SYRTO

Systemic Risk Tomography
*Signals, Measurements, Transmission channels and Policy Interventions*

# Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals

Roberto Savona, Marika Vezzoli

SYRTO WORKING PAPER SERIES
Working paper n. 4 | 2013

This project is funded by the European Union under the 7th Framework Programme (FP7-SSH/2007-2013) Grant Agreement n° 320270

This documents reflects only the author's view. The European Union is not liable for any use that may be made of the information contained therein

# Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals*

Roberto Savona† Marika Vezzoli‡

This draft: June 2013

## Abstract

In this paper, we try to realize the best compromise between in-sample goodness of fit and out-of-sample predictability of sovereign defaults. To do this, we use a new regression-tree based approach that signals impending sovereign debt crises whenever pre-selected indicators exceed specific thresholds. Using data from emerging markets and Greece, Ireland, Portugal, and Spain (GIPS) over the period 1975–2010, we show that our model significantly outperforms existing competing approaches (logit, stepwise logit, noise-to-signal ratio, and regression trees), while balancing in- and out-of-sample performance. Our results indicate that illiquidity (high short-term debt to reserves) and default history, together with real GDP growth and U.S. interest rates, are the main determinants of both emerging market country defaults and the recent European sovereign debt crisis.

**JEL classification numbers**: C53, F33, F34.

**Keywords**: Sovereign Debt Crisis; Data mining; Evaluating forecasts; Model selection; Probability forecasting.

*The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 320270 - SYRTO.

†*Department of Economics and Management, University of Brescia, c/da S. Chiara 50, 25122 Brescia, Italy, (e-mail: savona@eco.unibs.it)*

‡*Department of Molecular and Translational Medicine, University of Brescia, Viale Europa 11, 25123 Brescia, Italy, (e-mail: marika.vezzoli@med.unibs.it)*

# I. Introduction

The recent sovereign debt crisis in the Eurozone revived the debate on 'forecasting versus policy dilemma' introduced in Clements and Hendry (1998) and on the gap between models used for forecasting and models used for policy-making. Abundant empirical evidence proves that simple models are usually better than complex models in terms of forecast accuracy, but the latter provide a better description of past data. How should we combine the in-sample goodness of fit and out-of-sample predictability in the context of sovereign default? How should we evaluate model performance when jointly considering in- and out-of sample accuracy? Our objective is to give an answer to these questions by inspecting the sovereign defaults in emerging markets occurring between 1975 and 2010, and the recent Eurozone sovereign debt crisis.

The questions we face in this paper have achieved new relevance given the recent global financial crisis for different decision-maker categories. International investors, who are generally more focused on pure forecasting (i.e., expectation of risk/return profile), are showing different risk tolerance levels (from low to high risk aversion) depending on the increased sensitivity towards macroeconomic conditions after the Greek crisis[1]. Policy makers are concerned with realizing optimal early warning systems (EWSes) to provide risk signals with a sufficient lead time to implement adequate policy measures. In this perspective, first, it is preliminarily essential that stylized facts on crisis occurrence are well established based on past data; second, the EWSes should be conceived with the main objective of minimizing false alarms (type-II errors) while maintaining a high predictive ability of impending crises, rather than with the objective of controlling for missing defaults (type-I errors). The costs associated with false alarms are in fact potentially huge in terms of negative market sentiment, international reputation, contagion effects, and political interventions, which translates into a great concern towards type-II errors.

The literature on sovereign defaults is extensive in terms of early warning indicators and model specification. On the selection of best crisis predictors, the empirical evidence suggests that the probability of a debt crisis is positively correlated with higher levels of total (McFadden et al., 1985) and short-term debt (Detragiache and Spilimbergo, 2001), negatively correlated with GDP growth (Sturzenegger, 2004), and the level of international reserves (Dooley, 2000). Moreover, defaults are also related to more volatile and persistent output fluctuations (Catão and Sutton, 2002), less trade openness (Cavallo and Frankel, 2008), political conditions (Manasse *et al.*, 2003), previous history of de-

[1] De Grauwe and Ji (2012) found evidence that a large part of the surge in the government bond spreads of GIPS during 2010–2011 was a result of negative market sentiments that have become very strong since the end of 2010.

faults (Reinhart *et al.*, 2003), and contagion (Eichengreen *et al.*, 1996). Taken together, these papers contribute to our understanding of potential predictors of debt crises, which in turn, can be classified as follows: (1) *insolvency risk*, which includes capital and current account variables (international reserves, capital flows, short-term capital flows, foreign direct investment, real exchange rate, current account balance, and trade openness), and debt variables (public foreign debt, total foreign debt, short-term foreign debt, and foreign aid); (2) *illiquidity risk*, proxied by liquidity variables (short-term debt to reserves, debt service relative to reserves and/or exports, M2 to reserves); (3) *macroeconomic risk*, measured by macroeconomic variables (real GDP growth, inflation rate, exchange rate overvaluation, and international interest rates); (4) *political risk*, measured by institutional/structural factors (international capital market openness, financial liberalization, degree of political instability and political rights, and default history[2]; (5) *systemic risk*, namely the contagion variable usually proxied by the number/proportion of other debt crises [3] while focusing on the geographical localization of the countries[4].

As for model specification, different approaches have been explored based on the philosophical assumptions about the nature of sovereign default. One approach is based on reduced-form models, in which the default is assumed to be an inaccessible event whose probability is specified through a stochastic intensity process (Duffie *et al.*, 2003). Another approach is based on structural models, in which the default is explicitly modelled as a triggering event based on the balance-sheet notion of solvency (Gapen *et al.*, 2005). A third, and in some sense parallel, perspective is given by pure statistical approaches whose objective is mainly to predict defaults in a way that is only loosely connected to the theory. Here, the literature is extensive and focuses on, (i) logit/probit models; (ii) classification methods, namely cluster and discriminant analysis, and artificial neural networks; (iii) signal approach, which includes the noise-to-signal ratio approach and the regression tree analysis.

Many key studies exploring the issue of sovereign default using the three abovementioned statistical approaches complement our work in terms of empirical results and methodological procedures.

With regard to logit/probit models, McFadden *et al.* (1985) use both specifications and Oral et al. (1992) introduce a generalized logit model to link country risk rating and

[2]In this perspective, default history assumes a signalling role about the credibility of a sovereign to meet creditor needs, and this is coherent with the debt intolerance view introduced in Reinhart *et al.* (2003).

[3]This definition is in line with Eichengreen *et al.* (1996) who define contagion as a case where knowing that there is a crisis elsewhere increases the probability of a crisis at home, even after taking into account a country's fundamentals.

[4]The prevalent literature assumes that contagion is regionally-based (Kaminsky and Reinhart, 2000).

political-economic indicators. Moreover, Ciarlone and Trebeschi (2005) apply a multinomial model to develop an EWS for emerging markets over 1980–2002 predicting crises 76 per cent of the times and raising false alarms 36 per cent of the times. Fuertes and Kalotychou (2006) prove that out-of-sample, simple pooled logit models outperform complex logit specifications when using panel data.

With regard to classification methods, Frank and Cline (1971) and Taffler and Abassi (1984) apply discriminant analysis to predict whether a country will experience debt servicing difficulties, while Fioramanti (2008) uses artificial neural networks to realize an EWS for sovereign debt crises.

With regard to the signal approach, Kaminsky, Lizondo, and Reinhart (1998) (KLR) introduce the noise-to-signal ratio approach, also used in Kaminsky (1998), Goldstein *et al.* (2000), and Alessi and Detken (2011). Instead, Manasse *et al.* (2003) and Manasse and Roubini (2009) propose regression tree analysis to realize EWSes for debt crises finding that while on the one hand, regression trees show very strong crisis prediction ability, on the other, they send out more false alarms relative to the logit model (Manasse et al., 2003). The authors are also able to identify the following three major types of risks (Manasse and Roubini, 2009): (i) solvency, characterized by high external debt over GDP together with monetary or fiscal imbalances, as well as large external financing needs; (ii) illiquidity, identified by moderate debt levels, but with short-term debt in excess of 130 per cent of reserves coupled with political uncertainty and tight international capital markets; and (iii) macro-exchange rate risks, which arise from the combination of low growth and relatively fixed exchange rates. In terms of methodology, empirical analysis, dataset, and results, Manasse *et al.* (2003) and Manasse and Roubini (2009) are the closest papers to our work. The results obtained in our empirical analysis complement and generalize their findings, since we offer a better explanation of the sovereign defaults that occurred over the inspected time period, we better identify the main commonalities and differences in sovereign debt crises, and better predict out-of-sample defaults.

Such improvements are obtained by using the new regression tree-based algorithm introduced in Vezzoli and Stone (2007), which allows us to remove some limitations of traditional regression trees when dealing with panel data. In fact, traditional regression trees do not pay attention to autocorrelations among covariates and country-specificities. In other terms, the classical approach explores the data as if they were a collection of independent observations in both time and spatial dimensions (e.g., the real GDP growth measured in 2010 for Greece is completely independent from its past data as well as from contemporaneous and past observations of the same variable measured for other countries). The algorithm proposed by Vezzoli and Stone (2007) is instead devised to cope with the fitting versus forecasting paradox taking into account country-specificities by preserving the information structure contained in the panel data.

Computationally, the procedure is in two steps: (1) in the *first step*, we estimate a number of regression trees by removing one country at a time from the dataset, thereby obtaining multiple predictions (and taking into account for country-specificities); (2) in the *second step*, we fit a single final regression tree (FRT) using the average of the predictions obtained in the first step in place of the original dependent variable.

We show that our FRT is a parsimonious model, with good predictability (accuracy), better interpretability, and minimal instability. In the first step, the model is constructed in a forward-looking basis while allowing for forecasting averaging, which is particularly useful in improving accuracy and reducing the variance of forecasting errors, as discussed in Fuertes and Kalotychou (2007). In the second step, the replacement of $y$ with $\hat{y}$ mitigates the effects of noisy data on the estimation process that affect both the predictors and the dependent variable itself. As shown in Debashis et al. (2008), this replacement is, in essence, a sort of de-noising procedure with which the outcome should help reduce the variance in the model selection process.

The data used in the empirical analysis are from S&P's, World Bank's Global Development Finance (GDF), IMF, Government Finance Statistics database (GFS), and Freedom House (2002), and include annual observations over 1975–2010 for 66 emerging economies together with GIPS, the European countries that experienced an actual crisis episode and/or exhibited a large surge in government bond spreads driven by market sentiment (De Grauwe and Ji, 2012). We conduct a horse race of our base model with competing EWSes: (i) logit; (ii) stepwise logit; (iii) noise-to-signal ratio, introduced in KLR; (iv) regression tree analysis, used in Manasse and Roubini (2009). The in-sample analysis is performed over the entire time horizon 1975–2010, with 122 debt crisis episodes, while the out-of-sample analysis is carried out one-step-ahead from 1991 to 2010, including 49 debt crisis episodes, and focusing on the models' performance during the 'big three' crises (Mexican, Asian, and 2007–2010 global financial crises).

Our results prove that short-term debt to reserves and default history are the most significant variables in predicting a debt crisis, which basically identify: (i) episodes with low illiquidity problems where, (a) the risk of a debt crisis is the lowest for countries with no bad default history, while (b) the risk is high for countries with bad default history and strong negative real GDP growth when U.S. interest rates are low (as it is the case for the Greek and Irish crises of 2010); (ii) episodes with high illiquidity problems and bad default history, where the probability of default is high.

The several metrics run to assess the model accuracy in-sample show that our model provides an accurate description of past data, and near to the best model, while out-of-sample the diagnostics (root mean square error, Brier score, logarithmic probability score, Diebold and Mariano (1995) test, and AUC-based tests) document that our model produces the best forecasts, while also adapting to different risk aversion targets. Interestingly, in the clinical study of major crises occurring over 1991–2010, we find that only the so-called 'algorithmic modelling' approaches (FRT, regression tree, and KLR) are able to identify the common latent root of the recent global financial crisis in the Eurozone.

Finally, to compare alternative EWSes considering both in- and out-of-sample accuracy, we introduce a 'two-dimensional' loss function attaching: (a) a cost to missed defaults (type-I errors) relative to false alarms (type-II errors); (b) a weight to in-sample relative to out-of-sample type-I and type-II errors. In this way we evaluate an EWS in relation to a decision-maker's objective function defined in the spirit of the forecasting versus policy dilemma. Using this new metric, we show that our classifier strongly dominates competing EWSes while exhibiting stable accuracy. The rest of the paper is organized as follows. Section II discusses the methodology and Section III describes the data. Section IV reports the results and Section V concludes.

# II. Methodology

In this section we describe the methodology used in this paper, which is based on the signal approach, first giving some preliminaries and basic notation on regression trees, and then presenting the algorithm underlying our EWS. Next, we introduce the other competing approaches used in the empirical analysis: (i) logit; (ii) stepwise logit; and (iii) KLR.

The methodological notations we present are based on the issue of sovereign default prediction, by letting $Y$ be the observed indicator variable that takes the values 1 and 0 for default- and non-default, respectively, and $\mathbf{X} = (X_1, X_2, \ldots, X_R)$ be a collection of $r = 1, 2, \ldots, R$ predictors. The relationship between $Y$ and $\mathbf{X}$ is specified as:

$$y_{jt} = f\left(\mathbf{x}_{jt\text{-}1}\right) + \varepsilon_{jt}, \tag{1}$$

where $f\left(\mathbf{x}_{jt\text{-}1}\right)$ is an unknown functional form of predictors $\mathbf{X}$ measured in $t-1$ and parameterized by $\theta$: $f\left(\mathbf{x}_{jt\text{-}1}\right) = \theta'\mathbf{x}_{jt\text{-}1}$, where $\varepsilon$ is the random term for which some distributional assumption can be specified. The objective is to estimate $\theta$ with and without making an assumption about the random term distribution.

## Regression trees

Regression trees are nonparametric (or model free) approaches that look for the best local prediction of a response variable[5] $y$. The data consists of $R$ inputs and a continuous response, $Y$, for each of the $N$ observations. The algorithm needs to decide on the splitting variables and split points, and also what topology (shape) the tree should have. The algorithm recursively partitions the input space $\mathcal{S}$, which is the set of all possible values of $\mathbf{X}$ ($\mathbf{X} \in \mathcal{S}$), into disjoint regions $\widetilde{T}_k$ with $k = 1, 2, \cdots, K$. More precisely,

$$\mathcal{S} \subseteq \bigcup_{k=1}^{K} \widetilde{T}_k. \tag{2}$$

A tree $T$ can be formally expressed as $T(Y, \mathbf{X}, \theta)$ with $Y$ the vector of the dependent variable, $\mathbf{X} = (X_1, X_2, \cdots, X_R)$, and $\theta = \{\widetilde{T}_k, g_k\}_1^K$, where $g_k$ is a piecewise constant for each $k$. The predictive model is given by

$$f(\mathbf{X}) = \sum_{k=1}^{K} g_k I(\mathbf{X} \in \widetilde{T}_k), \tag{3}$$

where $I(\mathbf{X} \in \widetilde{T}_k)$ is an indicator function and $g_k = average(y_{jt} \mid \mathbf{X} \in \widetilde{T}_k)$. Hence, $Y$ is predicted as the average of the dependent variable observations within the corresponding region (terminal node) of the regression tree.

The partition is realized keeping the objective of obtaining maximum homogeneity within the regions, which is achieved by minimizing an impurity index measured by the Gini index for classification trees, or by the sum of squared errors for regression trees[6]. Furthermore, regression trees are conceived with the aim of improving out-of-sample predictability, and hence, they are estimated through a rotational estimation procedure—cross-validation—with which the sample is partitioned into subsets such that the analysis is initially performed on a single subset (the training set), while the other subsets are retained for subsequent use in confirming and validating the initial analysis (the validation or testing sets).

[5] We distinguish between regression trees, when the dependent variable is continuous, and classification tree, when the response variable is categorical.

[6] Due to technical difficulties in solving such a minimization process, many researchers use a greedy algorithm to grow the tree, by sequentially choosing splitting rules for nodes based upon some maximization criterion, and then controlling for overfitting by pruning the largest tree according to a specific model choice rule such as cost-complexity pruning (i.e., cross-validation or multiple tests for the hypothesis that two adjoining regions should merge into a single one). See Hastie et al. (2009) for technical details.

