How to evaluate an Early Warning System ?

Bertrand Candelon, Elena-Ivona Dumitrescu, Christophe Hurlin

How to Evaluate an Early Warning System ?

Towards a Unified Statistical Framework for Assessing Financial Crises Forecasting Methods

Bertrand Candelon [*] Elena-Ivona Dumitrescu [†] Christophe Hurlin [‡§]

September 2011

Abstract

This paper proposes an original and unified toolbox to evaluate financial crisis Early Warning Systems (EWS). It presents four main advantages. First, it is a model free method which can be used to asses the forecasts issued from different EWS (probit, logit, markov switching models, or combinations of models). Second, this toolbox can be applied to any type of crisis EWS (currency, banking, sovereign debt, etc.). Third, it does not only provide various criteria to evaluate the (absolute) validity of EWS forecasts but also proposes some tests to compare the relative performance of alternative EWS. Fourth, our toolbox can be used to evaluate both in-sample and out-of-sample forecasts. Applied to a logit model for twelve emerging countries we show that the yield spread is a key variable to predict currency crises exclusively for South-Asian countries. Besides, the optimal cut-off correctly allows us to identify now on average more than 2/3 of the crisis and calm periods.

Key words: Currency Crisis, Early Warning System, Credit-Scoring.

J.E.L Classification : C33, F37.

[*]b.candelon@maastrichtuniversity.nl, Maastricht University, School of Business and Economics, Department of Economics,

[†]elena.dumitrescu@univ-orleans.fr, University of Orléans and Maastricht University, Laboratoire d'Economie d'Orléans (LEO),

[‡]christophe.hurlin@univ-orleans.fr, University of Orléans, Laboratoire d'Economie d'Orléans (LEO)

[§]The authors thank two anonymous referees as well as Pierre-Olivier Gourinchas, the editor of the IMF Economic Review for stimulating comments. We are also indebted to the participants at the IMF institute Seminar, European Economic Association (Glasgow), the Econometric Society World Meeting (Shanghai), the 2010 meeting of the Association Francaise de Sciences Economiques (Paris), the 2010 business cycle meeting at Eurostat (Luxemburg) and the MIFN meeting (Luxemburg) for their questions and reactions. The usual disclaimers apply.

1 Introduction

Early Warning Systems (EWS) constitute a crucial tool for authorities to implement optimal policies to prevent or at least attenuate the impact of a financial turmoil. The first EWS was proposed by Kaminsky, Lizondo and Reinhart (1998) (KLR hereafter) relying on a signaling approach. They use a large database of 15 indicators covering the external position, the financial sector, the real sector, the institutional structure and the fiscal policy of a particular country. An indicator signals a crisis when it exceeds a particular cut-off. The estimation of this threshold is at the core of such an analysis. KLR determine it so as to minimize the noise-to-signal ratio (NSR), such that the probability of occurrence of a crisis is at its maximum after exceeding the cut-off. The EWS for country j is then built as the weighted-sum of the individual indicators, the weights being given by the inverse of the NSR. Berg and Patillo (1999) use panel probit models as EWS and show that their forecasting ability outperforms the one obtained using a signaling based model. This analysis hence paved the way for several other studies (Kumar et al., 2003, Fuertes and Kalotychou, 2007, Berg et al., 2008). These EWS do not exploit the fact that financial turmoils refer to specific regimes structurally different from the ones observed during tranquil periods. Hence, Bussiere and Fratzscher (2006) propose a multinomial logit EWS, whereas other studies use Markov-Switching models (see Abiad, 2003, Martinez-Peria, 2002 and Fratzcher, 2003).

Nevertheless, even if these approaches seem to be different, they suffer from similar drawbacks in their evaluation strategies. First, they all use the NSR measure (or sometimes similar measures of correct identification of crisis and calm periods) based on *ad hoc* cut-offs as *in fine* comparison criterion. Yet, as noticed by Bussiere and Fratzscher (2006 p.957) the choice of the cut-off is crucial: if it is low, crises will be more accurately detected (*i.e.,* the type I error will decrease), but at the same time, the number of false alarms will increase (*i.e.,* the type II error) leading to an efficiency cost in terms of economic policy. Second, no statistical inference is provided to test for the forecasting superiority of an EWS compared to another one. This absence represents an important issue, in particular when one has to choose between an EWS model exhibiting low type I and high type II errors and another one with different features.

Therefore, we argue that the evaluation of the forecasting abilities of EWS has not been sufficiently exploited, even though it is essential for crisis forecast. This paper aims at filling this gap by proposing an original evaluation methodology. First, our toolbox is model free, *i.e.* it can be applied to any EWS, whatever the model considered. This characteristic is essential given the great diversity of econometric approaches used in the EWS literature. Second, it can be applied to any type of crisis (currency, banking, debt, etc.) forecasting models. Third,

we not only provide various criteria to evaluate the (absolute) validity of EWS forecasts but also propose some tests to compare the relative performance of alternative EWS. Fourth, our toolbox can be used to evaluate both in-sample and out-of-sample forecasts.

Our evaluation methodology is based on two steps. In a first step, for a given EWS model, we determine optimal cut-off points, *i.e.* thresholds, that best discriminate between crisis and calm periods. Elaborating on the traditional credit-scoring measure (Basel Committee on Banking Supervision, 2005 and Lambert and Lipkovich, 2008 inter alii), it goes beyond a simple analysis of the NSR, by determining the optimal threshold for each country as the value of the cut-off that maximizes (minimizes) different measures balancing type I and type II errors, *i.e. sensitivity-specificity* and *accuracy measures*. In a second step, various criteria and tests are proposed to compare alternative models.

The main finding of our paper is that a correct EWS evaluation requires to take into account the cut-off in the model comparison step and then to determine an optimal crisis forecast. We show that traditional QPS-type criteria tend to conclude to the superiority of a model, even though the two alternative EWS considered have identical forecasting abilities. On the contrary, the criteria integrating the cut-off, *i.e.* AUC, behave correctly in this case. Furthermore, we argue that inference for nested and non-nested hypotheses is essential to identify the optimal specification. The choice of the outperforming model should thus rely on proper statistical tests, which check the significance of the difference between the evaluation criteria associated with two alternative models. To this aim, the classic Diebold-Mariano (1995) test (and its nested version, Clark-West, 2007), as well as an AUC comparison test that takes into account the cut-off are proposed.

To empirically illustrate the utility of such an evaluation toolbox, we propose an application which aims at assessing the relevance of the yield spread in currency crises EWS. For economic theory, yield spreads are usually associated with credit growth sustained by excessive monetary expansion as well as investors' anticipations which can result in capital flight. Hence they may contain information on a potential future distress in the balance of payment. This economic reasoning can be considered as a special case of an EWS specification issue that gages the importance of a leading indicator for correctly forecasting crises. To this aim, we consider two EWS models, *one including the spread, the other without the spread*, in a fixed-effects panel framework. We assess their forecasting performances for six Latin-American and six South-Asian countries. We show that the criteria and tests including the cut-off should be favored as they allow us to refine the forecasting abilities of EWS. Indeed, the yield spread appears to be an important indicator of currency crises in half of the countries when we rely on tests including the cut-off such as the *Area under the ROC test*, whereas it first appears to be essential in almost all the countries in the sample if we consi-

der the general *Clark-West test* based on standard QPS. The outperforming model (with or without spread) for each country is then used to forecast crises by relying on the optimal cut-off. It turns out that the optimal cut-off is quite different from the NSR one, and more importantly it leads to a better trade-off between the two types of errors. In particular, the optimal cut-off correctly identifies on average more than 2/3 of the crisis and calm periods, in contrast with the NSR one, that correctly forecasts all the calm periods at the expense of most of the crisis ones. Our findings seem robust to changes in the crises dating method.

The paper is organized as follows. Sections 2 to 4 present our new evaluation framework. More exactly, we tackle the determination of the optimal cut-off in Section 2. Section 3 introduces the evaluation criteria whereas Section 4 presents the comparison tests. Section 5 is devoted to the empirical application, which reveals the role played by the yield spread in currency crises EWS. Section 6 concludes.

2 Optimal Cut-off

The aim of any EWS is to forecast crisis and calm periods as correctly as possible, so that the appropriate policy measures can be taken in both tranquil and tumultuous situations. In this section we thus propose to quantify how well an EWS discriminates between the two types of periods by identifying the optimal cut-off.

2.1 How important is the cut-off choice ?

EWS deliver probabilities indicating the chance for a specific crisis to occur in a certain period. Therefore, the in-sample (or out of-sample) evaluation of an EWS relies on the direct comparison of these crisis probabilities with an original crisis dating, which constitutes the benchmark. [1] This comparison implies two inputs of different nature: a sequence of probabilities and a crisis dating that takes the form of a dichotomic variable, labelled y_t. By convention, we assume that y_t takes a value equal to one if a crisis is identified at time t and zero otherwise. [2]

The forecasted probabilities are thus transformed into a dichotomic variable, known as crisis forecast. Formally, if we denote $\hat{p}_t$ the estimated (or forecasted) crisis probability at

1. We do not tackle here the pertinence of the crisis dating. We assume that economic experts are able (ex-post) to precisely date the crisis periods. Nevertheless, a robustness analysis with respect to the potential inaccuracy of the crisis dating will be performed in the last section.

2. It can also be assumed that y_t equals one if a crisis occurs in a certain time horizon (6, 12, 24 months, etc.), so as to forecast the approximate timing of a crisis some periods before it actually occurs (see KLR, Berg et al., 1999). This approach presents the advantage of giving the authorities the time necessary to implement appropriate policies to avoid an economic crash.

time t issued from an EWS model, the crisis forecast variable $\hat{y}_t$ is computed as follows:

$$\hat{y}_t(c) = \begin{cases} 1, & \text{if } \hat{p}_t > c \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where $c \in [0,1]$ represents the cut-off. In this perspective, the first step of any EWS evaluation consists in determining an optimal cut-off c that discriminates between predicted crisis periods ($\hat{y}_t(c) = 1$) and predicted calm periods ($\hat{y}_t(c) = 0$). The choice of the cut-off has strong implications for both forecasts evaluation and economic analysis. Obviously, the cut-off determines type I and type II errors, *i.e.* the errors associated to a misidentified crisis or to a false alarm. The type I error (or false negative) corresponds to a case in which the estimated (or forecasted) probability of crisis is smaller than the cut-off, but a crisis occurs. On the contrary, the type II error (also known as false alarm) corresponds to a situation in which the estimated (or forecasted) probability of crisis is larger than the cut-off whereas no crisis occurs. Ceteris paribus, the higher the cut-off is, the more type II (respectively type I) errors are frequent (respectively infrequent).

The optimal cut-off also contains economic interpretation in terms of vulnerability. The higher the probabilities during observed calm periods, the larger the optimal cut-off is and the more the country is vulnerable. This is particularly accurate if the crisis probability series is volatile during the calm periods, i.e. sharp risk exposure. Besides, since the cut-off increases until it finds the optimal balance between type I and type II errors, the closer the level of the crisis probabilities during crisis and calm periods, the more difficult it is to get a correct diagnosis. By contrast, a lower cut-off and lower volatility of the crisis probabilities during calm periods characterize solid economies. However, we have to keep in mind that the exchange market pressure revealed by the optimal cut-off depends on the underlying model and the decision-maker's risk aversion reflected in the method chosen to compute the cut-off. These vulnerability results should hence be interpreted with caution. By contrast, the evolution of the indicators entering the exchange market pressure index, contains useful information relative to a country's resilience to crisis (see Lau et al. 2003). Adequate macroeconomic policies should then lead to a steady evolution of these indicators during the volatile periods and thus to a currency crisis pressure index as close to its average as during the pre-crisis period.

Given the cut-off's importance, it is surprising that the methods used to determine it are so over-arbitrary. At the same time, it exists a very rich literature devoted to the EWS specification topic, stressing the choice of the most pertinent explanatory variables, the consequences of particular events on the models, etc. (see Jacobs et al., 2004 for a survey on

financial crises EWS). And yet, to the best of our knowledge, no paper has been devoted to EWS evaluation, and more specifically to the choice of an optimal cut-off.

At the time being, two types of cut-offs have been used in this literature. In most papers, the cut-off is arbitrarily fixed, generally to 0.5 or 0.25. This approach is economically nonsensical, since it means to arbitrarily determine type I and type II nominal risks. In other ones, it relies on the "Noise to Signal Ratio" (NSR) criterion proposed by KLR. In order to define the NSR cut-off, let us consider a sequence $\{y_t, \widehat{y}_t(c)\}_{t=1}^{T}$.

