Senoussi, Houcine

**Article**

# Inflation and inflation uncertainty in growth model of Barro: An application of Random Forest method

International Econometric Review (IER)

**Provided in Cooperation with:**
Econometric Research Association (ERA), Ankara

*Suggested Citation:* Senoussi, Houcine (2021) : Inflation and inflation uncertainty in growth model of Barro: An application of Random Forest method, International Econometric Review (IER), ISSN 1308-8815, Econometric Research Association (ERA), Ankara, Vol. 13, Iss. 1, pp. 4-23, https://doi.org/10.33818/ier.854697

This Version is available at:
https://hdl.handle.net/10419/290159

# Inflation and Inflation Uncertainty in Growth Model of Barro: An Application of Random Forest Method

**Houcine Senoussi®**

CY-Tech
CY Cergy Paris Université
Cergy
France

## ABSTRACT

One of the major problems of the empirical economists while building an economic model is the selection of variables which should be included in the true regression model. Conventional econometrics use several model selection criteria to determine the variables. Recent years' developments in Machine Learning (ML) approaches introduced an alternative way to select variables. In this paper, I have an application of ML to select variables to include for a nonlinear relationship between inflation and economic growth. Among ML methodologies, Random Forest (RF; Breiman, 2001) approach is one of the most powerful to capture nonlinear relationships. Therefore, I applied RF and found that both high and low inflation can be the cause of low economic growth which is a major contribution of the paper to economic literature. This observation produces clear suggestions for central bank inflation targeting policies. Moreover, in the paper, as an outcome of RF there are other variables effecting economic growth with an order of importance.

**Key words:** *Growth, Inflation, Machine Learning, Random Forest.*
JEL Classifications: C18, E31, E58, O49.

## 1. INTRODUCTION

Empirical economists before going on estimation, hypothesis testing, direction of effect or prediction, initially, face the problem of variable selection which should be included in the true regression model. Conventional econometrics use model selection criteria to determine the variables. Akaike Information Criterion (AIC; Akaike, 1974), a biased corrected version of AIC (AICC; Hurvich and Tsai, 1989), Schwarz Criterion (SC; Rissanen, 1978 and Schwarz, 1978) and Hannan-Quinn Criterion (HQC; Hannan and Quinn, 1979 and Quinn, 1980) are the most popular ones. Basci et al. (2010) suggested the usage of cross validation of variance estimation in these criteria rather than standard variance estimation, namely, predictive residuals sum of squares (PRESS), and showed that their performance improved especially for large samples with such a replacement.

In his seminal 1991 paper, Barro (1991) explained the growth rate of real per capita Gross Domestic Product (GDP) for 98 countries for the period 1960 - 1985 with a regression

® Computer Science department, CY Tech, CY Cergy Paris Université, France. Email : hse@cy-tech.fr. Tel : +33134258418..

analysis. Following this paper, many other researchers focused on the topic as well. In all of them, the following cross-sectional regression exists.

$$\gamma = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \quad (1.1)$$

Here, $\gamma$ is the vector of economic growth (depending on the choice of the researcher definition of economic growth can change) and $x_1$, $x_2$, ..., $x_n$ are vectors of explanatory variables. Different growth theories use different $x_j$, $j = \{1,2,...,n\}$ in the 'true' regression model. Decision of which variables to include in the true regression model is the main problem for the empirical economists. A variable $x_i$ can be significant in a regression where variables $x_j$ is included but whenever $x_k$ is included it may become insignificant. As a consequence of these different conclusions, contrary results emerge. Goldfarb (1995, 1997) documented this contrary results phenomenon.

Sala-i-Martin (1997) studied all these possible variables and after running two million regressions found out 62 variables which were significant at least in one of the regressions. Hendry and Krolzig (2004) criticizes Sala-i-Martin's work by stating that in fact there is no need to run two million regressions but it is enough to run just one regression and apply general-to-specific approach. Here the only regression is the general unrestricted model.

There are several interesting results in Sala-i-Martin's work but the one for inflation is the concern of this paper. Inflation was not among the significant variables in this work but economic literature states that there should be a relation between inflation and economic growth. Sala-i-Martin (1997) explains his conclusion by stating that the analysis he makes is a linear one but most probably, the relation between inflation and economic growth is not linear.

A Machine Learning (ML) methodology, Random Forest (RF; Breiman, 2001) captures nonlinear relationships. In this paper, we use this methodology, rather than using nonlinear regression, to analyze inflation and economic growth relation for the period 1961 - 2016 for 54 countries and 16 variables.

The paper is organized as follows. In Section 2, we present our contributions both from econometric and ML point of views. In Section 3, we describe RF methodology. In Section 4, we present related work. In Section 5, we describe the data we used in our analysis. In Section 6, we explain our setting and methodology. In Section 7, we comment on our results. In the last Section, we have the conclusion.

## 2. CONTRIBUTIONS

The contributions of this paper are:

- Econometrics point of view:
  - Prediction and explanation of the GDP growth rate: The set of rules we obtain by RF allows us to predict the value of GDP growth rate and to determine the main groups of explanatory variables.

  - Analysis of the effect of inflation: Empirical economics literature has the consensus that when inflation is high economic growth tends to be low and our rules obtained by RF are consistent with this. The solutions to this problem had

been studied extensively in economics literature. However, as shown in Basci et al. (2020), focus on the effects of low inflation on economic growth is lacking in economics literature although this became the recent decades' problem. Japan is a very good example. Our rules obtained by RF show that when inflation is low, it pulls GDP growth rate to lower values.

- Analysis of the effect of standard deviation of inflation: Our rules obtained by RF are similar to the rules that we obtained for inflation.

- Suggestion of inflation target rates for the central banks: according to our rules obtained by RF if inflation is in the range 5.84 and 25.07 (corresponding to the middle value in our analysis), then the expected GDP growth rate is above 0.04 (corresponding to the high value in our analysis). Therefore, depending on country specific environments, central banks of the countries should set their targets within this range to be able to achieve high growth rates.

- Machine learning point of view:
  - Extension of the use of ML methods in econometrics: In the past few years, we have seen the publication of several works that combine ML and empirical econometrics as a result of emerging big data. (see examples in the related work section).

  - Studying RF properties: In this paper we studied experimentally the sensitivity of the accuracy and stability of the method to its main parameters. We also examined the behavior of the two main variable importance measures.

