Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 23 (2007) 85–100
international journal of forecasting
www.elsevier.com/locate/ijforecast

Optimal design of early warning systems for sovereign debt crises

Ana-Maria Fuertes *, Elena Kalotychou

Faculty of Finance, Cass Business School, City University, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom

Abstract

This paper tackles the design of an optimal early warning system (EWS) for sovereign default from two distinct angles: the choice of the econometric methodology and the evaluation of the EWS itself. It compares K-means clustering of macrodata, a logit regression for macrodata, a logit regression for credit ratings, and the combined forecasts from all three methods. The optimal choice of forecast method is shown to depend on the desired trade-off between missed defaults and false alarms. Hence, it is crucial to account for the decision-maker's preferences which are characterized through a loss function and risk-aversion parameter. Recursive forecast combining generally yields a better balance of type I and type II errors than any of the individual forecasting methods, and outperforms the naïve predictions.

JEL classification: C15; C22; C52

Keywords: Country risk analysis; Clustering; Default prediction; Emerging markets; Forecast combining; Logit forecast; Loss function

1. Introduction

The financial turmoil that hit emerging markets in recent decades has triggered the need for accurate country risk assessment. A number of studies have focused on the development of empirical models for explaining and predicting banking and currency crises (Berg & Pattillo, 1999; Frankel & Rose, 1996; Kaminsky & Reinhart, 1999; Kumar, Moorthy, & Perraudin, 2003). As more countries move toward flexible exchange rates, twin crises are becoming less frequent. But sovereign debt crises remain a matter of concern for international financial markets and economic policymakers.

The process of building an Early Warning System (EWS) can be broadly divided into four decision stages: the sample (country and time span), the input variables, the econometric approach, and the evaluation of the EWS in relation to its end use by the decision-maker. The first two have by now received extensive attention in the sovereign default literature. Most studies have focused on identifying the nature–region, country, or period specific–of debt crises, or their main determinants among domestic fundamentals and indicators of the international business-cycle and market sentiment. For this purpose, different classification techniques have been used. However, the

* Corresponding author. Tel.: +44 20 7040 0186; fax: +44 20 7040 8881.
E-mail addresses: a.fuertes@city.ac.uk (A.-M. Fuertes), e.kalotychou@city.ac.uk (E. Kalotychou).

0169-2070/$ - see front matter © 2006 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.ijforecast.2006.07.001

empirical literature on EWSs with an explicit forecasting objective is relatively young. Scant attention has been paid to forecasting issues and to the design and validation of an EWS tailored to the decision-maker's preferences. The aim of this paper is to contribute to filling this gap.

Several studies have applied discriminant analysis (Frank & Cline, 1971; Taffler & Abassi, 1984), whereas more recent research has been based on panel logit models (Peter, 2002). Non-parametric classification techniques such as clustering and recursive tree analysis, albeit popular in other areas, have received little attention in this context. There is evidence that country credit ratings have predictive power regarding sovereign debt crises and that they Granger-cause sovereign bond spreads (Cantor & Packer, 1996; Reinhart, 2001; Rojas-Suárez, 2001). Moreover, the New Basel Accord allows banks to use internal ratings for calculating capital requirements. The *Institutional Investor* ratings can be regarded as consensus internal ratings from major international banks. The upshot is that it is unclear which method and information set one should adopt in developing an EWS for sovereign default. In this respect, forecast combining may be fruitful.

This paper presents a novel framework for the optimal design of an EWS focusing on methodological issues. The contribution is twofold. First, it assesses alternative forecasting techniques in the light of the decision-maker's degree of risk-aversion towards default. These are: (i) a multivariate logit model based on macrodata, (ii) a univariate logit model based on the *Institutional Investor* ratings, (iii) K-means clustering of macrovariables, and (iv) a combination of the above three forecasting methods (or classifiers) using a parametric regression. In the present context, clustering has not been utilized as yet and issues of forecast combination have barely been addressed. The analysis is based on a sample of 75 emerging/developing economies over the 1983–2000 period.

Second, the paper explores the evaluation of an EWS in relation to the decision-maker's objective function. We show how the latter can be taken into account to choose the classifier and its embedded parameters. The classifiers are shown to have different strengths in terms of missed defaults and false alarms. Furthermore, their forecast ranking is unstable over the holdout years. On the one hand, these findings imply that the user's loss function and degree of risk-aversion are critical inputs in the assessment of an EWS. On the other hand, they motivate forecast combining. It is shown that a relatively better balance of missed default and false alarms is achieved by combining the classifiers. Our framework can be easily adapted to distinct classifiers and loss functions. Finally, as a by-product of our analysis, some lessons emerge for practitioners in the area of sovereign default prediction. First, optimal recursive in-sample calibration of the classifiers is worthwhile. Second, given the persistence of sovereign default events, it seems sensible to gear the out-of-sample assessment of forecast ability toward default *entries* rather than continuing defaults.

Section 2 outlines the background literature. Section 3 describes the methodology and Section 4 introduces the data. Section 5 illustrates several issues regarding the optimal, recursive calibration of classifiers. The forecast combining analysis is presented in Section 6 before concluding.

2. Elements in the design of an optimal EWS

The goal of an EWS is to issue signals of pending debt repayment difficulties. Hence, the variable of interest takes a value of one at year t if a default occurs any time within an h-length window

$$Y_{it} = \begin{cases} 1 & \text{if } d_{i,t+k} = 1 \text{ at any } k = 0, 1..., h-1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and the classification problem at hand is formalized as $y_{it}=f(x_{i,t-1})$ where $x_{i,t-1}$ represents the available predictors at $t-1$. The forward-looking variable y_{it} is called the EWS indicator. The warning horizon, h, is the time interval within which the EWS should anticipate the occurrence of a crisis. If the warning horizon chosen is, say, $h=3$ years, then the forecast $\hat{y}_{i,t+1}=1$ indicates that a debt crisis will occur sometime during $[t+1, t+3]$. Currency crisis studies typically focus on $h=2$ years (Berg & Pattillo 1999; Kamin, 1999; Kumar et al., 2003), whereas in the debt crisis literature most studies use $h=1$ year. Choosing the optimal h requires a trade-off. The longer h is, the fewer missed defaults but the more false alarms, and vice versa. Bussière and Fratzscher (2002) show how to find the optimal warning horizon empirically according to a loss function. In order to assess the adequacy of an EWS, the probability forecasts are usually

transformed into event forecasts and compared with the EWS indicator y_{it}. For this purpose, the decision-maker should adopt a cut-off or threshold probability λ that is consistent with his loss function.

2.1. Decision-maker's loss function

The loss function facilitates the expected cost of mispredicting. In the present context, type II errors may be a matter of less concern than type I errors for two reasons. First, the costs of missed investment opportunities or those of adopting pre-emptive policies after *false warnings* are often less severe than the losses (reflected in the lender's balance sheet and reserves level) or the nation's welfare cost implied by *missed defaults*. Second, false alarms are not always 'mistakes' in that they may not stem from predictive failure of the EWS but simply reflect that, although there were severe economic weaknesses, suitable policy actions were taken and a crisis was avoided.

Suppose that an EWS is developed using the warning horizon h and the cut-off λ. On the basis of its forecasts, different error measures can be computed. Let $E_0(\lambda, h)$ and $E_1(\lambda, h)$ denote the number of false warnings ($\hat{y}_{it}=1\,|\,y_{it}=0$) and missed defaults ($\hat{y}_{it}=0\,|\,y_{it}=1$), respectively. Let $C_0(h)$ and $C_1(h)$ denote the total number of tranquil ($y_{it}=0$) and debt crisis ($y_{it}=1$) cases, respectively. The available sample has $C=C_0+C_1=NT$ cases, where T denotes time periods and N denotes countries. The type I error probability (P_{I} hereafter) is estimated as the percentage of missed defaults, $E_1(\lambda, h)/C_1(h)$. The type II error probability (P_{II}) gives the likelihood of a false alarm, and can be estimated as $E_0(\lambda, h)/C_0(h)$. Finally, let θ denote the decision-maker's degree of risk-aversion toward missing a crisis.

A typical loss function, which we call investor's loss (IL), is defined as

$$\mathrm{IL}(\theta,\lambda,h)=\theta P_{\mathrm{I}}(\lambda,h)+(1-\theta)P_{\mathrm{II}}(\lambda,h),\quad \mathrm{IL}\in[0,1] \tag{2}$$

and can be estimated by

$$\widehat{IL}(\theta,\lambda,h)=\theta\frac{E_1(\lambda,h)}{C_1(h)}+(1-\theta)\frac{E_0(\lambda,h)}{C_0(h)}.$$

The cost attached to a missed default relative to that of a false alarm is captured by the risk-aversion parameter, e.g. $\theta=0.8$ reflects a cost ratio for the decision-maker of 4 to 1. Thus, Eq. (2) represents a family of loss functions, parameterized by θ, which presumes that correct alarms have negligible costs, and so is thought to be more typical of investors.[1] To simplify the exposition, the ensuing analysis focuses on y_{it} with $h=1$ as the event to be forecasted and IL(θ, λ) as the objective function.[2]

It is assumed that the forecaster 'knows' the appropriate loss function and risk-aversion level that characterize the decision-maker. An analysis of the process by which the decision-making problem is shaped into a loss function and risk-aversion parameter goes beyond the scope of this study (for a discussion, see Abhyankar, Sarno, & Valente, 2005; Granger & Pesaran, 2000).