Definition 1. *The optimal cut-off according to KLR minimizes the NSR criterion and is defined as*:

$$c_{NSR} = \underset{c\in[0,1]}{\arg\min} NSR(c), \tag{2}$$

where $NSR(c)$ represents the ratio of the false alarms (type II error or false positive) to the number of crises correctly identified (true positive) by the EWS for a given cut-off.

$$NSR(c) = \frac{\sum_{t=1}^{T} \mathbb{I}_{(\hat{y}_t(c)=1)} \times \mathbb{I}_{(y_t(c)=0)}}{\sum_{t=1}^{T} \mathbb{I}_{(\hat{y}_t(c)=1)} \times \mathbb{I}_{(y_t(c)=1)}}, \tag{3}$$

where $\mathbb{I}_{(z)}$ denotes an indicator function that takes the value 1 if z is true and 0 otherwise.

However, this criterion omits type II error, and assumes that the costs related to the occurrence of a misidentified crisis (type I error) overweight the ones inflicted by a false alarm (type II error). This clearly constitutes an important constraint one may want to rule out, since type I error is usually the main element we try to control, as generally done in other statistical literatures. We thus propose two methods which identify the optimal cut-off by taking into account both types of errors.

2.2 A credit-scoring approach

The first method is based on the traditional credit-scoring notions of *sensitivity* and *specificity* (Basel Committee on Banking Supervision, 2005). We thus label it the *credit-scoring* approach.

Definition 2. *The optimal credit-scoring cut-off is the threshold that minimizes the absolute value of the difference between sensitivity and specificity*:

$$c^*_{CSA} : \underset{c\in[0,1]}{\arg\min} |Se(c^*_{CSA}) - Sp(c^*_{CSA})|, \tag{4}$$

where sensitivity (Se), also known as hit rate, represents the proportion of crisis periods correctly identified by the EWS, while specificity (Sp) is the proportion of calm periods correctly identified by the model:

$$Se\,(c) = \frac{\sum_{t=1}^{T} \mathbb{I}_{(\hat{y}_t(c)=1)} \times \mathbb{I}_{(y_t(c)=1)}}{\sum_{t=1}^{T} \mathbb{I}_{(y_t(c)=1)}}, \tag{5}$$

$$Sp\,(c) = \frac{\sum_{t=1}^{T} \mathbb{I}_{(\hat{y}_t(c)=0)} \times \mathbb{I}_{(y_t(c)=0)}}{\sum_{t=1}^{T} \mathbb{I}_{(y_t(c)=0)}}. \tag{6}$$

The underlying idea is that variation in the cut-off leads to higher values of *sensitivity* corresponding to lower values of *specificity*. Figure 1 displays the *specificity* and *sensitivity* of an hypothetical EWS as functions of the cut-off c. The *sensitivity* is a decreasing function of c, since a rise in c results in decreasing the number of crisis signals, $\hat{y}_t\,(c) = 1$, and thus in the percentage of crises correctly predicted. On the contrary, the specificity is an increasing function of c. The higher c is, the higher the number of calm signals, $\hat{y}_t\,(c) = 0$, and hence, the larger the proportion of calm periods correctly identified. The general form of both curves depends on the specification of the EWS. Even so, an optimal cut-off can be found at the intersection of both curves, as shown in Figure 1.

Insert Figure 1

The main advantage of this *credit-scoring* identification method is that, in contrast to the NSR criterion, it relies on both type I and type II errors (see Engelmann et al., 2003, Renault et al., 2004 and Stein, 2005). It assigns equal weight to both types of errors. However, this assumption should be relaxed if we assume that the identification of a crisis is more costly for an economy than a false alarm (or vice versa). A possible extension of our method can be envisaged so as to take into account the costs c_1 and c_2 associated to the non-predicted crises and to a false alarm respectively. It simply consists in determining the optimal cut-off as the threshold that minimizes the difference between the weighted *sensitivity* and the weighted *specificity*, where the weights are defined by c_1 and c_2 respectively. For example, the costs of the misidentification of a crisis -in terms of GDP- (c_1) can be approximated by an econometric evaluation of GDP gap during crisis periods. By contrast, the costs linked to false alarms (c_2) cannot be assessed easily, because they consist of costs incurred as a result of the reaction of monetary and/or banking authorities to an unfounded crisis announcement. If the policymaker can estimate these costs, however, this weighted method of identification of the optimal cut-off should be privileged.

Alternatively, instead of directly arbitrating between type I and type II errors, the optimal cut-off can be determined as the one that maximizes some accuracy measures or the one that minimizes the misclassification error measures respectively (Lambert and Lipkovich, 2008).

2.3 Accuracy Measures

The second approach consists in aggregating the number of crisis and calm periods correctly identified by the EWS in an accuracy measure. c is thus obtained by the maximization of the corresponding accuracy measure. The simplest measure, named *Total Accuracy* (TA), is defined as the ratio of cases correctly predicted to the total number of periods. Maximizing the TA measure is thus equivalent to maximize the number of correctly identified periods, whatever their type (crisis or calm). This measure does not arbitrate between type I and type II errors as the two types of periods are not considered separately (the denominator represents the total number of periods in the sample). We can thus be confronted with an undesirable situation in which the optimal cut-off correctly identifies all calm periods, but only a few, or none of the crisis periods. We hence propose another measure, which arbitrates between the two types of errors.

Definition 3. *According to this accuracy measure, the optimal cut-off satisfies*:

$$c^*_{AM} = \underset{c\in[0,1]}{\arg\max} J(c), \tag{7}$$

where $J(c)$ *denotes the Youden Index, defined as* $J = Se(c) + Sp(c) - 1$.

The *Youden Index* ranges between 0 and 1; the higher the proportion of calm and crisis periods correctly identified (relatively to the number of crisis and calm periods) by the model, the greater the J-measure. This optimal cut-off c^*_{AM} also corresponds to the cut-off that minimizes the misclassification error measure (also called *Total Error* measure) defined as the sum of the ratios of misidentified crises and false alarms to the number of crisis and calm periods respectively. More formally, the *Total Error* is defined as $TME(c) = 2 - Se(c) - Sp(c)$ and corresponds to $1 - J(c)$.

3 Evaluation Criteria

Traditionally, the forecasting abilities of an EWS are assessed only on the basis of the crisis probabilities p_t, *i.e.* independently of the cut-off. In fact, two criteria are generally used, namely the *Quadratic Probability Score* (QPS) and the *Log Probability Score* (LPS).[3] The QPS statistic is simply a mean square error measure comparing the crisis probability

3. The *Log Probability Score* (LPS), corresponds to a loss function that penalizes large errors more heavily than QPS, with $LPS = -1/T \sum_{t=1}^{T} \left[(1 - y_t)\ln(1 - \hat{p}_t) + y_t \ln(\hat{p}_t)\right]$. This score ranges from 0 to ∞, with $LPS = 0$ being perfect accuracy.

(the prediction) with an indicator variable for the crisis. It is defined as:

$$QPS = \frac{2}{T}\sum_{t=1}^{T}(\hat{p}_t - y_t)^2, \tag{8}$$

where $\hat{p}_t$ represents the *ex-ante* forecast probability of crisis at time t and y_t is a dummy variable taking the value one when a crisis occurs at time t. QPS takes values from 0 to 1, with 0 indicating perfect accuracy. This metric originated in weather forecasting and has been introduced by Diebold and Rudebusch (1989). It relies on the sum of squared residuals as in a standard linear model.

However, these traditional criteria evaluate the EWS only on the basis of the gap between the crisis probability and the realization of the observed crisis variable. The cut-off, essential for an EWS, is not taken into account.

We thus propose to include the cut-off in the validation of EWS via two main ways. On the one hand, the EWS (*i.e.* the crisis probabilities) and the optimal cut-off can be jointly validated. Still, the optimal cut-off has been identified so as to maximize (minimize) the accuracy (misclassification error) measures. Hence, we cannot use the same or similar criteria to jointly assess the probabilities and the optimal cut-off. We draw an analogy with calibration and validation of DSGE models (Kydland and Prescott, 1991), where the moments of interest ("validation" step) are different from the auxiliary moments used in calibration ("estimation" step). This approach is thus unfeasible in our context. On the other hand, we can assess the forecasting abilities of an EWS *conditionally to all the values of the cut-off*, *i.e.* from 0 to 1, in a similar vein to robustness analysis. This constitutes the main advantage of this approach, as the predictive abilities of a "good" EWS should not break down for reasonable changes in the value of the cut-off.

3.1 Cut-off based criteria

In this context, we propose an original evaluation criterion, the *Area under the ROC Curve* (*AUC*), first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battle fields, and then used in medicine, machine learning and credit scoring literature.

Definition 4. *The ROC (Receiving Operating Characteristic) curve is a graphical tool which reveals the predictive abilities of an EWS. More exactly, it represents the trade-off between sensitivity and* $1 - specificity$ *for every possible cut-off. The ROC curve is thus obtained by representing all the couples* $\{Se(c); 1 - Sp(c)\}$ *corresponding to each value of the cut-off* c *ranging from 0 to 1 (see Figure 2).*

For a perfect EWS model, the *ROC* curve passes through the point (0,1), indicating that it correctly recognizes all crisis and non-crisis periods. On the contrary, a completely random guess about crisis would give a point along a diagonal line (the so-called line of no-discrimination) from the left bottom to the top right corners.

Insert Figure 2

This criterion can be summarized by the *Area Under the ROC* curve *(AUC)* defined as:

$$AUC = \int_0^1 \left[Se\left(c\right) \times \left(1 - Sp\left(c\right)\right)\right] \; d(1 - Sp\left(c\right)). \tag{9}$$

An area under the *ROC* curve approaching 1 indicates that the EWS is getting closer to the perfect classification. In contrast, the expected value of the *AUC* statistic for a random ranking is 0.5.

The *AUC* is straightforward to implement, since it can be estimated by using an average of a number of trapezoidal approximations. Another way to obtain the *AUC* consists in using a non-parametric kernel estimator, as follows:

$$AUC = \frac{1}{T_1 \times T_0} \sum_{j:y_j=0} \sum_{i:y_i=1} K(\hat{p}_j, \hat{p}_i), \tag{10}$$

where T_1 (T_0) is the number of crisis (calm) periods in the sample, and $K\left(.\right)$ denotes a kernel function that depends on the estimated crisis probabilities in crisis periods ($p_i, \forall i: \; y_i = 1$) and calm periods ($p_j, \forall j: \; y_j = 0$) defined by:

$$K(\hat{p}_j, \hat{p}_i) = \begin{cases} 1, & \text{if } \hat{p}_i < \hat{p}_j \\ \frac{1}{2}, & \text{if } \hat{p}_i = \hat{p}_j \\ 0, & \text{if } \hat{p}_i > \hat{p}_j. \end{cases} \tag{11}$$

Our toolbox allows us to evaluate EWS models by taking into account the cut-offs (and therefore the crisis forecasts, the most important output of such a model) apart from relying solely on the crisis probabilities. The best model according to the *AUC* criterion is the outperforming EWS whatever the cut-off, and implicitly it is the best one conditional to the choice of the optimal cut-off. Indeed, taking into account the cut-off in the evaluation of an EWS can turn out to be crucial for an EWS with respect to a simple analysis based on *QPS*-type criteria.

3.2 Example

Let us consider a simple example of two EWS (denoted by A and B) with exactly the same forecasting abilities, but different estimated crisis probabilities. To be more exact, we suppose that the series of probabilities associated with model B, *i.e.* p_B, corresponds to an upward shift of the sequence of probabilities associated with model A, p_A, by a constant α: $p_A = p_B + \alpha$. It results that the two models have the same forecasting abilities, since the optimal cut-off for model B differs from that of model A only by α. The *sensitivity* and *specificity* (and implicitly type I and type II errors) series associated with the two EWS are thus identical. And yet, in this context, QPS-type criteria wrongfully privilege one of the models. On the contrary, when taking into account the cut-off in the evaluation, *e.g.* AUC, we can conclude that the two models are utterly equivalent in terms of crisis forecasts.

Let us assume that $\alpha > 0$. In such a case, the crisis probabilities obtained from model B are always higher than those of model A, as illustrated in figure 3.

