Ca' Foscari
University
of Venice

Department
of Economics

Working Paper

Roberto Savona and
Marika Vezzoli

Fitting and Forecasting Sovereign Defaults Using Multiple Risk Signals

ISSN: 1827/3580
No. 26/WP/2012

Working Papers
Department of Economics
Ca' Foscari University of Venice
No. 26/WP/2012
ISSN 1827-3580

Fitting and Forecasting Sovereign Defaults Using Multiple Risk Signals

Roberto Savona and Marika Vezzoli
University of Brescia

This draft: October 2012

Abstract: In this paper we face the fitting versus forecasting paradox with the objective of realizing an optimal Early Warning System to better describe and predict past and future sovereign defaults. We do this by proposing a new Regression Tree-based model that signals a potential crisis whenever preselected indicators exceed specific thresholds. Using data on 66 emerging markets over the period 1975-2002, our model provides an accurate description of past data, although not the best description relative to existing competing models (Logit, Stepwise logit, Noise-to-Signal Ratio and Regression Trees), and produces the best forecasts accommodating to different risk aversion targets. By modulating in- and out-of sample model accuracy, our methodology leads to unambiguous empirical results, since we find that illiquidity (short-term debt to reserves ratio), insolvency (reserve growth) and contagion risks act as the main determinants/predictors of past/future debt crises.

Keywords: Data mining; Evaluating forecasts; Model selection; Panel data; Probability forecasting.

JEL Codes: C14, C23, G01, H63.

Address for correspondence:
Roberto Savona
Department of Business Studies
University of Brescia
Contrada S. Chiara 50
25122 Brescia - Italy
Phone: (++39) 030 2988557
Fax: (++39) 030 295814
e-mail: savona@eco.unibs.it

This Working Paper is part of the preparatory works for the project FP7-SSH-2012-2 "SYRTO Systemic risk tomography: signals, measurement, transmission channels, and policy interventions" submitted at the European Commission, in which Prof. Roberto Savona (project coordinator) and Prof. Monica Billio (local coordinator) are together the scientific coordinators.

The Working Paper Series
is available only on line
(http://www.unive.it/nqcontent.cfm?a_id=86302)
For editorial correspondence, please contact:
wp.dse@unive.it

Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta San Giobbe
30121 Venice Italy
Fax: ++39 041 2349210

1 Introduction

Forecasting financial crises is a crucial professional and scientific issue with the end of providing warnings to be used in preventing impending abnormalities, and taking action to minimize the negative externalities that could propagate on a systemic level. For emerging markets, this topic is of central importance due to their increasing vulnerability to the financial turmoil exhibited in recent decades and to their strong interrelations with developed countries (which became substantial starting from the 90s). Academic and policy circles are focused on methods of monitoring that could help detect symptoms of economic weaknesses and on providing possible "preventative treatments" in due time, since ex-post actions have, as is obvious, higher economic and social costs.

From an empirical perspective, ample evidence proves that crisis prediction is usually inaccurate even for sophisticated models which fit past data quite well. Indeed, when comparing simple models to complex ones, what we observe is that, on the one hand, simple models provide better prediction than complex models even if they do not necessarily fit the past data well; on the other hand, sophisticated models fit past data quite well but do not predict accurately the future. This fact creates what Clements and Hendry (1998) defined as the 'forecasting versus policy dilemma' to indicate the separation between models used for forecasting and models used for policy-making. Presumably, the reason for this problem is because having a model that over-fits in-sample when past data could be noisy leads to the retention of variables that are spuriously significant, which produces severe deficiencies in forecasting. The noise could also affect the dependent variable when the definition of 'crisis event' is unclear or when, notwithstanding a clear and accepted definition of crisis, the event itself is misclassified due to a sort of noisy transmission of the informational set used to classify that event. Moreover, the problem is complicated by the fact that, as observed by Fuertes and Kalotychou (2006, 2007), scant attention has been paid to sovereign default prediction, since most studies compare alternative models on the basis of their in-sample fitting ability. In this paper we face the fitting versus fore-

casting paradox with the objective of realizing an optimal Early Warning System (EWS henceforth) that best fits past data and accurately predicts sovereign defaults. The contribution of this paper is both methodological and empirical. The methodology we propose is conceived with the end of realizing a novel EWS that signals a potential crisis whenever a group of indicators exceed specific thresholds. This approach belongs to the Regression Trees analysis, popularized in the statistical community by the seminal work of Breiman et al. (1984), and recently applied to crisis prediction in Manasse and Roubini (2009) and Kaminsky (2006). Unlike the traditional Regression Trees approach, we use a new algorithm introduced in Vezzoli and Stone (2007) and Vezzoli and Zuccolotto (2011) aiming to remove severe limitations when inspecting panel data. Specifically, the methodology tries to solve the problem of fitting versus forecasting paradox in a two-step procedure:

1. In the *first step*, we run a new Regression Trees-based algorithm conceived to deal with panel data obtaining multiple predictions.

2. In the *second step*, we fit a single tree using the average of the predictions obtained in the first step in place of the original dependent variable.

We show that this two-step procedure represents a possible reconciling solution to our problem, since we obtain a parsimonious model, with good predictions (accuracy), better interpretability and minimal instability. In the first step, the model is constructed in a forward-looking basis also allowing the forecasting averaging, which is particularly useful in improving accuracy and reducing the variance of forecasting errors as recently discussed in Fuertes and Kalotychou (2007) and Makridakis and Taleb (2009). 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 dependet variable itself. As recently discussed 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 paper comes 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 the period 1975–2002 for 66 emerging economies. The in-sample analysis is performed over the entire time horizon, while out-of-sample analysis is carried out one-step-ahead from 1990 to 2002. To define a sovereign default, we followed Manasse and Roubini (2009), classifying a country as in debt crisis using Standard & Poor's criterion, or if it receives a large (in excess of 100 percent) non-concessional IMF loan. We also handle the problem of missing data using the multiple imputation technique suggested in King et al. (2001) which has proven to be well suited for panel data. Our empirical findings offer new insights on different angles of the issue involving sovereign default and are summarized as follows.

Predicting Factors. Six out of twenty-six potential predictors appeared to be optimal to fit the data in-sample: (1) Short-Term Debt to Reserves; (2) Reserve growth; (3) Short-Term Debt to GDP; (4) Openness, namely the sum of exports and imports divided by GDP; (6) Current account to GDP, and (6) Contagion, namely the total number of countries that go into default in a given year. When estimating the model one-step-ahead to predict the data out-of-sample, the number of variables changed over time while Short-Term Debt to Reserves, Contagion and Reserve growth were selected in each 'step' across the holdout periods. We thus conclude that although debt crisis determinants could differ over time, the risk factors that play a central role in explaining and predicting the sovereign defaults are, in order: (1) illiquidity risk (Short-Term Debt to Reserves); (2) insolvency risk (Reserve growth), and (3) systemic risk (Contagion).

Risk Stratification. Our EWS stratifies the country risk within four main categories on the basis of multiple risk signals: (1) *Higher Risk*, in which Short-Term Debt to Reserves is high together with negative Reserve growth and relatively high Contagion; (2) *Medium-High Risk*, in which Short-Term Debt to Reserves is high and Reserve growth is strongly negative; (3) *Medium Risk*, in which Short-Term Debt to Reserves is high; (4) *Low Risk*, in which Short-Term Debt to Reserves is low. At a two-state default classification based on probability estimates, Short-Term Debt to Reserves is the most important variable

since it actually discriminates between non-default (low default probability) and default (high default probability).

Model Accuracy. Several metrics were run to assess the adequacy of our model to fit and predict the data relative to traditional and advanced competing EWSs, namely: (i) Logit; (ii) Stepwise logit (sLogit); (iii) Noise-to-Signal Ratio (NSR), introduced in Kaminsky, Lizondo and Reinhart (KLR henceforth) (1998); (iv) Regression Trees (RT), recently used in Manasse and Roubini (2009). The statistical tests indicate that our model describes the data quite well, though Logit and Stepwise logit accuracy are slightly, while not statistically, greater. Again, in the forecasting analysis we prove that the accuracy of our methodology is statistically greater than that of competing EWSs.

Model Selection. To compare alternative EWSs both considering 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. We show that our classifier strongly dominates competing EWSs also exhibiting stable accuracy. These findings together seem to suggest that our model may be considered as a possible solution to the fitting versus forecasting paradox, at least when exploring the issue of emerging market debt crises. This paper is organized as follows. Section 2 describes the predictors and the methodological issues. Section 3 introduces the methodology and Section 4 describes the data. Section 5 reports the results and Section 6 concludes.

2 Predictors and Methodological Issues

Most of the attention of the existing literature on the determinants of debt crises mainly focuses on solvency, liquidity, and, say, willingness to pay proxies. By and large, statistical evidence suggests that the probability of a debt crisis is positively associated with higher

levels of total (McFadden et al., 1985) and short-term debt (Detragiache and Spilimbergo, 2001), negatively associated with GDP growth (Sturzenegger, 2004; De Paoli et al., 2006) and the level of international reserves (Dooley, 2000; Greenspan, 1999). Moreover, defaults are also related to more volatile and persistent output fluctuations (Catão and Kapur, 2006; Catão and Sutton, 2002), less trade openness (Cavallo and Frankel, 2008), political conditions (Block, 2003; Manasse et al., 2003; Van Rijckeghem and Weder, 2004), a previous history of defaults (Reinhart et al., 2003), and contagion (Eichengreen et al., 1996; Glick and Rose, 1999; Van Rijckghem and Weder, 2001).

Such literature helps to identify a list of potential predictors accounting for the strong connection with the other types of financial crises (currency, banking, twin, and capital account crises) also suggesting fundamental risk sources of debt crises. Specifically:

1. **Insolvency Risk**, which includes *capital and current account variables*, such as international reserves, capital flows, short-term capital flows, foreign direct investment, real exchange rate, current account balance, trade openness, and *debt variables*, namely public foreign debt, total foreign debt, short-term foreign debt and foreign aid.

2. **Illiquidity Risk**, proxied by *liquidity variables* such as short-term debt to reserves, debt service relative reserves and/or exports, M2 to reserves.

3. **Macroeconomic Risk**, measured by *macroeconomic variables*: real GDP growth, inflation rate, exchange rate overvaluation, international interest rates.

4. **Political Risk**, measured by *institutional/structural factors* for which we could use international capital market openness, financial liberalization, degree of political instability and political rights and default history[1].

5. **Systemic Risk**, namely the *contagion variable*, which could be proxied by the

[1]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).

number/proportion of the other debt crises also focusing on the geographical localization of the countries[2]. 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.

On the question of the relationship between observable predictors and country risk, different methodologies have been explored based on the philosophical assumptions about the nature of default. A first 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 (Claessens and Pennacchi, 1996; Bhanot, 1998; Merrik, 2001; Duffie, et al. 2003; Pan and Singleton, 2006). A second 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; Keswani, 2005; Gray et al., 2006). 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 three sub-categories: (i) Logit/Probit models (McFadden et al., 1985; Frankel and Rose, 1996; Demirg*üç*-Kunt and Detragiache, 1998, 1999, 2000; Milesi-Ferretti and Razin, 1998; Oral et al. 1992; Berg and Pattillo, 1999; Kalotychou and Staikouras, 2005; Fuertes and Kalotychou, 2006); (ii) classification methods, namely cluster and discriminant analysis (Taffler and Abassi, 1984; Burkart and Coudert, 2002; Fuertes and Kalotychou, 2007); (iii) signal approach. Such a last category starts with Kaminsky at al. (1998), in which a crisis is signaled when preselected leading economic indicators exceed some thresholds to be estimated according to a minimization procedure of the false alarm-to-good signal ratio; that is, the ratio of

[2]The prevalent literature assumes indeed that contagion is regionally-based (Gerlach and Smets, 1995; Glick and Rose, 1999; Eichengreen and Rose, 1999; Kaminsky and Reinhart, 2000). Moreover, some (e.g., Masson, 1998) distinguish between spillover, in which one crisis spreads to other countries via financial and/or trade linkages, and contagion, where the domino effect arises despite the absence of any economic relationship.

false signals to good signals. This method is further implemented in a composite way by Goldstein et al. (2000), which is an elaborated version of Kaminsky (1998), who proposes a composite indicator approach as a weighted sum of individual indicators. Kaminsky (1998) uses the signal approach to inspect currency and banking crises and Berg et al. (2004) compare the Probit model introduced in Berg et al. (1999) with the KLR and Logit models proposed in Ades et al. (1998), Roy and Tudela (2000), and Garber et al. (2000), finding mixed results on in- and out-of sample predictability even though it seems that their Probit model and KLR perform better than alternative methods out-of-sample.

More recently, Manasse et al. (2003) develop an EWS by using a Regression Trees approach significantly outperforming the Logit model in predicting debt crisis. Extending such an approach, Manasse and Roubini (2009) propose a collection of rules of thumbs that help predict potential crisis in the spirit of KLR, while simultaneously using the pre-selected indicators. Except for these two papers, Regression Trees are employed only to explore currency crises as in Ghosh and Ghosh (2002) and Frankel and Wei (2004). Also in Kaminsky (2006), Regression Trees are used to explore differences among currency crises, pointing to the use of second-generation EWS such as Regression Trees or parametric multiple-regime models in order to capture the broad spectrum of crises, thus leaving the door open to other default types.