## CRAGGING

The classical regression tree algorithm was designed for cross-sectional data, assuming $Y$ to be i.i.d. within each region and independent across the regions. Unfortunately, neither the first nor the second assumption applies for panel data.

As discussed in the introduction, in order to remove these limitations and preserve the information contained in the data, Vezzoli and Stone (2007) proposed the CRAGGING (CRoss-validation AGGregatING) algorithm[7]. In a nutshell, the idea underlying the CRAGGING is to repeatedly rotate the subsets in which the analysis is initially performed to such an extent as to, first, generate multiple predictors and, second, combine them to obtain a univariate and stable tree. This is the reason why CRAGGING can be viewed as a generalization of regression trees.

Let $(Y, \mathbf{X})$ be panel data with $N$ observations and suppose, for simplicity, that each unit $j$, with $j = 1, \cdots, J$, has the same number of years $t$, with $t = 1, \cdots, T_j$, (balanced panel data) and $J \cdot T_j = N$. Use $\mathcal{L} = \{1, 2, \cdots, J\}$ to denote the set of units and $\mathbf{x}_{jt-1} = (x_{1jt-1}, x_{2jt-1}, \cdots, x_{rjt-1}, \cdots, x_{Rjt-1})$ to denote the vector of predictors of unit $j$ observed at time $t-1$ where $j \in \mathcal{L}$. The procedure is in two steps and runs as follows.

In the *first step*, by using the $V$-fold cross-validation, $\mathcal{L}$ is randomly partitioned into $V$ subsets[8] denoted by $\mathcal{L}_v$, with $v = 1, 2, \cdots, V$, each containing $J_v$ units and $N_v$ observations[9]. We then randomly select one of the $\mathcal{L}_v$ sets, reserved for testing, and the corresponding training set, denoted by $\mathcal{L}_v^c$, is obtained as $\mathcal{L} - \mathcal{L}_v$, which contains $J_v^c$ units and $N_v^c$ observations. Next, by removing one unit (country) $\ell$ from $\mathcal{L}_v^c$, we get a perturbed training set denoted by $\mathcal{L}_{v \setminus \ell}^c$.

A regression tree is now trained on the dataset $\{y_{jt}, \mathbf{x}_{jt-1}\}_{j \in \mathcal{L}_{v\setminus\ell}^c,\ t=1,2,\cdots,T_j}$ and pruned with a cost-complexity parameter $\alpha \geq 0$. We thus proceed to compute predictions in the test set as follows:

$$\hat{y}_{jt,\alpha\ell} = \hat{f}_{\alpha,\mathcal{L}_{v\setminus\ell}^c}(\mathbf{x}_{jt-1}) \quad \text{with} \quad j \in \mathcal{L}_v, \quad \text{and} \quad t = 1, 2, \cdots, T_j, \tag{4}$$

where $\hat{f}_{\alpha,\mathcal{L}_{v\setminus\ell}^c}(\cdot)$ is the prediction function of the single tree.

Once the predictions are obtained, the $\ell$-th country is reinserted in the training set and the same procedure is repeated for each $\ell$ in $\mathcal{L}_v^c$, thus obtaining $J_v^c$ predictions of debt crisis probabilities for each country-year belonging to the test set.

To improve the accuracy of predictions, these estimated default probabilities are finally

[7] This algorithm is also used in Savona and Vezzoli (2012).

[8] In the partition, it is necessary that $V < J$ in order to preserve the structure of the data.

[9] The dimension of each $V$ subset is of as equal size as possible.

combined through averaging by running the following equation:

$$\hat{y}_{jt,\alpha} = \frac{1}{J_v^c} \sum_{\ell \in \mathcal{L}_v^c} \hat{f}_{\alpha,\mathcal{L}^c_{v\setminus\ell}}(\mathbf{x}_{jt-1}) \quad \text{with} \quad j \in \mathcal{L}_v \quad \text{and} \quad t = 1, 2, \cdots, T_j, \tag{5}$$

which is the average[10] of the functions (4) fitted over the units contained within the test set $\{y_{jt}; \mathbf{x}_{jt-1}\}_{j \in \mathcal{L}_v,\ t=1,2,\cdots,T_j}$.

The perturbation procedure just described, through which we remove one unit at a time within the training set, is the leave-one-*unit*-out cross-validation introduced in Vezzoli and Stone (2007) in order to preserve the structure of the data.

Finally, a second cross-validation, the well-known *v-fold cross-validation*, is carried out over the test sets with $v = 1, \ldots, V$. The objective of this second cross-validation is to find the optimal tuning parameter, $\alpha^*$, namely the cost complexity parameter that minimizes the prediction errors on all the test sets. In essence, equations (4) and (5) are run by arbitrarily changing the value for $\alpha$ and solving the following objective function:

$$\alpha^* = \arg\min_{\alpha} LF(y_{jt}, \hat{y}_{jt,\alpha}) \quad \text{with} \quad j \in \mathcal{L}, \quad t = 1, 2, \cdots, \sum_{j=1}^{J} T_j, \tag{6}$$

where $LF(\cdot)$ is a generic loss function.

The entire procedure described before is finally run $M$ times so as to minimize the generalization error, which is the prediction error over an independent test sample, and then averaging the results in order to get the CRAGGING predictions to be used in the second step. Using the Strong Law of Large Numbers[11], Breiman (2001) has indeed shown that as the number of trees increases ($M \to \infty$), the generalization error has a limiting value and the algorithm does not overfit the data. As a result, the CRAGGING predictions are given by

$$\tilde{y}_{jt}^{\text{crag}} = M^{-1} \sum_{m=1}^{M} \hat{y}_{jt,\alpha^*} \quad \text{with} \quad j \in \mathcal{L}, \quad t = 1, 2, \cdots, \sum_{j=1}^{J} T_j. \tag{7}$$

In the *second step*, we estimate FRT, namely a regression tree fitted on the CRAGGING predictions $(\tilde{Y}^{\text{crag}}, \mathbf{X})$ with cost complexity parameter $\alpha^{**} = M^{-1} \sum_{m=1}^{M} \alpha^*$. Here, through the replacement of $Y$ with CRAGGING predictions, (1) we mitigate the effects of noisy data on the estimation process that affect both the predictors and the dependent

[10] The base learners $\hat{f}_{\alpha,\mathcal{L}^c_{v\setminus\ell}}(\cdot)$ are linearly combined so that $\hat{y}_{jt,\alpha}$ will act as a good predictor for future $(y|\mathbf{x})$ in the test set.

[11] The theorem proves that the average computed on a large number of sequences of variates will be much closer to the expected value if the number of trials carried out is large.

variable itself[12]; (2) we realize a tree-based model that encompasses the overall forecasting ability arising from multiple trees, thus obtaining a parsimonious model with good predictability (accuracy), better interpretability, and minimal instability.

In the context of sovereign default predictability, where the $Y$s are 'crisis/non-crisis' indicators (binary variable), the computational problem for CRAGGING is the same as that for regression trees. In fact, the main aim is to find 'criteria' (expressed as inequalities) for $\mathbf{X}$ such that as many 'crisis' as possible fall in one partition, and as many 'non-crisis' as possible fall in a different one. To provide an example to illustrate how the results can be interpreted intuitively, suppose you realize an FRT using lagged data for $\mathbf{X}$ to predict $Y$ up to $t$. Suppose that the FRT may indicate that a particular combination of individual characteristics, such as high debt (say, more than 49.7 per cent of GDP) and high inflation (say, larger than 10.5 per cent) incur the largest risk node with a probability of, say, 66.8 per cent[13]. We are now in $t$ and you make predictions for $t+1$. In $t$, when a country's debt over GDP and inflation are both larger than the corresponding thresholds, we get the probability of a crisis occurring in $t+1$ to be 66.8. To better illustrate how CRAGGING is computed in practice, in Appendix B, we describe each step to obtain the FRT, considering only 10 countries for simplicity.

## Competing models

### *Logit and stepwise logit*

In the logistic regression technique, the posterior probabilities of observing a default case are modelled by means of linear functions in $\mathbf{X}$ assuming a standard logistic distribution for the random term $\varepsilon$ in (1). Then the functional approximation, assuming country and time homogeneity, has the following linear basis expansion

$$f(\mathbf{x}_{jt-1}) = \Pr(y_{jt} = 1|\mathbf{x}_{jt-1}) \equiv p_{jt}(\mathbf{x}_{jt-1}) = \frac{1}{1 + \exp -(\iota + \mathbf{x}'_{jt-1}\boldsymbol{\beta})}. \tag{8}$$

This is the pooled logit model specification with the $(1+R) \times 1$ parameter set vector $\theta = (\iota, \boldsymbol{\beta}')'$ estimated by maximum likelihood, using the conditional likelihood of $Y$ given $\mathbf{X}$.

[12] As recently proven in Debashis *et al.* (2008).

[13] The example is from Manasse and Roubini (2009), since they realize a tree structure with the highest risk node (with a probability of 66.8 per cent), when debt over GDP is greater than 49.7 per cent and inflation is higher than 10.5 per cent.

The second related model we use in the comparative analysis is the backward stepwise logit. Starting with the full model in (8) that includes all candidate variables, backward elimination tests the deletion of each variable using the Akaike Information Criterion (AIC), deleting one variable per time in order to minimize the AIC score and repeating this process until no further improvement is possible.

### *Noise-to-signal ratio approach*

Discrete choice models (logits) fit a specific stable relation between a set of covariates and a latent variable that translates to crisis probability. The underlying, rigid assumption is that such a latent variable is both linearly dependent on the indicators and strictly monotonously related to crisis probability. Without imposing such rigid assumptions, the KLR approach aims at sending a warning signal for an impending crisis through a non-parametric threshold approach. The behavior of single variables is analysed as sending a warning signal for an impending crisis if the corresponding value exceeds some threshold to be chosen to minimize the probability of failing to call crises and the probability of false alarms simultaneously. The optimal cut-off point is estimated by minimizing the 'false alarm-to-good signal ratio', namely the type-II errors (noise or $1 - specificity$) over the 1 *minus* type-I errors (good signals or *sensitivity*). The procedure is repeated for all $r$ predictors with $r = 1, \ldots, R$, and then a weighted sum of the 0–1 signals by individual predictors is computed while excluding those having a noise-to-signal ratio greater than 1 and using the inverse of the optimal noise-to-signal ratio as weight. Therefore, such a composite index (CI) gives more weight to better performing (smaller minimum noise-to-signal ratios) indicators. Formally, let $\omega_r = \frac{b}{1-a}$ be the noise-to-signal ratio of the $r$-th variable with $a$ and $b$ denoting the type-I and type-II errors, respectively; let $\omega^*_{r,\mathfrak{c}_r} = \arg\min_{\mathfrak{c}_r} \omega_r$ with $\omega^*_{r,\mathfrak{c}_r} < 1$ be the optimal noise-to-signal ratio of the $r$-th variable, computed in correspondence of the threshold $\mathfrak{c}_r$. As a result, the CI for unit $j$ at time $t$ is computed as

$$CI_{jt}(\mathbf{x}_{jt-1}) = \sum_{r=1}^{R} \frac{1}{\omega^*_{r,\mathfrak{c}_r}} I_{\mathfrak{c}_r}\left(x_{r,jt-1}\right) \quad \text{with} \quad \omega^*_{r,\mathfrak{c}_r} < 1, \tag{9}$$

where

$$I_{\mathfrak{c}_r}\left(x_{r,jt-1}\right) = \begin{cases} 1 & if\ |x_{r,jt-1}| > \mathfrak{c}_r \\ 0 & if\ |x_{r,jt-1}| \leq \mathfrak{c}_r. \end{cases} \tag{10}$$

Once the CI has been obtained, the probabilities of observing default-cases[14], that is, the functional approximation in (1), are estimated as the number of times where

[14] The probabilities obtained through the KLR procedure are constant in each time $t$ with $t = 1, 2, \cdots, T$, and across all units $j$. For this reason, we remove the subscript $j$ in CI.

CI exceeds a certain threshold $\mathfrak{C}$ and a crisis occurred, divided by the total number of observations in which $CI > \mathfrak{C}$. Formally,

$$f(\mathbf{x}_{t-1}) = \Pr(\mathbf{x}_{t-1}) = \frac{\sum_t I_{\mathfrak{C}}(CI|y_t = 1)}{\sum_t (I_{\mathfrak{C}}(CI))} \quad \text{with} \quad t = 1, 2, \cdots, T, \tag{11}$$

where

$$I_{\mathfrak{C}}(CI) = \begin{cases} 1 & if\ CI > \mathfrak{C} \\ 0 & if\ CI \leq \mathfrak{C}, \end{cases} \tag{12}$$

with $\Pr(\mathbf{x}_{t-1}) = 0$ when $y_t = 0$. To compute the threshold $\mathfrak{C}$, we used a similar procedure as for single predictors, but instead of selecting the threshold that minimizes the noise-to-signal ratio of CI, we referred to the Youden index (YI), a diagnostic test for accuracy widely used in clinical applications involving the receiver operating characteristic curve that we will discuss in the next section. As will be shown, YI is simply the sum of sensitivity $(1-a)$ and specificity $(1-b)$ *minus* 1 using a specific threshold $\mathfrak{C}$, and gives us a summary measure about the classification ability of a model considering both default and non-default classifications. Hence, the objective is to find the optimal $\mathfrak{C}$ so as to maximize YI. As opposed to the noise-to-signal ratio, YI is quite robust to extreme type-I and type-II errors giving an optimal trade-off between good signals and false alarms being also directly related to the area under the curve. On the other hand, as pointed out in Mulder *et al.* (2002), the minimization of the noise-to-signal ratio could lead to extreme thresholds for which the default is hardly signalled while false signals tend to zero.

## Model accuracy

### *In-sample diagnostics*

We use several metrics to assess models' accuracy to fit the in-sample data based on the difference between $Y$ and $\hat{Y}$ (or prediction errors). First, we use the root mean square error (RMSE) to assess the standard model fitting quality:

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^{J} \sum_{t=1}^{T_j} (\hat{y}_{jt} - y_{jt})^2}. \tag{13}$$

Then, we apply the Bayesian Information Criterion (BIC) that has been shown to be asymptotically consistent as a model selection criterion as $N \to \infty$, and that also gives the approximate Bayesian posterior probability of the true model among possible alternatives:

$$BIC = N \ln\left(\frac{RSS}{N}\right) + \delta \ln(N), \tag{14}$$

where RSS is the residual sum of square errors $\sum_{j=1}^{J}\sum_{t=1}^{T_j}(\hat{y}_{jt} - y_{jt})^2$ and $\delta$ is the number of parameters of the estimated model with $\delta \in [1,\, R]$.

Second, we turn to scoring rules based on probability estimates, namely the Brier score:

$$BS = \frac{1}{N}\sum_{j=1}^{J}\sum_{t=1}^{T_j} 2\,(\hat{y}_{jt} - y_{jt})^2, \quad BS \in [0,\, 2], \tag{15}$$

and the related logarithmic probability score that penalizes large errors more than Brier score:

$$LPS = -\frac{1}{N}\sum_{j=1}^{J}\sum_{t=1}^{T_j} y_{jt}\ln(\hat{y}_{jt}) + (1 - y_{jt})\ln(1 - \hat{y}_{jt}), \quad LPS \in [0,\, \infty]. \tag{16}$$

Third, we rely on signal-based diagnostic tests that provide a tool for model selection, focusing on the classification ability of default and non-default cases using the receiver operating characteristic (ROC) curve.

The ROC curve is a monotone increasing function mapping $(1-a) = sensitivity$ onto $b = 1 - specificity$, where sensitivity is computed as the fraction of the default cases correctly classified over total defaults (true positives), and specificity, as the fraction of non-defaults correctly classified over total non-defaults. Defaults are classified according to different cut-off points $\mathfrak{C} \in [0,1]$, which results in an ROC curve that is a function of $\mathfrak{C}$, namely $ROC(\mathfrak{C})$. The diagnostics based on the ROC used in this paper are (1) AUC and pairwise test on AUC differences; (2) YI; (3) loss function.

