Sarlin, Peter

Working Paper

On policymakers' loss function and the evaluation of early warning systems

ECB Working Paper, No. 1509

Provided in Cooperation with:
European Central Bank (ECB)

Suggested Citation: Sarlin, Peter (2013) : On policymakers' loss function and the evaluation of early warning systems, ECB Working Paper, No. 1509, European Central Bank (ECB), Frankfurt a. M.

This Version is available at:
https://hdl.handle.net/10419/153942

EUROPEAN CENTRAL BANK

EUROSYSTEM

WORKING PAPER SERIES

NO 1509 / FEBRUARY 2013

ON POLICYMAKERS' LOSS FUNCTIONS AND THE EVALUATION OF EARLY WARNING SYSTEMS

Peter Sarlin

In 2013 all ECB publications feature a motif taken from the €5 banknote.

MACROPRUDENTIAL RESEARCH NETWORK

Macroprudential Research Network

This paper presents research conducted within the Macroprudential Research Network (MaRs). The network is composed of economists from the European System of Central Banks (ESCB), i.e. the 27 national central banks of the European Union (EU) and the European Central Bank. The objective of MaRs is to develop core conceptual frameworks, models and/or tools supporting macro-prudential supervision in the EU.

The research is carried out in three work streams: 1) Macro-financial models linking financial stability and the performance of the economy; 2) Early warning systems and systemic risk indicators; 3) Assessing contagion risks.

MaRs is chaired by Philipp Hartmann (ECB). Paolo Angelini (Banca d'Italia), Laurent Clerc (Banque de France), Carsten Detken (ECB), Cornelia Holthausen (ECB) and Katerina Šmídková (Czech National Bank) are workstream coordinators. Xavier Freixas (Universitat Pompeu Fabra) and Hans Degryse (Katholieke Universiteit Leuven and Tilburg University) act as external consultant. Angela Maddaloni (ECB) and Kalin Nikolov (ECB) share responsibility for the MaRs Secretariat.

The refereeing process of this paper has been coordinated by a team composed of Cornelia Holthausen, Kalin Nikolov and Bernd Schwaab (all ECB).

The paper is released in order to make the research of MaRs generally available, in preliminary form, to encourage comments and suggestions prior to final publication. The views expressed in the paper are the ones of the author(s) and do not necessarily reflect those of the ECB or of the ESCB.

Acknowledgements

The paper is a longer version of a paper forthcoming in Economics Letters. The author wants to thank two anonymous referees, Barbro Back, Andrew Berg, Frank Betz, Bertrand Candelon, Carsten Detken, Elena Dumitrescu, Marco Lo Duca, Silviu Oprica, Tuomas Peltonen, Samuel Rönnqvist, Pierre-Daniel Sarte, Gregor von Schweinitz, and seminar participants at the ECB Financial Stability seminar on 16 May 2012 in Frankfurt am Main for useful comments, discussions and assistance. The financial support of the Academy of Finland (grant no. 127592) and the hospitality of the ECB DG Financial Stability, where part of this work was completed, is gratefully acknowledged. The views presented in the paper are those of the author and do not necessarily represent the views of the European Central Bank or the Eurosystem.

Peter Sarlin

Åbo Akademi University; e-mail: psarlin@abo.fi

Address	Kaiserstrasse 29, 60311 Frankfurt am Main, Germany
Postal address	Postfach 16 03 19, 60066 Frankfurt am Main, Germany
Telephone	+49 69 1344 0
Internet	http://www.ecb.europa.eu
Fax	+49 69 1344 6000

ISSN	1725-2806 (online)
EU Catalogue No	QB-AR-13-006-EN-N (online)

Abstract

This paper introduces a new loss function and Usefulness measure for evaluating early warning systems (EWSs) that incorporate policymakers' preferences between issuing false alarms and missing crises, as well as individual observations. The novelty derives from three enhancements: *i*) accounting for unconditional probabilities of the classes, *ii*) computing the proportion of available Usefulness that the model captures, and *iii*) weighting observations by their importance for the policymaker. The proposed measures are model free such that they can be used to assess signals issued by any type of EWS, such as logit and probit analysis and the signaling approach, and flexible for any type of crisis EWSs, such as banking, debt and currency crises. Applications to two renowned EWSs, and comparisons to two commonly used evaluation measures, illustrate three key implications of the new measures: *i*) further highlights the importance of an objective criterion for choosing a final specification and threshold value, and for models to be useful *ii*) the need to be more concerned about the rare class and *iii*) the importance of correctly classifying observations of the most relevant entities. Beyond financial stability surveillance, this paper also opens the door for cost-sensitive evaluations of predictive models in other tasks.

JEL Codes: E44, E58, F01, F37, G01.

Keywords: Early warning systems, policymakers' preferences, misclassification costs

Non-technical summary

The high real costs of crises have stimulated research on predicting financial instabilities. These models are oftentimes called Early Warning Systems (EWSs). There is a broad literature on predicting vulnerabilities prior to various types of crises, such as currency crises (e.g. Kaminsky *et al*., 1998), debt crises (e.g. Manasse *et al*. 2003) and banking crises (e.g. Demirgüç-Kunt and Detragiache, 2000). More recent EWSs have focused on broader, systemic financial crises (e.g. Lo Duca and Peltonen, 2013).

Rather than adding yet another model to the already extensive list of EWSs, this paper has its core in an evaluation and validation of EWSs tailored to the needs for policymaking and the properties of the underlying data. Hence, it is of central importance that a policymaker designing an EWS acknowledges that financial crises are oftentimes outlier events in three aspects: *i*) dynamics at the time of crisis differ significantly from tranquil times, *ii*) crises are commonly more costly and *iii*) crises occur more rarely.

The literature on evaluations has attempted to design a policymaker's loss-function that measures the loss for a policymaker of an EWS in a two-class setting (tranquil vs. crisis), in which the loss generally stems from false alarms and missed crises. Demirgüç-Kunt and Detragiache (2000) introduced the concept of a policymaker's loss-function that describes mainly a trade-off between Type 1 and Type 2 errors (probability of not receiving a warning conditional on a crisis occurring and of receiving a warning conditional on no crisis occurring). Adaptations of this type of loss functions have been introduced to EWSs for other types of crisis, e.g. debt crises (Fuertes and Kalotychou, 2007), currency crises (Bussière and Fratzscher, 2008), and asset price boom/bust cycles (Alessi and Detken, 2011). In addition to a loss function, Alessi and Detken (2011) also propose a Usefulness measure that compares the loss of the model to the loss of disregarding the model. The model is Useful, if the loss of the model is smaller than the loss of disregarding it. However, while the above evaluation frameworks have become state-of-the-art, they fail to account for characteristics of imbalanced data. Rather than the share of errors in relation to class size, the relevant measure for a policymaker to be concerned about is the absolute number of errors. Thus, assuming that tranquil and pre-crisis periods are of similar frequency imposes not only a bias on the weighting of type 1 and type 2 errors, but also on the derived Usefulness measure, as a best guess of always or never signaling when disregarding a model is highly affected by the frequency of the classes.

Another shortcoming of EWS evaluations stems from the fact that the models oftentimes utilize pooled panel data, i.e. data with both a cross-sectional and time dimension. This can be motivated by the relatively small number of crisis events in individual countries and by the strive to capture a wide variety of crisis types, as well as a global policy approach. However, while pooling panel data may be motivated, the importance of a single country in the evaluation phase may vary. In an evaluation framework, this leads to a need for weighting observations in terms of their importance, such as systemic relevance.

To this end, this paper proposes a new policymaker's loss function and Usefulness measure that depend on unconditional probabilities, as well as misclassification costs of not only classes but also observations. The novelty of the measures derives from three enhancements: *i*) by accounting for unconditional probabilities of the classes, *ii*) by computing the proportion of available Usefulness that the model captures, and *iii*) by weighting observations by their importance for the

policymaker. The new policymaker's loss function and the Usefulness measures are illustrated and compared to commonly used evaluation measures on replicas of a seminal EWS by Berg and Pattillo (1999b) and a recent one by Lo Duca and Peltonen (2013). The measures and experiments highlight three key implications: *i*) further accentuate the importance of an objective criterion for choosing a final specification and threshold value, and for models to be Useful *ii*) the need to be more concerned about the rare class and *iii*) the importance of correctly classifying observations of the most relevant entities as defined by the policymaker.

1. Introduction

The still ongoing financial crisis, and crises' generally high costs to welfare and economic growth (Cardiarelli *et al*., 2011), have stimulated research on predicting financial instabilities. Rather than attempting to predict the exact timings of crises and the triggers that lead to a crisis, the early warning system (EWS) literature focuses on detecting vulnerabilities prior to a crisis. The seminal EWSs focused mainly on predicting vulnerabilities prior to currency crises (Frankel and Rose, 1996; Kaminsky *et al*., 1998; Berg and Pattillo, 1999a). While the literature at the turn of the century concentrated its attention on debt and banking crises (Fuertes and Kalotychou, 2006; Manasse *et al*. 2003; Demirgüç-Kunt and Detragiache, 2000), more recent EWSs have focused on broader, systemic financial crises (Alessi and Detken, 2011; Sarlin and Peltonen, 2011; Lo Duca and Peltonen, 2013).

