Finding early warnings of systemic risk in the South African financial system

Ivan van der Merwe

Abstract

To maintain financial stability macroprudential policymakers should aim to mitigate systemic risk. The extent to which policymakers can achieve this goal relies on the ability to detect, with sufficient lead time, warning signals for pending periods of elevated systemic stress. The aim of this paper is to explore the process for identifying warning signals and to establish if promising indicators exist in the South African case. The methodology deviates from traditional early warning literature in that the focus is not on periods of currency, banking or dual crises, but rather on the early detection of financial stress periods. The paper argues that initial indicator selection should be rigorous to ensure that only indictors with consistent signalling ability are considered. In support of this argument several signal evaluation criteria are used and a novel signalling score system is proposed for improved indicator selection (in-sample) and performance measurement (in-and-out-of-sample). The importance of measuring an indicator's overall signalling ability is emphasised by comparing signalling results at various thresholds and the case is made that over-reliance on the noise-to-signal ratio could cause useful signals to be ignored. The paper further evaluates in- and-out-of-sample signalling results from several composite indicators to test the assertion that multivariate indicators are superior to univariate indicators in issuing early warnings. The results suggest that this type of signalling methodology would be a useful addition to the macroprudential policy toolkit.

1. Introduction

Systemic risk is complicated and, as illustrated in Figure 1, it originates from numerous sources and manifests through several channels. This justifies the development of a macroprudential policy framework for South Africa. The ultimate goal of macroprudential policy is to mitigate systemic stress before it occurs by detecting early warning signals (EWSs) (i.e. an EWS as represented by the oval superimposed over section A of Figure 1). Recognising this need, the study on which this article reports, aimed to establish whether warning signals can be identified for pending periods of financial stress in South Africa. The signal extraction approach, introduced by Kaminsky, Lizondo and Reinhart (1998), was adapted to ascertain if there are indicators that constantly exhibit changing behaviour prior to periods of financial stress.

The contribution of this article is six-fold. First, it supplements the relatively little EWS research available for South Africa. Secondly, the methodology deviates from traditional EWS literature in that the focus is not on currency, banking or dual crises, but rather periods of financial stress as identified by a financial stress index. To date, very little research has been done on EWSs for financial stress in South Africa. Thirdly, the case is made that initial indicator selection should be rigorous to ensure that only good and consistent indictors are considered for in-depth analysis. To do this, the researcher combined several signalling criteria into a novel aggregate scoring system that improves performance measurement. Furthermore, since existing literature seldom measure indicators' overall signalling ability, this article contributes by presenting such results. A fourth contribution of this article entails the assessment of indicator signalling ability at various potential optimal thresholds. Several measures can identify optimal thresholds but EWS research relies mostly on a noise-to-signal ratio (NTSR) (see Detken *et al.*, 2014) for this purpose. This leaves a void in the literature since signalling performance at other thresholds are ignored. Accordingly, the researcher compared optimal thresholds as identified by various evaluation criteria to test for consistency, and reports the results in this article. Fifthly, the article presents a combination of individual indicators into three types of composite indicators to establish whether such multivariate indicators yield superior in-sample signalling results. Lastly, the out-of-sample signalling performance of univariate and multivariate indicators was assessed and this is reported.

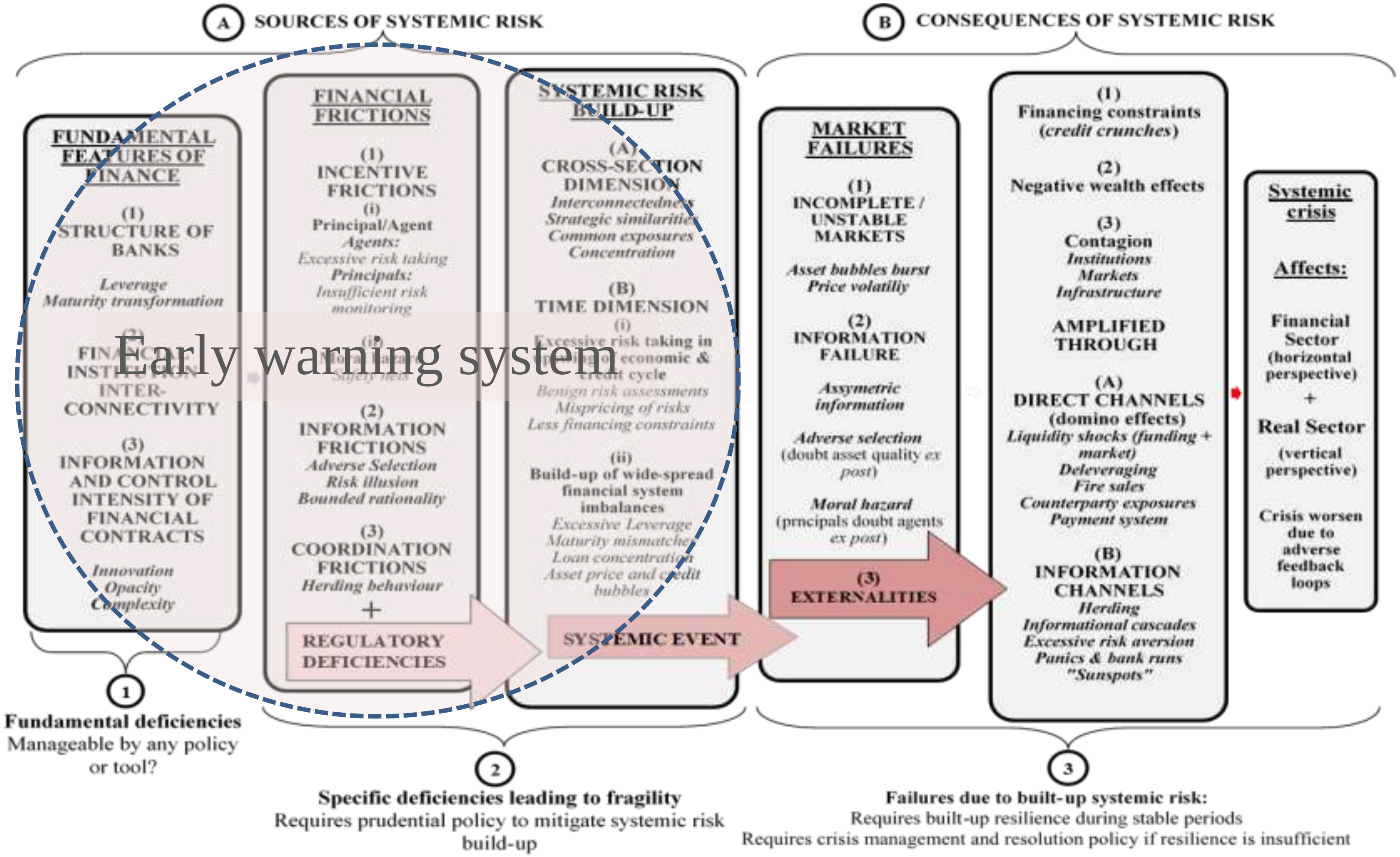
The outline of the article is as follows. Section 2 briefly reviews the EWS literature, while section 3 explains how and why a financial stress index is used to apply the signal extraction approach to South Africa. Section 4 reviews the signals approach methodology before introducing an aggregate scoring system for better measurement of indicator signalling ability. Section 5 advocates the measurement of indicator signalling ability over the entire threshold spectrum. Section 6 presents in- and out-of-sample signalling results for both univariate and multivariate indicators, while the article is concluded with section 7.

2. Early warning systems – a brief literature review

Early warning system (EWS) literature experienced a popular growth phase between the 1990s and early 2000s (Yucel, 2011). This was an automatic response to several banking and currency crises during that period, including major crises in European countries (e.g. the exchange rate mechanism [ERM] crisis and crises in the Nordic and East European countries), Latin America and East Asia (Laeven & Valencia, 2008; Vidal-Abarca & Ruiz, 2015). Pioneering EWS work by Kaminsky *et al*. (1998) and Berg and Pattillo (1999a; 1999b) focused on currency crises in emerging markets, and research soon extended to identify problems in banking systems also (e.g. Demirgüç-Kunt & Detragiache, 1998; Eichengreen & Rose, 1998) or dual (banking and currency) crises (Kaminsky, 1999; Kaminsky & Reinhart, 1999). The appeal of these models lies in their ability to use financial and real economy indicators to extract signals that indicate potential financial and economic vulnerability. In theory, such indicators provide policymakers with enough time to take appropriate action to avoid a crisis or to mitigate its severity. The 2007–2009 financial crisis caused a resurgence in EWS research, with forums such as the Group of Twenty (G20) summit (held in London, April 2009) clearly expressing the need for enhanced early warning tools to detect macroeconomic and financial risks (G20, 2009). Most of the early warning literature is based either on qualitative indicators of distress or on limited dependent regressions or on signal extraction methods (Abiad, 2003; Collins, 2003; Edison, 2003; Gaytán & Johnson, 2002; Hawkins & Klau 2000).[1] Earlier EWS research on currency crisis relied on qualitative analysis to identify causes of such events (Kaminsky *et al.*, 1998).

[1] More recent models include so-called ‘decision tree models’ where algorithms allocate a set of explanatory variables into decision trees to calculate optimal thresholds (e.g. Alessi & Detken, 2018).

Figure 1: Elements of systemic risk: Conceptual summary


A SOURCES OF SYSTEMIC RISK
B CONSEQUENCES OF SYSTEMIC RISK
Early warning system
FUNDAMENTAL FEATURES OF FINANCE
(1) STRUCTURE OF BANKS
Leverage
Maturity transformation
(2) FINANCIAL INSTITUTION INTER-CONNECTIVITY
(3) INFORMATION AND CONTROL INTENSITY OF FINANCIAL CONTRACTS
Innovation
Opacity
Complexity
FINANCIAL FRICTIONS
(1) INCENTIVE FRICTIONS
(i) Principal/Agent
Agents:
Excessive risk taking
Principals:
Insufficient risk monitoring
(ii) Moral hazard
Safety nets
(2) INFORMATION FRICTIONS
Adverse Selection
Risk illusion
Bounded rationality
(3) COORDINATION FRICTIONS
Herding behaviour
+
REGULATORY DEFICIENCIES
SYSTEMIC RISK BUILD-UP
(A) CROSS-SECTION DIMENSION
Interconnectedness
Strategic similarities
Common exposures
Concentration
(B) TIME DIMENSION
(i) Excessive risk taking in upswing of economic & credit cycle
Benign risk assessments
Mispricing of risks
Less financing constraints
(ii) Build-up of wide-spread financial system imbalances
Excessive Leverage
Maturity mismatches
Loan concentration
Asset price and credit bubbles
SYSTEMIC EVENT
MARKET FAILURES
(1) INCOMPLETE / UNSTABLE MARKETS
Asset bubbles burst
Price volatiliy
(2) INFORMATION FAILURE
Assymetric information
Adverse selection
(doubt asset quality ex post)
Moral hazard
(prncipals doubt agents ex post)
(3) EXTERNALITIES
(1) Financing constraints (credit crunches)
(2) Negative wealth effects
(3) Contagion
Institutions
Markets
Infrastructure
AMPLIFIED THROUGH
(A) DIRECT CHANNELS (domino effects)
Liquidity shocks (funding + market)
Deleveraging
Fire sales
Counterparty exposures
Payment system
(B) INFORMATION CHANNELS
Herding
Informational cascades
Excessive risk aversion
Panics & bank runs
"Sunspots"
Systemic crisis
Affects:
Financial Sector (horizontal perspective)
+
Real Sector (vertical perspective)
Crisis worsen due to adverse feedback loops
1
Fundamental deficiencies
Manageable by any policy or tool?
2
Specific deficiencies leading to fragility
Requires prudential policy to mitigate systemic risk build-up
3
Failures due to built-up systemic risk:
Requires built-up resilience during stable periods
Requires crisis management and resolution policy if resilience is insufficient

Some are more narrative in nature with little or no formal testing for the relative usefulness of indicators (e.g. Dornbusch, Goldfajn & Valdés, 1995; Frankel & Rose, 1996; Kaminsky & Reinhart, 1996; Krugman, 1996).Other research assessed whether fundamentals, used as leading indicators, behave differently in pre- and post-crisis periods or between crisis-prone and non-crisis countries (e.g. Aldasoro, Borio & Drehmann, 2018; Aziz, Caramazza & Salgado, 2000; Borio & Drehmann, 2009a; Carramazza, Ricci & Salgado, 2000; Eichengreen, Rose & Wyplosz, 1995; Frankel & Rose, 1996; Glick & Moreno, 1999; Kaminsky & Reinhart, 1996; 1999). These studies often employed graphical analysis and other parametric and non-parametric tests to establish the usefulness of indicators before and after the crisis.[2]

A popular approach often used in conjunction with these qualitative comparisons is to use statistical methods, such as linear regression analysis and the limited dependent variable probit/logit technique, also referred to as the discrete choice approach. The statistical approach establishes which indicators provide statistically significant contributions to the probability of possible future crises and even allows for variable transformation into a jointly determined continuous crisis probability indicator (see Detken *et al.*, 2014: 13). The ERM crisis in Europe in 1992–1993 motivated researchers such as Eichengreen *et al.* (1995; 1996) to use probit and multinomial logit models while the Mexican crisis motivated research by Sachs, Tornell and Velasco (1996) to identify possible reasons for the occurrence of financial crises in 20 developing countries by means of linear regression models. Similarly, the study by Frankel and Rose (1996) employed a multivariate probit model to assess which factors rendered 105 developing countries more susceptible to currency crises between 1971 and 1992. Berg and Pattillo (1999a; 1996b) applied a similar approach to test the predictive ability of general probit models, compared to the signal extraction approach, and to compare models of Frankel and Rose (1996), Sachs *et al.* (1996) and Kaminsky *et al.* (1998) to ascertain whether these authors' models were able to predict the 1997 Asian crisis. Other research applying the probit/logit techniques were that of Alessi and Detken (2018), Baron and Xiong, (2017), Barrell, Davis, Karim and Liadze (2010), Behn, Detken, Peltonen and Schudel (2013) Bussiere and Fratzscher (2002), Danielsson, Valenzuela and Zer (2018),

[2] See Kaminsky *et al.* (1998) and Abiad (2003) for extensive literature reviews.

Davis and Karim (2008a; 2008b), Holopainen and Sarlin (2016), Lo Duca and Peltonen (2011) and Lund-Jensen (2012).

The signal extraction (also called indicator or signalling) approach, originally popularised by Kaminsky *et al.* in 1998, focused on identifying signals in various real and financial variables that pre-empt the occurrence of currency crises. Subsequent research (e.g. Kaminsky, 1999; Kaminsky & Reinhart, 1999) replicated this procedure for early detection of banking and currency or dual crises, while others (e.g. Bruggemann & Linne, 2002; Edison, 2003) refined the signalling method by including more variables, countries or regions. More recent studies are those done by Aldasoro *et al*. (2018), Babecky *et al*. (2012; 2013), Behn *et al*. (2013), Borio and Drehmann (2009a), Drehmann, Borio, Gambacorta, Jiminez and Trucharte (2010), Drehmann, Borio and Tsatsaronis (2011), Drehmann and Juselius, (2014) and Holopainen and Sarlin (2016). In studies employing the signalling approach, the analysis of the signalling ability of selected variables requires a four-step approach. First, the incidence of currency, banking or dual crises in the selected sample of countries is identified. Crisis periods are usually identified by reviewing the historical narrative or crisis survey data or by calculating an index value representing large movements in exchange rates, interest rates, and currency reserves. Secondly, the choice of variables to be used as early warning indicators is based on evidence from prior literature and data availability (Kaminsky & Reinhart, 1999: 480) while variable transformations are often used to enhance signal extraction (Kaminsky *et al.*, 1998: 9). Thirdly, unusual behaviour of an early warning indicator is identified as a warning signal whenever such unusual behaviour occurs within a period of 12 to 36 months prior to a crisis (see Christensen & Li, 2014). Unusual behaviour, in turn, is identified as periods when a variable achieves levels that are relatively high (or low) to its historic levels. The level of deviation is measured against a predefined threshold, which, when exceeded, prompts a signal to be issued.

Signalling thresholds are determined by observing the empirical distribution of the variable and identifying those observations that exceed a certain percentile. The percentile cut-off choice varies for each variable and is affected by Type I and Type II measurement errors (described in detail in Section 4.1 below). Type I errors occur when an indicator does not issue a signal prior to a crisis period occurring. This often happens when the threshold is set too high. Conversely, a Type II error often occurs when the threshold is set too low and many false crisis warning signals are issued (Comelli, 2013: 9). The optimal threshold is usually identified at a level that minimises the relative occurrence of Type I and II errors (Aldasoro *et al.*, 2018: 35). To identify the optimal threshold, the relative frequency of each error type can be expressed as a ratio or, if policymaker preference exists in terms of Type I or II errors, a loss-function reflecting such a preference can be used (Ferrari & Pirovano, 2015: 4). Fourthly, the efficacy of signals is assessed by determining whether a crisis occurs within the predefined period (e.g. 24 months as was used in the current research), both in and out of sample. Although Type I and II errors are important, signalling performance can be measured by an array of criteria, obtainable from a table known as a 'contingency' or 'confusion' matrix (see Alessi, et al., 2015: 5; Alessi, & Detken, 2014: 339), and section 4 gives a detailed description of such performance measures. However, before these measures are presented, section 3 briefly describes why the signal extraction approach is used and how it was adapted in this study to incorporate financial stress periods for South Africa.

3. Adapting the signal extraction approach by using a financial stress index

Although both the signalling or discrete (regression-type) approaches are applied in the EWS literature, the current research applied the signalling approach for two reasons. First, as suggested by Davis and Karim (2008a), the signalling approach tends to be statistically superior when it comes to EWS focusing on individual countries, while the probit/logit approach is preferred for multi-country studies. Since South Africa was the focus of this research, the signalling approach seemed more appropriate. Secondly, the discrete approach also has some drawbacks in terms of robustness and statistical significance of variables in small samples and, as a result, problems with out-of-sample testing (Cihák & Schaeck, 2010: 136; Detken *et al.*, 2014: 14). Furthermore, including unimportant variables in multivariate regressions could increase the regression standard errors and possibly render misleading results (Babecky *et al.*, 2012: 19). Finally, the signalling approach allows thresholds to be set and trade-offs to be made between true and false signals, which make it appealing from a policymaker perspective.

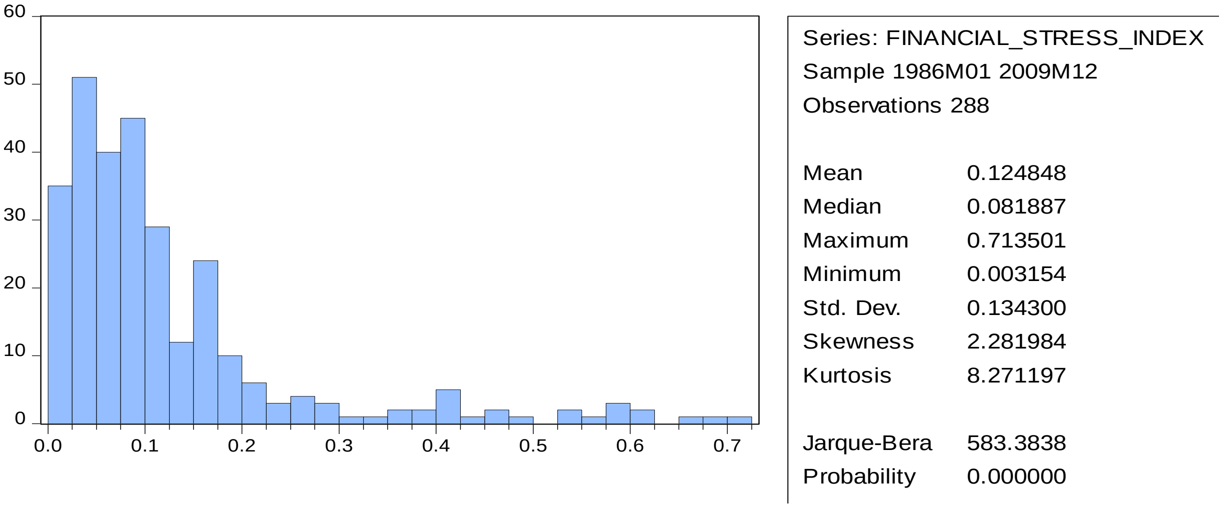
Most EWS literature focuses on detecting currency, banking or dual crises since the aim is to find early signals of imminent crisis periods. The identification of currency crises is often measured by the excessive movement in a composite measure representing foreign exchange market pressure (e.g. Berg & Pattillo, 1999a; 1999b; Bussiere & Mulder, 2000; Collins, 2003; Corsetti, Pesenti & Roubini, 1999; Eichengreen, Rose & Wyplosz, 1996; Frankel & Wei, 2005; Fratzcher, 1998; Herrera & Garcia, 1999; Kaminsky *et al.*, 1998; Sachs *et al.*, 1996; Tornell, 1999). Such a composite measure (often called the exchange market pressure index [see Patnaik, Felman & Shah, 2017]) usually includes measures of exchange rate depreciation, foreign reserve depletion and, if available, interest rate volatility, and is useful since it identifies pressure periods irrespective of the exchange rate system used. Defining banking crises is more subjective and might depend on aspects such as loan losses and bank capital erosion (e.g. Caprio & Klingebiel, 1996) or more specific criteria to be met, such as percentage of non-performing loans and/or bank nationalisation or bank runs (e.g. Demirgüç-Kunt & Detragiache, 1998). For most countries, the occurrence of crisis periods is not common and thus EWS literature often employs multi-country panel data to assess the signalling ability of indicators for the pooled sample (e.g. Aldasoro *et al.*, 2018).