The area under ROC (i.e., AUC) gives a measure of a model's discrimination power and can be interpreted as the probability of assigning higher and lower estimates for defaults and non-defaults, respectively. Formally,

$$AUC = \int_0^1 ROC(\mathfrak{C})d\mathfrak{C}. \tag{17}$$

In our analysis, we use the trapezoidal rule with which (17) is approximated summing the areas of the trapezoids formed after dividing the area into a number of strips of equal width. As shown in Bamber (1975), when calculated by the trapezoidal rule, AUC has been shown to be identical to the Mann–Whitney U-statistic for comparing distributions. This intuition is formalized in DeLong *et al.* (1988) who propose a nonparametric test for the AUC differences we use for ranking models on the basis of pairwise AUC differences. Letting $\hat{\mathbf{U}}$ be the vector of AUC estimates, $\mathbf{L}$ is a suitable contrast matrix for (i.e., $H_0 : \mathbf{LU} = \mathbf{0}$, where $\mathbf{0}$ is the zero matrix) and $\mathbf{S}$ is the covariance matrix for AUC estimates[15]; then, the statistic for a pair of classifiers is

$$\frac{(\mathbf{L}\hat{\mathbf{U}})^2}{(\mathbf{LSL}')} \sim \chi^2_{(\mathrm{rank}(\mathbf{L}))} \tag{18}$$

which follows a chi-square distribution with rank($\mathbf{L}$) degrees of freedom.

YI is a diagnostic accuracy measure which has been proven to be effective in finding the optimal cut-off point in order to maximize the overall classification ability, thus minimizing both type-I and type-II errors. Mathematically,

$$YI = \arg\max_{\mathfrak{C}} [(1-a) + (1-b) - 1] \quad \text{with} \quad \mathfrak{C} \in [0,1]. \tag{19}$$

YI is the point on the ROC curve farthest from chance, that is, the diagonal line of the ROC space, the so-called line of no-discrimination for which the classification is equivalent to random guessing. Note also that with two states, as in our study, YI has been shown to be a linear transformation of AUC with $YI = 2 \cdot \triangle - 1$ and the approximated AUC $\triangle = [(1-a) + (1-b)]\,/2$[16].

Using the best cut-off point $\mathfrak{C}^*_{YI}$ obtained from (19), we finally compute the loss function for each classifier as the weighted sum of the missed default and non-default probabilities with cost for type-I and type-II errors $\zeta$ and $(1-\zeta)$, respectively:

$$LF = [\zeta \cdot a_{\mathfrak{C}^*_{YI}} + (1-\zeta) \cdot b_{\mathfrak{C}^*_{YI}}], \quad LF \in [0,1]\,. \tag{20}$$

The cost $\zeta$ reflects the risk-aversion for the decision-makers who presumably can be more sensitive to missing defaults (which is also coherent with the Neyman–Pearson decision rule[17]), thus yielding $\zeta > 0.5$. Decision-makers could be also less risk-averse, as is the case for investors looking for high-yield (and high-risk) investments, and this appears as $\zeta < 0.5$. To take into account the two perspectives, we computed the loss function for values of $\zeta$ arbitrarily ranging from 0.1 to 0.9 and ranked the models for each of these cost values.

[15]See DeLong *et al.* (1988) for further details on the mathematical derivation and parameter computation for the test.

[16]See Hilden and Glasziou (1996).

[17]The Neyman–Pearson decision rule, commonly used in signal processing applications, is to minimize the type-I error subject to some constant type-II error which implies more sensitivity towards the type-I error.

***Out-of-sample diagnostics***

To evaluate forecasting accuracy, we rely on an out-of-sample exercise where the models were recursively computed using the most recent observations available and were forecast one year ahead. For each year of the out-of-sample period[18] $t^{\text{out}}$, we add one-$t$-ahead observations to the previous fit period $t^{\text{in}}$ and we use the new fitting period for updating the model estimates; next, these new estimates are used to make predictions for the following year. As a result, we provide forecasts dynamically for the holdout sample that can be evaluated using the same battery of diagnostic tests employed in-sample. We then replicate the tests (13)-(20) only excluding the BIC for computational convenience[19].

Furthermore, we include the Diebold–Mariano forecasting test to assess whether our model is significantly better than competing models controlling for non-normality of forecasting errors and serial correlation. As discussed in Diebold and Lopez (1996), the test is applicable to a wide class of loss functions and can readily accommodate the non-normality of forecast errors, as well as ordinal and probability forecasts.

Define $d_{jt^{\text{out}}} = [(\hat{y}^A_{jt^{\text{out}}} - y_{jt^{\text{out}}})^2 - (\hat{y}^B_{jt^{\text{out}}} - y_{jt^{\text{out}}})^2]$, the square error difference of models $A$ and $B$[20] for the observation of unit $j$ at time $t^{\text{out}}$ in the holdout sample, and let $\overline{d} = \frac{\sum_{j=1}^{J}\sum_{t^{\text{out}}=T_j^{\text{in}}+1}^{T_j} d_{jt^{\text{out}}}}{N^{\text{out}}}$, where $N^{\text{out}}$ is the number of observations in the out-of-sample test. The Diebold–Mariano forecasting test is as follows:

$$DM = \frac{\overline{d}}{\sqrt{\frac{\widehat{var}(\overline{d})}{N^{\text{out}}}}} \backsim N(0,1), \tag{21}$$

where $\widehat{var}(\overline{d})$ is a consistent estimate of the variance of $\overline{d}$ (see Diebold and Mariano, 1995). In our analysis, we use only one-step ahead in computing the Diebold–Mariano test, since it is common practice to update forecasts on an annual basis, namely when new values for economic variables are added to past data in order to recalibrate the EWS predictions. Note also that in this case, the potential error autocorrelation becomes an issue to deal with, as pointed out by Fuertes and Kalotychou (2007).

[18]Note that $t^{\text{in}} = 1, \cdots, T_j^{\text{in}}$ and $t^{\text{out}} = T_j^{\text{in}} + 1, \cdots, T_j$ denote the time period for in- and out-of-sample tests, respectively.

[19]As is known (see equation 14), BIC introduces a penalty term for the number of parameters in the model ($\delta$ in equation 14), since it is possible to increase the likelihood by adding parameters, thereby yielding the overfitting problem. In the out-of-sample analysis, the models were recursively computed from 1991 to 2010, and the number of selected predictors changed in each estimation for all models with the only exception being the logit model. As a result, the computation of BIC over the holdout sample would be reflected by a change in $\delta$ in equation (14).

[20]In the empirical analysis $A$ denotes our model and $B$ the competing model.

***Two-dimensional loss function***

The forecasting versus policy dilemma highly complicates the global evaluation of the models when jointly considering in- and out-of-sample accuracy. In one extreme, in-sample accuracy could be good (bad) while out-of-sample bad (good), thus requiring to trade between fitting ability and forecasting ability *together with* missed defaults and false alarms. To put the discussion into perspective, consider that (i) on the one hand, decision-makers can be more or less risk-averse, namely more or less sensitive towards type-I errors (this is, say, the first dimension of the problem); (ii) on the other hand, decision-makers can be either more interested in the data generation process (thus showing more sensitivity towards in-sample errors), or more interested in forecasting activity (the second dimension of the problem). As a result, we have a two-dimensional problem we propose to handle with a corresponding two-dimensional loss function, the $2^{\mathrm{D}}LF$, by attaching (a) a cost to missed defaults (type-I errors) relative to false alarms (type-II errors); (b) a weight to in-sample relative to out-of-sample type-I and type-II errors. In this way, we evaluate EWSes in relation to a decision-maker's objective function conceived in the spirit of the forecasting versus policy dilemma.

Using $\varrho$ and $(1-\varrho)$ to denote the weights for in- and out-of-sample errors, respectively, and referring to the notation in (20), our $2^{\mathrm{D}}LF$ becomes

$$\begin{aligned} 2^{\mathrm{D}}LF = \zeta \cdot [\varrho \cdot (a_{\mathfrak{C}^{*\mathrm{in}}_{YI}})^{\mathrm{in}} + (1-\varrho) \cdot (a_{\mathfrak{C}^{*\mathrm{out}}_{YI}})^{\mathrm{out}}] + \\ +(1-\zeta) \cdot [\varrho \cdot (b_{\mathfrak{C}^{*\mathrm{in}}_{YI}})^{\mathrm{in}} + (1-\varrho) \cdot (b_{\mathfrak{C}^{*\mathrm{out}}_{YI}})^{\mathrm{out}}], \quad 2^{\mathrm{D}}LF \in [0,1], \end{aligned} \tag{22}$$

where $\mathfrak{C}^{*\mathrm{in}}_{YI}$ and $\mathfrak{C}^{*\mathrm{out}}_{YI}$ denote the optimal cut-off points identified by the YI in- and out-of-sample, while $(a_{\mathfrak{C}^{*\mathrm{in}}_{YI}})^{\mathrm{in}}$ and $(a_{\mathfrak{C}^{*\mathrm{out}}_{YI}})^{\mathrm{out}}$ denote the type-I errors in- and out-of-sample computed in correspondence of $\mathfrak{C}^{*\mathrm{in}}_{YI}$ and $\mathfrak{C}^{*\mathrm{out}}_{YI}$, respectively. Analogously, $(b_{\mathfrak{C}^{*\mathrm{in}}_{YI}})^{\mathrm{in}}$ and $(b_{\mathfrak{C}^{*\mathrm{out}}_{YI}})^{\mathrm{out}}$ denote the type-II errors in- and out-of-sample computed in correspondence of the two YI-based cut-off points.

$2^{\mathrm{D}}LF$ helps select the best model for given $\zeta$ and $\varrho$ which also identify the key features for major classes of decision-makers. In fact, first, investors are generally more focused on future risk adjusted returns of their investments (low $\varrho$) while showing different risk aversion levels based on their utility function: aggressive investors exhibit high-return targets, while conservative investors show low-risk targets. As a result, on the one hand, the costs of missed investment opportunities after false warning are on average higher than losses due to defaults, thus translating into low $\zeta$. On the other hand, losses from sovereign debt crises are clearly greater than missed high yield opportunities, which implies high $\zeta$.

Second, policy-makers and macro-financial supervisors are more focused on detecting impending risk signals in order to take adequate policy interventions. In doing so, first, they must explain the reasons for past crises, and second, realize optimal EWSes that have the objective of minimizing false alarms while maintaining a high predictive ability of impending crises. Thus, they should realize models to forecast future crises, by calibrating in- vs. out-of-sample predictability while minimizing false alarms. A collateral issue in our study concerns the inspection of $2^{\mathrm{D}}LF$ within a more formal framework, in which different decision-makers optimize their utility functions based on in- vs. out-of-sample and low vs. high risk aversion targets. We leave this question for future research.

# III. Data

The dataset used in this paper updates and extends the data used in Manasse and Roubini (2009), since it includes annual observations for 66 emerging economies over the period 1975–2010 together with GIPS countries that were more affected by the global financial crisis of 2008–2010. Data on predictors are from GDF, IMF, GFS, and Freedom House (2002), and are grouped according to the five categories outlined before by including (1) capital, current account, and debt variables; (2) liquidity measures; (3) macroeconomic factors; and (4) political risk factors[21] also including default history measured as the sum of past debt crises; (5) systemic risk, namely the contagion variable measured as the number of other debt crises occurring in the same year, distinguishing between total (the overall number of debt crises) and regional contagion (the number of debt crises within the same region). We also included dummies for oil producing nations as defined by WEO where fuel is the main export (DOIL), access to international capital markets (MAC), IMF lending (IMF), and EU membership (EU), thereby taking into account the economic and political status of EU countries. Except for contagion and dummies for oil and international capital markets, all the predictors are lagged one year, which is in the spirit of any predicting model[22].

Sovereign defaults are defined following Manasse *et al.* (2003) who consider a country to be experiencing a debt crisis if, (a) it is classified as being in default by S&P's, that is, when government fails to meet a principal or interest payment on an external obligation on the due date (including exchange offers, debt-equity swaps, and buyback for cash); (b) it has access to a large nonconcessional IMF loan in excess of 100 per cent of quota. As discussed by the authors, by using such a definition, we capture cases of outright

[21] We used, in particular, the index of political rights compiled by Freedom House (2002) that takes values on a scale from one (most free) to seven (least free).

[22] It is common procedure in the literature to use a proxy for contagion which is contemporaneous to the default indicator.

default or semicoercive restructuring together with near-to-defaults avoided through large financial packages from the IMF. Information was collected by S&P's and IMF's Finance Department, which relates in particular to Stand By Arrangements (SBA) and Extend Fund Facilities (EFF). Furthermore, we also referred to the countries that had access to the Emergency Financing Mechanism (EFM) used during the global financial crisis of 2008–2010. With the aim of realizing an EWS to predict a default *entry* rather than a *continuing* default, we included all the default events for each country subject to the fact that the country in $t-1$ was not in default. Otherwise, we excluded the observations for the default indicator as well as those for predictors.

In Appendix A, we provide data descriptions of debt crises and the 27 predictors used in the empirical analysis. As a whole, we analysed 122 crisis episodes reported in Table A1, of which 49 are used in the out-of-sample analysis over the sub-period 1991–2010.

# IV. Empirical results

## In-sample fitting accuracy

### *Predictors and risk stratification*

We run our procedure, as outlined in Section II, over the entire period 1975–2010, obtaining the risk stratification reported in Figure 1, which is the FRT run over the CRAGGING predictions. Computationally, the panel data containing 70 countries was first randomly divided into 10 sets with each containing 7 countries, i.e. the 10% of the overall number of units in the panel data. Hence, the training set contains 63 countries and the test set contains 7 countries. Since CRAGGING repeatedly perturbs the training sets by removing 7 units at a time,in correspondence of the optimal cost-complexity parameter $\alpha^*$ (6), we obtain 63 probability estimates for each observation. As discussed before, such a procedure was run $M$ times (7) so as to minimize the generalization error. Computationally, we run $M = 50$ times with corresponding $50 \times 63 = 3,150$ probability estimates, which allowed us to obtain a generalization error with a limiting value with no overfitting problems.

By using the 27 potential predictors discussed in Section III, only 5 variables are selected by FRT: (1) short-term debt to reserves; (2) default history (number of past defaults); (3) U.S. Treasury Bill rates; (4) real GDP variations; (5) exchange rate overvaluation. Hence, the economic process underlying a sovereign debt crisis can be explained using a parsimonious number of suitable proxies for illiquidity, macroeconomic, and political risks. From a statistical viewpoint, having only 5 out of 27 variables, which reflects the trade-off between complexity and accuracy implied by FRT, is particularly useful for realizing a model that is as simple as possible while providing a reasonable explanation for past data. Indeed, if on the one hand, by increasing the complexity of a model, we provide a better fit to the data, on the other hand, having too many parameters would reflect a large sensitivity to small changes, which in turn implies that the model will not distinguish between true dynamics and fluctuations due to measurement error and/or noise (Orrell and McSharry, 2009).

The EWS we realize partitions the predictor space into 9 terminal nodes according to specific splitting rules, thus obtaining a country risk stratification using multiple risk signals, while providing probability estimates of debt crises conditional on predictors and terminal nodes.

Short-term debt to reserves and default history are the most significant variables in predicting a debt crisis, with values of the corresponding thresholds of 84.8355 and 0.5, respectively. Together, the two predictors basically split the overall sample into (i) episodes with low illiquidity problems (smaller or equal than 84.8355 per cent) for which the probability of default is low for non-serial defaulters, while it is high for countries with bad default history and negative real GDP growth or high international interest rates (U.S. Treasury Bill rates); (ii) episodes with high illiquidity problems (greater than 84.8355 per cent) and bad default history where the probability of default is high. An in-depth analysis of the tree structure gives interesting insights about the determinants of sovereign debt crises. If indeed we focus on the main risk clusters of the tree, we can identify the following five major categories:

- *Higher Risk*, in which short-term debt to reserves is high (greater than 84.8355 per cent), the country experienced at least one default (history greater than 0.5), and strong exchange rate devaluations (OVER below $-99.89$) are accompanied by high U.S. Treasury Bill rates (greater than 9.795 per cent). As we note, along this path, the default probability is the highest with a value of 79.23;

- *Medium-High Risk*, where U.S. Treasury Bill rates play a key role both when short-term debt to reserves is low and illiquidity problems are high. In the first scenario (node 4), serial defaulters (history greater than 0.5) are exposed to risk with high U.S. Treasury bill rates (UST greater than 9.795 per cent). In the second (node 6), the path is that observed for higher risk but with low interest rates (UST below 9.795 per cent);

- *Medium Risk*, where short-term debt to reserves is significantly high (greater than 128.257 per cent) (node 9), or notwithstanding low values for short-term debt to reserves, real GDP growth is strongly negative (RGRWT less than $-1.5963$ per cent) (node 2);

- *Moderate Risk*, in which short-term debt to reserves is high but not as we have in the medium risk case ($84.8355 \leq STDR < 128.257$), the country experienced at least one default and exchange rate is not strongly undervaluated (OVER greater than $-99.8874$) (node 8);

- *Low Risk*, for countries that never suffered from sovereign defaults (nodes 1 and 5), or which exhibit non-negative real GDP growth (RGRWT greater than $-1.5963$ per cent) during periods of moderate U.S. interest rate trends (UST less than 9.795 per cent) (node 3).