3 Methodology

Except for the signal approach, most of the models briefly discussed in the previous section rely on what Breiman (2001b) referred to as the data modelling culture, namely they assume that the data are generated by a given stochastic data model. The problem is that sovereign defaults have multiple sources of risk that do not conform to the underlying distributional assumption upon which these models rely. As a result, the conclusions may be misleading, relying essentially on the model's mechanism instead of sovereign default nature's mechanism. The signal approach belongs instead to the algo-

rithmic modelling culture, through which the data are democratically processed trying to speak about sovereign default, not about a priori theories (which, in the data modelling approach, are usually pre-specified). This is the philosophical perspective underlying the methodology proposed in this paper. We first give some preliminaries and basic notations on Regression Trees and then present the algorithm, so as to make clear the main differences between simple Regression Trees and our approach. Next, we introduce the other competing approaches used in the empirical analysis to assess the model's accuracy, namely: (i) Logit; (ii) Stepwise logit; (iii) KLR.

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

$$y_{jt} = f\left(\mathbf{x}_{jt\text{-}1}\right) + \varepsilon_{jt} \tag{3.1}$$

where $f\left(\mathbf{x}_{jt\text{-}1}\right)$ is an unknown functional form of predictors $\mathbf{X}$ measured in $t-1$ parameterized by θ and ε is the random term for which some distributional assumptions can or cannot be specified. The objective is to estimate $f\left(\mathbf{x}_{jt\text{-}1}\right)$ making assumption or not about the random term distribution.

3.1 Regression Trees

Regression Trees are nonparametric models that look for the best local prediction of a response variable[3] y. Consider the issue of growing a Regression Tree. The data consists of R inputs and a continuous response, Y, for each of N observations. The algorithm needs to decide on the splitting variables and split points, and also what topology (shape) the tree should have. The recursive partitioning partitions the input space $\mathcal{S}$, which is

[3]When the dependent variable is continuous a Regression Trees is grown. On the contrary, when the response variable is categorical, a Classification Tree is grown.

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:

$$\mathcal{S} \subseteq \bigcup_{k=1}^{K} \widetilde{T}_k. \tag{3.2}$$

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 $\theta = \{\widetilde{T}_k, g_k\}_1^K$. In the Regression Tree, the underlying response-predictor structure $f(\mathbf{X})$ is represented by the piecewise constant g_k's fitted over the input subspace:

$$f(\mathbf{X}) = \sum_{k=1}^{K} g_k I(\mathbf{X} \in \widetilde{T}_k) \tag{3.3}$$

The sum of squares $\sum(Y - f(\mathbf{X}))^2$ is used as criterion of minimization[4]. As a result, the approach gives a rating mapping which can be considered as optimal for both the number of classes and the corresponding probabilities of default. The maximum homogeneity within the regions is indeed obtained by minimizing an impurity index [5] thus delivering a rating system which is validated by construction, since the partitions are realized in terms of maximum predictability. Again, Regression Trees are conceived with the end of improving the out-of-sample predictability and to do this they are estimated through a rotational estimation procedure, the cross-validation, with which the sample is partitioned into subsets such that the analysis is initially performed on a single subset (the training sets), while the other subset(s) are retained for subsequent use in confirming and validating the initial analysis (the validation or testing sets).

These are essentially the main motivations of the paper by Manasse and Roubini (2009), who referred to the Classification Trees in predicting the sovereign debt crisis, since the dependent variable was assumed to be qualitative[6].

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

[5]Frequently measured by the Gini index, for Classification Trees, or by the sum of squared errors, for Regression Trees.

[6]Vezzoli (2007) shows that if Y is a dummy variable, as is the case for sovereign defaults, regression and classification trees converge, thus allowing the use of regression-type models also for dummy variables.

3.2 The Model

The main drawback of Regression Trees (RT henceforth) is the two basic assumptions on Y which is supposed 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, which is the typical data structure of sovereign defaults. Hence, RT models basically look like a technique that does not pay attention to the intrinsic structure of the data, wherein autocorrelations and other latent dependencies could play a major role.

Considering these limitations, Vezzoli and Stone (2007) and Vezzoli and Zuccolotto (2011) introduced a new algorithm to remove such problems proposing a generalization of the basic RT in order to obtain better and more robust results when dealing with structured data. This new procedure is called CRAGGING, namely CRoss-validation AGGregatING, which is designed to reconcile the accuracy, stability and interpretability of a prediction system. It is in fact well known in the statistical community that, although recent technical improvements have led to new methods achieving better predictions and controlling for instability, some problems still remain concerning the pervasiveness these methods have on the data structure as well as their inner complexity, which usually provides multiple outputs. Within this context, the CRAGGING represents a possible reconciling solution in the spirit of the Occam's razor principle, since it delivers, in a sense, a data compressor which is also a scientific explanation/formulation generator.

In a nutshell, the idea underlying this algorithm is to repeatedly rotate the subsets in which the analysis is initially performed (the training sets) to such an extent as to, first, generating multiple predictors and, second, combine them to obtain a univariate and stable tree. It is for this reason that the CRAGGING can be viewed as a generalization of the RT.

3.2.1 First Step: CRAGGING

Let $(Y, \mathbf{X})$ be a panel data with N observations. For simplicity, suppose 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 $J \cdot T_j = N$. Denote with $\mathcal{L} = \{1, 2, \cdots, J\}$ the set of units and with $\mathbf{x}_{jt-1} = (x_{1jt-1}, x_{2jt-1}, \cdots, x_{rjt-1}, \cdots, x_{Rjt-1})$ the vector of predictors of unit j observed at time $t-1$ where $j \in \mathcal{L}$. As the name CRAGGING suggests, using the V-fold cross-validation, $\mathcal{L}$ is randomly partitioned into V subsets[7] denoted by $\mathcal{L}_v$, with $v = 1, 2, \cdots, V$, each containing J_v units and N_v observations[8]. Denote with $\mathcal{L}_v^c$ the complementary set of $\mathcal{L}_v$ containing J_v^c units and N_v^c observations, and $\mathcal{L}_{v\setminus\ell}^c$ the set where the ℓ-th unit is removed by $\mathcal{L}_v^c$ ($\ell \in \mathcal{L}_v^c$ and $\mathcal{L}_{v\setminus\ell}^c \bigcup \ell = \mathcal{L}_v^c$).

The cost complexity parameter $\alpha \geq 0$, is the tuning parameter of the cross-validation. Hence, for a fixed α, for each $\mathcal{L}_v$ and for each $\ell \in \mathcal{L}_v^c$ let

$$\hat{f}_{\alpha, \mathcal{L}_{v\setminus\ell}^c}(\cdot) \tag{3.4}$$

be the prediction function of a single tree (base learner) trained on data $\{y_{jt}, \mathbf{x}_{jt-1}\}_{j \in \mathcal{L}_{v\setminus\ell}^c,\ t=1,2,\cdots,T_j}$ and pruned with cost-complexity parameter α. The corresponding prediction in the test set is

$$\hat{y}_{jt,\alpha\ell} = \hat{f}_{\alpha, \mathcal{L}_{v\setminus\ell}^c}(\mathbf{x}_{jt-1}) \quad \text{with} \quad j \in \mathcal{L}_v, \quad \text{and} \quad t = 1, 2, \cdots, T_j. \tag{3.5}$$

Therefore, at every step, one unit is deleted from the training set and a tree is grown on it. If this perturbation causes significant changes in the obtained J_v^c trees, the accuracy of the predictors is improved by running the following equation:

$$\hat{y}_{jt,\alpha} = \frac{1}{J_v^c} \sum_{\ell \in \mathcal{L}_v^c} \hat{f}_{\alpha, \mathcal{L}_{v\setminus\ell}^c}(\mathbf{x}_{jt-1}) \quad \text{with} \quad j \in \mathcal{L}_v \quad \text{and} \quad t = 1, 2, \cdots, T_j \tag{3.6}$$

[7] In the partition, it is necessary that $V < J$ for preserving the particular structure of the data.
[8] The dimension of each V subset is of as nearly equal size as possible.

which is the average[9] of the functions (3.5) 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 objective of the CRAGGING is to improve the variance reduction of predictions by reducing the correlation between the trees and this is achieved through a double cross-validation. The first, called leave-one-*unit*-out cross-validation, is used for perturbing the training set removing one unit per time. Furthermore, we have to note that such a type of cross-validation does not destroy the structure of the data, differently from the common cross-validation that partitions randomly the observations. Hence, the CRAGGING tries to solve the sampling of the observations in the case of dataset with time-varying predictors.

The second cross-validation used by the CRAGGING is the well-known *v-fold cross-validation* on the test sets with $v = 1, \ldots, V$. The purpose is to find the optimal tuning parameter, α^*, that minimizes the estimate of the prediction error on all the test sets. Formally,

$$\alpha^* = \arg\min_{\alpha} LF(y_{jt}, \hat{y}_{jt,\alpha}) \quad \text{with} \quad j \in \mathcal{L}, \quad t = 1, 2, \cdots, \sum_{j=1}^{J} T_j \tag{3.7}$$

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

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

$$\tilde{y}_{jt}^{\text{crag}} = M^{-1} \sum_{m=1}^{M} \hat{y}_{jt,\alpha^*} \quad \text{with} \quad j \in \mathcal{L}, \quad t = 1, 2, \cdots, \sum_{j=1}^{J} T_j. \tag{3.8}$$

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

3.2.2 Second Step: Final Model

In the second step, a single tree, we name as Final Model (FM henceforth), is fitted on $(\tilde{Y}^{\text{crag}}, \mathbf{X})$ with cost complexity parameter $\alpha^{**} = M^{-1} \sum_{m=1}^{M} \alpha^{*}$. Here, through the replacement of Y with CRAGGING predictions we do two things: (1) we mitigate the effects of noisy data on the estimation process that affect both the predictors and the dependent variable itself[10]; (2) we realize a final RT which encompasses the overall forecasting ability arising from multiple trees. In doing this we obtain a parsimonious model, with good predictions (accuracy), better interpretability and minimal instability. In other words, the second step of our procedure is conceived to deliver a single tree to better understand the complex CRAGGING predictions. This is in line with the idea of assigning the simplest representations to the most accurate models suggested by many authors (Catlett, 1991; Quinlan, 1993; Evans and Fisher, 1994; Fayyad et al., 1996).

3.3 Competitive Models

3.3.1 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 ε in (3.1). Then the functional approximation, assuming country and time homogeneity, has the following linear basis expansion

$$f(\mathbf{x}_{jt-1}) = \Pr(y_{jt} = 1 | \mathbf{x}_{jt-1}) \equiv p_{jt}(\mathbf{x}_{jt-1}) = \frac{1}{1 + \exp -(\iota + \mathbf{x}'_{jt-1}\boldsymbol{\beta})}. \tag{3.9}$$

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 $\mathbf{X}$.

The second related model we use in the comparative analysis is the backward Stepwise

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

logit. Starting with the full model in (3.9), the backward stepwise sequentially deletes the predictor that has the least explanatory power. In more detail, the stepwise method we use is based on the Akaike Information Criterion (AIC) dropping at each step one variable at time in order to minimize the AIC score.

3.3.2 KLR

The KLR approach extracts signals of impending crises on the basis of a threshold criterion with which the predictor space is partitioned into crisis and non-crisis regions according to an optimal cut-off point to be 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, then computes a weighted average of the 0–1 signals by individual predictors while excluding those having a Noise-to-Signal Ratio (NSR henceforth) greater than 1 and using the inverse of the optimal NSR as weight. Therefore, such a composite index (CI) gives more weight to better performing (smaller minimum NSRs) indicators. Formally, let $\omega_r = \frac{b}{1-a}$ be the NSR 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 NSR 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

$$CI_{jt}(\mathbf{x}_{jt-1}) = \sum_{r=1}^{R} \frac{1}{\omega^*_{r,\mathfrak{c}_r}} I_{\mathfrak{c}_r}(x_{r,jt-1}) \quad \text{with} \quad \omega^*_{r,\mathfrak{c}_r} < 1, \tag{3.10}$$

where

$$I_{\mathfrak{c}_r}(x_{r,jt-1}) = \begin{cases} 1 & if\ |x_{r,jt-1}| > \mathfrak{c}_r \\ 0 & if\ |x_{r,jt-1}| \leq \mathfrak{c}_r. \end{cases} \tag{3.11}$$

Once the CI has been obtained, the probabilities of observing a default-case[11], i.e., the

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

functional approximation in (3.1), are estimated as the number of times where CI exceeds a certain threshold $\mathfrak{C}$ and a crisis occurred, divided by the total number of observations in which $CI > \mathfrak{C}$. Formally,

$$f(\mathbf{x}_{t-1}) = \Pr(\mathbf{x}_{t-1}) = \frac{\sum_t I_{\mathfrak{C}}(CI|y_t = 1)}{\sum_t (I_{\mathfrak{C}}(CI))} \quad \text{with} \quad t = 1, 2, \cdots, T \tag{3.12}$$

where

$$I_{\mathfrak{C}}(CI) = \begin{cases} 1 & if\ CI > \mathfrak{C} \\ 0 & if\ CI \leq \mathfrak{C} \end{cases} \tag{3.13}$$

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 NSR of CI we referred to the Youden Index, a diagnostic test for accuracy widely used in clinical application involving the receiver operating characteristic curve which we will discuss in the next section. As it will be shown, the Youden Index (YI henceforth) 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, both considering default and non-default classifications. Hence, the objective is to find the optimal $\mathfrak{C}$ so as to maximize the YI. As opposed to the NSR, the 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 NSR could lead to extreme thresholds for which the default is hardly signalled while false signals tend to zero.