In this article, the focus falls on financial stress as measured by a financial stress index (FSI) and not on currency or banking crises. Article 3 in this thesis elucidated why it is important to understand how financial stress leads to real economic consequences. Periods of high financial stress, and not just outright crisis periods, also cause systemic risk that could lead to detrimental real economic consequences. Cardarelli, Elekdad and Lall (2011) show that financial stress often precedes an economic slowdown in advanced economies, while financial vulnerability (e.g. credit and asset price growth), in turn, precedes financial stress periods. Thus, an EWS for financial stress is a valuable macroprudential policy tool. For policymakers, it will be preferable if factors causing periods of financial stress could be identified timeously, thereby enabling the implementation of prudential tools to limit the frequency and severity of such periods. For instance, Juks and Melander (2012) illustrate how an FSI for Sweden could have provided useful warning information to policymakers regarding the decision to release countercyclical capital buffers prior to the global financial crisis.

Another reason for focusing on signals for financial stress is that, unlike most EWS research that use multi-country samples, the focus here is only on South Africa. From a macroprudential point of view, country-specific early warnings are essential for credible policymaking (BIS, 2018: 65). Like other countries, the number of banking, currency or dual crises for South Africa is limited, while periods of financial stress had a higher frequency over the last 30 years. By using the FSI as 'crisis indicator' it becomes possible to identify a sufficient number of stress periods, thereby making a country-specific EWS study for South Africa viable. Although FSIs are contemporaneous measures of stress, they are also useful in early warning research, since they can be used as the dependent variable in EWS models (Oet, Ong & Gramlich, 2013: 21). Constructing a South African FSI for the period 1986–2018 therefore makes it possible to use the FSI as proxy for periods of systemic stress during that period.

The FSI is constructed by including eight basic financial market variables that are grouped into four sub-indices before amalgamating them into an overall index. These variables and sub-indices represent the most important elements of the South African financial system, namely the money, bond, equity and foreign exchange markets. Secondly, the FSI includes a conditional-rolling volatility measure for each of the eight basic variables to reflect stress that is only on the down-side and persists for a long-enough period. These volatility measures are then grouped together with their basic variable counterparts in the four market sub-indices. Thirdly, the FSI also includes correlations between the four sub-indices since these correlations become more pronounced during financial stress periods. As in portfolio theory, the overall level of the FSI is less than the weighted sum of its sub-indices, unless the correlation between all four sub-indices is perfectly positive (see Article 3 of the thesis). Accordingly, the FSI will approach the square of its weighted composite value during times of extreme stress when correlations approach +1. In the results section below (section 6), this FSI is used to identify periods of high financial stress (*hfs*), with such periods coded as a 'binary crisis' indicator in the EWS model. Only a few EWS studies exist were an FSI was used as 'dependent crisis' variable (cf. Christensen & Li, 2014; Lo Duca *et al.*, 2017; Pasricha *et al.*, 2013). These studies make use of variance-weighted FSIs and, as proposed by Illing and Liu (2003) and Cardarelli *et al*. (2009), identify high-stress periods when the FSI is more than 1 or 2 standard deviations (SDs) away from its sample mean. Since the 16-variable FSI in this article is not normally distributed (see Figure 2) using SDs to identify stress periods is not optimal.

Figure 2: Descriptive statistics of the FSI constructed for South Africa



Series: FINANCIAL_STRESS_INDEX
Sample 1986M01 2009M12
Observations 288

Mean	0.124848
Median	0.081887
Maximum	0.713501
Minimum	0.003154
Std. Dev.	0.134300
Skewness	2.281984
Kurtosis	8.271197
Jarque-Bera	583.3838
Probability	0.000000

Accordingly, the level of the index over the in-sample period (1986–2009) is used to construct a cumulative distribution function (CDF), and a percentile threshold is set to reflect non-stress periods, with values above this threshold considered as stress periods. The following condition must be satisfied to qualify as a period of high financial stress (*hfs*):

$$hfs_t = \begin{cases} 1 \quad if\ FSI_t > \ \mu_{FSI} + k\mu_{FSI}\ and\ \frac{FSI_t}{FSI_t^2} > \frac{1}{n}\sum_{i=1}^{n}\frac{FSI_{t-i}}{FSI_{t-i}^2}, where\ n = 12 \\ 0 \ \ otherwise \end{cases} \qquad [1]$$

High financial stress in period t occurs when the current FSI level is higher than its sample mean plus a pre-defined fraction k of the sample mean. The level of *k* is set so that *hfs* makes up no more than a certain percentage of observations. For the purposes of this study, *k* was set to identify 15% of observations as high-stress periods. For policymaking, the level of *k* can easily be increased or reduced. A lower value for *k* reduces the number of high-stress periods, which also limits the number of signals to be used for predictive purposes. However, a level of *k* = 15% is in line with estimates by Cambón and Estévez (2015: 38), who identified 12% of their Spanish FSI observations as being high stress.

4. Measures of indicator signalling ability

The signal extraction approach relies on the principle that real, financial and combined indicators issue signals prior to crisis periods. The value of an early warning indicator is converted into a binary variable that either issues an EWS or does not. Formally, at any time *t* during the in-sample period, an indicator *i*, with a CDF-normalised value X_t^i, issues a binary signal S_t^i that either denotes a calm period ($S_t^i = 0$) or issues a warning ($S_t^i = 1$). The warning signal occurs only when X_t^i exceeds a specified CDF threshold θ^i, representing the percentage of in-sample observations lying in the relevant extreme tail of the empirical CDF of the variable.[3] Because some indicators have thresholds in the upper tail, while others issue signals in the lower tail, the more general notation reflects absolute values of X_t^i as follows: ($S_t^i = 1\ if\ |X_t^i| > \theta^i$, otherwise $S_t^i = 0$). Determining θ^i for every indicator depends on the ability of that variable to correctly signal periods of financial stress over the in-sample period. In most signal extraction studies, signalling ability is measured by analysing the contingency matrix illustrated in Table 1 (e.g. see Alessi *et al.*, 2015: 3).

Table 1: Signalling performance evaluation (contingency) matrix for EWS indicators

Signal issued by indicator?	**Crisis occurs within defined period after signal (or no signal)**	**Crisis does not occur within defined period after signal (or no signal)**
Yes	**A** **[Correct Signals]**	**B** **[Incorrect Signals]** **{Type II error}**
No	**C** **[Missing Signals]** **{Type I error}**	**D** **[Correct No Signals]**

Source: Author's own compilation

For each indicator i, the in-sample observations X_t^i are grouped into one of the quadrants in Table 1. If ($|X_t^i| > \theta^i$), then ($S_t^i = 1$), and if a high-stress period then occurred within a defined period (e.g. 6 quarters or 24 months), the observation is counted as a correct signal and placed in quadrant A (in Table 1). If ($S_t^i = 1$) and no stress followed, the signal is incorrect noise and placed in quadrant B. Similarly, if ($|X_t^i| < \theta^i$), then ($S_t^i = 0$), and if stress then occurred within the defined period, the observation is classified as a missing signal under quadrant C, or correct non-signal under quadrant D if no stress follows. A perfect indicator would have all its observations counted under A and D and none under B and C. Since no indicator can achieve this perfection, a trade-off between the relative frequencies of values in Table 1 is required. This choice is known as threshold selection and,

[3] In the current study, empirical CDF rankings and threshold identification were done recursively, meaning that rankings at time *t* were based on observations up to time *t* and not over the whole sample period. This ensured that observations have appropriate rankings, given available data at time *t*. In general, non-recursive rankings lead to lower optimal threshold identification.

akin to statistical hypothesis testing, it reflects the probability of making Type I and Type II errors. Here, 'Type I errors' refers to no signal being issued prior to a stress period (i.e. false negatives as measured in C and expressed as C/(A+C)) while Type II errors occur when a signal is issued but no stress follows (i.e. false positives as measured in B and expressed as B/(B+D)) (Ferrari & Pirovano, 2015: 4).

A choice to minimise Type I errors implies that lower thresholds are used, leading to more false alarms. From a policymaking perspective, this might lead to unnecessary interventions or costly investigations (Davis & Karim, 2008a: 100). Alternatively, minimising Type II errors implies the use of higher thresholds with fewer signals and policy interventions, but also more frequent failures to identify stress periods. To minimise both Type I and Type II errors is ideal, but in practice, a trade-off must be made (Aldasoro *et al.*, 2018: 35). This trade-off is often expressed as a loss function that identifies optimal values for Type I and II errors by using one of the performance measurement criteria described next.

4.1. Loss functions to measure signalling ability

The quadrants of the contingency matrix in Table 1 make it possible to calculate various contingency matrix-based measures that could be used to perform two important functions, namely to evaluate the in- and out-of-sample signalling ability of an indicator (i.e. performance measurement criteria) and (ii) to identify the optimal threshold of an indicator where signalling performance is maximised (i.e. threshold selection criteria). Although threshold selection receives most attention in EWS literature, the measures described below were used in this study to perform both functions. Early warning studies (e.g. Kaminsky *et al.*, 1998: 21) rely on identifying a certain signalling threshold that maximises signalling performance. This is known as the optimal threshold, and it is usually identified by minimising a certain contingency matrix-based loss function for the in-sample period (Alessi *et al.*, 2015: 3). The most prominent loss function is known as the NTSR, which compares the ratio of false signals as a fraction of all possible false signals (B/(B+D)) to the ratio of good signals as a fraction of all possible good signals (A/(A+C)), or the Type II error/(1-Type I error) ratio (Davis & Karim, 2008a: 100). Formally, this first loss function (L_1) is expressed as:

$$min_{\theta}[L_1] = min_{\theta}\left[\frac{Type\ II}{1-Type\ I}\right] \qquad [2]$$

Noisy indicators will have many observations in B and/or C, and the NTSR for such an indicator will be more than 1, making it no better than a random signal (Alessi & Detken, 2009: 12). For each indicator i it becomes possible to find its optimal threshold θ^i that minimises the NTSR by making use of a grid search. Using the NTSR in isolation does have potential drawbacks though. First, for loss to be minimised, both Type I and Type II errors should be small, and this often favours trade-off positions where the noise is smaller, compared to the signal. This is usually achieved by setting higher thresholds where fewer signals are issued, but this might suggest that policymakers are extremely averse to false alarms and not to missing financial crises (Detken *et al.*, 2014: 15).[4] From a policymaking perspective, this might not be optimal, and other measures of indicator usefulness should also be considered. Alessi and Detken (2011) measure indicator usefulness with a loss function that incorporates policymaker preference in terms of Type I and Type II errors. The authors define this loss function as:

$$\min_{\theta}[L_2] = \min_{\theta}[\mu(Type\ I) + (1-\mu)(Type\ II)] \qquad [3]$$

The parameter μ reflects the policymaker's relative preference between Type I and II errors such that a value of (μ = 0.5) indicates no preference, while (μ > 0.5) indicates that the policymaker is more concerned with missing signals. Throughout this study, the value for μ was set at 0.5, since policyholder preferences fell outside the scope of this study. This loss function can also be used to indicate the relative usefulness U of an indicator by comparing the loss function (at various levels of θ) with the loss a policymaker incurs by ignoring the signal and relying on a random guess instead (Sarlin & Peltonen, 2011: 12). Loss from ignoring the signal occurs by assuming a signal is never issued [i.e. (C+D)/(A+B+C+D) = 1, and thus C/(A+C) = 1 and B/(B+D) = 0], or always issued [i.e. (A+B)/(A+B+C+D) = 1, and thus C/(A+C) = 0 and B/(B+D) = 1]. Either case encourages the signal to be disregarded and leads to a benchmark loss function of min[μ; 1 – μ] and the usefulness (U) of a variable is thus determined as:[5]

[4] The criticism expressed here applies in the opposite direction as well, since low NTSRs could also occur at low variable thresholds. This happens with variables that give signals below a certain threshold. In such cases, Type I errors decrease, Type II errors increase and the NTSR increases as threshold levels increase.

[5] Sarlin (2013) proposes a similar loss function, except that it includes unconditional probabilities for crisis periods (i.e. (A+C)/(A+B+C+D)) and calm periods (i.e. (B+D)/(A+B+C+D)). Sarlin (2013: 8) argues that conditional probabilities should be included to account for 'class size' differences in the contingency matrix.

$$U = \min[\mu;\ 1 - \mu] - \{\min_{\theta}[\mu(Type\ I) + (1-\mu)(Type\ II)]\} \qquad [4]$$

If (U > 0), the early warning indicator is useful, has signalling ability and can be included as an EWS indicator. This loss function is appealing, but to be applied in practice it requires the preference value μ to be known, which is difficult (Aldasoro *et al*., 2018: 36; Borio & Drehmann, 2009a: 32). Accordingly, Borio and Drehmann (2009a) define the following related loss function that minimises the NTSR, subject to a minimum percentage of crises periods X correctly predicted:

$$min_{\theta}[L_3] = \ min_{\theta}\left[\frac{Type\ \mathrm{II}}{1-Type\ \mathrm{I}}|(1 - Type\ \mathrm{I}) \geq X\right] \qquad [5]$$

For this loss function, Borio and Drehmann (2009a) propose a value of X ranging between 60% and 75%. The high value for X was applied since the authors used annual data and focused only on banking crisis periods, which are rare but extreme occurrences (i.e. only 13 crises in their dataset for 18 countries). The authors also classified signals to be correct if issued for periods up to three years prior to and during the year that crises occur. As described in detail in section 3, this study used financial stress periods as proxy for crisis periods. Accordingly, if financial stress periods are to be signalled, the requirement of X ≥ 60–75% is far too strict. Financial stress is measured with monthly data, and occurs far more frequently than bank crisis periods. Since the number of financial stress periods in this study could be selected and because the appropriate value for X was not clear, its value was set at 15%. This minimum value was in line with what other research (cf. Christensen & Li, 2014) found for individual indicators.[6] In the special case where (X = 0), loss functions 1 and 3 (defined by equations 2 and 5) are identical. Pasricha, Roberts, Christensen and Howell (2013: 15) propose a loss function (equation 6), which specifies the relative importance of conditional to unconditional signals rather than the minimisation of Type I and II errors:

$$min_{\theta}[L_4] = \ min_{\theta}\left[\frac{1}{\mathrm{A+B+C+D}} \times \left[\frac{CD}{C+D} + \frac{AB}{A+B}\right]\right] \qquad [6]$$

However, Alessi and Detken (2014) suggest that this approach by Sarlin (2013) is not robust to changes of the preference parameter (μ) and is difficult to use for policy purposes.

[6] The better indicators used in this study reached a maximum value for X of between 40 and 60%.

An omission in existing EWS literature is that at least three issues arise from using these contingency matrix-based measures:

1. Since each measure focuses only on a specific signalling aspect, it is plausible that combining them will result in a more robust performance measurement and threshold selection. However, such combinations are not found in EWS literature.
2. Most EWS studies only rely on optimal thresholds to evaluate signalling performance, although this gives no indication of the overall signalling ability of an indicator over its entire threshold spectrum. This could result in over-reliance on variables with inconsistent signalling ability, meaning that they perform better than pure chance at some thresholds, while at others they do not. Yet, contingency matrix measures can measure signalling performance at any threshold, which makes it possible to measure overall signalling ability and consistency. Since it is important for policymakers to establish an indicator's overall signalling ability before incorporating it into EWS models, this omission in the literature is surprising.
3. Justification for using one contingency matrix-based measure, such as the NTSR, over another for performance measurement or threshold selection, is mostly ignored in the literature. Since various measures are available, and because some could perhaps outperform others, this omission in the literature is also surprising.

The current study attended to these three issues as follows. Issue one is covered in section 4.2 where an aggregate scoring system that incorporates ten signal evaluation criteria is introduced. Results indicate that this scoring system is robust and superior to individual evaluation criteria in terms of performance measurement and threshold selection. The second issue is covered in section 5 where the merits of measuring indicators' overall signalling performance are detailed, with supporting evidence also presented in section 6. Lastly, section 6 also compares results from various contingency matrix criteria in terms of their ability to measure overall signalling performance and identify optimal thresholds.

4.2. An aggregate scoring system for better performance measurement and threshold selection

As described in 4.1, contingency matrix measures can be used to measure signalling performance and to identify optimal thresholds. However, from the EWS literature it is not clear why one measure should be preferred to another and, in principle, it would be better if several such measures are combined. Therefore, this section introduces an aggregate scoring system in Table 2, which incorporates ten contingency matrix measures (criteria). This scoring system facilitates more comprehensive assessment of indicators' signalling ability, and more robust optimal threshold identification.

An example of this scoring system is shown in Table 2 where the signalling performance of a potential early warning indicator used in this study (i.e. the 24-month moving average (MA) of the FSI in this case) is summarised at a threshold level of 88%. The ten contingency matrix criteria are listed in rows 1–10 in column 2 of Table 2. Their individual performance is recorded in column 3 and a combined score, called the aggregate signalling score measure (ASSM for short), is shown in row 11. This ASSM score is obtained by adding the 10 criteria values (rows 1–10 in column 3) together, with the aggregation sign for each criterion shown in column 4. Five of the 10 criteria measure lack of signalling ability (e.g. the NTSR) and their scores are entered as a negative value (i.e. negative aggregation sign). The ASSM is simple to interpret, with a bigger value indicating a more useful indicator. Conversely, if the indicator is no better than random (i.e. the number of observations in cells A, B, C and D in Table 1 are equal) the ASSM in row 11 will have a total of zero, thereby confirming a lack of superior signalling ability.

The example in Table 2 shows that, at a threshold of θ = 88%, the 24-month MA of the FSI is a useful early warning indicator, since its ASSM of 352% is far above 0%. As the frequency of A and D (correct signals) changes relative to B and C (incorrect signals), the ASSM will change, and this happens with every adjustment of θ. For instance, if θ = 64%, the ASSM in Table 2 reduces to 222%, indicating worse performance than at 88%.