Rather than adding yet another model to the already extensive list of EWSs, this paper has its core in the explicit forecasting objectives and validations of EWSs tailored to the needs for policymaking and the complex nature of the problem. While a particular strand of literature has focused on the evaluation of EWSs, the utilized measures seldom cover the wide spectrum of factors that may concern a policymaker. The seminal study by Kaminsky *et al*. (1998) utilized the noise-to-signal ratio, a simple ratio of the probability of receiving a signal conditional on no crisis occurring to the probability of receiving a signal conditional on a crisis occurring, to set an optimal threshold value.[1] While the comprehensive toolbox for evaluating EWSs by Candelon *et al*. (2012) provides significant contributions to statistical inference for testing the superiority of one EWS over another, they lack an explicit focus on variations in misclassification costs and imbalanced data.[2] A crucial characteristic of measures attempting to grasp a problem of this order of complexity is to explicitly tailor forecasting objectives and validations to the preferences of a decision-maker and the properties of the underlying data. Hence, it is of central importance that a policymaker designing an EWS acknowledges that financial crises are oftentimes outlier events in three aspects: *i*) dynamics at the time of crisis differ significantly from tranquil times, *ii*) crises are commonly more costly and *iii*) crises occur more rarely.

While the first of the above three aspects has been tackled with a number of enhancements to EWS specifications and modeling, such as a crisis and post-crisis bias (Bussière and Fratzscher, 2006), a literature on the derivation of a policymaker's loss-function has focused on the two latter particularities of crisis data, i.e. so-called low-probability, high-impact events. Demirgüç-Kunt and Detragiache (2000) introduced the notion of a policymaker's loss-function in a banking crisis context, where the policymaker has a cost for preventive actions and type 1 and type 2 errors (probability of not receiving a warning conditional on a crisis occurring and of receiving a warning conditional on no crisis occurring). Later, adaptations of this type of loss functions have been

[1] Demirgüç-Kunt and Detragiache (2000) and El-Shagi *et al*. (2012) showed that minimizing the noise-to-signal ratio could lead to a relatively high share of missed crisis episodes (or only noise minimization) if crises are rare and the cost of missing a crisis is high. Lund-Jensen (2012) concludes the same, and chooses not to use the measure, while Drehman *et al*. (2011) choose to minimize the noise-to-signal-ratio subject to at least two thirds of the crises being correctly called.

[2] Based upon Receiver Operating Characteristics (ROC) curves and the area below them, measures applied by Sarlin and Marghescu (2011) to EWS evaluations, Jordà and Taylor (2011) formulated and Jordà *et al*. (2011) applied a correct classification frontier (CCF) with advantages like providing visual means and summarizations of results for all possible thresholds. In practice, it is a plot of the share of correct alarms to the share of false alarms for all threshold values. Yet, the measures do not properly pay regard to varying misclassification costs and imbalanced data, and suffer from the fact that all thresholds may be far from policy relevant (e.g. both ends of the CCF).

introduced to EWSs for other types of crisis, e.g. debt crises (Fuertes and Kalotychou, 2007), currency crises (Bussière and Fratzscher, 2008), and asset price boom/bust cycles (Alessi and Detken, 2011). While Bussière and Fratzscher still focused on costs of preventive actions, the later literature has mainly focused on the trade-off between type 1 and 2 errors. There are two key motivations for focusing on relative preferences between the errors: *i*) the possibility of incorporating the costs of actions and no actions in preferences between type 1 and 2 errors as unrealized benefits can be "rolled up" into error costs (Fawcett, 2006), and *ii*) the uncertainty of exact costs associated with preventive actions, false alarms and missing crises. In addition to a loss function, Alessi and Detken (2011) also propose a Usefulness measure that indicates whether the loss of the prediction is smaller than the loss of disregarding the model. However, while the above evaluation frameworks have become state of the art, they fail to account for characteristics of imbalanced data.[3] Rather than the share of misclassifications in relation to class size (i.e., type 1 and 2 errors), the relevant measure for a policymaker to be concerned about is the absolute number of errors. Assuming that tranquil and pre-crisis periods are of similar frequency by relating them to class size imposes a bias on the weighting of type 1 and 2 errors in the loss function. Likewise, as a best guess of always or never signaling when disregarding a model is highly affected by the frequency of the classes, a bias is also introduced in previous Usefulness measures, where the loss of disregarding a model only depends upon the preferences between type 1 and 2 errors.

Another shortcoming of EWS evaluations stems from the fact that the models oftentimes utilize pooled panel data. Generally, results indicate that accounting for country-specific and time-specific effects lead to an improved in-sample fit, while it decreases predictive performance on out-of-sample data (e.g. Fuertes and Kalotychou, 2006). This can be further motivated by the relatively small number of crisis events in individual countries and by the striving to capture a wide variety of crisis types, as well as the requirement of a global policy approach. However, while these conditions motivate using pooled panel data, the importance of a single country in the evaluation phase may vary depending on the objectives of the policymaker. In an evaluation framework, this leads to a need for weighting entities in terms of their importance, such as systemic relevance or size. The entity-level importance is, however, also a time-varying parameter, and should thus more preferably be defined on the observation level.

The main innovation of the present paper is a policymaker's loss function and a Usefulness measure that both depend upon unconditional probabilities of classes. The absolute Usefulness measure is further augmented by also computing the proportion of available Usefulness that the model captures, a concept we coin as relative Usefulness, rather than only providing a Usefulness number difficult to judge. To address the observation-varying importance in evaluations, a third contribution of this paper is to introduce misclassification costs that are observation-specific, in addition to preferences between the classes. Hence, we extend previous measures in three aspects: *i*) by accounting for unconditional probabilities of the classes, *ii*) by computing the proportion of available Usefulness that the model captures, and *iii*) by weighting individual observations by their importance for the policymaker. The new policymaker's loss function and the Usefulness measures are illustrated and compared to commonly used evaluation measures on replicas of a seminal EWS

[3] While the seminal loss function of Demirgüç-Kunt and Detragiache (2000) accounts for unconditional probabilities, they do not propose a Usefulness measure for the function. Given their complex definition of loss, deriving the Usefulness would not be an entirely straightforward exercise. Further, the version applied in Bussière and Fratzscher (2008) neither accounts for unconditional probabilities nor distinguishes between losses from correct and wrong calls of crisis..

by Berg and Pattillo (1999b) and a recent one by Lo Duca and Peltonen (2013). The measures and experiments highlight three key implications: *i*) further accentuate the importance of an objective criterion for choosing a final specification and threshold value, and for models to be Useful *ii*) the need to be more concerned about the rare class and *iii*) the importance of correctly classifying observations of the most relevant entities as defined by the policymaker.

The paper is organized as follows. First, we present the new policymaker's loss function and Usefulness measures. The measures are derived for policymakers of three kinds: a cost-ignorant policymaker, a cost-aware policymaker with fixed but unequal preferences and a cost-aware policymaker with observation-specific costs. Second, we apply the policymaker's loss function and the Usefulness measures to a seminal EWS by Berg and Pattillo (1999b) and a recent one by Lo Duca and Peltonen (2013). Finally, we summarize the key findings and their implications for designing EWSs.

2. A policymaker's loss function and Usefulness measure

Crisis data require evaluation criteria that account for their complex nature. Crises are oftentimes outlier events in three aspects: *i*) they differ significantly from tranquil times, *ii*) they are commonly more costly and *iii*) they occur more rarely. Given these properties, especially the two latter ones, we put forward an evaluation framework that resembles the decision problem faced by a policymaker. We first discuss a general framework for deriving a policymaker's loss function and the Usefulness of a model, and then provide variations depending on the costs for the policymaker.

The occurrence of a crisis can be represented with a binary state variable $I_j(0) \in \{0,1\}$ (where observation j=1,2,…,N). Signaling the contemporaneous occurrence of distress does not, however, provide enough reaction time for a policymaker. The wide variety of triggers may also complicate the task of identifying exact timings. To enable policy actions for decreasing further build up of vulnerabilities and strengthening the financial system, the focus should rather be on identifying pre-crisis periods $I_j(h) \in \{0,1\}$ with a specified forecast horizon h. Let $I_j(h)$ be a binary indicator that equals one during pre-crisis periods and zero otherwise. Using univariate or multivariate data, various methods can be used for turning indicators into estimated probabilities of an impending crisis $p_j \in [0,1]$ (probability forecasts). To mimic the ideal leading indicator $I_j(h)$, the probability p_j is transformed into a binary point forecast P_j that equals one if p_j exceeds a specified threshold λ and zero otherwise. The correspondence between P_j and I_j can be summarized into a so-called contingency matrix (frequencies of prediction-realization combinations).

