Sovereign risk ratings' country classification using machine learning

Diego Ramon Bezerra da Silva* Thaís Gaudêncio do Rêgo† Bruno Ferreira Frascaroli‡

July 20, 2019

Abstract

In this paper, we present new empirical evidence about sovereign risk ratings provided by credit rating agencies. They are important because they indicate the risk assumed by foreign investors when acquiring debt bonds from any country. Our empirical strategy has four steps. First, we built a database using the observed sovereign ratings provided by Fitch, Moody's and Standard & Poor's, the World Development Indicators and reports from the World Economic Situation and Prospects report. The dataset of 137 countries collected from 1958 to 2017 was firstly composed of 3,596 instances with 22 different sovereign classifications. Second, we manually processed this dataset using clustering and principal component analysis for comparison purposes. Therefore, the number of instances was reduced to 1,597, increasing the prediction likelihood. Then, we used a machine learning framework, more specifically, the Random Forest algorithm to predict the sovereign ratings. Lastly, we used econometric models of truncated response to test if sovereign ratings patterns changed i) in the aftermath of the subprime crisis in 2008 and ii) according to the level of development of countries. Also, the model was able to predict up to 98.28% of the ratings. The clustering indicated four large groups of countries that share some characteristics among them. In addition, the crisis that took place in 2008 represents a structural break in the ratings, as well as in the level of development of countries.

Keywords: Sovereign risk ratings; countries; big data; machine learning; random forest.

Resumo

Neste artigo são apresentadas novas evidências empíricas sobre ratings de risco soberano fornecidos por agências de classificação de risco de crédito. Eles são importantes porque indicam o risco assumido pelos investidores estrangeiros ao adquirir títulos de dívida de qualquer país. A estratégia empírica tem quatro etapas. Primeiro, foi construído um banco de dados usando os ratings soberanos observados fornecidos pela Fitch, Moody's e Standard & Poor's, os Indicadores de Desenvolvimento Mundial e os relatórios do relatório World Economic Situation and Prospects. O conjunto de dados de 137 países coletados entre 1958 e 2017 foi composto inicialmente por 3.596 instâncias, com 22 diferentes classificações soberanas. Depois, esse conjunto de dados foi processado manualmente usando a análise de clustering e de componentes principais para fins de comparação. Portanto, o número de instâncias foi reduzido para 1.597 para aumentar a acurácia da previsão do modelo. Então, foi utilizada uma estrutura de aprendizado de máquinas, mais especificamente, o algoritmo Random Forest para prever os ratings soberanos. Por último, utilizou-se modelos econométricos de

*Graduate Program of Informatics - Federal University of Paraíba. Email: diegoramon95@gmail.com.
†Graduate Program of Informatics - Federal University of Paraíba. Email: gaudenciothais@gmail.com.
‡Department of Economics - Federal University of Paraíba. Email: frascaroli.b@gmail.com.

resposta truncada para testar se os padrões de ratings soberanos mudaram i) em função da crise do subprime em 2008 e ii) de acordo com o nível de desenvolvimento dos países. Além disso, o modelo foi capaz de prever até 98,28% das classificações. O agrupamento indicou quatro grandes grupos de países que compartilham algumas características entre eles. Além disso, a crise que ocorreu em 2008 representa uma quebra estrutural na atribuição de ratings, bem como no nível de desenvolvimento dos países.

Palavras-chave: Ratings de risco soberano; países; big data; aprendizado de máquinas; random forest.

JEL Code: G24, E44, C45.

Área 4: Macroeconomia, Economia Monetária e Finanças.

1 Introduction

Sovereign risk ratings are among the most important indicators used in the international financial market to reduce information asymmetry (Poor's, 2011). They summarize information about the future creditworthiness of the debt issuers, i.e., the risk incurred by foreign investors when acquiring securities of some issuer. Provided by the credit rating agencies (CRAs), they are divided into corporate and sovereign risk. This last one is especially important because some of the largest debt issuers are sovereign states. Therefore, ratings provided by CRAs impact their bonds issuing and contracts in worldwide financial markets.

The greater the risk which investors assume when acquiring some bond from a sovereign government, the lower the government's ability to make this acquisition attractive and thus attract foreign investors. As a consequence, higher is the reward paid to investors to compensate them for assuming this risk (Basu et al., 2013; Seetharaman et al., 2017). In addition, ratings are based on credit scores of each country and, for this reason, they have a severe impact on the corporate ratings of companies located in their respective sovereign territories (Borensztein et al., 2013). Sovereign ratings reflect, by consequence, quantitative and qualitative analysis of economic and political risks.

Specifically, they involve judgment of internal and external macroeconomic variables, as well as the prediction of their expectations in the future (Moody's Investors Service, 2016). Considering that financial transactions are intrinsically marked by information asymmetries and spillover effects, the CRAs could help reduce international financial crisis. They present themselves as independent companies of any interest, either by governments or private companies. This feature allows them to have as principles: independence, objectivity, credibility and freedom of disclosure of ratings regarding issuers credit quality and debt issuance (Poor's, 2015).

However, there are several reported problems associated with ratings classification (Cantor and Packer, 1996; Ferri and Stiglitz, 1999; Partnoy et al., 2002; Sy, 2009; Utzig, 2010; Božović et al., 2011; Alsakka and Gwilym, 2013; Giacomino, 2013; Doluca, 2014). The high market shares of more than 80% of the global top three markets, e.g., Standard & Poor's, Moody's Investor Service and Fitch, also draws attention to several studies, such as Asiri and Hubail (2014), Rowland (2004), Andritzky et al. (2005), Mora (2005), Vij (2005), Cruces (2006), Host et al. (2012), Kiff et al. (2012), Frascaroli and Oliveira (2017) and Malliaropulos and Migiakis (2018). Thus, from pioneering studies such as Cosset and Roy (1991), Oral et al. (1992), Haque et al. (1998) and Bhatia (2002), for example, there are a variety of empirical studies on this theme, mainly investigating the impact of ratings, as well as the role of CRAs.

Artificial Intelligence (AI), for instance, may be an important empirical strategy to be used to learn how CRAs rank credit risk for countries, as already investigated in pioneering studies, such as Yim and Mitchell (2005), Bennell et al. (2006) and Frascaroli et al. (2009). The Machine Learning (ML) model used in this paper has five advantage points among models which descend from AI: it has accuracy, it also may be automated, it is fast, customizable and scalable (Haykin, 2009; Brink et al., 2017). The Supervised Machine Learning (SML)

algorithm is able to handle large data, so it is very useful when doing the analysis of ratings. In this sense, the objective of this paper is to investigate the classification of sovereign risk ratings of countries using big data, i.e., the maximum of available information.

Four steps are proposed as the empirical strategy of this paper: 1) construct the database from the observed sovereign ratings and the information provided by the World Development Indicators database (WDI) (Bank, 2016); 2) handle and process the database, choosing attributes manually and automatically, by using clustering and principal component analysis (PCA) for comparison; 3) Use the Random Forest (RF) model to train, learn, and predict sovereign risk ratings; 4) Estimate econometric truncated response models from clustered data to test statistical significance of a) the structural change of ratings caused by the 2008 American subprime crisis and b) the level of development of countries on sovereign ratings explanation.

