 months Original serie based on monthly mean	Datastream code: FRGBOND
LIR12	Long interest rate ln difference on 12 months Original serie based on monthly mean	Datastream code: FRGBOND
EXR	Exchange rate USD to EURO ln difference on 1 month Original serie based on monthly mean	Datastream code: FRXRUSD
EZCLI3	Euro-Zone aggregate composite leading indicator lagged 3 months	OECD
EZCLI12	Euro-Zone aggregate composite leading indicator lagged 12 months	OECD
OECDCLI3	OCDE aggregate composite leading indicator lagged 3 months	OECD
OECDCLI12	OCDE aggregate composite leading indicator lagged 12 months	OECD
USCLI3	US composite leading indicator lagged 3 months	OECD
USCLI12	US composite leading indicator, lagged 12 months	OECD

References

P. Aghion and P. Howitt. A model of growth through creative destruction. *Econometrica*, 60(2):323,351, 1992. Econometric Society.

K. Akamatsu. A historical pattern of economic growth in developing countries. *Journal of Developing Economies*, 1(1):3, 25 1962.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545,1558, 1997.

K.J. Archer and RV. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, 52(4):2249,2260, 2008.

B.W. Arthur. Complexity and the economy. *Science*, 284:107,109, 1999.

R.J. Barro. Government spending in a simple model of endogenous growth. *Journal of Political Economy*, 98(5):103,125, october 1999.

G. Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(5):1063,1095, 2012.

G. Biau and E. Scornet. A random forest tour. *TEST An Official Journal of the Spanish Society of Statistics and Operations Research*, 25(2):197,1227, 2016. DOI 10.1007/s11749-016-0481-7.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307,327, 1986.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123,140, 1996.

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

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199,231, 2001b.

L. Breiman. *Manual on setting up, using, and understanding random forests v3.1.*, 2002. http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf.18,19.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman & Hall, 1984.

R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979. DOI 10.2307/2286407.

W.S. Cleveland and S.J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596,610, 1988. DOI 10.2307/2289282.

S. Clémençon, M. Depecker, and N. Vayatis. Ranking forests. *Journal of Machine Learning Research*, 14, 2013.

C. Cobb and P. Douglas. A theory of production. *American Economic Review*, 18 (1):139,165, 1928.

CMS Collaboration. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716:1,254, 2012. URL http://www.sciencedirect.com/science/article/pii/S0370269312008581.

R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223,236, 2001.

V. Coudert and V. Mignon. Reassessing the empirical relationship between the oil price and the dollar. *Energy Policy*, 95:147,157, 2016. http://dx.doi.org/10.1016/j.enpol.2016.05.002.

M. Denil, D. Matheson, and N. de Freitas. Consistency of online random forests. In *International Conference on Machine Learning (ICML)*, 2013.

T.G. Dietterich. Ensemble methods in machine learning. volume 1857. 2000. Lecture Notes in Computer Science.

L. Einav and J. Levin. The data revolution and economic analysis. *Innovation Policy and the Economy*, 346(6210):1,24, 2014a.

L. Einav and J. Levin. Economics in the age of big data. *Science*, 346(6210), 2014b. DOI: 10.1126/science.1243089.

R.F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of u.k. inflation. *Econometrica*, 50(4):987,1008, 1982. DOI: 10.2307/1912773.

C. Fernández, E. Ley, and M.F.J. Steel. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563,576, 2001.

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133,3181, 2014.

Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148,156. L. Saitta, ed, 1996.

J.H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189,1232, 2001.

R. Frisch. Editor's note. *Econometrica*, 1(1):1,2, 1933.

X. Gabaix. Power laws in economics: An introduction. *Journal of Economic Perspectives*, 30(1):185,206, 2016.

A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 2014.

B.A. Goldstein, A.E. Hubbard, A. Cutler, and L.F. Barcellos. An application of random forests to a genome-wide association dataset: Methodological considerations and new findings. *BMC Genetics*, 11(49), 2010.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, pages iii,115, 1944.