## 3. RANDOM FOREST

RF is a supervised learning technique introduced by L. Breiman (2001) in the early 2000's and used for classification and regression. It aims to build a classifier consisting of a collection of decision trees grown on subsets of the original data. Each subset is defined by two random selections: a vertical selection (on variables) and a horizontal selection (on observations). The classifier's prediction is obtained by taking the majority vote of the trees in the case of classification and the average over their predictions in the case of regression. In this work, we have a classification problem in which all the variables are categorical. In this case, the idea of RF can be formalized as follows :

- Input:
  - The training set $D = \{O_i = (X_i,Y_i)\ i = 1,...,n\}$. The instances $X_i$ are described by $p$ features (categorical or binary variable) $F_1,..,F_p$. The class $Y$ is a categorical or binary variable which takes its value in a set $\{C_1,...,C_s\}$.
  - *ntree* : the number of the trees to build.
  - *mtry* : the number of features to try in each node of each tree.
- Output:
  - The forest RF: a set of *ntree* trees.
- Algorithm:

- Draw *ntree* bootstrap samples $D_k$ from $D$.
- For $k = 1,...,ntree$ use $D_k$ to build the decision tree $T_k$ by recursively repeating the following steps for each node until the stopping criterion is met :
  - Randomly select *mtry* features.
  - Pick the best feature according to the splitting criterion (see below) among the *mtry*.
  - Use this feature to split the node into two nodes.

- To make a prediction at a new instance $x$:
  - $RF(x)$ = majority vote $\{T_k(x)\}$

### 3.1. Splitting And Stopping Criteria

Let us first recall that during the building of a tree, each node $N$ represents a subset $D_N$ of $D$. Splitting $N$ means partitioning $D_N$ into two subsets $D_{NL}$ and $D_{NR}$, each one corresponding to some values of $F$, the feature we use in splitting. As we said above, $F$ is the ”best” feature with respect to splitting criterion. Since we aim to reach leaves of the tree, which are nodes corresponding to ”pure” subsets (subsets in which majority or totality of observations belong to the same class), we will define a ”good” feature as a feature that improves purity. In RF, purity is measured by Gini index.

Stopping criterion is simply defined by the minimum node size (i.e cardinality of the set associated with the node). This parameter will be noted *ndsize*.

### 3.2. Outputs

In addition to the classifier itself, RF algorithm has two outputs:

- *OOB* error : for each observation $O_i = (X_i,Y_i)$, let us aggregate the votes only over those trees $T_k$ whose bootstrap sample $D_k$ does not contain $O_i$. The classifier thus obtained is called the *out-of-bag* (*OOB*) classifier(Breiman, 2001). The error rate of this classifier on the training set is used to estimate the prediction error of the RF classifier, and is called the *out-of-bag* error.
- Variable importance: There are four ways to measure a variable importance, i.e. its relevance to the problem we are dealing with (Breiman 2002), but the two most commonly used ones are:
  - *Mean decrease accuracy* (MDA, Measure 1 of Breiman (2002)) also called *permutation importance*. When the tree $T_k$ is created, its prediction accuracy is estimated using its OOB sample. Then, the values for each feature $F$ are randomly permuted and the new prediction accuracy of $T_k$ is computed. The measure of importance of $F$ is obtained by averaging the decrease in accuracy due to these permutations.
  - *Mean decrease Gini* (MDG, Measure 4 of Breiman (2002)). At every inner node of each tree, a randomly selected variable is used to do the split. This split results in decreasing the Gini. The sum of all decreases due to a given variable, normalized by the number of trees, is the measure of importance of this variable.

  These measures complement each other (Breiman, 2002). The first measure is more intuitive. According to many authors (Behnamian et al., 2017 and Call and Urrea, 2011), the second is more stable.

### 3.3. Extracting Rules From RF

A set of If/Then rules can be extracted from each tree in the forest by traversing each path from root to leaves. The obtained rules have the form

$$(F_1 = v1) \wedge \ldots \wedge (F_m = vm) \rightarrow Y = C.$$

where $F_1$, $F_2$, ..., $F_m$ are features (explanatory variables) and $Y$ is the class (dependent variable).
It is important to remember that in RF, rules are only weak classifiers whose decision contributes by only one vote to the final decision (that of the forest).

### 3.4. Evaluating an RF Classifier

Evaluating data mining models helps to predict how well they will work in the future and to refine their parameters. In this work, we mainly use two criteria to evaluate a Random Forest: accuracy and stability.

- Prediction accuracy is characterized by two errors: the OOB-error explained above and the Test-error obtained by applying the model to a test set.
- Stability: Since it uses randomness in sampling training set and in selecting features in each node of each tree, RF method is not deterministic: different runs with the same values of *ntree* and *mtry* usually produce different forests, i.e. different predictions for the same data. There are many ways to measure this variability/stability (see for example Bryan et al., 2017). In this work we characterize it by the fact that when we run RF building algorithm many times with the same values of *ntree* and *mtry*, we have only "small changes" in the error and the ranking of the variables with respect to their importance.

Let us add that, as outlined by Bryan et al. (2017), there is a third criterion to consider: the computational cost. We show below how to find a tradeoff between a low error, a high stability and an acceptable computational cost.

let us also notice that we also have the possibility to measure the quality of the weak classifiers forming the forest (individual trees and the rules extracted from them). Two metrics are used for that:

- The frequency: it measures rule's popularity and is defined as the proportion of data instances satisfying the rule condition.
- The error: it is defined as the number of incorrectly classified instances by the rule divided by the number of instances satisfying the rule condition.

## 4. RELATED WORK

Historically, linear modelling was the first approach of econometrics to explain growth. Combining factor analysis with linear modelling is an extension of the approach and it is especially important for forecasting. Then, nonlinear modelling became important where Osin´ska et al. (2018), Khan and Hanif (2018) and Omay et al. (2017) can be given as

examples. A wide review of articles can be found in (Akinsola and Odhiambo, 2017; Eklund and Kapetanios, 2008; Osin´ska et al., 2018). However, Varian(2014) states that with big data a need to detect and summarize more complex models than the linear ones emerged and historical methods are not sufficient any more. ML offers a set of tools like neural network, decision trees and random forests to analyze these complex models. In the following, we present some representative examples of these works.

Biau and D'elia (2010) forecast quarterly GDP growth for a large data set for Euro area. The authors present and evaluate a two-step strategy where firstly RF variable importance measure is used to identify relevant variables and then a linear model is built using these selected variables as input. Biau et al. (2007) also uses RF to reduce dimensionality while forecasting French manufacturing output growth based on a firm level survey data. Lkonen (2016) analyzes distribution of income problem by using a dataset of 43 variables and more than 1,200,000 observations. Again RF variable importance measures are used to determine the main distinguishing variables. Then the effects of these important variables are analyzed by the 'two trees algorithm' (Athey and Imbens, 2015). To compare ML approaches with traditional econometric models, the authors also ran a logistic regression. More recently, Minhas (2018) applied the same approach to analyze the effect of financial factors on employment rate for manufacturing firms.