		Actual class I_j	
		Crisis	No crisis
Predicted class P_j	Signal	**A** *True positive*	**B** *False positive*
	No signal	**C** *False negative*	**D** *True negative*

From the elements of the above matrix, one can then define various goodness-of-fit measures. We approach the problem from the viewpoint of a policymaker.[4] In a two-class prediction problem, policymakers can be assumed to have relative preferences of conducting two types of errors: issuing false alarms and missing crises. Type 1 errors represent the probability of not receiving a warning conditional on a crisis occurring $T_1 = P\left(p_j \le \lambda \middle| I_j(h) = 1\right)$ and type 2 errors the probability of receiving a warning conditional on no crisis occurring $T_2 = P\left(p_j > \lambda \middle| I_j(h) = 0\right)$. Given probabilities p_j of a model, the policymaker should focus on choosing a threshold λ such that her loss is minimized. The loss of a policymaker consists of T_1 and T_2 weighted according to her relative preferences between missing crises $(\mu \in [0,1])$ and issuing false alarms $(1-\mu)$. The preference parameters may also be derived from a benefit/cost matrix that matches the contingency matrix. A standard 2x2 benefit/cost matrix may easily be manipulated to only include error costs by scaling and shifting entries of columns without affecting the decisions (Elkan, 2001; Fawcett, 2006). A

[4] A further discussion on shaping decision-makers' problems through loss functions, as well as on the relation between statistical and economic value of predictions, can be found in Granger and Pesaran (2000) and Abhyankar et al. (2005).

benefit may be treated as a negative error cost and hence unrealized benefits can be "rolled up" into error costs. For instance, the costs c for the elements of the matrix with two degrees of freedom can be reduced to a simpler matrix of class-specific costs c_1 and c_2 with one degree of freedom: $c_1 = c_C - c_A$ and $c_2 = c_B - c_D$. Most likely, c_B and c_C have a non-negative cost, while c_A and c_D have a non-postive cost. From this, we can derive the relative preferences $\mu = c_1/(c_1 + c_2)$ and $1-\mu = c_2/(c_1 + c_2)$. However, when only using T_1 and T_2 weighted according to relative preferences, we fail to account for imbalances in class size.[5] By accounting for unconditional probabilities of crises $P_1 = P(I_j(h) = 1)$ and tranquil periods $P_2 = P(I_j(h) = 0) = 1 - P_1$, a loss function can be written as follows:

$$L(\mu) = \mu T_1 P_1 + (1-\mu) T_2 P_2 \qquad (1)$$

As the parameters are unknown *ex ante*, we can use in-sample frequencies to estimate them. Given a threshold λ and forecast horizon h, P_1 and P_2 are estimated with the frequency of the classes ($P_1 = (A+C)/(A+B+C+D)$ and $P_1 = (B+D)/(A+B+C+D)$) and T_1 and T_2 with the error rates ($T_1 \in [0,1] = C/(A+C)$ and $T_2 \in [0,1] = B/(B+D)$). Using the loss function $L(\mu)$, we can then define the Usefulness of a model. A policymaker could achieve a loss of $\min(P_1, P_2)$ by always issuing a signal of a crisis if $P_1 > 0.5$ or never issuing a signal if $P_2 > 0.5$. However, by weighting with policymakers' preferences, as she may be more concerned about one of the classes, we achieve the loss $\min(\mu P_1, P_2(1-\mu))$ when ignoring the model. This differs from Alessi and Detken (2011), as they only account for policymakers' preferences $(\min(\mu, 1-\mu))$. First, we derive the absolute Usefulness $U_a(\mu)$ of a model by computing the loss generated by the model subtracted from the loss of ignoring it:

$$U_a(\mu) = \min(\mu P_1, P_2(1-\mu)) - L(\mu). \qquad (2)$$

This measure highlights the fact that achieving well-performing, useful models on highly imbalanced data is a difficult task. It is also worth noting that already an attempt to build an EWS with imbalanced data implicitly necessitates a policymaker to be more concerned about the rare class. With a non-perfectly performing model, it would otherwise easily pay off for the policymaker to always signal the high-frequency class. Second, we compute the share of $U_a(\mu)$ to the maximum possible Usefulness of the model with a measure that we coin relative Usefulness:

[5] The loss function used by Alessi and Detken (2011) differs from the one introduced here as it assumes equal class size. Their Usefulness measure does, similarly, not account for imbalanced classes, as the loss of disregarding a model depends solely on the preferences. Usefulness measures close to that in Alessi and Detken (2011) have been applied in a large number of works, such as Lo Duca and Peltonen (2012), Sarlin and Peltonen (2011), Sarlin and Marghescu (2011) and El-Shagi *et al*. (2012). Similar loss functions have been applied in Fuertes and Kalotychou (2007), Candelon *et al*. (2012) and Lund-Jensen (2012).

$$U_r(\mu) = \frac{U_a(\mu)}{\min(\mu P_1, P_2(1-\mu))}. \quad (3)$$

That is, $U_r(\mu)$ reports $U_a(\mu)$ as a percentage of the Usefulness that a policymaker would gain with a perfectly performing model. This derives from the fact that if $L(\mu) = 0$ then $U_a(\mu) = \min(\mu P_1, P_2(1-\mu))$. The $U_r(\mu)$ provides means for representing the Usefulness as a ratio rather than only reporting a number difficult to judge. In particular, it facilitates comparisons of models for policymakers with different preferences.

Within the above framework, we can generate policymakers of different kinds depending on their preferences. Below, we provide loss functions and Usefulness measures for policymakers of three kinds: a cost-ignorant policymaker, a cost-aware policymaker with fixed but unequal preferences and a cost-aware policymaker with observation-specific costs.

2.1. Cost-ignorant policymaker

A cost-ignorant policymaker assumes the cost of missing a crisis and issuing a false alarm to be equal. This leads to the more frequent class being the best guess of the policymaker when disregarding the model. By setting $\mu = 0.5$, we can use equations (1)-(3) for computing $L(\mu)$, $U_a(\mu)$ and $U_r(\mu)$. The preferences of this policymaker resemble the preferences of a policymaker that uses the noise-to-signal ratio (Kaminsky *et al*., 1998), in which type 1 and 2 errors are weighted equally. However, in domains with low-probability, high-impact events, such as settings for EWSs generally are, one would assume a policymaker to also have imbalanced misclassification costs.

2.2. Fixed cost-aware policymaker

A fixed cost-aware policymaker has fixed misclassification costs for all observations, but may want to set them unequally. By varying the preference parameter $\mu \in [0,1]$, she can set the preferences to approximate the cost of misclassifying one class relative to the other. Again, by inserting μ into equations (1)-(3), we can derive $L(\mu)$, $U_a(\mu)$ and $U_r(\mu)$ for the given preferences μ. While introducing the unconditional probabilities to equations (1) and (2), the preferences between type 1 and 2 errors in $L(\mu)$ and $U_a(\mu)$ of this policymaker resemble those in Alessi and Detken (2011).

2.3. Flexible cost-aware policymaker

A flexible cost-aware policymaker may still want to augment the specification by accounting for observation-specific differences in costs. In the case of EWSs for financial crises, the entity-level misclassification costs are highly related to the systemic or contagious relevance of an entity.[6] Let

[6] While relevance is related to interconnectedness, such as various network measures, a simplified indicator of relevance for the system is the size of the entity (e.g. assets for a bank or stock-market capitalization for an economy). From a utilitarian perspective, one could also weight Usefulness based upon population such that average Usefulness per capita would be maximized. Another perspective would be that of a national central bank that weights by threats to the domestic real economy, where the largest weight is on correctly signaling events in the home country while the

w_j be an observation-specific weight that approximates the importance of correctly calling observation j specified by the policymaker and let A_j, B_j, C_j and D_j be binary vectors of combinations of predicted and actual classes rather than only their sums. By multiplying each binary element of the contingency matrix by w_j, the elements of the contingency matrix become importance-weighted sums which may used similarly as non-weighted sums for computing T_1 and T_2. Thus, following the above framework, we can derive a policymakers' loss function with observation and class-specific misclassification costs. Let the elements of T_1 and T_2 be weighted by w_j to have weighted type 1 and 2 errors:

$$T_{w1} = \sum_{j=1}^{N} w_j C_j \Big/ \left(\sum_{j=1}^{N} w_j A_j + \sum_{j=1}^{N} w_j C_j \right), \tag{4}$$

$$T_{w2} = \sum_{j=1}^{N} w_j B_j \Big/ \left(\sum_{j=1}^{N} w_j B_j + \sum_{j=1}^{N} w_j D_j \right) \tag{5}$$

As $T_{w1} \in [0,1]$ and $T_{w2} \in [0,1]$ are ratios of sums of weights rather than sums of binary values, they can replace T_1 and T_2 in equations (1)-(3). Hence, by inserting μ, T_{w1}, T_{w2}, P_{w1} and P_{w2} into equations (1)-(3), we can derive the measures $L(\mu, w_j)$, $U_a(\mu, w_j)$ and $U_r(\mu, w_j)$ for given preferences μ and weights w_j.

importance of neighbors would be diminishing based upon some measure of closeness. It is worth to note that defining the importance of an observation in relation to the cross section is not always, however, an easy task.