2.2. Forecast combining schemes

Bates and Granger (1969) set out the concept of forecast combining in their seminal paper. The basic idea is that, when several forecasts are available, it may pay to combine this information rather than to opt for one of the alternatives. Combination has been shown to be effective not only when the forecasts are obtained from widely heterogenous methods but also more generally (Clemen, Winkler, & Murphy, 1995; Montgomery, Zarnowitz, Tsay, & Tiao, 1998; Winkler & Makridakis, 1983). Forecast-ranking instability provides another rationale for combining (Aiolfi & Timmermann, 2003; Stock & Watson, 2001).

The extensive and continued interest in forecast combination is largely due to the wealth of evidence from empirical studies on its merits (Newbold & Harvey, 2004). Surprisingly, the concept has received scant attention in the sovereign default literature. Sommerville and Taffler (1995) compare judgmental forecasts (bankers' ratings) and parametric forecasts from logit and discriminant analysis (both based on macrodata), but do not assess the merits of combining them. Mascarenhas and Sand (1989) investigate the accuracy of discriminant analysis forecasts based on

[1] A correct default warning entails some transaction costs for investors.

[2] Another loss function used in the broad financial crisis literature is the *noise-to-signal ratio* introduced by Kaminsky et al. (1998), defined as the probability of a false alarm over the probability of a correct crisis warning. Other studies have employed what we call the *policymaker's loss* function, defined as a weighted sum of the missed default probability and the probability of issuing an early warning. For further discussion and an extension of the present analysis to these loss functions, see Fuertes and Kalotychou (2004).

credit ratings or macrovariables and find that combining them using Gupta and Wilton's (1988) Bayesian odds-matrix method outperforms the individual forecasts. But they do not explore other classifiers nor alternative weighting schemes in order to find the 'optimal' combination for the decision problem at hand. Manasse, Roubini, and Schimmelpfennig (2003) compare the forecasts from macrodata using a logit regression and a non-parametric recursive tree. They find that the latter yields fewer missed defaults but more false alarms and that accuracy is improved by combining the forecasts.

3. Competing classifiers

Three forecast approaches are considered which differ either in their underlying principle (classification method) or in the information set used. The forecasts are then combined by means of a parametric regression to assess the potential improvement in predictive power.

3.1. The LOGIT-M and LOGIT-R approaches

The first forecast approach is a pooled logit model for macrodata (LOGIT-M)

$$\log = \left[\frac{p_{it}}{1-p_{it}}\right] = \alpha + \beta' x_{i,t-1}, \quad i = 1,...,N, t = 1,...,T \qquad (3)$$

which implies the nonlinear relationship $p_{it} = \frac{e^{\alpha+\beta' x_{i,t-1}}}{1+e^{\alpha+\beta' x_{i,t-1}}}$, where $p_{it} \equiv \Pr(y_{it}=1|x_{i,t-1})$ and $x_{i,t-1}$ is an $s \times 1$ vector. The coefficient β_j, $j=1,...,s$, estimated by maximum likelihood, represents the marginal effect of the jth macrovariable on the log-odds ratio, $\log\left[\frac{p_{it}}{1-p_{it}}\right]$, ceteris paribus.[3]

Second, our analysis draws on sovereign credit ratings (z_{it} hereafter) which reflect a consensus of bankers' judgment. Several studies have found these internal ratings to be correlated with default signals such as GDP per capita, inflation, external debt, economic development, and default history (Cantor & Packer, 1996; Lee, 1993). Futhermore, these credit ratings incorporate important qualitative information on default risk such as the effects of social, political, and cultural conditions, and market expectations. The logit mapping, $\log\left[\frac{p_{it}}{1-p_{it}}\right] = \gamma + \phi z_{i,t-1}$, facilitates debt-crisis forecasts from the bankers' ratings. We refer to the latter as LOGIT-R forecasts.[4] In this framework, a cutoff probability λ is required to transform the probability estimates into EWS signals, i.e. $\hat{y}_{it}=1$ if $\hat{p}_{it}>\lambda$ and $\hat{y}_{it}=0$ if $\hat{p}_{it} \le \lambda$. Hence, optimal calibration implies finding the cut-off rate, λ^*, that is 'best' according to the decision-maker's preferences, namely, her loss function and risk-aversion.[5]

3.2. The K-clustering approach

The third classification technique we employ is K-means clustering.[6] The inputs or cases are the observation vectors, $x_{it}=(x_{it,1}, x_{it,2}, \ldots, x_{it,s})$, where s is the number of macrovariables. K-means clustering consists of comparing the distances of each observation vector from the mean vector of each of K clusters in the sample. The observation x_{it} is assigned to the cluster with the nearest mean vector. The distances are recomputed and reassignments are made as necessary. This process continues until all observations are in clusters with minimum distances to their mean vectors. Essentially, cases are allocated in clusters so as to maximise within-cluster similarity and between-cluster discrepancy. The K-clustering algorithm in steps is as follows:

1. Take the first K sample cases as the initial cluster centroids ($c_1^0 \equiv x_{11}$, $c_2^0 \equiv x_{21}$, ..., $c_K^0 \equiv x_{K1}$).
2. Assign case x_{it} to the cluster whose centroid is closest

$$c_j^0 = \underset{q=1,2,\ldots,K}{\operatorname{argmin}} \quad D(x_{it}, c_q^0)$$

[3] The forecasts from the pooled logit model are shown to outperform those from more sophisticated specifications such as a random coefficients logit under several loss functions (see Fuertes & Kalotychou, in press-a).

[4] The finite-sample properties of different estimation approaches for generating rating migration probabilities from Moody's external ratings are explored by simulation in Fuertes and Kalotychou (in press-b).

[5] See Fuertes and Kalotychou (2004) for details on how to embed the choice of warning horizon h and cut-off probability λ into the optimal design of an EWS.

[6] K-means clustering was chosen over hierarchical clustering techniques, such as nearest neighbour or average linkage, because for large datasets like ours these are computationally rather expensive.

$i=1, 2, \ldots, N$, $t=1, 2, \ldots, T$, where $D(x_{it}, c_q^0)$ denotes the Euclidean distance between the itth case and the qth cluster centroid, given by

$$D(x_{it}, c_q^0) = \sqrt{\sum_{l=1}^{s} (x_{it,l} - c_{q,l}^0)^2}$$

Thus, the outcome of step 2 is a set of K clusters of observation vectors. Let $m_1, m_2, \ldots, m_K$ denote the number of observation vectors in each cluster such that $\sum_{q=1}^{K} m_q = NT$.

3. The centroid of the qth cluster is given by its mean observation vector. The latter is defined as

$$c_q^1 = \left[\frac{1}{m_q}\sum\nolimits_{it} x_{it,1}, \ldots, \frac{1}{m_q}\sum\nolimits_{it} x_{it,s}\right]'$$

which facilitates a measure of the change in the cluster centroids, $\Delta S_q = D(c_q^1, c_q^0)$, $q=1, 2, \ldots, K$.

4. If $\Delta S_q < \varepsilon$ for all $q=1, 2, \ldots, K$ the algorithm terminates. Otherwise, a new iteration starts at step 2. We set $\varepsilon = 0.01$. The algorithm's output is the set of K clusters obtained at iteration j such that $\Delta S_q = D(c_q^{j+1}, c_q^j) < \varepsilon$ for all $q=1, 2, \ldots, K$.

The final K clusters are labelled as either default ($\hat{y}=1$) or non-default ($\hat{y}=0$) according to a user-specified assignment rule. An unseen or out-of-sample case x_{it} is classified to the cluster whose centroid is closest. The choice of K does not follow from the algorithm itself and so is often made subjectively. Hence, optimal calibration requires us to find the 'best' assignment rule and value of K according to the decision-maker's preferences. We deploy the following calibration approach.

For a given K, the assignment rule can be optimized as follows. Let $n_q(1)$ be the number of default cases (vectors $x_{i,t-1}$ such that $y_{it}=1$) in cluster q, and likewise for $n_q(0)$. Let C_1 (and C_0) denote the total number of default (non-default) cases. The loss implied by labelling cluster q as non-default is $\hat{\mathrm{L}}_{0,q}(\theta) = \theta \times \hat{\mathrm{P}}_{\mathrm{I}}$, where $\widehat{P}_{\mathrm{I}} = \frac{n_q(1)}{C_1}$ is the estimated probability that a default case falls in cluster q. Likewise, $\widehat{L}_{1,q}(\theta) = (1-\theta) \times \widehat{P}_{\mathrm{II}} = (1-\theta)\frac{n_q(0)}{C_0}$. The optimal rule for cluster q is

$$\widehat{y}_q^* = \underset{\widehat{y}\in\{0,1\}}{\operatorname{argmin}} L_{\widehat{y},q}(\theta) \tag{4}$$

with loss $L_q^*(\theta) = \min\{L_{0,q}, L_{1,q}\}$. The minimal overall loss is $L(\theta, K) = \sum_{q=1}^{K} L_q^*(\theta)$.