3.4 Model Accuracy

3.4.1 In-Sample Tests

We use several metrics to assess the models' accuracy to fit the data in-sample based on errors between Y and $\hat{Y}$. First, we use the Root Mean Square Error ($RMSE$) to assess

the standard model fitting quality,

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{J}\sum_{t=1}^{T_j}\left(\hat{y}_{jt} - y_{jt}\right)^2} \tag{3.14}$$

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

$$BIC = N \ln\left(\frac{RSS}{N}\right) + \delta \ln\left(N\right) \tag{3.15}$$

where RSS is the Residual Sum of Square errors $\sum_{j=1}^{J}\sum_{t=1}^{T_j}\left(\hat{y}_{jt} - y_{jt}\right)^2$and δ 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),

$$BS = \frac{1}{N}\sum_{j=1}^{J}\sum_{t=1}^{T_j} 2\left(\hat{y}_{jt} - y_{jt}\right)^2, \quad BS \in [0,\, 2], \tag{3.16}$$

and the related Logarithmic Probability Score (LPS) which penalizes large errors more than the BS,

$$LPS = -\frac{1}{N}\sum_{j=1}^{J}\sum_{t=1}^{T_j} y_{jt} \ln\left(\hat{y}_{jt}\right) + \left(1 - y_{jt}\right)\ln\left(1 - \hat{y}_{jt}\right), \quad LPS \in [0,\, \infty]. \tag{3.17}$$

Third, we rely on signal-based diagnostic tests providing 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 fraction of the default cases correctly classified over the total defaults (true positive), and specificity as 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 a ROC curve which is a function of $\mathfrak{C}$, namely $ROC(\mathfrak{C})$. The diagnostics based on the ROC we use in the paper are: (1) the AUC and pairwise test on AUC differences; (2) the YI; (3) the Loss Function.

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

$$AUC = \int_0^1 ROC(\mathfrak{C}) d\mathfrak{C}. \tag{3.18}$$

In our analysis we use the trapezoidal rule with which (3.18) 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, the 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, $\mathbf{L}$ a suitable contrast matrix (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[12], the statistic for a pair of classifiers is

$$\frac{(\mathbf{L}\hat{\mathbf{U}})^2}{(\mathbf{LSL}')} \sim \chi^2_{(\text{rank}(\mathbf{L}))} \tag{3.19}$$

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

The 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,

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

$$YI = \arg\max_{\mathfrak{C}}[(1-a)+(1-b)-1] \quad \text{with} \quad \mathfrak{C} \in [0,1]. \tag{3.20}$$

The YI is indeed the point on the ROC curve farthest from chance, i.e., 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 the AUC with $YI = 2 \cdot \triangle - 1$ and $\triangle = [(1-a)+(1-b)]/2$ the approximated AUC[13].

Using the best cut-off point $\mathfrak{C}^*_{YI}$ obtained from (3.20), we finally compute the Loss Function (LF) for each classifier as the weighted sum of the missed default and non-default probabilities with cost for type I and type II error ζ and $(1-\zeta)$, respectively:

$$LF = [\zeta \cdot a_{\mathfrak{C}^*_{YI}} + (1-\zeta) \cdot b_{\mathfrak{C}^*_{YI}}], \quad LF \in [0,1]\,. \tag{3.21}$$

The cost ζ 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[14]) thus having $\zeta > 0.5$. 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 LF for values of ζ arbitrarily ranging from 0.1 to 0.8 and ranked the models for each of these cost values.

3.4.2 Out-Of-Sample Tests

To evaluate forecasting accuracy we rely on one-step-ahead analysis, recalibrating the models in each rolling-time window. For each year of the out-of-sample period[15] t^{out}, we add one-t-ahead observations to the previous fit period t^{in} and we use the new fitting

[13] See Hilden and Glasziou (1996).

[14] 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 type I error.

[15] 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 test, respectively.

period for updating the model estimates; next, these new estimates are used to make prediction for the following year (see Algorithm 1). 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 firstly replicate the tests (3.14)-(3.21) only excluding the *BIC* for computational convenience[16].

Algorithm 1 One-Step-Ahead Analysis

Require: A panel data $(Y_{jt}, \mathbf{X}_{jt})$ with $j = 1, \cdots, J$ and $t = 1, \cdots, T_j$
Require: The time period t divided in $t^{\text{in}} = 1, \cdots, T_j^{\text{in}}$ and $t^{\text{out}} = T_j^{\text{in}} + 1, \cdots, T_j$

1: **for** $\tau = 0$ **to** $T_j - T_j^{\text{in}} - 1$ **do**
2: fit the model on $(Y_{jt^{\text{in}}}, \mathbf{X}_{jt^{\text{in}}})$ with $j = 1, \cdots, J$ and $t^{\text{in}} = 1, \cdots, T_j^{\text{in}} + \tau$
3: predict on $(Y_{jt^{\text{out}_\tau}}, \mathbf{X}_{jt^{\text{out}_\tau}})$ with $j = 1, \cdots, J$ and $t^{\text{out}_\tau} = T_j^{\text{in}} + \tau + 1$
4: **end for**

Ensure: predictions $Y_{jt^{\text{out}}}$ with $j = 1, \cdots, J$ and $t^{\text{out}} = T_j^{\text{in}} + 1, \cdots, T_j$

Furthermore, we include the Diebold and Mariano (1995) (*DM* henceforth) forecasting test to assess whether our model is significantly better than competing models controlling for non-normality of forecasting errors and serial correlation. 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[17] for the observation of unit j at time t^{out} in the holdout sample and letting $\overline{d} = \frac{\sum_{j=1}^{J} \sum_{t^{\text{out}}=T_j^{\text{in}}+1}^{T_j} d_{jt^{\text{out}}}}{N^{\text{out}}}$ where N^{out} is the number of observations in the out-of-sample test, the *DM* test is as follows

$$DM = \frac{\overline{d}}{\sqrt{\frac{\widehat{var}(\overline{d})}{N^{\text{out}}}}} \backsim N(0,1) \tag{3.22}$$

and $\widehat{var}(\overline{d})$ 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 *DM* test, since it is common practice to update forecasts on a annual basis, namely when new values for economic variables are added to past data in order to recalibrate the EWS predictions. Note that

[16] We exclude the *BIC* since the models are dynamically estimated over rolling time windows which means possibly different models with different numbers of parameters.

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

also in this case the potential error autocorrelation become an issue to deal with, as pointed out by Fuertes and Kalotychou (2007).

3.4.3 Two Dimensional Loss Function

The forecasting versus policy dilemma highly complicates a 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 was bad (good); in the other, we could have bad (good) in-sample together with good (bad) out-of-sample accuracy. In these circumstances, we are obliged to trade off between fitting and forecasting ability *together with* missed defaults and false alarms. To put the discussion into perspective consider that:

- On the one hand, decision makers can be more or less risk averse, namely more or less sentitive towards type I errors. This is, say, the first dimension of the problem.
- On the other, 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 (when out-of-sample errors will be the target variables). This is 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, say $2^{\mathrm{D}}LF$, 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 EWSs in relation to a decision-maker's objective function conceived in the spirit of the forecasting versus policy dilemma.

Denoting by ϱ and $(1-\varrho)$ the weights for in- and out-of-sample errors and referring to the notation in (3.21) our $2^{\mathrm{D}}LF$ becomes

$$2^{\mathrm{D}}LF = \zeta \cdot [\varrho \cdot (a_{\mathfrak{C}^{*\mathrm{in}}_{YI}})^{\mathrm{in}} + (1-\varrho) \cdot (a_{\mathfrak{C}^{*\mathrm{out}}_{YI}})^{\mathrm{out}}] + \tag{3.23}$$

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

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.

The $2^{\mathrm{D}}LF$ helps select the best model for given ζ and ϱ also allowing a global ranking based on dominance criteria by comparing the bivariate shapes depicted by each classifier.

4 Data

The data used in this paper include annual observations for 66 emerging economies over the period 1975–2002. This dataset derives from Manasse and Roubini (2009), who use information on 47 economies with market access for the period 1970 to 2002, though we extended the overall number of sovereigns to 66 and reduced the time interval to 1975–2002, since in the subperiod 1970–1974 the number of defaults was virtually zero. Data on predictors are collected by World Bank's Global Development Finance (GDF), IMF, Government Finance Statistics database (GFS) and Freedom House (2002). These are grouped along the five categories outlined before and include: (1) capital, current account and debt variables; (2) liquidity measures; (3) macroeconomic and (4) political risk factors[18] 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 occurred in the same year, distinguishing between total (the overall number of debt

[18]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").

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), and IMF lending (IMF). 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[19].

Debt crises are defined following Manasse et al. (2003), thus using both the S&P's definition and the 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 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—these relating to Stand By Arrangements (SBA) and Extend Fund Facilities (EFF). With the end of realizing an EWS to predict a default *entry* rather than a *continuing* default, we included all the default events for each t and for each country subject to the fact that the country in $t-1$ was not in default, otherwise we excluded the observation for the default indicator as well as those for predictors.

As a whole, we analysed 112 crisis episodes reported below in Table 1, of which 40 are used in the out-of-sample analysis over the sub-period 1990–2002. The data in the table seem to suggest that over time, crises exhibit a cyclical effect, where much of the defaults were in the period 1981–1986, and a new cycle commenced starting from 1996, i.e., starting from the Asian crisis of 1997. Columns 6–8 of Table 2 report the sample mean for non-crisis and crisis states and their t/z-statistic[20] in order to get preliminary results on the discriminatory power of each predictor. 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

[19] While it is common procedure in the literature to use a proxy for contagion which is contemporaneous to the default indicator, to be more realistic and to provide a pure forecasting model we should use an expectation of contagion for period t observed in $t-1$. To this end, using a probability estimate of contagion should be effective. To obtain such a measure, contagion should be explored, first, as a dependent variable, and second, as a potential predictor. But this is our future research.

[20] z-test was performed for dummy variables, namely for MAC, DOIL and IMF.

multicollinearity problems. Based on these values, we note that TEDY and PEDY suffer from multicollinearity, thus reflecting a potential misinterpretation about the impact such variables exert on dependent variable while controlling for the others.

For 20 out of 26 variables the mean differential is statistically significant thus providing some 'univariate' ability in signalling sovereign defaults. (1) Market Access, (2) Oil Producing, (3) Exports, (4) FDI inflows variations, (5) Exchange rate overvaluation (OVER), and (6) Political Rights, have low power in discriminating between non-crisis and crisis. Interestingly, all five categories representing the fundamental risk sources of sovereign countries exhibit statistical significance, as shown by at least one variable pertaining to each category.

4.1 Missing Values

The original dataset had a number of missing values for many of the variables used as predictors. To control for such a problem, instead of using the common practice of listwise deletion or complete observations elimination which usually lead to significant biases in empirical analyses, we resorted to the multiple imputation technique proposed in King et al. (2001). This is a Bayesian algorithm which involves imputing a number of values for each one missing, ranging from 5 to 10, then averaging them to obtain the point estimates to fill in the missing cell.

Computationally, we carried out a multiple imputation technique using Autoregressive Distributed Lag (ADL) $(1, 1)$ model, which allows the controlling for time series cross sectional, also imposing empirical beliefs so as to shrink the posterior of the point estimate of each missing cell and for each country to within the specific historical range.

Summary statistics of missing values are in Table 2. Columns 3–5 report for each predictor the missing values' proportions, the mean of the observed values μ, and the mean of the missing value estimates $\hat{\mu}$. Note that the percentage of missing values never exceeds 0.3 over total of 1402 observations and that the means of missing estimates are quite sim-

ilar to those of the observed values for most predictors, except for FDIG, TEDX and INF for which unreported standard deviations were extremely high thus reflecting significant differences between the mean of the observed data and the mean of the estimated missing values.

5 Empirical Results

5.1 Fitting Model Accuracy

5.1.1 Predictors and Risk Stratification

We run our procedure, as outlined in Section 3.2, over the entire period 1975–2002, obtaining the risk stratification reported in Figure 1. This is a single tree (the FM) run over the CRAGGING predictions obtained along the line described in the methodological section. The panel data containing 66 countries was first randomly divided in 11 sets each one containing 6 countries, i.e. the 10% of the overall number of units in the panel data. Hence, the training set contains 60 countries and the test set contains 6 countries. Since the CRAGGING repeatedly perturbs the training sets removing 6 units per time, in correspondence of the optimal cost-complexity parameter α^* (3.7) we obtain 60 probability estimates for each observation. As discussed before, such a procedure was run M times (3.8) so as to minimize the generalization error. Computationally, we run $M = 50$ times with corresponding $50 \times 60 = 3000$ probability estimates, which allowed us to obtain a generalization error with a limiting value with no overfitting problems.

By inspecting Figure 1, we note that using the 26 potential predictors discussed in Section 4, only 6 variables are selected by the FM. These are: (1) Short-Term Debt to Reserves; (2) Reserve growth; (3) Short-Term Debt to GDP; (4) Openness, namely the sum of exports and imports divided by GDP; (5) Current account to GDP, and (6) Contagion, namely the total number of countries that go into default in a given year. Hence, the economic process underlying a sovereign debt crisis can be explained using

a parsimonious number of suitable proxies for illiquidity, insolvency, and systemic risk. From a statistical viewpoint, having only 6 out 26 variables, which reflects the trade-off between complexity and accuracy implied by the FM, is particularly useful for realizing a model that is as simple as possible while providing a reasonable explanation of past data. Indeed, if, on the one hand, by increasing the complexity of a model we provided a better fit to the data, on the other hand, having too many parameters would reflect a large sensitivity to small changes which in turns implies that the model will not distinguish between the generative 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. What we obtain is thus a country risk stratification using multiple risk signals also providing probability estimates of debt crises conditional on predictors and terminal nodes.