Table 2: Aggregate scoring system to measure signalling ability of EWS indicators

<table>
<tr><th colspan="6">Potential indicator being evaluated: 24-month moving average of the FSI</th></tr>
<tr><th colspan="6">Signalling threshold level currently evaluated: θ = 88%</th></tr>
<tr><th colspan="2">Contingency-matrix-based criteria used to evaluate indicator</th><th>Criteria value at current threshold</th><th>Aggregation sign</th><th>Criteria benchmark to qualify as optimal threshold</th><th>Does indicator meet criteria benchmark?</th></tr>
<tr><td>1</td><td>Noise to signal</td><td>4%</td><td>-</td><td><60%</td><td>Yes</td></tr>
<tr><td>2</td><td>A/(A+C) (signal)</td><td>22%</td><td>+</td><td>>15%</td><td>Yes</td></tr>
<tr><td>3</td><td>B/(B+D) (noise)</td><td>1%</td><td>-</td><td><25%</td><td>Yes</td></tr>
<tr><td>4</td><td>A/(A+B)</td><td>98%</td><td>+</td><td>>50%</td><td>Yes</td></tr>
<tr><td>5</td><td>(A+C)/(A+B+C+D)</td><td>62%</td><td>+</td><td>>20%</td><td>Yes</td></tr>
<tr><td>6</td><td>$\left[\frac{1}{A+B+C+D}\times\left[\frac{CD}{C+D}+\frac{AB}{A+B}\right]\right]$</td><td>22%</td><td>-</td><td><25%</td><td>Yes</td></tr>
<tr><td>7</td><td>[1] + [6] < 85%</td><td>26%</td><td>-</td><td><85%</td><td>Yes</td></tr>
<tr><td>8</td><td>(A+D) / (B+C)</td><td>106%</td><td>+</td><td>>100%</td><td>Yes</td></tr>
<tr><td>9</td><td>(A/(A+B))/[(A+C)/(A+B+C+D)]</td><td>158%</td><td>+</td><td>>100%</td><td>Yes</td></tr>
<tr><td>10</td><td>(μ)[C/(A+C)] + (1-μ)[B/(B+D)] weight of C/(A+C) = 50%, weight of B/(B+D) = 50%</td><td>40%</td><td>-</td><td><50%</td><td>Yes</td></tr>
<tr><td>11</td><td><b>Aggregate signalling score measure (ASSM): Total of rows 1-10</b></td><td><b>352%</b></td><td></td><td><b>>0%</b></td><td><b>Yes</b></td></tr>
<tr><td>12</td><td>[7] and [8] beat benchmarks?</td><td>TRUE</td><td></td><td>TRUE</td><td>Yes</td></tr>
<tr><td>13</td><td>[9] and [12] beat benchmarks?</td><td>TRUE</td><td></td><td>TRUE</td><td>Yes</td></tr>
<tr><td>14</td><td>[5] > [10]?</td><td>TRUE</td><td></td><td>TRUE</td><td>Yes</td></tr>
<tr><td>15</td><td>Total [1-10] > 0?</td><td>TRUE</td><td></td><td>TRUE</td><td>Yes</td></tr>
<tr><td>16</td><td>Loss function 1 (equation 2)</td><td>4%</td><td></td><td><60%</td><td>Yes</td></tr>
<tr><td>17</td><td>Loss function 2 (equation 3)</td><td>40%</td><td></td><td><50%</td><td>Yes</td></tr>
<tr><td>18</td><td>Loss function 3 (equation 5)</td><td>40%</td><td></td><td><50%</td><td>Yes</td></tr>
<tr><td>19</td><td>Loss function 4 (equation 6)</td><td>22%</td><td></td><td><25%</td><td>Yes</td></tr>
<tr><td>20</td><td><b>Total of 4 loss functions</b></td><td><b>105%</b></td><td></td><td><b><185%</b></td><td>Yes</td></tr>
</table>

Source: Author's own compilation

Since an indicator's performance can be measured at any threshold between 0% and 100%, this scoring system also identifies an optimal threshold at the level of θ where the ASSM in row 11 is maximised. However, not all indicators will necessarily have good EWS performance at their optimum score. This is due to two reasons. First, as detailed in section 6.1, some indicators simply have bad overall signalling ability, which translates into bad optimal threshold performance. Secondly, indicators (even better ones) could have unsatisfactory optimal threshold signalling performance if some of the evaluation criteria in rows 1–10 in Table 2 show poor results (e.g. high noise level or low signal ratio or low good-to-bad signal ratio). As a result, additional checks are required to verify that an indicator has robust signalling performance at its optimal threshold. This additional requirement is that each of the evaluation criteria in rows 1–10 must meet a benchmark value shown in column 5. Most benchmark values (i.e. for criteria 4, 6, 8, 9 and 10) are based on the performance of a random variable, but to ensure that noisy indicators are rejected, criteria 1, 3 and 7 are stricter. Since stress periods only comprise fraction *k* of the sample, the frequency for A in Table 1 is often small, and therefore less strict values apply to criteria 2 and 5. Column 6 indicates whether the indicator meets these criteria benchmarks at the current threshold. In Table 2, the FSI indicator meets all ten benchmark values, again confirming its good signalling ability at the 88% threshold. For added rigour, rows 12–15 of column five impose additional requirements that must be met to qualify as potential optimal threshold. Lastly, the four loss functions from section 4.1 and their combined total appear in rows 16–20 and they too must meet their benchmark requirements in column 5.

In summary, to qualify as optimal threshold requires the highest possible aggregate signalling score where all 20 criteria benchmarks in column 5 are met. If not all 20 criteria are met, the level with the next highest score is selected as optimal threshold, provided that all 20 requirements are met. In Table 2, the threshold of 88% also turns out to be the optimal threshold of the FSI indicator, since it is the maximum aggregate score where all 20 benchmarks are satisfied. Overall, this scoring system is very robust and performs well not only in identifying optimal thresholds, but particularly also in measuring overall signalling performance – the importance of which is motivated next.

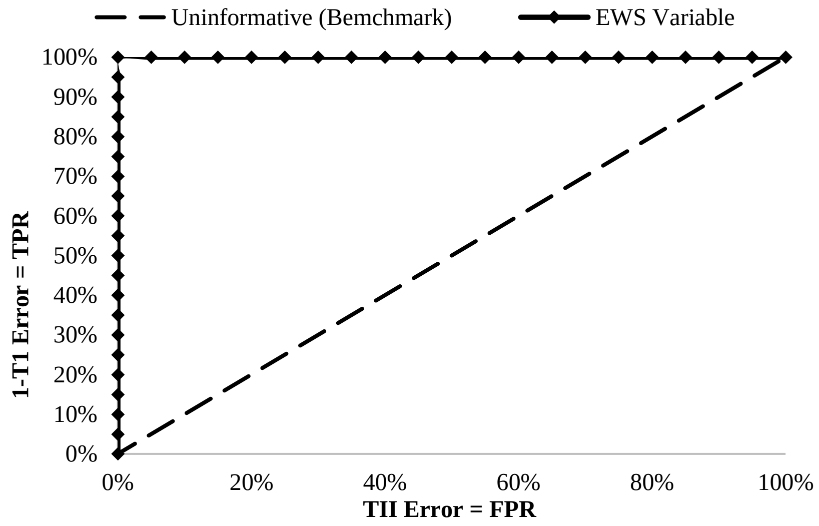
5. Receiver operating characteristic (ROC) curves and the merits of measuring overall signalling performance

Most previous EWS research measure indicator signalling performance only at an optimal threshold, as identified by some evaluation measure, such as the NTSR. Usually, the optimal threshold is identified by means of a grid search, with the search mostly limited to the top (or bottom) 20–30% of the applicable distribution tail. A drawback of this is that the stability of the trade-off between Type I and Type II errors is not sufficiently emphasised, since the goal is solely to find the optimal threshold for a specific evaluation measure (e.g. a loss function). This might ignore the 'bigger picture', not just in terms of the error trade-off but also in terms of signalling consistency over the entire threshold spectrum. For instance, at high thresholds, some indicators might have a good NSTR and ASSM, but the results could become worse and even volatile at thresholds ignored during the grid search. This inconsistency is problematic, especially when policymaker preferences for Type I and II errors must be considered (Aldasoro *et al.*, 2018: 34). Therefore, evaluating signal quality at any threshold is incomplete when it is not known what the costs and benefits of implementing policy, based on that threshold's signals, will be (Drehmann & Juselius, 2014: 762).

To minimise this potential inconsistency, only indicators with good overall signalling performance during the in-sample period should be considered for further analysis, and this process was followed in the current research. To identify such indicators, values for some evaluation measures from sections 4.1 and 4.2 are calculated over the entire threshold spectrum, and the results are then averaged to assess 'overall' signalling performance for each potential indicator. This overall performance measurement not only excels at identifying good indicators, but also makes it possible to compare the performance measurement ability of different evaluation criteria. However, using contingency matrix-based criteria to measure overall signalling performance is not common in EWS literature, and proper evaluation of this process requires that an existing benchmark be used for comparison. For this purpose, a performance measure known as the receiver operating characteristic (ROC) curve can be used (Aldasoro *et al.*, 2018: 34). Accordingly, the rest of this section describes what ROC curves entail, while the results section includes ROC curve analysis as part of the relevant performance comparisons.

Gaining popularity with the development of radar during World War II, a ROC curve illustrates the trade-off between true and false radar signals (Streiner & Cairney, 2007: 122). An increase in the gain (akin to the volume button on a radio) of the radar set allows for more signals to be picked up, but this is also accompanied by an increase in the number of false signals. The ROC curve thus illustrates the extent to which good signals outperform bad signals at every gain level. Although more prominent in medical research, ROC curves are also useful in EWS research (e.g. (Aldasoro *et al.*, 2018; Berge & Jordà, 2011; Drehmann & Juselius, 2014; Candelon, Dumitrescu & Hurlin, 2012) since they indicate how true positive signals (i.e. 1-Type I errors) outweigh false positive signals (i.e. Type II errors) at every threshold level (i.e. θ). The ROC curve of a purely random indicator with no signalling ability is represented by the 45-degree diagonal (benchmark) line in Figure 3. This reflects the fact that a random variable is equally likely to signal a true or a false signal at any threshold level.

Figure 3: ROC curve of a perfect EWS indicator



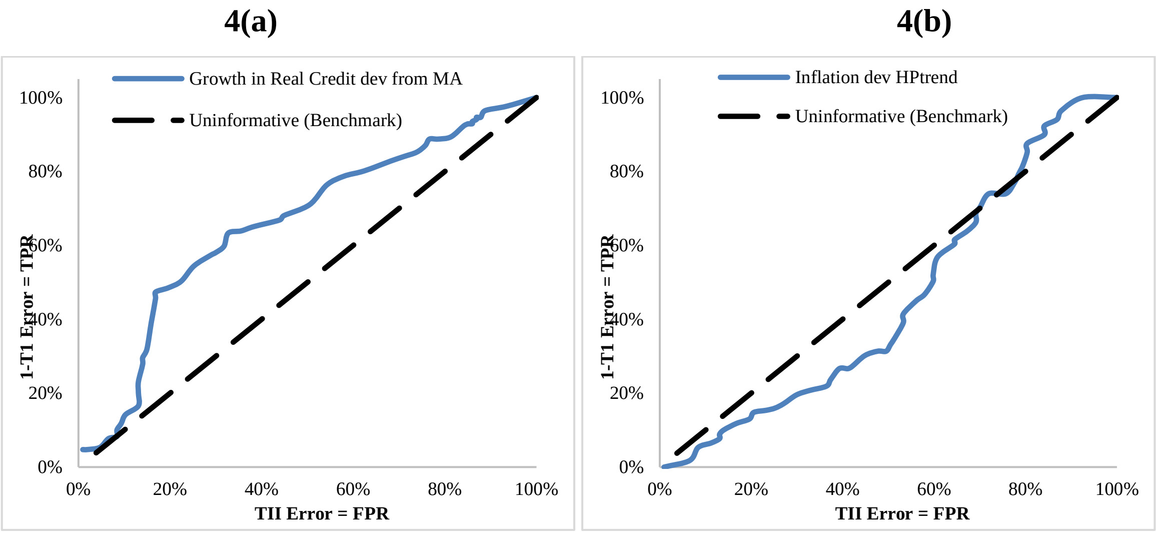
Source: Author’s own compilation

Just like this benchmark line, ROC curves have a positive slope, representing the trade-off between Type I and II errors. For an indicator to classify as having signalling ability, its ROC curve should lie above the diagonal benchmark, hence reflecting that true signals outnumber bad signals at any threshold. This applies to variables that issue signals in the upper tail of their distribution. If a variable issues signals in the lower tail, its ROC will lie below the diagonal.[7] More formally, the ROC curve measures the conditional probability of the true positive rate (TPR) to the conditional probability of a false positive rate (FPR) over the threshold spectrum (Sarlin & Peltonen, 2011: 13) and values above the diagonal confirm that an indicator’s distribution is stochastically larger in one state (i.e. prior to high-stress periods) than in the normal state (i.e. prior to no-stress periods) (Drehmann & Juselius, 2014: 762). The ROC curve of a perfect indicator in Figure 3 starts at the origin (i.e. FPR and TPR are at 0% when threshold θ = 100%). As the threshold decreases, the curve extends upwards towards the maximum TPR, while the x-axis value remains at 0%. Once the TPR of 100% is reached, the perfect ROC curve will make a 90-degree turn and form a horizontal line,

[7] ROC curves below the diagonal can be converted to also lie above the diagonal by changing the sign of the indicator.

with only the FPR increasing as θ decreases further (Lo Duca *et al.*, 2017: 35). Figures 4(a) and 4(b) provide examples of good and bad EWS indicators used in this study for South Africa. The ROC curve for real private sector credit growth in Figure 4(a) shows that it is a good EWS indicator, while the ROC curve in Figure 4(b) suggests that inflation above trend is a bad indicator.

Figure 4: ROC curves of two potential EWS indicators



Source: Author’s own compilation

ROC curves measure Type II errors (i.e. the FPR) on the x-axis, while 1-Type I errors (i.e. the TPR) are measured on the y-axis. At high threshold levels near the origin (e.g. $\theta = 99\%$), both the TPR and FPR are low (Detken *et al.*, 2014: 15). Since few signals are then issued, the NTSR is usually low in this area, which is why EWS studies tend to identify θ at a level where the NTSR is minimised (Dawood, Horsewood & Strobel, 2017: 23). However, as θ is reduced, the TPR and FPR increase, causing the NTSR to increase, and thus the rest of the threshold spectrum is ignored for signalling purposes. The result is that the signalling ability of the indicator over the entire threshold spectrum is then ignored, while policymaker choice regarding Type I and II errors is also not accommodated. This deficiency justifies the use of a ROC analysis, since the ROC curve has several attributes that make it appealing. First, the ROC curve augments grid searches by showing an indicator’s signalling ability at all thresholds and without specifying any loss function that is contingent on policymaker preferences for Type I and II errors (Detken *et al.*, 2014; Drehmann & Juselius, 2014).

Second, the ROC curve exhibits several other useful properties, namely:

- it does not depend on measurement unit and is invariant to monotone increasing transformations of the variable;
- it provides a common scale for comparing signalling ability of variables;
- it provides a visual representation of trade-offs between true and false signals at all threshold levels; and
- the area under the ROC (AUROC) can be interpreted as a kind of probability that the signalling distribution of an indicator during crisis periods is higher than during tranquil periods (Drehmann & Juselius, 2014; Pepe, Longton & Janes, 2009).

The last two attributes are useful when evaluating indicators, since some have good evaluation results for one section of the CDF distribution, with more unstable results over the rest of the threshold spectrum. In addition, a visual representation of the AUROC not only indicates the average signalling ability of an indicator but also allows for visual comparison of signalling ability between indicators. Average signalling ability is determined by measuring the AUROC (Drehmann & Juselius, 2014: 762). The AUROC can be measured in various ways (cf. Candelon *et al.*, 2012), and in this study, the average of 50 trapezoidal approximations was used to estimate the AUROC as follows (cf. Savona & Vezzoli, 2015):[8]

$$AUROC = \sum_{\theta=2\%}^{100\%} \left\{ \left[\frac{(1-T1_{\theta})+(1-T1_{\theta-2\%})}{2} \right] \times (T2_{\theta} - T2_{\theta-2\%}) \right\} \qquad [7]$$

where θ is the threshold ranging between 0% and 100%, and *T1* and *T2* are Type I and II errors respectively. As the ROC curve of a perfect indicator covers the whole area above the 45-degree diagonal (see Figure 3), it will have an AUROC of 100%, with 50% above and the other 50% below the diagonal respectively. Higher AUROCs indicate a superior TPR to FPR rate over all threshold levels (Brave & Butters, 2012: 3; Hsieh & Turnbull, 1996: 27) and any AUROC above 50% means the indicator is useful, compared to a random indicator with an AUROC of 50% (Aldasoro *et al.*, 2018; Berge & Jordà, 2011). In Figure 4(a), the AUROC is 71%, meaning that the extent to which this indicator outperforms a random variable (i.e. with no signalling ability and benchmark surface area of 50%) is 21 percentage

[8] Type I and II errors are measured at intervals of 2% for θ, which leads to 50 trapezoidal approximations per indicator.

points. Put differently, the probability that the signalling distribution of the indicator in Figure 4(a) is higher prior to high-stress periods than during tranquil periods is 71%. Conversely, the AUROC of the indicator in Figure 4(b) is at 46%, implying it is worse than a random indicator.

Lastly, it is also possible to determine an optimal threshold level on the ROC curve where the TPR to FPR ratio is maximised (Aldasoro *et al.*, 2018: 35; Baker & Kramer, 2007: 344).[9] This is the point on the ROC curve that is highest above the benchmark diagonal. In this regard, interpreting the ROC curve is like the efficiency frontier in modern portfolio theory (see Markowitz, 1952), where the optimal trade-off between risk and return is also found at the upper-left point on the frontier. In the current study, the ROC curve was primarily used to measure the overall signalling performance of indicators and to compare the results (see section 6) with those of other contingency matrix criteria. However, since ROC curve analysis can also identify an optimal threshold, it is included when comparing several optimal threshold identification criteria in section 6.

6. Empirical results

Section 4.1 reflected three neglected issues in EWS literature. First, the absence of a performance measure that consists of multiple measurement criteria is surprising. Second, measuring the overall signalling ability of indicators is often omitted. Lastly, EWS literature neglects to justify why a measure, such as the NTSR, is preferred, yielding little comparative evidence with other measures to test for consistency. The results presented in this section reflect issues as follows.

1. The overall signalling ability of individual indicators was assessed for the in-sample period. Five contingency matrix performance measures were used to test for consistency, including the newly proposed aggregate signalling score measure. Based on these results, only the best indicators were considered for subsequent univariate and multivariate analysis at optimal thresholds.

[9] Optimal thresholds on ROC curves can also be optimised to reflect policymaker preferences. Such thresholds are calculated at points where the expected marginal rate of substitution between the net marginal utilities of accurate prediction of non-stress and stress periods equals the slope of the ROC curve (Baker & Kramer, 2007:345; Drehmann & Juselius, 2014: 762). However, such policy preferences are not easily determined, and fell beyond the scope of the current study and were therefore ignored (i.e. the value of μ in equation 3 is constant at 0.5).

2. Once good indicators had been identified, several potential optimal thresholds were calculated for those indicators, and the results were compared for consistency before selecting a final optimal threshold.
3. These final optimal thresholds were used to assess the optimal threshold signalling performance of univariate indicators for the in-sample period. For robustness, ten evaluation measures (including the ASSM) were used for this purpose.
4. Three types of composite indicators were constructed from univariate indicators and the in-sample signalling performance of these multivariate indicators was analysed.
5. Lastly, out-of-sample signalling performance of univariate and multivariate indicators was assessed.

6.1. In-sample overall signalling performance of individual indicators

This section reports on the overall (i.e. average) in-sample performance of individual indicators to identify good indicators. To avoid the inconsistency issue raised in section 5, overall signalling ability was first assessed before subjecting indicators to further performance evaluation at optimal thresholds. Indicator evaluation was based on in-sample data between January 1986 and December 2009, while data for 2010 to 2018 was used to test out-of-sample signalling ability.**[10]** Indicator selection was influenced by EWS literature and data availability and included the following groupings:[11]

- external sector measures (e.g. current account/GDP, financial account/GDP, REER, bilateral exchange rate levels and trends and forex reserves ratios);
- financial sector measures (e.g. M2 and M3 ratios, total and private sector domestic credit ratios, domestic and US interest rates [real and nominal], bank reserves/assets and bank liquid assets);
- real sector measures (e.g. real GDP growth, inflation, gross fixed capital formation, recession and business cycle indicators); and
- asset price measures (e.g. equity prices movements [real and nominal] and correlations, share price volatility, domestic residential real estate prices [real and nominal], gold

[10] The in-sample dates were specifically chosen to include the global financial crisis period (2007–2009).

[11] Indicators were tested in various transformations, including levels, first differences, percentage changes, deviation from trends or MAs of varying lengths. This is standard practice in EWS literature (Sarlin & Peltonen, 2011: 11).

and oil prices).

The overall signalling performance of over 100 possible indicators was assessed to identify those with above average signalling ability. Usually, the AUROC is used to assess overall signalling performance, but section 5 clarified that other measures can, in principle, also perform this function. Therefore, five contingency matrix criteria were used for this purpose, with the comparative results shown in Table A1 in the Appendices section. Four criteria in Table A1 are existing contingency matrix criteria namely AUROC, NTSR (i.e. equation 2), relative usefulness (i.e. equation 4) and the conditional probability measure (i.e. A/(A+B)). The aggregate signalling score (i.e. ASSM in row 11 in Table 2) was used as fifth performance measure to establish whether it had superior evaluation properties.[12] Each of the five criteria in Table A1 reflects the average score that an indicator obtained over all threshold levels between 0 and 100%.[13] To simplify analysis, variables were ranked according to each of the five criteria. Due to some ranking inconsistency, the overall ranking (i.e. average of the five criteria rankings) in the last column of Table A1 was used as final filter for selecting the best indicators.