Large K-clustering characterizes the sample rather well, but not necessarily the population, and so it may yield poor out-of-sample forecasts. The optimal K can be found by a method introduced by Frydman, Altman, and Kao (1985) to correct for an overfitting bias in recursive partitioning. Consider $K \in \{2, \ldots, K_{\max}\}$ and let $L(\theta, K, \delta) = L(\theta, K) + \delta \times K$, where $L(\theta, K)$ is the overall minimal loss for a given K, defined above, and $\delta \geq 0$ is an overfitting penalty. For each $\delta \in \{\delta_1, \ldots, \delta_n\}$, we find $\widetilde{K}_\delta = \underset{K}{\operatorname{argmin}} L(\theta, K, \delta)$. This yields a set $\{\tilde{\mathrm{K}}_{\delta_1}, \ldots, \tilde{\mathrm{K}}_{\delta_n}\}$ from which K^* is found using cross-validation. To conduct the latter, a random partition of $\{x_{it}\}$ into V equally sized groups is made. For δ_1, we leave out one group and cluster the remaining cases using the above $\tilde{\mathrm{K}}_{\delta_1}$ selected from the complete $\{x_{it}\}$. The left-out cases are then assigned to the nearest existing cluster. The procedure is iterated by excluding a different group each time. The cross-validated loss associated with $\tilde{\mathrm{K}}_{\delta_1}$ is the average loss over all iterations, $\mathrm{CV}\left[L\left(\theta, \widetilde{K}_{\delta_1}\right)\right] = \frac{1}{V}\sum_{i=1}^{V} L\left(\theta, \widetilde{K}_{\delta_1)_i}\right)$, where i signifies the group left out in iteration i. The optimal K^* minimizes the cross-validated loss

$$K^* = \underset{\widetilde{K}_{\delta_j}}{\operatorname{argmin}}\, \mathrm{CV}[L(\theta, \widetilde{K}_\delta)], j = 1, 2, \ldots, n$$

For the analysis below, we set $K_{\max}=10$, $\delta \in \{0.001, 0.002, \ldots, 0.01\}$ and $V=5$.[7]

The main advantage of clustering over logit is its non-parametric nature, namely, it does not require the forecaster to formalize the relationship between the exogenous variables and the default event. But clustering has some pitfalls. First, it does not provide a continuous scoring scale such as the posterior probability of default, and so it cannot produce the rankings of the susceptibility of different countries to crisis which are key for international investors. Second, the main aspects of the default clusters (e.g. low trade/GDP) are often not clear-cut, particularly when many variables are used, and so one cannot identify the key determinants of default which are crucial for policymakers.

[7] $V=5$ has been shown to produce good calibration of the number of nodes in classification trees (Breiman, Friedman, Olsen, & Stone, 1984). Our choice of range for δ is driven by the order of magnitude of the IL changes for successive K.

How does the K-clustering work in practice? For simplicity, let us focus on a two-variable setting, external debt to GDP and GDP growth, and $K=4$ clusters. Suppose we want to classify the observations at $t=1995$ as default or non-default. In the present sample, there are 76 countries with data for 1995, 22 of which are default incidents ($C_1=22$), while the remaining 54 are non-defaults ($C_0=54$). Thus, the in-sample default probability is 29%. Table 1 sets out the results of applying the four steps of the aforementioned clustering algorithm.

The final cluster centroids or variable means are examined to assess how distinct the four clusters are and provide a broad characterization of each cluster: 'low debt–high growth', 'high debt–medium growth', and so forth. Ideally, one would obtain very different means for most, if not all, of the variables used in the analysis. The 'low debt–high growth' cluster ($q=1$) contains 5 default and 38 non-default cases, and the within-cluster default probability is 12% (=5/43). The latter gives the default probability conditional on low debt and high growth rates. In the 'high debt–medium growth' cluster, there are two default and one non-default observations, and the conditional default probability rises to 67% (=2/3). Interestingly, the 'medium debt–medium growth' cluster splits the cases virtually evenly (13 default and 12 non-default) so that the default probability conditional on medium debt and growth ratios is 52%. Finally, the 'low debt–low growth' cluster contains two defaults and three non-defaults, lowering the conditional probability of default to 40%.

Table 1
Example of clustering analysis for 1995

Cluster	1	2	3	4
A: *Characteristics*				
Centroid	(0.120, 0.748)	(0.788, 0.649)	(0.376, 0.634)	(0.143, 0.481)
External debt/GDP	Low	High	Medium	Low
GDP growth	High	Medium	Medium	Low
Total cases	43	3	25	5
Default	5	2	13	2
Non-default	38	1	12	3
Default prob	12%	67%	52%	40%
B: *Assignment rule*				
Default loss, $L_1(0.5)$	35.2%	**0.9%**	**11.1%**	**2.8%**
Non-default loss, $L_0(0.5)$	**11.4%**	4.5%	29.5%	4.5%
Label	non-default	default	default	default

Clustering is carried out for two variables, external debt to GDP and GDP growth, over the year 1995. Centroid refers to the within-cluster variable averages. $L_i(0.5)$ is the loss of labelling a cluster as event $i=0,1$ at a risk-aversion level of 0.5. Bold denotes the minimum loss. For each variable, the clusters are denominated low, medium or high on the basis of how the centroid compares with the overall variable average.

The next step is to label the clusters using an optimal assignment rule as described above. For each cluster, the loss implied by labelling it as default, $L_{1,q}(\theta=0.5)$, and non-default, $L_{0,q}(\theta=0.5)$, is computed. For instance, the implied losses for cluster $q=1$ are $L_{1,1}\ (\theta=0.5)=(1-\theta)\frac{n_1(0)}{C_0}=0.5\times\frac{38}{54}=35.2\%$ and $L_{0,1}\ (\theta=0.5)=(\theta)\frac{n_1(1)}{C_1}=0.5\times\frac{5}{22}=11.4\%$, respectively. The cluster is then assigned to the loss-minimising category. Thus, cluster 1 is labelled 'non-default', whereas 2, 3, and 4 are the 'default' clusters. The minimal overall clustering loss is the sum of the individual clustering losses, that is $L(\theta=0.5, K=4)=0.262$. Out-of-sample forecasts for $t=1996$ can now be generated by classifying each observation vector to the nearest of the four clusters formed.

3.3. Combining the forecasts from LOGIT-M, LOGIT-R, and K-clustering

Let $\{\hat{y}^m_{i,t+1}\}^M_{m=1}$ denote M rival forecasts formed at period t, and $\hat{y}^C_{i,t+1}=\mathbf{R}(\hat{y}^1_{i,t+1}, \ldots, \hat{y}^M_{i,t+1})$ the combined forecast where $\mathbf{R}$ is a mapping or transformation. We consider a parametric mapping based on a logit regression on the individual forecasts. Kamstra and Kennedy (1998) [KK] propose a logit approach that is simple to apply and can combine probability and event forecasts, or a mix. They show by simulation that, for large samples, it is superior to the equal-weights scheme. In contrast to the latter approach, KK-logit accounts for the historical, in-sample forecast ability. Moreover, the KK-logit principle can be extended to polychotomous and ordered classification problems using multinomial or ordered logits, respectively.[8]

According to the KK-logit approach, we fit a logit regression of the EWS indicator (y_{it}) on a constant, the

[8] We also deployed two non-parametric voting rules: (*i*) the Majority Rule and (*ii*) the Unanimous Rule, according to which the combined forecast is the event predicted by the majority of the classifiers or by all of them, respectively. Hereafter, the discussion focuses on *KK*-logit because overall it was found to beat these voting rules.

log-odds ratio forecasts $\left(\frac{\log \hat{p}_{it}}{1-\hat{p}_{it}}\right)$ from LOGIT-M and LOGIT-R, and the event forecasts ($\hat{y}_{it}$) from K-clustering. The coefficient estimates are the combining weights. To allow for time variation, this approach is recursively applied in-sample over a 12-year rolling window. Thus, we obtain weights $\mathbf{w}_\tau \equiv (w_\tau^1, w_\tau^2, w_\tau^3)$ for each set of out-of-sample forecasts, $\tau = 1996, \ldots, 2000$.

A nice property of KK-logit is that it enables forecast encompassing tests (Fair & Shiller, 1990). In our analysis, the mth individual forecast ($m = 1, 2, 3$) for the year τ is discarded if the combining weight w_τ^m is statistically insignificant. Finally, the combined forecast, $\hat{p}_{i\tau}^C$, is transformed into an event forecast by means of a cut-off λ_τ^* which is chosen optimally for each $\tau = 1996, \ldots, 2000$.