Short-Term Debt To Reserves appears as the most significant variable in predicting a crisis. The value of the corresponding threshold is 1.81 which basically splits the overall sample into: (i) episodes with low illiquidity problems (smaller or equal than 1.81) for which the probability of default is low; (ii) episodes with high illiquidity problems (greater than 1.81) where the probability of default is high.

An in depth analysis of the tree structure gives some 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 four major categories:

- *Higher Risk*, in which Short-Term Debt to Reserves is high (greater than 1.81), Reserve growth is negative (below than -35.53 per cent) and Contagion is high (greater than 5 other contries that go into default in a given year). Furthermore, when we also observe a strong deterioration of Current Account (below than -4.7 per cent of GDP), the default probability is the highest and near 0.65[21];

[21] In the figure we highlight the relative path with bold face.

- *Medium-High Risk*, in which Short-Term Debt to Reserves is high (greater than 1.81 per cent) and Reserve growth is negative (below than -35.53). Note that when Reserve growth is strongly negative (-69.23 per cent) the risk tends to increase;

- *Medium Risk*: in which Short-Term Debt to Reserves is high (greater than 1.81 per cent) and Reserve growth can be also moderately negative (but however greater than -35.53). In this scenario the other two insolvency proxies play a key role in grading the risk of a country: Short-Term Debt to GDP and Openness. When short indebtness relative to the GDP is less than about 9 points, the risk is moderate but tends to increase with the short-term debt (greater than 8.92). Interestingly, whenever a country enters into such a scenario, having less trade openness leads to higher risk;

- *Low Risk*, in which Short-Term Debt to Reserves is low (less than 1.81 per cent).

The main conclusion we obtain here is that debt crises are mainly driven by liquidity concerns. As is obvious, this result reflects the debt crisis definition used in this paper, which takes into account the near-to-defaults avoided through IMF financial aids and for which countries with severe illiquidity other than structural (insolvency) problems are the most usual users. Another interesting point to note is that contagion effects play a key role in explaining debt crises, since they are associated with the higher probability of default. Specifically, when liquidity and insolvency concerns arise, contagion has a disruptive effect especially when the current account deficit become larger. Reducing the problem to a two-states default classification based on probability estimates, Short-Term Debt to Reserves appears as the most important variable since it actually discriminates between low risk (Short-Term Debt to Reserves less than 1.81 per cent) and high risk (Short-Term Debt to Reserves less than 1.81 per cent).

5.1.2 Alternative Models and Different Explanations

Logit and Stepwise Logit Coloumns 2–3 of Table 3 report Logit and Stepwise logit model estimates. To make more informative the results obtained through logistic regression we also run a variable selection process so as to better explain the economic message of the models. Specifically, we used the boostrap method introduced in Austin and Tu (2004), taking 3,000 randomly selected sub-samples each one constituted by 90 per cent of the total observations, running the Stepwise logit on each bootstrap sample including all the 26 candidate variables; then, the predictors are ordered according to their importance, where the variable chosen most frequently is ranked first and so on. The results are in column 4. Arbitrarily putting at 90 per cent the cut-off point to select the most important predictors, 10 variables appear as the most relevant. Insolvency risk proxies are the major factors, almost all exhibiting statistical significance. These are (with corresponding estimated sign): Total External Debt to GDP (+), Openness (−), Market Access (+), FDI inflows to GDP (−), IMF lending (−) which coefficient does not reach statistical significance, Short-Term Debt to GDP (+), and Total Debt to Exports (−) which is in contrast with the expected sign and difficult to explain from an economic perspective. Furthermore, one proxy for macroeconomic and one for illiquidity risk factors are also important, namely: Real GDP growth (−) and M2 to Reserves (+) both bearing the expected sign of the slope.

The major difference with the FM is that logistic regressions attribute to pure solvency concerns the main reason why sovereign countries go into default, while macroeconomic and illiquidity conditions only marginally impact on debt crises.

KLR In Table 4 we report the NSR for each predictor according to the KLR methodology, also showing 1 *minus* type I (sensitivity) and 1 *minus* type II (specificity) errors. As discussed in Section 3.3.2, the inverse of the optimal NSR is the weight to be used in calculating the *CI* index according to (3.10). Then, such a weight gives us the variable

importance of each predictor attributed by the KLR. The methodology excluded only the exchange rate overvaluation (OVER) while other non-dummy variables[22] all show $\omega^*_{r,\mathfrak{c}_r} < 1$, which is the constraint used for dropping noisy predictors. According to KLR, the risk signals implied in the Short-Term Debt to GDP are the most informative as documented by the relative weight that accounts for about 21 per cent with respect to the remaining predictors. M2 to Reserves, Total External Debt to Exports, and Inflation are, in that order, the other relevant variables, all showing similar weights, together with long-term Debt Service to Reserves, Reserve growth, Short-Term Debt to Reserves and Contagion (total and regional), which are the remaining most relevant predictors used by KLR. These 9 variables account for 82 per cent in computing the *CI* index.

While it is difficult to compare KLR with FM since the first uses each predictor one at time lacking their interactions, the economic explanation implied in the NSR ranking variables depict a picture which is in part similar to what we have seen with the FM. Indeed, the KLR results seem to suggest that insolvency (SEDY, TEDX, ResG) and illiquidity (M2R, DSER, STDR) factors are the most informative risk signals, and that contagion, both at a global and regional level, play a key role as well together with inflation, thus indicating that systemic and macroeconomic risk factors also really matter.

RT Figure 2 reports the tree structure obtained using the RT approach outlined in Section 3.1. As in the FM, we realize a risk stratification using multiple risk signals also providing probability estimates of a debt crisis conditional on predictors and terminal nodes. Specifically, the RT selected 9 out 26 variables: (1) Short-Term Debt to Reserves; (2) Reserve growth; (3) Short-Term Debt to GDP; (4) Openness; (5) Contagion; (6) External Public Debt to GDP; (7) FDI inflows to GDP; (8) long-term Debt Service to Reserves; (9) M2 to Reserves. The risk stratification implied in the RT is more complex to understand with respect to the FM not only for the number of predictors but also for the economic interpretation of the splitting rules. Note, in particular, that higher and lower

[22]Indeed, KLR requires that the variables should be quantitative.

risk nodes are identified using a similar path: having high STDR, SEDY and Openness, together with low FDIY and DSER leads to the higher probability of default (0.86), which in turns becomes the lower (0.00) when, *ceteris paribus*, DSER is high with low M2R. As is clear, this partition is not efficient since the distance between high and low risk nodes is not maximized as would be required for minimizing the prediction error using the splitting rules implied in the RT. Secondly, Reserve growth seems to discrimate among risk grades although when combined with other predictors the relationship with debt crisis changes, possibly due to high unconditional and conditional non-linearity relative to other variables. Indeed, we observe that whenever Reserve growth is negative together with Contagion and high Short-Term Debt to Reserves *or* Reserve growth is strongly negative with high Short-Term Debt to Reserves, the probability of default is in both cases extremely high (nodes 2 and 5); again, as previously noted high Reserve growth can lead alternatively to higher and lower risk nodes (nodes 8 and 9) when combined with other insolvency (SEDY, FDIY, OPEN) and liquidity (DSER, M2R) risk factors.

As in the FM the economic explanation of debt crises implied in the RT emphasizes the role played by illiquidity, insolvency and systemic risks, while the risk stratification appears more complex and sometimes potentially erratic with large shifts in probability estimates due to minor changes in the splitting rules.

5.1.3 In-Sample Model Ranking

Table 5 reports the battery of statistical tests used to assess how the five different models describe the data in-sample. *BIC* and *RMSE* rank the RT best, and the KLR the worst. This is because *BIC* penalizes heavily for the number of parameters and because the probability estimates of RT are not as scattered as other models thus reflecting a minor error dispersion. When using the scoring rules *BS* and *LPS*, again RT is ranked best and KLR worst, while between the two extremes we note that FM is slightly penalized when using *LPS* since it moves from the second (*BIC*, *RMSE* and *BS*) to the third rank, due

to larger errors relative to the Logit.

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 Table 5. Based on AUC values the best model is the Logit, while subsequent classifiers are, in order, (2) Stepwise logit; (3) FM; (4) RT; (5) KLR. AUC pairwise differences and corresponding p-value computed according to (3.19) are in Table 6. Based on these data, RT and KLR accuracy appears as significantly lesser than Logit, Stepwise logit and FM. KLR is the worst classifier exhibiting a value for AUC which is indeed statistically lesser than that shown by the RT. Again, our FM while ranked as third seems to be a good descriptor of past data and in line with competing Logit and Stepwise logit which AUC values are not statistically greater than that computed for the FM.

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 the maximized YI and reported in Table 5. In general, we note that Logit, Stepwise logit and FM obtain higher sensitivity than specificity by trading off between type I and II errors while maintaining good performance in classifying defaults and non-defaults. Conversely, KLR and RT appear as good non-default classifiers while the performance in classifying default episodes is extremely modest. 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 A of Table 7 reports the loss function values computed for each model using risk-aversion weights ranging from 0.1 to 0.8. For each value of this weight, panel B shows the best and the worst classifier based on min and max LF values. By assuming higher risk aversion, the Stepwise logit seems to be the best classifier although it is ranked as worst when the weight is set at 0.1, namely when assuming the perspective of a high yield investor. Again, the worst classifier is the KLR which shows higher LF values for weights greater

than 0.1.

5.2 Forecasting Model Accuracy

To compare the models on the basis of their ability to forecast out of the estimation sample we focused on the entire period 1990–2002 and the two sub-periods 1990–1995 and 1996–2002, thus exploring how the 'new cycle' of sovereign defaults starting in 1996 impacted on forecasting performance. How predictors change their importance over time and how reliable are the models in forecasting future sovereign defaults based on past data are the two key questions we address in this section.

5.2.1 Time-Varying Predicting Factors

The one-step-ahead analysis and the corresponding recalibration of each model for all the rolling windows provide a first interesting result on whether predictors change over time when they are asked to make forecasts on debt crisis. Table 8 lists the variables selected by FM, Stepwise logit, KLR, and RT according to their importance (from higher to lower). The Logit model is implicitly considered by the Stepwise logit which is essentially its nested version. However it is important to note that when considering the Logit compared to its stepwise specification we have to keep in mind that it includes all the covariates without making a variable selection, presumably reflecting on forecasting performance. In this sense, the pure Logit can capture only in part the time dynamics of predictors, being partially implied in changing coefficient estimates.

Computationally we proceed as follows: (i) for Stepwise logit, RT and FM we use the number of times that the variables were selected over the total number of the estimation samples (from 1990 to 2002 we have 13 samples); (ii) for KLR, first, we listed the variables according to their weight used in computing the *CI*, as in (3.10), second, we computed the number of times the variables were ranked within the highest percentile accounting

for 80 per cent[23] of the total weights, third, such a number was expressed relative to the total number of the estimation samples.

Let us look at the data in Table 8 and consider, firstly, the results of the FM. The number of predictors selected by the model changed over time, thus proving that variable selection is dependent upon the time period used in estimating the model. Short-Term Debt to Reserves, Contagion and Reserve growth were selected in each 'step' across the out-of-sample period, thus acting as the key proxies for the main sovereign risk sources. Indeed, although debt crisis determinants could differ over time and, possibly, across regions, liquidity, insolvency, and systemic risk appear as the major determinants of sovereign defaults. The same conclusion holds for RT, for which Short-Term Debt to Reserves, and Reserve growth were selected in each step while Contagion in 12 out of 13 estimations, namely in 92 per cent of the rolling windows.

The results of Stepwise logit are coherent with the Austin and Tu (2004) procedure run over the entire sample and previously discussed in Section 5.1.2. Indeed, pure solvency proxies (MAC, IMF, FDIY, TEDY, OPEN, SEDY) appear as more important relative to illiquidity (M2R) and macroeconomic (UST) proxies, all selected in more than 90 per cent of the 13 regressions.

Consider, now, the results of the KLR. As discussed in the methodological section, this methodology uses all predictors excluding those having $\omega^*_{r,\mathfrak{c}_r} > 1$. The corresponding variable selection is thus quite conservative and indeed in our analysis we found that only OVER has been excluded for each rolling window. This suggests that KLR captured the time dynamics of default predictors only marginally, thus showing a modest adaptability to changes occured in the economic environments. The importance attributed to the predictors slightly changed over time, although illiquidity (DSER, M2R), insolvency (SEDY, TEDY, TEDX) and macroeconomic (INF) risk proxies appeared constantly ranked as first.

[23]We chose 80 per cent since we used the same percentage in commenting on the results in Section 5.1.2.

5.2.2 Out-Of-Sample Model Ranking

Table 9 presents different metrics computed over the entire holdout sample as well as for the two sub-periods 1990–1995 and 1996–2002. Together with the same tests used to assess the models' reliability in-sample except for the *BIC*, we computed the *DM* test comparing the forecasting errors of the FM with alternative EWSs along the lines described in Section 3.4.2.

Inspecting *RMSE*, *BS* and *LPS* we note that FM outperforms competing models both considering the entire holdout sample and the two sub-periods. Similarly, when testing the forecasting ability of FM using the *DM* test the results indicate that our model significantly exceeds the Stepwise logit and RT, while for KLR the forecasting superiority is near to significance. When compared with the Logit model the statistics do not reach statistical significance, although the FM forecasting errors are on average lesser than those of the Logit. Again, by inspecting the two sub-periods, FM seems to be stronger in the years 1990–1996, showing *DM* statistics which appear statistically significant for all the models except for KLR.