The results indicate that most variables showed poor overall signalling ability and thus they were excluded, thereby leaving only 41 potential indicators shown in Table A1. Compared to the overall ranking in the last column, the ASSM (i.e. criterion 5) is most consistent, confirming that it is robust and marginally superior to the other individual criteria in identifying average signalling ability. These results also confirm the expectation that an aggregate performance measure should outperform individual measures, and it provides policymakers with a robust indicator filtering mechanism. Aside from ASSM superiority, the rankings in Table A1 show that overall signalling performance was measured consistently by all five criteria and, accordingly, there was no reason to omit this step before undertaking optimal threshold analysis. The results in Table A1 also suggest that many indicators from traditional EWS literature (e.g. current account or GDP, house price growth, domestic credit to private sector or GDP) are not consistent or useful indicators of pending periods of financial stress in South Africa, and they were therefore excluded from further

[12] The five criteria used represent a variety of evaluation measures found in the literature. The aim was not to perform an exhaustive comparison of all possible measures, but rather to see whether different measures yield similar results.

[13] To ease the computational burden, threshold increments of 2% were used.

analysis. Indicators with good overall signalling ability are mostly financial variables, such as share market returns, the effective exchange rate, interest rates and previous levels of the FSI. The lower-ranked indicators in Table A1 compare badly to the top-ranked indicators. Accordingly, only the top-ranked indicators in Table A1, with an ASSM above 200%, were considered for more thorough analysis at their optimal thresholds, as indicated in section 6.2.

6.2. In-sample optimal threshold signalling performance of individual indicators

In most previous EWS studies, signalling performance of an indicator was assessed at its optimal threshold. This optimal threshold first had to be identified by using a performance measure, and although the NTSR is often preferred, several other contingency matrix criteria can perform this function. This implies that the NTSR might not necessarily be the best measure to identify optimal thresholds. However, predominant use of the NTSR means that threshold identification, using other measures, is largely unexplored. Accordingly, for consistency purposes, the current study contributes by comparing optimal thresholds identified by several such measures. Table 3 lists eight potential optimal threshold measures for the best-performing indicators identified in Table A1. The results suggest that threshold selection is not a clear-cut process and that the NTSR is not necessarily the best threshold selection measure. The first six measures in Table 3 are popular performance criteria as described in section 4.1.[14] The seventh criterion is the ASSM as introduced in row 11 of Table 2. The eighth criterion, called ‘T1 = T2 error’, identifies optimal thresholds where Type I and II errors are equal, which is insightful from a neutral policymaker perspective (cf. Candelon *et al.*, 2012). In Table 3, the first three criteria constantly identified the thresholds that tended towards the extreme section of the distributions of indicators. As described in section 4.1, this result was expected for the NTSR. The conditional probability of stress measure (i.e. A/(A+B)) mirrored the NTSR at all threshold levels and therefore identified the same optimal thresholds. In turn, the ratio of conditional to unconditional probability of stress {(A/A+B)/[(A+C)/(A+B+C+D)]} correlated highly with the conditional probability measure and therefore it also identified the same optimal thresholds.

[14] Loss function 3 (i.e. equation 5 in section 4.1) was excluded, since it yielded the same results as the NTSR. This occurred because the requirement in equation 5 that (1-Type I) > X, with X = 15% was met for all variables in Table 3.

Table 3: Various potential optimal thresholds identified for top indicators

Variable name	NTSR	A÷ (A+B)	(A/A+B) ÷ [(A+C)/ (A+B+C+D)]	Loss function 2	Loss function 4	ROC curve	ASSM	T1=T2 error	**Final optimal threshold**
24-month ALSI avg. real return	72%	72%	72%	34%	34%	34%	34%*	58%	**58%**
36-month ALSI avg. real return	92%	92%	92%	70%	70%	70%	70%	32%	**70%**
Federal funds rate (Y-o-Y % change)	52%	52%	52%	40%	40%	40%	40%	40%	**40%**
Business cycle indicator (2-year change)	78%	78%	78%	32%	32%	32%	32%*	44%	**78%**
Lend-dep rate (Y-o-Y % change)	18%	18%	18%	34%	28%	34%	28%	52%	**28%**
REER deviate from 36-month MA	4%	4%	4%	46%	46%	46%	46%*	48%	**40%**
Imports (Y-o-Y % change)	98%	98%	98%	62%	76%	62%	76%	50%	**76%**
Manufacturing production (Y-o-Y % change)	98%	98%	98%	48%	24%	48%	48%*	60%	**70%**
REER (Y-o-Y % change)	2%	2%	2%	42%	42%	42%	42%	46%	**42%**
Growth in nom credit dev from MA	96%	96%	96%	50%	34%	50%	50%*	52%	**66%**
Real M3 (Y-o-Y % change)	98%	98%	98%	58%	58%	58%	58%*	68%	**84%**
16 CDF PCA FSI (Y-o-Y % change)	100%	100%	100%	72%	72%	72%	72%	54%	**72%**
Growth in real total credit extended to private sector	84%	84%	84%	64%	64%	64%	64%*	62%	**72%**
FSI_16CDF_Ew_> 24-month avg.	92%	92%	92%	82%	84%	82%	88%	60%	**88%**
ALSI (Y-o-Y % change)	98%	98%	98%	70%	70%	70%	82%	40%	**78%**
Exports (Y-o-Y % change)	92%	92%	92%	54%	54%	54%	54%	50%	**64%**
6-mth avg. % change in R$	100%	100%	100%	40%	12%	40%	40%*	52%	**64%**

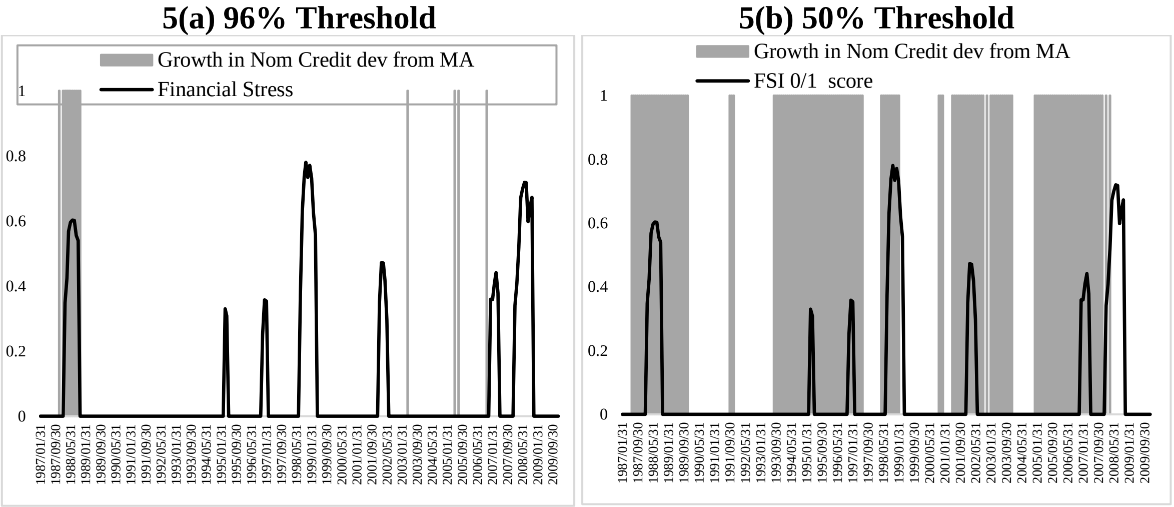
Note: A * indicates that not all 20 benchmarks were satisfied at that threshold.

Conversely, the next four criteria identified less-extreme thresholds and were often similar to each other.[15] Lastly, thresholds selected by the T1 = T2 error fell in the middle of the threshold spectrum, which was expected since Type I and II errors are usually similar in this region for most indicators. Results in Table 3 suggest that optimal threshold selection is not as clear-cut as simply minimising the NTSR. For instance, for South Africa, extreme

[15] Although the ROC curve and loss function 2 (see equation 3) criteria indicate the same thresholds in Table 3, this only occurs when policymaker preference is neutral (i.e. policymaker neutrality of $\mu = 0.5$ in equation 3 applies). If this preference is adjusted, the loss function 2 criterion will assign different preferences for Type I and II errors and it will identify different thresholds. For instance, if policymakers prefer few missed stress periods (i.e. $\mu = 0.9$) or minimal false signals (i.e. $\mu = 0.1$), the optimal threshold indicated by loss function 2 will be lower and higher respectively. Although excluded in Table 3, the same applies to thresholds identified by the relative useful measure (i.e. equation 4).

thresholds identified by the first three criteria often lead to few signals being issued by individual indicators and therefore high Type I errors. Figure 5(a) provides an example of this for the nominal credit growth variable, with a 96% optimal threshold as identified by the NTSR.[16] Alternatively, criteria that identify less-extreme thresholds (i.e. loss functions 2 and 4, the ROC curve or ASSM) can cause too many noisy signals. This is illustrated in Figure 5(b), where numerous signals were issued by the same nominal credit growth indicator at an optimal threshold of 50%, as suggested by criteria like the ROC curve.

Figure 5: Signalling frequency of nominal credit growth indicator at different thresholds



5(a) 96% Threshold

5(b) 50% Threshold

Source: Author's own compilation

Given the novelty of an EWS model that signals financial stress, the lack of evidence favouring the use of a particular criterion in Table 3 justified a 'middle ground' approach to be followed. In this regard, the ASSM again proved to be valuable. Section 4.2 clarified that the highest ASSM usually identifies the optimal threshold, but that 20 additional benchmark values (rows 1–20 in column 5) also had to be satisfied. If any of these benchmarks were not satisfied, that threshold was not considered. Instead, the optimal threshold was identified at the next highest ASSM where all 20 benchmarks were satisfied. For example, in Table 3, it is shown that the highest ASSM for the nominal credit growth indictor was achieved at the 50% threshold. This level was also indicated by the ROC and loss function 2 criteria, while

[16] However, there is merit in using higher thresholds for multivariate indicators. Evidence is presented in section 6.

52% was identified by the T1 = T2 error. However, at 50%, none of the 20 benchmarks from Table 2 were met and thus the next highest score where all benchmarks were satisfied was at 66%. This level was selected as ‘final’ optimal threshold and, as indicated in the last column of Table 3, this process often narrows the divide between threshold measures. Accordingly, the optimal signalling performance of indicators in Table 4 was measured at these ‘final’ thresholds.

Table 4 summarises the in-sample performance of indicators at their optimal thresholds. Ten contingency matrix performance measures were used to assess performance more robustly and to test for consistency. As depicted in Table A1, indicators in Table 4 were again ranked, but in this case, according to optimal performance. On average, indicators in Table 4 performed relatively well in the sample, with correct signals making up 47% of all signals, while the NTSR average was only 29%.

Compared to other EWS studies (e.g. Christensen & Li, 2014) these results were good, even though perhaps not directly comparable due to different variables (and transformations) and because those earlier results were based on cross-section data for 14 countries. The probability of a high financial stress period following within 24 months of a signal being issued averages 85%, with some indicators lying above 95% for this measure. The leading business cycle indicator has several measures equalling zero (i.e. noise, NTSR and B/(A+B)) because it issues very few signals (good or bad) in the sample. Accordingly, it has a high NTSR ranking, while the lower ASSM ranking recognises the lack of signals. Indicators in Table 4 have an average conditional probability of financial stress without a signal being issued (C/(C+D)) of 49%, which is high but acceptable for an individual indicator. The ratio of conditional to unconditional probability of correctly signalling a stress event {A/(A+B)/[(A+C)/(A+B+C+D)]} averages far above 100%, which confirms the usefulness of these indicators. Usefulness is also confirmed by the ratio of good-to-bad signals ((A+D)/(B+C)) of above 100% for all variables. Similar results were confirmed by other measures not shown in Table 4, including the measures of usefulness and relative usefulness and the false alarm ratio (B/(A+B)) that averaged only 15% of total signals issued.

Table 4: EWS indicator performance at optimal thresholds (in-sample period Jan. 1986–Dec. 2009)

Variable name	A/(A+C) (signal) [1]	B/(B+D) (noise) [2]	NTSR [3]	A/ (A+B) [4]	CP/UP of Crisis[17] [5]	C/ (C+D) [6]	(A+D)/ (B+C) [7]	Loss function 2 [8]	Loss function 4 [9]	ASSM [10]	Rankings										
											[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	Avg. Rank
24-month ALSI avg. real return	59%	9%	16%	91%	148%	42%	245%	25%	18%	501%	5	7	4	4	4	3	2	2	2	2	4
Federal funds rate (Y-o-Y % change)	75%	23%	31%	84%	136%	34%	312%	24%	18%	523%	1	12	8	8	8	1	1	1	1	1	4
36-month ALSI avg. real return	44%	5%	11%	94%	153%	48%	179%	30%	19%	437%	12	3	3	3	3	6	6	4	3	3	5
REER (Y-o-Y % change)	67%	24%	36%	81%	133%	40%	241%	28%	20%	420%	2	14	10	10	10	2	3	3	4	4	6
Lend-dep rate (Y-o-Y % change)	46%	8%	18%	90%	146%	48%	176%	31%	20%	403%	10	6	6	6	6	7	7	6	5	5	6
Business cycle indicator (2-year change)	26%	0%	0%	100%	163%	54%	121%	37%	21%	393%	15	1	1	1	1	15	15	10	7	6	7
REER deviate from 36-month MA	63%	23%	37%	81%	132%	43%	214%	30%	20%	382%	3	12	11	11	11	4	4	5	6	7	7
Growth in real total credit extended to private sector	60%	22%	38%	81%	132%	45%	200%	31%	21%	364%	4	10	12	12	12	5	5	7	8	8	8
Imports (Y-o-Y % change)	36%	7%	18%	90%	147%	52%	140%	35%	21%	353%	14	5	5	5	5	11	12	9	8	10	8
FSI_16CDF_Ew_> 24-month avg.	22%	1%	4%	97%	159%	55%	108%	40%	22%	355%	17	2	2	2	2	16	16	16	11	9	9
16 CDF PCA FSI (Y-o-Y % change)	48%	15%	31%	84%	136%	49%	165%	34%	21%	342%	9	8	9	9	9	8	8	8	8	11	9
ALSI (Y-o-Y % change)	22%	5%	21%	88%	144%	56%	103%	41%	22%	287%	16	3	7	7	7	17	17	17	16	12	12
Real M3 (Y-o-Y % change)	44%	19%	43%	79%	129%	52%	140%	37%	22%	266%	13	9	13	13	13	13	12	13	12	13	12
Manufacturing production (Y-o-Y % change)	50%	24%	49%	76%	125%	51%	149%	37%	22%	257%	6	14	15	15	15	9	9	11	13	14	12
6-month avg. % change in R$	50%	24%	49%	76%	125%	51%	149%	37%	22%	257%	6	14	15	15	15	9	9	11	13	14	12
Growth in nom credit dev. from MA	46%	22%	49%	76%	125%	52%	140%	38%	22%	246%	10	11	14	14	14	14	12	14	15	16	13
Exports (Y-o-Y % change)	48%	24%	51%	76%	124%	52%	142%	38%	22%	242%	8	14	17	17	17	12	11	15	17	17	15
Average	**47%**	**15%**	**29%**	**85%**	**139%**	**49%**	**172%**	**34%**	**21%**	**355%**											

[17] Conditional probability of a crisis (A/(A+B)) divided by the unconditional probability of a crisis (A+C)/(A+B+C+D).

Type II errors averaged only 15% and were comparable to results from Kaminsky (1999), while Type I errors were generally less than those obtained by other EWS research, although it was still not ideal from a policymaker perspective.[18] These low errors also allow for excellent policymaker loss minimisation as indicated by low loss functions (criteria 8 and 9). For each indicator in Table 4, individual criteria rankings often differ, although most generally indicate the same top performers.

An interesting result is that the NTSR criterion did not consistently provide the same rankings as the average of all ten rankings in the last column. Contrary to this, the ASSM (criteria 10) identified top indicators most consistently and again confirmed its superior evaluation properties as was also shown in Table A1. Overall, indicators in Table 4, as first filtered according to overall signalling ability in Table A1, exhibit relatively good optimal signalling performance, as indicated in Table 4. These results therefore add weight to the assertion made in section 5 that such an initial performance filter should be implemented.[19]

Apart from average and optimal performance analysis reported in Table A1 and Table 4 respectively, visual analysis of indicator behaviour, before and after high-stress periods, also provided evidence of signalling performance. Credit growth, for instance, is often cited as a good indicator, and in Figure 6(a), the signalling ability of real credit growth to the private sector in South Africa confirms this. The left-hand graph indicates that credit growth accelerates prior to high-stress periods and subsides thereafter. Similarly, the right-hand graph shows how a growing credit growth trend in the months prior to stress events reverses in the post-stress months.

Care should however be taken when interpreting indicator signals in EWS models. Finding good indicators often requires that various variable transformations, trend lengths and deviations from those trends be considered, which might have implications in terms of interpreting the results. For instance, some of the thresholds in Table 3 fall in what might be considered the opposite distributional tail than reported in earlier bank or currency crisis EWS literature.

[18] It should be noted that stress periods made up 15% of the sample in the current study and therefore the study was probably not directly comparable to other bank or currency crisis research.

[19] Indicators with low average scores in Table A1 also show worse optimal performance than the top performers.

Figure 6: Examples of indicator performance before and after high-stress periods

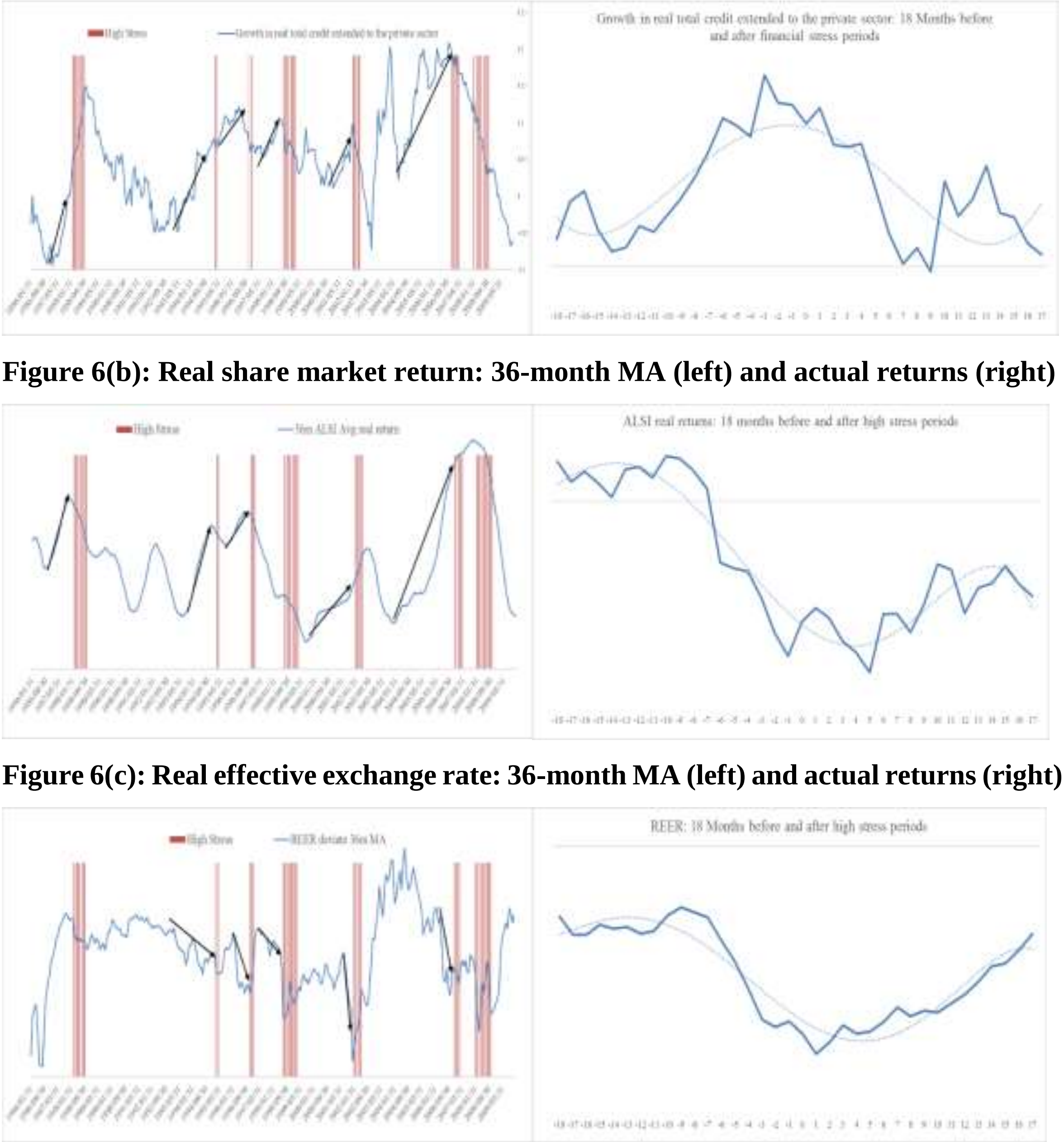


Figure 6(a): Growth in real total credit extended to the private sector

Figure 6(b): Real share market return: 36-month MA (left) and actual returns (right)

Figure 6(c): Real effective exchange rate: 36-month MA (left) and actual returns (right)

Note: The three right-hand graphs show indicator performance in a 36-month window around high-stress periods (divided equally into 18 months before and after high-stress periods). Indicator levels in both 18-month periods are measured against an indicator's average level during tranquil periods (i.e. excluding those months around high-stress periods).