The big picture coming from FRT is that debt crises are mainly driven by liquidity concerns together with the worsening of macroeconomic conditions. Illiquidity problems are exacerbated by strong exchange rate undervaluation for countries with bad default history during times of high interest rates. This finding is economically consistent for two reasons: (i) exchange rate undervaluation usually reflects current account deterioration; (ii) high interest rates in U.S., which in turn reflects tight monetary conditions, may suggest that capital flows to emerging markets are expected to reduce, thus contributing to debt servicing difficulties (Manasse and Roubini, 2009)[23].

Focusing on the Greek and Irish crises of 2010, the strong contraction in GDP growth together with low interest rates and a bad default history have been the major drivers. Both crises are clustered together with other defaults, such as Turkey 2002, Ukraine 1998, and Venezuela 1995, proving that the root of the recent Greek and Irish sovereign debt crises has been the same as that of other emerging market crises that occurred in the past. This point is of particular interest, since it complements Reinhart and Rogoff (2010), who proved that in both advanced and emerging countries, high debt/GDP levels are associated with notably lower growth outcomes. Putting together the two things, we may thus conjecture that the risk threshold we found for real growth of GDP may encompass excessive indebtedness, and this was the case for many, but not all, crises clustered within the same node, namely (in parenthesis the value of public debt over GDP): Jamaica 2010 (145 per cent), Greece 2010 (127 per cent), Hungary 1991 (117 per cent), Jordan 1989 (98 per cent), Turkey 2002 (78 per cent), Venezuela 1995 (72 per cent), and Ireland 2010 (65 per cent). The tree structure of FRT thus realize a risk partition controlling for different and significant country idiosyncrasies. Indeed, on the one hand,

[23]this is what happened during the crisis of 1980-1983, for which serial defaulters experienced debt problems because of high interest rates notwithstanding low levels of short-term debt to reserves

Greece and Ireland were classified as medium risk, as it shows a probability estimate of 20.21 per cent. On the other hand, Portugal and Spain were clustered within node 1 which is the lesser risky node, as it shows a probability estimate of 0.527 per cent, thus proving that FRT is also efficient in detecting which of the four EU countries included in the sample were risky and which were not.

Estimates of competing models (logit, stepwise logit, KLR, and Regression Tree) together with their different economic explanations of the sovereign default are in Appendix A.

***In-sample model ranking***

Panel A of Table 1 reports the battery of statistical tests used to assess how the five different models describe the data in-sample, where we note that BIC and RMSE rank regression tree as the best, and KLR as the worst. This is because BIC penalizes heavily for the number of parameters and because the probability estimates of regression tree are not as scattered as other models, thus reflecting a minor error dispersion. When using the scoring rules, the Brier score and the logarithmic probability score, regression tree is again ranked the best, and KLR again the worst. Using these diagnostics, FRT is ranked second, thus proving the model superiority of the regression tree approach (FRT and regression tree) relative to logistic regressions and KLR.

Signal-based diagnostic tests computed using the ROC curve provide better information about model reliability in classifying default and non-default episodes. The results are in columns 6–10 of Panel A of Table 1 and show that, based on AUC values, the best model is FRT, while the subsequent classifiers are, in order, regression tree, stepwise logit, logit, and KLR. AUC differences from FRT with corresponding $p$-value computed according to (18) (last column of the table) lead us to conclude that only KLR accuracy appears to be significantly lesser than FRT, while other competing models (although showing less in-sample accuracy) are not statistically outperformed by our base model.

To better understand the classification ability of the models implied by AUC, let us look at sensitivity (Sens) and specificity (Spec) computed using the best cut-off point ($\mathfrak{C}^*_{YI}$) based on maximized YI, that is reported in Panel A of Table 1. Except for KLR, which shows the lowest AUC with a better ability in predicting defaults (sensitivity), all other models obtain higher specificity than sensitivity by trading off between type-I and type-II errors, while maintaining good performance in classifying defaults and non-defaults. As is clear and pointed out in many studies (e.g., Fuertes and Kalotychou, 2007), validating an EWS strictly depends on the decision-maker's preferences. To this end, Panel B of Table 1 reports the loss function values computed for each model using risk-aversion weights ranging from 0.1 to 0.9. For each value of this weight, panel C shows the best and the worst classifiers based on the min and max loss function values. By assuming a range of values for risk aversion from 0.4 to 0.9, FRT is the best classifier while having low risk-aversion, and assuming from 0.1 to 0.3, regression tree is the best model. The worst classifiers are KLR and regression tree, depending on the risk-aversion level: from low to average risk-aversion KLR shows a higher loss function, and for high risk-aversion, regression tree is the worst model. Regression tree thus shows great instability depending on the type of target error. Further, using signal-based diagnostic FRT appears to be the best approach although this model is statistically significant only relative to KLR. Indeed, while showing better performance than regression tree, stepwise logit, logit, and KLR, our base model is statistically better only when compared to KLR.

## Forecasting accuracy

To compare the models on the basis of their ability to forecast out of the estimation sample, we focused on 1991–2010 while inspecting in more depth how the models performed during the 'big three' crises, namely the Mexican crisis of 1995, the Asian crisis of 1997–1998, and the global financial crisis of 2007–2010.

### ***Out-of-sample model ranking***

Table 2 presents different metrics computed over the entire holdout sample and a loss function analysis based on different risk-aversion targets. Using the same tests as those used to assess the models' reliability in-sample except for BIC, we computed the Diebold–Mariano forecast to compare the forecasting errors of FRT with alternative EWSes as in Section II.

Inspecting RMSE, Brier score, logarithmic probability score, and Diebold–Mariano forecast, we found strong evidence of FRT's superiority relative to competing models over the entire holdout sample. Based on these statistics, subsequent classifiers are, in order, stepwise logit, logit, KLR, and regression tree. Logistic regressions perform quite similarly, while regression tree and KLR seem to be less efficient.

AUC-based tests provide a clear view about FRT reliability in making predictions. Indeed, in the overall holdout period, FRT shows an area under the ROC curve of 0.8353 against values ranging from 0.6690 (KLR) to 0.7780 (logit). From RMSE, Brier score, logarithmic probability score, and Diebold–Mariano forecast, FRT outperforms logit, regression tree, and KLR. Sensitivity and specificity computed using the YI criterion show that our approach predicts 88 per cent of the default episodes and about 64 per cent of the non-defaults occurring in 1991–2010. On the other hand, competing classifiers are less efficient, except for regression tree which shows a very high value for sensitivity near to 0.94 but at the cost of poor non-default predictability. Sensitivity ranges from 71 (stepwise logit) to 78 (KLR) per cent, while specificity varies from 46 (regression tree) to 75 (stepwise logit) per cent (i.e., higher than FRT).

The difference between the AUC values and the corresponding $p$-values confirm that over the period 1991–2010, FRT significantly outperforms competing EWSes. This is certainly true for regression tree and KLR, for which the AUC differences are strongly significant, while for logistic regressions, FRT's superiority is very near to significance ($p$-values are 0.146 against logit and 0.116 against stepwise logit).

The loss function analysis extends these results providing some interesting insights on how accuracy perception changes with decision-makers' targets. Panel B of Table 2 reports the value for the loss function, assuming the same range for the risk-aversion level as in Section IV. When risk aversion is low ($\zeta < 0.4$), stepwise logit is the best model while a higher cost is associated with regression tree and KLR. However, as we move from average to high risk-aversion ($0.5 \leq \zeta \leq 0.7$), FRT dominates the competing EWSes over the entire holdout period, although for higher risk-aversion ($\zeta \geq 0.8$), regression tree is the best classifier, as it is obvious given its value for sensitivity. Interestingly, changes in the decision-makers' perspective make the performance of stepwise logit and regression tree significantly unstable, potentially moving from the best to the worst classifier and vice versa. Indeed, we observe that these models are alternatively ranked as best/worst performers depending on the risk-aversion level.

***Big crisis prediction***

Our out-of-sample analysis also includes a clinical study of major crises that occurred over the period 1991–2010. We focused on probability estimates of single big sovereign debt crises realized by the 5 competing models, and then on inspecting their ability in forecasting the actual defaults based on optimal cut-off points obtained through YI (the values of the best thresholds are reported in Panel A of Table 2). Table 3 reports which models correctly predicted the sovereign debt crises, as grouped in the following three clusters: (i) Mexican crisis of 1995; (ii) Asian crisis of 1997–1998; (iii) 2007–2010 global financial crisis. Table A5 in Appendix A reports in more detail the probability estimates for all the 49 crises occurring in the period 1991–2010. Looking at major crises, we note in Table 5 that FRT correctly forecasted all single events thus proving to be the best model also in this clinical study. The Mexican crisis was predicted by all models excluding KLR, and the 1995 Venezuela crisis was missed by stepwise logit. For the Asian crisis, all models were able to predict the single entry crisis except for Indonesia 1997 and Sri

Lanka 1997, which were correctly predicted only by FRT, regression tree, and KLR, and Korea 1997 that was missed by stepwise logit. Interestingly, the sovereign defaults that occurred during the 2007–2010 global financial crisis in Europe (Hungary 2008, Latvia 2008, Ukraine 2008, Greece 2010, Ireland 2010) were all predicted by FRT, regression tree (with the exception of Latvia 2008), and KLR, while logit regressions correctly forecasted only Latvia 2008. This point is relevant, since FRT, regression tree, and KLR are non-parametric models, and this signifies that only the approaches pertaining to the so-called 'algorithmic modelling' were able to identify the common latent root of the recent global financial crisis in Europe.

As discussed earlier for in-sample model estimates, FRT shows that strong negative real GDP growth together with low U.S. interest rates were the reason for the Greek and Irish crises. The unreported tree structure for the last FRT realized in the out-of-sample analysis (i.e., using data up to 2009 to make a prediction for 2010), confirmed the same path for both cases with the same splitting values for real GDP growth ($-1.5963$ per cent) and U.S. Treasury Bill rates (9.795 per cent).

## Fitting versus forecasting model accuracy

What we found in the empirical analysis is that on the one hand, our FRT appears to be quite as good a descriptor of past data, although the model's superiority is statistically significant only against KLR. On the other hand, when testing the models out-of-sample, FRT significantly outperforms competing EWSes, which are unstable when moving from in- to out-of-sample analysis (regression tree) and when risk-aversion targets change (regression tree and stepwise logit). Such a problem not only reflects on the use of the models (fitting vs. forecasting model separation), but also on a coherent evaluation procedure that would take into account both fitting and forecasting ability. By reconciling the 'two-sides' of model reliability, the question is how to provide a general framework in which in-sample and out-of-sample accuracy are balanced on the basis of the possible different targets of decision-makers.

As argued in Section II, to do this, we use $2^{\mathrm{D}}LF$ by simply computing a weighted average of loss function in- and out-of-sample with weights reflecting the decision-makers' objective function (data generating process vs. forecasting activity).

Table 4 reports the best model (Panel A) and the worst model (Panel B) based upon the values for $2^{\mathrm{D}}LF$ using (22) where $0.1 \leq \zeta \leq 0.9$ and $0.1 \leq \varrho \leq 0.9$ with step 0.1. To make the model comparison easier, we also report in Appendix A Figure A2, the bivariate function generated by $2^{\mathrm{D}}LF$ for each model. When moving from modest to high risk-aversion ($0.4 \leq \zeta \leq 0.9$), FRT appears to be the best model to use when exploring the data generating process and when making forecasts of future debt crises. Regression tree is ranked first only when decision-maker's function is strongly focused on forecasting targets ($0.1 \leq \varrho \leq 0.2$) with extreme risk-aversion. When moving from low to modest risk-aversion ($0.1 \leq \zeta \leq 0.3$), stepwise logit is the finest model for both fitting and forecasting sovereign defaults ($0.1 \leq \varrho \leq 0.8$) except for pure fitting targets where regression tree is the best performer ($0.8 \leq \varrho \leq 0.9$). On the other hand, excluding higher risk-aversion ($\varrho \geq 0.8$) in which regression tree and stepwise logit are worst for forecasting debt crises, KLR shows the highest cost for low and high risk-aversion, thereby yielding a different trade-off between fitting and forecasting ability.

The bivariate distribution depicted by $2^{\mathrm{D}}LF$ is thus particularly useful for comparing the models based on preferences expressed by the combinations of $\zeta$ and $\varrho$[24]. To put the issue into perspective, we ordered the values of $2^{\mathrm{D}}LF$ based on the combinations of $\zeta$ and $\varrho$ for each model, converting to the corresponding rank order the loss values of the models from 1 (best) to 5 (worst). In this way, we obtained the matrix $\mathbf{Q}$ with $E$ rows, which are the number of $\zeta - \varrho$ combinations (in our case $E = 9 \cdot 9 = 81$), and $H$ columns which are the number of models involved in the analysis (in our case $H = 5$). Each element of the matrix $\mathbf{Q}$ is denoted by $\mathfrak{r}_{eh}$ with $e = 1, \cdots, E$ and $h = 1, \cdots, H$; $\mathfrak{r}_{eh}$ is the rank for the $h$-th model based on the $e$-th combination of weights. Hence, for each $\zeta$ and $\varrho$, the model ranked as first takes the value 1 and so on to the worst model, which scores 5. To inspect the matrix $\mathbf{Q}$, we followed the common non-parametric statistics for ranks (Gibbons and Chakraborti, 2003). Specifically, we first computed a synthetic indicator to rank the models, and then used the paired Wilcoxon signed-rank test providing statistical significance to the model ranking obtained through such an indicator. The synthetic indicator for each model is

$$\pi_h = \frac{E \cdot H - \mathfrak{R}_h}{E \cdot H - E}, \quad \pi_h \in [0, 1], \tag{23}$$

where $\mathfrak{R}_h = \sum_{e=1}^{E} \mathfrak{r}_{he}$ is the sum of the ranks for model $h$. Dividing (23) by $E$ yields

$$\pi_h = \frac{H - \overline{\mathfrak{R}}_h}{H - 1}, \quad \pi_h \in [0, 1], \tag{24}$$

where $\overline{\mathfrak{R}}_h$ is the mean rank for model $h$, and $1 \leq \overline{\mathfrak{R}}_h \leq H$. If a model was rated the best for each combination of weights, $\overline{\mathfrak{R}}_h$ would be equal to 1. On the contrary, if a model was the worst for each combination of $\zeta$ and $\varrho$, $\overline{\mathfrak{R}}_h$ would be equal to $H$. In this way, whenever a model is rated as the best, (24) takes value 1, and takes 0 when the model is

[24] In our analysis, $\zeta$ and $\varrho$ range from 0.1 to 0.9 with step 0.1, thus having $9 \cdot 9 = 81$ different combinations of weights.

rated as the worst.

In Table 5, we report the value for $\pi_h$ with corresponding $\overline{\mathfrak{R}}_h$ together with paired Wilcoxon statistics. FRT is ranked first, and paired comparison through Wilcoxon statistic shows strong significance against all competing models. Logit and stepwise logit are ranked second and third, respectively, while regression tree is ranked fourth. KLR is ranked fifth, and clearly exhibits the worst performance relative to other classifiers.

The main message coming from the $2^{\mathrm{D}}LF$ analysis is that FRT seems to be the best model for both fitting and predicting debt crises while exhibiting quite stable performance by changing possible decision-makers' targets. On this point, see Appendix A Figure A2, in which we report the box-plots using the values for $2^{\mathrm{D}}LF$. As we note, the median cost for FRT is the lowest, and as such, FRT exhibits low dispersion relative to competing models.