Precision metrics were used to test the accuracy and robustness of the model, seeking to test the highest number of hypotheses and comparisons with other results. By employing data mining techniques, e.g., automated indicator selection, we test some attributes required for an efficient prediction of sovereign ratings Han et al. (2012). The implications of such empirical findings will allow us to understand which indicators are most closely related to the decisions of the CRAs (Singh and Dharmaraja, 2017). In addition, it is important to have more information available to make it easier for emerging countries to attract international investments through better sovereign ratings (Elkhoury, 2008).

The focus of our main hypothesis is testing i) if without any prior knowledge it is possible for ML to process big data and predict sovereign ratings classification; and ii) if the subprime crisis of 2008 and the level of development of countries are significant to understand the sovereign ratings. Our paper is divided into five sections, in addition to this brief section 1. In section 2, the literature of ratings is explored. In section 3, we present the model setup. The data sample and processing are described in section 4. The estimation procedures are presented in section 5 and empirical findings are discussed in section 6. Final considerations are drawn in section 7.

2 Literature

Considering that obtaining and analyzing data about debt issuers is costly, the CRAs could be key to reduce asymmetric information in global financial markets. These are independent companies, aware of public or private sector interests, and their profits come from charging debt issuers for rating their risk assets. CRAs send important signals to market participants, thus, their analysis must be credible and not lead to biased or unreliable ratings. The latter ones are potentially based on multiple factors which indicate the sovereign debt solvency of countries, however there is no clear pattern for ratings determinants or framework. They are characterized by judgments of internal and external factors, as well as their expectation (Seetharaman et al., 2017).

Meanwhile, the legal rules and regulations are substantially conditioned to CRAs, specifically, those members of the Nationally Recognized Statistical Rating Organizations (NRSROs). The decisions of portfolio managers, who represent institutional investors' funds, banks, receivables securitization, for example, are all limited to internal procedures and rules based on ratings classification. Firstly investigated by Cosset and Roy (1991) and Cantor and Packer (1996), sovereign ratings are surrounded by answers and puzzles, barriers to entry and lack of competition, as well as conflicts of interest, transparency and accountability (Elkhoury, 2008). Since pioneer contributions, the literature of empirical evidence covers multiple countries and distinct periods of time Bennell et al. (2006) (1989-1999), Mora (2005) (1989-1999), E Bissoondoyal-Bheenick (2005) (1995–1999), Arezki et al. (2011) (2007–2010), Giacomino (2013) (2001-2011), Basu et al. (2013), Beirne and Fratzscher (2013) and Fatnassi et al. (2014) (2008–2012).

Some literature indicate that despite the credibility and importance of CRAs 1) their inability to predict economic crises, such as the 1997 Asian crisis (Ferri and Stiglitz, 1999) and the 2008 subprime crisis, is

undeniable 2) there is the risk of moral hazard (Božović et al., 2011), 3) presence of spillover effects (Alsakka and Gwilym, 2013), and 4) the ratings are biased (Doluca, 2014), 5) procyclical (Ferri and Stiglitz, 1999; Giacomino, 2013) and 6) opaque to regulators (Sy, 2009; Utzig, 2010). The empirical evidence also supports that there are structural breaks in ratings (corporative and sovereign) provided by CRAs from risk perspective (Basu et al., 2013), global perspective (Afonso, 2003; Frenkel et al., 2004; Host et al., 2012; Maltritz and Berlemann, 2013; Malliaropulos and Migiakis, 2018) and also from the point of view of the emerging markets (EMEs) (Frascaroli et al., 2009; Amstad and Packer, 2015; Frascaroli and Oliveira, 2017).

A new chapter was recently added to this literature, which points out the European Union (EU) debt crisis and their consequences to ratings and risk contagion (Afonso et al., 2012; Host et al., 2012; Beirne and Fratzscher, 2013; Baum et al., 2016; Reusens and Croux, 2016). As noted by Checherita and Rother (2010) and Lane (2012) to the EU, government budget deficits and changes in debt ratios are found to be linearly and negatively associated with economic growth activity. Following studies which have been used to investigate sovereign ratings (Yim and Mitchell, 2005; Bennell et al., 2006; Frascaroli et al., 2009), our contribution is to predict and understand some of their patterns using big data. For this, it will be necessary to reduce the dimensionality of variables through some data mining techniques, such as PCA (Johnson and Wichern, 2007) and clustering algorithms as Simple K Means, to pre-select those most important to predict ratings.

In addition, these algorithms will also be useful for identifying important attributes regarding the patterns of sovereign ratings provided by the CRAs across country categories. These data mining techniques consist of SML's detection of groups of potentially useful input examples to make their learning process more efficient (Lee and Monard, 2000). Hence, such datasets of variables, whose attributes were statistically important, after being processed, will be used as inputs to predict the sovereign ratings. Still in the direction of pursuing more robust results and contributing to the literature, it will also be tested whether the 2008 subprime crisis brings structural break in ratings, as well as if they are linked to the level of development of countries. For these goals, truncated response models will be estimated as well.

3 Empirical strategy

3.1 Principal component analysis (PCA)

Excessively large dimension datasets may lead to a low robustness of the predictions of the model, as well as make computational processing very intensive (Han et al., 2012). Therefore, it may be pertinent to perform a dimensionality reduction, i.e., the extraction of the most important variables to understand the ratings. The PCA are among several important attribute extraction methods described in the data mining literature. Also known as the Karhunen-Loève transform, it was first described by Pearson (1901). It consists of transforming a set of original variables into another set of variables of the same dimension called principal components.

The variables in the new dataset are linear combinations of all original variables, independent of each other, and carry the maximum information in terms of the total variation contained in the data. The main components are linear combinations of p random variables $X_1, X_2, ..., X_p$ (Johnson and Wichern, 2007). The main components represent a new coordinate system, obtained by a rotation of the original system. This new system provides the directions of maximum variability and provides a simpler and more efficient description of the covariance structure of the data.