Source: Author's own compilation

An example is the lower-tail REER variable thresholds in Table 3, which suggest that rand weakness is an EWS. This is contrary to the other EWS research (e.g. Kaminsky & Reinhart, 1999; Frankel & Saravelos, 2012), which found currency overvaluation a warning signal.[20] Similarly, the ALSI moving average indicators have thresholds in the upper tail, implying that persistently high average real equity returns precede periods of financial stress. This again seems contradictory to conventional EWS results that classify lower tail returns (i.e. bursting equity bubbles) as warning indicators (e.g. Kaminsky & Reinhart, 1999).

Potential explanations for these apparent anomalies can be given though. First, financial stress, as measured by the FSI in this study, included share returns and the REER as input variables, meaning that when share prices or the REER fall there is a contemporaneous rise in the FSI. Therefore, when share prices fall or the currency depreciates dramatically, such data does not pre-empt periods of financial stress, since it is already too late. Second, interpretation of variables also depends on the timeframe of analysis and method of variable transformation. For instance, the left-hand graph in Figure 6(b) indicates that the 36-month MA of the ALSI usually increases in the months prior to financial stress. This suggests that the asset return build-up phase should be analysed for warning signals, and provides some confirmation of the notion that periods of financial stability might be misleading (i.e. Minsky's financial instability hypothesis [see Minsky, 1977] or the paradox of instability as phrased by Borio & Drehmann, 2009b).

In particular, fragility can build up during tranquil periods, as good and sustained asset market performance leads to excessive leverage build-up, risk taking and unsustainable periods of irrational exuberance that sow the seeds of impending financial stress (Bhattacharya, Goodhart, Tsomocos & Vardoulakis, 2015; Borio & Drehmann, 2009a; Lo Duca & Peltonen, 2011: 22; Patel & Sarkar 1998).[21] Rose and Spiegel (2009) also found that countries experiencing large share market growth (relative to output) between 2003 and 2006 were more likely to be hit by the 2008 crisis. Since the 36-month ALSI average is a longer-term measure, it indicates changing share market performance only gradually, but when shorter performance is measured (i.e. ALSI annual real returns) the results look different. Accordingly, the right-hand graph in Figure 6(b) shows that South Africa is characterised by marginal positive real returns in the months leading up to financial stress, while it

[20] EWS research often uses the bilateral real exchange rate against the dollar (not the REER) to measure deviation from trend.

[21] It is also possible that EWS research using data prior to the 2000s could yield different results to data after 2000, which would confirm the concern raised by Borio and Drehmann (2009b: 16) and Aldasoro *et al*., (2018: 43) that past relationships might not hold in future.

deteriorates rapidly from about nine months prior to stress events. This graph therefore supports the notion expressed in EWS literature that falling share prices precede crisis periods.[22]

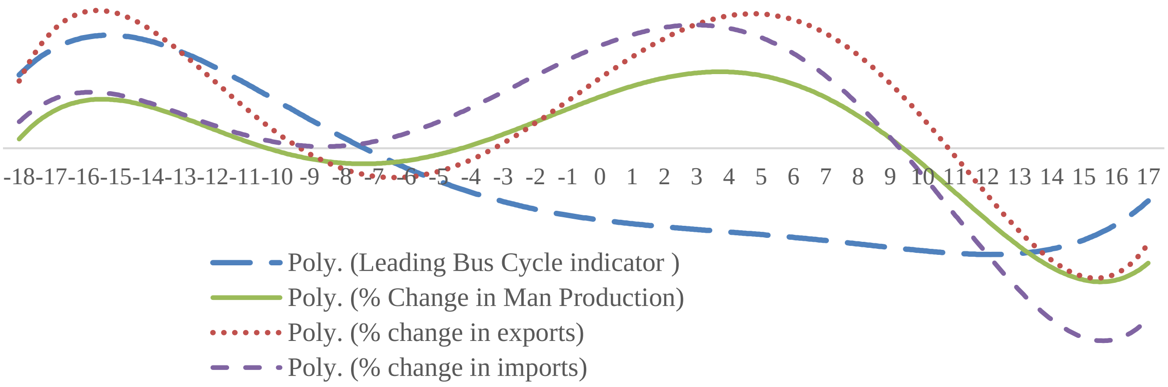
In terms of interpreting a weakening REER as a signal, the conventional view suggests that currency undervaluation might prevent financial crises due to export promotion and limiting of current account deficits (e.g. Gala, 2008). However, in South Africa, currency depreciation and current account deficits often have a positive relationship (Bank for International Settlements [BIS], 2014: 324) and therefore both can contribute to financial stress. Gadanecz and Jayaram (2009: 368) indicate that exchange rate under- and overvaluations could lead to financial instability, which suggests that both results could provide EWS signals. For South Africa, the left-hand graph in Figure 6(c) confirms that REER deterioration (i.e. deviation below the 36-month MA) usually precedes high-stress periods.[23] Again, this is a long-term measure that shows changes only slowly. However, the right-hand graph in Figure 6(c), which is a short-term measure of REER performance, also indicates that REER deterioration gathers momentum about nine months prior to stress periods, before recovering rapidly afterwards.[24] The evidence in Figures 6(b) and 6(c) suggests that variable transformation plays an important role in EWS research, and the interpretation of these measures should not be simply mechanical, but should include policymaker judgement (also see Pasricha *et al.*, 2013: 18).

However, visual analysis does not always present clear results. As shown in Figure 7, the signalling ability of real sector indicators like imports, manufacturing and exports is unclear for South Africa, which again emphasises the importance of judgment. These indicators have relatively good overall and optimal results depicted in Table A1 and Table 4 respectively, but their smoothed values in Figure 7 suggest that they do not exhibit clear behavioural changes in the 18-month period before and after financial stress events. The leading business cycle (two-year change) seems to reach high growth rates approximately 16 months prior to stress periods, after which it decreases for roughly a two-year period. However, the threshold adjustment for this indicator (in Table 3) is large, which casts doubt regarding its consistent

[22] If the period prior to high-stress is extended (e.g. to 28 months) the decreasing trend is even more visible.
[23] Also see De Jager (2012: 15) for estimates of REER undervaluation of the rand during similar periods as identified in the left-hand panel in Figure 6(c).
[24] Furthermore, EWS research such as by Frankel and Saravelos (2012), considered REER values in a five-year period prior to crisis, while the focus in this study was limited to 18 months prior to a crisis.

signalling ability. Similarly, the lack of signals (shown in Table 4) suggests that the business cycle, together with other real sector variables reflected in Figure 7, should be treated with circumspection in South African early warning models.

Figure 7: Real sector variables: Behaviour before and after high-stress periods



Source: Author’s own compilation

Although some individual indicators in Table 4 have good signalling ability, it will be unwise to make policy decisions based solely on univariate indicator results. Kaminsky (1999) suggests that signalling ability of multivariate indicators is likely to be more credible and informative than univariate indicators and the next section describes the composition and in-sample performance of three types of composite indicators.

6.3. Constructing composite EWS indicators

The probability of a crisis occurring increases when multiple indicators issue signals simultaneously. Therefore, using a variable composite indicator (VCI) is a well-established process in EWS models (see Edison, 2003; Kaminsky 1999; Kaminsky *et al*., 1998). The construction of these multivariate models does not have to be complicated, and often a simple aggregation of signals can prove useful. This section describes the construction and performance evaluation of three VCIs originally proposed by Kaminsky *et al.* (1998) and adapted here for use with a financial stress index. The first composite indicator (VCI_t^1) is a simple aggregate indicator, consisting of *n* individual indicators *i* at time *t*. Formally, the index is defined in equation 8 as:

$$VCI_t^1 = \sum_{i=1}^{n} S_t^i \qquad \textbf{[8]}$$

As before, an indicator *i* only signals a warning $\left(S_t^i = 1\right)$ when its historical observations X_t^i satisfy $(\left|X_t^i\right| > \theta^i)$. Accordingly, the value of VCI_t^1 can range between a minimum of 0, when all $S_t^i = 0$, and a maximum of *n*, when all $S_t^i = 1$. A potential drawback of VCI_t^1 is that it assigns equal weight to every indicator. Therefore, excellent indicators and extreme signals are equally important as poorer indicators and milder signals. To adjust for this, two additional composite indicators are constructed (see equations 9 and 10). The second composite indicator distinguishes between mild and extreme signals and assigns extra weight to extreme measures as follows:

$$VCI_t^2 = \sum_{i=1}^{n}(SM_t^i + 2SE_t^i) \qquad \textbf{[9]}$$

In this case, mild signals ($SM_t^i = 1$) are issued if an indicator value lies between its specified mild and extreme thresholds ($\left|\theta_M^i\right| < \left|X_t^i\right| < \left|\theta_E^i\right|$). Here θ_M^i is the same optimal threshold as in I_t^1, while ($\left|X_t^i\right| > \left|\theta_E^i\right|$) are extreme observations in the distribution of the indicator (e.g. top 5%), so that ($SE_t^i = 1$). Lastly, composite indicator 3 distinguishes between poor and excellent indicators as follows:

$$VCI_t^3 = \sum_{i=1}^{n}\frac{S_t^i}{\omega^i} \qquad \textbf{[10]}$$

This indicator differs from VCI_t^1 in that it assigns a bigger weight to indicators with better forecasting accuracy, as measured by its NTSR (ω^i). Evidence in section 6.3.1 indicate that these three composite indicators do in fact significantly improve in-sample signalling ability.

6.3.1. In-sample overall signalling performance of composite indicators

For more robust results, two versions are constructed for each type of composite indicators described above (i.e. six VCIs in total). First, a group of ten indicators from Table 4 is used to construct the first set of VCIs and are called 10VCIs for short. This is done to test the signalling ability of a relatively simple composite indicator, consisting of limited but diverse inputs, that do not use only the best in-sample univariate indicators. To avoid signal duplication, the ten indicators are selected to avoid the inclusion of similar indicators (i.e. duplication of share return or credit growth measurement).

Second, for comparative purposes, composite indicators consisting of all seventeen top-performing variables in Table 4 (called 17VCIs for short) are also constructed. The 17VCIs indicate how signalling ability might change if more variables are included and signal duplication is ignored.

The overall in-sample performance of all the VCIs is exceptional, as illustrated by the ROC curves in Figure A1 in the Appendices section. Compared to individual indicators (see Table 3) all VCIs have superior overall signalling ability, with higher AUROCs (average of 94%), lower NTSRs (average of 29%) and smaller loss functions. The relative usefulness of VCIs is also superior (average of 42%) and they have higher probabilities of correctly signalling periods of financial stress (A/(A+B) of 87% on average). As for individual indicators, the best indication of the VCIs overall signalling performance is their aggregate signalling score measure (ASSM average of 517%). The overall in-sample performance of all six VCIs suggests that the 17VCIs are not superior to the 10VCIs, even though they include all the best individual indicators. However, optimal threshold performance is more important and, as indicated below, all six VCIs performed quite well.

6.3.2. In-sample optimal signalling performance of composite indicators

As was the case for individual indicators (see Table 3), eight threshold selection criteria identify potential optimal thresholds for VCIs as depicted in Table 5. Again, the first three criteria in Table 5 identify higher optimal thresholds than the last five and, as elaborated below, this has implications for policymaking. The robustness of the ASSM again makes it the preferred criterion to select the final threshold (see last column in Table 5) for optimal performance analysis as depicted in Table 6.

As with individual indicators (see Table 4), the optimal threshold performance of VCIs is summarised in Table 6. The first nine criteria in Table 6 indicate that VCIs perform better than individual indicators, although the variability of results obtained by individual indicators as reflected in Table 3 complicates comparison. For example, the measures for Type II error, NTSR, A/(A+B) and conditional probability for VCIs are similar to those of some top-performing individual indicators.

Table 5: Potential optimal thresholds for composite indicators (in-sample period Jan. 1986–Dec. 2009)

Variable name	NTSR	A/(A+B)	(A/A+B) ÷ [(A+C)/ (A+B+C+D)]	Loss function 2	Loss function 4	ROC curve	ASSM	T1 = T2 error	Final optimal threshold
10-Composite Indicator 1 (10VCI^1)	90%	90%	90%	46%	46%	46%	46%	46%	46%
10-Composite Indicator 2 (10VCI^2)	84%	84%	84%	44%	44%	44%	44%	46%	44%
10-Composite Indicator 3 (10VCI^3)	84%	84%	84%	52%	52%	52%	52%	44%	52%
17-Composite Indicator 1 (17VCI^1)	82%	82%	82%	46%	46%	46%	46%	46%	46%
17-Composite Indicator 2 (17VCI^2)	90%	90%	90%	52%	52%	52%	52%	46%	52%
17-Composite Indicator 3 (17VCI^3)	86%	86%	86%	48%	46%	48%	46%	44%	46%

Table 6: Performance of composite indicators at optimal thresholds (in-sample period Jan. 1986–Dec. 2009)

Variable name	A/(A+C) (signal)	B/(B+D) (noise)	NTSR	A/(A+B)	Conditional / unconditional probability of crisis	C/(C+D)	(A+D)/(B+C)	Loss function 2	Loss function 4	ASSM
10-Composite Indicator 1 (10VCI^1)	85%	15%	18%	90%	147%	22%	573%	15%	12%	867%
10-Composite Indicator 2 (10VCI^2)	89%	20%	22%	88%	143%	18%	590%	15%	12%	867%
10-Composite Indicator 3 (10VCI^3)	79%	8%	11%	94%	153%	27%	513%	15%	13%	830%
17-Composite Indicator 1 (17VCI^1)	84%	13%	16%	93%	152%	23%	573%	15%	12%	875%
17-Composite Indicator 2 (17VCI^2)	78%	9%	12%	87%	142%	28%	487%	16%	13%	796%
17-Composite Indicator 3 (17VCI^3)	89%	9%	11%	94%	153%	16%	852%	10%	9%	1190%

The superiority of VCIs is more clearly visible in their smaller loss functions and lower conditional probabilities (C/(C+D)) and a higher ratio of good-to-bad signals (i.e. (A+D)/(B+C) > 480%). However, VCI superiority is most clearly visible in their ASSM values (i.e. shown in the last column of Table 6), which are more than double the score achieved by most individual indicators. The latter result not only confirms the usefulness of VCIs, but again emphasises the usefulness of this aggregate scoring mechanism for EWS analysis as well.

Although the optimal performance of VCIs as shown in Table 6 is promising, the two different optimal threshold groupings identified in Table 5 are problematic from a policymaker perspective. As shown in Figure A2 in the Appendices section, VCIs issue a multitude of signals (i.e. grey areas) at optimal thresholds derived from the lower threshold set depicted in Table 5 (i.e. columns 5–9). These signals are too noisy for policymaking since, contrary to portfolio management where optimal portfolios are preferred to risk-minimised portfolios, prudential policymakers cannot afford too much noise. Hence, policymakers will likely prefer higher thresholds that minimise noise and not thresholds that optimise the true-to-false signal ratio. Therefore, signals at higher thresholds, as identified by the first three criteria depicted in Table 5, should also be considered in EWS models. The merits of this are visible in Figure A3 in the Appendices section, where fewer signals are issued by VCIs at thresholds where the NTSR is minimised. In this case, most signals are good in-sample policy triggers, since VCIs tend to issue them only before high-stress periods. As with individual indicators, these high thresholds cause fewer VCI signals, but unlike individual indicators, the VCIs still manage to issue numerous signals due to enhanced signalling ability from being multivariate. The results suggest that stricter threshold criteria shown in Table 5 (i.e. columns 2–4) are perhaps more appropriate for signal analysis with composite indicators, while the other criteria shown in Table 5 (i.e. columns 5–9) are more appropriate for identifying optimal thresholds for individual indicators.

The difference in signalling frequency depicted in Figures A2 and A3 also suggests that information can be gained from both, which implies that a relatively wider threshold range could be considered for policymaking purposes. A similar proposal by Borio and Drehmann (2009a: 44) states that policymakers could use such a range to determine appropriate policy intervention triggers. Figures A2 and A3 confirm that policymakers could perhaps use a threshold range between the two sets of thresholds in Table 5. For example, policymakers could construct a type of ‘heat map’ where signals issued at higher thresholds are indicated as ‘hotter’ than those issued at lower thresholds (see Aldasoro *et al*. [2018: 41]). This process of issuing signals along a threshold continuum could also be a useful communication tool in the macroprudential policy toolkit. Furthermore, as shown in the out-of-sample results below, signals from a range of thresholds are also useful during tranquil periods when few periods of financial stress are predicted. For policymakers, the issue of threshold choice complicates policymaking. However, compared to individual indicators, VCIs have a significant advantage in that the optimal threshold is not even required to extract useful signalling information. This attribute is discussed next where it is indicated that a VCI also provides early warning information purely from its composite value, with larger values signifying a higher conditional probability of imminent financial stress.

6.3.3. Composite indicators and the conditional probability of financial stress

At the start of section 6.3 the assertion was made that more signals occurring simultaneously should, in theory, provide more accurate probability forecasts of financial stress. Accordingly, if a composite indicator *VCI* has a high value at time *t*, the probability of imminent stress should be higher than if it has a low value. This should be true for all three types of VCIs as indicated in section 6.3 above. To test this, each type of VCI is analysed according to the following conditional probability, as adapted from Kaminsky (1999), to forecast financial stress:

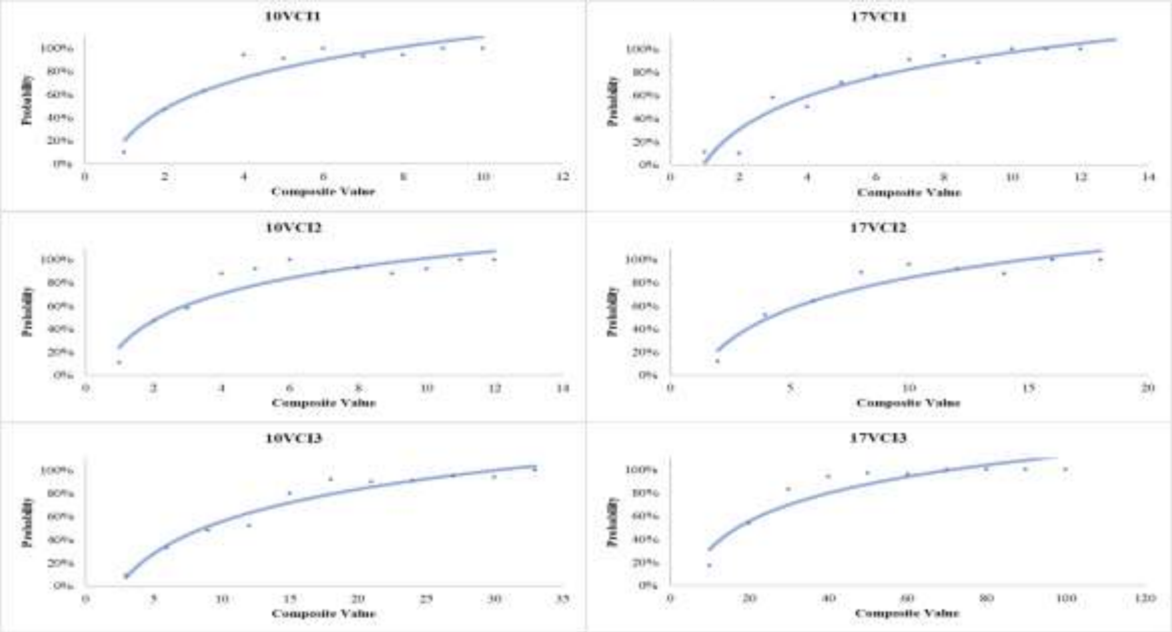
$$P\left(hfs_{t,t+h} \middle| a_k < VCI_t^k < b_k\right) = \frac{\text{months } with\ a_k{<}VCI_t^k{<}b_k\ and\ hfs_{t,t+h}}{\text{months } with\ a_k{<}VCI_t^k{<}b_k} \qquad \textbf{[11]}$$

where *P* is the probability that a high financial stress period, *hfs*, occurs within the time interval *t*,*t*+*h*, conditional on the type of composite indicator VCI^k, with *k* = 1, 2 and 3, issuing a signal at time *t* when its value lies within an interval a_k–b_k. Here, a_k and b_k denote a specific lower and upper bound for each type of indicator VCI^k and the time interval *h* equals 24 months. Due to difference in construction the values of VCI_t^k differ and their maximum values will result in $VCI_t^1 < VCI_t^2 < VCI_t^3$. Accordingly, the size and number of

a_k–b_k intervals for each VCI_t^k differs.[25] Nevertheless, the value of each VCI^k increases as more signals are issued by its component indicators (X_t^i). As the value of VCI^k increases, it moves to a higher *a_k–b_k* interval, which should increase the conditional probability *P* of financial stress. This probability is measured as the frequency of *hfs* months that follows within *h* periods after indicator VCI^k issues a signal, relative to the total number of months that VCI^k issues a signal. Figure 8 illustrates the conditional probabilities for all three types of VCIs (both 10- and 17-variable versions).