AUC-based tests provide a clear view about FM reliability in making predictions. Indeed, in the overall holdout period the FM shows an area under the *ROC* curve near 0.71 against values ranging from 0.6028 (RT) to 0.6614 (Logit). In line with *RMSE*, *BS*, *LPS* and *DM* test, in the two sub-periods, FM seems significantly outperform in the years 1990–1995 while in the sub-period 1996–2002 it is again ranked as first but the performance of the Logit model is virtually the same. Sensitivity and specificity computed using the *YI* criterion also show that our approach predicts 70 per cent of the default episodes and about 67 per cent of the non-defaults occurring in the period 1990–2002, becoming less accurate from the first to the second sub-period. On the other hand, competing classifiers are less sensitive, ranging from 35 (RT) to 67.5 (Logit) per cent together with values for specificity which is in some cases higher than that shown by the FM, varying from 61.76 (Logit) to 89 (KLR) per cent.

Table 10 sets out AUC differences with corresponding p-values for the entire period as well as for the two sub-period 1990–1995 and 1996–2002. The data in the table confirm that over the period 1990–2002, FM significantly outperforms competing EWSs except for Logit model, in which the AUC of the FM while greater is not statistically higher than that of the Logit. In the sub-period 1990–1995 the FM is again ranked as first based on the higher AUC although significant superiority is achieved only against Stepwise and pure Logit models. Finally, over the sub-period 1996–2002, FM strongly outperforms RT while when comparing against other EWSs the difference between the AUC values is not so high as to conclude a clear superiority for the FM. This is particularly true for the Logit, of which the AUC value is virtually aligned to that of the FM, as mentioned before. Interestingly, only RT and FM show significant differences between their AUC values for the first and second sub-periods, missing their accuracy in the years 1996–2002. This conclusion does not hold for other models which instead exhibit quite constant values for the AUC. Based on these findings, it seems that RT and FM are more sensitive towards the supposed new default cycle occurring over the years 1996–2002.

The LF analysis extends these results providing some interesting insights on how accuracy perception changes with different decision maker's targets. Table 11 reports the value for the LF assuming the same range for the risk aversion level used in Section 5.1.3. When risk aversion is low ($\zeta < 0.3$) the RT is the best model while the higher cost is associated with Logit and Stepwise logit, when splitting the analysis between 1990–1995 and 1996–2002. However, from modest to high risk aversion ($0.3 \leq \zeta \leq 0.8$) the FM dominates the competing EWSs over the entire holdout period, although when focusing on the two subperiods, Stepwise logit (1990–1995) and especially Logit (1996–2002) are better than our FM. Interestingly, by changing the decision maker's perspective the performance exhibited by Logit, Stepwise logit and RT are significantly unstable: potentially moving from best to worst classifier and vice versa. Indeed, we observe that these models are alternatively ranked as worst performers depending on risk aversion level

and time period.

As a whole, the out-of-sample analysis indicates that the FM exhibits higher accuracy in making predictions than the competing models. On average, FM predicted 70 per cent of the total defaults occurred over the period 1990–2002 versus 67.5 and 60 per cent, for Logit and Stepwise logit, and 37.5 and 35 per cent, for KLR and RT. The higher accuracy of the FM model is also confirmed by *RMSE* and scoring-based tests (*BS* and *LPS*) as well as by the *DM* statistic, which provides statistical evidence about model superiority in making one-step ahead predictions. Assuming a subjective evaluation of the models' performance through the imputation of a cost to both missing defaults and non-defaults, again we prove that the FM is the best model for different risk aversion levels, particularly when missing defaults (type I errors) become the primary issue. On the other hand, RT seems to be the worst performer in particular when sensitivity towards type I errors tend to be higher. Hence, what we note is that, on the one hand, our FM appears to be quite as good a descriptor of past data, while it is not the finest descriptor, being dominated by logistic regression in-sample, in particular the Stepwise alternative. On the other hand, when testing the models out-of-sample, the FM significantly outperforms competing EWSs, and logistic regression becomes a poor forecaster, thus suffering from the so-called fitting versus forecasting paradox. 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 fitting and forecasting ability jointly. By reconciling the 'two-sides' of the 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 the decision makers could have relative to the issue of debt crisis. This is what we try to do in the next section.

5.3 Fitting Versus Forecasting Model Accuracy

As argued in Section 3.4.3, in order to provide a global evaluation of the models jointly considering in- and out-of-sample accuracy, we use the $2^{\mathrm{D}}LF$ simply computing a weighted average of LF in- and out-of-sample with weights reflecting the decision-maker's objective function (data generating process vs. forecasting activity).

Table 13 reports the best model (panel A) and the worst model (panel B) based upon the values for the $2^{\mathrm{D}}LF$ using (3.23) where $0.1 \leq \zeta \leq 0.8$ and $0.1 \leq \varrho \leq 0.8$ with step 0.1. To make the model comparison easier, we also report in Figure 3 the bivariate function generated by the $2^{\mathrm{D}}LF$ for each model. From modest to high risk aversion ($0.4 \leq \zeta \leq 0.8$), the FM appears to be the best model to use when exploring the data generating process and when making forecasts of future debt crises. Stepwise logit is ranked first only when decision maker's function is strongly focused on policy issues ($0.7 \leq \varrho \leq 0.8$), to the detriment of forecasting ability. From low to modest risk aversion ($0.1 \leq \zeta \leq 0.3$), the RT is the finest model for both fitting and forecasting sovereign defaults ($0.3 \leq \varrho \leq 0.8$) except for pure forecasting targets where KLR appears as the best performer ($0.1 \leq \varrho \leq 0.2$). On the other hand, excluding pure forecasting targets ($0.1 \leq \varrho \leq 0.2$) in which decision makers obtain the worst performance using Logit and RT model to forecast debt crises, Stepwise logistic regression and KLR show the highest cost for low and high risk aversion, respectively.

The bivariate distribution depicted by the $2^{\mathrm{D}}LF$ is thus particularly useful for comparing the models based on preferences expressed by the combinations of ζ and ϱ [24]. To put the issue into perspective, we ordered the values of the $2^{\mathrm{D}}LF$ based on the combinations of ζ and ϱ for each model, converting to 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 $E = 8 \cdot 8 = 64$), and H columns which are the number of models involved in the analysis (in our case $H = 5$).

[24] In our analysis ζ and ϱ range from 0.1 to 0.8 with step 0.1, thus having $8 \cdot 8 = 64$ different combinations of weights.

Each element of the matrix $\mathbf{Q}$ is denoted by $\mathfrak{r}_{eh}$ with $e = 1, \cdots, E$ and $h = 1, \cdots, H$ be the rank for the h-th model based on e-th combination of weights. Hence, for each ζ and ϱ 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, second, we 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,

$$\pi_h = \frac{E \cdot H - \mathfrak{R}_h}{E \cdot H - E}, \quad \pi_h \in [0, 1], \tag{5.24}$$

where $\mathfrak{R}_h = \sum_{e=1}^{E} \mathfrak{r}_{he}$ is the sum of the ranks for model h. Dividing the (5.24) by E we obtain

$$\pi_h = \frac{H - \overline{\mathfrak{R}}_h}{H - 1}, \quad \pi_h \in [0, 1], \tag{5.25}$$

where $\overline{\mathfrak{R}}_h$ is the mean rank for model h, and $1 \leq \overline{\mathfrak{R}}_h \leq H$. If a model were the best one for each combination of weights, $\overline{\mathfrak{R}}_h$ would be equal to 1. On the contrary, if a model were the worst for each combination of ζ and ϱ, $\overline{\mathfrak{R}}_h$ would be equal to H. In this way, whenever a model is rated as first the (5.25) will take the value 1, otherwise 0 will be the value whether the model is rated as worst.

In Table 14 we report the value for π_h with corresponding $\overline{\mathfrak{R}}_h$ together with paired Wilcoxon statistics. The FM is ranked first and paired comparison through Wilcoxon statistic shows strong significance against all competing models. The Logit is ranked second while RT exhibits a mean ranking which is not statistically different from it; thus the two models seems to perform quite similarly although the π_h index is higher for Logit than RT. Stepwise logit and KLR are, in order, the fourth and fifth models which clearly exhibit the worst performance relative to other classifiers. This results is also robust from a statistical viewpoint as implied by the p-values of the Wilcoxon statistics, which are all near zero except when comparing Stepwise logit and KLR each other, thus indicating

that the worst performance is almost aligned for the two models.

The main message coming from the $2^{\mathrm{D}}LF$ analysis is that the FM seems to be the best model for both fitting and predicting debt crises also exhibiting stable performance by changing possible decision makers' targets. See on this point Figure 4, in which we report the Box-plots using the values for the $2^{\mathrm{D}}LF$. As we note, the median cost for the FM is lesser than that shown by alternative models, also exhibiting low dispersion relative to competing models.

As a result, it seems that with the FM we may provide a possible reconciling solution to the fitting versus forecasting paradox. Indeed, through the trade-off between fitting 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 FM:

- provides an accurate description of past data, although not the best description;
- 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, say, averaging fitting and predicting ability together with low and high risk aversion. In this sense, the meaning we attribute to the term *best* has to be interpreted as *best average model*.

6 Conclusion

In this paper we considered the problem of fitting and predicting sovereign debt crisis in light of the forecasting versus policy dilemma exposited 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 RT based

EWS using a two-step procedure. In the first step, we generate multiple predictions by cross-validating the model on rotated sub-samples until the average of the estimates stabilizes. In the second step, we fit a RT using such an average as dependent variable. This two-step procedure entails a trade-off between fitting and forecasting ability, also imposing a penalization for the model complexity, 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 over the period 1975–2002, the several statistical metrics run to assess the model reliability in- and out-of-sample relative to the existing state-of-the-art models (Logit, Stepwise logit, NSR, RT) indicate 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, while not the best description among the alternative EWSs (Logit and Stepwise logit are the best classifiers). Our model thus leads to an unambiguous interpretation of the in-sample and out-of-sample results. And indeed, we find that illiquidity (short-term debt to reserves ratio), insolvency (reserve growth) and contagion risks act as the main determinants and predictors of past and future debt crises.

Acknowledgments

We thank Charles Stone (University of Berkeley) for unvaluable suggestions and stimulating discussions on the project, Gianni Amisano (ECB and University of Brescia), Pierluigi Balduzzi (Boston College), Maurizio Carpita (University of Brescia), Paolo Manasse (University of Bologna), and seminar participants at Bicocca University, University of Brescia, Financial Management Association (FMA) 2008 European Conference, Royal of Statistical Society (RSS) 2008 Conference, GRETA Conference 2008 (Venice), European Central Bank, SoFiE Conference 2011 (Chicago) for useful comments.

References

Ades, A., Masih, R. and Tenengauzer, D. (1998), *Gs-Watch: A New Framework for Predicting Financial Crisis in Emerging Markets*, Goldman Sachs.

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

Bamber, D. (1975), "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Carachteristic Graph," *Journal of Mathematical Psycology*, 12, pp. 387-415.

Berg, A., Borensztein, E., Milesi-Ferretti, G.M. and Pattillo, C. (1999), "Anticipating Balance of Payments Crises: The Role of Early Warning Systems," *IMF Occasional Paper*, 186.

Berg, A., Borensztein, E. and Pattillo, C. (2004), "Assessing Early Warning Systems: How Have They Worked in Practice?," *IMF Working Paper*, 04/52.

Berg, A. and Pattillo, C. (1999), "Predicting Currency Crises: The Indicators Approach and an Alternative," *Journal of International Money and Finance*, 18, pp. 561-586.

Bhanot, K. (1998), "Recovery and Implied Default in Brady Bonds," *Journal of Fixed Income*, 8, pp. 47-51.

Block, S. A. (2003), "Political Conditions and Currency Crises in Emerging Markets," *Emerging Markets Review*, 4(3), pp. 287-309.

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

Breiman, L. (2001b), "Statistical Modeling: the Two Cultures," *Statistical Science*, 16(3), pp. 199-231.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth Inc., California.

Burkart, O. and Coudert, V. (2002), "Leading Indicators of Currency Crises for Emerging Countries," *Emerging Markets Review*, 3, pp. 107-133.

Catão, L. and Kapur, S. (2006), "Volatility and the Debt-Intolerance Paradox," *IMF Staff Papers*, 53(2), pp. 195-218.

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

Catlett, J. (1991), "Megainduction: Machine Learning on Very Large Databases," *PhD thesis*, Basser Department of Computer Science, University of Sidney, Sydney, Australia.

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*, 27(8), pp. 1430-1452.

Claessens, S. and Pennacchi, G. (1996), "Estimating the Likelihood of Mexican Default from the Market Prices of Brady Bonds," *Journal of Financial and Quantitative Analysis*, 31, pp. 109-126.

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

De Paoli, B., Hoggarth, G. and Saporta, V. (2006), "Costs of Sovereign Default," *Bank of England Financial Stability Paper*, 1.

Debashis, P., Bair, E., Hastie, T. and Tibshirani, R. (2008), "'Preconditioning' for Feature Selection and Regression in High-Dimensional Problems," *Annals of Statistics*, 36(4), 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*, 44, pp. 837-845.

Demirgüç-Kunt, A. and Detragiache, E. (1998), "The Determinants of Banking Crises in Developing and Developed Countries," *IMF Staff Papers*, 45(l), pp. 81-109.

Demirgüç-Kunt, A. and Detragiache, E. (1999), "Financial Liberalization and Financial Fragility," in *Annual World Bank Conference on Development Economics 1998*, eds. B. Pleskovic and J. Stiglitz, World Bank.