As a result, FRT may provide a possible reconciling solution to the fitting versus forecasting paradox. Indeed, through the trade-off between fitting ability and forecasting ability implied in the cross-validation estimation technique, together with the penalization imposed for model complexity, which in turns reflects a simple model structure and a parsimonious number of parameters, the FRT: (i) provides an accurate description of past data, and near to be best description; (ii) produces the best forecasts, while also adapting to different risk aversion targets.

Note that this is a 'global' and *objective* evaluation of the model obtained using *subjective* preferences. In other words, starting from subjective evaluations about in- and out-of sample model reliability, we come to select the best model by averaging fitting and forecasting ability together with low and high risk-aversion. In this sense, the meaning we attribute to the term *best* has to be interpreted as the *best average model*.

# V. Conclusion

In this paper, we consider the problem of fitting and predicting sovereign debt crises in light of the forecasting versus policy dilemma introduced in Clements and Hendry (1998). The accepted wisdom is that simple models outperform more complex models in terms of forecast accuracy although the latter provide a better description of sovereign debt default data (Fuertes and Kalotychou, 2006). To this end, we introduce a regression tree-based model using a two-step procedure in which, in the first step, we generate multiple predictions by cross-validating the model on rotated sub-samples until the average of the estimates stabilizes, and in the second step, we fit a regression tree using such an average as the dependent variable. This two-step procedure entails a trade-off between fitting ability and forecasting ability, while imposing penalization for model complexity, thereby producing a simple model structure with a parametric parsimony that provides an accurate description of past crises and good forecasts of future defaults.

Using data from emerging markets and GIPS over 1975–2010, we run several statistical metrics to assess the model reliability in- and out-of-sample relative to the existing state-of-the-art models (logit, stepwise logit, regression tree, KLR), and show that our methodology significantly outperforms competing models when in-sample and out-of-sample accuracy are jointly considered. The trade-off between fitting and forecasting ability translates into a compromise that favours forecasting ability while maintaining a good description of the data generating process.

The investigation of the economic side of our results supports three main findings. First, we found that illiquidity (short-term debt to reserves), macroeconomic (U.S. Treasury Bill rate, real GDP growth, and exchange rate undervaluation), and political (default history) risks are the main determinants and predictors of past and future debt crises. Second, we proved that the root of the recent Greek and Irish sovereign debt crises has been the same as that of other emerging market crises that occurred in the past (such as Turkey in 2002, Ukraine in 1998, and Venezuela in 1995), namely the strong contraction in GDP growth together with low interest rates and a bad default history. Third, the European sovereign defaults of the 2007–2010 global financial crisis were predicted only by non-parametric approaches, while traditional logit regressions failed to signal the deterioration of economic conditions in those regions which then went into default.

Last, we comment on the contagion variable used in our empirical analysis, since we used a proxy which was contemporaneous to the default indicator, according to prevalent literature. To be more realistic and to provide a pure forecasting model, we should use an expectation of contagion for period $t$ observed in $t-1$. Hence, a contagion should be explored, first, as a dependent variable, and second, as a potential predictor. This is left our future research.

# References


Alessi, L. and Detken, C. (2011). 'Quasi Real Time Early Warning Indicators for Costly Asset Price Boom/Bust Cycles: A Role for Global Liquidity', *European Journal of Political Economy*, Vol. 27, pp. 520-533.

Austin, P. and Tu, J. (2004). 'Bootstrap Methods for Developing Predictive Models', *The American Statistician*, Vol. 58, pp. 131-137.

Bamber, D. (1975). 'The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph', *Journal of Mathematical Psychology*, Vol. 12, pp. 387-415.

Breiman, L. (2001). 'Random Forests', *Machine Learning*, Vol. 45, pp. 5-32.

Catão, L. and Sutton, B. (2002). 'Sovereign Defaults: The Role of Volatility', *IMF Working Paper*, 02/149.

Cavallo, E. A. and Frankel, J. A. (2008). 'Does Openness to Trade Make Countries More Vulnerable to Sudden Stops, or Less? Using Gravity to Establish Causality', *Journal of International Money and Finance*, Vol. 27, pp. 1430-1452.

Clements, M. and Hendry, D. (1998). *Forecasting Economic Time Series*, Cambridge University Press, Cambridge.

De Grauwe, P. and Ji, Y. (2012). 'Mispricing of Sovereign Risk and Multiple Equilibria in the Eurozone', *CEPS Working Document*, 361.

Debashis, P., Bair, E., Hastie, T. and Tibshirani, R. (2008). "Preconditioning' for Feature Selection and Regression in High-Dimensional Problems', *Annals of Statistics*, Vol. 36, pp. 1595-1618.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988). 'Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach', *Biometrics*, Vol. 44, pp. 837-845.

Detragiache, E. and Spilimbergo, A. (2001). 'Crises and Liquidity: Evidence and Interpretation', *IMF Working Paper*, 01/2.

Diebold, F. and Lopez, J. (1996). 'Forecast Evaluation and Combination', in G. Maddala and C. Rao (ed.), *Statistical Methods of Finance, Handbook of Statistics Series*, Elsevier, North-Holland, Amsterdam, New York, pp. 241-268.

Diebold, F. and Mariano, R. (1995). 'Comparing Predictive Accuracy', *Journal of Business and Economic Statistics*, Vol. 13, pp. 253-263.

Dooley, M. (2000). 'A Model of Crises in Emerging Markets', *The Economic Journal*, Vol. 110, pp. 256-272.

Duffie, D., Pedersen, L. and Singleton, K. (2003). 'Modeling Sovereign Yield Spreads: A Case Study of Russian Debt', *Journal of Finance*, Vol. 58, pp. 119-159.

Eichengreen, B. and Rose, A. (1999). 'The Empiric of Currency and Banking Crises', in *NBER Reporter*.

Eichengreen, B., Rose, A. and Wyplosz, C. (1996). 'Contagious Currency Crises: First Tests', *Scandinavian Journal of Economics*, Vol. 98, pp. 463-84.

Fuertes, A. and Kalotychou, E. (2006). 'Early Warning Systems for Sovereign Debt Crises: The Role of Heterogeneity', *Computational Statistics and Data Analysis*, Vol. 51, pp. 1420-1441.

Fuertes, A. and Kalotychou, E. (2007). 'Optimal Design of Early Warning Systems for Sovereign Debt Crises', *International Journal of Forecasting*, Vol. 23, pp. 85-100.

Gapen, M., Gray, D., Lim, C. and Xiao, Y. (2005). 'Measuring and Analyzing Sovereign Risk with Contingent Claims', *IMF Working Paper*, 05/155.

Gibbons, J. and Chakraborti, S. (2003). *Nonparametric Statistical Inference*, Marcel Dekker.

Goldstein, M., Kaminsky, G. and Reinhart, C. (2000). *Assessing Financial Vulnerability: An Early Warning System for Emerging Markets*, Institute for International Economics.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer, Berlin.

Hilden, J. and Glasziou, P. (1996). 'Regret Graphs, Diagnostic Uncertainty, and Youden's Index', *Statistics in Medicine*, Vol. 15, pp. 969-986.

Honaker, J. and King, G. (2010). 'What to Do about Missing Values in Time-Series Cross-Section Data', *American Journal of Political Science*, Vol. 54, pp. 561-581.

Kaminsky, G. L. (1998). 'Currency and Banking Crises: The Early Warnings of Distress', *FED International Finance Discussion Papers*, 629.

Kaminsky, G. L., Lizondo, S. and Reinhart, C. (1998). 'Leading Indicators of Currency Crises', *IMF Staff Papers*, 45, pp. 1-48.

Kaminsky, G. L. and Reinhart, C. (2000). 'On Crises, Contagion and Confusion', *Journal of International Economics*, Vol. 51, pp. 145-168.

King, G., Honaker, J., Joseph, A. and Scheve, K. (2001). 'Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation', *American Political Science Review*, Vol. 95, pp. 49-69.

Manasse, P. and Roubini, N. (2009). 'Rules of Thumb for Sovereign Debt Crises', *Journal of International Economics*, Vol. 78, pp. 192-205.

Manasse, P., Roubini, N. and Schimmelpfennig, A. (2003). 'Predicting Sovereign Debt Crises', *IMF Working Paper*, 03/221.

McFadden, D., Eckaus, R., Feder, G., Hajivassiliou, V. and O'Connell S. (1985). 'Is There Life after Debt? An Econometric Analysis of the Creditworthiness of Developing Countries', in A. Smith, and J. T. Cuddington (ed.), *International Debt and the Developing Countries*, pp. 179-209.

Mulder, C., Perilli, R. and Rocha, M. (2002). 'The Role of Corporate, Legal, and Macroeconomic Balance Sheet Indicators in Crisis Detection and Prevention', *IMF Working Paper*, 02/59.

Oral, M., Kettani, O., Cosset, J. and Daouas, M. (1992). 'An Estimation Model for Country Risk Rating', *International Journal of Forecasting*, Vol. 8, pp. 583-593.

Orrel, D. and McSharry, P. (2009). 'System Economics: Overcoming the Pitfalls of Forecasting Models via a Multidisciplinary Approach', *International Journal of Forecasting*, Vol. 25, pp. 734-743.

Reinhart, C. and Rogoff, K. (2010). 'Growth in a Time of Debt', *NBER Working Paper* No 15639, Cambridge, MA.

Reinhart, C., Rogoff, K. and Savastano, M. (2003). 'Debt Intolerance', *Brookings Papers on Economic Activity*, Vol. 1, pp. 1-74.

Savona, R. and Vezzoli, M. (2012). 'Multidimensional Distance to Collapse Point and

Sovereign Default Prediction', *Intelligent Systems in Accounting, Finance and Management*, Vol. 19, pp. 205-228.

Sturzenegger, F. (2004). 'Toolkit for the Analysis of Debt Problems', *Journal of Restructuring Finance*, Vol. 1, pp. 201-203.

Taffler, R. and Abassi, B. (1984). 'Country Risk: A Model for Predicting Debt-Servicing Problems in Developing Countries', *Journal of the Royal Statistical Society - Series A*, Vol. 147, pp. 541-568.

Vezzoli, M. and Stone, C. J. (2007). 'Cragging', *Book of Short Papers CLADAG 2007*, University of Macerata, September 12-14.

## TABLE 1

*In-Sample Model Accuracy*

| Panel A: Diagnostics | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | *BIC* | *RMSE* | *BS* | *LPS* | *YI* | *Cut-off** | *Sens* | *Spec* | *AUC* | *AUC diff* |
| FRT | -6259.493 | 0.2048 | 0.0839 | 0.1528 | 0.6374 | 9.90% | 0.7869 | 0.8505 | 0.8914 | - |
| RT | -6726.621 | 0.1817 | 0.0661 | 0.1300 | 0.5710 | 14.40% | 0.6230 | 0.9480 | 0.8855 | 0.0059 (0.436) |
| S_Logit | -6083.459 | 0.2112 | 0.0892 | 0.1641 | 0.5906 | 8.30% | 0.7213 | 0.8692 | 0.8729 | 0.0185 (0.223) |
| Logit | -6011.092 | 0.2094 | 0.0877 | 0.1626 | 0.6086 | 8.30% | 0.7705 | 0.8382 | 0.8725 | 0.0189 (0.209) |
| KLR | -5369.139 | 0.2480 | 0.1230 | 0.2595 | 0.3492 | 12.40% | 0.7951 | 0.5541 | 0.7345 | 0.1569 (0.000) |

| Panel B: Loss Values | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Risk Aversion Parameter | | | | | | | | |
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| FRT | 0.1559 | 0.1622 | 0.1686 | 0.1750 | 0.1813 | 0.1877 | 0.1940 | 0.2004 | 0.2068 |
| RT | 0.0845 | 0.1170 | 0.1495 | 0.1820 | 0.2145 | 0.2470 | 0.2795 | 0.3120 | 0.3445 |
| S_Logit | 0.1456 | 0.1603 | 0.1751 | 0.1899 | 0.2047 | 0.2195 | 0.2343 | 0.2491 | 0.2639 |
| Logit | 0.1686 | 0.1754 | 0.1821 | 0.1889 | 0.1957 | 0.2024 | 0.2092 | 0.2160 | 0.2227 |
| KLR | 0.4218 | 0.3977 | 0.3736 | 0.3495 | 0.3254 | 0.3013 | 0.2772 | 0.2531 | 0.2290 |

| Panel C: Best vs. Worst | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Risk Aversion Parameter | | | | | | | | |
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| Min Loss | RT | RT | RT | FRT | FRT | FRT | FRT | FRT | FRT |
| Max Loss | KLR | KLR | KLR | KLR | KLR | KLR | RT | RT | RT |

*Notes*: The table shows the diagnostics used to assess the models' accuracy over the entire period 1975–2010. *BIC* is the Bayesian information criterion, *RMSE* is the root mean squared error, *BS* is the Brier Score, *LPS* is the Logarithmic Probability Score, *YI* is the Youden Index and *Cut-off** is the corresponding probability value used to maximize the *YI*. *Sens* and *Spec* are the sensitivity (1 *minus* type I error) and the specificity (1 *minus* type II error) computed using the *Cut-off**. *AUC* is the area under the *ROC* curve and the *AUC diff* are pairwise differences with corresponding *p*-values in parentheses computed according to DeLong et al. (1988). In panel B we report the loss values (*LF*) over the entire period 1975–2010, while panel C reports the best and worst model conditional on specific risk aversion level, i.e., the models showing the lesser (Best) and the higher (Worst) value of the *LF*.

## TABLE 2

*Out-Of-Sample Model Accuracy*

| Panel A: Diagnostics | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | *RMSE* | *DM* | *BS* | *LPS* | *YI* | *Cut-off** | *Sens* | *Spec* | *AUC* | *AUC diff* |
| FRT | 0.1920 | - | 0.0737 | 0.1467 | 0.5134 | 5.50% | 0.8776 | 0.6359 | 0.8353 | - |
| Logit | 0.2140 | -2.6503 (0.008) | 0.0916 | 0.2121 | 0.4731 | 3.60% | 0.7551 | 0.7179 | 0.7720 | 0.0633 (0.146) |
| S_Logit | 0.2126 | -2.5812 (0.010) | 0.0904 | 0.2104 | 0.4621 | 3.60% | 0.7143 | 0.7479 | 0.7661 | 0.0692 (0.116) |
| RT | 0.2978 | -8.3716 (0.000) | 0.1774 | 0.5466 | 0.3995 | 5.00% | 0.9388 | 0.4607 | 0.6768 | 0.1585 (0.000) |
| KLR | 0.2101 | -4.6074 (0.000) | 0.0883 | 0.2262 | 0.3029 | 5.88% | 0.7755 | 0.5274 | 0.6690 | 0.1663 (0.000) |

| Panel B: Loss Values | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Risk Aversion Parameter | | | | | | | | |
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| FRT | 0.3399 | 0.3158 | 0.2916 | 0.2674 | 0.2433 | 0.2191 | 0.1949 | 0.1708 | 0.1466 |
| Logit | 0.2783 | 0.2746 | 0.2709 | 0.2672 | 0.2635 | 0.2598 | 0.2560 | 0.2523 | 0.2486 |
| S_Logit | 0.2555 | 0.2589 | 0.2622 | 0.2656 | 0.2689 | 0.2723 | 0.2756 | 0.2790 | 0.2824 |
| RT | 0.4915 | 0.4437 | 0.3959 | 0.3481 | 0.3003 | 0.2525 | 0.2047 | 0.1568 | 0.1090 |
| KLR | 0.4478 | 0.4230 | 0.3982 | 0.3734 | 0.3486 | 0.3238 | 0.2989 | 0.2741 | 0.2493 |

| Panel C: Best vs. Worst | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Risk Aversion Parameter | | | | | | | | |
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| Min Loss | S_Logit | S_Logit | S_Logit | S_Logit | FRT | FRT | FRT | RT | RT |
| Max Loss | RT | RT | KLR | KLR | KLR | KLR | KLR | S_Logit | S_Logit |

*Notes*: The table shows the same diagnostics used in Table 3 to assess the models' accuracy over the out-of-sample 1991–2010, also including the Diebold-Mariano test (column *DM*).