Figure 8: Conditional probability of VCIs correctly forecasting financial stress

8(a) 10-VCIs 1, 2 and 3 **8(b) 17-VCIs 1, 2 and 3**

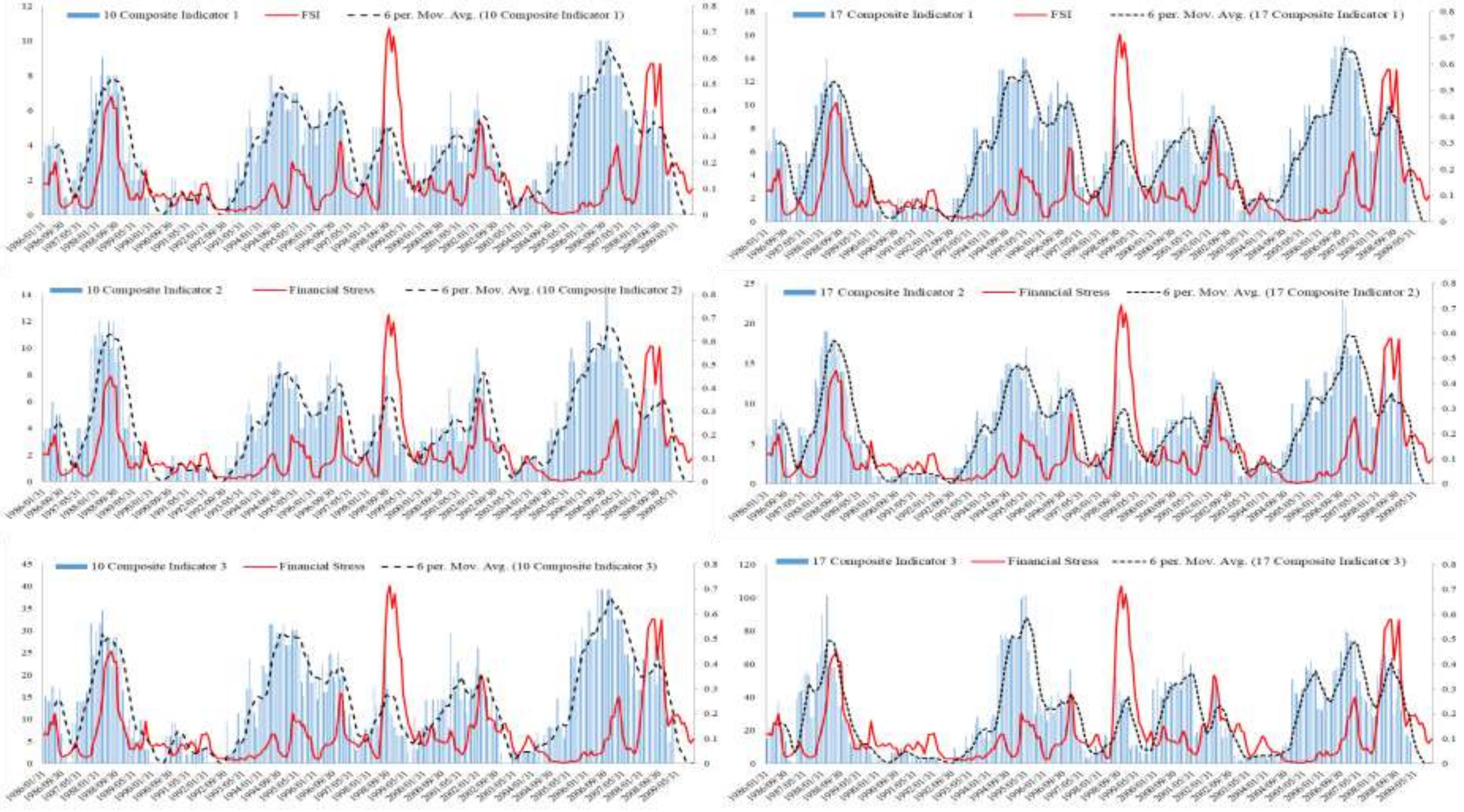


Source: Author’s own compilation

Policymakers should not only predict stress periods timely and accurately, but also communicate such predictions throughout. An example of this in Figure 9 shows how the VCIs and their n-period MA (6-month MA in this case) can be superimposed onto the financial stress index. During the in-sample period, higher VCI values and MAs tend to precede financial stress periods, which justifies their use as a macroprudential communication tool. Importantly, optimal VCI thresholds are not required to construct the VCI graphs as illustrated in Figure 9.

[25] Similarly, the maximum values of the 17VCIs will differ from those of the 10VCIs since more individual component indicators are included in the 17VCIs, thereby increasing their potential maximum value.

Figure 9: EWS composite indicator values and MAs compared to financial stress index (in-sample period Jan. 1986–Dec. 2009)



9(a) 10-VCIs 1, 2 and 3

9(b) 17-VCIs 1, 2 and 3

Another advantage of this is that, again without selecting a threshold, a long-term average for any VCI can also be used to distinguish strong signal periods (i.e. VCI value above its long-term average) from weak signal periods. Combined with a threshold range in Figures A2 and A3, Figure 9 can be used to create a signal 'heat map', as mentioned in section 6.3.2, with hotter signals indicated as the VCI moves further above its average.[26] Although the in-sample results above are promising, a good EWS model must also perform well out of sample. Accordingly, the out-of-sample predictive ability of VCIs is assessed next.

6.4. Out-of-sample indicator performance

The out-of-sample period fell between January 2010 and September 2018 and, as indicated in Figure 11, only two periods of financial stress were observed during this period (2011–2012 and 2015–2016). Based on financial stress index (FSI) values, these are perhaps better described as periods of elevated stress, with the 2011–2012 period not even registering when stress periods are limited to 10% of observations. Given the absence of high financial stress (*hfs*) periods, it is expected that a good EWS model will issue few signals, and the a priori expectation here is thus for relatively weak signalling results. Since the in-sample results suggested that multivariate indicators outperform individual indicators, the out-of-sample results shown in Table 7 mostly reflect overall performance of the VCIs. However, for comparative purposes, results for three individual indicators are also shown.

The results for all indicators in Table 7 are clearly worse than the in-sample results. The average AUROC is about 20% less, while the NTSR is 10% higher and the conditional probability of correctly signalling stress (A/(A+B)) is about 20% lower. On average, the VCIs are only slightly useful, as confirmed by high loss functions and low usefulness scores. Although only three individual indicators are shown in Table 7, the performance of all individual indicators was also inferior to in-sample results. Based on overall performance, the VCIs in Table 7 are not superior to individual indicators as was the case in the sample. The optimal threshold performance results in Table 8 also suggest that VCIs perform similar to the individual indicators, although this assessment differs, depending on which criterion is used. As before, the ASSM, introduced in section 4.2, gives the clearest indication of signalling performance, with $10VCI^3$ and $17VCI^3$ performing slightly better than the others.

[26] This long-term average would be similar to the relatively low VCI optimal thresholds indicated in Table 6.

Table 7: Out-of-sample performance of individual and composite indicators over entire threshold spectrum (Jan. 2010–Sept. 2018)

Variable Name	**Selection Criteria** **Scores for criteria (2-6) are calculated as average over all thresholds**				
	AUROC	**NTSR**	**Relative useful**	**A/A+B**	**ASSM**
10-Composite Indicator 1 (10VCI^{1})	70%	38%	13%	58%	212%
10-Composite Indicator 2 (10VCI^{2})	70%	38%	13%	63%	221%
10-Composite Indicator 3 (10VCI^{3})	74%	37%	17%	64%	247%
17-Composite Indicator 1 (17VCI^{1})	70%	34%	13%	57%	217%
17-Composite Indicator 2 (17VCI^{2})	70%	33%	13%	58%	223%
17-Composite Indicator 3 (17VCI^{3})	85%	27%	23%	67%	322%
Average	**73%**	**35%**	**15%**	**61%**	**240%**
ALSI 24-mth avg. real return	69%	34%	18%	55%	237%
FSI > 24-month average	60%	69%	8%	66%	141%
% change in FSI	64%	55%	14%	66%	191%

Table 8: Out-of-sample performance of individual and composite indicators at final optimal threshold (Jan. 2010–Sept. 2018)

Variable Name	**A/(A+C) (signal)**	**B/(B+D) (noise)**	**NTSR**	**A/(A+B)**	**Cond prob. / uncond. crisis prob.**	**C/(C+D)**	**B/(A+B)**	**Loss function 2**	**Loss function 4**	**ASSM**	**(C+D)/ (A+B)**	**(A+D)/ (B+C)**
10-Composite Indicator 1 (10VCI^{1})	57%	25%	44%	71%	136%	39%	29%	34%	22%	314%	137%	190%
10-Composite Indicator 2 (10VCI^{2})	57%	31%	54%	67%	128%	41%	33%	37%	23%	253%	123%	170%
10-Composite Indicator 3 (10VCI^{3})	51%	13%	25%	82%	155%	38%	18%	31%	21%	418%	205%	214%
17-Composite Indicator 1 (17VCI^{1})	33%	16%	50%	69%	131%	47%	31%	42%	24%	211%	300%	132%
17-Composite Indicator 2 (17VCI^{2})	21%	4%	17%	87%	165%	48%	13%	41%	23%	332%	673%	132%
17-Composite Indicator 3 (17VCI^{3})	57%	4%	6%	95%	180%	33%	5%	23%	17%	626%	214%	314%
ALSI 24-mth avg. real return	11%	0%	0%	100%	190%	50%	0%	44%	23%	378%	1557%	115%
FSI > 24-month average	25%	4%	15%	88%	168%	46%	12%	40%	23%	357%	582%	142%
% change in FSI	33%	9%	28%	80%	152%	45%	20%	38%	23%	321%	364%	152%

Although all VCIs performed worse out of the sample, their results in Table 8 are comparable to results presented by Christensen and Li (2014) for 13 OECD countries and some positive aspects are noticeable, namely that, the VCIs:

- have high conditional probabilities of correctly signalling stress;
- are useful as measured by conditional to unconditional probability of correctly signalling a stress event (A+C)/(A+B+C+D);
- have a low percentage of false alarms (B/(A+B));
- have a good-to-bad signal ratio ((A+D)/(B+C)) above 100%; and
- mostly show low Type II errors and a high ratio of no signals to signals ((C+D)/(A+B)) and given that VCIs have low optimal thresholds, this low signal frequency confirms the relative tranquil out-of-sample period.

The out-of-sample performance of the VCIs in Table 8 is not exceptional though, and can be attributed to the following three important factors.

- The fact that few *hfs* periods occurred during this period implied that early warning signals had to be rare during this period. Given this interpretation, it is not surprising that the VCIs generally exhibited a high ratio of no signals to signals in Table 8.

- No EWS system will accurately predict all stress periods, particularly in the case of a small, open country, such as South Africa, where external factors often influence domestic financial stability. It would therefore be naïve to expect VCIs to predict such externally generated or politically driven, stress events by using only real economic and financial variables. The most that can be expected from such EWS models is that they would signal potential *hfs* periods at extreme thresholds, but also possibly less severe stress at lower thresholds. Examples of the 10VCIs in Figure 10 suggest that the VCIs manage to do this. In particular, during this tranquil sample period, the 10VCIs (and 17 VCIs) issued signals only near their historical optimal thresholds, which were quite low (see Table 5). Conversely, only a few signals were issued at higher thresholds (e.g. 60%), while no signals were issued at even higher thresholds in the range usually identified by the NTSR (e.g. 80%).[27] Since signals were mostly issued at low thresholds, as shown in Figure 10, it suggests that the probability for financial stress

[27] The 17VCIs have similar results and are therefore omitted from Figure 10.

was also low during this out-of-sample period. Accordingly, if stress was to occur, the severity would have been expected to be limited as well. Similarly, the absence of high-threshold warning signals during the out-of-sample period is a good result, because no serious stress events occurred. Another promising result obvious from Figure 10 is that no false positive signals were issued at higher thresholds. However, some were issued at lower historical optimal thresholds, which was also the third reason for the relatively weak signalling results indicated in Table 8.

- As mentioned above, low optimal VCI thresholds are problematic for policymaking. Policymakers should take care not to over-fit or rely solely on a single threshold, as this might lead to spurious precision and disappointing out-of-sample performance. A more prudent and credible approach for policymakers would be to analyse out-of-sample signalling information over a threshold range (cf. Borio & Drehmann, 2009a: 34, 44). Doing this for South Africa improves the out-of-sample performance of all VCIs significantly. For instance, if the threshold for $10VCI^1$ is adjusted to 60%, its out-of-sample results improve, with the NTSR dropping to 8%, A/(A+B) increasing to 96% and the ASSM almost doubling. Figure 10 also indicates that at the 60% threshold, all three 10VCIs perform better, with less noisy signals.[28]

Results for the three 10VCIs shown in Figure 11 confirm the lack of extreme signals during the out-of-sample period, although the 6-month MAs of all three 10VCIs seem to edge upwards in the months prior to the 2015–2016 stress period.[29] Figure 11 indicates that the VCIs increased marginally and would not have provided policymakers with enough evidence to justify major policy intervention. Compared to the historical conditional probabilities indicated in Figure 8, the VCIs in Figure 11 rarely reached levels that would provide policymakers with the required proof to make significant policy changes. For this period, the only action they could perhaps have taken was to communicate that a gradual increase in the VCIs had been detected (e.g. October 2014–July 2015, 2011–2012 and 2015–2016) and to encourage participants to apply more measured discretion in transactions and risk-taking.

[28] Except for using a threshold range, another potential solution for low historical VCI thresholds might be to update their optimal thresholds dynamically as new data becomes available, but this fell outside the scope of the current study.

[29] Again, the 17VCIs had similar results and are therefore omitted from Figure 11.

Figure 10: 10VCI out-of-sample signals at various upper threshold ranges (Jan. 2010–Sept. 2018)

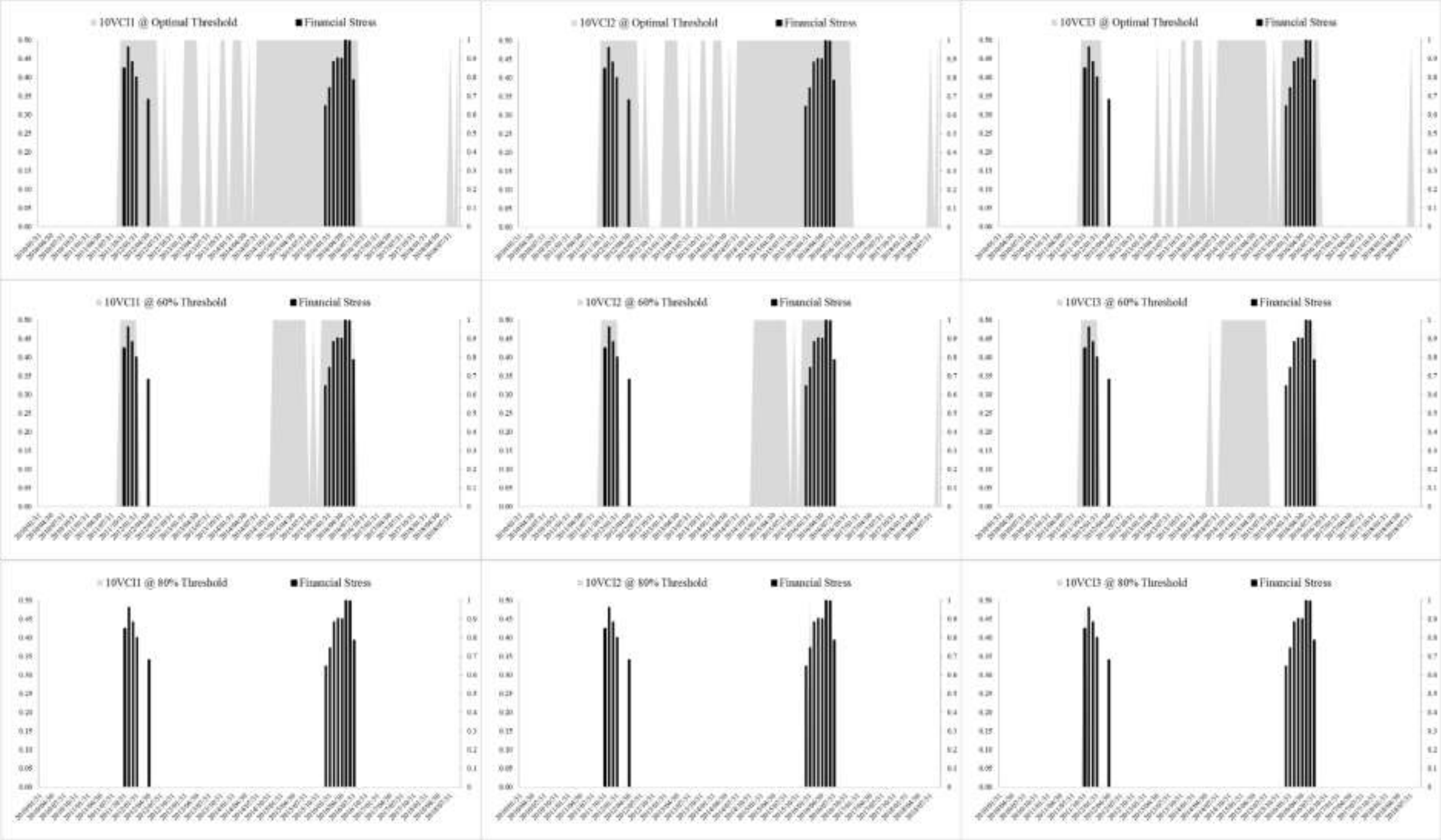
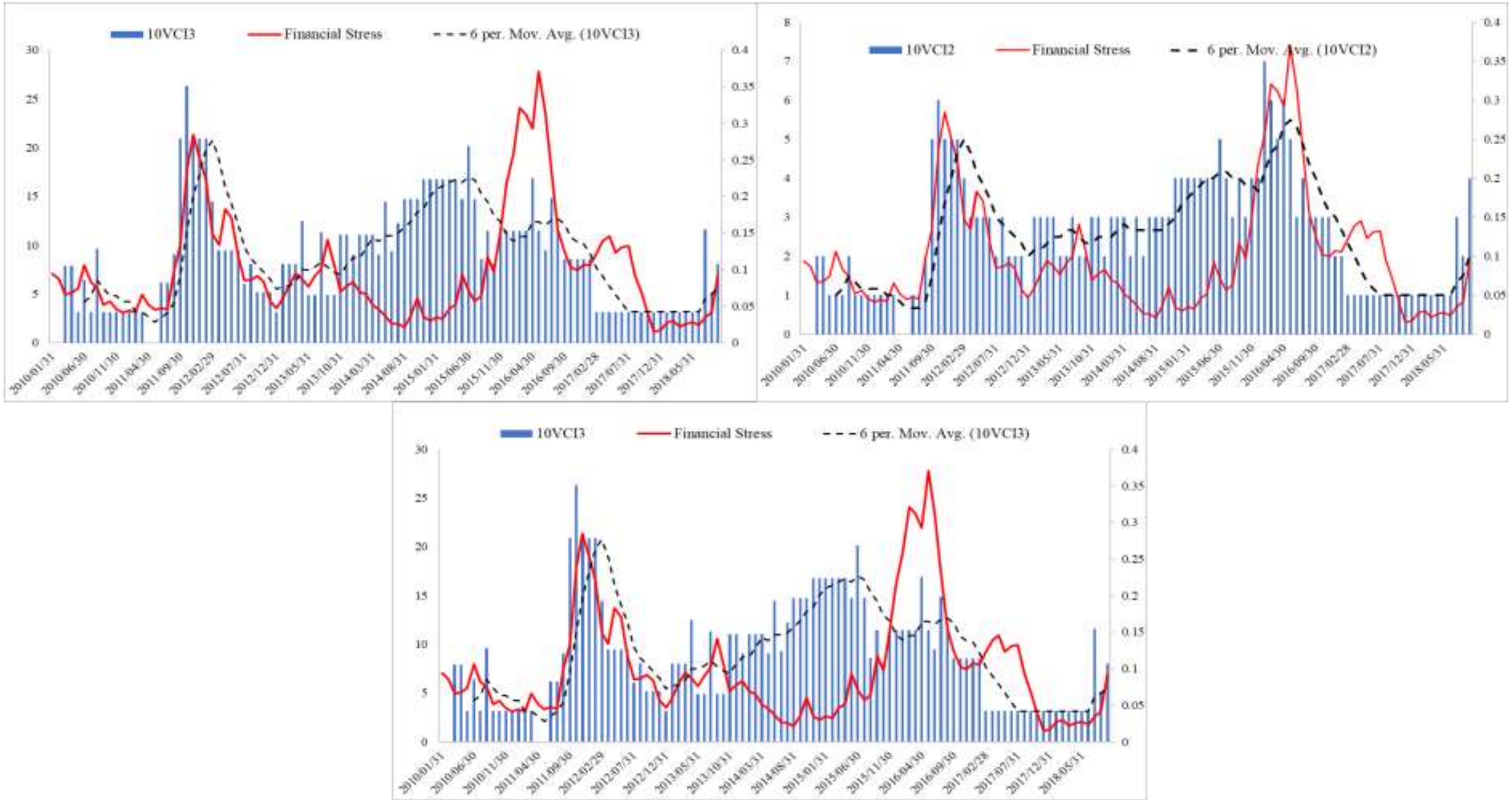