Zhao et al. (2017) proposes a bankruptcy prediction model on the kernel extreme learning machine and compares this with support vector machines, extreme learning machine, RF, particle swarm optimization enhanced fuzzy k-nearest neighbor and Logit model methods. Medeiros et al. (2018) use several machine learning methods like LASSO, bagging, boosting, RF and so on to forecast US inflation. It is shown that these methods give more accurate results than the conventional methods. To analyze money laundering detection Zhang et al. (2018) use ML and sampling schemes in empirical analysis of money laundering detection algorithms for US financial institutions. Five major ML algorithms, namely Bayes logistic regression, decision tree, RF, support vector machine, and artificial neural network are used in the paper.

Ahmed et al. (2019) investigate predicting stock trends over the short term for a specific company. To train, test and validate the system, a dataset stretching over a duration of ten years is used. The results show improvement over the efficient market hypothesis and the system has a significant improvement on the predictive power.

## 5. DATA

In Sala-i-Martin (1997), 62 variables were studied for the period 1961-1992. Among them, there were variables related to regions, religions and level of democracy. We did not include these variables to our analysis since they do not serve the purpose of our paper. Therefore, we ended up with a set of 15 variables. We consider the period 1961-2016.

We downloaded the data from the World Bank database[1]. In order to have a full data, we did not include the countries which have missing data within the period studied for our analysis. Therefore, we ended up with 49 countries[2].

We transformed numerical variables into categorical/ordinal variables having two or three values (Low/High or Low/Middle/High). For the variables included in the study, except

[1] https://data.worldbank.org/indicator , The data is available if requested.

[2] Algeria, Argentina, Australia, Austria, Belgium, Canada, Colombia, Costa Rica, Cyprus, Denmark, Ecuador, El Salvador, Finland, France, Germany, Ghana, Greece, Guatemala, Honduras, India, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kenya, Korea, Malawi, Malaysia, Mexico, Netherlands, New Zealand, Nicaragua,
Norway, Pakistan, Panama, Paraguay, Peru, Philippines, Senegal, Spain, Sri Lanka, Sweden, Switzerland,
Tanzania, Thailand, Tunisia, Turkey, Uganda, United Kingdom, United States, Uruguay and Venezuela

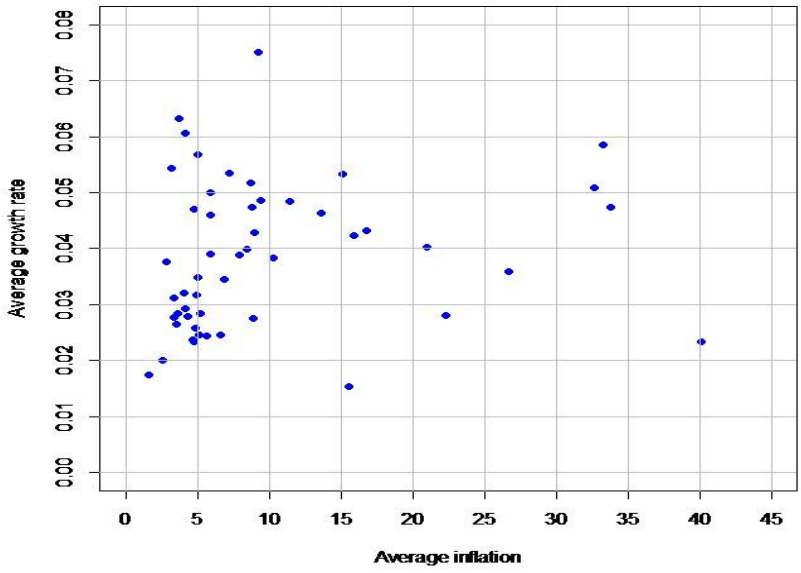
*average GDP growth rate*, *log of average GDP per capita* and *average rate of population growth*, we determined the ranges of low, medium and high values for the variables with the following rule:

- If *value* < (*min* + $\mu$)/2 then *Low*
- Else If *value* > ($\mu$ + *max*)/2 then *High*
- Else *Middle*.

where $\mu$ is the mean of the data, *min* is the minimum value in the data and *max* is the maximum value in the data. However, for *average GDP growth rate*, *log of average GDP per capita* , and *average rate of population growth*, mean of the data is the point where there is a shift from *Low* to *High*[3].

In Figure 5.1, each dot represents the average inflation and average GDP growth of the countries. The nonlinear relationship can be realized from this figure. Most of the inflation rates are within the range of 0−15%. The GDP growth rate goes up to 7%.

**Figure 5.1** Average Inflation vs. Average GDP Growth (1961 - 2016)




We report the descriptive statistics of average GDP growth and inflation for the period in table 5.1.

[3] The names of the variables, the short names we gave them ($\gamma$, Infl, X1, X2, ...), and their ranges of low, middle and high values are presented in the Appendix.

| *1961-2016* | | |
|---|---|---|
| | GDP | Inflation |
| Mean | 0.04157 | 13.367 |
| Median | 0.04045 | 7.719 |
| Minimum | 0.01526 | 3.606 |
| MaximuM | 0.07808 | 93.072 |
| Std. Dev | 0.01490924 | 16.34126 |

**Table 5.1** Descriptive statistics of data.

## 6. EXPERIMENTAL SETTING/METHODOLGY

Two parameters have to be chosen: the number of trees to build (*ntree*) and the number of variables to try at each node (*mtry*). Following (Breiman, 2002), The *R* package that we use proposes default values for these parameters (Liaw and Wiener, 2002) : *ntree*=500 and *mtry*= $\sqrt{p}$ (*p* is the total number of variables). But there is no theoretical results to support these values which have been shown in many works to be not often optimal (see for example (Genuer et al., 2008)).

About the parameter *ntree*, we know that it has to be large enough to create diversity among the base classifiers, but giving it a very high value would be computationally expensive and useless: several works (e.g. (Latinne et al., 2001; Oshiro et al., 2012)) showed that there exists a value of this parameter, depending on the dataset, beyond which we don't have improvement of the Random Forest quality, whatever the criterion used to measure this quality. In (Latinne et al., 2001), the authors propose a procedure based the McNemar test (Salzberg, 1997) to choose a value of *ntree* for a given data set. This test compares the number of examples misclassified by two classifiers, and the authors apply it between pairs of random forests that differ only by their number of trees. Applied to 5 datasets having different values of *n* (number of observations), *p* (number of variables) and *s* (number of classes), this method suggests an optimal value of *ntree* varying from 60 to 200. The work presented in (Oshiro et al., 2012) introduces the so-called *density* of a dataset and gives three metrics to measure it. Then, analyzing the area under the ROC curve (AUC) and comparing its values for a number of trees varying from 2 to 4096, and 29 datasets, the authors conclude that low-density datasets may require a higher value of *ntree* than high-density ones, and they suggest a range between 64 and 128 trees in a random forest.