Insert Figure 3

Consider that the frequency of crisis occurrence is low and that ex-post we observe two crises of different but limited durations (which is the more common case in EWS literature). QPS is based on the difference between the probabilities outputted by the EWS and unit (zero) during crisis (calm) periods. Accordingly, it corresponds to the sum of squared differences, depicted by hatched areas in figure 3. It is thus clear that QPS for model B is higher than the one for model A if crises are not frequent, pointing out that model A has better predictive abilities. Still, the result depends on the frequency of observed crises (the length of hatch areas 1 and 2 in figure 3). Formally, model A is preferred to model B if and only if:

$$QPS_B - QPS_A > 0 \quad \Leftrightarrow \quad 2\sum_{t=1}^{T}[\alpha^2 - 2\alpha(y_t - \hat{p}_{A,t})] > 0. \tag{12}$$

This condition is fulfilled if:

$$\sum_{t=1}^{T} y_t < \frac{T\alpha + 2\sum_{t=1}^{T}\hat{p}_{A,t}}{2}, \tag{13}$$

As a result, if the proportion of crisis periods, *i.e.* $1/T\sum_{t=1}^{T} y_t$, is small relatively to the ratio of the sum of probabilities to the number of periods, *i.e.* $1/T\sum_{t=1}^{T} p_{A,t}$ and the constant α, for $\alpha > 0$, the first model, A, is improperly considered to be more parsimonious than the latter, B. Besides, if the frequency of crises is below $\alpha/2$, eq. 13 is fulfilled independently of the sum of probabilities $\{p_{A,t}\}_{t=1}^{T}$. QPS privileges either model A or model B, even though

they have identical forecasting abilities, except for the case where $\alpha = 0$, *i.e.* the two series of probabilities are identical. This finding is true provided that the cut-off has not been taken into account in the evaluation of the EWS. By contrast, the ROC evaluation criteria allow us to confirm this equivalence.

To illustrate this theoretical findings, let us consider the series of estimated probabilities, p_A, for two countries, Brazil (over the period 1994-2010) and Indonesia (over the period 1986-2009) obtained by estimating a logit model (see the empirical section for more details). Denote by p_B the sequence of probabilities obtained by shifting p_A upwards by $\alpha = 0.2$. The left part of table 1 presents the QPS, and the AUC for each country and model.

Insert Table 1

Notice that QPS differs from one model to another, leading to the improper conclusion that the first model is slightly better than the second one and should be privileged for taking policy decisions. By comparison, the criteria based on the ROC curve are identical for the two models. To be more precise, not only the areas under the ROC, *i.e.* AUC, are equivalent, but also the ROC *curves* themselves.

Better still, apart from the aforementioned criteria ($QPS$and AUC), the comparison of two EWS models must rely on statistical test that we present in the next section.

4 Comparison Tests

Usually, the EWS literature aims to propose new econometric specifications (panel logit, Markov switching model, time varying probabilities Markov switching model etc.) or more frequently, new choices of explanatory variables in order to improve the crisis forecast ability (out-of-sample analysis) or the explanation of the crisis origins (in-sample analysis). These choices cannot be reduced to simple tests of significance, even if we are interested in the influence of a given variable on the crisis probability. It is well known that the significance of a parameter associated to a particular economic variable (in-sample) does not necessarily mean that this variable is able to improve the forecast ability of the EWS.

Hence, the EWS literature should be based on the comparison of forecasts issued from alternative models. However, this comparison is usually conducted according only to simple criteria such as the QPS (with the drawbacks previously mentioned) without any statistical inference (see, for example, Kaminski, 2003, Arias and Erlandsson, 2005, Jacobs et al., 2008) even if they already exist in the statistical literature.

Accordingly, in the last step of our evaluation procedure, we propose to use a set of (model free) comparison tests in order to test the differences in crisis forecasts obtained

from alternative EWS models. For that, let us consider two EWS models, denoted 1 and 2. Denote by $\{y_t\}_{t=1}^T$ the sequence of observed crises series and $\{\hat{p}_{j,t}(c_1)\}_{t=1}^T$ the sequence of probabilities obtained from the EWS model j for $j = 1, 2$.

The first test we propose is a non parametric test of comparison of *ROC* curves (DeLong et al., 1988). It is based on the comparison of the areas under the ROC curves associated with the two EWS models, denoted AUC_1 and AUC_2. The null of the test corresponds to the equality of areas under the *ROC* curves *i.e.*, $H_0 : AUC_1 = AUC_2$; in other words, neither of the models performs better than the other. DeLong et al. propose a test statistic based on the difference of AUC and use the theory on generalized U-statistics to propose an estimator of the variance of the difference (see Appendix 1 for technical details).

Definition 5. *Under the null* $H_0 : AUC_1 = AUC_2$, *the two EWS forecasts are equivalent, and the AUC curve test statistic satisfies (DeLong et al., 1988)*:

$$W_{AUC} = \frac{(AUC_1 - AUC_2)^2}{\mathbb{V}(AUC_1 - AUC_2)} \xrightarrow[T\to\infty]{d} \chi^2(1), \tag{14}$$

The second test is the seminal test of comparison of forecast accuracy proposed by Diebold-Mariano (1995) and its specific version for nested models, proposed by Clark and West (2007). Both tests are based on the forecast errors of the two models, denoted $\{e_{1,t}\}_{t=1}^T$, and $\{e_{2,t}\}_{t=1}^T$, with $e_{j,t} = y_t - \hat{p}_{j,t}$ for $j = 1, 2$. The null corresponds to the hypothesis of equal forecasting accuracy, conditionally to a particular loss function $g(.)$. These tests are very general and can be applied with any type of loss function including MSFE, MAE, etc.. Since there is no specific loss function in the case of EWS models, we propose to use the MSFE, with $g(e_{j,t}) = (y_t - \hat{p}_{j,t})^2$. This loss function corresponds to half the QPS standard criterion, since $QPS = 2/T \sum_{t=1}^T g(e_{j,t})$.

Definition 6. *Under the null hypothesis of equal predictive accuracy of both EWS,* H_0: $\mathbb{E}[(y_t - \hat{p}_{1,t})^2] = \mathbb{E}[(y_t - \hat{p}_{2,t})^2]$, *the Diebold and Mariano (1995) test statistic DM verifies*

$$DM = \frac{\sqrt{T}\bar{d}}{\sigma_{\bar{d},0}} \xrightarrow[T\to\infty]{d} N(0,1), \tag{15}$$

where d_t *denotes the loss differential,* $d_t = (y_t - \hat{p}_{1,t})^2 - (y_t - \hat{p}_{2,t})^2$, $\bar{d}$ *is the loss differential mean,* $\bar{d} = (1/T)\sum_{t=1}^T d_t$ *and* $\sigma^2_{\bar{d},0}$ *is the asymptotic long run variance of the loss differential.*

Following standard practice, the long run variance $\sigma^2_{\bar{d},0}$ can be estimated with a Kernel estimator as a weighted sum of the available sample autocovariances of the loss differentials $\{d_t\}_{t=1}^T$. Given the definition of the DM statistic, it is obvious that it fails to converge

when the models are nested (since the denominator converges to zero). However, in EWS literature we often come across cases requiring to compare nested models. An appropriate test for nested models has been suggested by Clark and McCracken, (2001) and Clark and West, (2007). [4]

Note that the test of comparison of ROC curves relies on the AUC criterion whereas the DM test-statistic is based on the same loss function (MSE) as the QPS criterion. It follows that the ROC test takes into account not only the observed crises periods and the crises probabilities issued form two EWS specifications, as the DM (and its nested alternative CW) but also all the values of the cut-off.

Let us return to the previous example of Brazil and Indonesia in which we compare two EWS that have the same forecasting abilities. In the right part of table 1 we present the test statistic and p-value for Clark-West (1997)'s test, based on a QPS-type loss function, and DeLong (1988) test, relying on AUC differences. It results that for these two countries, when the two series of probabilities are different enough ($\alpha = 0.2$), the W_{AUC} test does not reject the null hypothesis of equal forecasting abilities. On the contrary, the CW test leads to the rejection of the null hypothesis and consequently to an improper choice of model A over B. These findings emphasize the importance of the three evaluation tests for comparing EWS models.

5 Empirical Application

We now propose an empirical application to illustrate the importance of the EWS evaluation procedure. This application focuses on the role of the yield spread (*i.e.*, long term government bonds minus the short term money rate) as a forward-looking indicator in the construction of EWS models. In a more general perspective, it can be viewed as an example of EWS specification where the main issue consists in assessing the importance of a leading indicator for correctly forecasting crises.

The yield spread can be considered as a forward interest rate that can be decomposed following the expectation hypothesis theory into an expected real interest rate and an expected inflation component (Estrella and Mishkin, 1996). It is hence linked to both changes

4. Let us assume that model 1 is the parsimonious model and model 2 is the larger one, that reduces to model 1 if some of its parameters are set to 0. The corrected statistic proposed by Clark and West (2007), denoted CW, is defined as follows:

$$CW = \frac{\sqrt{T}\bar{f}}{\sigma_{\bar{f},0}} \xrightarrow[T\to\infty]{d} N(0,1), \tag{16}$$

where $\hat{f}_t = (y_t - \hat{p}_{1,t})^2 - [(y_t - \hat{p}_{2,t})^2 - (\hat{p}_{2,t} - \hat{p}_{1,t})^2]$, $\bar{f}$ is the sample average of $\hat{f}_t$ and $\sigma^2_{\bar{f},0}$ is the sample variance of $\hat{f}_t - \bar{f}$.

in investors' expectations and expectations of future monetary policy. Since currency crises have been associated with credit growth sustained by excessive monetary expansion in many countries and investors anticipations can result in capital flight, aggravating a potential crisis, yield spreads can be assumed to reflect distress in the balance of payment. Moreover, since the yield spread seems to outperform other variables at long term forecasting horizons that are relevant from an investor's point of view, this variable is more forward-looking than other leading indicators (Estrella and Hardouvelis, 1991, Estrella and Trubin, 2006). Consequently, the use of the yield spread as a forecasting tool is even more compelling since it can signal the occurrence of a crisis in advance.

In this context, we propose to apply our evaluation methodology to asses the genuine usefulness of yield spread in forecasting currency crises.

5.1 EWS Specification

In order to asses the influence of yield spread, we consider a simple Logit EWS. More formally, let y_{it} represent the binary crisis variable for country $i \in \{1, ..N\}$ at time $t \in \{1, ..T_i\}$. T_i denotes the number of time periods considered for the i^{th} country (unbalanced panel). For each country, the crisis (logit) probability is defined as follows:

$$\Pr(y_{it} = 1) = \frac{\exp(\alpha_i + \beta_i' x_{it})}{1 + \exp(\alpha_i + \beta_i' x_{it})}, \ i = 1, ..., N, \tag{17}$$

where x_{it} denotes a vector of macroeconomic indicators, that includes yield spread. α_i denotes a constant and β_i the vector of slope parameters. In this first specification, all parameters are country specific.

The approach generally used in the literature consists in estimating the binary EWS model in a panel framework by imposing some restrictions on the β_i parameters (see Berg and Patillo, 1999, Kumar et al., 2003 *inter alii*). It is well know that the panel approach is a way to reveal unobservable country heterogeneity and to increase the information set. This last point is particularly important in the specific context of currency crisis, given the relative scarcity of such events. However, this advantage has an obvious limit: the more heterogeneous countries are pooled in a "meta" model, the less the restrictions ($\beta_i = \beta$, for all i) on the slope parameters β_i are likely to make sense (even if we introduce individual effects α_i). This trade-off between more information and heterogeneity of slope parameters β_i can be summarized by a simple question, to pool or not to pool ? To overcome this issue, Berg et al. (2008) recommend to construct country clusters, for which the slope parameters can be assumed to be homogeneous. There are two main approaches to construct such country clusters. The

first one, proposed by Kapetanios (2003) is a pure statistical method based on an iterative procedure of homogeneity tests. The second approach focuses on macroeconomic similarities, crisis transmission mechanisms, contagion, etc.. We favor here the latter and consider two regional clusters, the first one including the Latin-American countries and the second the South-Asian ones. [5] For each country the crisis probability is then defined as follows:

$$\Pr(y_{it} = 1) = \frac{\exp(\alpha_i + \beta' x_{it})}{1 + \exp(\alpha_i + \beta' x_{it})} \forall i \in \Omega_h, \tag{18}$$

where Ω_h is the h^{th} regional cluster, $h \in \{1, ..., H\}$, and $\dim(\Omega_h) = N_h$, so that $\sum_{h=1}^{H} N_h = N$, and $\dim(\Omega_h)$ is the number of countries in the h^{th} cluster. α_i represents the fixed effects (*i.e.,* the constant term specific to each country).

5.2 Data and Estimation

We consider a sample of twelve countries [6] for the period January 1980 to December 2010 extracted from the IMF-IFS database as well as the national banks of the countries under analysis via Datastream.The currency crisis indicator (the dependent variable), representing crises in the coming 24 months, is obtained by implementing the Kaminski, Lizondo and Reinhart *(1998)* modified dating method, thereafter $KLRm$, proposed by Lestano and Jacobs (2004) (see Appendix 2 for more details). Note that in the case of binary EWS models there is a debate related to the crisis dating quality, contrary to Markov-based models which do not require an *a-priori* identification of crises in the estimation step. However, the dating method impacts not only the estimation of an EWS, but also its evaluation, which means that the evaluation of Markov-based EWS depends on the dating method too. To check the sensitivity of our results to the dating method, we perform a robustness analysis based on the pressure index proposed Zhang (2001), instead of $KLRm$.