## TABLE 3

### *Big Crisis Prediction*

| Debt Crisis | Correctly Predicted Defaults – Models | | | | |
|---|---|---|---|---|---|
| *Mexican Crisis* | | | | | |
| Mexico 1995 | FRT | Logit | S_Logit | RT | |
| Venezuela 1995 | FRT | Logit | | RT | |
| *Asian Crisis* | | | | | |
| Indonesia 1997 | FRT | | | RT | KLR |
| Korea, Rep. 1997 | FRT | Logit | | RT | KLR |
| Sierra Leone 1997 | FRT | Logit | S_Logit | RT | KLR |
| Sri Lanka 1997 | FRT | | | RT | KLR |
| Thailand 1997 | FRT | Logit | S_Logit | RT | KLR |
| Argentina 1998 | FRT | Logit | S_Logit | RT | KLR |
| Brazil 1998 | FRT | Logit | S_Logit | RT | KLR |
| Moldova 1998 | FRT | Logit | S_Logit | RT | KLR |
| Pakistan 1998 | FRT | Logit | S_Logit | RT | KLR |
| Philippines 1998 | FRT | Logit | S_Logit | RT | KLR |
| Ukraine 1998 | FRT | Logit | S_Logit | RT | KLR |
| *2007-2010 Crisis* | | | | | |
| Ecuador 2008 | FRT | Logit | S_Logit | | KLR |
| Hungary 2008 | FRT | | | RT | KLR |
| Latvia 2008 | FRT | Logit | S_Logit | | KLR |
| Pakistan 2008 | FRT | Logit | S_Logit | | KLR |
| Ukraine 2008 | FRT | | | RT | KLR |
| Greece 2010 | FRT | | | RT | KLR |
| Ireland 2010 | FRT | | | RT | KLR |
| Jamaica 2010 | FRT | Logit | S_Logit | RT | KLR |

*Notes*: The table reports the models that correctly predicted the Mexican, Asian and 2007-2010 crises.

## TABLE 4

### *$2^D$ Loss Function*

| Panel A: Best Models | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| In- vs. Out | Risk Aversion Parameter | | | | | | | | |
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| 0.9 | RT | RT | RT | FRT | FRT | FRT | FRT | FRT | FRT |
| 0.8 | RT | S_Logit | S_Logit | FRT | FRT | FRT | FRT | FRT | FRT |
| 0.7 | S_Logit | S_Logit | S_Logit | FRT | FRT | FRT | FRT | FRT | FRT |
| 0.6 | S_Logit | S_Logit | S_Logit | FRT | FRT | FRT | FRT | FRT | FRT |
| 0.5 | S_Logit | S_Logit | S_Logit | FRT | FRT | FRT | FRT | FRT | FRT |
| 0.4 | S_Logit | S_Logit | S_Logit | FRT | FRT | FRT | FRT | FRT | FRT |
| 0.3 | S_Logit | S_Logit | S_Logit | FRT | FRT | FRT | FRT | FRT | FRT |
| 0.2 | S_Logit | S_Logit | S_Logit | FRT | FRT | FRT | FRT | FRT | RT |
| 0.1 | S_Logit | S_Logit | S_Logit | S_Logit | FRT | FRT | FRT | RT | RT |
| Panel B: Worst Models | | | | | | | | | |
| In- vs. Out | Risk Aversion Parameter | | | | | | | | |
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| 0.9 | KLR | KLR | KLR | KLR | KLR | KLR | KLR | RT | RT |
| 0.8 | KLR | KLR | KLR | KLR | KLR | KLR | KLR | RT | RT |
| 0.7 | KLR | KLR | KLR | KLR | KLR | KLR | KLR | RT | RT |
| 0.6 | KLR | KLR | KLR | KLR | KLR | KLR | KLR | KLR | S_Logit |
| 0.5 | KLR | KLR | KLR | KLR | KLR | KLR | KLR | S_Logit | S_Logit |
| 0.4 | KLR | KLR | KLR | KLR | KLR | KLR | KLR | S_Logit | S_Logit |
| 0.3 | KLR | KLR | KLR | KLR | KLR | KLR | KLR | S_Logit | S_Logit |
| 0.2 | KLR | KLR | KLR | KLR | KLR | KLR | KLR | S_Logit | S_Logit |
| 0.1 | RT | KLR | KLR | KLR | KLR | KLR | KLR | S_Logit | S_Logit |

*Notes*: In this table we report the best and the worst model based on minimum and maximum values of the $2^D LF$ computed for different in- vs. out-of-sample weights and risk aversion parameter combinations.

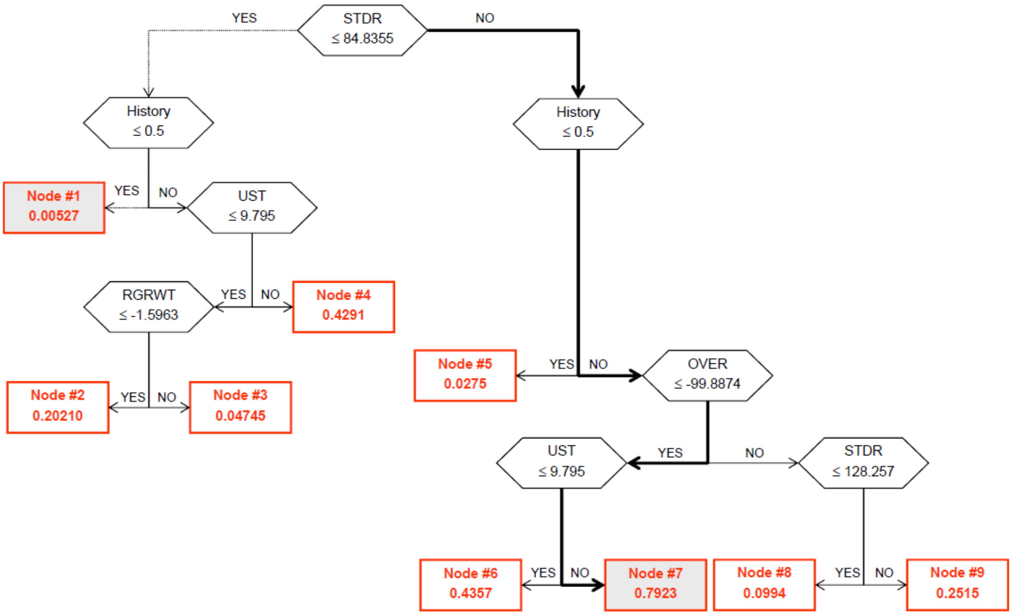
TABLE 5

*Rank Comparison*

| Model | R-mean | π | W-stat |
|---|---|---|---|
| FRT | 1.7037 | 0.8241 | - |
| Logit | 2.3951 | 0.6512 | -4.58*** |
| S_Logit | 2.7531 | 0.5617 | -3.96*** |
| RT | 3.4198 | 0.3951 | -7.26*** |
| KLR | 4.7284 | 0.0679 | -7.82*** |

*Notes*: The table reports the value for the mean rank for each model (R-mean) and corresponding $\pi$ computed according to (4.24). W-stat is the paired Wilcoxon statistics with ***, **, * denoting significance at 0.01, 0.05, 0.1 levels.

Figure 1: FRT




*Notes*: The figure depicts the structure of the FRT estimated over the period 1975–2010. For each split, we specify the variable and the corresponding threshold. The values within each terminal node are the estimated probabilities of default (PD). The most risky and the safest nodes are indicated by the grey area also highlighting the paths towards the higher (bold line) and the lesser (dashed line) PD.

# Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals*

# –Appendix –

Roberto Savona† and Marika Vezzoli‡

This draft: June 2013

* The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 320270 - SYRTO.

† *Department of Economics and Management, University of Brescia, c/da S. Chiara 50, 25122 Brescia, Italy, (e-mail: savona@eco.unibs.it)*

‡ *Department of Molecular and Translational Medicine, University of Brescia, Viale Europa 11, 25123 Brescia, Italy, (e-mail: marika.vezzoli@med.unibs.it)*

## Supplementary Material

Appendix A presents the data used in the paper and additional results of empirical analysis, while Appendix B describes the CRAGGING algorithm used to realize our early warning system.

## Appendix A

### A.I. Data Description

In Table A1 we reports the 122 crisis episodes used in the empirical analysis, of which 49 are used in the out-of-sample analysis over the sub-period 1991–2010.

Columns 6–8 of Table A2 report the sample mean for non-crisis and crisis states and their $t$– and $z$–statistic [1] in order to get preliminary results on the discriminatory power of each predictor, and the last column reports the variance inflation factor (VIF) to check for multicollinearity. Heuristically, it is common practice in the statistical community to consider VIFs greater than 5 or 10 as an indicator of multicollinearity problems. Based on these values, we note that TEDY suffers from multicollinearity, thus reflecting a potential misinterpretation about the impact such a variable exerts on the dependent variable while controlling for the others.

Except for FDI inflow variations (FDIG) and exchange rate overvaluation (OVER), the mean differential is statistically significant for all predictors, thus providing strong 'univariate' ability in signalling sovereign defaults.

The original dataset had a number of missing values for many of the independent variables used in the analysis. To control for such a problem, we used the multiple imputation technique proposed in Honaker and King (2010) which updated the original approach introduced in King et al. (2001). The procedure (Amelia II) uses the expectation-maximization (EM) algorithm on multiple bootstrapped samples of the original incomplete data to draw values of the complete-data parameters. Once all relevant information from observed data is extracted, the algorithm draws imputed values from each set of bootstrapped parameters, replacing the missing values for each one missing, ranging from 5 to 10, then averaging them to obtain the point estimates to fill in the missing cell[2].

---

[1]The $z$–test was performed for dummy variables, namely for MAC, DOIL, and IMF.

[2]The routine is written in R code and is available at http://gking.harvard.edu/amelia/.

Computationally, we carried out a multiple imputation technique controlling for time series cross sectionally, while imposing empirical beliefs so as to shrink the posterior of the point estimate of each missing cell and the point estimate for each country within the specific historical range (for technical details, see Honaker and King (2010) and King et al. (2001)).

Summary statistics of missing values too are given in Table A2. Columns 3–5 report for each predictor, the missing values' proportions, the mean of the observed values $\mu$, and the mean of the missing value estimates $E(\mu)$. Note that the percentage of missing values never exceeds 0.13 over the 1988 observations and that the means of missing estimates are quite similar to those of the observed values for all predictors.

TABLE A1

*Debt Crisis*

| Year | # of Crises | Countries |
|---|---|---|
| 1975 | 2 | Kenya, Zimbabwe, |
| 1976 | 1 | Peru |
| 1977 | 2 | Jamaica, Mexico |
| 1978 | 4 | Egypt, Peru, Turkey, Zambia |
| 1979 | 5 | Honduras, Kenya, Malawi, Mauritius, Nicaragua |
| 1980 | 8 | Bangladesh, Bolivia, Costa Rica, Korea, Madagascar, Morocco, Pakistan, Philippines |
| 1981 | 10 | Dominican Republic, El Salvador, Ethiopia, Honduras, India, Jamaica, Poland, Romania, Thailand, Zambia |
| 1982 | 10 | Argentina, Ecuador, Haiti, Hungary, Kenya, Malawi, Mexico, Nigeria, Peru, Turkey |
| 1983 | 12 | Brazil, Burkina Faso, Chile, Korea, Mauritius, Niger, Philippines, Sierra Leone, Uruguay, Venezuela, Zambia, Zimbabwe |
| 1984 | 1 | Egypt |
| 1985 | 3 | Cameroon, South Africa, Thailand |
| 1986 | 7 | Bolivia, Gabon, Madagascar, Morocco, Paraguay, Romania, Sierra Leone |
| 1987 | 2 | Jamaica, Uruguay |
| 1988 | 3 | Malawi, Trinidad and Tobago, Tunisia |
| 1989 | 2 | Jordan, South Africa |
| 1990 | 1 | Uruguay |
| 1991 | 3 | Algeria, Ethiopia, Hungary |
| 1992 | 1 | Zimbabwe |
| 1993 | 1 | South Africa |
| 1994 | 4 | Kenya, Lithuania, Philippines, Turkey |
| 1995 | 2 | Mexico, Venezuela |
| 1996 | 3 | Jordan, Kazakhstan, Moldova |
| 1997 | 5 | Indonesia, Korea, Sierra Leone, Sri Lanka, Thailand |
| 1998 | 6 | Argentina, Brazil, Moldova, Pakistan, Philippines, Ukraine |
| 1999 | 4 | Ecuador, Gabon, Mexico, Turkey |
| 2000 | 3 | Argentina, Uruguay, Zimbabwe |
| 2001 | 1 | Brazil |
| 2002 | 6 | Gabon, Indonesia, Moldova, Paraguay, Turkey, Uruguay |
| 2004 | 1 | Cameroon |
| 2005 | 1 | Venezuela |
| 2008 | 5 | Ecuador, Hungary, Latvia, Pakistan, Ukraine |
| 2010 | 3 | Greece, Ireland, Jamaica |
| 1975–2010 | 122 | |
| 1991–2010 | 49 | |

*Notes*: The table reports the sovereign defaults analysed over the period 1975–2010 while specifying for each year, the number of debt crises as well as the countries classified as defaulters.

TABLE A2

*Predictors*

| Variable | Definition | Missing Value Statistics | | | Mean | | t/z-test | VIF |
|---|---|---|---|---|---|---|---|---|
| | | Missing | $\mu$ | E($\mu$) | Non-Crisis | Crisis | | |
| Insolvency Risk Factors | | | | | | | | |
| MAC | Market access dummy | - | 0.702 | - | 0.697 | 0.779 | 1.251 | - |
| IMF | IMF lending dummy | - | 0.009 | - | 0.009 | 0.000 | -11.401*** | - |
| CAY | Current account (% of GDP) | 0.122 | -2.853 | -3.054 | -2.738 | -5.018 | -8.658*** | 1.280 |
| ResG | Reserves % change | 0.040 | 21.949 | 29.944 | 23.579 | 1.832 | -12.702*** | 1.029 |
| XG | Exports % change | 0.108 | 5.852 | 5.441 | 6.005 | 2.790 | -8.226*** | 1.214 |
| WX | Exports in USD billions | 0.109 | 23.831 | 15.172 | 23.285 | 16.770 | -6.864*** | 1.268 |
| TEDX | Total debt to exports (%) | 0.102 | 170.849 | 205.454 | 171.539 | 217.877 | 9.970*** | 2.619 |
| MG | Imports % change | 0.095 | 6.383 | 5.764 | 6.465 | 4.179 | -4.715*** | 1.384 |
| FDIY | FDI inflows to GDP (%) | 0.078 | 2.391 | 2.864 | 2.461 | 1.670 | -3.929*** | 1.369 |
| FDIG | FDI inflows % change | 0.090 | 43.182 | 154.852 | 54.743 | 46.415 | -0.935 | 1.027 |
| TEDY | Total ext. debt to GDP (%) | 0.029 | 39.708 | 33.103 | 38.901 | 48.966 | 10.876*** | 5.054 |
| SEDY | Short term debt to GDP (%) | 0.039 | 4.687 | 4.984 | 4.489 | 7.912 | 11.891*** | 1.748 |
| PEDY | Public debt to GDP (%) | 0.058 | 52.697 | 40.153 | 51.722 | 55.790 | 4.082*** | 2.470 |
| OPEN | Exports + imports to GDP (%) | 0.041 | 71.792 | 66.078 | 72.288 | 60.402 | -13.345*** | 1.616 |
| Illiquidity Risk Factors | | | | | | | | |
| STDR | Short term debt to reserves (%) | 0.018 | 72.746 | 82.569 | 64.352 | 203.954 | 14.002*** | 2.666 |
| M2R | M2 to reserves | 0.035 | 37.428 | 35.194 | 34.438 | 81.890 | 3.100*** | 1.019 |
| DSER | Debt service on L-T debt to reserves | 0.025 | 0.598 | 0.664 | 0.535 | 1.586 | 8.155*** | 2.591 |
| Macroeconomic Risk Factors | | | | | | | | |
| DOil | Oil producing dummy | - | 0.092 | - | 0.093 | 0.074 | -2.264** | - |
| INF | Inflation (%) | 0.118 | 41.823 | 38.379 | 36.228 | 120.753 | 3.435*** | 1.016 |
| RGRWT | Real GDP % change | - | 3.866 | - | 3.990 | 1.971 | -8.193*** | 1.401 |
| OVER | Exch. rate residual over linear trend | - | -21.888 | - | -22.067 | -19.148 | 0.367 | 1.010 |
| UST | US Treasury Bill | - | 5.359 | - | 5.228 | 7.362 | 12.193*** | 2.093 |
| EU | EU dummy | - | - | - | - | - | - | - |
| Political Risk Factors | | | | | | | | |
| PR | Index of political rights | 0.034 | 3.616 | 3.743 | 3.602 | 3.910 | 2.329** | 1.267 |
| History | # of past defaults | - | 0.999 | - | 0.946 | 1.820 | 8.770*** | 1.365 |
| Systemic Risk Factors | | | | | | | | |
| Cont_tot | Contagion | - | 3.219 | - | 3.102 | 5.016 | 11.238*** | 2.147 |
| Cont_area | Regional contagion | - | 0.724 | - | 0.689 | 1.262 | 5.819*** | 1.640 |

*Notes:* This table reports summary statistics of the potential predictors of debt crises. Missing denotes the percentage of missing values over the total number of observations (1988), $\mu$ is the mean of each predictor computed using the observed data, and E($\mu$) is the mean computed using the point estimates obtained through the multiple imputation technique. Mean is the average conditional upon the default state (Non-crisis and Crisis) and t/z-statistic is computed on the mean difference between Crisis and Non-crisis: the z test is for dummy variables (CAY, IMF, DOIL). ***, **, * denote significance at the 0.001, 0.05, and 0.1 levels. VIF is the variance inflation factor obtained as $1/(1-R^2)$, where $R^2$ is obtained by regressing each predictor one at a time using the remaining ones as explanatory variables. VIF values exceeding 5 or 10 indicate a multicollinearity problem.