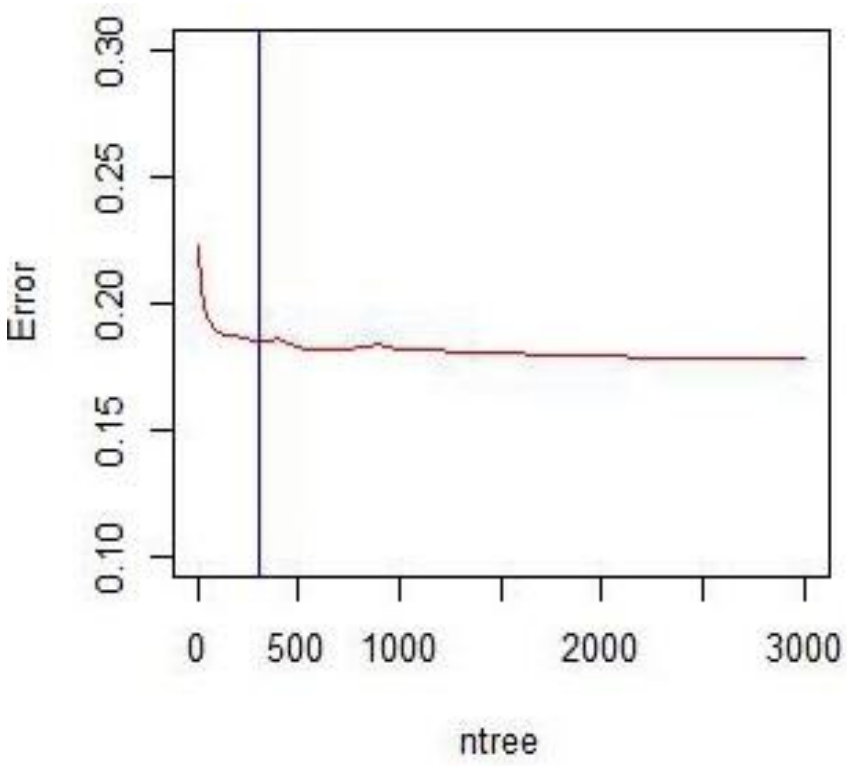
In this work we consider that accuracy is not the only criterion to characterize a 'good' classifier. Therefore, the experiments described in the following sections aim to find a couple (or a set of couples) (*ntree,mtry*) giving 'as good as possible' random forests with respect to two criteria : prediction accuracy and stability. We characterize prediction accuracy by OOB-error described in the previous section. Stability we want to achieve is that of both accuracy and variable importance.

### 6.1. How Many Trees To Minimize The Error ?

This first experiment aims to know if there is a tree number beyond which prediction accuracy remains the same or decreases. For that we considered a value of *ntree* varying

from 10 to 3000 and for each one of these values we considered all the possible values of *mtry* ({1,...,15}). For each couple (*ntree*,*mtry*), OOB-error is averaged over 100 runs. Figure 6.1 represents the best error (minimum over mtry) in function of *ntree*. We notice an asymptotic behavior: there is a first interval in which the error decreases, then a second one in which the error is almost constant. The vertical line indicates the border between the two parts. This value of *ntree* has been calculated by considering that the error is constant if its relative variation is under some threshold, set at 0.05 in this experiment. Since it is the value beyond which prediction accuracy can not be improved, this value can be considered as the optimal value with respect to this criterion. This optimal value is equal 300. In the following sections we combine this criterion with other ones.

**Figure 6.1.** OOB error against ntree




## 6.2. How Many Trees To Stabilize The Prediction Accuracy?

Since it uses randomness in sampling training set and in selecting features in each node of each tree, random forest method is not deterministic: different runs with the same values of *ntree* and *mtry* usually produce different forests, i.e. different predictions for the same data. There are different methods to measure this variability (see for example (Bryan et al., 2017)). In this work we characterize it by prediction error standard deviation over a fixed number, *nbr*, of runs. Figure 6.2 shows the maximum, the average and the minimum standard deviation of the OOB error in function of *ntree* for *nbr* = 100.

In order to find the ”good values” of *ntree* with respect to accuracy variability/stability, we define the stability level of a random forest by a couple $(\alpha,l)$ as follows :

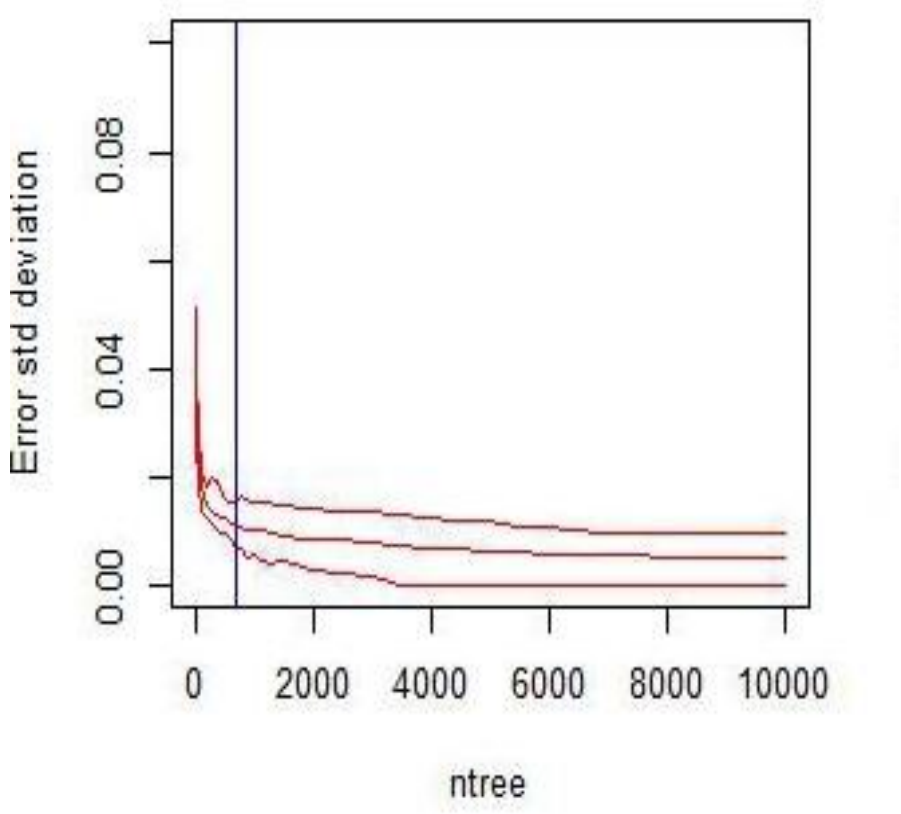
$$Pr(Error \in I_l) \geq \alpha$$

where $I_l$ is an interval of width $l$, $l$ is a real and $\alpha$ is a real belonging to [0,1]. High stability corresponds to low values of $l$ and a high values of $\alpha$. To obtain a stability level of a random forest, we use the Tchebychev inequality in the following form:
Let $X$ be a random variable with a mean value $\mu$ and a standard deviation $\sigma$, we have for any real number $k>0$ :