In all the estimated models (regional and pooled panel logit as well as country-by-country logit), the set of explicative variables includes growth of international reserves, growth of exports, growth of domestic credit over GDP, first difference of lending over deposit rate, first difference of industrial production index and yield spread. All variables are lagged one period. Among them, the yield spread plays a key role in our analysis since we aim to gauge its contribution to the improvement of the EWS. The other predictors are classic leading indicators for currency crises, associated with devaluation pressure, loss of competitiveness,

5. Note that, as a robustness check, we have also considered the pooled logit model as well as the optimal clusters derived from the Kapetanios procedure.

6. Argentina, Brazil, Mexico, Peru, Uruguay, Venezuela, Indonesia, South Korea, Malaysia, Philippines, Taiwan and Thailand.

indebtedness, loan quality and recessions, respectively (see KLR ; Berg and Patillo, 1999 ; *inter alii*). The procedure used to select these leading indicators is described in Appendix 2. A thorough attention has been given to the stationarity of the series, outliers and especially to the possible correlation among leading indicators (see Appendix 2 for more details). We also take into account the potential presence of serial correlation [7] using the sandwich estimator (Williams, 2000). This method is described in Appendix 3.

insert table 2

The estimation results for the model including the yield spread are presented in table 2. The yield spread is one of the most important explanatory variables both in panel and time-series models. It is significant at a 5% level for the Latin-American cluster as well as for 11 out of 12 countries. The results for the pooled panel model confirm these findings. This first result implies that in all the countries considered, a higher short term interest rate with respect to the long term one, *i.e.* a negative slope in yield spread, signals future balance of payment problems that lead to currency crises. The other explanatory variables generally have the correct sign too, but their significance plummets when regrouping the countries in a panel set without accounting for the heterogeneity of the estimated parameters.

5.3 EWS Evaluation

To analyze the importance of the yield spread in forecasting currency crises, we consider two specifications of our EWS models, *one that includes the yield spread indicator and another that does not include it* and compare their forecasting abilities. These models are estimated for the two optimal regional clusters (South America and South Asia). First, let us compare the two specifications (with and without spread) using a QPS-type criteria, as it is usually done in the literature. The left part of table 3 displays the QPS corresponding to the two models for each of the twelve countries in our sample. The QPS criteria seems to confirm that the spread improves the forecasting abilities of the EWS for almost all the countries (10 out of 12).

Insert Table 3

While most of the papers in the literature would stop here and conclude to the importance of the yield spread for all countries, we propose to go further on and to test if the spread is *significantly* important for a currency crisis. As shown in section 4, since both specifications

7. Berg and Coke (2004) show, that considering a forecast horizon larger than 1 leads to autocorrelation in the crisis variable. This stylized fact is confirmed by Harding and Pagan (2006).

are nested, (the former can be reduced to the latter by imposing the nullity of the spread parameter), Clark-West's (2007) CW test is used to compare the forecasting ability of the two logit models. The results are displayed in the right part of table 3. The tests roughly confirm the importance of spread's contribution to currency-crises forecasting but only at 10% level, offering hence a more refined diagnostic than QPS-criteria.

In particular, in the case of Brazil, Malaysia and Venezuela, the test rejects the null of equal forecasting abilities only at 10% significance level, whereas for Peru it cannot reject it at all. This step provides evidence that the introduction of the spread in the model does not provide sufficient information to improve the model's ability to correctly forecast currency crises. Figure 4 depicts the crisis probabilities for Brazil issued from the model with spread (EWS1) and the model without spread (EWS2). Graphically, the series of probabilities seem almost identical, confirming the results of the CW test. These findings support the use of statistical tests to compare EWS instead of relying on simple $QPS - type$ criteria.

Insert Figure 4

In the next step we propose to check this diagnostic by relying on criteria integrating the cut-off. We thus use the AUC evaluation criterion (see the left part of table 3) and find evidence of a positive effect of the yield spread on the forecasting abilities of the EWS. For most of the countries, the AUC is always higher for the logit with yield spread relatively to the logit without yield spread. Still, relying on AUC criterion instead of QPS, we identify one country, namely Argentina, for which the yield spread does not contribute to the improvement of the EWS.

However, the main impact of the cut-off is that the differences between criteria are generally proven to be statistically insignificant. The right part of table 3 displays the DeLong's (1988) test statistics W_{AUC} and their p-values. Under the null, the areas under the ROC of both logit models are identical. These tests conclude to the rejection of the null only for 6 countries out of 12. In particular, for Argentina, Brazil, Peru, Philippines, Uruguay and Venezuela no gain in terms of sensitivity and specificity - no improvement in the type I and type II errors - results from the introduction of the spread in the EWS model. Indeed, this test leads to the conclusion that spread is important mostly for South-Asian countries. It appears that 5 out of the 6 countries for which this leading indicator improves the EWS belong to this cluster. For the South-American cluster, it seems that the yield spread does not significantly improve the crisis forecasts. Taking into account the cut-off in the evaluation leads hence to relativize the importance of spread in crisis forecasting. It is thus evident that this comparison of the areas under the ROC test allows us to better grasp the significance of the difference between the forecasting abilities of the two specifications. This evidence

empirically supports the use of AUC type criteria and tests, that are able to provide a more reliable diagnostic for any EWS.

Insert Table 3

5.4 Optimal cut-off

Previous results suggest that for some countries, *e.g.* Indonesia, Malaysia, Mexico, South Korea, Taiwan, and Thailand, the yield spread has a significant impact on the forecasting abilities of the EWS, whereas for others the influence of this variable is not significant. We now investigate the forecast accuracy of the outperforming model (with or without spread) for each country by identifying the optimal cut-off and calculating the associated percentage of correctly identified crisis (respectively calm) periods, *i.e. sensitivity* (respectively *specificity*). To emphasize the importance of the optimal cut-off in crisis forecasting, the time-series results are also presented in a similar vein to robustness check.

Insert Table 4

The right part of table 4 displays some descriptive statistics of the three cut-offs considered, *i.e.* credit-scoring, c^*_{CSA}, accuracy measures, c^*_{AM} and noise-to-signal ratio, c_{NSR} for the country-by country analysis. The major insight is that the NSR cut-off proposed by Kaminski et al. (1998) is always larger than the optimal cut-off we propose. The difference is significative, as the mean ratio is of 2 to 1 for the panel analysis and 3.3 to 1 for the country-by-country analysis. Besides, c_{NSR} is characterized by a larger dispersion relative to the other two cut-off. It results that the average forecast performance of the EWS model are quite different given the cut-off choice. The use of an optimal cut-off leads on average to a correct identification of at least 2/3 of the crisis and calm periods, as sensitivity exceeds 72% on average and specificity outruns 71%.

On the contrary, the NSR cut-off leads to a perfect identification of calm periods (*specificity* = 100%). However, this accuracy with respect to calm periods is possible only to the detriment of the crisis ones, since the average sensitivity is equal to 5.2%. In other words, the average type I error (false negative) is larger when using c_{NSR}, while the average type II error (false alarms) is lower compared to c^*_{CSA} and c^*_{AM}. Thus, the optimal cut-off are less sensitive to false alarms compared with missed crises. One rationale behind this could be that policy makers and enterprises are possibly willing to take a 'crisis insurance' and to accept a possible false alarm rather than be taken by surprise by a crisis, especially since the costs of a false alarm are thought to be inferior to those engendered by an unexpected crisis.

Insert Table 5

The performance of the optimal model at the country level confirms our previous findings i.e. the use of an optimal cut-off improves significantly currency crises forecasts (see table 5). Take the example of Argentina. [8] Using the NSR criteria leads to the correct identification of all the calm periods, but only 8.1% of the crisis periods. Assuming that a crisis occurs each ten years on average, it would take the model 123.46 years to correctly predict a currency crisis. At the same time, with our c^*_{CSA} cut-off the probability to correctly identify crises periods rises to 70.3% and the time necessary to correctly identify a crisis reduces to 14.2 years. This increase in sensitivity (and drop in misidentified crises) is possible only at the cost of more false alarms. However, if we admit that the cost of a false alarm is lower than the one of a misidentified crisis, the gain becomes evident, especially since a large increase in the proportion of crises correctly identified (from 8.1% to 70.3%) is associated with a smaller reduction of *specificity* (from 100% to 64.8%). An extreme case is that of Malaysia. To correctly identify all calm periods, the value of the NSR cut-off rises until reaching the value of 0.245, missing all the crisis periods (*sensitivity* = 0). By contrast, the optimal cut-offs correctly identify not only the crisis periods (*sensitivity* = 0.839) but also the calm ones (*specificity* = 0.810). The time-series results confirm this forecast gain associated with the computation of the optimal cut-off.

Insert Figure 5

Figure 5 depicts the cut-offs (z axis) against their associated *sensitivity* (x axis) and *specificity* (y axis) in a 3D scatterplot. The points on this figure correspond to the three types of cut-off estimated (CSA, NSR and AM) for the twelve countries in the sample. We observe that the values of optimal cut-off CSA and AM are relatively low (z axis) compared to the NSR ones. Besides, they are concentrated in a region with large *specificity* and *sensitivity*, indicating that these cut-offs correctly identify most of the crisis and calm periods. On the contrary, the NSR cut-off correctly identifies all calm periods at the expense of most of the crises. Graphically, these cut-off are clustered around specificity equal to one (y axis) and small values of sensitivity (x axis).

As mentioned in section 2.1, the optimal cut-off can also be analyzed in terms of vulnerability to crisis. Figures 6 and 7 depict the currency crisis probability series issued from the optimal EWS model as well as the optimal credit-scoring cut-off c^*_{CSA}. Notice that the crisis probabilities during the calm periods in the second half of the sample are as elevated as before, suggesting that the forecasting abilities of macroeconomic indicators have not been improved recently.

8. The results for the other countries are available upon request.

Insert Figures 6 and 7

We find that crisis probabilities during calm periods do not exhibit a downward trend, revealing a certain constant pressure in the exchange market. The highest pressure, as the optimal cut-off indicates, corresponds to Brazil, Peru, Philippines and Venezuela. Notice, however, that Brazil and Peru are characterized by a lower volatility than the other two countries during calm periods, while for Venezuela the model does not seem to perform too well. By contrast, the rest of the countries are characterized by a lower cut-off (around 0.2). Furthermore, the variance of the crisis probability series during observed crises is higher than the one characterizing calm periods in most countries. This could be explained by alternating moments of extreme vulnerability and short periods of recovery all the way through the crisis. Among others, the recent financial crisis has left two of the twelve countries in our analysis, namely Peru, and South Korea on the verge of a currency crisis (the $KLRm$ and Zhang dating methods identify both events). At the same time, we also identify risky periods in the recent years that are not considered as *crises* by the dating method. It is particularly the case of Indonesia, Thailand, and Venezuela, for which crisis probabilities soar, indicating balance-of-payment vulnerability. How comes that for these countries our dating method does not identify a currency event after 2007 ? Looking at the three indicators on which the pressure index relies, *i.e.* relative changes in exchange rate, relative changes in international reserves and absolute changes in interest rate, we ascertain the idea that in these countries the drop in reserves, the exchange rate depreciation and the rise in interest rate are incomparably lesser than those registered during previous currency crises, and they are not simultaneous. On the contrary, Peru knows one of its largest sudden drops in the growth of international reserves, whereas South Korean interest rates soar. It is thus clear that most of these emerging markets have become more resilient to currency crises over time, since in spite of their vulnerability, they do not face extreme movements in the balance of payment during the crisis period $2007 - 2009$.

5.5 Robustness Check

We now propose a sensitivity analysis of our results to the choice of crisis dating method. In a binary model (logit, probit, etc.), this choice impacts not only the evaluation results, but also the estimation of the parameters of the EWS. In the worst case, we could argue that it is useless to asses the EWS forecast with respect to a dating scheme that could be invalid. The aim of our paper is neither to provide a new dating methodology, nor to show that the $KLRm$ is the best dating procedure. We consider only that it is possible to identify the crisis and the calm periods according to a binary scheme. This identification can be the result of

economic experts' analyses or can be done through a pressure index approach. Whatever the methodology used, in this paper, we simply assume that it is possible *ex-post* to well identify the crisis. If this assumption is not satisfied, no EWS evaluation is possible whatever the model used (logit, markov, etc.).