4. The data

The analysis is based on annual data for 75 emerging and developing countries 1983–2000.[9] The predictors (explanatory variables) are lagged 1 year and so the effective sample period for y_{it} is 1984–2000: the initial 12-year rolling window is 1984–1995 and the holdout period is 1996–2000.[10]

The default indicator $\{d_{it}\}_{t=1984}^{2000}$ is based on *World Bank* data for external debt, principal, and interest arrears to official/private creditors, and principal rescheduled. A given country-year is a 'default' case ($d_{it} = 1$) if: (a) arrears increase over a threshold percentage of external debt, $\Delta A_{it} > \delta D_{it}$, where δ is the overall sample mean of $\Delta A_{it}/D_{it}$ at 2.26%; and (b) a rescheduling agreement is reached and the total amount of debt rescheduled exceeds the decrease, if any, in the arrears stock. The default events thus obtained correspond closely to those in Standard and Poor's (2001).

Data on 24 macroeconomic and financial ratios is collected from the *World Bank*. These include short-term financial solvency and liquidity proxies and long-term structural economic signals. To reduce the degree of skewness and kurtosis in the ratios and the number of outliers, these are logged using $\text{sign}(x)\ln(1+|x|)$. Any remaining outlier in each default/non-default group is tackled by windsorizing the log variables as follows. A data point x_{it} is indexed by $c \in \{0,1\}$ according to whether it pertains to a tranquil ($y_{it} = 0$) or default ($y_{it} = 1$) episode. If x_{it}^c falls outside $\bar{x}^c \pm 4\hat{\sigma}^c$, it is replaced by the appropriate interval limit. Classification methods have been shown to work better when the variables are mapped onto the [0,1] interval. Thus, the N points per year for each variable, $\{x_{it}\}_{i=1}^N$, are further transformed using $\tilde{x}_{it} = (x_i - \min\{x_{it}\}) / (\max\{x_{it}\} - \min\{x_{it}\})$.

This large regressor set is shrunk separately for each classifier by means of an in-sample jacknife or cross-validation approach.[11] Table 2 reports the 'optimal' regressor sets, $x_{it} \equiv (x_{it,1}, \ldots, x_{it,s})'$ with $s = 13$ for LOGIT-M and $s = 15$ for K-clustering.

The jacknife results are quite robust in that for about 70% of the ratios there is agreement for LOGIT-M and K-clustering: 10 ratios are jointly selected and 6 ratios are jointly discarded. There is unanimity for the external economic activity signals: volatility of export growth and trade balance/GDP are relevant for both classifiers while the remaining three ratios are discarded in both. Other jointly relevant ratios are: three external credit exposure signals (external debt to GDP, short-term to total debt and IMF credit to exports), four measures of domestic conditions (private credit to GDP, growth in GDP, growth in GNP volatility, real exchange rate), and a global link signal, the degree of trade openness. Other variables like short-term debt to reserves, government expenditure and gross domestic savings are found to be fruitful for clustering but not for logit.

Country credit ratings are obtained from the *Institutional Investors* database.[12] These ratings are an index based on the weighted scores from the 100 largest international banks and seek to capture the perception of bankers worldwide regarding a country's ability and willingness to service its financial obligations. The latter closely monitor the observance of

[9] The regions (number of countries in parenthesis) are: East Europe (7), Asia (12), Latin America (22), Middle East/North Africa (9), and Africa (25).

[10] LIMDEP 8 and SPSS 10 are used in the empirical analysis.

[11] The jacknife is based on recursive logit/clustering predictions 1984–1995 (λ is set at the default frequency in each iteration and, to reduce computational costs, $K = 2$). A variable is dropped if in doing so the cross-validated loss, a conservative IL measure for $\theta = 1$, does not increase (for details, see Fuertes & Kalotychou, in press-a).

[12] External credit ratings from Moody's and S&P's were unavailable for many countries in our sample.

standards—whether a country has published an IMF Article IV or ROSC and met the SDDS specifications.[13] More specifically, the rating series used, $\{z_{it}\}_{t=1984}^{2000}$ varies, on a 1–100 scale with 100 representing low default-risk countries. The ratings are updated semi-annually and our LOGIT-R classifier is based on end-of-year scores. To avoid sample selection bias in comparing the three classifiers, the country–period cases subsequently used in this study are those for which both x_{it} and z_{it} are available.

We should stress that sovereign debt crises typically last longer than 1 year, in contrast with banking crises. About 30% of all country-period cases over 1984–2000 are defaults ($d_{it}=1$), whereas about 10% are default entries ($\Delta d_{it}=1$). The average length of a debt crisis is around 3 years. Hence, the real challenge for an EWS is to predict a default *entry* rather than a continuing default. In order to develop a powerful EWS in the above sense, it is sensible to evaluate the loss functions over an entry set defined as follows. Year t is excluded for country i if this sovereign was already in default at year $t-1$, i.e. $d_{it}=1$ is excluded if $d_{i,t-1}=1$.

5. Optimal calibration of forecasting tools

This section discusses the in-sample calibration of the classifiers over the 12-year window 1984–1995.

5.1. Balancing missed defaults and false alarms

The cut-off rate (λ) and warning horizon (h) parameters of an EWS are often chosen subjectively. An objective choice should take into account the decision-maker's desired trade-off between type I and type II errors. Fig. 1, panel A, illustrates this point for the LOGIT-M classifier.[14]

A higher λ or a lower h yield fewer false alarms at the cost of more missed defaults. From the perspective of creditors or investors, the latter will imply realised losses, rising reserve holdings and adversely affected cash flows and asset values. From the perspective of policymakers, the experience of the 1990s has highlighted falling output, increasing unemployment, and poverty rates as some of the notable repercussions of

Table 2
In-sample jacknife variable selection 1983–1995

	LOGIT-M		K-Clustering
	Coefficient	*t*-ratio	Outcome
External credit exposure			
Total external debt/GDP	6.96*	7.19	✓
Official debt/total debt	2.47*	2.03	×
Short-term debt/reserves[a]	×		✓
Short-term debt/total debt	−0.07	−0.06	✓
Debt service/exports	−1.94*	−3.13	×
IMF credit/exports	−1.21	−1.78	✓
External economic activity			
Export growth[b]	×		×
Volatility of export growth[c]	1.42*	2.18	✓
Trade balance[d]/GDP	−1.20	−1.34	✓
Reserves growth[a,b]	×		×
Reserves/imports[a]	×		×
Domestic conditions			
Credit to private sector/GDP	−1.80*	−3.31	✓
GDP growth[b]	−0.81	−1.39	✓
GNP per capita	1.91*	3.31	×
Volatility of GNP p.c. growth[b,c]	1.68*	3.08	✓
Government expenditure/GDP[e]	×		✓
Inflation	×		×
M2/reserves[a]	×		×
Real exchange rate[f]	0.44	0.76	✓
Gross capital formation/GDP	×		✓
Gross domestic savings/GDP	×		✓
Global links			
Trade[g]/GDP	−2.75*	−3.99	✓
Net bond flow[f,h]	×		✓
Net equity flow[f,h]	×		×

A cross indicates that the variable is not retained by the jacknife procedure.
[a] FX reserves, excluding gold.
[b] Annual growth (%).
[c] Volatility proxied by StDev over [t, $t-4$] years.
[d] Exports–imports.
[e] Expenditure on consumption, national security and defense.
[f] Deviation from long-run trend.
[g] Exports+imports.
[h] US$ billion.
* 1% level significant, two-sided test.

[13] The Special Data Dissemination Standards (SDDS) was designed for countries with/seeking access to international capital markets. It sets data definitions, particularly for reserves, and minimum timeliness and frequency standards for data releases. The Reports on the Observance of Standards and Codes (ROSC) are voluntary and refer to transparency, financial market regulation and corporate governance issues (see Glennerster, 2004).

[14] The LOGIT-R calibration raises similar issues to that of the LOGIT-M so we just focus on the latter.

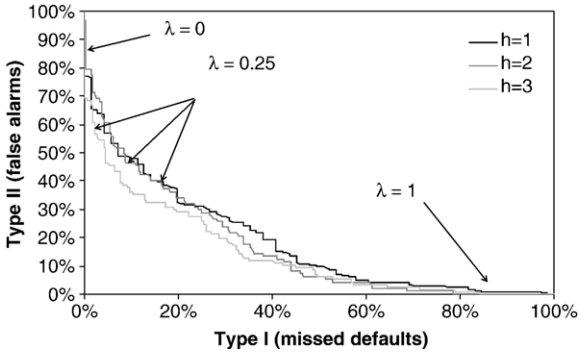

Fig. 1. Types I and II error and the cut-off rate (LOGIT-M).

sovereign debt crises. Lowering λ or raising h will induce the opposite effect: fewer missed crises at the cost of more false alarms. The latter means foregone profit opportunities for investors and unnecessary policy actions which may be costly, for instance, in terms of social unrest. Hence, forecasters should use the (λ, h) combination that is 'best' according to the decision-maker's preferences. Full consideration of all the aforementioned choices (logit cut-off, number of clusters, warning horizon, risk-aversion, loss function) has too many dimensions to be tractable. To keep the analysis focused, a warning horizon $h=1$ year is adopted and the main task is to incorporate the decision-maker's degree of risk-aversion (θ) into the optimal calibration of the cut-off and number of clusters.