3. The Usefulness of two EWSs

This section computes the new loss functions and Usefulness measures, as well as two commonly used measures, for two well-known EWSs and tests to which extent and for which policymakers' preferences they yield Usefulness. In the evaluations, we use the non-weighted and weighted Usefulness measures introduced in Section 2. We replicate the EWS for currency crises in Berg and Pattillo (1999b) (henceforth BP) and the EWS for systemic financial crises in Lo Duca and Peltonen (2013) (henceforth LDP).

3.1 The BP model

The probit model in BP uses monthly data for 23 emerging market economies for the period from 1986:1–1998:12.[7] The explanatory variables are the following five indicators: foreign exchange reserve loss, export loss, real exchange-rate overvaluation relative to trend, current account deficit as a percentage of GDP, and short-term debt in relation to reserves. The indicators and their transformations, and the countries in the dataset are shown in Table A in the Appendix. The predicted variable is a binary indicator that indicates the occurrence of a currency crisis within 24 months. The occurrence of a crisis $I_j(0) \in \{0,1\}$ is defined as the sum of a weighted average of monthly percentage depreciation in the exchange rate and monthly percentage declines in reserves exceeds its mean by more than three standard deviations (weighted so that the variances of the two components are equal).[8] Pre-crisis periods $I_j(h) \in \{0,1\}$, where h=24, can then be defined as the 24 months preceding $I_j(0)$.

For computing the weighted Usefulness measures, data on country-specific population is used. We assign the weight w_j to be the share of population of country i in period t of the sum of population in the sample in period t. This takes a utilitarian perspective of an external observer interested in maximizing average Usefulness per capita. To test the predictive performance of the model on the Asian financial crisis of the late 1990s, the dataset is divided into in-sample data (1986:1–1995:4) and out-of-sample data (1995:5–1996:12). The estimates of the replicated probit regression shown in Table 1 refer to in-sample data and are applied to the out-of-sample data. As in BP, the probit estimates are used for generating probabilities p_j and transformed into a binary point forecast P_j that equals one if p_j exceeds a threshold λ and zero otherwise. Here, however, the λ is set to optimize in-sample absolute Usefulness $U_a(\mu)$ for given preferences, and then later applied to the out-of-sample data.

[7] The model in BP (Berg and Pattillo, 1999b) is a simplified specification of that in Berg and Pattillo (1999a) in terms of explanatory variables. We use this parsimonious version for two reasons: it was found to be more accurate and the focus is on the probabilities p_j rather than what method and data lie behind them.

[8] The definition of a crisis is originally from Kaminsky *et al*. (1998).

Table 1. The estimates for the BP model

Variables	*Estimates*	*Std. error*	*Z*	*Sig.*
Intercept	-2.475	0.132	-18.715	0.000 ***
Reserve loss	0.007	0.001	5.608	0.000 ***
Export loss	0.002	0.001	1.485	0.138
Real exchange rate deviation	0.001	0.001	3.964	0.000 ***
Current account deficit to GDP	0.011	0.001	8.903	0.000 ***
Short-term debt to reserves	0.004	0.001	3.541	0.000 ***

Notes: Significance levels: 1%, ***; 5 %, **; 10 %, *.

Table 2. In-sample (a) and out-of-sample (b) performance of the BP model for policymakers' preferences $\mu = 0.0,0.1,...,1.0$

a) In-sample performance

Preferences	λ	*TP*	*FP*	*TN*	*FN*	T_1	T_2	*Accuracy*	$U_a(\mu)$	$U_r(\mu)$	$U_a(\mu,w_j)$	$U_r(\mu,w_j)$
μ=0.0	0.48	13	0	2104	357	96.49 %	0.00 %	85.57 %	0.00	NA	0.00	NA
μ=0.1	0.48	13	0	2104	357	96.49 %	0.00 %	85.57 %	0.00	3.51 %	0.00	4.54 %
μ=0.2	0.48	13	0	2104	357	96.49 %	0.00 %	85.57 %	0.00	3.51 %	0.00	4.54 %
μ=0.3	0.48	13	0	2104	357	96.49 %	0.00 %	85.57 %	0.00	3.51 %	0.00	4.54 %
μ=0.4	0.44	24	5	2099	346	93.51 %	0.24 %	85.81 %	0.00	4.46 %	0.00	5.31 %
μ=0.5	0.43	27	7	2097	343	92.70 %	0.33 %	85.85 %	0.00	5.41 %	0.01	6.74 %
μ=0.6	0.38	57	41	2063	313	84.59 %	1.95 %	85.69 %	0.01	8.02 %	0.02	20.41 %
μ=0.7	0.29	107	153	1951	263	71.08 %	7.27 %	83.19 %	0.01	11.20 %	0.04	37.32 %
μ=0.8	0.20	216	416	1688	154	41.62 %	19.77 %	76.96 %	0.04	30.27 %	0.05	41.62 %
μ=0.9	0.14	288	854	1250	82	22.16 %	40.59 %	62.17 %	0.02	24.33 %	0.02	29.38 %
μ=1.0	0.01	370	2104	0	0	0.00 %	100.00 %	14.96 %	0.00	NA	0.00	NA

b) Out-of-sample performance

Preferences	λ	*TP*	*FP*	*TN*	*FN*	T_1	T_2	*Accuracy*	$U_a(\mu)$	$U_r(\mu)$	$U_a(\mu,w_j)$	$U_r(\mu,w_j)$
μ=0.0	0.48	5	0	321	115	95.83 %	0.00 %	73.92 %	0.00	NA	0.00	NA
μ=0.1	0.48	5	0	321	115	95.83 %	0.00 %	73.92 %	0.00	4.17 %	0.00	0.00 %
μ=0.2	0.48	5	0	321	115	95.83 %	0.00 %	73.92 %	0.00	4.17 %	0.00	0.00 %
μ=0.3	0.48	5	0	321	115	95.83 %	0.00 %	73.92 %	0.00	4.17 %	0.00	0.00 %
μ=0.4	0.44	19	1	320	101	84.17 %	0.31 %	76.87 %	0.01	13.18 %	0.00	8.29 %
μ=0.5	0.43	24	1	320	96	80.00 %	0.31 %	78.00 %	0.01	18.23 %	0.01	14.46 %
μ=0.6	0.38	43	6	315	77	64.17 %	1.87 %	81.18 %	0.03	28.75 %	0.02	26.64 %
μ=0.7	0.29	84	41	280	36	30.00 %	12.77 %	82.54 %	0.04	38.87 %	0.05	47.52 %
μ=0.8	0.2	93	111	210	27	22.50 %	34.58 %	68.71 %	0.03	28.34 %	0.03	28.88 %
μ=0.9	0.14	110	205	116	10	8.33 %	63.86 %	51.25 %	0.02	22.95 %	0.02	21.03 %
μ=1.0	0.01	120	321	0	0	0.00 %	100.00 %	27.21 %	0.00	NA	0.00	NA

Notes: The in-sample dataset spans from 1986:1–1995:4 and the out-of-sample dataset from 1995:5–1996:12. Models for each μ set λ to optimize in-sample absolute Utility $U_a(\mu)$, while the same λ is applied to the out-of-sample data. The abbreviations are as follows: λ, threshold; *TP*, true positives; *FP*, false positives; *TN*, true negatives; *FN*, false negatives; T_1, type 1 errors; T_2, type 2 errors; $U_a(\mu)$, absolute Usefulness; $U_r(\mu)$, relative Usefulness; $U_a(\mu,w_j)$, weighted absolute Usefulness; $U_r(\mu,w_j)$, weighted relative Usefulness. *Accuracy* refers to $(TP+TN)/(TP+TN+FP+FN)$. The weights w_j represent the share of population of country i in period t to the sum of population of the sample in period t.