Figure 11: 10VCI values and MAs compared to financial stress (Jan. 2010–Sept. 2018)



Lastly, as with any policy framework, it is essential to have both transparency and accountability. From a macroprudential perspective, the evidence presented in this article emphasised why including EWS models in the analytical toolkit is beneficial, since they facilitate both. At a minimum, diagrammatical representations throughout this article served as transparent communication tools to support prudential decision-making. For instance, based on the out-of-sample evidence, South African policymakers would probably have made limited or no major prudential intervention decisions for the 2010–2018 period, given no concrete evidence to support the contrary. Results in this article also facilitate better *ex post* performance evaluation, since policymakers can be held accountable for implementing policy based on the *ex ante* evidence presented here. Again, in this out-of-sample case, the decision for inaction would have been justified.

7. Conclusion

The need for a better macroprudential policy analytical toolkit is clear and an important challenge is to combine the measurement of both causes and consequences of systemic risk. The current study contributes to enhancing the analytical toolset for macroprudential policy by assessing the ability of various indicators to detect systemic risk for South Africa, as measured by a financial stress index. Justification for this stems from the view that periods of financial vulnerability tend to precede financial stress periods which, in turn, lead to real economic consequences. The signal extraction approach was chosen as early warning method and although well established in banking and currency crisis literature, its application to detect financial stress in a single country is less researched. This study made the case that indicator selection should undergo a rigorous evaluation process to ensure that good and consistent indictors are identified prior to further analysis at optimal thresholds. A variety of evaluation criteria was used to assess the in-sample signalling ability of univariate indicators robustly. However, since these criteria can yield different results and because the relative importance of these criteria is unclear, a multi-criteria aggregate scoring system was also introduced. This system yielded excellent results and clearly identified good indicators based on overall signalling performance. The in-sample results indicated that few variables had exceptional overall signalling ability and only 17 were evaluated further at their optimal thresholds.

Various evaluation criteria were also used to identify and compare potential optimal thresholds, and since the proposed aggregate signalling score measure (ASSM) was robust, it was used to identify the final threshold. The in-sample optimal threshold results suggested that univariate indicators have acceptable signalling ability, but that reliance on them for policymaking purposes is not prudent. Accordingly, these individual indicators were combined into three composite indicators that managed to yield far superior in-sample signalling results. The aggregate scoring system again excelled at evaluating overall performance of composite indicators and identified similar optimal thresholds as several other contingency matrix-based measures. In-sample, all composite indicators showed good signalling ability and the conditional probability of correctly forecasting stress increased significantly with higher composite indicator scores. Similarly, the composite indicator MAs often increased prior to high-stress periods. The results suggest that the signals approach worked well for South Africa during the in-sample period and that composite indicators were useful for macroprudential purposes. The out-of-sample results were less impressive, but not surprising in the absence of high-stress periods during this period. Both univariate and multivariate indicators performed worse, composites did not clearly outperform individual indicators and signals were issued only at lower thresholds close to historical optimal levels. Given the absence of high-stress periods, these out-of-sample results were to be expected, and the fact that few signals were issued is a good result. Furthermore, the out-of-sample results are encouraging in that no false positive signals (i.e. Type II errors) were issued at higher thresholds, while only a few occurred at lower thresholds. This provides support for the notion that composite indicators can be interpreted in a versatile way, and policymakers are encouraged to rather assess signals over a threshold range. However, the relative tranquil out-of-sample period does mean that the efficacy of the model, especially in terms of Type I errors, must still be tested during more turbulent future periods. From a transparency and accountability perceptive, the fact that out-of-sample signals only occurred at low thresholds would justify inaction by policymakers. Overall, results from this study suggest that the signal extraction approach adds value and it should therefore be included in the SARB's macroprudential analytical framework.

8. References

Abiad, A. 2003. *Early warning systems: A survey and a regime-switching approach*. Working paper no. 03/32. Washington, DC: International Monetary Fund.

Aldasoro, I., Borio, C. & Drehmann, M. 2018. Early warning indicators of banking crises: Expanding the family. *BIS Quarterly Review*, March: 29–45.

Alessi, L., Antunes, A., Babecký, J., Baltussen, S., Behn, M., Bonfim, D., Bush, O., Detken, C., Frost, J., Guimarães, R., Havránek, T., Joy, M., Kauko, K., Matějů, J., Monteiro, N., Neudorfer, B., Peltonen, T., Rodrigues, P., Rusnák, M., Schudel, W., Sigmund, M., Stremmel, H., Šmídková, K., Van Tilburg, R., Vašíček, B. & Žigraiová, D. 2015. *Comparing different early warning systems: Results from a horse race competition among members of the Acro-prudential Research Network*. Paper no. 62194. Munich: MPRA, University Library of Munich.

Alessi, L. & Detken, C. 2009. *"Real time" early warning indicators for costly asset price boom/bust cycles: A role for global liquidity*. Working paper no. 1039. Frankfurt: European Central Bank.

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.

Alessi, L. & Detken, C. 2014. On policymakers' loss functions and the evaluation of early warning systems: Comment. *Economics Letters*, 124(3):338–340.

Alessi, L. & Detken, C. 2018. Identifying excessive credit growth and leverage. *Journal of Financial Stability*, 35:215–225.

Aziz, J., Caramazza, F. & Salgado, R. 2000. *Currency crises: In search of common elements*. Working paper no. 00/67. Washington, DC: International Monetary Fund.

Babecky, J., Havránek, T., Matějů, J., Rusnák, M., Šmídková, K. & Vašíček, B. 2012. *Banking, debt, and currency crises: Early warning indicators for developed countries*. Working paper no. 1485. Frankfurt: European Central Bank.

Babecky, J., Havránek, T., Matějů, J., Rusnák, M., Šmídková, K. & Vašíček, B. 2013. Leading indicators of crisis incidence: Evidence from developed countries. *Journal of International Money and Finance*, 35(C):1–19.

Baker, S.G. & Kramer, B.S. 2007. Peirce, Youden, and receiver operating characteristic curves. *American Statistician*, 61(4):343–346.

Baron, M. & Xiong, W. 2017. Credit expansion and neglected crash risk. *The Quarterly Journal of Economics*, 132(2):713–764.

Barrell, R., Davis, E.P., Karim, D. & Liadze, I. 2010. Bank regulation, property prices and early warning systems for banking crises in OECD countries. *Journal of Banking & Finance*, 34(9):2255–2264.

Beckmann, D., Menkhoff, L. & Sawischlewski, K. 2005. *Robust lessons about practical early warning systems.* Discussion paper no. 322. Hannover: University of Hannover.

Behn, M., Detken, C., Peltonen, T. & Schudel, W. 2013. *Setting countercyclical capital buffers based on early warning models: Would it work?* Working paper series no. 1604. Frankfurt: European Central Bank.

Berg, A. & Pattillo, C. 1999a. Are currency crises predictable? A test. *IMF Staff Papers*, 46(2):107–138.

Berg, A. & Pattillo, C. 1999b. Predicting currency crises: The indicators approach and an alternative. *Journal of International Money and Finance*, 18(4):561–586.

Berge, T.J. & Jordà, Ò. 2011. Evaluating the classification of economic activity into recessions and expansions. *American Economic Journal: Macroeconomics,* 3(2):246–277.

Bhattacharya, S., Goodhart, C., Tsomocos, D. & Vardoulakis, A. 2015. A reconsideration of Minsky's financial instability hypothesis. *Journal Money Credit and Banking*, 47(5):931–973.

BIS (Bank for International Settlements). 2014. *The international transmission of monetary policy – lessons learnt in South Africa.* Paper no. 78. Basel.

BIS (Bank for International Settlements). 2018. *Annual economic report*. Basel.

Borio, C. & Drehmann, M. 2009a. Assessing the risk of banking crises – revisited. *BIS Quarterly Review*, March: 29–46.

Borio, C. & Drehman, M. 2009b. *Towards an operational framework for financial stability: 'Fuzzy measurement and its consequences.* Working paper no. 127. Basel: Bank for International Settlements.

Brave, S. & Butters, R.A. 2012. *Detecting early signs of financial instability*. Essays on Issues, 305. Chicago, IL: Federal Reserve Bank of Chicago.

Brüggemann, A. & Linne, T. 2002. *Are the Central and Eastern European transition countries still vulnerable to a financial crisis? Results from the signals approach.*

Discussion paper no. 5. Helsinki: Institute for Economies in Transition, Bank of Finland.

Bussière, M. & Fratzscher, M. 2002. *Towards a new early warning system*. Working paper no. 145. Frankfurt: European Central Bank.

Bussière, M. & Mulder, C. 1999. *External vulnerability in emerging market economies: How high liquidity can offset weak fundamentals and the effects of contagion*. Working paper no. 99/88. Washington, DC: International Monetary Fund.

Cambón, M. & Estévez, L. 2015. A *Spanish Financial Market Stress Index (FMSI).* Working paper no. 60. Madrid: Comisión Nacional del Mercado de Valores.

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.

Caprio, G. & Klingebiel, D. 1996. *Bank insolvencies: Cross-country experience*. Policy research working paper no. 1620. Washington, DC: The World Bank.

Cardarelli, R., Elekdad, S. & Lall, S. 2011. Financial stress and economic contractions. *Journal of Financial Stability*, 7(2):78–97.

Carramazza, F., Ricci, L. & Salgado, R. 2000. *Trade and financial contagion in currency crises*. Working paper no. 00/55. Washington, DC: International Monetary Fund.

Christensen, I. & Li, F. 2014. Predicting financial stress events: A signal extraction approach. *Journal of Financial Stability*, 14:54–65.

Cihák, M. & Schaeck, K. 2010. How well do aggregate prudential ratios identify banking system problems? *Journal of Financial Stability,* 6:130–144.

Collins, S. 2003. *Probabilities, probits and the timing of currency crises*. Washington, DC: Georgetown University, The Brookings Institution and NBER.

Comelli, F. 2013. *Comparing parametric and non-parametric early warning systems for currency crises in emerging market economies*. Working paper no. 13/134. Washington, DC: International Monetary Fund.

Corsetti, G., Pesenti, P. & Roubini, N. 1999. Paper tigers? A model of the Asian crisis. *European Economic Review*, 43(7):1211–1236.

Danielsson, J., Valenzuela, M. & Zer, I. 2018. Learning from history: Volatility and financial crises. *The Review of Financial Studies*, 31(7):2774–2805.

Davis, E.P. & Karim, D. 2008a. Comparing early warning systems for banking crises. *Journal of Financial Stability,* 4:89–120.

Davis, E.P. & Karim, D. 2008b. Could early warning systems have helped to predict the sub-prime crisis? *National Institute Economic Review*, 206(1):35–47.

Dawood, M., Horsewood, N. & Strobel, F. 2017. Predicting sovereign debt crises: An early warning system approach. *Journal of Financial Stability*, 28:16–28.

De Jager, S. 2012. *Modelling South Africa's equilibrium real effective exchange rate: A VECM approach.* Working paper no. 12/02. Pretoria: South African Reserve Bank.

Demirgüç-Kunt, A. & Detragiache, E. 1998. The determinants of banking crises in developed and developing countries. *IMF Staff Papers*, 45(1):81–109.

Detken, C., Weeken, O., Alessi, L., Bonfim, D., Bouchina, M., Castro, C., Frontczak, S., Giordana, G., Giese, J., Jahn, N., Kakes, J., Klaus, B., Lang, J.H., Puzanova, N. & Welz, P. 2014. *Operationalising the countercyclical capital buffer: Indicator selection, threshold identification and calibration options*. Occasional paper series, no. 5. Frankfurt: European Systemic Risk Board.

Dornbusch, R., Goldfajn, I. & Valdés, R.O. 1995. Currency crises and collapses. *Brookings Papers on Economic Activity*, 26(2):219–294.

Drehmann, M., Borio, C., Gambacorta, L., Jiminez, G. & Trucharte, C. 2010. *Countercyclical capital buffers: Exploring options*. Working paper no. 317. Basel: Bank for International Settlements.

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

Drehmann, M. & Juselius, M. 2014. Evaluating early warning indicators of banking crises: Satisfying policy requirements. *International Journal of Forecasting*, 30(3):759–780.

Edison, H.J. 2003. Do indicators of financial crises work? An evaluation of an early warning system. *International Journal of Finance and Economics*, 8(1):11–53.

Eichengreen, B. & Arteta, C. 2000. *Banking crises in emerging markets: Presumptions and evidence*. Working paper no. C00-115. Berkeley: Centre for International and Development Economics Research.

Eichengreen, B. & Rose, A. 1998. *Staying afloat when the wind shifts: External factors and emerging-market banking crises.* Working paper no. 6370. Cambridge, MA: National Bureau of Economic Research.

Eichengreen, B., Rose, A. & Wyplosz, C. 1995. Exchange market mayhem: The antecedents and aftermath of speculative attacks. *Economic Policy*, 21:249–312.

Eichengreen, B., Rose, A. & Wyplosz, C. 1996. *Contagious currency crises*. Working paper no. 5681. Cambridge, MA: National Bureau of Economic Research.

ESRB (European Systemic Risk Board). 2014. *Operationalising the countercyclical capital buffer: Indicator selection, threshold identification and calibration options*. Occasional paper no. 5. Frankfurt.

Ferrari, S. & Pirovano, M. 2015. *Early warning indicators for banking crises: A conditional moments approach.* Paper no. 62406. Munich: MPRA, University Library of Munich.

Frankel, J. & Rose, A. 1996. Currency crashes in emerging markets: An empirical treatment. *Journal of International Economics*, 41(3/4):351–366.

Frankel, J. & Saravelos, G. 2012. Are leading indicators of financial crises useful for assessing country vulnerability? Evidence from the 2008-09 global crisis. *Journal of International Economics*, 87(2):216–231.

Frankel, J. & Wei, S. 2005. Managing macroeconomic crises. In J. Aizenman & B. Pinto (eds.). *Managing economic volatility and crises: A practitioner's guide.* Cambridge: Cambridge University Press, 315–405.

Fratzscher, M. 1998. Why are currency crises contagious? A comparison of the Latin American crisis of 1994–1995 and the Asian crisis of 1997–1998. *Weltwirtschaftliches Archiv/Review of World Economics*, 134(4):664–691.

G20 (Group of Twenty). 2009. *London Summit – Leaders' statement, 2 April 2009*. Retrieved from https://www.imf.org/external/np/sec/pr/2009/pdf/g20_040209.pdf [Accessed 2 September 2018].

Gadanecz, B. & Jayaram, K. 2009. Measures of financial stability – a review. In *Proceedings of the IFC Conference on "Measuring financial innovation and its impact", IFC Bulletin*, 31:365–380.

Gala, P. 2008. Real exchange rate levels and economic development: Theoretical analysis and econometric evidence. *Cambridge Journal of Economics*, 32(2):273–288.

Gaytán, A. & Johnson, C.A. 2002. *A review of the literature on early warning systems for banking crises*. Working paper no. 183. Santiago: Central Bank of Chile.

Glick, R. & Moreno, R. 1999. *Money and credit, competitiveness, and currency crises in Asia and Latin America*. Working paper no. PB99-01. San Francisco, CA: Centre for

Pacific Basin Monetary & Economic Studies, Federal Reserve Bank of San Francisco.

Hawkins, J. & Klau, M. 2000. *Measuring potential vulnerabilities in emerging market economies*. Working paper no. 91. Basel: Bank for International Settlements.

Herrera, S. & Garcia, C. 1999. *User's guide to an early warning system for macroeconomic vulnerability in Latin American countries*. Policy research working paper no. 2233. Washington, DC: The World Bank.

Holopainen, M. & Sarlin, P. 2016. *Toward robust early-warning models: A horse race, ensembles and model uncertainty.* Working paper series no. 1900. Frankfurt: European Central Bank.

Hsieh, F. & Turnbull, B. 1996. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, 24(1):25–40.

Illing, M. & Liu, Y. 2003. *An index of financial stress for Canada.* Working paper no. 2003-14. Ottawa: Bank of Canada.

Juks, R. & Melander, O. 2012. *Countercyclical capital buffers as a macroprudential instrument*. Riksbank Occasional Studies. Stockholm.

Kaminsky, G. 1999. *Currency and banking crises: The early warnings of distress*. Working paper no. 99/178. Washington, DC: International Monetary Fund.

Kaminsky, G., Lizondo, S. & Reinhart, C. 1998. *Leading indicators of currency crises*. Staff papers no. 45(1). Washington, DC: International Monetary Fund.

Kaminsky, G. & Reinhart, C. 1996. *The twin crises: The causes of banking and balance-of-payment problems.* International finance discussion paper no. 544. Washington, DC: Board of Governors of the Federal Reserve System.

Kaminsky, G. & Reinhart, C. 1999. The twin crises: The causes of banking and balance-of-payment problems. *American Economic Review*, 89(3):473–500.

Krugman, P.R. 1996. Are currency crises self-fulfilling? *NBER Macroeconomics Annual*, 11:345–407.

Laeven, L. & Valencia, F. 2008. *Systemic banking crises: A new database*. Working paper no. 08/224. Washington, DC: International Monetary Fund.

Lo Duca, M., Koban, A., Basten, M., Bengtsson, E., Klaus, B., Kusmierczyk, P., Lang, J.H., Detken, C. (ed.) & Peltonen, T. (ed.). 2017. *A new database for financial crises in European countries*. Occasional paper series no. 194. Frankfurt: European Central Bank.

Lo Duca, M. & Peltonen, T.A. 2011. *Macro-financial vulnerabilities and future financial stress: Assessing systemic risks and predicting systemic events*. Working paper no. 1311. Frankfurt: European Central Bank.