$$Pr(k\sigma \leq X - \mu \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

In this work we took $k$ = 3 which corresponds to $\alpha$=0.9 and $l$ = 0.05. We know that the bounds of confidence intervals given Tchebychev inequality are quite loose. We deduce that this couple ($l$,$\alpha$) corresponds to a good level of stability. Vertical line in figure 6.2 indicate the minimum value of *ntree* for which we obtain this level of stability. This value value is *ntree* = 700.

**Figure 6.2.** Standard deviation of OOB error against ntree




### 6.3. How Many Trees To Stabilize Variables Importance (VI)?

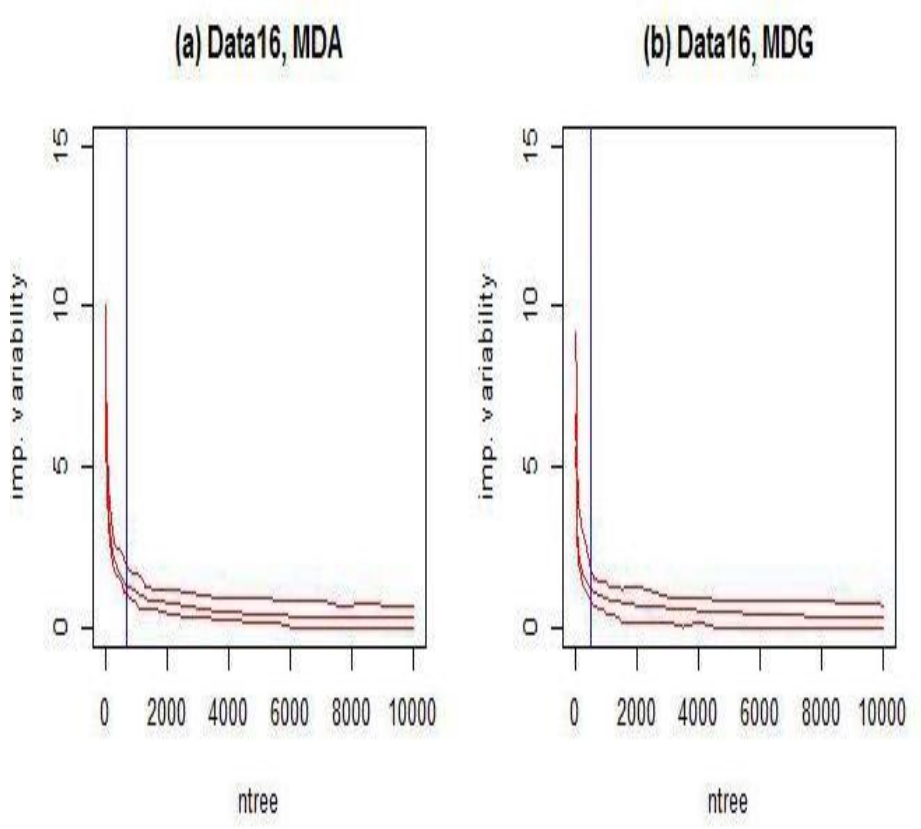
To study the sensitivity of variable importance to the parameter *ntree* we proceed as follows :

- For each value of *ntree*, for each value of *mtry*, we ranke variables with respect to their importance measure. Then, for each variable $v$ we compute the interquantile

range of level $\alpha$ (here $\alpha$=0.05) of its rank over *nbr* runs (here *nbr*=100). Let us call *IQR*(ntree,mtry,*v*) this value.

- For each couple (*ntree*, *mtry*), we compute the average of *IQR*() over the set of variables.
- For each value of *ntree* we compute the minimum, the maximum and the average values of this value over the values of *mtry*. Let us call IQRmax(*ntree*), IQRmin(*ntree*) and IQRmean(*ntree*) these three function.
- Figure 6.3 represents these three function for the two measures of variable importance.
- In order to obtain good values of *ntree* with respect to variable importance stability we consider that variable importance is stable when variable ranks belong to a *thr* width interval. Here we take *thr*=2, which means the rank of each variable is between $n - 1$ and $n + 1$ for some rank $n$. The vertical line of figure 6.3 indicate minimum values of *ntree* corresponding to this constraint. These values are *ntree* = 700 for MDA and *ntree* = 500 for MDG. We notice that MDA need higher values of *ntree* to be stabilized. This seem to confirm conclusions of Call and Urrea (2011).

**Figure 6.3** Variation of VI against ntree



### 6.4. How Many Variables To Test At Each Node?

According to the late Professor Breiman (2002), random forests are not too sensitive to mtry's value "as long as it's in the right ball park." The advice he gives is to begin with $mtry=\sqrt{p}$, a value that has been found to generally give near optimum values, then to continue with a value half as low and twice as high checking the *OOB*-error. But, as pointed out in (Scornet, 2018), there are no theoretical findings to support this default value.
In (Genuer et al., 2008), the authors empirically study the *OOB*-error in function of *mtry* for classification and regression problems. For that, they distinguish two kinds of problems depending on the values of $n$ and $p$: standard problems for which we have $n >> p$ and high dimensional problems corresponding to $p >> n$. In the standard classification problems, they consider 13 datasets (9 real datasets and 4 simulated datasets) and three values of *ntree* (100, 500 and 1000). Their results show that the *OOB*-error curve shape depends on the dataset and the value of *ntree*: its minimum is actually reached around $mtry=p$ for real datasets (especially when $ntree=500$) and it shows an increasing function for simulated datasets.

In (Dìaz-Uriarte and Alvares de Andrés, 2006), an investigation of the use of random forest for classification of microarray data includes the effects of *mtry* on error rate. Nine datasets have been considered and many values of *ntree* and *mtry* have been tried. The authors' conclusion is that the relation of OOB-error rate with *mtry* is largely independant of *ntree* and that the default setting of *mtry* often "a good choice", even if, in some cases, increasing it can lead to "small decreases" in error rate.

The effect of *mtry* value on OOB error and variable importance is examined in Genuer et al. (2017). The authors show that taking larger values of this parameter (with respect to the default value) allows to obtain a small gain in error and an important improvement of the magnitude of VI.