The length of the time series used in sovereign default studies is typically long enough for autocorrelation to become an issue in logit models. The debt crisis indicator y_{it} is often a multiple-period outcome (overlapping problem when $h>1$) or, more generally, autocorrelated (even for $h=1$), which implies moving average prediction errors. Furthermore, the typical macroeconomic and financial ratios employed are persistent, which induces autocorrelated errors—a sluggish, high external debt/GDP indicator that predicts a debt crisis at t is likely to also do so at $t+1$. Hence, the usual logit maximum likelihood standard errors are biased and tests based on them have incorrect size. Some studies use 'exclusion windows' whereby consecutive default years within a certain time window are excluded (Detragiache & Spilimbergo, 2001; Frankel & Rose, 1996). However, in using a reduced sample for model specification and estimation, one may discard important information. Hence, we exploit all the in-sample cases for the logit estimation and clustering, but the out-of-sample forecast accuracy is gauged on the basis of the default entry set, in which case forecast errors are likely to be independent. Nevertheless, the present goal is developing an optimal EWS (with a focus on the choice of econometric methodology and the evaluation of the EWS itself) as opposed to hypothesis testing, and so the bias in standard errors is irrelevant.

5.2. Optimal cut-off and number of clusters

What are the optimal cut-off and number of clusters for an EWS of sovereign default? To answer this question, we deploy the optimization approaches outlined in Section 3. Fig. 2, panel A, illustrates the in-sample calibration of LOGIT-M for representative decision-makers with low ($\theta=0.2$), moderate ($\theta=0.5$), and high ($\theta=0.8$) risk-aversion levels. Fig. 2, panel B, shows the optimal cut-off rate for different degrees of risk-aversion.

Overall, these findings illustrate that the optimal cut-off rate, λ^*, decreases with the decision-maker's degree of risk-aversion towards missing a debt crisis.

We now turn to the optimal calibration of K-clustering. The results suggest that, for the low risk-aversion level $\theta=0.3$, the number of clusters yielding the minimal loss is $K^*=8$, whereas for $\theta=0.5$ and $\theta=0.8$ the best choices are $K^*=7$ and $K^*=6$, respectively. Fig. 3(A) illustrates the calibration exercise under the IL objective function with $\theta=0.5$.

Fig. 3(B) gives the relationship between the optimal K and the risk-aversion level which, interestingly, is roughly V-shaped. The main result is that the optimal number of clusters depends on the decision-maker's degree of risk-aversion. The assignment or labelling of the final clusters as 1 (default) or 0 (non-default) is akin to the selection of the cut-off in the logit. As θ increases, the optimal rule is such that the default state is assigned to relatively more clusters.

6. Optimal forecast combination

The above results suggest that the decision-maker's preferences should be accounted for in the design of an EWS, namely, in calibrating parameters such as the cut-off rate and number of clusters. As competing

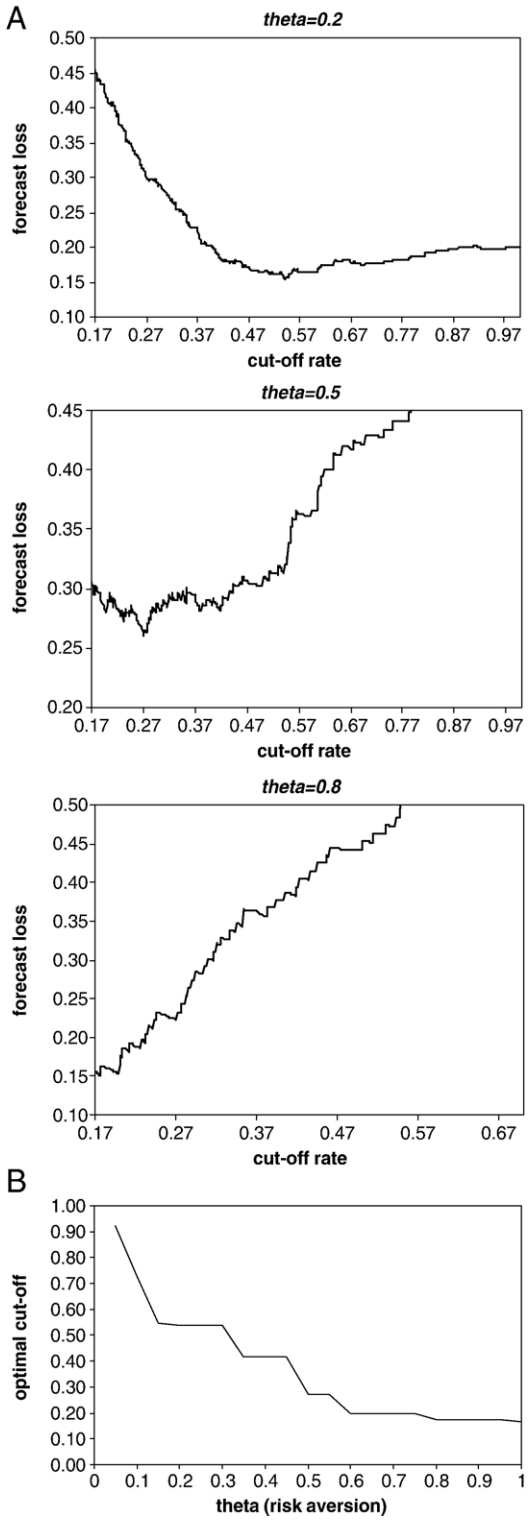

Fig. 2. Calibration of the cut-off rate (LOGIT-M). Panel A: loss and cut-off rate. Panel B: optimal cut-off rate and risk aversion (LOGIT-M).

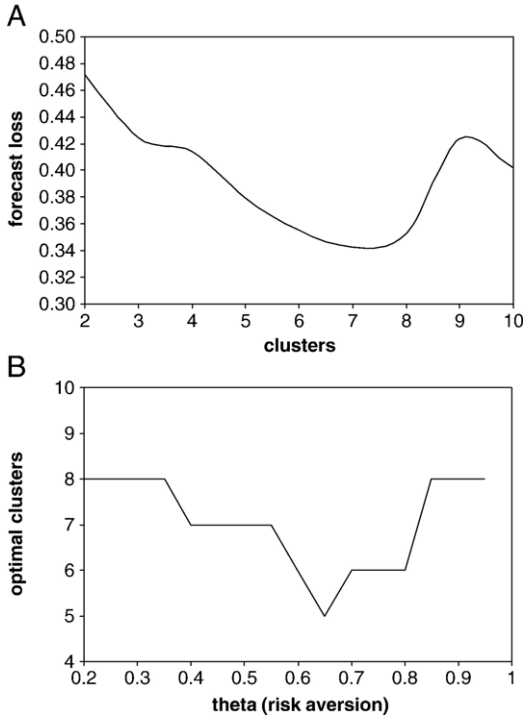

Fig. 3. Calibration of the number of clusters. Panel A: loss and number of clusters ($\theta = 0.5$). Panel B: optimal number of clusters and risk aversion.

econometric methods may have different strengths in terms of type I and II errors, another potential way to improve the performance of an EWS is by forecast combining.

We start by assessing the stability of the out-of-sample forecast ranking. Table 3 indicates the best classifier year by year and the associated minimal loss (in parentheses) for each risk-aversion level.