Demirgüç-Kunt, A. and Detragiache, E. (2000), "Monitoring Banking Sector Fragility: A Multivariate Logit Approach," *The World Bank Review*, 14(2), pp. 287-307.

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

Diebold, F. and Mariano, R. (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13(3), pp. 253-63.

Dooley, M. (2000), "A Model of Crises in Emerging Markets," *The Economic Journal*, 110(460), pp. 256-272.

Duffie, D., Pedersen, L. and Singleton, K. (2003), "Modeling Sovereign Yield Spreads: A Case Study of Russian Debt," *Journal of Finance*, 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*, 98(4), pp. 463-84.

Evans, B., Fisher, D., (1994), "Process Delay Analysis Using Decision Tree Induction," *IEEE Expert*, 9, 1, pp. 60-66.

Fayyad, U. M., Djorgovski, S. G. and Weir, N. (1996), "From Digitized Images to Online Catalogs: Data Mining a Sky Survey," *AI Magazine*, 17, 2, pp. 51-66.

Frankel, J. and Rose, A. (1996), "Currency Crashes in Emerging Markets," *Journal of International Economics*, 41, pp. 351-366.

Frankel, J. and Wei S-J. (2004), "Managing Macroeconomic Crises: Policy Lessons," *NBER Working Paper*, 10907.

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

Fuertes, A. and Kalotychou, E. (2007), "Optimal Design of Early Warning Systems for Sovereign Debt Crises," *International Journal of Forecasting*, 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.

Garber, P. M., Lumsdaine, R. L. and van der Leij, M. (2000), *Deutsche Bank Alarm Clock: Forecasting Exchange Rate and Interest Rate Events in Emerging Markets*. Deutsche Bank.

Gerlach, S. and Smets, F. (1995), "Contagious Speculative Attacks," *European Journal of Political Economy*, 11, pp. 45-63.

Ghosh, S. and Ghosh, A. (2002), "Structural Vulnerabilities and Current Crises," *IMF Working Paper*, 02/9.

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

Glick, R. and Rose, A. (1999), "Contagion and Trade: Why Are Currency Crises Regional?," *Journal of International Money and Finance*, 18(4), pp. 603-17.

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

Gray, D., Merton, R., and Bodie, Z. (2006), "A New Framework for Analyzing and Managing Macrofinancial Risks of an Economy," *NBER Working Paper Series*, 12637.

Greenspan, A. (1999), "Currency Reserves and Debt," *Remarks Before the World Bank Conference on Recent Trends in Reserves Management*, Washington, D.C., April 29.

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

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

Kalotychou, E. and Staikouras, S. (2005), "The banking exposure to international lending: Regional differences or common fundamentals," *Financial Markets, Institutions and Instruments*, 14(4), pp. 187-214.

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

Kaminsky, G. L. (2006), "Currency Crises: Are They All the Same," *Journal of International Money and Finance*, 25, pp. 503-527.

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*, 51(1), pp. 145-168.

Keswani, A. (2005). "Estimating A Risky Term Structure Of Brady Bonds," *The Manchester School*, 73(1), pp. 99-127.

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*, 95(1), pp. 49-69.

Makridakis, S. and Taleb, N. (2009), "Living in a World of Low Levels of Predictability," *International Journal of Forecasting*, 25(4), pp. 840-844.

Manasse, P. and Roubini, N. (2009), "Rules of Thumb for Sovereign Debt Crises", *Journal of International Economics*, 78(2), pp. 192-205.

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

Masson, P. (1998), "Contagion: Monsoonal Effects, Spillovers and Jumps Between Multiple Equilibria," *IMF Working Paper*, 142.

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 *International Debt and the Developing Countries*, eds. A. Smith and J. T. Cuddington, pp. 179-209.

Merrik, J. (2001), "Crisis Dynamics of Implied Default Recovery Ratios: Evidence from Russia and Argentina", *Journal of Banking and Finance*, 25(10), pp. 1921-1939.

Milesi-Ferretti, G.and Razin, A. (1998), "Current Account Reversals and Currency Crises: Empirical Regularities," *IMF Working Paper*, 98/89.

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*, 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 Forecast-*

ing, 25, pp. 734-743.

Pan, Y. and Singleton, K. (2006), "Default and Recovery Implicit in the Term Structure of Sovereign CDS Spreads," *Working Paper Graduate School of Business*, Stanford University.

Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*, San Mateo, CA, Morgan Kaufmann.

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

Roy, A. and Tudela, M. (2000), *Emerging Market Risk Indicator (Emri): Re-Estimated Sept 00*, Credit Suisse/First Boston.

Sturzenegger, F. (2004), "Toolkit for the Analysis of Debt Problems," *Journal of Restructuring Finance*, 1(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*, 147, pp. 541-568.

Van Rijckghem, C., Weder, B. (2001), "Sources of Contagion: Is it Finance or Trade?," *Journal of International Economics* 54, pp. 293-308.

Van Rijckghem, C. and Weder, B. (2004), "The Politics of Debt Crises," *Centre for Economic Policy Research Discussion Paper*, 4683.

Vezzoli, M. (2007), *Recent Advances in Classification and Regression Trees*, unpublished PhD Thesis, University of Milano Bicocca, Italy.

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

Vezzoli, M. and Zuccolotto, P. (2011), "CRAGGING Measures of Variable Importance for Data with Hierarchical Structure", in *New Perspectives in Statistical Modeling and Data Analysis*, eds. S. Ingrassia, R. Rocci and M. Vichi, Springer, forthcoming.

Table 1: Debt Crises

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	Rep Dominicana; 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; Ukraine; Philippines
1999	4	Ecuador; Gabon; Mexico; Turkey
2000	3	Argentina; Uruguay; Zimbabwe
2001	1	Brazil
2002	6	Gabon; Indonesia; Moldova; Paraguay; Turkey; Uruguay
Total crises		
1975–1989	72	
1990–2002	40	
1975–2002	**112**	

The table reports the sovereign defaults analysed over the period 1975–2002 also specifying for each year the number of debt crises as well as the countries classified as defaulters.

Table 2: Predictors

Variables	Definition	Missing Value Statistics			Mean		t/z-stat	VIF
		Missing	μ	$\hat{\mu}$	Non Crisis	Crisis		
1. Insolvency Risk Factors								
MAC	Market Access Dummy	-	-	-	0.709	0.777	1.632	-
IMF	IMF Lending Dummy	-	-	-	0.007	0.000	-3.01***	-
CAY	Current Account in % of GDP	-	-	-	-3.697	-5.334	-1.821*	1.060
ResG	Reserves % Change	0.114	28.957	45.729	33.633	2.351	-1.812*	1.065
XG	Exports % Change	0.254	9.263	9.856	9.929	3.733	-4.295***	1.084
WX	Exports in USD Billions	-	-	-	7.761	9.934	1.364	1.296
TEDX	Total Debt to Exports (in %)	0.128	275.583	140.132	250.057	324.091	2.799***	2.358
MG	Imports % Change	0.253	10.636	11.869	11.238	8.134	-1.957*	1.201
FDIY	FDI inflows to GDP (in %)	0.207	1.785	1.985	1.903	1.015	-3.459***	1.302
FDIG	FDI inflows % Change	0.223	62.140	137.948	86.025	27.321	-0.596	1.046
TEDY	Total ext. Debt to GDP (in %)	0.129	45.097	34.760	42.870	51.728	3.185***	18.484
SEDY	Short Term Debt to GDP (in %)	0.168	6.050	4.386	5.478	8.662	6.624***	2.070
PEDY	Public Debt to GDP (in %)	0.130	33.470	21.262	31.300	35.906	1.832*	17.575
OPEN	Exports+Imports to GDP (in %)	0.221	77.006	88.837	81.341	64.148	-4.409***	1.594
2. Illiquidity Risk Factors								
STDR	Short Term Debt to Reserves	0.186	1.884	1.542	1.627	3.946	4.44***	3.811
M2R	M2 to Reserves	0.111	10.544	6.417	9.137	20.232	5.507***	2.932
DSER	Debt service on L-T Debt to Reserves	0.156	1.299	0.940	1.135	2.393	4.854***	2.765
3. Macroeconomic Risk Factors								
DOil	Oil Producing Dummy	-	-	-	0.094	0.071	-0.872	-
INF	Inflation (in %)	0.007	37.386	64.836	29.648	129.205	2.856***	1.029
RGRWT	Real GDP % Change	-	-	-	3.679	1.853	-2.817***	1.141
OVER	Exch. Rate residual over linear trend	-	-	-	-388.235	-100.495	0.211	1.006
UST	US Treasury Bill	-	-	-	6.574	7.845	4.995***	1.903
4. Political/Institutional Risk Factors								
PR	Index of Political Rights	0.119	4.164	3.084	3.999	4.232	1.260	1.300
History	# of past defaults	-	-	-	0.670	0.848	1.834*	1.483
5. Systemic Risk Factors								
Cont_tot	Contagion	-	-	-	2.921	5.232	8.056***	2.042
Cont_area	Regional Contagion	-	-	-	0.867	1.321	3.959***	1.502

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 (1402), μ is the mean of each predictor computed using the observed data and $\widehat{\mu}$ 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-stat is the $t(z)$ statistic computed on mean difference between Crisis and Non Crisis; z-test is for dummy variables (CAY, IMF, DOIL). ***, **, * denote significance at 0.001, 0.05, and 0.1 levels. VIF is the Variance Inflation Factor obtained as $\frac{1}{1-R^2}$ where R^2 is the R-squared obtained by regressing each predictor one at time using the remaining ones as expalanatory variables. VIF values exceeding 5 or 10 indicate a multicollinearity problem.

Table 3: Logit and Stepwise Logit Estimates: 1975–2002

Variable	Logit	Stepwise logit	Austin–Tu Ranking
Intercept	-3.2626 (0.0000)	-3.396 (0.0000)	
CAY	-0.0192 (0.2145)	-	1) TEDY (1.000)
Contagion_area	0.0606 (0.5475)	-	2) OPEN (1.000)
Contagion_tot	0.0476 (0.358)	-	3) XG (1.000)
DOil	-0.249 (0.5862)	-	4) MAC (0.999)
DSER	0.0679 (0.1245)	0.0632 (0.1438)	5) FDIY (0.999)
FDIG	0 (0.7393)	-	6) IMF (0.997)
FDIY	-0.226 (0.0063)	-0.18 (0.0191)	7) SEDY (0.976)
History	0.0529 (0.6679)	-	8) RGRWT (0.961)
IMF	-32.1982 (1)	-14.56 (0.9746)	9) TEDX (0.950)
INF	0.0002 (0.5819)	-	10) M2R (0.936)
M2R	0.0134 (0.0364)	0.0119 (0.0528)	11) UST (0.684)
MAC	0.8195 (0.0125)	0.7877 (0.0124)	12) ResG (0.637)
MG	-0.0032 (0.6682)	-	13) STDR (0.620)
OPEN	-0.0203 (0.0001)	-0.0212 (0.0000)	14) DSER (0.572)
OVER	0 (0.9930)	-	15) WX (0.466)
PEDY	-0.0186 (0.3134)	-	16) PEDY (0.410)
PR	-0.0339 (0.6109)	-	17) Contagion_tot (0.357)
ResG	-0.003 (0.1731)	-0.0038 (0.0885)	18) CAY (0.273)
RGRWT	-0.0378 (0.0352)	-0.0378 (0.0230)	19) Contagion_area (0.078)
SEDY	0.0662 (0.0219)	0.0909 (0.0000)	20) DOil (0.054)
STDR	-0.0482 (0.0806)	-0.0407 (0.1227)	21) PR (0.016)
TEDX	-0.0012 (0.0623)	-0.0012 (0.0384)	22) History (0.008)
TEDY	0.0355 (0.0358)	0.0201 (0.0004)	23) MG (0.004)
UST	0.0856 (0.1266)	0.1329 (0.0008)	24) INF (0.000)
WX	0.009 (0.1882)	0.0103 (0.0997)	25) FDIG (0.000)
XG	-0.0209 (0.0115)	-0.0214 (0.0076)	26) OVER (0.000)

The table reports Logit and Stepwise logit estimates with p-values in parentheses. In the last column we report the Austin and Tu (2004) boostrap method to assess the variable importance. Specifically, we randomly selected 3,000 sub-samples each one constituted by 90 per cent of the total observations, running the Stepwise logit on each bootstrap sample including all 26 candidate variables. The predictors are ordered according to the frequency with which the variable is chosen over the 3,000 regressions. This frequency is reported in parentheses.