In this work, once determined the "good" values of *ntree*, we studied the effect of *mtry* on the error rate. We noticed that the lowest *OOB* error rates are obtained for small values of *mtry* ($mtry \in \{2,3\}$).

## 7. EXPERIMENTAL RESULTS/EVALUATION

Once the optimal values of *ntree* and *mtry* determined, we created the *RF* classifier with $ntree = 700$, $mtry = 2$. The OOB accuracy of this classifier is 81%. In the following we describe and analyze rules extracted from this classifier, and variable importance values given by the algorithm.

### 7.1. Rules

Table 7.1 reports the sixteen rules that include inflation but not standard deviation of inflation. Rules 2, 3, 4, 5 and rules from 10 to 16 involve middle values of inflation (between 5.338 and 25.086). For these rules, there is high GDP growth (greater than 0.038). On the other hand, for rules where inflation is either low (smaller than 5.338) or high (greater than 25.086), we end up with low GDP growth (smaller than 0.038). Therefore, contrary to the result of Sala-i-Martin (1997), where inflation is not a significant variable, there is a relation between inflation and growth. This difference is due to the fact that, as also stated by Sala-i-Martin (1997), the analysis there is a linear one but ours is a nonlinear

one. Since RF captures nonlinear relations as well, we could end up with a relation between inflation and growth.

| | Infl | X1 | X7 | X2 | X3 | X4 | X8 | X9 | X10 | X12 | X13 | X14 | *y* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | L | H | L | | | | | | | L | | | L |
| 2 | L/M | L | L | L/M | | | | | | | | | H |
| 3 | L/M | L | H | | | | M | | M | | | M/H | H |
| 4 | L/M | L | H | M/H | | | | | | M | | | H |
| 5 | L/M | L | H | | | | | | | | | | H |
| 6 | L/H | H | L | | | | | L/M | M/H | | | | L |
| 7 | L/H | H | L | | | | | | | L/M | | H | L |
| 8 | L/H | H | L | | | | | | | L/M | | L/M | L |
| 9 | L/H | H | L | | | | | | M/H | | | | L |
| 10 | M | L | H | | | | | | M/H | | | | H |
| 11 | M | L | H | | | | | L/M | | | M | | H |
| 12 | M | L | H | | | | | | | | | | H |
| 13 | M | L | H | | | | M | | M/H | | | M | H |
| 14 | M | L | H | | M/H | | | L/M | | | M | | H |
| 15 | M | L | H | | | M | | | | | | | H |
| 16 | M | L | H | L/M | | M | | | | | | | H |

**Table 7.1** Rules containing only inflation.

In Table 7.2, there are the eight rules which involve both inflation and standard deviation of inflation. Similar to the previous case there is again a nonlinear relationship between inflation and growth. As can be seen in the Table, except rules 2, 3 and 4, middle values of inflation (between 5.838 and 25.086) exist and these end up with high GDP growth (greater than 0.038). On the other hand, inflations which are either low or high (smaller than 5.838 or greater than 25.086) ends up with low GDP growth (smaller than 0.038). The results for standard deviation of inflation are parallel to the results for inflation due to the fact that inflation and standard deviation of inflation are positively correlated.

| | Std Dev. (X5) | Infl | X1 | X7 | X3 | X4 | X8 | X10 | X12 | $\gamma$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | L/M | M/H | L | H | | | | | | H |
| 2 | L/H | L | H | L | | | | M | | L |
| 3 | L/H | L | H | L | M | | | | L | L |
| 4 | L/H | L/H | H | L | | | | L/H | | L |
| 5 | M | L/M | L | H | | M | M | | | H |
| 6 | M | L/M | L | H | | | | | M | H |
| 7 | M | M | L | H | | | | M | | H |
| 8 | M/H | L/M | L | H | M | | | | | H |

**Table 7.2** Rules containing both inflation and standard deviation of inflation.

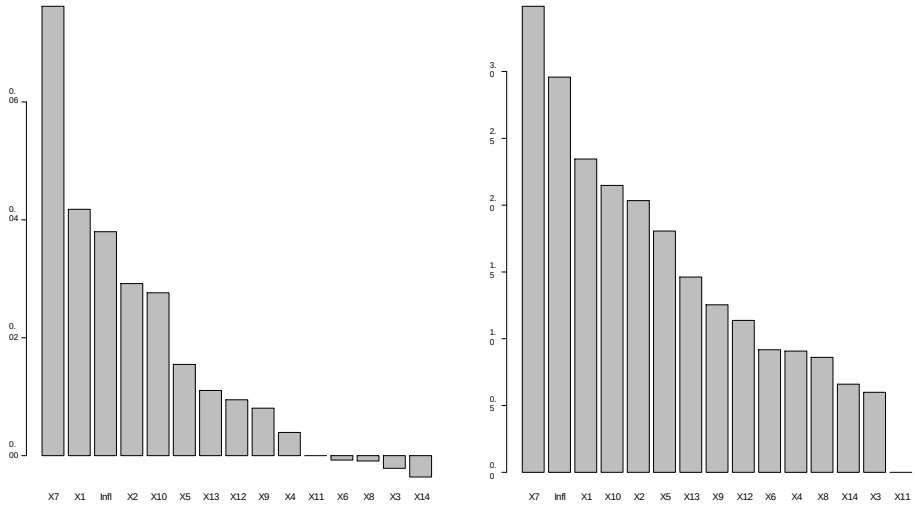
## 7.2. Variable Importance

As stated in Section 3.2, two most commonly used ways to measure variable importance are mean decrease accuracy (MDA) and mean decrease Gini (MDG). When we applied these measures we obtained the importance values and ranks reported in figure 7.1 and table 7.3. As can be noticed, from figure 7.1 and table 7.3 the most important variables are *average rate of population growth*, *Inflation*, *GDP per capita*, *Average life expectation at birth*, *Average urban population growth* and *Standard deviation of inflation.*

| Rank | MDA | MDG |
|---|---|---|
| 1 | Average Rate of Population Growth | Average Rate of Population Growth |
| 2 | Log of Average GDP per Capita | Average Inflation Rate |
| 3 | Average Inflation Rate | Log of Average GDP per Capita |
| 4 | Average Life Expectancy at Birth | Average Urban Population Growth |
| 5 | Average Urban Population Growth | Average Life Expectancy at Birth |
| 6 | Standard Deviation of the Inflation Rate | Standard Deviation of the Inflation Rate |
| 7 | Industry Value Added | Industry Value Added |
| 8 | Agriculture Value Added | Average Rate of Imports of Goods and Service |
| 9 | Average Rate of Imports of Goods and Services | Agriculture Value Added |
| 10 | Domestic Credit | Total Area of the Country |
| 11 | Labor Force | Domestic Credit |
| 12 | Total Area of the Country | Average Rate of Exports of Goods and Service |
| 13 | Average Rate of Exports of Goods and Services | Services Value Added |
| 14 | Average Primary School Enrollment | Average Primary School Enrollment |
| 15 | Services Value Added | Labor Force |