The performance of the BP model for $\mu = 0.0,0.1,...,1.0$ is shown in Table 2. The results in Table 2a refer to in-sample performance with a λ that optimizes absolute Usefulness $U_a(\mu)$ for the

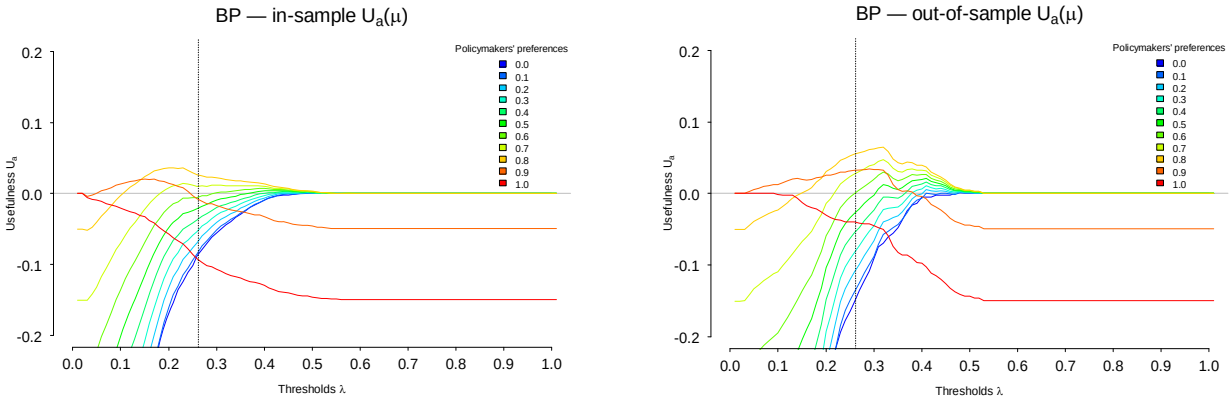
given preferences, while Table 2b shows the out-of-sample performance per μ for the λ specified in Table 2a. The tables show that in general the optimal thresholds λ decrease with increases in preferences μ, implying more signals with larger μ. This reflects the fact that when the loss of missing crises is larger, then it is optimal to signal more. This is also shown by the variation of T_1 and T_2 for $\mu = 0.0,0.1,...,1.0$. The in-sample results show that the relative Usefulness $U_r(\mu)$ is above 10% only for $\mu = [0.7,0.9]$, while out-of-sample performance fulfils the same criterion for $\mu = [0.4,0.9]$. Another general result of the performance table is that results for weighted Usefulness $U_a(\mu, w_j)$ and $U_r(\mu, w_j)$ are better than those for non-weighted on in-sample data, while the out-of-sample results are in general poorer. As the Usefulness per economy (non-weighted) is shown to be larger than the per capita results (weighted), the predictive performance is shown to be poorer for high-populated economies (except for $\mu = 0.7,0.8$).

Figure 1a shows $U_a(\mu)$ and Figure 1b shows $U_r(\mu)$ for the BP model with $\lambda \in [0,1]$ and $\mu = 0.0,0.1,...,1.0$. The vertical line represents the threshold $\lambda = 0.262$ that BP somewhat arbitrarily applied.[9] The in-sample $U_a(\mu)$ and $U_r(\mu)$ illustrate that for most μ, an optimal λ would have been lower than that in BP, while the out-of-sample Usefulness is higher for larger λ. In Figure 1c, we also assess optimal calibration that the noise-to-signal ratio, measured by $NtS = T_2/(1-T_1)$, would lead to, where the vertical line represents the optimal λ. Our findings corroborate those in e.g. Lund-Jensen (2012), Drehman *et al*. (2011) and El-Shagi *et al*. (2012) as the noise-to-signal ratio leads to an optimal calibration where very few crises are correctly called $(\lambda = 0.48)$. Further, as the λ chosen by BP is shown not to be optimal for any μ, we have to consider for what μ the λ would have been closest to optimum. To this end, Table 3 shows Usefulness values for three different specifications: *i*) μ for which $U_a(\mu)$ is closest to its optimum with the λ pre-defined by BP, *ii*) optimal λ for the previously defined μ, and *iii*) optimal λ for a policymaker with so-called balanced preferences $\mu = 1 - P_1 = 0.8$. The results in Tables 3a-b imply that the λ specified by BP is useful, but could still be improved by optimizing λ. While Figures 1a and 1b may advocate setting $\mu = 0.8$ to have a large in-sample Usefulness, Table 3c shows that its out-of-sample performance is significantly worse than that of the BP specification.

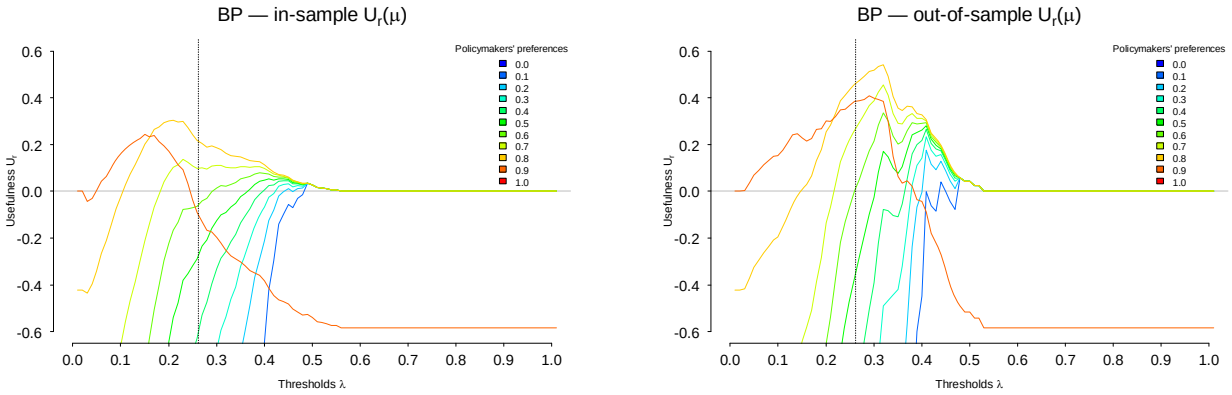
[9] Berg and Pattillo (1999b) set λ to generate $T_2 = 0.10$. Similarly, Berg and Pattillo (1999a) set in a related study λ at 50% and 25% of the crisis probability estimated with probit analysis.

Figure 1. $U_a(\mu)$ *(a),* $U_r(\mu)$ *(b) and NtS (c) for the BP model for* $\lambda \in [0,1]$ *and* $\mu = 0.0,0.1,...,1.0$

(a)



(b)



(c)

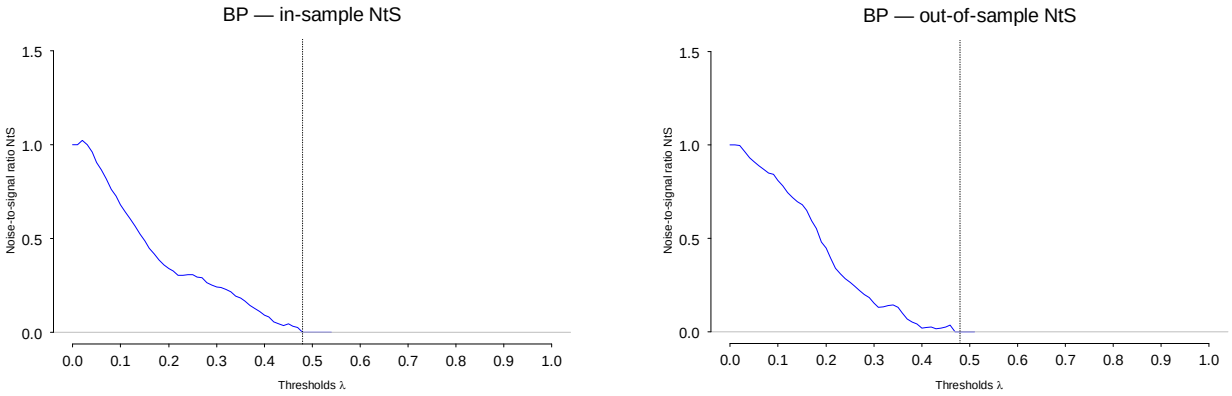


Table 3. Performance of the BP model for three different specifications

a) $\mu = 0.7$, $\lambda = 0.262$ and threshold specified by BP.