The forecast ranking changes over time, which further motivates the forecast combination exercise. For instance, LOGIT-R dominates in 1996, 1998, and 2000, whereas LOGIT-M stands out in 1997. The forecast instability pattern can be explained in terms of the different type I and II error rates of the classifiers, which are detailed below. For example, LOGIT-R emerges as quite 'pessimistic' toward country credit-worthiness, and so it does quite well in 1996 and 1998

Table 3
Stability of forecast ranking over holdout period 1996–2000

	Risk-aversion level (θ)														
Year	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
1996	Cluster	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-M	LOG-R	LOG-R	LOG-R	LOG-R	LOG-M
	(0.189)	(0.138)	(0.148)	(0.158)	(0.124)	(0.272)	(0.260)	(0.249)	(0.240)	(0.210)	(0.187)	(0.172)	(0.138)	(0.103)	(0.069)	(0.041)
1997	LOG-M	LOG-R	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-R
	(0.167)	(0.203)	(0.120)	(0.129)	(0.138)	(0.159)	(0.170)	(0.153)	(0.136)	(0.119)	(0.102)	(0.154)	(0.123)	(0.093)	(0.062)	(0.034)
1998	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R
	(0.198)	(0.217)	(0.256)	(0.295)	(0.268)	(0.263)	(0.284)	(0.302)	(0.280)	(0.240)	(0.247)	(0.223)	(0.198)	(0.191)	(0.160)	(0.130)
1999	Cluster	Cluster	Cluster	LOG-R	LOG-R	LOG-R	LOG-R	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-M	LOG-R
	(0.200)	(0.250)	(0.300)	(0.244)	(0.277)	(0.140)	(0.128)	(0.048)	(0.102)	(0.112)	(0.096)	(0.080)	(0.064)	(0.073)	(0.049)	(0.043)
2000	LOG-M	LOG-M	LOG-M	Cluster	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	LOG-R	Cluster	LOG-R
	(0.160)	(0.211)	(0.245)	(0.178)	(0.135)	(0.160)	(0.191)	(0.186)	(0.181)	(0.192)	(0.137)	(0.114)	(0.109)	(0.098)	(0.076)	(0.040)

The investor's loss (IL) metric is evaluated over the holdout default entry sample 1996–2000. For each out-of-sample year, we indicate the best model, and the minimal $\widehat{IL}(\theta) = \theta\,\hat{P}_{\mathrm{I}} + (1-\theta)\,\hat{P}_{\mathrm{II}}$ is reported in parenthesis. The out-of-sample forecasts are generated recursively from classifiers calibrated over a 12-year rolling window. LOG-M and LOG-R denote the logit classifier based on macrovariables and credit ratings, respectively. Cluster denotes the K-means clustering classifier.

when a relatively large number of defaults occurred, but it does worse in 1997 which had relatively few default entries. To combine the forecasts, we use the (parametric) KK-logit regression discussed earlier. Table 4 reports the KK-logit combining weights for the individual out-of-sample forecasts for the year 1996.

These weights are obtained via a 1984–1995 logit regression of y_{it} on the LOGIT-M, LOGIT-R, and K-clustering forecasts. Since the latter are event forecasts, they change with the optimal number of clusters and the assignment rule, which in turn depend on the decision-maker's risk-aversion level, θ, as the foregoing analysis has shown. Thus, each θ implies a different set of K-clustering forecasts, and so the forecast combining weights also vary. The cut-off rate (for LOGIT-M and LOGIT-R) plays no role in this combining exercise because $\hat{\mathrm{p}}_{it}$ is used directly. It is worth noting that the combining weights on the LOGIT-M and LOGIT-R forecasts are significant for all risk-aversion levels, whereas those of K-clustering are only significant for moderate levels, $\{0.4<\theta<0.6\}$.

We now compare the out-of-sample predictive ability of the classifiers. First, the calibration of parameters–cut-off for LOGIT-M and LOGIT-R and number of clusters and assignment rule for clustering–is carried out over the 1984–1995 window. The logit estimates and final clusters thus obtained are used to generate forecasts for 1996. Next, this calibration and estimation/clustering is reconducted for 1985–1996 to generate out-of-sample forecasts for 1997 and so forth. A country–mean forecast loss is obtained for each out-of-sample year and then averaged over years.

Table 5 sets out the holdout-sample comparison across individual and combined forecasts. The entries are the type I and type II error rates and the overall forecast loss. At a moderate risk-aversion level ($\theta=0.5$), the LOGIT-R achieves the minimal type I error rate, missing only 18.7% of default entries, as opposed to 44.3% and 49.3% by the LOGIT-M and the K-clustering, respectively. But the satisfactory performance of LOGIT-R in terms of correctly predicting crises comes at the cost of 22.6% false alarms. On the other hand, the LOGIT-M compensates for its relatively higher type I error rate by signalling fewer false alarms, 20.5%, while the K-clustering again ranks last with 33.7% of the tranquil periods wrongly predicted. The findings can be generalized for virtually all realizations of θ. More specifically, at $\theta=0.3$, LOGIT-R has a type I error rate of 72.1%

Table 4
Combination weights based on KK-logit regression 1984–1995

Weights (ω_τ^m)	Risk-aversion level (θ)									
	0.2	0.25–0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7–0.8	0.85–0.95
LOGIT-M	0.985	1.016	0.989	0.959	0.977	0.959	1.002	1.007	1.005	0.991
	(5.66)	(5.61)	(5.69)	(5.65)	(5.72)	(5.65)	(5.85)	(5.86)	(5.90)	(5.77)
LOGIT-R	0.591	0.605	0.577	0.579	0.577	0.579	0.573	0.590	0.565	0.614
	(3.39)	(3.47)	(3.24)	(3.36)	(3.33)	(3.36)	(3.17)	(3.28)	(3.08)	(3.54)
Clustering	0.390	−0.023	0.204	0.672	0.400	0.672	0.303	0.168	0.357	27.304
	(0.74)	(−0.05)	(0.60)	(2.03)	(2.26)	(2.03)	(0.566)	(0.29)	(0.61)	(0.00)
Intercept	−0.395	−0.330	−0.430	−0.817	−0.614	−0.817	−0.628	−0.500	−0.685	−27.640
	(−1.90)	(−1.53)	(−1.72)	(−2.64)	(−2.08)	(−2.64)	(−1.14)	(−0.82)	(−1.12)	(0.00)

The results are for the $\tau=1996$ forecasts. The weights for the 1997 forecasts (unreported) come from a 1985–1996 logit and so forth. The *t*-ratio for the coefficient significance is in parentheses.

compared to 73% from LOGIT-M, while at $\theta=0.8$ the difference is slightly more pronounced with the figures being 7.0% and 10.9%, respectively, for the two classifiers.

The type I error rate of K-clustering is essentially higher than that of LOGIT-R or LOGIT-M for low to moderate risk-aversion levels $\theta<0.55$. For instance, the clustering misses as many as 90.1% of crisis entries at a risk-aversion level of $\theta=0.3$. For higher values of θ, clustering gives few missed defaults (2–6%), albeit at the cost of too many false alarms (69–88%).

Regarding the false alarms, at $\theta=0.3$ LOGIT-M mispredicts 4.2% of all tranquil periods, whereas LOGIT-R yields 6.1%. At $\theta=0.8$, the LOGIT-M still outperforms the LOGIT-R with 52% and 58.2% type II error rates, respectively. The same ranking is observed

Table 5
Out-of-sample forecast ability of individual and combined classifiers 1996–2000

Classifier	Risk-aversion level (θ)															
	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
A: Type I error (missed defaults)																
LOGIT-M	**0.844**	0.824	0.730	0.667	0.667	0.607	0.443	0.243	0.137	0.137	0.137	0.137	0.109	0.089	0.089	0.060
LOGIT-R	1.000	**0.779**	**0.721**	**0.564**	**0.357**	**0.207**	**0.187**	**0.187**	0.139	0.099	0.070	0.070	0.070	0.070	0.020	0.020
Clustering	0.971	0.901	0.901	0.687	0.786	0.603	0.493	0.423	**0.060**	**0.060**	**0.060**	**0.060**	**0.060**	**0.020**	**0.020**	**0.020**
KK-logit	0.710	0.681	0.681	0.563	0.563	0.443	0.343	0.157	0.109	0.109	0.109	0.109	0.109	0.109	0.109	0.109
B: Type II error (false alarms)																
LOGIT-M	0.031	**0.035**	**0.042**	**0.070**	**0.070**	**0.084**	**0.205**	**0.214**	0.350	**0.367**	**0.367**	**0.449**	**0.520**	**0.572**	**0.572**	**0.773**
LOGIT-R	**0.009**	0.048	0.061	0.100	0.131	0.239	0.266	0.266	**0.337**	0.494	0.564	0.564	0.582	0.669	0.788	0.833
Clustering	0.026	0.086	0.086	0.128	0.069	0.190	0.337	0.408	0.690	0.683	0.794	0.794	0.794	0.834	0.834	0.877
KK-logit	0.044	0.048	0.061	0.075	0.070	0.101	0.145	0.270	0.367	0.380	0.393	0.402	0.419	0.419	0.419	0.419
C: Overall loss (IL)																
LOGIT-M	**0.193**	0.232	**0.248**	0.279	0.309	0.319	0.324	0.230	0.222	**0.218**	**0.206**	0.215	0.191	0.161	0.137	0.096
LOGIT-R	0.207	**0.231**	0.259	**0.262**	**0.221**	**0.227**	**0.227**	**0.223**	**0.218**	0.237	0.218	**0.194**	**0.172**	0.160	**0.097**	**0.061**
Clustering	0.215	0.290	0.330	0.324	0.356	0.376	0.415	0.416	0.312	0.278	0.280	0.243	0.207	**0.142**	0.101	0.063
KK-logit	0.177*	0.207*	0.247	0.245	0.267	0.255	0.244	0.208*	0.212	0.204*	0.194*	0.182*	0.171	0.155	0.140	0.124

The reported type I error is $\widehat{P}_{\mathrm{I}}=\frac{E_1}{C_1}$; the type II error is $\widehat{P}_{\mathrm{II}}=\frac{E_0}{C_0}$; the overall loss is $\widehat{IL}(\theta)=\theta\,\widehat{P}_{\mathrm{I}}+(1-\theta)\,\widehat{P}_{\mathrm{II}}$. E_i and C_i (event i=0,1) are the number of prediction errors and sample cases, respectively. All metrics are evaluated over the holdout default entry sample 1996–2000. The out-of-sample forecasts are generated recursively from classifiers calibrated over a 12-year rolling window. Bold denotes the best individual outcome. KK-logit is the regression based combining scheme. The forecast ability (based on the IL function) of the combined classifier is compared with that of the best individual classifier using a Diebold-Mariano test. *Denotes significantly better performance at the 5% level.

for virtually all risk-aversion levels. The upshot is that LOGIT-R outperforms LOGIT-M regarding missed default entries, whereas the opposite holds for false alarms. LOGIT-M classifies the non-defaults relatively well and so it generally dominates the other classifiers in terms of the type II error rate, whereas LOGIT-R is superior in terms of missed defaults at the cost of more false alarms. This suggests that the bankers' judgments (ratings) are relatively pessimistic about country creditworthiness. Evidence on the latter has also been provided by Sommerville and Taffler (1995).