The transformation is defined by a set of p-dimensional vectors $w_{(k)} = (w_1, ..., w_p)_{(k)}$ which maps each line of vector $X_{(i)}$ in a new principal component vector $t_{(i)} = (t_1, ..., t_m)_{(i)}$. It is given by $t_{k(i)} = X_{(i)} \cdot W_{(k)}$ for every $i = 1, ..., n$ and $k = 1, ..., n$ in such way that the individual variables t considered on the dataset successively inherit the maximum possible variance of X with each weight vector w, constrained to be a unit vector. To maximize the variance, the first vector of weights w, also called the first component, must satisfy:

$$w_{(1)} = \underset{\|w\|=1}{\arg\max} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \underset{\|w\|=1}{\arg\max} \left\{ \sum_i \left(x_{(i)} \cdot w \right)^2 \right\} \tag{1}$$

Writing in matrix form:

$$w_{(1)} = \underset{\|W\|=1}{\arg\max} \{ \|Xw\|^2 \} = \underset{\|w\|=1}{\arg\max} \left\{ w^T X^T X w \right\} \tag{2}$$

since $w_{(1)}$ has been defined by a unit vector, then satisfies the equivalence:

$$w_{(1)} = \arg\max \left\{ \frac{w^T X^T X w}{w^T w} \right\} \tag{3}$$

A standard result for a symmetric matrix such as $X^T X$ is that this maximum quotient value is the largest eigenvalue of the matrix, which occurs when w is the corresponding eigenvector. With $w_{(1)}$ obtained, the first component of a vector of observations $x_{(i)}$ may be given as a score $t_{1(i)} = X_i \cdot w_{(1)}$ in the transformed coordinates or as the corresponding vector in the original variables, $\left\{ X_i \cdot w_{(1)} \right\} \cdot w_{(1)}$. The k-th component may be found by subtracting the first $k-1$ main major components from X.

$$\hat{X}_k = X - \sum_{s=1}^{k-1} X w_{(s)} w_{(s)}^{\mathrm{T}} \tag{4}$$

Then we find the vector of weights that extracts the maximum variance from this new matrix of observations:

$$w_{(k)} = \underset{\|w\|=1}{\arg\max} \left\{ \|\hat{X}_k w\|^2 \right\} = \arg\max \left\{ \frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w} \right\} \tag{5}$$

The result is that the weight vectors are eigenvectors $X^T X$. Thus, the k-th component of a vector of observations $x_{(i)}$ may therefore be given as a score $t_{k(i)} = X_i \cdot w_{(k)}$ in the transformed coordinates, or as the corresponding vector in the space of the original variables $\left\{ X_i \cdot w_{(k)} \right\} \cdot w_{(k)}$, where $w_{(k)}$ is the k-th eigenvector of $X^T X$. The total decomposition of the main components of X may be written as:

$$T = XW \tag{6}$$

where W is a matrix p x p whose columns are the eigenvectors of $X^T X$. The transposition of W is called whitening or transformation. Furthermore, $X^T X$ may be recognized as proportional to the covariance matrix of the empirical sample of the X dataset. The covariance of the sample Q between two of the main components of the data set is given by:

$$\begin{aligned} Q(PC_{(j)}, PC_{(k)}) &\propto (X w_{(j)})^T (X w_{(k)}) \\ &= w_{(j)}^T X^T X w_{(k)} \\ &= w_{(j)}^T \lambda_{(k)} w_{(k)} \\ &= \lambda_{(k)} w_{(j)}^T w_{(k)} \end{aligned} \tag{7}$$

where the eigenvalue property of $w_{(k)}$ was used to move between the lines. However, the eigenvectors $w_{(i)}$ e $w_{(k)}$ corresponding to the eigenvalues of a symmetric matrix are orthogonal (if the eigenvalues are different); or they may be orthogonalized (if the vectors share a same repeated value). Hence, there is sample covariance between different major components throughout the dataset. Another way to characterize the transformation of the main

components is the transformation of the coordinates that diagonalize the covariance matrix of the sample. The empirical covariance matrix for the original variables may be written as:

$$Q \propto X^T X = W \Lambda W^T," \tag{8}$$

The empirical covariance matrix between the main components becomes:

$$W^T Q W \propto W^T W \, \Lambda \, W^T W = \Lambda \tag{9}$$

where Λ is the diagonal of the eigenvalue matrix Λ_k of $X^T X$. The transformation $T = Xw$ maps a data vector $w_{(i)}$ from an original space of variables p to a new space of variables $p*$ which are not correlated throughout the dataset. However, not all major components need to be sustained. Keeping only the first $p*$ principal components, produced using only the first vectors of weights, provide the following truncated transformation:

$$T_L = X W_L \tag{10}$$

where T_L is the array now with nxL. In other words, the PCA consists of a linear transformation $t = W^T x, x \in R^p, t \in R^L$, where the columns of p x L of matrix w form an orthogonal basis for the L attributes (the components of the representation t) that are not correlated (Bengio et al., 2013). By constructing i) all data matrices transformed with only L columns, this scoring matrix maximizes the variance in the original data that has been preserved, while ii) minimizes the total squared reconstruction error:

$$\| T W^T - T_L W_L^T \|_2^2 \tag{11}$$

3.2 Correlation matrix

Although PCA is a robust technique and widely used in many types of data problems, it may lose strength in specific cases. Therefore, due to the fact that the database has many attributes with missing value, the matrix of correlations and variance-covariance was estimated. Differently from the ML case, the imputation of values using the mean, or median, is not indicated, as it would bias the variance and is not useful for statistical inference. In addition, inherent to computational processing, there may be destruction of patterns in the original variables, making impossible a post-prediction analysis on these variables and their impact on the model and problem studied.

Hence, in contrast of the use of PCA, the Correlation-Based Feature Selection (CFS), described for the first time in Hall (2000), was also used to reduce the number of attributes of the database for comparison. This algorithm consists of the selection of variables based on the correlation between them, being the central hypothesis with which they are correlated with their categories, and not with each other. The operation of the algorithm has two steps: 1) correlation analysis between variables and categories; and 2) the search for subsets compatible with the central hypothesis. The correlation analysis follows the central hypothesis formalized through the relation (12), which calculates the merit of a subset of variables $k \in S$.

$$\text{Merit}_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}. \tag{12}$$

where $k\overline{r_{cf}}$ is the mean value of all correlations of rank variables, and $\overline{r_{ff}}$ is the mean value of all correlations between the variables in the dataset. The CFS is defined by the equation (13):

$$\text{CFS} = \max_{S_k} \left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1 f_2} + \cdots + r_{f_i f_j} + \cdots + r_{f_k f_1})}} \right]. \tag{13}$$

This equation may be rewritten as a mixed integer linear programming problem which is able to be solved by branch and bound algorithms:

$$\text{CFS} = \max_{x \in \{0,1\}^n} \left[\frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n} x_i + \sum_{i \neq j} 2 b_{ij} x_i x_j} \right]. \tag{14}$$

3.3 Supervised machine learning

After preprocessing and extracting variables, the data could be used in the ML model. In SML, one has the figure of an external teacher, represented through a set of inputs and outputs (Haykin, 2009). The machine will learn from a set of previously labeled examples, called a training set, and make predictions of sovereign ratings. Each example of the training set consists of a set of World Bank development indicators and sovereign ratings, called a label.

Given a training set with input and output examples $(x_1, y_1), ..., (x_n, y_n)$, where each output y_i was generated by an unknown function $y = f(x)$, process gets a function h which approximates the true function f (Mitchell, 1997; Russell and Norvig, 2010). Since the output y consists of a finite set of sovereign ratings, the learning problem will be classification, the chosen model being determined by its fit, i.e., originated from the algorithm's accuracy. Once defined and validated, the model will be used for ratings forecast and to test controls and find evidence on ratings.