**Table 7.3** Variable Importance.

**Figure 7.1** Variable Importance




## 8. CONCLUSION

Variable selection which should be included in the true regression model is the initial decision that should be given by empirical economists. Conventional econometrics use several different model selection criteria to determine the variables. Sala-i-Martin (1997) studied all variables that are included in research papers after running two million regressions found out 62 variables which were significant at least in one of the regressions. However, he could not find a relation between inflation and economic growth which is a conclusion that contradicts with economic theory. He explains this contradiction by stating that he made a linear analysis and most probably the relation between inflation and economic growth is a nonlinear one.

In this paper, we used *RF*, which is a powerful *ML* methodology for nonlinear analysis. Our methodology captures the relation between inflation and economic growth. If inflation is lower than 6% or higher than 25%, then growth is lower than 4%. If inflation is in the range 6% and 25%, then growth is above 4%. This result shows that not only high inflation but also low inflation can be a problem for economic growth. There is a consensus among economist about the tools to solve high inflation problem but low inflation problem is a new one that should be studied. According to these results, we can suggest that inflation targets should be set around 6% to achieve economic growth. Hammond (2012) and Roger (2010) reports inflation targets of twenty-eight countries. Among them eight countries have inflation targets of around 6% but the rest of them have targets of around 3%. Moreover, in our analysis we found out that *average rate of population growth*, *GDP per capita*, *inflation*, *average life expectation at birth*, *average urban population growth* and *standard deviation of inflation* are the most important variables for explaining growth. Therefore, in a further study a Monte Carlo simulation can be made comparing models obtained with conventional model selection criteria and the ones obtained with ML.

**Appendix A. List of variables**

- Dependent Variable: Average GDP Growth Rate.
  - Short name: $\gamma$.
  - The Ranges of Low and High Values:
    - If Average GDP Growth Rate < 0.038, then Low.
    - If Average GDP Growth Rate > 0.038, then High.
- Average Inflation Rate.
  - Short name: Infl.
  - The Ranges of Low, Middle and High Values:
    - If Average Inflation Rate < 5.838, then Low.
    - If 5.838 < Average Inflation Rate < 25.086, then Middle.
    - If Average Inflation Rate > 25.086, then High.
- Log of Average GDP per Capita (Constant 2010 US Dollar).
  - Short name: X1.
  - The Ranges of Low and High Values:
    - If Log of GDP per Capita < 8.851, then Low.
    - If Log of GDP per Capita > 8.851, then High.
- Average Life Expectancy at Birth.
  - Short name: X2.
  - The Ranges of Low, Middle and High Values:
    - If Average Life Expectancy at Birth < 57.809, then Low.
    - If 57.809 < Average Life Expectancy at Birth < 73.244, then Middle.
    - If Average Life Expectancy at Birth > 73.244, then High.
- Average Primary School Enrollment.
  - Short name: X3.
  - The Ranges of Low, Middle and High Values:
    - If Average Primary School Enrollment < 80.243, then Low.
    - If 80.243 < Average Primary School Enrollment < 105.303, then Middle.
    - If Average Primary School Enrollment > 105.303, then High.
- Domestic Credit Provided by Financial Sector (% of GDP)
  - Short name: X4.
  - The Ranges of Low, Middle and High Values:
    - If Domestic Credit Provided by Financial Sector < 33.726, then Low.
    - If 33.726 < Domestic Credit Provided by Financial Sector < 63.803, then Middle.
    - If Domestic Credit Provided by Financial Sector > 63.803, then High.
- Standard Deviation of the Inflation Rate.
  - Short name: X5.
  - The Ranges of Low, Middle and High Values:
    - If Standard Deviation of the Inflation Rate < 6.228, then Low.
    - If 6.228 < Standard Deviation of the Inflation Rate < 38.963, then Middle.
    - If Standard Deviation of the Inflation Rate > 38.963, then High.
- Total Area of the Country.
  - Short name: X6.

  - The Ranges of Low, Middle and High Values:
    - If Total Area of the Country < 100, then Low.
    - If 100 < Total Area of the Country < 1000, then Middle.
    - If Total Area of the Country > 1000, then High.
- Average Rate of Population Growth.
  - Short name: X7.
  - The Ranges of Low and High Values:
    - If Average Rate of Population Growth < 1.634, then Low.
    - If Average Rate of Population Growth > 1.634, then High.
- Average Rate of Exports of Goods and Services (% of GDP).
  - Short name: X8.
  - The Ranges of Low, Middle and High Values:
    - If Average Rate of Exports of Goods and Services < 19.348, then Low.
    - If 19.348 < Average Rate of Exports of Goods and Services < 50.696, then Middle.
    - If Average Rate of Exports of Goods and Services > 50.696, then High.
- Average Rate of Imports of Goods and Services (% of GDP)
  - Short name: X9.
  - The Ranges of Low, Middle and High Values:
    - If Average Rate of Imports of Goods and Services < 21.051, then Low.
    - If 21.051 < Average Rate of Imports of Goods and Services < 54.952, then Middle.
    - If Average Rate of Imports of Goods and Services > 54.952, then High.
- Average Urban Population Growth.
  - Short name: X10.
  - The Ranges of Low, Middle and High Values:
    - If Average Urban Population Growth < 1.438, then Low.
    - If 1.438 < Average Urban Population Growth < 4.425, then Middle.
    - If Average Urban Population Growth > 4.425, then High.
- Labor Force.
  - Short name: X11.
  - The Ranges of Low, Middle and High Values:
    - If Labor Force < 1000, then Low.
    - If 1000 < Labor Force < 10000, then Middle.
    - If Labor Force > 10000, then High.
- Agriculture Value Added (% of GDP).
  - Short name: X12.
  - The Ranges of Low, Middle and High Values:
    - If Agriculture Value Added < 4.472, then Low.
    - If 4.472 < Agriculture Value Added < 21.970, then Middle.
    - If Agriculture Value Added > 21.970, then High.
- Industry Value Added (% of GDP)
  - Short name: X13.
  - The Ranges of Low, Middle and High Values:
    - If Industry Value Added < 19.200, then Low.