How do these results compare with the literature? The answer is not straightforward because there are many dimensions to the problem and so the prediction error rates are not strictly comparable. The available studies are based on different samples regarding country coverage and time span. More importantly, different logit cut-off probabilities are used, which amounts to adopting different risk-aversion levels. Furthermore, the holdout sample size (number of years) varies and many available studies report in-sample forecast errors only. Nonetheless, it is interesting to get a rough indication of how the type I and type II error rates of our best models fare against those of key studies. For instance, Sommerville and Taffler (1995) report out-of-sample type I and type II error rates of 9% and 22%, respectively, for their best model (logit) based on a cut-off of 31.1%. The latter corresponds to a relative misclassification cost of between 0.60 and 0.85, which is not directly analogous to the risk-aversion parameter θ because their loss function is slightly different from (2). Manasse et al. (2003) report type I and type II error rates of 55% and 6%, respectively, based on a logit cut-off of 21% which represents their in-sample crisis frequency. Finally, McFadden et al. (1985) report in-sample forecast errors of 15% and 24%, respectively.

For each risk-aversion level, the ranking of the classifiers in terms of overall loss (IL) follows from their relative type I and type II errors. As the three classifiers have a different balance of missed defaults and false alarms, controlling for the decision-maker's degree of risk-aversion is key to any forecast evaluation exercise. For instance, at the low risk-aversion level $\theta=0.3$, the minimal loss is that of LOGIT-M because of its relatively small type II error rate, despite having a large type I error rate. However, as the risk-aversion penalty for type I errors increases, LOGIT-R generally beats LOGIT-M. The overall loss of the non-parametric (clustering) classifier is relatively large except at very high risk-aversion levels $\theta \geq 0.80$. These considerations prompt the thought that there may be gains from forecast combining.

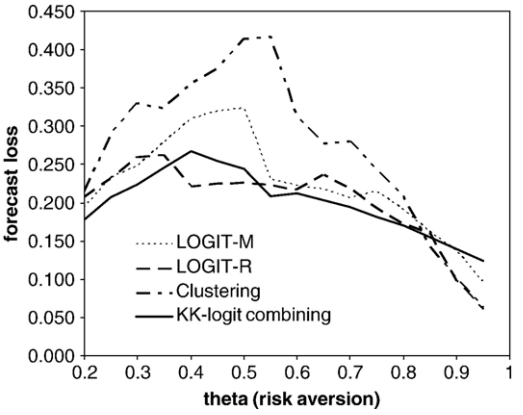

Fig. 4. Overall loss for individual and combined forecasts.

In terms of missed defaults, the combined forecast beats all the individual forecasts for low to moderate risk-aversions, $\{\theta \leq 0.35 \cup \theta = 0.55\}$. For instance, at $\theta=0.3$, the combination yields a type I error rate of 68.1%, whereas the best individual forecast, LOGIT-R, gives 72.1%. For a decision-maker with moderate risk-aversion, $\theta=0.55$, the figures are 15.7% for the combined forecast and 18.7% for the best individual forecast, again LOGIT-R. In terms of false alarms, forecast combining is fruitful for moderate to high risk-aversion levels, $\{\theta=0.5 \cup \theta>0.75\}$. At $\theta=0.5$, the combined scheme sends false crisis signals in 14.5% of cases whereas the best individual forecast, LOGIT-M, yields 20.5%. At higher risk-aversion levels, say $\theta=0.8$, the combined forecast achieves a 41.9% false alarm rate, whereas the best individual (LOGIT-M) model gives 52%.

Regarding the overall loss (IL), at first glance, the combining exercise appears worthwhile for most risk-aversion levels, $\{\theta \leq 0.35 \cup 0.55 \leq \theta \leq 0.8\}$. We conduct a Diebold and Mariano (1995) [DM] test to assess whether the combined forecast is significantly better than the best individual forecast over the 5 holdout sample years.[15] This confirms that, for several risk-

[15] We first compute DM_t, $t=1, \ldots, m$ ($m=5$), where $DM_t \overset{a}{\sim} N(0,1)$. The test statistic is $DM = \frac{1}{m}\sum_{t=1}^{m} DM_t$. Under independence between DM_t and DM_s for $t \neq s$, it follows that $DM \overset{a}{\sim} N\left(0, \frac{1}{m}\right)$.

Table 6
Ratio of the classifier's loss to the naïve predictor's loss 1996–2000

Classifiers	Risk-aversion level (θ)														
	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
LOGIT-M	0.967	0.930	0.826	0.798	0.773	0.709	0.648	0.511	0.555	0.622	0.687	0.860	0.954	1.074	1.369	1.913
LOGIT-R	1.036	0.924	0.865	0.750	**0.553***	**0.500***	**0.453***	0.495	0.545	0.677	0.728	0.774	0.862	1.065	**0.968**	**1.213**
Clustering	1.074	1.158	1.101	0.925	0.889	0.835	0.830	0.925	0.780	0.794	0.934	0.974	1.034	**0.947**	1.014	1.257
KK-logit	**0.886***	**0.827***	**0.745***	**0.701***	0.668	0.566	0.487	**0.462***	**0.530***	**0.582***	**0.646***	**0.727***	**0.853***	1.034	1.396	2.482

* Indicates that the overall accuracy (IL) of the best forecast, individual or combined, is significantly better than that of the naïve prediction at the 1% level on the basis of a Diebold–Mariano test. The naïve forecast is 0 for $\theta<0.5$, 1 for $\theta>0.5$ and the in-sample most frequent event for $\theta=0.5$. Bold denotes the best individual outcome.

aversion levels, it pays to combine the classifiers. For instance, for $\theta \in \{0.2, 0.25, 0.55, 0.65, 0.7, 0.75\}$, the minimal loss from the combined forecast is significantly smaller than that from any of the individual classifiers. The biggest improvements come at $\theta=0.25$ and $\theta=0.55$, with a loss differential between the individual and combined forecasts of 2.4% and 1.5%, respectively. Fig. 4 illustrates the central result of the paper graphically, namely, the gains from forecast combination at various risk-aversion levels. It shows that, overall, either the individual LOGIT-R forecasts or the combined forecast using the KK-logit scheme produce the best out-of-sample performance or minimum loss.

Finally, the forecasts are compared with naïve predictions. Take an uninformative, naïve model that always predicts 1 for $\theta>0.5$ (highly risk-averse decision-makers), 0 for $\theta<0.5$, and the most frequently observed event in-sample (here 0) for $\theta=0.5$. Table 6 reports the overall loss of each classifier (IL) relative to the naïve predictor's loss (IL^n). A DM test is conducted to compare the minimal loss among the classifiers, individual or combined, with that of the naïve predictor.

It turns out that for all risk-aversion levels the best forecasting method significantly outperforms the naïve predictor. The largest advantage comes from LOGIT-R for $\theta=0.5$ with the smallest ratio $\frac{IL}{IL^n}=0.453$, followed by the KK-logit combined forecast for $\theta=0.55$ with a ratio of 0.462. Interestingly, the gains of the best-performing model relative to the naïve predictor $\left(1-\frac{IL}{IL^n}\right)$ monotonically increase with θ up to $\theta=0.5$ and then decrease thereafter. This relates to the fact that the naïve forecast is always 0 for $\theta \leq 0.5$ and so IL^n rises with θ due to the increasing type I error penalty, whereas the naïve forecast is 1 for $\theta>0.5$ and so IL^n falls with θ.

7. Concluding remarks

An early warning system (EWS) for sovereign default provides a complementary tool to the analysis of decision-makers by facilitating objective measures of vulnerability. This paper investigates the optimal design of an EWS focusing on two aspects: the choice of the econometric methodology and the evaluation of the EWS itself. These two problems raise important issues which are seldom tackled in the literature. The analysis

is based on a sample of 75 emerging and developing economies 1983–2000. Forecasts are obtained from a logit regression (LOGIT-M) and K-means clustering, both based on macrovariables, and a logit regression based on internal-bank ratings (LOGIT-R). Clustering has not been used in this context to date.