3.3.1 Random forest

Random Forest is one of the parametric methods of SML which consists of a combination of decision trees from the random selection of attributes. Described initially by Ho (1995), it is an extension of the bagging technique (Breiman, 2013). In this model, each decision is composed from subsets of the training set, formed from the recomposition of samples of the original set. Each of these sets is created by bootstrapping Han et al. (2012). Therefore, each time a sample is selected, it is equally likely to be re-selected and added to the training set. Since the samples have the same characteristics as the original set distribution, Breiman (2013) demonstrates that this method brings substantial gains in accuracy in the classification process.

Since the basic constituents of the sets are tree-based predictors, and since each of these trees is constructed using random resampling, they are called RF (Biau, 2012). A RF consists of a collection of classifiers structured in trees $\{h(x, \Theta_k, k = 1...\}$, where $\{\Theta_k\}$ are vectors independent and identically distributed (i.i.d.) and each tree posts a voting unit to the most popular category in the x entry. Assuming that the training set is represented by T and even though there are M attributes, being $T = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, X_i is the input vector $x_{i1}, x_{i2}, ..., x_i n$ and y_i is the label or category of the instance.

Suppose that the forest has S , then new $S*$ randomized datasets with the same size as the original set will be created. This will result in $\{T_1, T_2, ..., T_s\}$ dataset. Each of these sets is called a bootstrap. In this path, each data set T_i may be duplicated or it may differentiate when compared to the original set. This process is called bagging. Therefore the RF will create $S*$ trees and will use $m = floor(lnM + 1)$ random sub-attributes of the possible M, to create each tree. This process is called the Random Subspace Method and its accuracy is highly dependent on the choice of these parameters.

Among the parameters, the size of the bag indicates the percentage referred to the size of the original set of data, created in its random rescheduling. The number of attributes indicates how many of them will be used to create each tree. Because it is an iterative process, it is also possible to control the maximum number of iterations of the algorithm. It is also possible to control the number of samples to be displayed before an update is performed through the batch size batch size, i.e., the batch size controls how many predictions will be made each time. Final predictions are obtained by aggregating the generated sets. The advantages of the RF method

are its relatively low computational cost, an indispensable feature when working with large amounts of data, in addition to avoiding overfitting and being less sensitive to noise Breiman (2013).

3.3.2 Confusion matrix

The validation step of the model is to evaluate its accuracy as a classifier. In this sense, it is possible to elaborate a confusion matrix, providing even more information about the model accuracy. The confusion matrix is a useful tool for analyzing how well a classifier may recognize tuples from different categories (Han et al., 2012). Being a matrix of m categories, a confusion matrix is a sequence of dimension m x m. An input, $CM_{i,j}$ in the first m rows and m columns indicates the number of tuples in the i category labeled by the classifier as i category. In order for a classifier to be accurate, most tuples must be represented along the diagonal of the confusion matrix, from the input $CM_{1,1}$ to the input $CM_{m,m}$, with the remainder of the inputs near zero. The sequence may have additional rows or columns to provide totals or recognition rates by category.

Through the confusion matrix it is possible to extract the metrics of sensibility and specificity. Sensitivity is also referred to as positive true rate, indicating the proportion of positive tuples correctly identified, while specificity is the true negative rate, indicating the proportion of negative tuples that are correctly identified. In addition, precision may be defined as the percentage of tuples labeled C_i which are correctly categorized as C_i. These metrics are defined in the equations:

$$sensibility = \frac{Tpositive}{positive} \tag{15}$$

$$Specificity = \frac{Tnegative}{negative} \tag{16}$$

Finally, it is possible to define accuracy in terms of sensitivity and specificity:

$$Accuracy = sensibility\frac{positive}{positive + negative} + specificity\frac{positive}{positive + negative} \tag{17}$$

From the values of sensibility and specificity it is possible to obtain the Receiver Operating Characteristic (ROC) and the F1-Score. The ROC curve aims to compare classification models for different thresholds for classification. F1 score is a harmonic mean between precision and sensitivity. It is especially used in models with disproportionate categories and it does not issue probabilities. To ensure that the accuracy values of a classifier are a reliable estimate, some evaluation techniques are used, such as the Holdout Method, Random Subsampling, and k-Fold Cross-Validation Han et al. (2012). Let a confusion matrix where sensitivity is indicated in the column, i.e., true positive rate but the false positive rate (1-specificity) is actually indicated on the line. Each cell in the matrix represents the respective values obtained.

		Predicted class	
		C_1	C_2
Current class	C_1	true positives	false negatives
	C_2	false positives	true negatives

3.3.3 Cross-validation k-group

K-fold cross validation is one of several methods used to measure the performance of a ML. The k-group cross validation technique begins by randomly dividing the data set into k disjoint subsets, with the value $k = 10$ being commonly used. For each cluster, or group, a model is trained with the total dataset, except the data from that group. After all groups are traversed in learning, the predictions for each group are aggregated and compared

with the actual variable to be predicted, thus evaluating the predictions Brink et al. (2017). Cross-validation k-group is also used for selecting ML methods or tuning specific parameters, where the one with the best performance is chosen (Shalev-Shwartz and Ben-David, 2014).

3.4 Econometric tests

In addition to the use of ML techniques, regression models with linear and binary functional forms were also estimated to answer some questions: i) whether the 2008 subprime crisis affected the sovereign ratings, and ii) whether the country's level of development is significant to explain sovereign ratings. To establish a comparison with the predictive accuracy of the model, we test if the variable sovereign ratings can be explained by the proposed variables. In this case, we took initially 1,334 variables, and after the analysis of correlation and variance, the matrix was reduced to 234 variables.

In this step, we estimated econometric models considering linear and truncated dependent variable 1) with a dummy to represent 2008 crisis and 2) another model with a categorical variable to represent the level of development of each sampled country. The statistical significance of these parameters will be verified to check if there is any evidence that the subprime crisis and the level of development of the countries affect the sovereign ratings.

4 Sample design and data preprocessing

4.1 Sample design

The sample is composed of three sources of data. First, we used the long-term foreign currency ratings assigned by Fitch, Moody's, Standard & Poors. Overall, 118 countries are considered since their very first appearance, e.g., Norway and France (1975), United Kingdom (1978), Netherlands and Belgium (1988), Italy and Finland (1994), Poland (1995), etc. To get access to a more complete historical series of ratings, it was used a web scraping script to extract and cluster this data automatically from web pages. Therefore, the unbalanced panel dataset used has 3,596 instances, with historical data from 1958 to 2017, i.e., 38 distinct years. Second, we obtained the WDI database. Third, the database containing the degree of economic development was obtained through the World Economic Situation and Prospects (WESP) report.

Next, we collapsed the three mentioned databases to build our panel. Then, we collapsed the three published databases to build our panel. The resulting pre-processing database to be used in the experiment is summarized in Table 1 and has 137 distinct countries, with about 57.94 % of the countries classified as developing economies (category C), 31.77% as countries with fully-developed economies (category A) and 10.28% as economies in transition (category B).