Dataset	λ	*TP*	*FP*	*TN*	*FN*	T_1	T_2	*Accuracy*	$U_a(\mu)$	$U_r(\mu)$	$U_a(\mu, w_j)$	$U_r(\mu, w_j)$
In-sample	0.26	126	211	1893	244	65.95 %	10.03 %	81.60 %	0.01	9.61 %	0.04	35.68 %
Out-of-sample	0.26	87	54	267	33	27.50 %	16.82 %	80.27 %	0.03	31.50 %	0.04	39.76 %

b) $\mu = 0.7$, $\lambda = 0.29$ and threshold optimizes preferences

Dataset	λ	*TP*	*FP*	*TN*	*FN*	T_1	T_2	*Accuracy*	$U_a(\mu)$	$U_r(\mu)$	$U_a(\mu, w_j)$	$U_r(\mu, w_j)$
In-sample	0.29	107	153	1951	263	71.08 %	7.27 %	83.19 %	0.01	11.20 %	0.04	37.32 %
Out-of-sample	0.29	84	41	280	36	30.00 %	12.77 %	82.54 %	0.04	38.87 %	0.05	47.52 %

c) $\mu = 0.8$, $\lambda = 0.20$ and threshold optimizes preferences

Dataset	λ	*TP*	*FP*	*TN*	*FN*	T_1	T_2	*Accuracy*	$U_a(\mu)$	$U_r(\mu)$	$U_a(\mu, w_j)$	$U_r(\mu, w_j)$
In-sample	0.20	216	416	1688	154	41.62 %	19.77 %	76.96 %	0.04	30.27 %	0.05	42.99 %
Out-of-sample	0.20	93	111	210	27	22.50 %	34.58 %	68.71 %	0.03	28.34 %	0.03	28.88 %

Notes: The in-sample dataset spans from 1986:1–1995:4 and the out-of-sample dataset from 1995:5–1996:12. Models a), b) and c) set λ based upon some measure on in-sample data (see above), while the same λ is applied on the out-of-sample data. The abbreviations are as follows: λ, threshold; *TP*, true positives; *FP*, false positives; *TN*, true negatives; *FN*, false negatives; T_1, type 1 errors; T_2, type 2 errors; $U_a(\mu)$, absolute Usefulness; $U_r(\mu)$, relative Usefulness; $U_a(\mu, w_j)$, weighted absolute Usefulness; $U_r(\mu, w_j)$, weighted relative Usefulness. *Accuracy* refers to $(TP+TN)/(TP+TN+FP+FN)$. The weights w_j represent the share of population of country i in period t to the sum of population of the sample in period t.

3.2 The LDP model

The logit model in LDP uses quarterly data for 28 countries, 10 advanced and 18 emerging economies, for the period 1990:1–2010:4. The dataset consists of 14 macro-financial indicators that proxy for asset price developments and valuations, credit developments and leverage, as well as traditional macroeconomic measures. The variables are defined both on a domestic and a global level, where the latter is an average of data for the Euro area, Japan, UK and US. These indicators are common in the macroprudential literature (see e.g. Borio and Lowe (2002, 2004) and Alessi and Detken (2010)). The indicators and their transformations (e.g. level, deviation from trend and rate of change), as well as the countries in the sample, are described in Table A in the Appendix. The occurrence of a systemic financial crisis is defined with a Financial Stress Index (FSI). The FSI captures distress in the main segments of the domestic financial market (money, equity and foreign exchange markets) into a country-specific composite index.[10] The FSI is transformed into the predicted variable by first defining crisis periods $I_j(0) \in \{0,1\}$ as those when the FSI exceeds the 90th percentile of the country-specific distributions, and then the pre-crisis periods $I_j(h) \in \{0,1\}$, where h=18, as the 18 months preceding the crises.

[10] The FSI consists of five components: the spread of the 3-month interbank rate over the rate of the 3-month government bill; quarterly equity returns; realized volatility of a main equity index; realized volatility of the exchange rate; and realized volatility of the yield on the 3-month government bill.

Table 4. The estimates for the LDP model

Variables	*Estimates*	*Std. error*	*Z*	*Sig.*	
Intercept	-6.744	0.612	-11.024	0.000	***
Inflation	-0.100	0.300	-0.334	0.738	
Real GDP growth	0.076	0.334	0.229	0.819	
Real credit growth	-0.001	0.001	-0.613	0.540	
Real equity growth	1.791	0.382	4.685	0.000	***
Leverage	0.003	0.001	3.204	0.001	***
Equity valuation	0.002	0.001	2.689	0.007	***
CA deficit	1.151	0.308	3.741	0.000	***
Government deficit	0.076	0.342	0.223	0.823	
Global inflation	0.207	0.341	0.608	0.543	
Global real GDP growth	1.156	0.419	2.761	0.006	***
Global real credit growth	0.685	0.381	1.799	0.072	*
Global real equity growth	0.832	0.419	1.985	0.047	**
Global leverage	0.712	0.427	1.668	0.095	*
Global equity valuation	0.959	0.472	2.029	0.042	**

Notes: Significance levels: 1%, ***; 5 %, **; 10 %, *.

The population-based weighted Usefulness measures in the BP model assume distress to solely be domestic, as misclassifications relate to the size of a economies' population rather than systemic relevance that might have effects outside domestic borders possible as well. Yet, a self-evident next step is to weight an entity by its importance to the system. By defining w_j to be the share of stock-market capitalization of country i in period t of the sum of stock-market capitalization in the sample in period t, we take the perspective of an external observer that assigns the importance of each entity based upon their relevance to the system. While this is an oversimplified measure of systemic relevance, it is out of the scope of this paper to create a more advanced measure, such as those of network models. The predictive performance of the model is tested on the Great Financial Crisis by dividing the dataset into in-sample data (1990:4–2005:1) and out-of-sample data (2005:2–2009:2). The estimates of the replicated LDP logit on the in-sample data are shown in Table 4 and are applied to the out-of-sample data.[11] The logit estimates are used for generating probabilities p_j and transformed into a binary point forecast P_j. The threshold λ is set to optimize in-sample Usefulness for given preferences μ, and later applied to the out-of-sample data.

[11] While the replica of the BP model is identical to the original one, we do not achieve identical results for the LDP model. The estimation procedure follows in general the specification in LDP and thus the model is still of similar nature. However, to better illustrate the Usefulness measures, the in-sample and out-of-sample datasets are static rather than recursive. In particular, it should be noted that more frequent real-time updates would improve the current discrepancy between the in-sample performance and the out-of-sample performance.

Table 5. In-sample (a) and out-of-sample (b) performance of the LDP model for policymakers' preferences $\mu = 0.0,0.1,...,1.0$

a) In-sample performance

Preferences	λ	*TP*	*FP*	*TN*	*FN*	T_1	T_2	*Accuracy*	$U_a(\mu)$	$U_r(\mu)$	$U_a(\mu,w_j)$	$U_r(\mu,w_j)$
$\mu=0.0$	0.78	0	0	1020	235	100.00 %	0.00 %	81.27 %	0.00	NA	0.00	NA
$\mu=0.1$	0.78	0	0	1020	235	100.00 %	0.00 %	81.27 %	0.00	NA	0.00	NA
$\mu=0.2$	0.78	0	0	1020	235	100.00 %	0.00 %	81.27 %	0.00	NA	0.00	NA
$\mu=0.3$	0.78	0	0	1020	235	100.00 %	0.00 %	81.27 %	0.00	NA	0.00	NA
$\mu=0.4$	0.61	28	14	1006	207	88.09 %	1.37 %	82.39 %	0.00	2.98 %	0.00	6.43 %
$\mu=0.5$	0.49	60	40	980	175	74.47 %	3.92 %	82.87 %	0.01	8.51 %	0.01	8.64 %
$\mu=0.6$	0.31	136	134	886	99	42.13 %	13.14 %	81.43 %	0.02	19.86 %	0.02	22.55 %
$\mu=0.7$	0.27	155	174	846	80	34.04 %	17.06 %	79.76 %	0.04	34.22 %	0.04	41.63 %
$\mu=0.8$	0.26	160	187	833	75	31.91 %	18.33 %	79.12 %	0.07	48.19 %	0.06	54.25 %
$\mu=0.9$	0.14	197	388	632	38	16.17 %	38.04 %	66.06 %	0.02	28.43 %	0.05	52.97 %
$\mu=1.0$	0.00	235	1020	0	0	0.00 %	100.00 %	18.73 %	0.00	NA	0.00	NA