Insert table 6

However, currency crises (and other crises) may not be precisely identified. That is why we recommend to conduct a robustness check of the evaluation procedure to the choice of the crisis dating method. Here, we propose to consider the $Zhang$ dating method instead of $KLRm$. The robustness check findings for our EWS evaluation are summarized in table 6. The QPS evaluation criterion indicates that the spread is important in forecasting currency crises for 10 out of 12 countries, and AUC confirms these findings. As for the comparison tests, CW generally supports the alternative hypothesis that the model with spread outperforms the one without this variable (7 out of 12 countries at the 5% significance level). Instead, the W_{AUC} test rejects the null hypothesis of equal forecasting abilities for 6 countries at the 5% level in favor of the model with yield spread. More precisely, for countries like Argentina, Indonesia, Malaysia, Peru, Philippines, South Korea, Taiwan, Thailand, Uruguay and Venezuela, the results obtained with the $Zhang$ dating method go along the lines of our previous analysis, based on the $KLRm$ dating method.

Nonetheless, changes in the dating method reflect in both the observed crisis series and the estimated crisis probabilities and thus in the relative comparison of the two models (with and without spread). More exactly, this time QPS and AUC favor the model without spread for three countries, *i.e.* Mexico, Peru and Philippines, while both comparison tests confirm this intuition.

Insert table 7

Moreover, the three cut-off considered (c^*_{CSA}, c^*_{AM} and c_{NSR}) have the same characteristics as in our previous analysis. Table 7 reports their descriptive statistics. For instance, the average NSR cut-off is at least twice the optimal ones, leading to the correct identification of all calm periods at the expense of most of the crisis ones, exactly as previously found. By contrast, the optimal cut-off lead to a better trade-off between type I and type II errors; by lowering the value of the cut-off, the number of crises correctly identified raises at a higher speed than the increase in false alarms. Our cut-off thus lead to an average correct identification of approximatively 2/3 of the crisis and calm periods.

6 Conclusion

In this paper, we propose an original, model-free, evaluation toolbox for EWS. This general approach not only assesses the validity of EWS forecasts, but also allows comparing the relative performance of alternative EWS. It is actually a two-step procedure combining the evaluation of the competing EWS and the comparison of their forecasting abilities. We show both theoretically and empirically that the cut-off has to be taken into account in EWS evaluation since existing QPS-type evaluation criteria often lead to diagnostic error. More importantly, we argue that the significance of the difference in evaluation criteria for two alternative models has to be tested in a statistical framework. To this aim, we introduce several comparison tests. Then, the optimal cut-off, the one that best discriminates between crisis and calm periods (by simultaneously minimizing type I and type II errors) is identified for the outperforming model. Therefore the cut-off appears as a key element in economic actors' decisions as it labels a country as vulnerable or not at a given moment. Additionally, we assert that the optimal cut-off is different from the NSR one, previously used in the literature, and on top of that, it leads to a better trade-off between the two types of errors.

Our new methodology has four main advantages. First, it is model-free (it can be applied to any EWS, independent of the underlying econometric model). Second, it can be used to assess EWS for any type of crises (currency, banking, debt, etc.). Moreover, it can be used not only for an in-sample evaluation, but also to assess the out-of-sample forecasts of an EWS. Besides, it covers both the selection of the outperforming model and the crisis forecast, thus proving to be extremely useful for researchers and economic actors as well.

Applying our evaluation toolbox to a sample of twelve emerging countries from 1980 to 2010, we show that the criteria and tests including the cut-off should be favored as they allow us to better refine the forecasting abilities of EWS. Indeed, the yield spread appears to be an important indicator of currency crises in half of the countries when we rely on the *Area under the ROC comparison test*, whereas it first seems to be essential in all the countries considered (when using QPS-based tests like $Clark\text{-}West$). Furthermore, the optimal cut-off correctly identifies on average more than 2/3 of the crisis and calm periods, in contrast with the NSR one, that correctly forecasts all the calm periods at the expense of most of the crisis ones.

Bibliography

1. Abiad, A., 2003, "Early Warning systems: A Survey and a Regime Switching Approach", IMF Working Paper.

2. Arias, G. and Erlandsson, U. G., 2005, "Improving early warning systems with Markov Switching model", C.E.F.I. Working Paper, 0502.
3. Barrios, S., Iversen, P., Lewandowska, M., Setzer, R., 2009, "Determinants of intra-euro area government bond spreads during the financial crisis", Economic Papers 388.
4. Basel Committee on Banking Supervision, 2005, "Studies on the Validation of Internal Rating Systems", working paper no.14, Bank for International Settlements.
5. Berg, A., and Pattillo, C., 1999, "Predicting Currency Crises: The Indicators Approach and an Alternative". *Journal of International Money and Finance*, Vol.18, pp. 561-586.
6. Berg, J.B., Candelon, B, and Urbain, J.P., 2008, "A Cautious Note on the Use of Panel Models to Predict Financial Crises", *Economics Letters*, Vol. 101, No. 1, pp. 80-83.
7. Berg, A., and Cooke, R., 2004, "Autocorrelation Corrected Standard Errors in Panel Probits: an Application to Currency Crisis Prediction", IMF Working Paper.
8. Bordo, M.D., Eichengreen, B., Klingebiel, D. and Martinez- Peria, S., 2001, "Is the crisis problem growing more severe ?", Economic Policy, Vol. 32, pp. 51 ?82.
9. Bussiere, M. and Fratzscher, M., 2006. "Towards a New Early Warning System of Financial Crises, Journal of International Money and Finance, Vol. 25, No. 6, pp. 953-973.
10. Clark, T. E., McCracken, M. W., 2001," Tests of Equal Forecast Accuracy and Encompassing for Nested Models", *Journal of Econometrics*, Vol. 105, No. 1, pp. 85-110.
11. Clark, T. E., West, K. D., 2007, "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models", *Journal of Econometrics*, Vol. 138, No. 1, pp. 291-311.
12. DeLong, E.R. , DeLong, D. M., Clarke-Pearson, D. L., 1988, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach", *Biometrics*, Vol. 44, No. 3, pp. 837-845.
13. Diebold, F. X., Mariano, S., 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*,Vol. 13, No. 3, pp. 253-63.
14. Diebold, F.X. and Rudebusch, G.D., 1989, "Scoring the Leading Indicators ", *The Journal of Business*, Vol. 62, No. 3, pp. 369-391.
15. Engelmann, B., Hayden, E. and Tasche, D., 2003, "Testing rating accuracy ?", *Risk* No. 16, pp. 82-86.
16. Estrella, A., and Hardouvelis, G. A., 1991, "The Term Structure as a Predictor of Real Economic Activity", *The Journal of Finance*, Vol. 46, No. 2, pp. 555-576.
17. Estrella, A., Mishkin, F.S., 1996, "The Yield Curve as a Predictor of US Recessions", *Current Issues in Economics and Finance*, Vol. 2, No. 7, pp. 1 - 5.

18. Estrella, A., and Trubin, M.R., 2006, "The Yield Curve as a Leading Indicator: Some Practical Issues", *Current Issues in Economics and Finance*, Vol. 12, No. 5, pp. 1 - 7.

19. Guidolin, M., Tam, Y. M., 2010, "A Yield Spread Perspective on the Great Financial Crisis: Break-Point Test Evidence", Federal Reserve Bank of St. Louis Working Paper No. 26.

20. Fratzscher,M., 2003, "On Currency Crises and Contagion", International Journal of Finance and Economics, Vol. 8, No. 2, pp. 109-129.

21. Fuertes, A.-M., Kalotychou, E., 2007, "Optimal Design of Early Warning Systems for Sovereign Debt Crises", International Journal of Forecasting, Vol. 23, No. 1, pp. 85-100.

22. Harding, D., Pagan, A., 2006, "The Econometric Analysis of Constructed Binary Time Series", Working Papers Series 963, The University of Melbourne.

23. Hoeffding, W. 1948, "A Class of Statistics with Asymptotically Normal Distributions", *Annals of Statistics*, Vol. 19, pp. 293-325.

24. Jacobs, J.P.A.M., Kuper, G.H., and Lestano, 2004, "Financial Crisis Identification: a survey", working paper, University of Groningen.

25. Jacobs, J.P.A.M., Kuper, G.H., and Lestano, 2008 "Currency crises in Asia : A multivariate logit approach", in *International Finance Review Asia Pacific Financial Markets*: *Integration, Innovation and Challenges Vol. 8, pp. 157-173.*, ed by S.J. Kim and M. McKenzie.

26. Kaminsky, G., Lizondo, S., Reinhart, C., 1998, "Leading Indicators of Currency Crises", *IMF Staff Papers*, Vol. 45, No. 1, pp. 1-48.

27. Kaminsky, G.L., 2003, "Varieties of Currency Crises", NBER Working Paper No. 10193.

28. Kapetanios, G., 2003, "Determining the Poolability of Individual Series in Panel Datasets", Working Paper 499.

29. Kraft H., Kroisandt G. and Muller M., 2004, "Redesigning Ratings: Assessing the Discriminatory Power of Credit Scores under Censoring", working paper.

30. Kumar, M., Moorthy, U. and Perraudin, W., 2003, "Predicting Emerging Market Currency Crashes", Journal of Empirical Finance, Vol. 10, pp. 427-454.

31. Kydland, F. E., and Prescott, E. C., 1991, "The Econometrics of the General Equilibrium Approach to Business Cycles", *Scandinavian Journal of Economics*, Vol. 93, No. 2, pp. 161-78.

32. Lambert, J. and Lipkovich, I., 2008, "A Macro for Getting more out of your ROC Curve", SAS Global forum, paper 231.

33. Lau, F., Yung, S., and Yong, I., 2003, "Introducing a Framework to Measure Resilience of an Economy", Hong Kong Monetary Authority Quarterly Bulletin.
34. Lestano, Jacobs, J., and Kuper, G. H., 2003, "Indicators of financial crises do work! An early-warning system for six Asian countries," Working Paper.
35. Martinez Peria, M.S., 2002, "A regime-switching approach to the study of speculative attacks: A focus on EMS crises", Empirical Economics, Vol. 27, No. 2, pp. 299-334.
36. Mitchener, K. J., Weidenmier,M. D., 2006, "The Baring Crisis and the Great Latin American Meltdown of the 1890s", NBER Working Paper No. 13403.
37. Renault, O., and De Servigny, A., 2004, *The Standard & Poor's Guide to Measuring and Managing Credit Risk,* 1^{st} ed. McGraw-Hill, 2004.
38. Stein, R. M., 2005, "The relationship between default prediction and lending profits: integrating ROC analysis and loan pricing.", *Journal of Banking & Finance*, Vol. 29, pp. 1213-1236.
39. Williams, R., 2004, "A Note on Robust Variance Estimation for Cluster-Correlated Data", *Biometrics*, Vol. 56, No. 2, pp. 645 - 646.
40. Wright, J., 2006, "The Yield Curve and Predicting Recessions," Finance and Economic Discussion Series No. 7, Federal Reserve Board.
41. Zhang, Z., 2001, "Speculative attacks in the Asian crisis", IMF Working Paper 189, International Monetary Fund, Washington, DC.

Appendix 1. Comparison of *ROC* Curves Test

The non-parametric *test of comparison of ROC curves* has been proposed by DeLong et al. (1988). It is based on the comparison of the areas under the ROC curves associated with the two EWS models, denoted AUC_1 and AUC_2. The null of the tests corresponds to the equality of areas under the *ROC* curves *i.e.*, $H_0 : AUC_1 = AUC_2$. The test statistic is defined as:

$$W_{AUC} = \frac{(AUC_1 - AUC_2)^2}{\mathbb{V}(AUC_1 - AUC_2)} \tag{19}$$

Under the null, it has an asymptotic $\chi^2(1)$ distribution. By definition the asymptotic variance of the difference $\mathbb{V}(AUC_1 - AUC_2)$ is equal to:

$$\mathbb{V}(AUC_1 - AUC_2) = \mathbb{V}(AUC_1) + \mathbb{V}(AUC_2) - 2cov\left(AUC_1, AUC_2\right) \tag{20}$$

Each of these three elements can be estimated using a non parametric kernel estimator. Let us consider $\mathbb{V}$ the variance covariance matrix of the vector $(AUC_1 \ AUC_2)'$. A non parametric

kernel estimator of $\mathbb{V}$, denoted $\widehat{\mathbb{V}}$, can be derived from the theory developed for generalized U-statistics by Hoeffding (1948) and Mann-Whitney statistics. Formally, we have:

$$\underset{(2,2)}{\widehat{\mathbb{V}}} = (T_1)^{-1} \underset{(2,2)}{\widehat{S}_1} + (T_0)^{-1} \underset{(2,2)}{\widehat{S}_0} \tag{21}$$

where T_1 (respectively T_0) is the number of crisis (respectively calm) periods in the sample, and $\widehat{S}_1$ (respectively $\widehat{S}_0$) denotes the estimated variance for the crisis (respectively calm) periods.