Lund-Jensen, K. 2012. *Monitoring systemic risk based on dynamic thresholds*. Working paper no. 12/159. Washington, DC: International Monetary Fund.

Markowitz, H. 1952. Portfolio selection. *Journal of Finance*, 7(1):77-91.

Minsky, H.P. 1977. A theory of systemic fragility. In E.I. Altman & A.W. Sametz (eds.). *Financial crises.* New York, NY: Wiley, 138–152.

Oet, M.V., Ong, S.J. & Gramlich, D. 2013. *Policy in adaptive financial markets: The use of systemic risk early warning tools.* Working paper no. 1309. Cleveland, OH: Federal Reserve Bank of Cleveland.

Pasricha, G., Roberts, T., Christensen, I. & Howell, B. 2013. Assessing financial system vulnerabilities: An early warning approach. *Bank of Canada Review*, Autumn: 10–19.

Patel, S. & Sarkar, A. 1998. Crises in developed and emerging stock markets. *Financial Analysts Journal*, 54(6):50–61.

Patnaik, I., Felman, J. & Shah, A. 2017. An exchange market pressure measure for cross country analysis. *Journal of International Money and Finance*, 73(A):62-77.

Pepe, M., Longton, G. & Janes, H. 2009. Estimation and comparison of receiver operating characteristic curves. *Stata Journal*, 9(1):1–16.

Rose, A. & Spiegel, M. 2009. *The causes and consequences of the 2008 crisis: Early warning.* Working paper no. 2009-17. San Francisco, CA: Federal Reserve Bank of San Francisco.

Sachs, J., Tornell, A. & Velasco, A. 1996. Financial crises in emerging markets: The lessons from 1995. *Brookings Papers on Economic Activity*, 27(1):147–199.

SARB (South African Reserve Bank). 2018. *Mandate.* Retrieved from http://www.resbank.co.za/Financial%20Stability/Domestic/Pages/Mandate.aspx [Accessed 21 September 2018].

Sarlin, P. 2013. *On policymakers' loss functions and the evaluation of early warning systems*. Working paper no. 1509. Frankfurt: European Central Bank.

Sarlin, P. & Peltonen, T.A. 2011. *Mapping the state of financial stability*. Working paper no. 1382. Frankfurt: European Central Bank.

Savona, R. & Vezzoli, M. 2013. Fitting and forecasting sovereign defaults using multiple risk signals. *Oxford Bulletin of Economics and Statistics*, 77(1):66–92.

Streiner, D.L. & Cairney, J. 2007. What's under the ROC? An introduction to receiver operating characteristics curves. *The Canadian Journal of Psychiatry*, 52(2):121–128.

Tornell, A. 1999. *Common fundamentals in the Tequila and Asian crises*. Working paper no. 7139. Cambridge, MA: National Bureau of Economic Research.

Vašíček, B., Žigraiová, D., Hoeberichts, M., Vermeulen, R., Šmídková, K. & De Haan, J. 2015. *Leading indicators of financial stress: New evidence*. Working paper no. 476. Amsterdam: De Nederlandsche Bank.

Vidal-Abarca, A.O. & Ruiz, A.U. 2015. *Introducing a new early warning system indicator (EWSI) of banking crises.* Working paper no. 15/02. Madrid: Banco Bilbao Vizcaya Argentaria.

Yucel, E. 2011. *A review and bibliography of early warning models*. Paper no. 32893. Munich: MPRA, University Library of Munich.

Zhuang, J. & Dowling, M. 2002. *Causes of the 1997 Asian financial crisis: What can an early warning system model tell us?* Working paper no. 26. Manila: ERD, Asian Development Bank.

Appendix

Table A1: Potential EWS indicators based on average performance over entire threshold spectrum (January 1986-December 2009)

Variable name	Selection criteria (criterium number in brackets) Scores for all 5 criteria are calculated as the average over all thresholds					Selection criteria ranking (by number)					Overall ranking
	AUROC (1)	NTSR (2)	Relatively useful[30] (3)	A/(A+B) (4)	ASSM[31] (5)	(1)	(2)	(3)	(4)	(5)	
24-month ALSI avg. real return	88%	30%	36%	86%	416%	1	1	1	1	1	1
36-month ALSI avg. real return	85%	37%	28%	83%	330%	2	2	3	2	2	2
Federal funds rate (Y-o-Y % change)	80%	49%	30%	77%	322%	4	4	2	4	3	3
Business cycle indicator (2-year change)	74%	44%	23%	80%	293%	9	3	7	3	4	5
Lend–dep rate (Y-o-Y % change)	78%	50%	25%	77%	287%	5	5	4	4	6	5
REER (deviate from 36-month MA)	73%	53%	25%	76%	290%	11	7	4	7	5	7
Imports (Y-o-Y % change)	74%	51%	22%	77%	269%	9	6	8	4	8	7
Manufacturing production (Y-o-Y % change)	76%	57%	24%	74%	278%	8	9	6	10	7	8
REER (Y-o-Y % change)	73%	58%	22%	74%	257%	11	10	8	10	10	10
real M3 (Y-o-Y % change)	82%	64%	21%	72%	258%	3	17	10	15	9	11
Growth in nominal credit (dev. from MA)	77%	61%	21%	73%	246%	6	13	10	13	11	11
16 CDF PCA FSI (Y-o-Y % change)	73%	58%	19%	74%	233%	11	10	13	10	13	11
Growth in real total credit extended to private sector	77%	65%	20%	72%	238%	6	18	12	15	12	13
FSI_16CDF_Ew_> 24-month avg.	71%	60%	16%	75%	216%	16	12	17	9	14	14
Exports (Y-o-Y % change)	69%	61%	18%	73%	215%	18	13	14	13	15	15
6-month avg. % change in R$	69%	63%	18%	72%	212%	18	15	14	15	16	16
ALSI (Y-o-Y % change)	65%	54%	14%	76%	212%	26	8	22	7	16	16

[30] Calculated from equation 3 as usefulness (U) divided by $(\min[\mu;\ 1 - \mu])$ and can be interpreted as signalling performance above a random variable.

[31] Aggregate signalling score measure refers to the signalling ability score introduced in row 11 of Table 2.

Growth in nominal total credit to private sector	69%	63%	16%	72%	197%	18	15	17	15	19	17
R$ dev from MA	72%	68%	17%	71%	198%	14	20	16	19	18	18
Vix (Y-o-Y change)	66%	70%	16%	70%	185%	24	22	17	23	21	21
3-mth avg. % change of R$	65%	66%	13%	71%	181%	26	19	26	19	22	22
M3/GDP (Y-o-Y % change)	72%	72%	13%	71%	178%	14	25	26	19	24	22
Growth in real credit (dev from MA)	71%	71%	14%	70%	175%	16	24	22	23	25	22
Gross fixed capital formation (Y-o-Y % growth)	68%	73%	16%	69%	194%	22	28	17	27	20	23
% change in real private sector credit/GDP	69%	72%	14%	69%	174%	18	25	22	27	26	24
M2/reserves	65%	73%	15%	69%	179%	26	28	21	27	23	25
% change in R$	64%	70%	13%	70%	166%	29	22	26	23	28	26
% change R$ (36-month MA)	61%	68%	11%	71%	163%	35	20	32	19	29	27
12-month growth in private sector debt/GDP	66%	76%	14%	68%	168%	24	31	22	32	27	27
% change R$ (dev. from MA)	63%	72%	11%	69%	153%	32	25	32	27	31	29
ABSA House price index growth (dev. from MA)	62%	74%	12%	70%	140%	33	30	31	23	32	30
M2/reserves (dev. from MA)	64%	78%	13%	67%	154%	29	33	26	33	30	30
3-month growth in private sector debt/GDP	64%	86%	13%	66%	140%	29	37	26	35	32	32
growth in real M3 (2-years)	67%	85%	9%	66%	128%	23	36	34	35	34	32
ALSI deviation from HP	62%	80%	9%	67%	122%	33	34	34	33	35	34
House price/disp. income (2-year % change)	56%	77%	5%	69%	102%	39	32	38	27	36	34
Real lend rate below trend for 12 months.	59%	82%	3%	65%	100%	36	35	40	37	37	37
Inflation (dev. from MA)	55%	87%	6%	65%	89%	40	38	37	37	38	38
ABSA House price index growth	58%	99%	8%	60%	75%	37	40	36	41	40	39
CA/GDP (quarterly growth)	57%	94%	1%	63%	79%	38	39	41	39	39	39
Growth in domestic credit to private sector/GDP	55%	100%	4%	62%	48%	40	41	39	40	41	40
Average	**69%**	**68%**	**16%**	**71%**	**197%**						

Figure A1: In-sample ROC curves for two sets of composite indicators

(a) 10-Composite Indicators 1, 2 and 3 **(b) 17-Composite Indicators 1, 2 and 3**

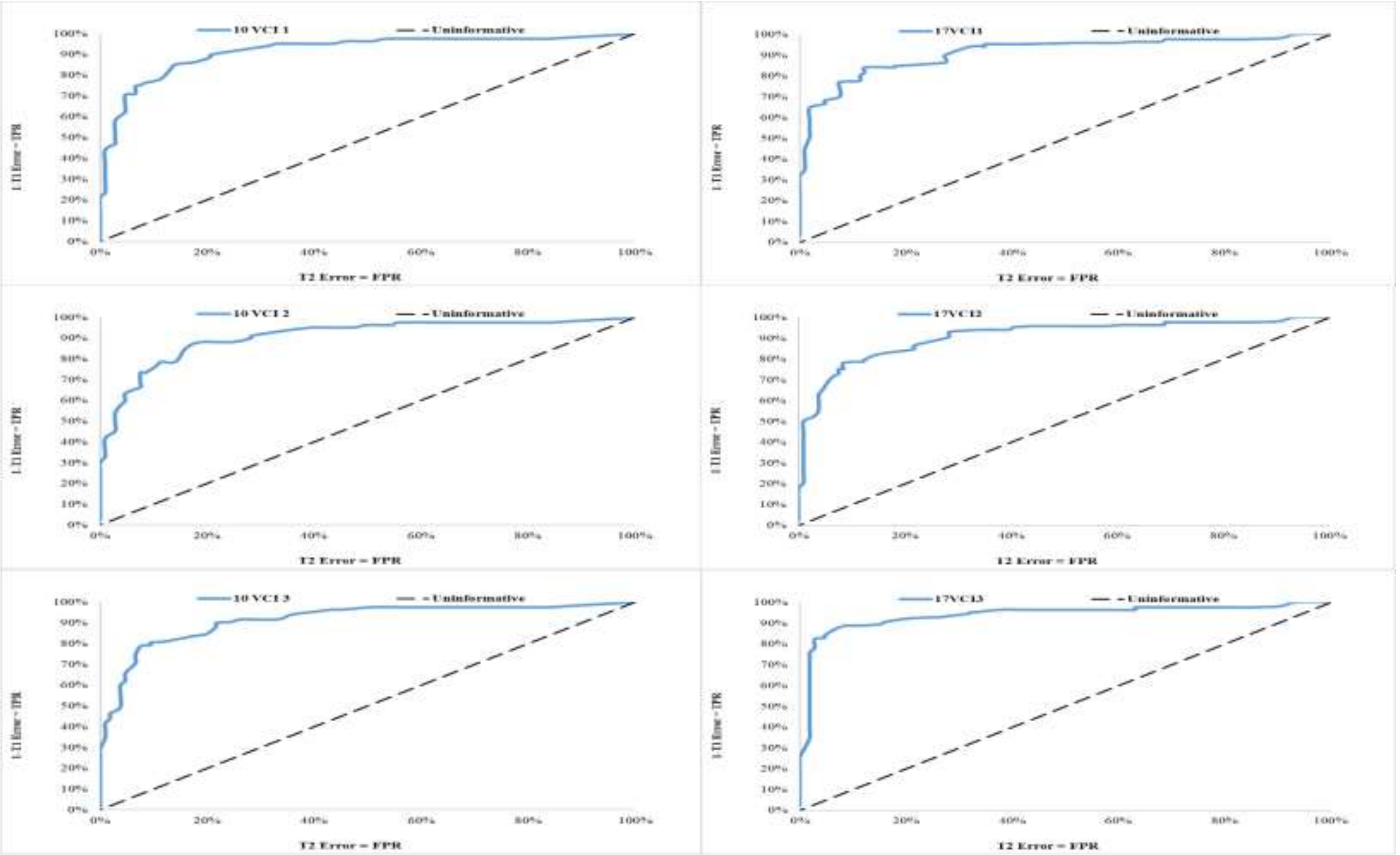
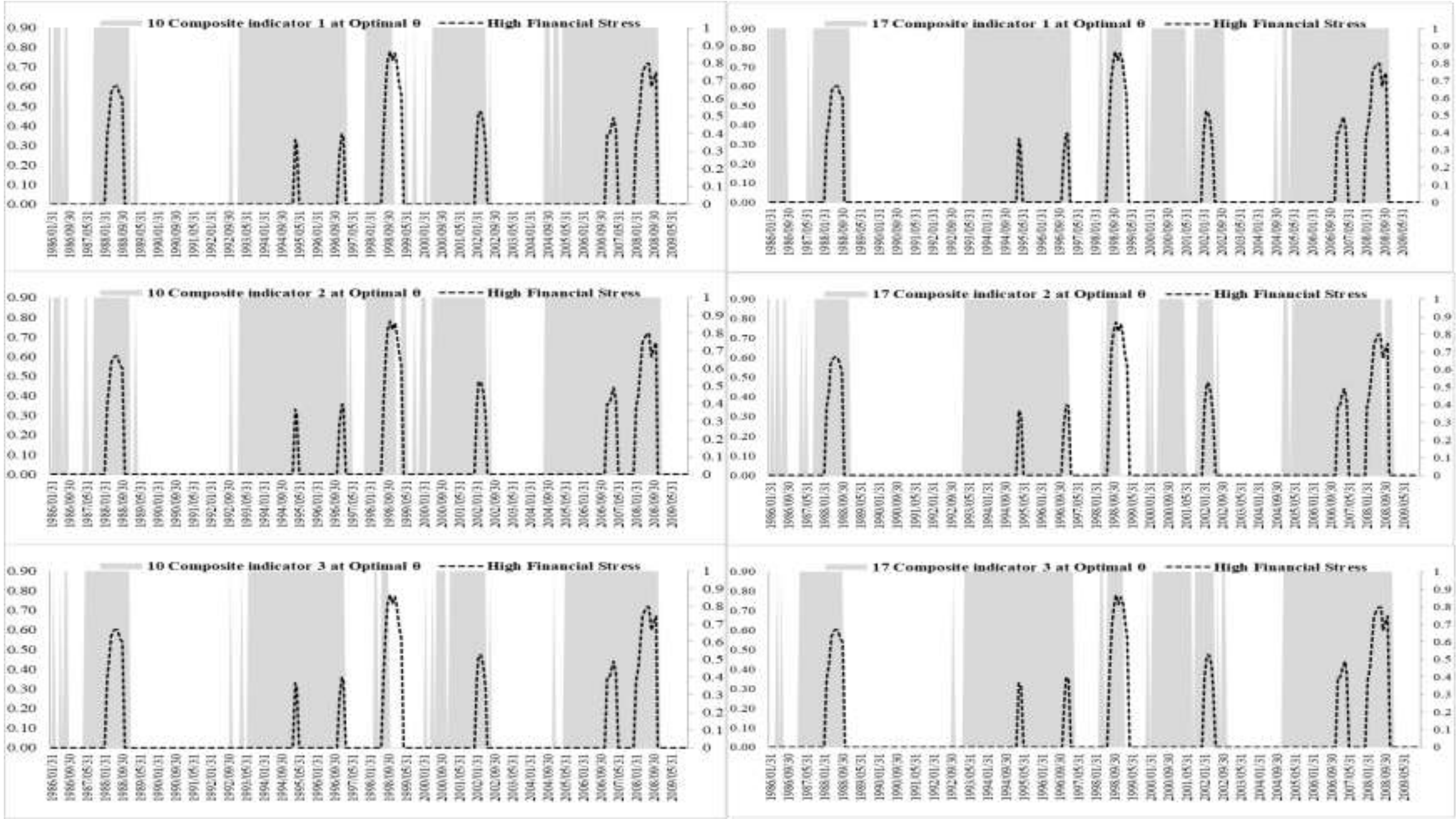


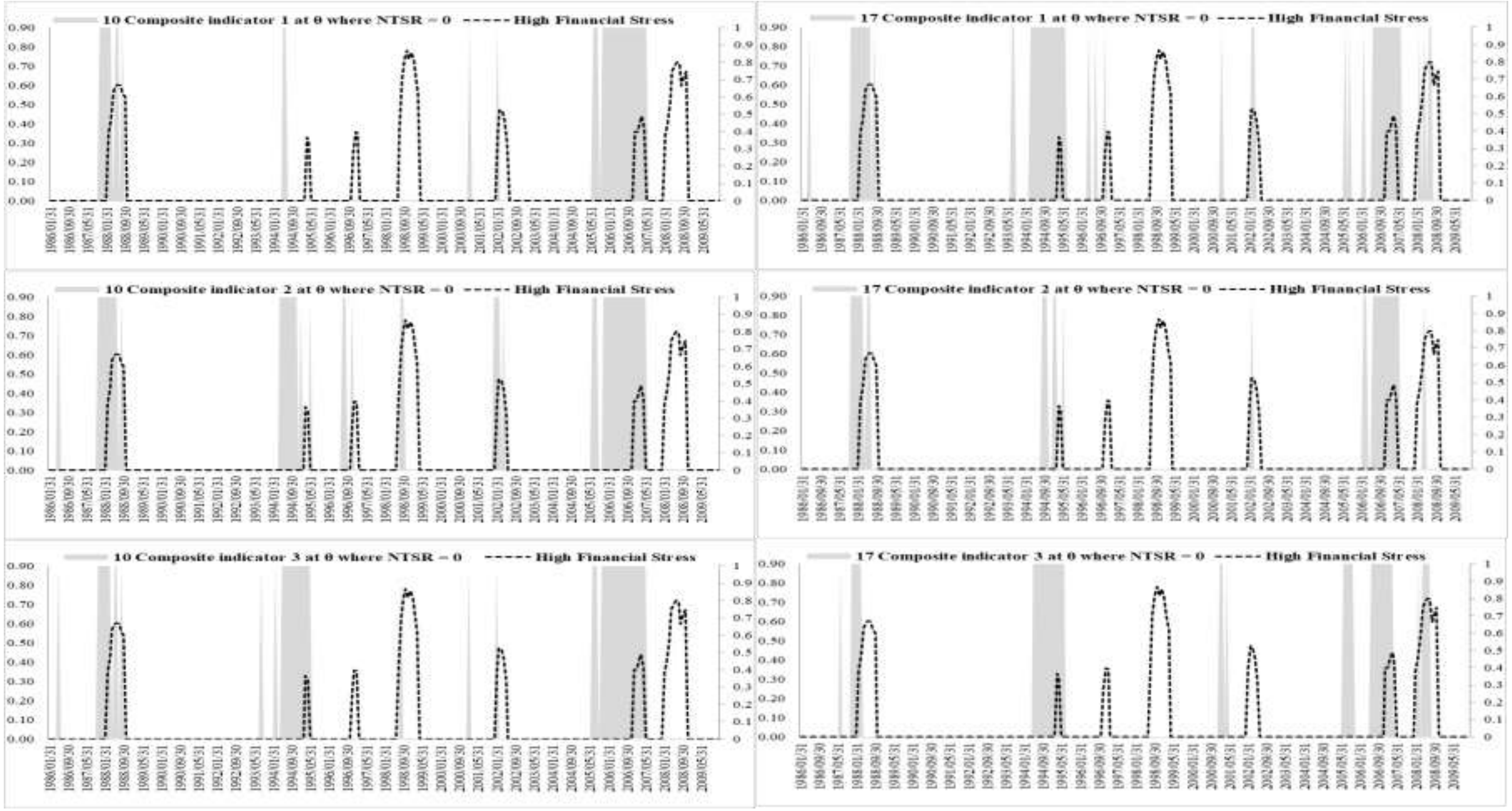
Figure A2: Signals issued by composite indicators at optimal threshold levels



(a) 10-Variable Composite Indicators 1, 2 and 3

(b) 17-Variable Composite Indicators 1, 2 and 3

Figure A3: Composite indicators in-sample signals at threshold where NTSR = 0%



(a) 10-Variable Composite Indicators 1, 2 and 3

(b) 17-Variable Composite Indicators 1, 2 and 3