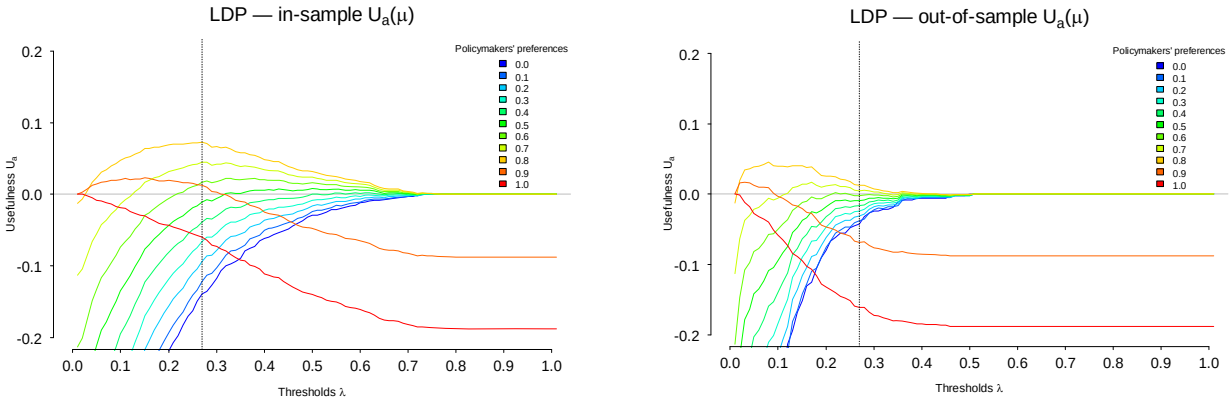
b) Out-of-sample performance

Preferences	λ	*TP*	*FP*	*TN*	*FN*	T_1	T_2	*Accuracy*	$U_a(\mu)$	$U_r(\mu)$	$U_a(\mu,w_j)$	$U_r(\mu,w_j)$
$\mu=0.0$	0.78	0	0	306	170	100.00 %	0.00 %	64.29 %	0.00	NA	0.00	NA
$\mu=0.1$	0.78	0	0	306	170	100.00 %	0.00 %	64.29 %	0.00	0.00 %	0.00	0.00 %
$\mu=0.2$	0.78	0	0	306	170	100.00 %	0.00 %	64.29 %	0.00	0.00 %	0.00	0.00 %
$\mu=0.3$	0.78	0	0	306	170	100.00 %	0.00 %	64.29 %	0.00	0.00 %	0.00	0.00 %
$\mu=0.4$	0.61	0	0	306	170	100.00 %	0.00 %	64.29 %	0.00	0.00 %	0.00	0.00 %
$\mu=0.5$	0.49	0	1	305	170	100.00 %	0.33 %	64.08 %	0.00	-1.42 %	0.00	-0.62 %
$\mu=0.6$	0.31	11	9	297	159	93.53 %	2.94 %	64.71 %	0.00	-2.04 %	-0.03	-7.00 %
$\mu=0.7$	0.27	24	16	290	146	85.88 %	5.23 %	65.97 %	0.01	4.39 %	0.00	1.14 %
$\mu=0.8$	0.26	24	17	289	146	85.88 %	5.56 %	65.76 %	0.01	8.09 %	0.01	8.07 %
$\mu=0.9$	0.14	84	63	243	86	50.59 %	20.59 %	68.70 %	-0.02	-25.48 %	0.02	20.25 %
$\mu=1.0$	0.00	170	306	0	0	0.00 %	100.00 %	35.71 %	0.00	NA	0.00	NA

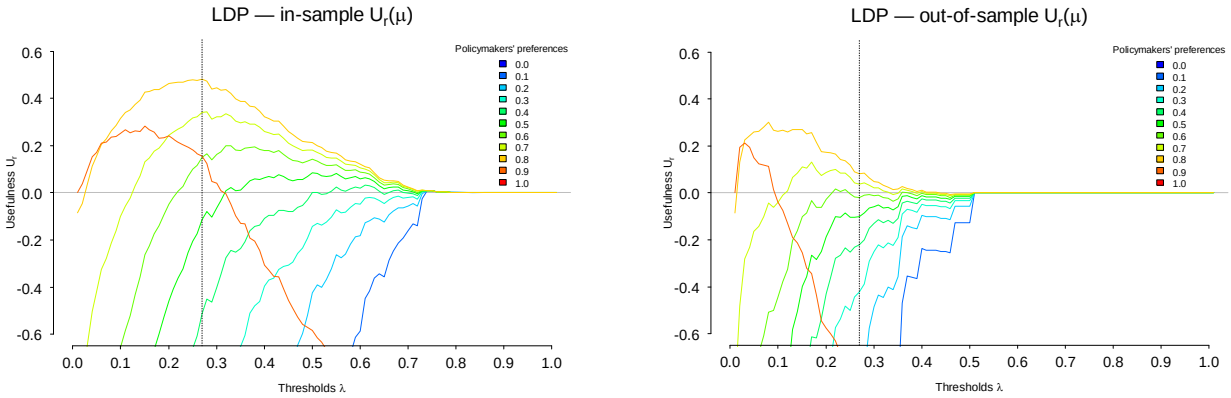
Notes: The in-sample dataset spans from 1990:4–2005:1 and the out-of-sample dataset from 2005:2–2009:2. Threshold λ is set to optimize in-sample $U_a(\mu)$, while the same λ is applied to the out-of-sample data. The abbreviations are as follows: λ, threshold; *TP*, true positives; *FP*, false positives; *TN*, true negatives; *FN*, false negatives; T_1, type 1 errors; T_2, type 2 errors; $U_a(\mu)$, absolute Usefulness; $U_r(\mu)$, relative Usefulness; $U_a(\mu,w_j)$, weighted absolute Usefulness; $U_r(\mu,w_j)$, weighted relative Usefulness. *Accuracy* refers to $(TP+TN)/(TP+TN+FP+FN)$. The weights w_j represent the proportion of stock-market capitalization of country i in period t to the sum of stock-market capitalization in the sample in period t.

Figure 2. $U_a(\mu)$ *(a),* $U_r(\mu)$ *(b), and* $U_{AD}(\mu)$ *(c) for the LDP model for thresholds* $\lambda \in [0,1]$ *and* $\mu = 0.0,0.1,...,1.0$

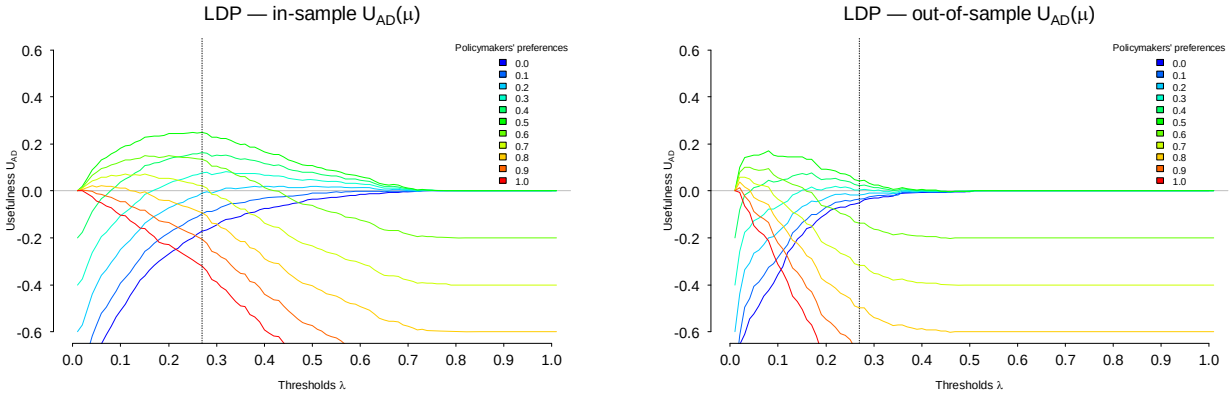
(a)



(b)



(c)



The performance of the LDP model for $\mu = 0.0,0.1,...,1.0$ is shown in Table 5. Table 5a shows in-sample performance with a λ that optimizes absolute Usefulness $U_a(\mu)$, and Table 5b shows the out-of-sample performance per μ for the λ specified in Table 5a. Figures 2a-b show $U_a(\mu)$ and

$U_r(\mu)$ for the LDP model with $\lambda \in [0,1]$ and $\mu = 0.0,0.1,...,1.0$. Table 5 and Figures 2a-b illustrate that the LDP model yields in-sample $U_r(\mu)$ above 10% for $\mu \in [0.6,0.9]$, while it is optimal to disregard the model for $\mu \in [0.0,0.3]$. However, only $\mu \in [0.7.0.8]$ yield positive Usefulness on out-of-sample data. The LDP authors show, however, that the out-of-sample Usefulness of the LDP model is robust to balanced preferences $(\mu \in [0.4,0.6])$ by using the Alessi-Detken (AD) evaluation framework.[12] The AD Usefulness $U_{AD}(\mu)$ for the model is computed in Figure 2c with $\lambda \in [0,1]$ and $\mu = 0.0,0.1,...,1.0$. As LDP use balanced preferences in their benchmark specification $(\mu = 0.5)$, but do not account for unconditional probabilities, this would translate to $\mu = 1 - P_1 = 0.8$ in our framework. Hence, the vertical lines in Figures 2a-c that represent the optimum for $U_a(\mu)$ and $U_r(\mu)$ with $\mu = 0.8$ and $U_{AD}(\mu)$ with $\mu = 0.5$ coincide. More generally, Figure 2c illustrates a shift in the preferences such that a lower μ is needed to offset the bias when not accounting for unconditional probabilities. While we show here that balanced preferences in the AD framework actually corresponded to a policymaker with $\mu = 0.8$, an in-depth discussion of optimal preferences is, however, out of the scope of this paper, such as the political economy aspects of maximization of a policymaker's utility vs. social welfare. These questions are to a large extent dependent on the case in question (e.g. bank-level distress vs. country-level debt crises or asset price busts), as well as the geographical perspective (e.g. a policymaker at a national central bank vs. one at an intergovernmental institution or organization) and the political role of a policymaker (e.g. one in charge of monetary policy or regulation vs. an external observer).