    - If 19.200 < Industry Value Added < 34.399, then Middle.
    - If Industry Value Added > 34.399, then High.
- Services Value Added (% of GDP).
    - Short name: X14.
    - The Ranges of Low, Middle and High Values:
        - If Services Value Added < 52.941, then Low.
        - If 52.941 < Services Value Added < 75.588, then Middle.
        - If Services Value Added > 75.588, then High.

**References**

Ahmed M., Sriram A., Singh S. (2019). Short Term Firm-Specific Stock Forecasting with BDI Framework. Computational Economics. https://doi.org/10.1007/s10614-019-09911-0.

Akaike H. (1974). A new look at statistical model identification. IEEE Transactions on Automatic Control. AC-19, 716-723.

Akinsola F., Odhiambo N. (2017). Inflation and Economic Growth: a Review of The International Literature. Comparative Economic Research 20(3), 42-56.

Athey S., Imbens G.W. (2015). Machine Learning for Estimating Heretogeneous Casual Effects. Stanford Graduate School of Business. https://www.gsb.stanford.edu/gsb-cmis/gsb-cmisdownload-auth/406621.

Barro R.J. (1991). Economic Growth in a Cross Section of Countries. The Quarterly Journal of Economics 106(2), 407-443.

Basci S., Zaman A., Kiraci A. (2010). Variance estimates and model selection. International Econometric Review. 2 (2), 57-72.

Behnamian A., Millard K., Banks S.N., White L., Richardson M., Pasher J. (2017). A Systematic Approach for Variable Selection With Random Forests: Achieving Stable Variable Importance Values. IEEE Geoscience and Remote Sensing Letters 14(11), 1988-1992.

Biau O., D'elia A. (2014). Euro Area GDP Forecast Using Large Survey Dataset- A Random Forest Approach. Eco Mod2010. https://EconPapers.repec.org/RePEc:ekd:002596:259600029.

Biau G., Biau O., Rouviere L. (2007). Nonparametric forecasting of the manufacturing output growth with firm-level survey data. Journal of Business Cycle Measurement and Analysis 3, 317-332.

Breiman L. (2001) Random Forests. Machine Learning 45(1), 5-32.

Breiman L. (2002). Manual On Setting Up, Using, And Understanding Random Forests V3.1. Machine Learning. http://oz.berkeley.edu/users/breiman/Using random forests V3.1.pdf.

Bryan Liu C.H., Chamberlain B.P., Little D.A., Cardoso A. (2017). Generalising Random Forest Parameter Optimisation to IncludeStability and Cost. CoRR. http://arxiv.org/abs/1706.09865.

Calle M.L., Urrea V. (2011). Letter to the Editor: Stability of Random Forest importance measures. Briefings in Bioinformatics 12(1), 86-89.

Dìaz-Uriarte R., de Andrés S.A. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(3).

Eklund J., Kapetanios G. (2008). A Review of Forecasting Techniques for Large Data Sets. Queen Mary University of London Working Papers 625.

Genuer M.R., Poggi J.M., Tuleau C. (2008). Random forests : some methodological insights. arXiv. https://arxiv.org/abs/0811.3619.

Genuer MR, Poggi JM, Tuleau-Malot C (2017). Variable selection using Random Forests. Pattern Recognition Letters 31(14), 2225-2236.

Goldfarb R. S. (1995). The economist-as-audience needs a methodology of plausible inference. Journal of Economic Methodology 2 (2), 201-222.

Goldfarb R. S. (1997). Now you see it, now you don't: emerging contrary results in economics. Journal of Economic Methodology 4 (2), 221-244.

Hammond G. (2012) State of the art of inflation-targeting. Centre for Central Banking Studies Handbook.London: Bank of England.

Hannan E. J., Quinn B. G. (1979). The determination of the order of an autoregression. Journal of the Royal Statistical Society B. 41, 190-195.

Hendry D.F., Krolzig H.M. (2004). We Ran One Regression. Oxford Bulletin of Economics and Statistics 66(5), 799 - 810.

Hurvich C. M., Tsai C. L. (1989). Regression and time series model selection in small samples. Biometrika. 76 (2), 297-307.

Khan M., Hanif W. (2018) Institutional quality and the relationship between inflation and economic growth. Empirical Economics 58, 627-649.

Latinne P., Debeir O., Decaestecker C. (2001). Limiting the Number of Trees in Random Forests. Multiple Classifier Systems (MCS), Second International Workshop, 2001 Cambridge, UK, July 2-4, 2001, Proceedings, 178-187.

Liaw A., Wiener M. (2002). Classification and Regression by RandomForest. R News 2/3(3), 18-22.

Lkonen H. (2016). Machine Learning in Applied Econometrics: Derining personal income drivers with randomized decision trees. Master's thesis, Aalto University School of Business.

Medeiros M.C., Vasconcelos G., Zilberman E., Veiga A. (2018). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. Available at SSRN: https://ssrn.com/abstract=3155480 or http://dx.doi.org/10.2139/ssrn.3155480 (Accessed August 17, 2021).

Minhas G. (2018). Essays in Applied Panel Data Econometrics and Machine Learning. Doctoral thesis, Department of Economics, University of Konstanz.

Omay T., van Eyden R., Gupta R. (2017). Inflation–growth nexus: evidence from a pooled CCE multiple-regime panel smooth transition model. Empirical Economics 54, 627-649.

Oshiro T.M., Perez P.S., Baranauskas J.A. (2012). How Many Trees in a Random Forest? Machine Learning and Data Mining in Pattern Recognition - 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings, 154-168.

Osin'ska M., Tadeusz K., Bazejowski M., Pawel K. (2018). Modeling mechanism of economic growth using threshold autoregression models. Empirical Economics 58, 1381-1430.

Quinn B. G. (1980). Order determination for a multivariate autoregression. Journal of the Royal Statistical Society B. 42, 182-185.

Rissanen J. (1978). Modelling by shortest data description. Automatica. 14, 465-471.

Roger S. (2010). Inflation Targeting Turns 20. Finance & Development 47(1), 46-49.

Sala-i-Martin X.X. (1997). I Just Run Two Million Regressions. The American Economic Review 87(2), 178-183.

Salzberg S. (1997). On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data Mining Knowledge Discovery1(3), 317-328.

Schwarz G. (1978). Estimating the dimension of a model. The Annals of Statistics. 6, 461-464.

Scornet E. (2018). Tuning parameters in random forests. ESAIM : Proceedings and surveys 60, 144-162.

Varian H.R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28(2), 3-28.

Zhang Y., Trubey P. (2018). Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection. Computational Economics. https://doi.org/10.1007/s10614018-9864-z (accessed august 17, 2021).

Zhao D.C., Huang C., Wei Y.(2017. AnEffective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach. Computational Economics. https://doi.org/10.1007/s10614-016-9562-7 (accessed august 17, 2021).