$$\begin{aligned}
\widehat{S}_1 \;=\; & \frac{1}{T_0^2\,(T_1-1)} \sum_{i:y_i=1} \left(\begin{array}{c} \left[\sum_{j:y_j=0} K(\hat{p}_j,\hat{p}_i) - T_0 \times AUC_1 \right]^2 \\ \left[\sum_{j:y_j=0} K(\hat{p}_j,\hat{p}_i) - T_0 \times AUC_1 \right] \times \left[\sum_{j:y_j=0} K(\hat{p}_j,\hat{p}_i) - T_0 \times AUC_2 \right] \\ \left[\sum_{j:y_j=0} K(\hat{p}_j,\hat{p}_i) - T_0 \times AUC_1 \right] \times \left[\sum_{j:y_j=0} K(\hat{p}_j,\hat{p}_i) - T_0 \times AUC_2 \right] \\ \left[\sum_{j:y_j=0} K(\hat{p}_j,\hat{p}_i) - T_0 \times AUC_2 \right]^2 \end{array} \right)
\end{aligned} \tag{22}$$

Similarly, we have:

$$\begin{aligned}
\widehat{S}_0 \;=\; & \frac{1}{T_1^2\,(T_0-1)} \sum_{j:y_j=0} \left(\begin{array}{c} \left[\sum_{i:y_i=1} K(\hat{p}_j,\hat{p}_i) - T_1 \times AUC_1 \right]^2 \\ \left[\sum_{i:y_i=1} K(\hat{p}_j,\hat{p}_i) - T_1 \times AUC_1 \right] \times \left[\sum_{i:y_i=1} K(\hat{p}_j,\hat{p}_i) - T_1 \times AUC_2 \right] \\ \left[\sum_{i:y_i=1} K(\hat{p}_j,\hat{p}_i) - T_1 \times AUC_1 \right] \times \left[\sum_{i:y_i=1} K(\hat{p}_j,\hat{p}_i) - T_1 \times AUC_2 \right] \\ \left[\sum_{i:y_i=1} K(\hat{p}_j,\hat{p}_i) - T_1 \times AUC_2 \right]^2 \end{array} \right)
\end{aligned} \tag{23}$$

where $K(.)$ denotes a kernel function of the estimated crisis probabilities in crisis periods ($y_i = 1$) and calm periods ($y_j = 0$) defined by:

$$K(\hat{p}_j,\hat{p}_i) = \begin{cases} 1, & \text{if } \hat{p}_i < \hat{p}_j \\ \frac{1}{2}, & \text{if } \hat{p}_i = \hat{p}_j \\ 0, & \text{if } \hat{p}_i > \hat{p}_j. \end{cases} \tag{24}$$

Appendix 2: Dataset

There is no official currency crisis dating method similar to the one NBER proposes for recessions. Therefore, a crisis episode is generally detected when an index of speculative pressure exceeds a certain threshold. Many alternative indexes have been developed and used for identifying currency crises. But they are all non-parametric termination rules that take into consideration the size of the movements in a combination of a number of series. Lestano and Jacobs, (2004) compare several currency crisis dating methods, aiming to identify the one that recognizes most of the crises categorized by the IMF for the 1997 Asian flu. They conclude that the KLR modified index, the Zhang original index (Zhang, 2001), and extreme values applied to the KLR modified index perform best.

Following their results, we identify crisis periods using the KLR modified pressure index (KLRm), which, unlike the KLR index, also includes interest rates:

$$\text{KLRm}_{it} = \frac{\Delta e_{it}}{e_{it}} - \frac{\sigma_e}{\sigma_r}\frac{\Delta r_{it}}{r_{it}} + \frac{\sigma_e}{\sigma_{ir}}\Delta ir_{it}, \tag{25}$$

where e_{it} denotes the exchange rate (*i.e.,* units of country i's currency per US dollar in period t), $r_{n,t}$ represents the foreign reserves, while ir_{it} is the interest rate. Meanwhile, the standard deviations σ_X are actually the standard deviations of the relative changes in the variables, $\sigma_{(\Delta X_{it}/X_{it})}$, where X denotes each variable separately, including the exchange rate and the foreign reserves, with $\Delta X_{it} = X_{it} - X_{i,t-6}$. For the interest rate, σ_{ir} is the standard deviation of the absolute changes in interest rate. For both subsamples, the threshold equals two standard deviations above the mean: [9]

$$Crisis_{it} = \begin{cases} 1, & \text{if } \text{KLRm}_{it} > 2\sigma_{\text{KLRm}_{it}} + \mu_{\text{KLRm}_{it}} \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

To check the robustness of our results to the dating method, we also consider the *Zhang* pressure index instead of the $KLRm$. It is defined as follows:

$$Crisis_{it} = \begin{cases} 1, & \text{if } \begin{cases} \frac{\Delta e_{it}}{e_{it}} > \beta_1 \sigma'_{e_{it}} + \mu_{e_{it}} & \text{or} \\ \frac{\Delta r_{n,t}}{r_{it}} < \beta_2 \sigma'_{r_{it}} + \mu_{r_{it}} \end{cases} \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

9. In the case of KLR the threshold equals three standard deviations; however, in this case, Taiwan would never register any currency crises, which is historically not accurate. For example, Taiwan was not exempted from the Asian crisis in 1997.

where $\sigma'_{e_{it}}$ is the standard deviation of $(\Delta e_{it}/e_{it})$ in the sample of (t-36, t-1), and $\sigma'_{r_{it}}$ is the standard deviation of $(\Delta r_{it}/r_{it})$ in the sample of (t-36, t-1). The thresholds are set to $\beta_1 = 3$ and $\beta_2 = -3$. Contrary to the KLRm index, the interest rates are excluded from the ZCC and the thresholds used are time-varying for each component.

From a macroeconomic point of view, it is more important to know if there will be a crisis in a certain horizon than in a certain month, because this time period allows the state to take steps to prevent the crisis. Consequently, we define for each country $C24_t$, which corresponds to y_t from our general framework and thus serves as the crisis dummy variable taking the value of 1 if there will be a crisis in the following 24 months and 0 otherwise:

$$C24_{it} = \begin{cases} 1, & \text{if } \sum_{j=1}^{24} Crisis_{it+j} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

At the same time, several explanatory variables from three economic sectors are considered (Lestano et al., 2003) on a monthly frequency and denoted in US dollars:
1. External sector: the one-year growth rate of international reserves, the one-year growth rate of imports, the one-year growth rate of exports, the ratio of M2 to foreign reserves, and the one-year growth rate of M2 to foreign reserves.
2. Financial sector: the one-year growth rate of M2 multiplier, the one-year growth rate of domestic credit over GDP, the one-year growth rate of real bank deposits, the real interest rate, the lending rate over deposit rate, and the real interest rate differential.
3. Domestic real and public sector: the industrial production index.

As in Kumar, (2003), we reduce the impact of extreme values by using the formula: $f(x_t) = sign(x_t) \times ln(1 + |x_t|)$. Traditional first generation (Im, Pesaran, Shin, 1997 and Maddala and Wu, 1999) and second generation (Bai and Ng, 2000 and Pesaran, 2003) panel unit root tests are performed, leading to the rejection of the null hypothesis of stochastic trend except for the lending rate over deposit rate and industrial production index indicators. Hence, these series are substituted by their first differences.

Finally, we identify the most correlated leading indicators for each country. Two indicators are considered as being correlated for a certain country if Pearson's correlation coefficient is higher than a 30% threshold. It seems that growth of real exchange rate and real interest rate are highly correlated with most indicators for all countries, whereas the first difference of lending rate over deposit rate, the first difference of the industrial production index and yield spread are the least correlated ones with all the other indicators for all the 12 countries. The competing models are defined such that no couple of indicators is correlated in more than 4 countries. We hence identify the leading indicators by minimizing the AIC and BIC

information criteria of the pooled panel data models, *i.e.,* growth of international reserves, growth of exports, growth of domestic credit over GDP, first difference of lending over deposit rate, first difference of industrial production index and yield spread. The missing values through the series are replaced using cubic splines interpolation, but when the series revealed missing values at the beginning of the sample, such as "the one-year growth of terms of trade" or "yield spread", the corresponding observations are dropped from the analysis, leading to an unbalanced panel framework. Table 8 shows the period covered by the leading indicators for each of the 12 countries.

Insert Table 8

Appendix 3: A Robust Estimator of the Variance of the Parameters

To compute robust estimators of the variance for logit models we use a sandwich estimator. Technically, variance-covariance matrix of the estimators is asymptotically equal to the inverse of the hessian matrix: $\mathbb{V}(\hat{\beta}) = -H(\hat{\beta})^{-1}$. However, this is appropriate only if we employ the real Data Generating Process (DGP). For a more permissive method from this point of view, we define the variance vector as follows:

$$\mathbb{V}(\hat{\beta}) = (-H(\hat{\beta})^{-1})\mathbb{V}(g(\hat{\beta}))(-H(\hat{\beta})^{-1}), \tag{29}$$

where $H(\hat{\beta})^{-1}$ is the inverse of the hessian matrix, and $\mathbb{V}(g(\hat{\beta}))$ is the variance of the gradient. Using the empirical variance estimator of the gradient we find that:

$$\mathbb{V}(\hat{\beta}) = -T/(T-1)H(\hat{\beta})^{-1}\{\sum_{t=1}^{T} g_t(\hat{\beta})g_t(\hat{\beta})'\}(-H(\hat{\beta})^{-1}), \tag{30}$$

which is a robust variance estimator for the time-series model.

The main advantage of this sandwich method is that it can also be applied in the case of grouped data, as in our case. It is important to note that in the current situation, each country from a cluster is a group of time-series observations that are correlated. Thus, the observations corresponding to a country are not treated as independent, but rather the countries themselves which form the clusters, are considered independent. Therefore, instead of using $g_t(\hat{\beta})$, we use the sum of $g_t(\hat{\beta})$ for each country, while T is replaced by the number of countries in a cluster. These changes ensure the independence of so-called "superobservations" entering the formula (Gould et al., 2005).