Table 1: General countries description according their development level

	Country	A	B	C
	Andorra	0	0	0
	United Arab Emirates	0	0	1
	Albania	0	1	0
	⋮	⋮	⋮	⋮
	Vietnam	0	0	1
	South Africa	0	0	1
Total	137	34	11	62

Source: World Bank.

4.2 Data preprocessing

The dataset of sovereign ratings and other indicators was processed using the software Waikato Environment for Knowledge Analysis (Weka). First, we collapsed the dataset of ratings issued by the three agencies, then we used the linearization to obtain a scale from 1 to 22 observed ratings. This means that countries located in the scale from 1 to 10 belong to the group of investment grade. As a complement, those in the scale above 10 are considered to be speculative, i.e., countries of greater risk, as summarized in Table 2:

Table 2: Ratings system

	Companies			
Classification	Moody's	S&P	Fitch	Scale
	Aaa	AAA	AAA	1
	Aa1	AA+	AA+	2
	Aa2	AA	AA	3
	Aa3	AA-	AA-	4
Investment grade	A1	A+	A+	5
	A2	A	A	6
	A3	A-	A-	7
	Baa1	BBB+	BBB+	8
	Baa2	BBB	BBB	9
	Baa3	BBB-	BBB-	10
	Ba1	BB+	BB+	11
	Ba2	BB	BB	12
	Ba3	BB-	BB-	13
	B1	B+	B+	14
	B2	B	B	15
Speculative	B3	B-	B-	16
	Caa1	CCC+	CCC+	17
	Caa2	CCC	CCC	18
	Caa3	CCC-	CCC-	19
	-	CC	CC	20
	-	C	C	21
	Ca	SD	DDD	22
	C	D	DD	
	-	-	D	

Source: Fitch, Moody's and Standard & Poors.

The application of preprocessing techniques enables obtaining more robust results, thus eliminating inconsistencies and labeling problems. The identification and attenuation of noise represented by distorted data was also treated. The resulting dataset has many values for the attributes considered for each instance. In this way, the treatment for missing values and the removal of duplicate and redundant attributes were also done. For some countries where there is no information, the missing values were input using the data median or the data average inputs.

We also observed a significative number of attributes with different scales, a situation that could negatively affect the robustness of the model. Typically, ML algorithms are most effective if the input attributes are on the same scale. Hence, we used standardization techniques and centralization of database attributes. The application of these two techniques together resulted in a standardized dataset, i.e., a uniform distribution $N \sim (0,1)$.

5 Estimation

The original database had many redundant attributes which can negatively affect the fit of the model. Since it culminates in a great heterogeneity among the observations, it decreases the accuracy of prediction and increases the computational cost. Therefore, it is necessary to apply attributes selection techniques that have more correlation with the estimation of sovereign ratings. These data mining techniques consist in finding subsets of attributes on which the ML algorithm will focus. In this way, after the PCA processing, the database size was reduced from 3,489 instances and 644 attributes to 1,597 instances and 191 attributes, respectively.

This step using the preprocessed database of selected attributes is divided into 1) training set, 2) test set, and 3) validation set. In this paper, the following proportions we used: 80% of data for training and 20% for tests. As the RF method has some adjustment parameters and their choices strongly affect the prediction accuracy, it was necessary to estimate empirical tests to avoid doubts and obtain robust models. For this experiment, the optimal parameters for the RF algorithm are shown in Table 3.

Table 3: Parameters that obtained better accuracy

Bag Size %	Batch Size	Max Depth	Num Features	Num Iterations
100	100	Max	Max	150

Source: Fitch, Moody's, Standard & Poors and WDI.

The Bag Size Percent (%) is the size of each bag as a percentage of the size of the training set; Batch Size is the desired batch size for batch forecasting; Max Depth is the maximum depth of the tree, and Num Features is the number of features used in the random selection. The clustering technique was also applied to the categories of the problem. The goal is to decrease the number of categories by clustering similar categories or ratings in the same clusters and have an eventual increase in accuracy.

6 Empirical findings

First, we report the results for the PCA and second, the clustering process. Next, the results of econometric tests are displayed and discussed. The other metrics, e.g., the concordance correlation coefficient (KAPPA), the mean absolute error (MAE), the root mean squared error (RMSE), the relative absolute error (RAE) and the root relative squared error (RRSE), are shown in Table 4 for the case of PCA, manual selection and clustering model, respectively.

Table 4: Cross-validation k-group by model

	Kappa	MAE	RMSE	RAE	RRSE
PCA	0.7705	0.033	0.1191	38.545 %	57.75 %
Manual	0.8654	0.0838	0.1945	24.80 %	47.3009 %
Clustering	0.9764	0.0309	0.1014	8.47 %	23.77 %

Source: Fitch, Moody's, Standard & Poors and WDI.

The PCA classifier was constructed from 1,597 instances, containing 191 artificial attributes generated, with accuracy predicting sovereign ratings of 78.52% with the use of cross validation with k = 10. Attempting to overcome this problem, the number of categories was reduced, and a combination of manual categories of similarity was performed. This procedure resulted in a decrease of 24 to only 4 categories, with a 90.91% accuracy in the prediction. The reduction of categories used the algorithm of clustering Simple K Means, also

resulting in 4 categories. There was an increase in the accuracy of the classifier to 98.28%. In Table 5 the statistics of categories $C_1, ..., C_4$ of the manual and the clustering models are displayed.

Table 5: Accuracy using manual selection and clustering by categories

		TP	FP	PPV	TPR	F1	ROC	Class
Manual		0.943	0.027	0.953	0.943	0.948	0.989	A
		0.869	0.042	0.869	0.869	0.869	0.971	B
		0.923	0.054	0.903	0.923	0.913	0.976	C
		0.651	0.007	0.75	0.651	0.697	0.935	D
	$\bar{x}$	0.909	0.04	0.909	0.909	0.909	0.978	
Clustering		0.995	0.003	0.99	0.995	0.992	1	C_1
		0.975	0.006	0.968	0.975	0.971	0.997	C_2
		0.981	0.005	0.987	0.981	0.984	0.998	C_3
		0.98	0.01	0.982	0.98	0.981	0.998	C_4
	$\bar{x}$	0.983	0.006	0.983	0.983	0.983	0.998	

Source: Fitch, Moody's, Standard & Poors and WDI.

All error metrics for the clustering model indicate improved values from the manual selection model. The categories $C_1 = C_A, ..., C_4 = C_D$ were created by the model and indicate that the ratings are better clustered and, therefore, better predicted. In Table 6 is the confusion matrix for the model with manual selection and clustering, respectively. The errors increase for clusters with lower ranks for both methods.

Table 6: Confusion matrix according manual and clustering selection

	a	b	c	d	Class
Manual	1229	53	21	0	C_A
	47	741	65	0	C_B
	13	58	1133	23	C_C
	0	1	36	69	C_D
Clustering	760	0	0	4	C_1
	0	509	5	8	C_2
	0	9	978	10	C_3
	8	8	8	1182	C_4

Source: Fitch, Moody's, Standard & Poors and WDI.