Again, in-sample results of the weighted in-sample Usefulness $U_a(\mu, w_j)$ and $U_r(\mu, w_j)$ are shown to be better than non-weighted, while the out-of-sample results are in general poorer. That is, the model fails in correctly classifying out-of-sample economies with large stock-market capitalization (except for $\mu = 0.9$). The failure of weighted out-of-sample Usefulness for both the BP and LDP models advocates accounting for the weights w_j not only when calibrating the thresholds of a model, but also when estimating the regression coefficients or optimizing the objective function of the method used for modeling the events.

[12] The definitions of the loss function and Usefulness measure in Alessi and Detken (2011) are as follows: $L_{AD}(\mu) = \mu T_1 + (1-\mu)T_2$ and $U_{AD}(\mu) = \min(\mu, 1-\mu) - L_{AD}(\mu)$.

4. Discussion and conclusion

This paper has derived a novel policymaker's loss function, and measures of absolute Usefulness and relative Usefulness for evaluating EWSs. In a descending order of contribution, the novelty derives from the following three enhancements: *i*) accounting for unconditional probabilities of the classes, *ii*) computing the proportion of available Usefulness that the model captures, and *iii*) weighting observations by their importance for the policymaker. The proposed loss functions and Usefulness measures are model free such that they can be used to assess signals issued by any type of model, e.g. logit and probit analysis and the signaling approach. The measures are also applicable to any types of crisis EWSs, e.g. banking, debt and currency crises. In this paper, the measures are derived for policymakers of three kinds: a cost-ignorant policymaker, a cost-aware policymaker with fixed but unequal preferences and a cost-aware policymaker with observation-specific costs. With the new evaluation measures, we have shown how a cost-aware policymaker would have perceived a seminal EWS for currency crises and a recent one for systemic financial crises. That is, we compare the new measures to two commonly used evaluation measures and assess to which extent and for which preferences the EWSs yield Usefulness.

The evaluation measures and the experiments lead to three key implications. First, we further highlight that the use of an objective criterion for setting the threshold and choosing the model specification is important as otherwise real-time models may lose in accuracy. Likewise, a pre-defined criterion increases objectivity of this types of *ex ante* performance tests. While optimization of the threshold is shown to improve the Usefulness of the BP results, an objective criterion for setting the threshold ought to have also improved the credibility of the exercise. Second, an artifact of incorporating unconditional probabilities is the need to be substantially more concerned of the rare class. For both the BP and the LDP models, the largest Usefulness is shown for so-called balanced preferences. Third, an obvious result of including observation-specific weights of entities is the importance of correctly classifying the most relevant ones. This self-evident implication derives from policymakers that have observation-specific preferences. Both EWSs in this paper were shown to be rather robust towards the weighted in-sample Usefulness, while their weighted performance on out-of-sample data is poorer. This opens the door for future studies. To better account for observation and class-specific costs, an interesting avenue for further research is to integrate the Usefulness measure into the objective function of classifiers, such as the maximum-likelihood function of probit/logit analysis. Another direction for further research is to use the weights in a weighted probit/logit model, or any other classifier where observations can be weighted, to give each observation its proper amount of influence over the parameter estimates.

Appendix

Table A. Variables and countries in the BP (a) and LDP (b) datasets

a) BP model with monthly data from 1986:1–1997:12.

Variables	*Countries*
Reserve loss[b]	Argentina
Export loss[b]	Bolivia
Real exchange rate deviation[a]	Brazil
Current account deficit to GDP[c]	Chile
Short-term debt to reserves[c]	Colombia
	India
	Indonesia
	Israel
	Jordan
	Korea
	Malaysia
	Mexico
	Pakistan
	Peru
	Philippines
	South Africa
	Sri Lanka
	Taiwan
	Thailand
	Turkey
	Uruguay
	Venezuela
	Zimbabwe

b) LDP model with quarterly data from 1990:1–2010:3.

Variables	*Countries*
Inflation[a]	Argentina
Real GDP[b]	Australia
Real credit to private sector to GDP[b]	Brazil
Real equity prices[b]	Canada
Credit to private sector to GDP[a]	China
Stock market capitalisation to GDP[a]	Czech Republic
Current account deficit to GDP[c]	Denmark
Government deficit to GDP[c]	Euro area
Global inflation[a]	Hong Kong
Global real GDP[b]	Hungary
Global real credit to private sector to GDP[b]	India
Global real equity prices[b]	Indonesia
Global credit to private sector to GDP[a]	Japan
Global stock market capitalisation to GDP[a]	Korea
	Malaysia
	Mexico
	New Zealand
	Norway
	Philippines
	Poland
	Russia
	Singapore
	South Africa
	Sweden
	Switzerland
	Taiwan
	Thailand
	Turkey
	UK
	US

Notes: Transformations: [a], deviation from trend; [b], annual change; [c], level.

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.

Alessi, L., Detken, C., 2011. Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity. European Journal of Political Economy 27(3), 520–533.

Berg, A., Pattillo, C., 1999a. Predicting currency crises – the indicators approach and an alternative. Journal of International Money and Finance 18, 561–586.

Berg A, Pattillo C., 1999b. What caused the Asian crises: an early warning system approach. Economic Notes 28, 285–334.

Bussière, M., Fratzscher, M., 2006. Towards a new early warning system of financial crises. Journal of International Money and Finance 25(6), 953–973.

Bussière, M., Fratzscher, M., 2008. Low probability, high impact: Policy making and extreme events. Journal of Policy Modeling 30, 111–121.

Candelon, B, Dumitrescu, E., Hurlin, C., 2012. How to Evaluate an Early-Warning System: Toward a Unified Statistical Framework for Assessing Financial Crises Forecasting Methods. IMF Economic Review 60(1), 75–113

Cardiarelli, R., Elekdag, S., Lall, S., 2011. Financial stress and economic contractions. Journal of Financial Stability 7(2), 78–97.

Demirgüç-Kunt, A., Detragiache, E., 2000. Monitoring Banking Sector Fragility. A Multivariate Logit. World Bank Economic Review 14(2) 287–307.

Drehmann, M., C. Borio and K. Tsatsaronis, 2011. Anchoring countercyclical capital buffers: the role of credit aggregates. International Journal of Central Banking 7(4), 189-240.

El-Shagi, M., Knedlik, T., von Schweinitz, G., 2012. Predicting Financial Crises: The (Statistical) Significance of the Signals Approach. IWH Discussion Papers No. 3.

Elkan, C., 2001.The foundations of cost-sensitive learning. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 01), pp. 973–978.

Frankel, J., Rose, A., 1996. Currency Crashes in Emerging Markets: An Empirical Treatment. Journal of International Economics 41(3–4), 351–366.

Fawcett, T., 2006. ROC graphs with instance-varying costs. Pattern Recognition Letters 27(8), 882-891.

Fuertes, A.-M., Kalotychou, E., 2006. Early Warning System for Sovereign Debt Crisis: the role of heterogeneity. Computational Statistics and Data Analysis 5, 1420-1441.

Fuertes, A.-M., Kalotychou, E., 2007. Towards the optimal design of an early warning system for sovereign debt crises. International Journal of Forecasting 23 (1), 85-100.

Granger, C., Pesaran, M., 2000. Economic and statistical measures of forecast accuracy. Journal of Forecasting 19, 537–560.

Jordá, O., Schularick, M., Taylor, A.M., 2011. Financial Crises, Credit Booms, and External Imbalances: 140 Years of Lessons. IMF Economic Review 59(2), 340–378.

Jordá, O., Taylor, A.M., 2011. Performance Evaluation of Zero Net-Investment Strategies. NBER Working Paper No. 17150.

Kaminsky, G., Lizondo, S., Reinhart, C.M., 1998. Leading Indicators of Currency Crises. IMF Staff Papers 45(1), 1–48.

Lo Duca, M., Peltonen, T.A., 2013. Assessing Systemic Risks and Predicting Systemic Events. Journal of Banking & Finance, forthcoming.

Lund-Jensen, K., 2012. Monitoring Systemic Risk Based on Dynamic Thresholds. IMF Working Paper, WP/12/159.

Manasse, P., Rubini N., Schimmelpfenning, A., 2003. Predicting sovereign debt crises. IMF Working Paper No. 221.

Sarlin, P., Marghescu, D., 2011. Neuro-Genetic Predictions of Currency Crises. Intelligent Systems in Accounting, Finance and Management 18(4), pp. 145-160.

Sarlin, P., Peltonen, T.A., 2011. Mapping the State of Financial Stability. ECB Working Paper No. 1382.