Table 4: KLR Ranking: 1975–2002

Variable	**Sens**	**Spec**	$\omega^*_{r,\mathfrak{c}_r}$	$\frac{1}{\omega^*_{r,\mathfrak{c}_r}}$	**Relative Weighs**
SEDY	0.0179	0.9992	0.0434	23.0357	0.2144
M2R	0.0268	0.9984	0.0579	17.2768	0.1608
TEDX	0.0089	0.9992	0.0868	11.5179	0.1072
INF	0.0089	0.9992	0.0868	11.5179	0.1072
DSER	0.0982	0.9798	0.2052	4.8729	0.0454
ResG	0.0179	0.9961	0.2171	4.6071	0.0429
STDR	0.1786	0.9589	0.2301	4.3464	0.0405
Contagion_tot	0.1071	0.9729	0.2532	3.9490	0.0368
Contagion_area	0.0536	0.9860	0.2605	3.8393	0.0357
UST	0.2857	0.9062	0.3283	3.0460	0.0283
TEDY	0.0625	0.9791	0.3349	2.9861	0.0278
WX	0.0625	0.9736	0.4217	2.3713	0.0221
History	0.0179	0.9922	0.4341	2.3036	0.0214
PEDY	0.0893	0.9581	0.4688	2.1329	0.0199
MG	0.0089	0.9946	0.6078	1.6454	0.0153
FDIG	0.1250	0.9132	0.6946	1.4397	0.0134
CAY	0.9018	0.3023	0.7737	1.2926	0.0120
PR	0.6518	0.4295	0.8754	1.1424	0.0106
RGRWT	0.2589	0.7659	0.9041	1.1060	0.0103
OPEN	1.000	0.0109	0.9891	1.0110	0.0094
XG	0.9911	0.0140	0.9949	1.0051	0.0094
FDIY	1.000	0.0008	0.9992	1.0008	0.0093
OVER*	0.9911	0.0000	1.009	-	-

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 ($\omega^*_{r,\mathfrak{c}_r}$). We report also the inverse of the optimal NSR $\left(\frac{1}{\omega^*_{r,\mathfrak{c}_r}}\right)$ which is the weight to be used in calculating the *CI* index according to (3.10). The last column is the weight of each predictor in computing the *CI* index expressed in relative terms $\left(\frac{\frac{1}{\omega^*_{r,\mathfrak{c}_r}}}{\sum_{r=1}^{R}\frac{1}{\omega^*_{r,\mathfrak{c}_r}}}\right)$.

*denotes non informative predictors due to a NSR greater than 1.

Table 5: In-Sample Model Accuracy

Model	***RMSE***	***BIC***	***BS***	***LPS***	***YI***	$\mathfrak{C}^*_{YI}$	**Sens**	**Spec**	***AUC***
Logit	0.2486	-3707	0.1236	0.2241	63.30%	8.10%	76.79%	86.51%	0.8135
Stepwise logit	0.2498	-3774	0.1248	0.2266	65.60%	4.60%	88.39%	77.21%	0.8103
KLR	0.3226	-3128	0.2082	0.8313	37.37%	10.70%	53.57%	83.80%	0.6891
RT	0.2303	-4052	0.1061	0.2097	50.40%	4.60%	58.04%	92.36%	0.7381
FM	0.2463	-3886	0.1213	0.2253	63.77%	6.00%	83.04%	80.74%	0.7861

The table shows the diagnostics used to assess the models' accuracy over the entire period 1975–2002 and computed according to (3.14)–(3.20). *RMSE* is the root mean squared error, *BIC* is the Bayesian information criterion, *BS* is the Brier score, *LPS* is the logarithmic probability score, *YI* is the Youden Index and $\mathfrak{C}^*_{YI}$ 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 $\mathfrak{C}^*_{YI}$. *AUC* is the area under the *ROC* curve.

Table 6: In-Sample *AUC* Differences

***AUC* diff→** ***p*-values↓**	Logit	Stepwise logit	FM	RT	KLR
Logit	-	0.0032	0.0274	0.0754	0.1244
Stepwise logit	0.5383	-	0.0241	0.0722	0.1212
FM	0.2138	0.2626	-	0.0480	0.0970
RT	**0.0072**	**0.0087**	**0.0064**	-	0.0490
KLR	**0.0000**	**0.0000**	**0.0008**	**0.0557**	-

The table shows the *AUC* pairwise differences above the diagonal and corresponding *p*-values under the diagonal computed according to DeLong et al. (1988) (3.19).

Table 7: In-Sample *LF*

Model	ζ							
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
Panel A: Loss Values								
Logit	0.1543	0.1641	0.1738	0.1835	0.1932	0.2030	0.2127	0.2224
Stepwise logit (sLogit)	0.2055	0.1944	0.1832	0.1720	0.1608	0.1496	0.1384	0.1273
KLR	0.1922	0.2225	0.2527	0.2829	0.3132	0.3434	0.3736	0.4038
RT	0.1450	0.1793	0.2137	0.2480	0.2823	0.3167	0.3510	0.3853
FM	0.1880	0.1857	0.1834	0.1811	0.1788	0.1765	0.1742	0.1719
Panel B: Best vs. Worst								
Min Loss	RT	Logit	Logit	sLogit	sLogit	sLogit	sLogit	sLogit
Max Loss	sLogit	KLR	KLR	KLR	KLR	KLR	KLR	KLR

In panel A we report the loss values computed using (3.21) over the entire period 1975–2002, while panel B reports the best and worst model conditional on specific risk aversion level ζ, i.e., the models showing the lesser (Best) and the higher (Worst) value of the *LF*.

Table 8: Time Varying Predictors

Stepwise logit	KLR	RT	FM
1) MAC (1.000)	1) DSER (1.000)	1) STDR (1.000)	1) STDR (1.000)
2) IMF (1.000)	2) INF (1.000)	2) ResG (1.000)	2) ResG (1.000)
3) M2R (1.000)	3) M2R (1.000)	3) Contagion_tot (0.9231)	3) Contagion_tot (1.000)
4) FDIY (1.000)	4) SEDY (1.000)	4) OVER (0.6154)	4) OVER (0.5385)
5) TEDY (1.000)	5) TEDX (0.8462)	5) M2R (0.6154)	5) FDIY (0.3077)
6) OPEN (1.000)	6) TEDY (0.8462)	6) INF (0.4615)	6) INF (0.3077)
7) SEDY (0.9231)	7) STDR (0.6923)	7) SEDY (0.3846)	7) SEDY (0.3077)
8) UST (0.9231)	8) ResG (0.6154)	8) UST (0.3846)	8) TEDX (0.3077)
9) ResG (0.8462)	9) Contagion_tot (0.5385)	9) TEDY (0.3077)	9) TEDY (0.3077)
10) TEDX (0.8462)	10) PEDY (0.5385)	10) MG (0.3077)	10) WX (0.3077)
11) STDR (0.8462)	11) WX (0.3077)	11) Contagion_area (0.3077)	11) DSER (0.2308)
12) WX (0.7692)	12) Contagion_area (0.2308)	12) FDIY (0.1538)	12) M2R (0.2308)
13) XG (0.5385)	13) UST (0.0769)	13) TEDX (0.1538)	13) XG (0.2308)
14) RGRWT (0.4615)	14) FDIY (0.000)	14) DSER (0.1538)	14) CAY (0.1538)
15) PEDY (0.4615)	15) XG (0.000)	15) CAY (0.1538)	15) History (0.1538)
16) CAY (0.1538)	16) OPEN (0.000)	16) OPEN (0.1538)	16) OPEN (0.1538)
17) History (0.0769)	17) PR (0.000)	17) MAC (0.1538)	17) PEDY (0.1538)
18) MG (0.0769)	18) RGRWT (0.000)	18) WX (0.0769)	18) RGRWT (0.1538)
19) DSER (0.0769)	19) FDIG (0.000)	19) XG (0.0769)	19) UST (0.1538)
20) Contagion_tot (0.000)	20) CAY (0.000)	20) History (0.000)	20) MG (0.0769)
21) Contagion_area (0.000)	21) History (0.000)	21) PEDY (0.000)	21) PR (0.0769)
22) DOil (0.000)	22) MG (0.000)	22) RGRWT (0.000)	22) Contagion_area (0.000)
23) INF (0.000)	23) OVER* (0.000)	23) PR (0.000)	23) DOil (0.000)
24) FDIG (0.000)		24) DOil (0.000)	24) FDIG (0.000)
25) OVER (0.000)		25) FDIG (0.000)	25) IMF (0.000)
26) PR (0.000)		26) IMF (0.000)	26) MAC (0.000)

In this table we show the variables selected by Stepwise logit, KLR, RT and FM in the out-of-sample analysis 1990–2002 recalibrating the models one-step-ahead. The variables are listed according to their importance (from higher to lower): (i) for Stepwise logit, RT and FM we use the number of times that variables were selected over the total number of the estimation samples (from 1990 to 2002 we have 13 samples); (ii) for KLR, first, we listed the variables according to their weight used in computing the *CI*, second, we computed the number of times the variables were ranked within the highest percentile accounting for 80 per cent of the total weights, third, such a number was expressed relative to the total number of the estimation samples. Such frequencies are reported for each variable in parentheses.

* denotes non informative predictors due to a NSR greater than 1 (KLR procedure).

Table 9: Out-Of-Sample Model Accuracy

Model	*RMSE*	*DM*	*BS*	*LPS*	*YI*	$\mathfrak{C}^*_{YI}$	Sens	Spec	*AUC*
				Logit					
1990–2002	0.2456	-1.1653 (0.2439)	0.1206	0.2692	0.2926	3.30%	67.50%	61.76%	0.6614
1990–1995	0.2149	-2.2189 (0.0265)	0.0923	0.2674	0.2571	2.70%	58.33%	67.37%	0.6093
1996–2002	0.2636	-1.0667 (0.2861)	0.1389	0.2704	0.3350	3.30%	78.57%	54.93%	0.6771
				Stepwise logit					
1990–2002	0.3596	-6.5416 (0.0000)	0.2586	1.3375	0.2565	4.60%	60.00%	65.65%	0.6384
1990–1995	0.2159	-2.2018 (0.0277)	0.0932	0.2710	0.2839	1.50%	75.00%	53.39%	0.6238
1996–2002	0.4277	-6.7526 (0.0000)	0.3658	2.0280	0.2495	4.20%	71.43%	53.52%	0.6208
				KLR					
1990–2002	0.2475	-1.5742 (0.1154)	0.1225	0.4338	0.2650	20.70%	37.50%	89.00%	0.6253
1990–1995	0.2169	-1.1224 (0.2617)	0.0941	0.3222	0.3277	26.40%	41.67%	91.10%	0.6587
1996–2002	0.2654	-1.1041 (0.2696)	0.1409	0.5060	0.2473	20.70%	35.71%	89.01%	0.6100
				RT					
1990–2002	0.2680	-3.3683 (0.0008)	0.1436	0.3371	0.2163	5.30%	35.00%	86.63%	0.6028
1990–1995	0.2278	-2.4235 (0.0154)	0.1038	0.2861	0.4562	8.60%	58.33%	87.29%	0.7060
1996–2002	0.2910	-3.2721 (0.0011)	0.1694	0.3701	0.1289	5.30%	25.00%	87.89%	0.5301
				FM					
1990–2002	0.2381	-	0.1134	0.2167	0.3684	6.50%	70.00%	66.84%	0.7077
1990–1995	0.2043	-	0.0835	0.1647	0.5184	14.60%	66.67%	85.17%	0.8030
1996–2002	0.2576	-	0.1327	0.2504	0.3330	6.50%	64.29%	69.01%	0.6798

The table reports the diagnostics used to assess the models' accuracy over the entire holdout sample 1990–2002 as well as in the two sub-periods 1990–1995 and 1996–2002. *RMSE* is the root mean squared error, *DM* is the Diebold and Mariano test computed according to (3.22), showing in parentheses the corresponding p-values, *BS* is the Brier score, *LPS* is the Logarithmic Probability Score, *YI* is the Youden Index and $\mathfrak{C}^*_{YI}$ 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 $\mathfrak{C}^*_{YI}$. *AUC* is the area under the *ROC* curve.

Table 10: Out-Of-Sample *AUC* Differences

From 1990 to 2002					
***AUC* diff→**	FM	Logit	Stepwise logit	KLR	RT
***p*-values↓**					
FM	-	0.0463	0.0693	0.0824	0.1049
Logit	0.2474	-	0.0230	0.0361	0.0586
Stepwise logit	**0.0815**	0.3843	-	0.0131	0.0356
KLR	**0.0587**	0.5674	0.8361	-	0.0225
RT	**0.0096**	0.3998	0.6091	0.7685	-
From 1990 to 1995					
***AUC* diff→**	FM	RT	KLR	Stepwise logit	Logit
***p*-values↓**					
FM	-	0.0969	0.1443	0.1792	0.1937
RT	0.2166	-	0.0473	0.0822	0.0968
KLR	0.1441	0.7418	-	0.0349	0.0494
Stepwise logit	**0.0360**	0.5785	0.7881	-	0.0145
Logit	**0.0214**	0.5237	0.7181	0.4021	-
From 1996 to 2002					
***AUC* diff→**	FM	Logit	Stepwise logit	KLR	RT
***p*-values↓**					
FM	-	0.0027	0.0590	0.0698	0.1496
Logit	0.9702	-	0.0563	0.0671	0.1469
Stepwise logit	0.3325	0.2070	-	0.0108	0.0906
KLR	0.2967	0.3297	0.8651	-	0.0799
RT	**0.0155**	**0.0376**	0.1802	0.3113	-

The table shows the AUC pairwise differences above the diagonal and corresponding p-values under the diagonal computed for the entire holdout period 1990–2002 and the two sub-periods 1990–1995 and 1996–2002.