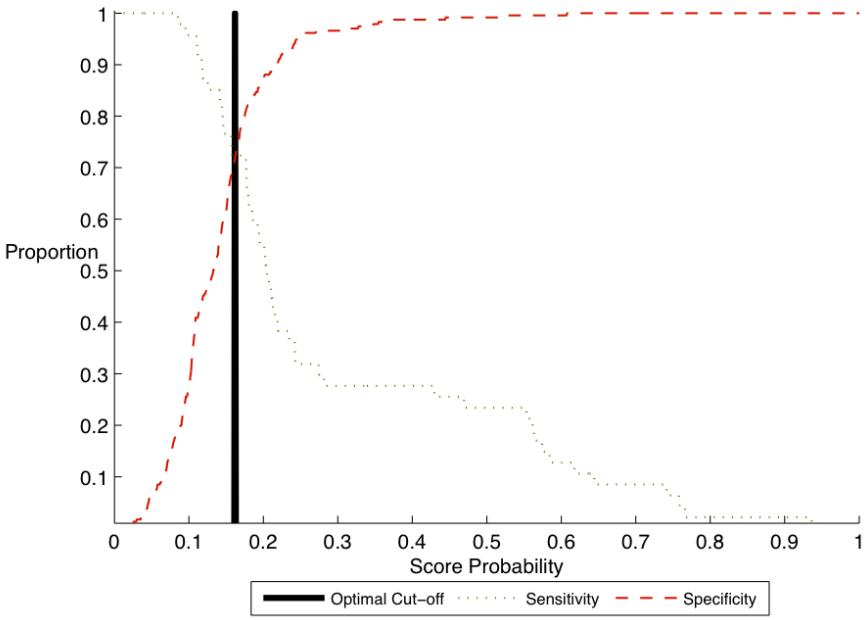

FIGURE 1 – Optimal Cut-off determination

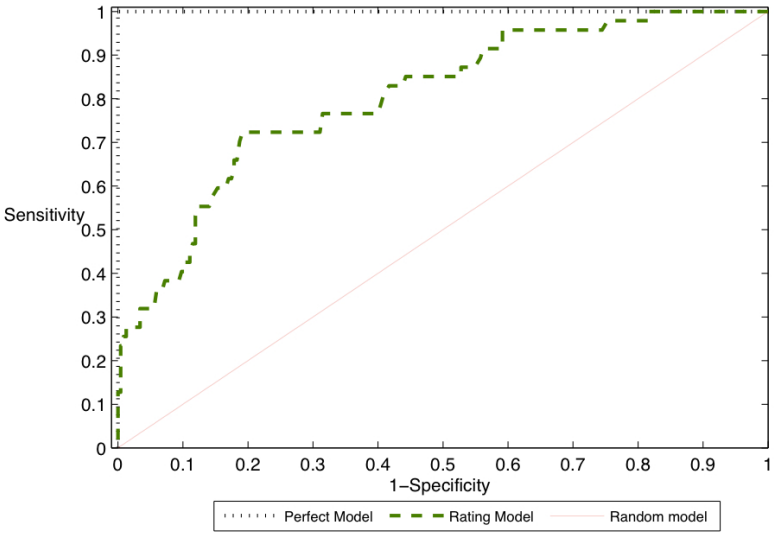

FIGURE 2 – The ROC curve

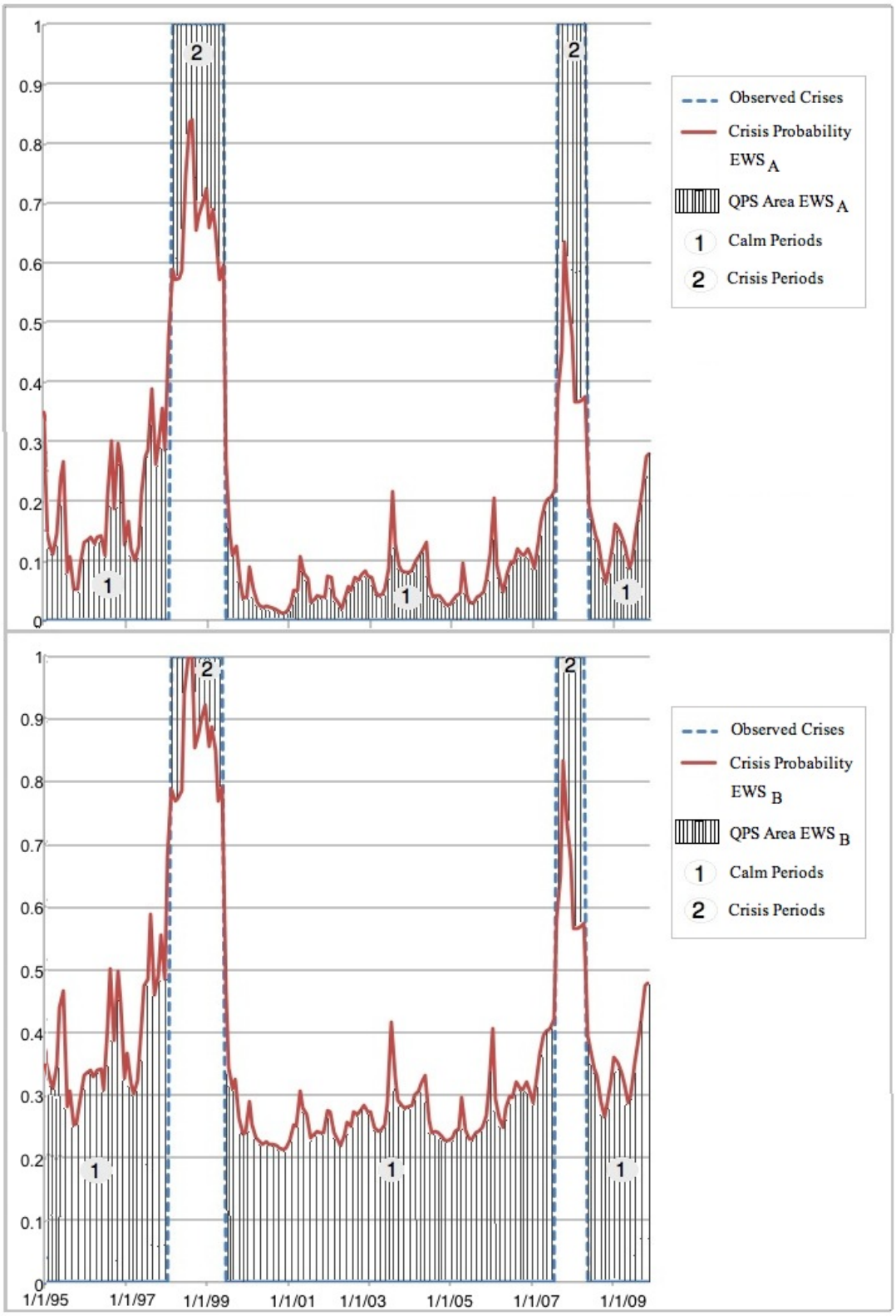

FIGURE 3 – QPS - Graphical Approach

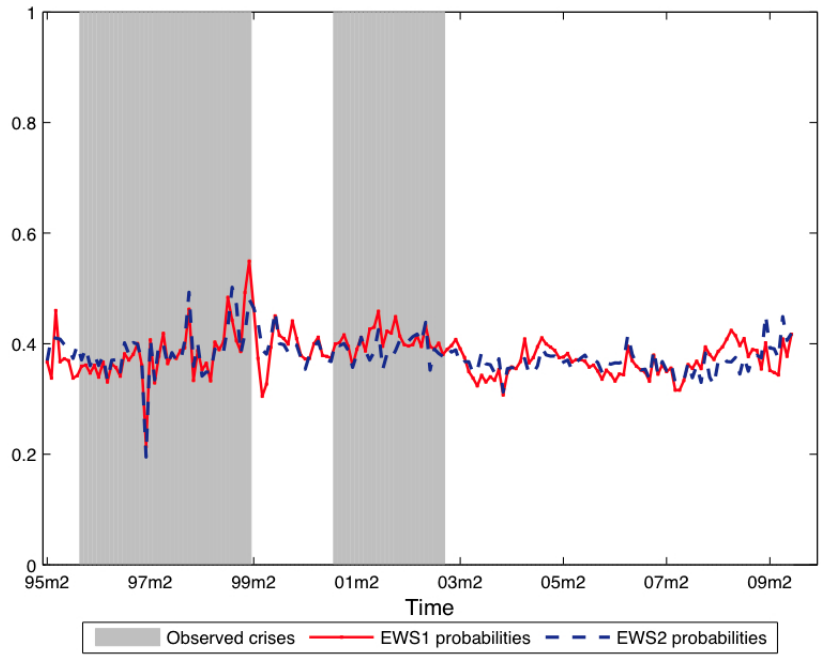

FIGURE 4 – Brazil - Crisis probabilities

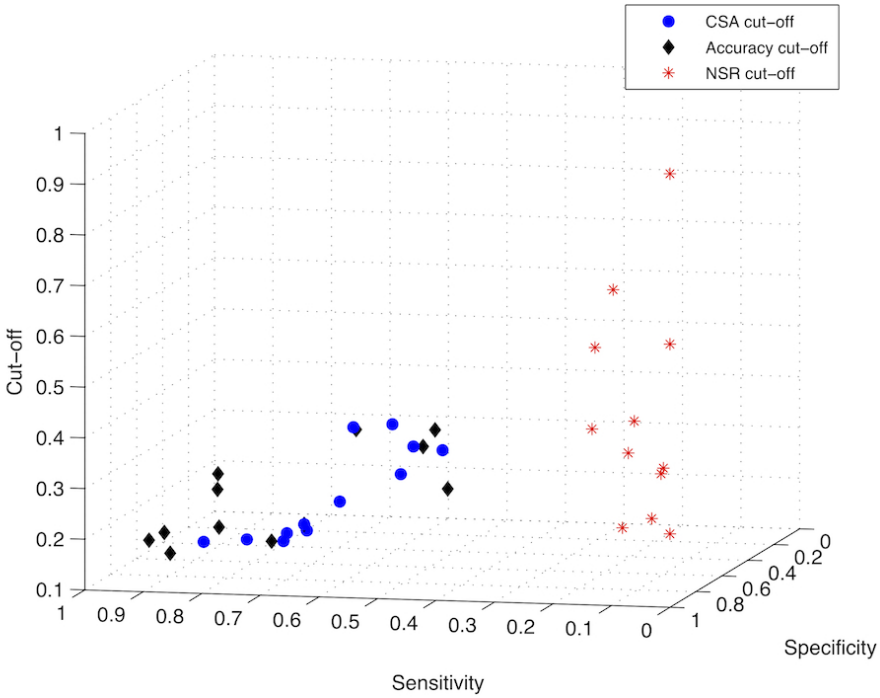

FIGURE 5 – Cut-off - Regional Panel Models

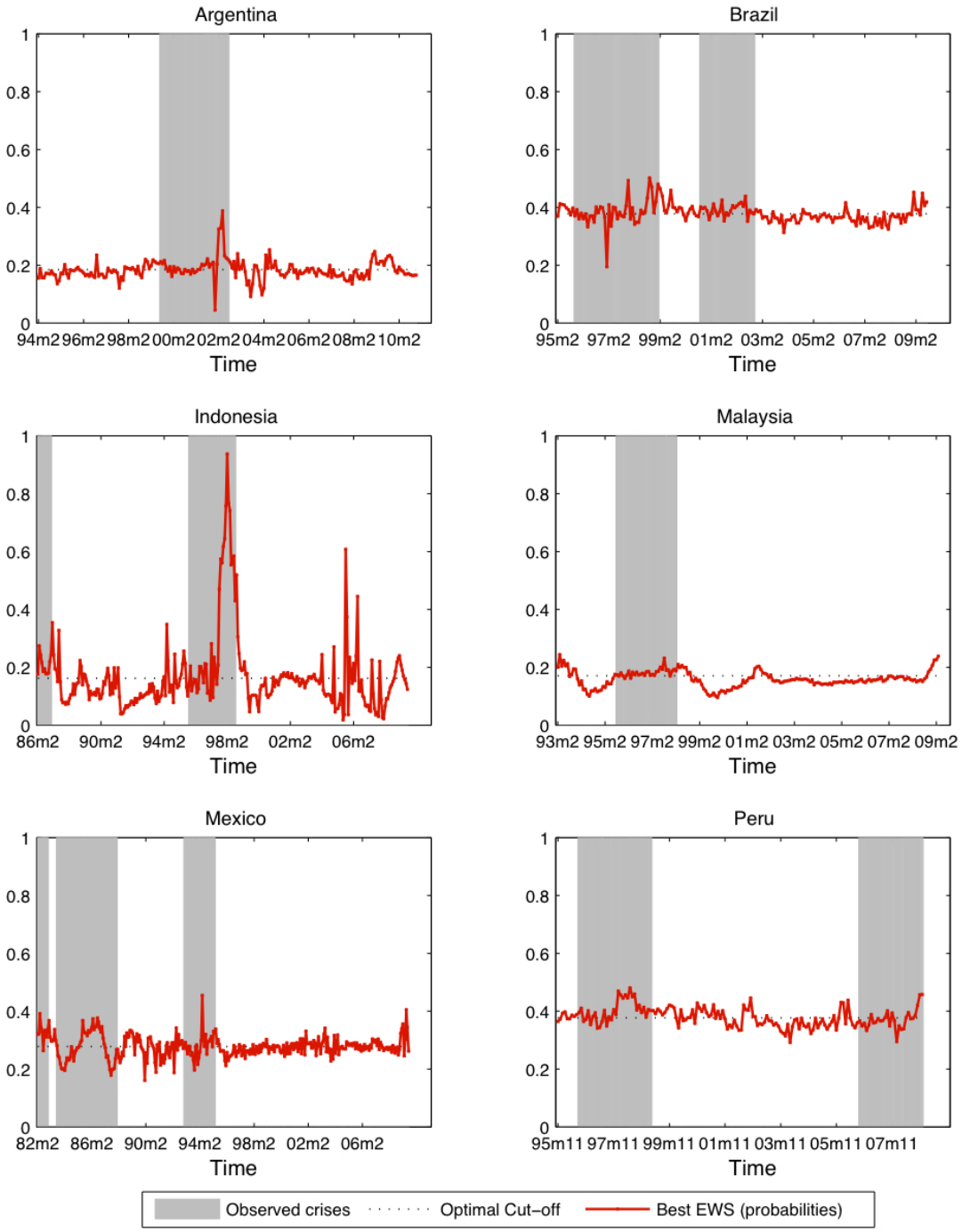

FIGURE 6 – Crisis Probabilities - Time-Series Models

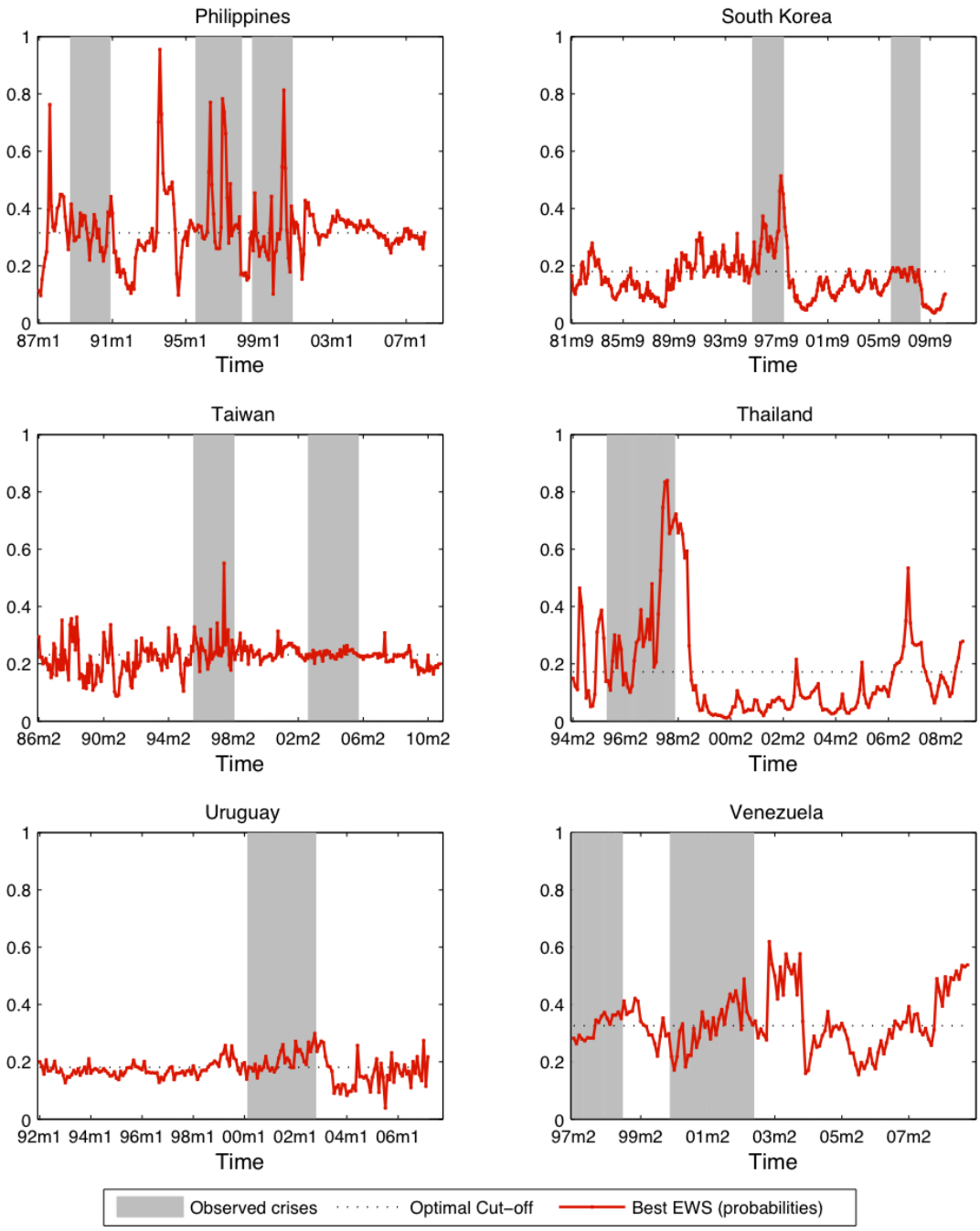

FIGURE 7 – Crisis Probabilities - Regional Panel Models (continued)

TABLE 1 – Example: Evaluation Criteria

Country	Model	QPS	AUC		CW	ROC
Brazil	EWS1	0.454030	0.639801	Statistic	5.537***	0
	EWS2	0.534030	0.639801	P-value	< 0.001	1
Indonesia	EWS1	0.231282	0.807832	Statistic	9.876***	0
	EWS2	0.311282	0.807832	P-value	< 0.001	1

Note: Two EWS with equal forecasting performance are compared by using evaluation criteria (QPS and AUC) as well as comparison tests (CW and ROC). The smaller the QPS the better the model; the larger the AUC, the better the model. The null hypothesis of both test is the equality of forecasting abilities of the two models. The alternative indicates that the non-constraint model (EWS1) is better than the other one. The asterisks *,**, and *** denote significance at the 90%, 95% and 99% level, respectively.

TABLE 2 – EWS Estimation

	Panel Model			Time-series Model	
	Pooled	Regional 1	Regional 2	Significance	
		(South America)	(South Asia)	at 5% level	at 1% level
Growth of international reserves	−1.438 (−0.500)	−0.800 (−0.210)	−3.793 (−0.770)	5	4
Growth of exports	−2.872 (−0.550)	−3.088 (−0.510)	−4.677 (−0.650)	7	4
Growth of domestic credit over GDP	1.342** (2.700)	−0.268 (−0.210)	1.634*** (3.590)	6	3
Lending rate over deposit rate (first difference)	−0.005 (−0.130)	−2.451 (−1.610)	0.025 (0.710)	1	
Growth of industrial production (first difference)	−0.211 (−0.600)	−0.328 (−0.480)	−0.068 (−0.110)		
Yield spread	−1.522** (−2.630)	−1.049** (−3.170)	−2.404* (−1.740)	11	5
Constant				7	6

Note: The table presents the estimation results for the pooled panel model, regional panel model and country-by-country (time-series) models. The figures between parentheses are t-statistics. The asterisks *,**, and *** denote significance at the 90%, 95% and 99% level, respectively. For the time-series models we present the number of countries for which a specific variable is significative at a given risk level.

TABLE 3 – EWS Evaluation: Regional Panel Model

	Evaluation Criteria			Comparison Tests	
	Model	QPS	AUC	CW test	W_{AUC} test
Argentina	with spread	0.276*	0.683	2.819** (0.005)	0.175 (0.676)
	without spread	0.288	0.705*		
Brazil	with spread	0.456*	0.659*	1.872* (0.061)	0.180 (0.672)
	without spread	0.461	0.646		
Indonesia	with spread	0.219*	0.803*	4.115 *** (<0.001)	17.02 *** (<0.0001)
	without spread	0.262	0.616		
Malaysia	with spread	0.256*	0.839*	1.792* (0.073)	5.645** (0.018)
	without spread	0.259	0.765		
Mexico	with spread	0.391*	0.611*	3.236 *** (<0.001)	4.970** (0.026)
	without spread	0.400	0.537		
Peru	with spread	0.467	0.559	0.510 (0.610)	0.670 (0.413)
	without spread	0.459*	0.601*		
Philippines	with spread	0.496	0.477	−2.056** (0.040)	2.379 (0.123)
	without spread	0.439*	0.537*		
South Korea	with spread	0.229*	0.829*	5.918 *** (<0.001)	15.51 *** (<0.001)
	without spread	0.260	0.691		
Taiwan	with spread	0.333*	0.687*	4.529 *** (<0.001)	34.09 *** (<0.001)
	without spread	0.351	0.443		
Thailand	with spread	0.230*	0.849*	4.879 *** (<0.001)	25.56 *** (<0.001)
	without spread	0.285	0.648		
Uruguay	with spread	0.223*	0.816*	3.367 *** (<0.001)	0.395 (0.530)
	without spread	0.270	0.786		
Venezuela	with spread	0.486*	0.544*	1.686∗ (0.092)	2.53 (0.112)
	without spread	0.488	0.471		

Note: QPS ranges from 0 to 2, the lower its level, the better the model. The AUC criteria takes values between 0.5 and 1, 1 being the perfect model. The best model according to each evaluation criteria is denoted by an asterisk (*). The null hypothesis of the comparison tests is the equality of predictive performance of the two models. The alternative of the Clark-West, CW, and Diebold-Mariano test, DM, is the statistical difference between the two criteria (it indicates that the model with the smaller QPS is better than the other one), while the alternative hypothesis of the DeLong, W_{AUC}, test is the statistical difference between the two areas (the model with a larger AUC is better). The asterisks *,**, and *** denote test significance at the 90%, 95% and 99% level, respectively.