D.S. Hamermesh. Six decades of top economics publishing : Who and how ? *Journal of economic literature*, 51(1):162,172, 2013.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. Springer Verlag, second edition edition, 2009.

D. Heath, S. Kasif, and S. Salzberg. Induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2(2):1,32, 1993.

D.F. Hendry and H.M. Krolzig. We ran one regression. *Oxford Bulletin of Economics and Statistics*, 66(5), 2004.

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

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651,674, 2006.

H Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841,860, 2008.

C.M. Jarque and A.K. Bera. Efficient test for normality, homoscedasticity and serial independence of residuals. *Economic Letters*, 6(3):255,259, 1980.

R. Kümmel, R.U. Ayres, and D. Lindenberger. Thermodynamic laws, economic methods and the productive power of energy. *Journal of Non-Equilibrium Thermodynamics*, 35(2):145,179, 2010.

R. Levine and D. Renelt. A sensitivity analysis of cross-country growth regressions. *American Economic Review*, 82(4):942,963, 1992.

W.Y. Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329,348, 2014. doi:10.1111/insr.12016.

R. Lucas. On the mechanics of economic development. *Journal of Monetary Economics*, 22(1):3,42, 1988.

B. Mandelbrot. New methods in statistical economics. *Journal of Political Economy*, 71(5):421,440, 1963.

N. Meinhausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983,999, 2006.

J. Mincer. *Schooling, Experience and Earnings*. Columbia University Press, 1974. National Bureau of Economic Research.

J.M. Morgan and J.A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415,434, 1963.

OCDE. Prudent debt targets and fiscal framework. OECD economic policy paper, OCDE, 2015.

D. Osborn, M. Sensier, and D van Dijk. *Predicting growth regimes for European countries*. Centre for Ecomomic Policy Research, 2005.

P.N. Peduzzi, J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(1):1373–1379, 1996.

J.R. Quinlan. *Expert systems in the micro electronic age*, chapter Discovering rules by induction from large collections of examples. Edinburgh University Press, 1979.

J.R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, 1993.

F.P. Ramsey. A mathematical theory of saving. *The Journal of Economic History*, 38(52):543,559, 1928.

P.M. Romer. Increasing returns and long-run growth. *The Journal of Political Economy*, 94(5):1002,1037, 1986.

X. Sala-I-Martin. I just ran two million regressions. *NBER Working Paper*, (6252), 1996.

X. Sala-I-Martin. I just ran two million regressions. *The American Economic Review*, 87(2):178,183, 1997b.

X. Sala-i Martin, G. Doppelhofer, and R.I. Miller. Determinants of long-term growth: A bayesian averaging of classical estimates (bace). *The American Economic Review*, 94(4):813,835, 2004.

T. Schmith, S. Johansen, and P. Thejll. Statistical analysis of global surface temperature and sea level using cointegration methods. *Journal of climate*, 25(22), 2012.

J. Shotton, A. Fitzgibbon, A. Blake, A. Kipman, M. Finocchio, R. Moore, and T. Sharp. Real-time human pose recognition in parts from a single depth image. In *CVPR*. IEEE, 2011. URL https://www.microsoft.com/en-us/research/publication/real-time-human-pose-recognition-in-parts-from-a-single-depth-image/.

R.M. Solow. A contribution to the theory of economic growth. *Quarterly Journal of Economics*, 70(1):65,94, 1956.

T.W. Swan. Economic growth and capital accumulation. *Economic record*, 32(2): 334,361, 1956.

T. Teräsvirta, D. Tjøstheim, and C.W.J. Granger. *Modelling nonlinear economic time series*. Oxford University Press, 2010.

H. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 2014.

A. Verikas, A. Gelzinis, and M Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330,349, 2011. doi:10.1016/j.patcog.2010.08.011.

S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forest : The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1):1625,1651, 2014.

W. Wang. *Stochasticity, nonlinearity and forecasting of streamflow processes*. IOS Press, 2006.

X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1): 37, 2008.

D. Yan, A. Chen, and M.I. Jordan. Cluster forests. *Computational Statistics and Data Analysis*, 66:178,192, 2013.