The metrics point out that the classifier has efficient performance when the number of categories is reduced. Specifically, the clustering techniques of the categories resulted in an increase in accuracy of classification of at least 12.36%. However, it was observed some loss of sensitivity. In the validation stage of the model, where the confusion matrix was used, it was revealed that Austria, Australia, Belgium, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, United Kingdom, Ireland, Italy, Japan, Luxembourg, Netherlands, Republic of Korea, Sweden and United States, in alphabetical order, comprehend category A.

Category B comprises Albania, Bosnia and Herzegovina, Brazil, Grenada, Greece, Lebanon, Philippines, Romania, Serbia, El Salvador and Venezuela. Meanwhile, Armenia, Angola, Argentina, Barbados, Bangladesh, Burkina Faso, Bulgaria, Benin, Bolivia, Botswana, Belarus, Belize, Cote d'Ivoire, Chile, Cameroon, Costa Rica, Cabo Verde, Democratic Republic of Congo, Dominican Republic, Ecuador, Egypt, Arab Republic, Ethiopia, Fiji, Gabon, Ghana, Gambia, Guatemala, Honduras, Indonesia, India, Iraq, Jamaica, Jordan, Kenya, Cambodia, Kazakhstan, Lesotho, Morocco, Moldova, Montenegro, Mali, Mongolia, Malawi, Mozambique, Nigeria, Nicaragua, Peru, Papua New Guinea, Pakistan, Paraguay, Republic of North Macedonia, Russian

Federation, Senegal, The Bahamas, Turkmenistan, Tunisia, Trinidad and Tobago, Ukraine, Uganda, Uruguay, Vietnam and Zambia, among other countries comprehend category C.

Category D comprises Andorra, Azerbaijan, Bahrain, Chile, China, Colombia, Cyprus, Czech Republic, Estonia, Georgia, Hong Kong Special Administrative Region, China, Croatia, Hungary, Iceland, Israel, Islamic Republic of Iran, Kuwait, Liechtenstein, Sri Lanka, Lithuania, Latvia, Libya, Malta, Mexico, Malaysia, New Caledonia, Norway, Oman, Panama, Poland, Qatar, Rwanda, Saudi Arabia, Seychelles, Singapore, Slovenia, Slovak Republic, San Marino, South Africa, Suriname, Thailand, Turkey and United Arab Emirates. These results from 137 countries indicate that they have similarities but may not be directly verified.

In this sense, we first noted that category A comprises those countries with the lowest sovereign ratings indicated in Table 2. It was also verified that the algorithm clustered two categories with greater number of countries of different continents, as D, with great heterogeneity among them. In addition, the model points to category C countries with lower GDP per capita. In category B, were selected some countries which had, as default, a common characteristic in their historical data. What is remarkable in this category is the presence of countries whose economies currently deal with problems related to public debt and crisis, such as Greece and Brazil, for example.

Inevitably, structural deficits and increases in debt/GDP ratio led to a credit ratings crisis in some of the studied countries. In order to properly understand what does it means, Figure 1 highlights some selected sovereign ratings' trajectories. They indicate the linearized sovereign ratings from 2000 to 2016, where the vertical line points to the 2008 subprime crisis.

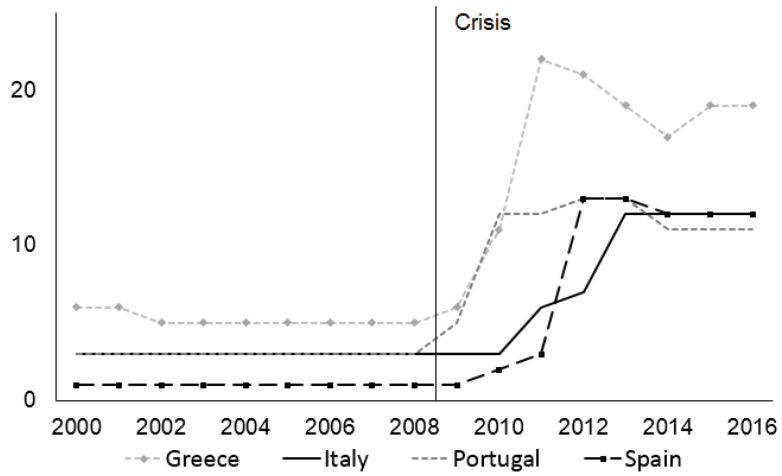

Figure 1: Trajectories of sovereign credit risk of selected countries
Source: Standard & Poors, Fitch and Moodys.

The figure points out that after the crisis, the observed trajectories for the EU case changed a lot, indicating significant increases, as also reported by (Lane, 2012). Furthermore, the Greek crisis made agencies consider Greek bonds as speculative position, e.g., $BBB - (Baa3) = 10$ in linearized scale. In order to test some hypotheses and to obtain additional empirical evidence, we also used econometric models using clustering data selection. First, the results of the model to test the subprime crisis are summarized in the Table 7:

Table 7: Estimated regression models considering 2008 subprime crisis

X_i	β	σ	t	P-value
X_1	0.0131	0.001	10.252	0.000
X_2	0.0012	0.006	0.219	0.827
X_3	-0.0250	0.011	-2.261	0.024
X_4	-0.0157	0.008	-2.029	0.042
⋮	⋮	⋮	⋮	⋮
X_{490}	0.0017	0.008	0.222	0.825
D	0.0384	0.011	3.611	0.000
R^2	0.915	$\hat{R}^2$	0.913	F Test 338.7

Source: Fitch, Moody's, Standard & Poors and WDI.

Alternatively to analyze the marginal effects of the parameters directly, since big data is used, we focus on fit measures R^2 and R^2 adjusted. They are 0.915 and 0.913, respectively, while for the linear model, were 0.914 and 0.912. Estimated for comparative terms, F-Test was 334.3. The p-value (< 0.05) reveals that the dummy variable (D) adopted to test the statistical significance of the subprime crisis is statistically significant. Therefore, our empirical evidence suggests that there is structural change of sovereign ratings before and after the 2008, confirming our hypothesis.

This outcome is also pointed by Basu et al. (2013) and Amstad and Packer (2015), which means their inability to signal risk increase. Another hypothesis raised in literature is that the classification of sovereign ratings is influenced by the level of development of the countries. Therefore, we estimated a regression model with a truncated variable with a categorical variable to find empirical evidence of this phenomenon. Table 8 summarizes the results of the estimated regression:

Table 8: Estimated regression models considering 2008 subprime crisis level of development of countries

X_i	β	σ	t	P-value
X_1	0.0131	0.001	10.252	0.000
X_2	0.0012	0.006	0.219	0.827
X_3	-0.0250	0.011	-2.261	0.024
X_4	-0.0157	0.008	-2.029	0.042
⋮	⋮	⋮	⋮	⋮
X_{490}	0.0008	0.008	0.101	0.919
C_A	-0.0236	0.006	-3.871	0.000
C_B	0.0088	0.002	5.028	0.000
C_C	-0.0093	0.006	-1.474	0.140
R^2	0.916	$\hat{R}^2$	0.913	F Test 338.7

Source: Fitch, Moody's, Standard & Poors and WDI.