Table 11: Out-Of-Sample *LF*

Model	ζ							
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
Panel A: Loss Values								
				Logit				
1990–2002	0.3709	0.3652	0.3594	0.3537	0.3480	0.3422	0.3365	0.3307
1990–1995	0.3444	0.3534	0.3624	0.3715	0.3805	0.3895	0.3986	0.4076
1996–2002	0.4034	0.3798	0.3561	0.3325	0.3089	0.2852	0.2616	0.2379
				Stepwise logit (sLogit)				
1990–2002	0.3548	0.3604	0.3661	0.3717	0.3774	0.3830	0.3887	0.3943
1990–1995	0.4229	0.4013	0.3797	0.3581	0.3364	0.3148	0.2932	0.2716
1996–2002	0.4290	0.4111	0.3932	0.3753	0.3573	0.3394	0.3215	0.3036
				KLR				
1990–2002	0.2130	0.2645	0.3160	0.3675	0.4190	0.4705	0.5220	0.5735
1990–1995	0.1879	0.2373	0.2867	0.3362	0.3856	0.4350	0.4845	0.5339
1996–2002	0.2165	0.2698	0.3231	0.3764	0.4297	0.4830	0.5363	0.5896
				RT				
1990–2002	0.2369	0.2886	0.3402	0.3918	0.4435	0.4951	0.5467	0.5984
1990–1995	0.1850	0.2140	0.2429	0.2719	0.3008	0.3298	0.3588	0.3877
1996–2002	0.2469	0.3098	0.3727	0.4356	0.4985	0.5613	0.6242	0.6871
				FM				
1990–2002	0.3253	0.3221	0.3190	0.3158	0.3127	0.3095	0.3063	0.3032
1990–1995	0.1853	0.2038	0.2223	0.2408	0.2593	0.2778	0.2963	0.3148
1996–2002	0.3193	0.3240	0.3288	0.3335	0.3382	0.3430	0.3477	0.3524
Panel B: Best vs. Worst								
Min Loss								
1990–2002	RT	RT	FM	FM	FM	FM	FM	FM
1990–1995	RT	FM	FM	FM	FM	FM	sLogit	sLogit
1996–2002	RT	RT	FM	Logit	Logit	Logit	Logit	Logit
Max Loss								
1990–2002	Logit	Logit	sLogit	RT	RT	RT	RT	RT
1990–1995	sLogit	sLogit	sLogit	Logit	KLR	KLR	KLR	KLR
1996–2002	sLogit	sLogit	sLogit	RT	RT	RT	RT	RT

In panel A we report the loss values computed using (3.21) over the periods 1990–2002, 1990–1995, and 1996–2002. For these same time intervals, panel B reports the best and worst model conditional on specific risk aversion level ζ, i.e., the models showing the lesser (Best) and the higher (Worst) value of the *LF*.

Table 13: $2^{\mathrm{D}}\boldsymbol{LF}$

Best									
		ζ							
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
ϱ	**0.8**	RT	RT	RT	FM	FM	sLogit	sLogit	sLogit
	0.7	RT	RT	RT	FM	FM	FM	FM	sLogit
	0.6	RT	RT	RT	FM	FM	FM	FM	FM
	0.5	RT	RT	RT	FM	FM	FM	FM	FM
	0.4	RT	RT	RT	FM	FM	FM	FM	FM
	0.3	RT	RT	RT	FM	FM	FM	FM	FM
	0.2	KLR	KLR	KLR	FM	FM	FM	FM	FM
	0.1	KLR	KLR	KLR	FM	FM	FM	FM	FM
Worst									
		ζ							
		0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
ϱ	**0.8**	sLogit	sLogit	KLR	KLR	KLR	KLR	KLR	KLR
	0.7	sLogit	sLogit	KLR	KLR	KLR	KLR	KLR	KLR
	0.6	sLogit	sLogit	sLogit	KLR	KLR	KLR	KLR	KLR
	0.5	sLogit	sLogit	sLogit	KLR	KLR	KLR	KLR	KLR
	0.4	sLogit	sLogit	sLogit	KLR	KLR	KLR	KLR	KLR
	0.3	sLogit	sLogit	sLogit	sLogit	KLR	KLR	RT	RT
	0.2	Logit	Logit	sLogit	sLogit	RT	RT	RT	RT
	0.1	Logit	Logit	Logit	sLogit	RT	RT	RT	RT

In this table we report the best and the worst model based on minimum and maximum values of the $2^{\mathrm{D}}LF$ computed using (3.23) for different combinations of ζ and ϱ .

Table 14: Rank Comparison

Panel A: Rank Comparison					
	FM	**Logit**	**RT**	**Stepwise logit**	**KLR**
$\overline{\mathfrak{R}}_h$	1.859	2.922	3.047	3.469	3.703
π_h	0.638	0.422	0.397	0.311	0.263
Panel B: Wilcoxon Test					
W-stat→	FM	Logit	RT	Stepwise logit	KLR
***p*-values**↓					
FM	-	-5.54	-3.95	-6.59	-5.19
Logit	**0.001**	-	-0.24	-2.59	-2.65
RT	**0.001**	0.8103	-	-1.59	-3.52
Stepwise logit	**0.001**	**0.0032**	0.1118	-	-0.34
KLR	**0.001**	**0.008**	**0.0004**	0.7339	-

Panel A reports the value for π_h computed according to (5.25) and the corresponding mean rank $\overline{\mathfrak{R}}_h$. Panel B reports the paired Wilcoxon statistics above the diagonal and corresponding p-values under the diagonal.

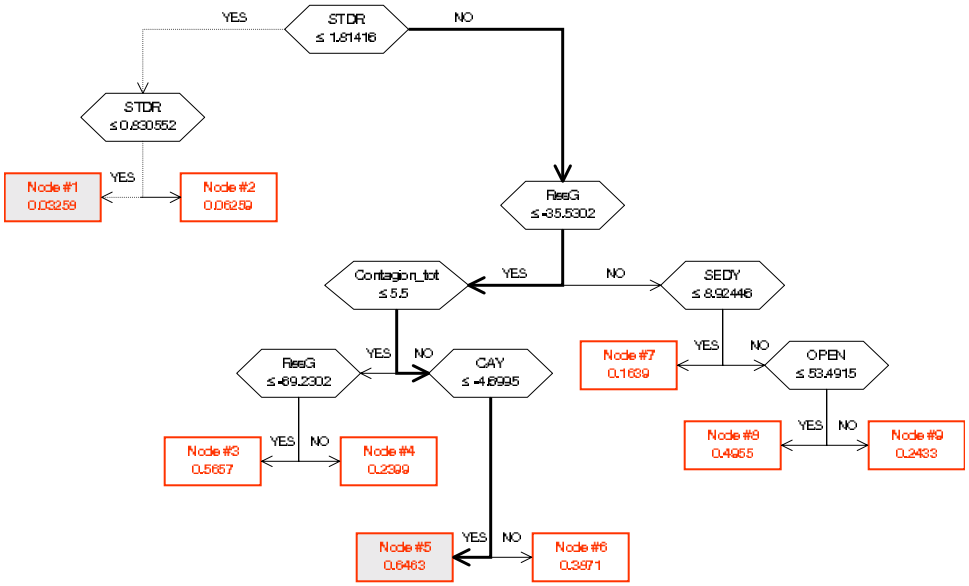

Figure 1: FM Structure

The figure depicts the structure of the tree grown on the CRAGGING estimates, namely the FM estimated over the entire period 1975–2002. 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. 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) default probability.

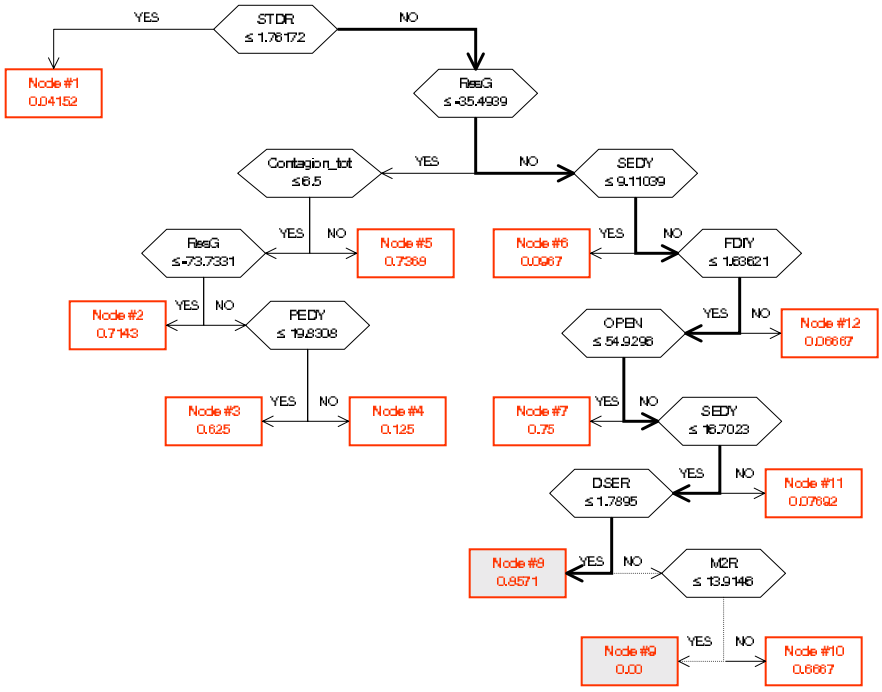

Figure 2: RT Structure

The figure depicts the RT estimated over the entire period 1975–2002. As for the FM, 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. 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) default probability.

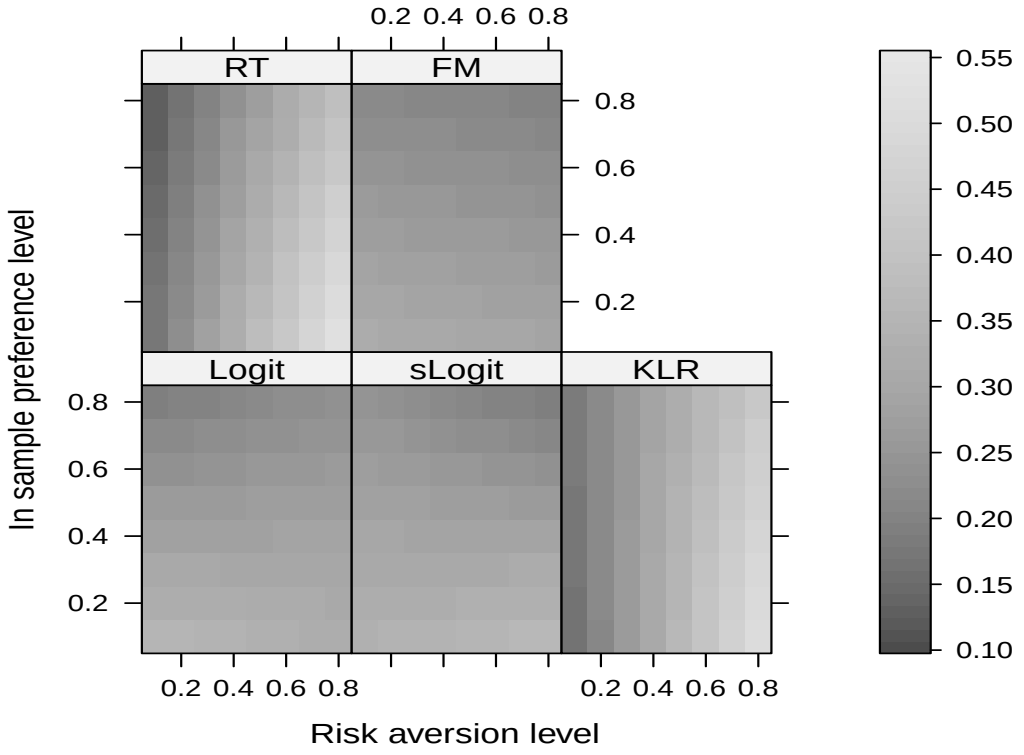

Figure 3: $2^{\mathrm{D}}LF$

In this figure we graphically report the $2^{\mathrm{D}}LF$ bivariate distribution for each model obtained through (3.23). 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.

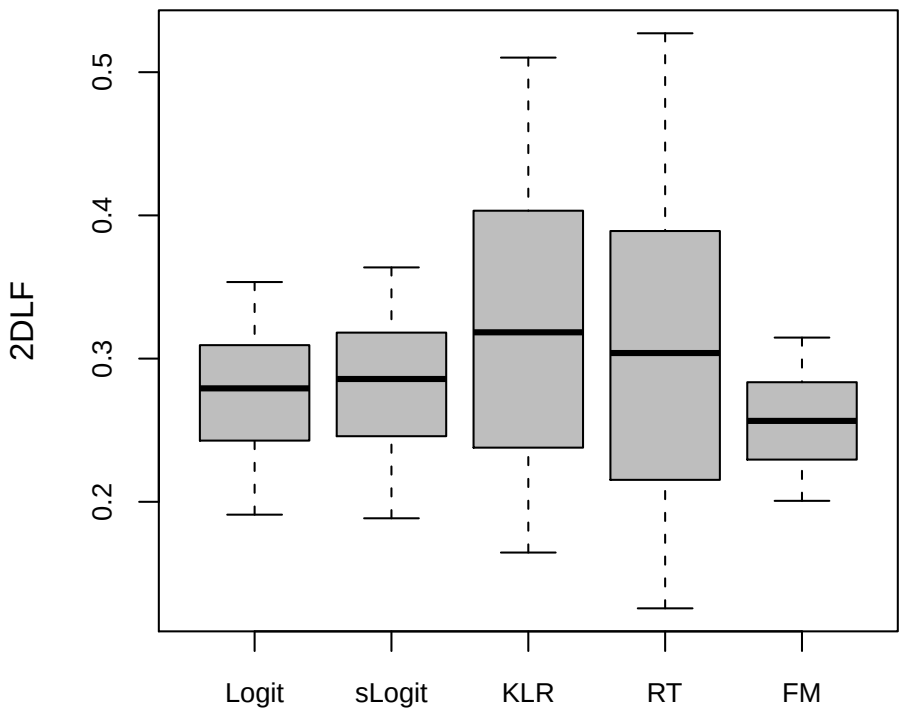

Figure 4: $2^{\mathrm{D}}LF$ Box-Plots

The figure shows the box plots for the models using the $2^{\mathrm{D}}LF$ 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.