## A.II. Additional Results

### A.II.1. Alternative Models and Different Explanations

Here we report in-sample estimates of logit, stepwise logit, KLR, and Regression Tree also discussing the different economic explanations of the sovereign default underlying each model.

### Logit and Stepwise Logit

Columns 2-3 of Table A3 report logit and stepwise logit model estimates. To make the results obtained through logistic regression more informative, we also run a variable selection process so as to better explain the economic message of the models. Specifically, we used the bootstrap method introduced in Austin and Tu (2004), taking 3,000 randomly selected sub-samples with each constituting 90 per cent of the total observations, running the stepwise logit on each bootstrap sample including all the 27 candidate variables; then, the predictors are ordered according to their importance, where the variable chosen most frequently is ranked first and so on. Arbitrarily putting the cut-off point at 90 per cent to select the most important predictors, 12 variables appear as the most relevant. Insolvency risk proxies are the major factors, all exhibiting statistical significance except for IMF lending. These are (with corresponding estimated signs): market access $(+)$, current account $(-)$, IMF lending $(+)$, short-term debt to GDP $(+)$, total ext. debt to GDP $(-)$, FDI inflows to GDP $(-)$, and exports $(-)$. Furthermore, other proxies for illiquidity, macroeconomic, and political risk factors are also important, namely, debt service on long-term debt to reserves $(+)$, U.S. Treasury Bill rates $(+)$, real GDP growth $(-)$, and finally, default history $(+)$ with index of political rights $(+)$. In general, logistic regression results are quite consistent with FRT, especially for the most important variables selected by the Austin-Tu procedure. Total ext. debt to GDP shows an anomalous negative sign, maybe due to the positive correlation with short-term debt to GDP, which instead shows a positive coefficient and is ranked as the best explanatory variable together with default history, current account to GDP, IMF lending, and U.S. Treasury Bill rates, all being selected in each of the 3,000 stepwise logit estimations.

## KLR

In Table A4, we report the noise-to-signal ratio for each predictor according to the KLR methodology, also showing 1 *minus* type-I errors (sensitivity) and 1 *minus* type-II errors (specificity). As discussed in Section 2.3.2, the inverse of the optimal noise-to-signal ratio is the weight to be used in calculating the CI index. Then, such a weight gives us the variable importance of each predictor attributed by KLR. The methodology did not exclude any non-dummy variables[3]: all show $\omega < 1$, which is the constraint used for dropping noisy predictors. According to KLR, the risk signals implied in the debt service on long-term debt to reserves are the most informative as documented by the relative weight that accounts for nearly 27 per cent with respect to the remaining predictors. Short-term debt to reserves is also important (accounting for 13 per cent), and next we have short-term debt to GDP, inflation, M2 to reserves, and FDI inflows to GDP (all showing the same variable importance weight). Together with contagion, which accounts for about 3 per cent, all these 7 variables account for nearly 80 per cent when computing the CI index.

While it is difficult to compare KLR with FRT since the former uses the predictors one at a time sans their interactions, the economic explanation implied in the noise-to-signal ratio ranking variables depicts a picture which is in part similar to that with FRT. Indeed, the KLR results suggest that illiquidity factors are the most informative risk signals followed by insolvency risk proxies, and that contagion plays a key role along with inflation, thus indicating that systemic and macroeconomic risk factors also matter.

## Regression Tree

Figure A1 reports the tree structure obtained using the regression tree approach outlined in Section 2.1. As in FRT, we realize a risk stratification using multiple risk signals while providing probability estimates of a debt crisis conditional on predictors and terminal nodes. Specifically, the regression tree selected 7 out of 27 variables: (1) default history; (2) U.S. Treasury Bill rates; (3) reserve growth; (4) contagion; (5) exchange rate overvaluation; (6) short-term debt to reserves; (7) current account over GDP. Default history is the variable that stays on the top of the tree and first splits between no-serial defaulters (which have very low probability of default, 0.22 per cent) and countries with bad default history with at least 1 past default. For serial defaulters, the risk stratification is quite complex and shows two main

[3]KLR requires that the variables should be quantitative.

regimes based on U.S. interest rates. The first regime is with low rates. Here, illiquidity problems (short-term debt over GDP approximately between 131 and 139 per cent) together with negative current account over GDP leads to maximum risk (probability of default is 100 per cent); however, without having strong deficits in current account, illiquidity problems do not matter at all, since the probability of default is set at the null. Within the same low interest rates regime, other two economic situations show high risk, namely when reserve growth is negative (69.23 per cent) and when exchange rate devaluations are strong (58.33 per cent). The second regime is with high U.S. interest rates and shows some seemingly nonsensical results. Here, we observe that whenever we have less than 9 other countries in defaults, the probability is 84.85 per cent, but having systemic risk situations, namely when contagion is greater than 9, the risk partition is quite anomalous due to the masking variable effect. According to this effect, regional contagion is, indeed, blind and nonlinearly related with overall contagion. This is the reason why we have cases with no risk (probability of default set at the null) and maximum risk (probability of 100 per cent). In more depth, regression tree perfectly splits cases for contagion with value equal to 10 (node 9), 11 (node 10), and 12 (node 11), in which the maximum risk was associated with high regional contagion (on average, the corresponding value was 3.5) and no risk with low regional contagion (on average, the values for regional contagion were 1.9 for node 9 and 1.7 for node 11). Unlike with FRT, the economic explanation of debt crises implied in regression tree emphasizes the role played by U.S. interest rates, and conditional on specific interest rate regimes: (i) illiquidity, insolvency, and macroeconomic risks, when interest rates are low; (ii) systemic risk, when interest rates are high. Compared to FRT, the risk stratification of regression tree appears more complex and sometimes potentially erratic with large shifts in probability estimates due to minor changes in the splitting rules.

TABLE A3

*Logit and Stepwise Logit Estimates: 1975–2010*

| Variable | Logit | Stepwise logit | Austin-Tu Ranking |
|---|---|---|---|
| Intercept | -7.2180 (0.000) | -7.2862 (0.000) | |
| History | 1.1020 (0.000) | 1.1123 (0.000) | 1) History (1.000) |
| CAY | -0.0870 (0.001) | -0.0872 (0.000) | 2) CAY (1.000) |
| IMF | -16.0500 (0.975) | -16.1161 (0.975) | 3) IMF (1.000) |
| UST | 0.2759 (0.000) | 0.3191 (0.000) | 4) UST (1.000) |
| SEDY | 0.0766 (0.001) | 0.0680 (0.000) | 5) SEDY (1.000) |
| DSER | 0.2063 (0.009) | 0.2090 (0.000) | 6) DSER (0.999) |
| RGRWT | -0.0531 (0.016) | -0.0528 (0.008) | 7) RGRWT (0.998) |
| TEDY | -0.0167 (0.025) | -0.0106 (0.014) | 8) TEDY (0.998) |
| MAC | 0.8171 (0.015) | 0.7586 (0.015) | 9) MAC (0.995) |
| WX | -0.0063 (0.087) | -0.0061 (0.093) | 10) WX (0.976) |
| PR | 0.1267 (0.071) | 0.1291 (0.048) | 11) PR (0.932) |
| FDIY | -0.0806 (0.163) | -0.0912 (0.085) | 12) FDIY (0.927) |
| XG | -0.0097 (0.381) | -0.0147 (0.157) | 13) XG (0.592) |
| OPEN | -0.0026 (0.613) | - | 14) OPEN (0.306) |
| ResG | -0.0021 (0.456) | - | 15) ResG (0.211) |
| Cont_tot | 0.0407 (0.427) | - | 16) Cont_tot (0.167) |
| PEDY | 0.0052 (0.290) | - | 17) PEDY (0.156) |
| INF | 4.224E-05 (0.642) | - | 18) INF (0.114) |
| TEDX | 0.0003 (0.635) | - | 19) TEDX (0.076) |
| Cont_area | 0.0351 (0.738) | - | 20) Cont_area (0.063) |
| EU | -0.3713 (0.657) | - | 21) EU (0.040) |
| OVER | 0.0002 (0.533) | - | 22) OVER (0.025) |
| MG | -0.0006 (0.952) | - | 23) MG (0.002) |
| STDR | -7.31E-06 (0.989) | - | 24) STDR (0.002) |
| DOil | 0.2979 (0.541) | - | 25) DOil (0.001) |
| M2R | 0.0002 (0.374) | - | 26) M2R (0.001) |
| FDIG | -1.63E-05 (0.758) | - | 27) FDIG (0.001) |

*Notes:* The table reports logit and stepwise logit estimates with *p*-values in parentheses. In the last column we report the Austin and Tu (2004) bootstrap method to assess the variable importance. Specifically, we randomly selected 3,000 subsamples each one constituted by 90 per cent of the total observations, running the stepwise logit on each bootstrap sample including all 27 candidate variables. The predictors are ordered according to the frequency (reported in parenthesis) with which the variable is chosen over the 3,000 regressions.

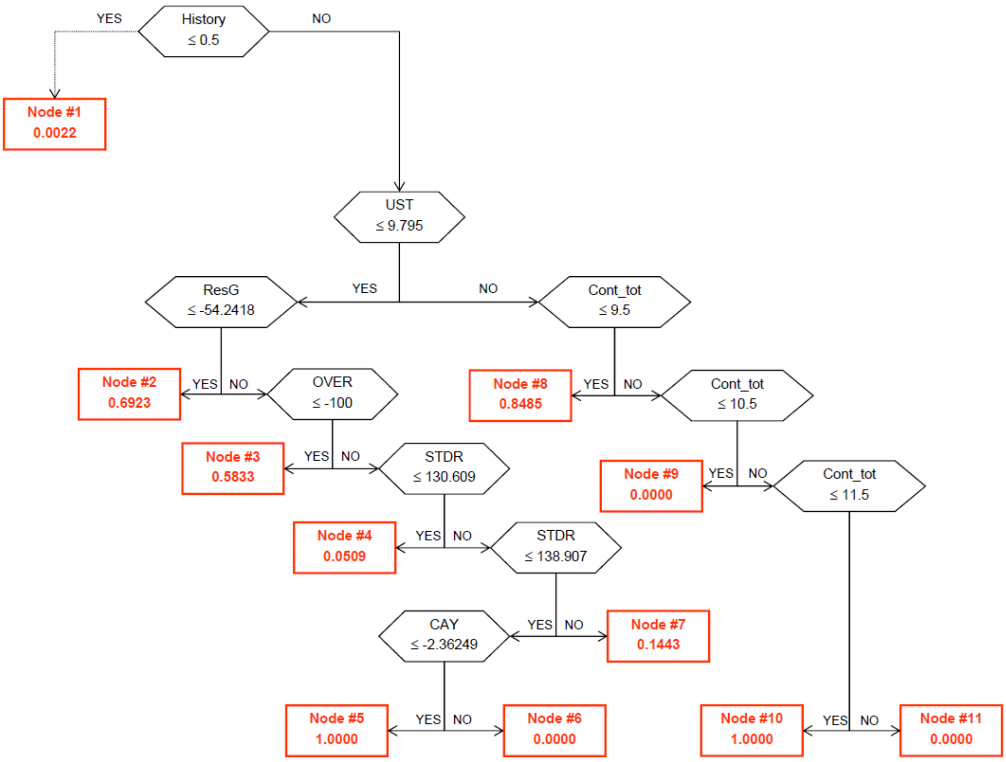
TABLE A4

*KLR Ranking: 1975–2010*

| Variable | Sens | Spec | w | 1/w | Relative weights |
|---|---|---|---|---|---|
| DSER | 0.0246 | 0.9995 | 0.0218 | 45.8852 | 0.2652 |
| STDR | 0.0246 | 0.9989 | 0.0436 | 22.9426 | 0.1326 |
| SEDY | 0.0082 | 0.9995 | 0.0654 | 15.2951 | 0.0884 |
| INF | 0.0082 | 0.9995 | 0.0654 | 15.2951 | 0.0884 |
| M2R | 0.0082 | 0.9995 | 0.0654 | 15.2951 | 0.0884 |
| FDIY | 0.0082 | 0.9995 | 0.0654 | 15.2951 | 0.0884 |
| Cont_tot | 0.0984 | 0.9834 | 0.1689 | 5.9207 | 0.0342 |
| MG | 0.0082 | 0.9984 | 0.1961 | 5.0984 | 0.0295 |
| Cont_area | 0.0492 | 0.9898 | 0.2070 | 4.8300 | 0.0279 |
| UST | 0.2623 | 0.9373 | 0.2390 | 4.1833 | 0.0242 |
| OVER | 0.0164 | 0.9952 | 0.2942 | 3.3989 | 0.0196 |
| ResG | 0.0164 | 0.9946 | 0.3269 | 3.0590 | 0.0177 |
| History | 0.0574 | 0.9780 | 0.3829 | 2.6114 | 0.0151 |
| PEDY | 0.0984 | 0.9502 | 0.5067 | 1.9736 | 0.0114 |
| TEDY | 0.0738 | 0.9598 | 0.5448 | 1.8354 | 0.0106 |
| TEDX | 0.5656 | 0.6688 | 0.5856 | 1.7077 | 0.0099 |
| CAY | 0.1639 | 0.8885 | 0.6800 | 1.4707 | 0.0085 |
| FDIG | 0.3033 | 0.7706 | 0.7563 | 1.3222 | 0.0076 |
| OPEN | 0.0246 | 0.9807 | 0.7846 | 1.2746 | 0.0074 |
| PR | 0.7459 | 0.3762 | 0.8363 | 1.1957 | 0.0069 |
| XG | 0.0082 | 0.9925 | 0.9153 | 1.0925 | 0.0063 |
| RGRWT | 0.8934 | 0.1324 | 0.9711 | 1.0297 | 0.0060 |
| WX | 0.8852 | 0.1168 | 0.9977 | 1.0023 | 0.0058 |

*Notes:* The table reports the results from the KLR procedure. Sens and Spec are the sensitivity (1 *minus* type I error) and the specificity (1 *minus* type II error) for each predictor obtained by minimizing the NSR (column w). We report also the inverse of the optimal NSR (column 1/w) which is the weight to be used in calculating the CI index. The last column is the weight of each predictor in computing the CI index expressed in relative terms, i.e., $w_r^{-1} / \sum_r w_r^{-1}$.

Figure A1: RT




*Notes:* The figure depicts the RT estimated over the period 1975–2010. For each split we specify the variable and the corresponding threshold also indicating the paths towards the terminal nodes. The values reported within each terminal node are the estimated probabilities of default.