The categorical variables C_A, C_B and C_C show the developed economies, in transition economies and in those considered in development process, respectively. While p-value shows that the C_C category is not statistically significant, the categories C_A and C_B are statistically significant. The provided outcomes suggest that there is statistical evidence that CRAs evaluate countries in transition and those considered developed in quite a different way. A possible explanation for this distinct risk evaluation is also the result of the episode of the 2008 subprime crisis, as displayed by Figure 1 in the case of selected countries of the EU. This last result may also be verified through 9. In this case, there is a considerable difference between the mean and the standard deviation, which indicates some volatility of ratings to developed economies (C_A).

Table 9: Additional statistics for developed countries

A	Sum	Average	Stand. Deviat.
0	2086	10.52	4.514
1	1403	6.374	4.431

Source: Fitch, Moody's, Standard & Poors and WDI.

The hypotheses favoring the differences in the sovereign ratings issued by the CRAs, as reported in the literature (Ferri and Stiglitz, 1999; Božović et al., 2011; Host et al., 2012; Alsakka and Gwilym, 2013; Basu et al., 2013; Giacomino, 2013; Maltritz and Berlemann, 2013; Doluca, 2014; Malliaropulos and Migiakis, 2018). The exception of certain cases such as those referring to Category B, e.g., Albania, Bosnia and Herzegovina, Brazil, Grenada, Greece, Lebanon, Philippines, Romania, Serbia, El Salvador and Venezuela. These countries need more research to subsidize policies for minimizing how the risks of these economies are perceived in the international financial market. In this sense, the confusion matrix is a useful tool to analyze how well a classifier may recognize tuples of different categories.

The dimensionality reduction improves the fit of the proposed models, indicating to be useful to process datasets containing high number of dimensions. The model retains as much of data variance as possible, i.e., the most important information for the analysis. However, beyond the reported problems, it is possible to use these models to optimize portfolios and decisions, as noted in Singh and Dharmaraja (2017). Finally, in contrast to more than 98.28% accuracy, we also reported some difficulty to fit the used data to sovereign ratings. This is an intrinsic characteristic verified through the good performance of prediction when a reduced number of categories is considered Yim and Mitchell (2005); Bennell et al. (2006); Frascaroli et al. (2009).

7 Final considerations

The robust prediction of sovereign credit risk ratings is important for many types of economic decisions, among the emerging countries that look to attract international investors. In this paper, we proposed an empirical strategy based on big data, divided into four stages: 1) construct a database; 2) preprocess the dataset; 3) use the Random Forest model to predict the sovereign risk ratings; 4) estimate truncated response econometric models to test the change caused by the subprime crisis and the level of development of countries on sovereign ratings.

The main finding of this paper is that it adds new insights for policymakers and international financial market participants. The main message is that is possible to train AI models using big data as input to robust prediction of ratings classification, without any prior knowledge. We also found some common problems of methodological limitations when 1) using the same indicators for all countries sampled, without controlling them, or without allowing them to have different weights for the countries, and 2) without considering properly the endogeneity between some variables.

In contrast to other research where estimators are previously selected and functional forms are assumed to be linear, the adjustment to the used data does not require more rigid hypotheses. However, to get access to in-depth empirical outcomes with some economic intuition, some econometric models of truncated responses with clustered datasets were used. This heavy focus on modeling outcomes confirmed that CRAs are severely exposed to puzzles. Among them, we have found that ratings issuing changed after the subprime crisis and that the level of development of the countries is important to drive CRAs judgments.

This empirical evidence suggests the CRAs inability to predict crises, which is very critical for the proper functioning of international financial markets. Despite the results obtained, future research could further reduce the size of the database, without losing information, in order to seek better visualization of the most important attributes to predict ratings. The large number of attributes, even with the use of dimensionality reduction

techniques, is still very large, especially in the context of using regressions for testing. This important aspect increases the costs of manipulating and analyzing these attributes.

In addition, it is possible to automate the generation of new databases, eliminating the manual work of data acquisition and pre-processing. It could be very useful to new analysis, e.g., using other classifications of development of economies, etc. Finally, the reliance on unconventional economic policies, mainly monetary and fiscal policy practiced by some of the studied countries, are among the fragilities to get access to better sovereign ratings. This, in most part of the cases, reflects each country's current institutional architecture. Therefore, macroprudential tools need to be used mainly among those countries in the category of greater sovereign risk, forcefully and without delay.

References

Afonso, A. (2003). Understanding the determinants of government debt ratings: Evidence for the two leading agencies. *Journal of Economics and Finance 27*(1), 56–74.

Afonso, A., D. Furceri, and P. Gomes (2012). Sovereign credit ratings and financial markets linkages: Application to European data. *Journal of International Money and Finance 31*(3), 606–638.

Alsakka, R. and O. Gwilym (2013). Rating agencies' signals during the European sovereign debt crisis: Market impact and spillovers. *Journal of Economic Behavior & Organization 85*, 144– 162.

Amstad, M. and F. Packer (2015). Sovereign ratings of advanced and emerging economies after the crisis. *BIS Quarterly Review 2015*(December), 1–15.

Andritzky, J. R., G. J. Bannister, and N. T. Tamirisa (2005). The impact of macroeconomic announcements on emerging market bonds. In *IMF Working papers*, Volume 05/83.

Arezki, R., B. Candelon, and A. N. R. Sy (2011). Sovereign rating news and financial markets spillovers: evidence from the european debt crisis. In *IMF Working papers*, Volume 68.

Asiri, B. K. and R. A. Hubail (2014). An empirical analysis of country risk ratings. *Journal of Business Studies Quarterly 5*(4), 52–67.

Bank, W. (2016). Wdi highlights. Technical report, World Bank.

Basu, K., S. DE, D. Ratha, and H. Timmer (2013). Sovereign ratings in the post-crisis world: An analysis of actual, shadow and relative risk ratings. In *Policy Research Working Paper Series*, Volume 6641. World Bank.

Baum, C. F., D. Schäfer, and A. Stephan (2016). Credit rating agency downgrades and the Eurozone sovereign debt crises. *Journal of Financial Stability 24*(June), 117–131.

Beirne, J. and M. Fratzscher (2013). The pricing of sovereign risk and contagion during the European sovereign debt crisis. *Journal of International Money and Finance 34*(2), 60—-82.

Bengio, Y., A. Courville, and P. Vincent (2013). Representation learning: A review and new perspectives pattern analysis and machine intelligence. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(8), 1798–1828.

Bennell, J., D. Crabbe, S. Thomas, and O. Gwilym (2006). Modelling sovereign credit ratings: Neural networks versus ordered probit. *Expert Systems with Applications 30*(3), 415–425.

Bhatia, A. V. (2002). Sovereign credit ratings methodology: An evaluation. In *IMF Working Papers*, 02/170. IMF.

Biau, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research 13*, 1063–1095.