The study has two main components. First, it incorporates the decision-maker's preferences (captured by a loss function and risk-aversion parameter) into the optimal calibration of the classifiers and the assessment of their out-of-sample forecasting properties. Second, it investigates forecast combining issues. For this purpose, a regression framework is adopted that exploits the distinct in-sample forecast ability of the individual methods. The issues of the objective function and forecast combination have received scant attention in the literature.

The results suggest that the decision-maker's preferences influence the choice of forecast methodology and its optimal calibration. LOGIT-M outperforms the non-parametric (clustering) and judgmental (LOGIT-R) classifiers by issuing fewer false alarms. But the latter two classifiers dominate LOGIT-M in missing fewer defaults. Moreover, the out-of-sample forecast ranking of the individual classifiers is unstable. These findings vindicate forecast combining. Both individual and naïve forecasts are outperformed by the combined forecasts for a range of risk-aversions. In this paper, the decision-maker's preferences have been formalized using loss functions that simply seek to reflect the desired trade-off between missed defaults and false alarms. Future work could extend this framework by aiming to better capture the economic or utility-based value of debt-crisis predictability.

Acknowledgement

This is a revised version of the Emerging Markets Group discussion paper 0404 and SSRN working paper 634063. We acknowledge the helpful comments of the editor, Michael Clements, an associate editor, two anonymous referees, Roy Batchelor, Chris Brooks, Jerry Coakley, Shelagh Heffernan, and Sotiris Staikouras. The paper has also benefited from the suggestions of Frank Packer at the *Bank of International Settlements*, participants at the 10th Meeting of the *Society for Computational Economics*, the 13th Meeting of the *Society for Nonlinear Dynamics and Econometrics*, the *Emerging Markets Finance* conference at Cass Business School, and a Bank of England seminar. We are responsible for any errors.

References

Abhyankar, A., Sarno, L., & Valente, G. (2005). Exchange rates and fundamentals: Evidence on the economic value of predictability. *Journal of International Economics*, *66*, 325−348.

Aiolfi, M. & Timmermann, A. (2003). Persistence in forecasting performance *Mimeo*, University of California at San Diego. Available at econweb.rutgers.edu.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, *20*, 451−468.

Berg, A., & Pattillo, C. (1999). Predicting currency crises: The indicators approach and an alternative. *Journal of International Money and Finance*, *18*, 561−586.

Breiman, L., Friedman, J. H., Olsen, E. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Bussière, M., & Fratzscher, M. (2002). Towards a new early warning system of financial crises. *European Central Bank Working Paper*, 02/145. Available at www.ecb.int.

Cantor, R., & Packer, F. (1996, October). Determinants and impact of sovereign credit ratings. *FRBNY Economic Policy Review*, 37−52.

Clemen, R. R, Winkler, R., & Murphy, A. (1995). Screening probability forecasts: Contrasts between choosing and combining. *International Journal of Forecasting*, *11*, 133−146.

Detragiache, E. & Spilimbergo, A. (2001). Crises and liquidity: Evidence and interpretation. *IMF Working Paper* 01/2.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 134−144.

Fair, R. C., & Shiller, R. J. (1990). Comparing information in forecasts from econometric models. *American Economic Review*, *80*, 375−389.

Frank, C. R., & Cline, W. R. (1971). Measurement of debt-servicing capacity: An application of discriminant analysis. *Journal of International Economics*, *41*, 327−344.

Frankel, J. A., & Rose, A. K. (1996). Currency crashes in emerging markets: An empirical treatment. *Journal of International Economics*, *41*, 351−366.

Frydman, H., Altman, E. I., & Kao, D. L. (1985). Introducting recursive partitioning for financial classification: The case of financial distress. *Journal of Finance*, *XL*, 269−291.

Fuertes, A. M., & Kalotychou, E. (2004). Towards the optimal design of an early warning system for sovereign debt crises, *EMG Working Paper 04-04.* Available at www.cass.city.ac.uk/emg

Fuertes, A. M. & Kalotychou, E. (in press-a). Early warning systems for sovereign debt crises: The role of heterogeneity. *Computational Statistics and Data Analysis*.

Fuertes, A. M. & Kalotychou, E. (in press-b). On sovereign credit migration: A study of alternative estimators and rating dynamics. *Computational Statistics and Data Analysis*.

Glennerster, R. (2004). *Transparency and standards: Evaluating the effect of institutions*. Unpublished PhD thesis, Birkbeck College, University of London.

Granger, C. W. J., & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*, 537−560.

Gupta, S., & Wilton, P. (1988). Combination of economic forecasts: An odds-matrix approach. *Journal of Business and Economic Statistics*, *6*, 373−379.

Kamin, S. B. (1999). The current international financial crisis: How much is new? *Journal of International Money and Finance*, *18*, 501−514.

Kaminsky, G., & Reinhart, C. M. (1999). The twin crises: The cause of banking and balance of payments problems. *American Economic Review*, *3*, 473−500.

Kaminsky, G. L., Lizondo, S., & Reinhart, C. M. (1998). Leading indicators of currency crises. *IMF Staff Papers*, *45*, 1−48.

Kamstra, M., & Kennedy, P. (1998). Combining qualitative forecasts using logit. *International Journal of Forecasting*, *14*, 83−93.

Kumar, M. S., Moorthy, U., & Perraudin, W. (2003). Predicting emerging market currency crashes. *Journal of Empirical Finance*, *10*, 427−454.

Lee, S. H. (1993). Are the credit ratings assigned by bankers based on the willingness of the LDC borrowers to repay? *Journal of Development Economics*, *40*, 349−359.

Manasse, P., Roubini, N., & Schimmelpfennig, A., (2003). Predicting sovereign debt crises. *IMF Working Paper* 03/221.

Mascarenhas, B., & Sand, O. C. (1989). Combination of forecasts in the international context: Predicting debt reschedulings. *Journal of International Business Studies*, *20*, 539−552.

McFadden, D., Eckaus, R., Feder, G., Hajivassiliou, V., & O'Connell, S. (1985). Is there life after debt? An econometric analysis of the creditworthiness of developing countries. In G. Smith & J. Cuddington (Eds.), *International Debt and the Developing Countries*. Washington: The World Bank.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., & Tiao, G. C. (1998). Forecasting the US unemployment rate. *Journal of the American Statistical Association*, *93*, 478−493.

Newbold, P., & Harvey, D. I. (2004). Forecast combination and encompassing. In M. P. Clements & D. F. Hendry (Eds.), *A companion to economic forecasting*. (pp. 268−283). UK: Blackwell.

Peter, M. (2002). Estimating default probabilities of emerging market sovereigns: A new look at a not-so-new literature, *HEI Working Paper* 02/6, Available at www.hei.unige.ch.

Reinhart, C. M. (2001). Default, currency crises and sovereign credit ratings. *World Bank Economic Review*, *16*, 151−170.

Rojas-Suárez, L. (2001). Rating banks in emerging markets, *Institute for International Economics Working Paper* 01/6. Available at www.ssrn.com.

Sommerville, R. A., & Taffler, R. J. (1995). Banker judgement versus formal forecasting models: The case of country risk assessment. *Journal of Banking and Finance*, *19*, 281−297.

Standard & Poor's (2001). Sovereign defaults decline through third-quarter 2000. In S&P's (Eds.), *Ratings Performance 2000*. Available at www.standardandpoors.com.

Stock, J. H., & Watson, M. (2001). A comparison of linear and non-linear univariate models for forecasting macroeconomic time series. In R. F. Engle & H. White (Eds.), *Cointegration, causality and forecasting: Festschrift in honour of Clive Granger*. (pp. 1−44). Oxford University Press.

Taffler, R. J, & Abassi, B. (1984). Country risk: A model for predicting debt-servicing problems in developing countries. *Journal of the Royal Statistical Society, Series A*, *147*, 541−568.

Winkler, R. L., & Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society, Series A*, *146*, 150−157.

Ana-Maria Fuertes (BSc Eng, MSc Control Eng, PhD Economics) is currently Reader in Econometrics at the Faculty of Finance, Cass Business School, City University, London. Her research interests are in time series analysis, panel data methods, forecasting and empirical finance. She has published in the *Journal of Economic Dynamics and Control, Economics Letters, Journal of International Money and Finance, Computational Statistics and Data Analysis* and *International Journal of Finance and Economics*.

Elena Kalotychou (BA/MA Cantab Maths, MSc Operational Research, PhD Finance) is currently Lecturer in Finance at the Faculty of Finance, Cass Business School, City University, London. Her research interests are in applied econometrics, credit risk, financial modelling and forecasting. She has published in the *Journal of Multinational Financial Management, Financial Markets Institutions and Instruments* and *Applied Economics*. Her work has been presented in seminars held at the *Bank of England* as well as international conferences.