## A.II.2. Out-Of-Sample PD and 2P Loss Function

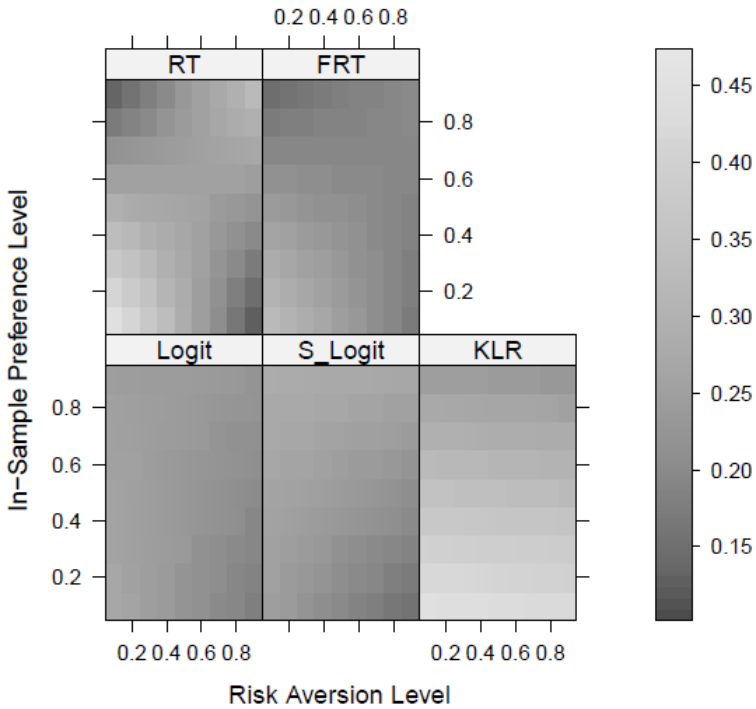
TABLE A5

*Out-Of-Sample PD*

| Crisis Episodes | FRT | Logit | S_Logit | RT | KLR |
|---|---|---|---|---|---|
| Algeria 1991 | 0.070 | 0.014 | 0.022 | 0.144 | 0.000 |
| Ethiopia 1991 | 0.266 | 0.121 | 0.200 | 0.167 | 0.000 |
| Hungary 1991 | 0.070 | 0.685 | 0.581 | 1.000 | 0.000 |
| Zimbabwe 1992 | 0.427 | 0.340 | 0.393 | 0.063 | 0.100 |
| South Africa 1993 | 0.058 | 0.604 | 0.710 | 0.060 | 0.100 |
| Kenya 1994 | 0.269 | 0.002 | 0.002 | 0.057 | 0.250 |
| Lithuania 1994 | 0.303 | 0.012 | 0.005 | 0.057 | 0.250 |
| Philippines 1994 | 0.064 | 0.332 | 0.375 | 0.057 | 0.250 |
| Turkey 1994 | 0.269 | 0.490 | 0.519 | 0.057 | 0.250 |
| Mexico 1995 | 0.206 | 0.288 | 0.291 | 0.066 | 0.050 |
| Venezuela 1995 | 0.059 | 0.042 | 0.028 | 0.066 | 0.050 |
| Jordan 1996 | 0.056 | 0.131 | 0.103 | 0.076 | 0.000 |
| Kazakhstan 1996 | 0.248 | 0.015 | 0.013 | 0.076 | 0.000 |
| Moldova 1996 | 0.056 | 0.040 | 0.044 | 0.076 | 0.000 |
| Indonesia 1997 | 0.196 | 0.008 | 0.007 | 0.076 | 0.188 |
| Korea, Rep. 1997 | 0.059 | 0.036 | 0.016 | 0.076 | 0.188 |
| Sierra Leone 1997 | 0.059 | 0.707 | 0.511 | 0.076 | 0.188 |
| Sri Lanka 1997 | 0.059 | 0.010 | 0.012 | 0.076 | 0.188 |
| Thailand 1997 | 0.059 | 0.532 | 0.341 | 0.076 | 0.188 |
| Argentina 1998 | 0.256 | 0.042 | 0.047 | 0.082 | 0.059 |
| Brazil 1998 | 0.056 | 0.081 | 0.096 | 0.082 | 0.059 |
| Moldova 1998 | 0.056 | 0.203 | 0.176 | 0.082 | 0.059 |
| Pakistan 1998 | 0.256 | 0.178 | 0.163 | 0.082 | 0.059 |
| Philippines 1998 | 0.256 | 0.850 | 0.847 | 0.082 | 0.059 |
| Ukraine 1998 | 0.286 | 0.046 | 0.041 | 0.082 | 0.059 |
| Ecuador 1999 | 0.164 | 0.096 | 0.086 | 0.088 | 0.130 |
| Gabon 1999 | 0.427 | 0.758 | 0.778 | 0.778 | 0.130 |
| Mexico 1999 | 0.053 | 0.602 | 0.628 | 0.088 | 0.130 |
| Turkey 1999 | 0.053 | 0.726 | 0.725 | 0.088 | 0.130 |
| Argentina 2000 | 0.060 | 0.064 | 0.067 | 0.092 | 0.063 |
| Uruguay 2000 | 0.060 | 0.775 | 0.766 | 0.092 | 0.063 |
| Zimbabwe 2000 | 0.231 | 0.416 | 0.421 | 0.092 | 0.063 |
| Brazil 2001 | 0.057 | 0.194 | 0.212 | 1.000 | 0.000 |
| Gabon 2002 | 0.590 | 0.960 | 0.914 | 1.000 | 0.286 |
| Indonesia 2002 | 0.054 | 0.043 | 0.046 | 0.086 | 0.286 |
| Moldova 2002 | 0.175 | 0.137 | 0.118 | 0.086 | 0.286 |
| Paraguay 2002 | 0.054 | 0.130 | 0.096 | 0.086 | 0.286 |
| Turkey 2002 | 0.249 | 0.787 | 0.732 | 0.086 | 0.286 |
| Uruguay 2002 | 0.249 | 0.891 | 0.847 | 0.086 | 0.286 |
| Cameroon 2004 | 0.053 | 0.007 | 0.009 | 0.065 | 0.000 |
| Venezuela 2005 | 0.054 | 0.007 | 0.006 | 0.084 | 0.042 |
| Ecuador 2008 | 0.059 | 0.137 | 0.154 | 0.035 | 0.167 |
| Hungary 2008 | 0.059 | 0.000 | 0.000 | 0.106 | 0.167 |
| Latvia 2008 | 0.213 | 0.112 | 0.158 | 0.035 | 0.167 |
| Pakistan 2008 | 0.059 | 0.155 | 0.125 | 0.035 | 0.167 |
| Ukraine 2008 | 0.059 | 0.009 | 0.011 | 0.106 | 0.167 |
| Greece 2010 | 0.199 | 0.000 | 0.000 | 0.051 | 0.200 |
| Ireland 2010 | 0.199 | 0.000 | 0.000 | 0.051 | 0.200 |
| Jamaica 2010 | 0.199 | 0.212 | 0.173 | 0.051 | 0.200 |

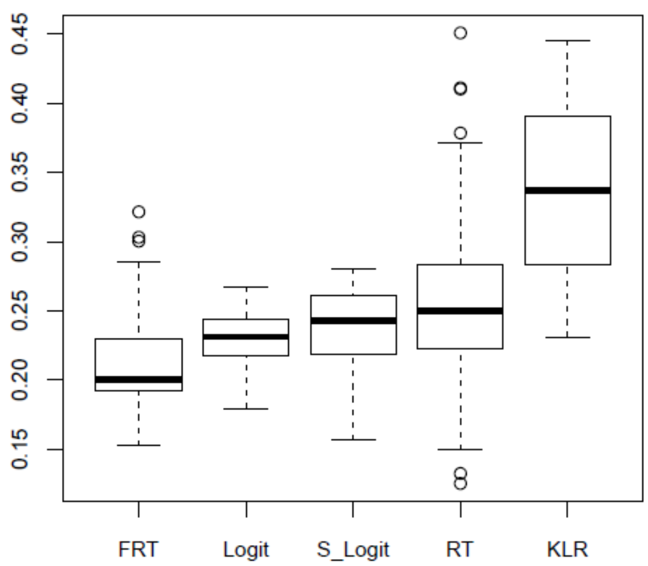
*Notes:* The table reports for each model, the estimated probability of default computed for all the 49 crises occurring over 1991–2010. Values in bold font and gray contour denote correctly predicted crises.

Figure A2: $2^D$ Loss Function




*Notes:* In this figure we graphically report the $2^D LF$ bivariate distribution for each model. The loss values are plotted over the risk aversion level ($x$ axis) and in-sample preference level ($y$ axis) space. The color scale is reported on the right.

Figure A3: Box Plots




*Notes:* The figure shows the box plots for the models using the $2^D LF$ values, depicting: (1) the sample minimum; (2) the lower quartile (Q1); (3) the median (Q2) which is the bold line within each box; (4) the upper quartile (Q3), and (5) the sample maximum.

## Appendix B

## CRAGGING and FRT

To better illustrate how CRAGGING works in practice, we now provide an illustration of the methodology supposing a panel data with $j = 1, \ldots, 10$ units (Algeria, Brazil, Egypt, Greece, Jamaica, Korea, Turkey, Uruguay, Venezuela, Zimbabwe) over 1975–2010 as follows:

| Algeria | | | | Brazil | | | | Egypt | | | | Greece | | | | Jamaica | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ |
| 1975 | 12.1 | | 0.12 | 1975 | 9.8 | | 0.12 | 1975 | 13.7 | | 0.11 | 1975 | 12.5 | | 0.02 | 1975 | 11.1 | | 0.13 |
| --- | --- | | --- | --- | --- | | --- | --- | --- | | --- | --- | --- | | --- | --- | --- | | --- |
| 2010 | 9.7 | | 1.1 | 2010 | 9.64 | | 2.14 | 2010 | 13 | | 2.13 | 2010 | 9.83 | | 0.13 | 2010 | 13 | | 0.08 |

| Korea | | | | Turkey | | | | Uruguay | | | | Venezuela | | | | Zimbabwe | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ |
| 1975 | 12.5 | | 0.04 | 1975 | 10.7 | | 0.13 | 1975 | 6 | | 0.03 | 1975 | 12.1 | | 0.12 | 1975 | 9.82 | | 0.09 |
| --- | --- | | --- | --- | --- | | --- | --- | --- | | --- | --- | --- | | --- | --- | --- | | --- |
| 2010 | 12.6 | | 1.12 | 2010 | 11 | | 1.05 | 2010 | 11 | | 3.1 | 2010 | 9.7 | | 1.1 | 2010 | 11.3 | | 0.14 |

In the *first step*, the units are randomly divided in $V = 5$ subsets consisting of $J_v = J/V = 2$ units each.

| Greece | | | | Korea | | | |
|---|---|---|---|---|---|---|---|
| Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ |
| 1975 | 12.5 | | 0.02 | 1975 | 12.5 | | 0.04 |
| --- | --- | | --- | --- | --- | | --- |
| 2010 | 9.83 | | 0.13 | 2010 | 12.6 | | 1.12 |

| Brazil | | | | Turkey | | | |
|---|---|---|---|---|---|---|---|
| Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ |
| 1975 | 9.8 | | 0.12 | 1975 | 10.7 | | 0.13 |
| --- | --- | | --- | --- | --- | | --- |
| 2010 | 9.64 | | 2.14 | 2010 | 11 | | 1.05 |

| Zimbabwe | | | | Jamaica | | | |
|---|---|---|---|---|---|---|---|
| Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ |
| 1975 | 9.82 | | 0.09 | 1975 | 11.1 | | 0.13 |
| --- | --- | | --- | --- | --- | | --- |
| 2010 | 11.3 | | 0.14 | 2010 | 13 | | 0.08 |

| Algeria | | | | Egypt | | | |
|---|---|---|---|---|---|---|---|
| Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ |
| 1975 | 12.1 | | 0.12 | 1975 | 13.7 | | 0.11 |
| --- | --- | | --- | --- | --- | | --- |
| 2010 | 9.7 | | 1.1 | 2010 | 13 | | 2.13 |

| Venezuela | | | | Uruguay | | | |
|---|---|---|---|---|---|---|---|
| Time | $X_1$ | ... | $X_n$ | Time | $X_1$ | ... | $X_n$ |
| 1975 | 12.1 | | 0.12 | 1975 | 6 | | 0.03 |
| --- | --- | | --- | --- | --- | | --- |
| 2010 | 9.7 | | 1.1 | 2010 | 11 | | 3.1 |

We then select the first test set ($v = 1$: Greece-Korea) that is taken out of the observations used for estimation and reserved for testing. Hence, the corresponding training (estimation) set contains 8 units (Brazil, Turkey, Zimbabwe, Jamaica, Algeria, Egypt, Venezuela, Uruguay). The training set is now perturbed by removing one unit at a time from Brazil to Uruguay, and then estimating a regression tree[4] on the resulting training set consisting of 7 countries. Each regression tree is tested on the same test set (Greece-Korea), resulting in $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_8$ estimates that are then used to compute $\bar{y} = \frac{\sum_{i=1}^{8} \hat{y}_i}{8}$, which is a single predicted probability of crisis obtained for each country-year of the test set.

[4] The regression tree is estimated in correspondence to the $\alpha$ tuning parameter that modulates the trade-off between the complexity and interpretability of the results.


This perturbation is repeated for the remaining 4 test sets, thus obtaining $\bar{y}$ estimates for the entire panel data, which are used to compute the final regression tree (FRT). Indeed, we replace the original binary (0, 1) definition of a crisis by the probabilities generated in the first step and run the regression tree analysis as usual, obtaining the best predictors, thresholds, and interactions.

| Greece | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 12.5 | | 0.02 |
| ... | ... | | ... |
| 2010 | 9.83 | | 0.13 |

| Korea | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 12.5 | | 0.04 |
| ... | ... | | ... |
| 2010 | 12.6 | | 1.12 |

| Brazil | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 9.8 | | 0.12 |
| ... | ... | | ... |
| 2010 | 9.64 | | 2.14 |

| Turkey | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 10.7 | | 0.13 |
| ... | ... | | ... |
| 2010 | 11 | | 1.05 |

| Zimbabwe | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 9.82 | | 0.09 |
| ... | ... | | ... |
| 2010 | 11.3 | | 0.14 |

| Jamaica | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 11.1 | | 0.13 |
| ... | ... | | ... |
| 2010 | 13 | | 0.08 |

| Algeria | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 12.1 | | 0.12 |
| ... | ... | | ... |
| 2010 | 9.7 | | 1.1 |

| Egypt | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 13.7 | | 0.11 |
| ... | ... | | ... |
| 2010 | 13 | | 2.13 |

| Venezuela | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 12.1 | | 0.12 |
| ... | ... | | ... |
| 2010 | 9.7 | | 1.1 |

| Uruguay | | | |
|---|---|---|---|
| Time | $X_1$ | ... | $X_R$ |
| 1975 | 6 | | 0.03 |
| ... | ... | | ... |
| 2010 | 11 | | 3.1 |

| Time | Country | $\bar{y}$ |
|---|---|---|
| 1975 | Greece | 0.01 |
| ... | ... | ... |
| 2010 | Greece | 0.16 |
| 1975 | Korea | 0.01 |
| ... | ... | ... |
| 2010 | Korea | 0.04 |
| 1975 | Brasil | 0.00 |
| ... | ... | ... |
| 2010 | Brasil | 0.04 |
| 1975 | Turkey | 0.00 |
| ... | ... | ... |
| 2010 | Turkey | 0.07 |
| 1975 | Zimbabwe | 0.00 |
| ... | ... | ... |
| 2010 | Zimbabwe | 0.17 |
| 1975 | Jamaica | 0.00 |
| ... | ... | ... |
| 2010 | Jamaica | 0.09 |
| 1975 | Algeria | 0.00 |
| ... | ... | ... |
| 2010 | Algeria | 0.04 |
| 1975 | Egypt | 0.00 |
| ... | ... | ... |
| 2010 | Egypt | 0.04 |
| 1975 | Venezuela | 0.01 |
| ... | ... | ... |
| 2010 | Venezuela | 0.13 |
| 1975 | Uruguay | 0.01 |
| ... | ... | ... |
| 2010 | Uruguay | 0.04 |

**Final Regression Tree**