Božović, M., B. Urošević, and B. Živković (2011). Credit rating agencies and moral hazard. *Panoeconomicus 2*, 219–227.

Borensztein, E. A., K. B. Cowan, and P. Valenzuela (2013). Sovereign ceilings "lite"? the impact of sovereign ratings on corporate ratings. *Journal of Banking & Finance 37*(11), 4014–4024.

Breiman, L. (2013). Bagging predictors. *Machine Learning 24*, 123—140.

Brink, H., J. Richards, and M. Fetherolf (2017). *Real-word machine learning*. Manning.

Cantor, R. and F. Packer (1996). Determinants and impact of sovereign credit ratings. *Economic Policy Review 2*(2), 37–53.

Checherita, C. and P. Rother (2010). The impact of high and growing government debt on economic growth an empirical investigation for the Euro area. Technical report, European Central Bank.

Cosset, J. C. and J. Roy (1991). The determinants of country risk ratings. *Journal of International Business Studies 22*(1), 135–142.

Cruces, J. J. (2006). Statistical properties of sovereign credit ratings. *Emerging Markets Review 7*(1), 27–51.

Doluca, H. (2014). Is there a bias in sovereign ratings due to financial reasons? *The Empirical Economics Letters 13*(7), 801–814.

E Bissoondoyal-Bheenick (2005). An analysis of the determinants of sovereign ratings. *Global Finance Journal 15*, 251–280.

Elkhoury, M. (2008). Credit rating agencies and their potential impact on developing countries. In *Discussion Papers*, 186. United Nations Conference on Trade and Development.

Fatnassi, I., Z. Ftiti, and H. Hasnaoui (2014). Stock market reactions to sovereign credit rating changes: Evidence from four European countries. *Journal of Applied Business Research 30*(3), 953–958.

Ferri, G. and L.-G. L. J. E. Stiglitz (1999). The procyclical role of rating agencies: Evidence from the East Asian crisis. *Economic Notes 28*(3), 335–355.

Frascaroli, B. F. and J. C. T. Oliveira (2017). Sovereign risk ratings and macroeconomic fundamentals and accountability: Evidence from developing countries. *Advances in Scientific and Applied Accounting 10*(3), 304–318.

Frascaroli, B. F., L. C. Silva, and O. C. S. Filho (2009). Classificação de ratings de risco soberano de países emergentes a partir de fundamentos macroeconômicos utilizando redes neurais artificiais [sovereign risk ratings of emerging countries based on macroeconomic fundamentals using artificial neural networks]. *Brazilian Review of Finance 7*(1), 73–106.

Frenkel, M., A. Karmann, and B. Scholtens (2004). *Sovereign risk and financial crises*. Springer International Publishing: New York.

Giacomino, P. (2013). Are sovereign credit ratings pro-cyclical? A controversial issue revisited in light of the current financial crisis. *Rivista di Politica Economica 4*(October/December), 79–111.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings of the 17th International Conference on Machine Learning 23*, 359—-366.

Han, J., M. Kamber, and J. Pei (2012). *Data mining: Concepts and techniques* (3 ed.). Elsevier Inc./ Morgan Kaufmann.

Haque, N. U., M. Mark, and D. J. Mathieson (1998). The relative importance of political and economic variable in creditworthiness ratings. In *IMF working papers*, Number 46 in 98.

Haykin, S. (2009). *Neural networks and learning machines*. Prentice-Hall.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition, Volume 1, ICDAR '95*. Page 278. IEEE Xplore.

Host, A., I. Cvečić, and V. Zaninović (2012). Credit rating agencies and their impact on spreading the financial crisis on the Eurozone. *Ekonomska misao i praksa 21*, 639–662.

Johnson, R. A. and D. W. Wichern (2007). *Applied multivariate statistical analysis* (6th ed.). Prentice-Hall.

Kiff, J., S. B. Nowak, and L. Schumacher (2012). Are rating agencies powerful? An investigation into the impact and accuracy of sovereign ratings. In *IMF Working papers*, Volume 23. IMF.

Lane, P. R. (2012). The European sovereign debt crisis. *Journal of Economic Perspectives 26*(3), 49–68.

Lee, H. D. and M. C. Monard (2000). Applying knowledge-driven constructive induction: Some experimental results. Technical Report 101, Institute of Mathematics and Computer Science, University of São Paulo.

Malliaropulos, D. and P. M. Migiakis (2018). The re-pricing of sovereign risks following the Global Financial Crisis. *Journal of Empirical Finance 49*, 39–56.

Maltritz, D. and M. Berlemann (2013). *Financial crises, sovereign risk and the role of institutions*. Springer International Publishing Switzerland: Switzerland.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Moody's Investors Service (2016). Moody's rating symbols and definitions. Technical report, Moody's Inc.

Mora, N. (2005). Sovereign credit ratings: Guilty beyond reasonable doubt? *Journal of Banking & Finance 30*, 2041–2062.

Oral, M., O. Kettani, J.-C. Cosset, and M. Daouas (1992). An estimation model for country risk rating. *International Journal of Forecasting 8*(4), 583–593.

Partnoy, F., R. Levich, G. Majnoni, and C. M. Reinhart (2002). *The paradox of credit ratings*. Ratings, rating agencies and the global financial system. Kluwer Academic Press.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2*(11), 559–572.

Poor's, S. . (2011). Sovereign government rating methodology and assumptions. Technical report, Standard & Poor's.

Poor's, S. . (2015). Sovereign ratings history since 1975. Technical report, Standard & Poor's.

Reusens, P. and C. Croux (2016). Sovereign credit rating determinants: The impact of the European debt crisis. In *Working paper 1425*. Faculty of Economics and Business, KU Leuven.

Rowland, P. (2004). Determinants of spread, credit ratings and creditworthiness for emerging market sovereign debt: A follow-up study using pooled data analysis. In *Working papers of Banco de la Republica de Colombia*. Banco de la Republica de Colombia.

Russell, S. and P. Norvig (2010). *Artificial intelligence: A modern approach* (3 ed.). Prentice-Hall.

Seetharaman, A., V. K. Sahu, A. S. Saravanan, J. R. Raj, and I. Niranjan (2017). The impact of risk management in credit rating agencies. *Risks 5*(4), 52.

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Singh, A. and S. Dharmaraja (2017). A portfolio optimisation model for credit risky bonds with Markov model credit rating dynamics. *International Journal of Financial Markets and Derivatives 6*(2), 102–119.

Sy, A. N. (2009). The systemic regulation of credit rating agencies and rated markets. *World Economics Data Papers 10*(4), 69–108.

Utzig, S. (2010). The financial crisis and the regulation of credit rating agencies: A European banking perspective. Technical Report 188, Asian Development Bank Institute.

Vij, M. (2005). The determinants of country risk analysis. *Journal of Management Research 5*(1), 20–31.

Yim, J. and H. Mitchell (2005). Comparison of country risk models: Hybrid neural networks, logit models, discriminant analysis and cluster techniques. *Expert Systems with Applications 28*(1), 137–148.