TABLE 4 – EWS Optimal Cut-off: Descriptive Statistics

	Time-Series			Regional Panel		
Optimal Cutoff	c^*_{CSA}	c^*_{AM}	c_{NSR}	c^*_{CSA}	c^*_{AM}	c_{NSR}
Average	0.245	0.239	0.815	0.246	0.233	0.477
Std-deviation	0.079	0.117	0.130	0.084	0.092	0.213
Minimum	0.147	0.115	0.567	0.162	0.118	0.245
Maximum	0.371	0.437	0.993	0.378	0.388	0.955
Average Sensitivity	0.72	0.819	0.048	0.656	0.745	0.052
Average Specificity	0.72	0.711	1.000	0.654	0.635	1.000

Note: This table includes some descriptive statistics for the cut-offs. We select the optimal cut-off for the best model by using two methods (credit-scoring - CSA-, and accuracy measures - AM-). For comparison reasons we also present KLR's NSR cut-off. The corresponding average levels of correctly identified crisis and calm periods are also included.

TABLE 5 – EWS Forecasting abilities

		Time-series			Regional Panel		
		cut-off	sensit	specif	cut-off	sensit	specif
Argentina	c^*_{CSA}	0.151	0.703	0.703	0.185	0.703	0.648
(without spread)	c^*_{AM}	0.120	0.892	0.624	0.185	0.703	0.648
	c_{NSR}	0.782	0.108	1.000	0.254	0.081	1.000
Malaysia	c^*_{CSA}	0.293	0.935	0.939	0.171	0.839	0.810
(mixt)	c^*_{AM}	0.293	0.935	0.939	0.169	0.935	0.798
	c_{NSR}	0.993	0.000	1.000	0.245	0.000	1.000

Note: The optimal model, as resulting from table 3, is chosen for each country. We select the optimal cut-off by using two methods (credit-scoring - CSA-, and accuracy measures - AM-). The corresponding proportion of correctly identified crisis (calm) periods, denoted $sensit$, ($specif$) are also presented. Note that type I and type II errors can be obtained as their complement. For comparison reasons, we also present the NSR cut-off.

TABLE 6 – EWS Evaluation: Regional Panel Models (Robustness check)

	Evaluation Criteria			Comparison Tests	
	Model	QPS	AUC	CW test	W_{AUC} test
Argentina	with spread	0.487*	0.564*	2.221** (0.026)	1.126 (0.289)
	without spread	0.500	0.508		
Brazil	with spread	0.485*	0.677*	4.503 *** (<0.001)	13.83 *** (<0.001)
	without spread	0.496	0.547		
Indonesia	with spread	0.361	0.565*	1.204 (0.228)	4.687** (0.030)
	without spread	0.356*	0.455		
Malaysia	with spread	0.419*	0.722*	5.132 *** (<0.001)	11.74 *** (<0.001)
	without spread	0.433	0.610		
Mexico	with spread	0.339	0.438	−3.644*** (<0.001)	15.03 *** (<0.001)
	without spread	0.334*	0.516*		
Peru	with spread	0.396	0.382	−2.770** (0.006)	19.21 *** (<0.001)
	without spread	0.375*	0.646*		
Philippines	with spread	0.334	0.540	−1.984** (0.047)	16.69 *** (<0.001)
	without spread	0.308*	0.692*		
South Korea	with spread	0.271*	0.759*	5.642 ** (<0.001)	21.13 *** (<0.001)
	without spread	0.289	0.565		
Taiwan	with spread	0.170*	0.669*	2.137** (0.033)	5.676*** (0.017)
	without spread	0.174	0.541		
Thailand	with spread	0.283*	0.895*	7.929 *** (<0.001)	45.47 *** (<0.001)
	without spread	0.396	0.559		
Uruguay	with spread	0.280*	0.708*	3.137 *** (<0.001)	2.126 (0.145)
	without spread	0.317	0.604		
Venezuela	with spread	0.386	0.540*	0.734 (0.463)	0.036 (0.849)
	without spread	0.385*	0.528		

Note: See note to table 3.

TABLE 7 – EWS Optimal Cut-off: Descriptive Statistics (Robustness check)

	Time-Series			Regional Panel		
Optimal Cutoff	c^*_{CSA}	c^*_{AM}	c_{NSR}	c^*_{CSA}	c^*_{AM}	c_{NSR}
Average	0.269	0.262	0.805	0.271	0.275	0.514
Std-deviation	0.115	0.144	0.120	0.119	0.123	0.209
Minimum	0.105	0.087	0.638	0.101	0.098	0.157
Maximum	0.501	0.496	0.957	0.498	0.514	0.848
Average Sensitivity	0.733	0.834	0.112	0.603	0.611	0.028
Average Specificity	0.732	0.694	1.000	0.606	0.689	1.000

Note: See note to table 4.

TABLE 8 – Database

Country	Period
Argentina	February 1994 - December 2010
Brazil	February 1995 - August 2009
Indonesia	February 1986 - August 2009
Malaysia	February 1993 - April 2009
Mexico	February 1982 - August 2009
Peru	November 1995 - January 2009
Philippines	January 1987 - February 2008
South Korea	September 1981 - December 2010
Taiwan	February 1986 - December 2010
Thailand	February 1994 - January 2009
Uruguay	January 1992 - April 2007
Venezuela	February 1997 - December 2008

Note: Data availability.