Forecast Evaluation in Macroeconomics and International Finance. Ph.D. thesis, George Washington University, Washington, DC, USA.

Bespalova, Olga

George Washington University

28 March 2018

Essays on Forecast Evaluation in Macroeconomics and International Finance

by Olga G. Bespalova

B.S. in Economics, Accounting, and Audit, June 2002, Astrakhan State Technical University
B.S. in Interpretation and Translation (English), May 2006, Astrakhan State Technical University
M.A. in Economics, May 2011, Kansas State University
M.Phil. in Economics, May 2015, The George Washington University

A Dissertation submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

May 20, 2018

Dissertation directed by

Robert F. Phillips
Professor of Economics

The Columbian College of Arts and Sciences of The George Washington University certifies that Olga Bespalova has passed the Final Examination for the degree of Doctor of Philosophy as of March 7, 2018. This is the final and approved form of the dissertation.

Essays on Forecast Evaluation in Macroeconomics and International Finance

Olga Bespalova

Dissertation Research Committee:

Robert Phillips, Professor of Economics, Director

Michael Bradley, Professor of Economics, Committee Member

Fred Joutz, Professor of Economics, Committee Member

Dedication

I dedicate this dissertation to my family, especially to my Mother Liubov, whose unconditional love and support helped me through this long and difficult journey far away from home. I am praying that God gives her many more years to come and that she will enjoy her retirement in good health and happy spirits, being able to travel and pursue her hobbies, and see her dreams to come true.

I also dedicate this dissertation in memory of my grandparents Mikhail and Valentina, and Aleksandr and Vera, who, unfortunately, passed away before they could see me to defend it. Three of them passed away last year, which made completion of this research twice harder for me. My grandpa Aleksandr passed away during my undergraduate studies - I will remember him as a cheerful and kind man, often playing his harmonica. My grandfather Mikhail encouraged me to learn English and obtain a doctorate degree. My grandmothers Valentina and Vera were very kind and spiritual – I felt their love even being across an ocean; just a thought about them lifts my spirits in any tough moment. I believe that my grandparents are now in heaven and guarding and helping me as my angels.

Additionally, I dedicate this dissertation to my God Parents Mikhail and Irina who often offered help and support when Mother and I needed it the most. I am also grateful to the rest of my family who contributed to my happy childhood memories and always supported and encouraged me in all my academic, professional, and personal endeavors. In any moment of doubt, I recall my family saying "Who if not you?" and continue to pursue my dreams. I would not be who I am now without such a wonderful family and God's love. Thank you very much!

Acknowledgments

First and foremost, I wish to express my sincerest gratitude to Herman Stekler, Emeritus Professor of Economics, who inspired me to undertake this research project and led me by hand through every stage of this research. He is my role model of a research economist – passionate and fully devoted to improving macroeconomic forecasting techniques, successful author and editor of the dozens of academic publications, and a joyful person with a great sense of humor, tact, and discipline. I appreciate the countless hours he spent discussing my research ideas, reading and commenting on my writing, which tremendously helped me to improve this thesis.

I acknowledge regular support I received from my Dissertation Director – Professor Robert Phillips who carefully guided novel applications of the econometrics methods developed in this work. His patience and encouragement helped me to overcome challenges in this research. I am thankful to Professors Michael Bradley and Fred Joutz for spending their precious time to serve on my committee and giving many insightful suggestions for my research.

This dissertation would not be completed without financial support, research and travel grants, teaching opportunities and excellent scholarship I received from the Columbian College of Arts and Sciences and the Elliott School of International Affairs at the George Washington University. I want to specially thank John Kendrick's family for their endowment and support of international students! I enjoyed the time I spent at the Department of Economics and thank its faculty and graduate students for their friendships and encouragement. I am also grateful for the comments I received from participants at the International Symposium of Forecasters, the Federal Forecasters Conference, the GWU Research Program on Forecasting, and SAGE seminars at Economics Department where I presented this research.

Finally, I owe success of this long journey to the excellent academic preparation I received in Russia, to the Fulbright program which brought me to the USA, and to many sincere friendships I developed over this time.

Abstract of Dissertation

Essays on Forecast Evaluation in Macroeconomics and International Finance

This dissertation shares three common themes: (i) forecasting rare macroeconomic events, i.e. GDP declines and currency crises; (ii) the use of non-parametric methods to evaluate binary indicators, in particular, the advantages of the analysis of the Receiver Operating Characteristic (ROC) curves; and (iii) value of qualitative information from expert surveys and textual analysis in macroeconomic forecasting.

Chapter 1 contributes to the literature on evaluation of the qualitative survey directional forecasts using the World Economic Survey (WES) for the U.S. economy in 1989q1-2015q4. I offer a methodology which combines the ROC curves analysis with the traditional analysis of the contingency tables. I propose criteria to assess in-sample and out-of-sample directional predictive value of the binary indicators, including directional forecasts from the qualitative surveys. I find that the WES has high out-of-sample value in forecasting movements in GDP and consumption, and moderate for imports, trade balance, inflation, and short-term interest rate. It has no value in predicting changes in investment and exports. I also motivate and confirm that the WES Economic Climate (EC) indicator is as a more accurate predictor of future movements in the real GDP than future expectations alone. Additionally, I show that the ROC-optimal thresholds yield more accurate predictions than their alternatives proposed by Hutson et al. (2014).

Chapter 2 re-examines indicators of currency crises suggested by Kaminsky and Reinhart (1999) and subsequent studies using the ROC curves analysis. I utilize a training set (1975-1995) to confirm a list of indicators with the in-sample predictive value, and test their out-of-sample using data for 1996-2002. Four variables have both in-sample and out-of-sample predictive value: the deviation of the real exchange rate (RER) from a trend, the foreign reserves, the ratio of broad money M2 to reserves, and the decline in exports. I show that the ROC-optimal thresholds issue more accurate signals than the minimum noise-to-signal ratio previously used in the literature. I

also employ modified ROC curves to display the relationship between the precision of sent signals and recall of crisis episodes. Finally, I propose forecast combinations using several ad-hoc rules which help to improve forecast accuracy.

Chapter 3 contributes to the discussion of asymmetric information about the U.S. economy between the Federal Reserve System (FRS) and the Survey of Professional Forecasters (SPF) via textual analysis of the Federal Open Market Committee (FOMC) minutes. It builds on Stekler and Symington (2016), who scored the texts of the FOMC minutes in 2006-2010 to produce the indexes for the current and future outlooks and their calibrations to the U.S. real GDP. I extend their time-series adding 26 years of observations to cover 1986Q1-2016Q4. Following Ericsson (2016), I interpret the derived calibrations (FMIs) as elicitcasts of the Greenbook (GB) forecasts. Results indicate that the FMIs are unbiased, efficient, rational, and contain the same informational advantage as the GB forecasts. The forecast encompassing tests suggest that both the FMIs and the SPF forecasts contain their own unique knowledge and can learn from each other. I find that the SPF forecasters already pay close attention to the FOMC minutes available to them at the time of forecast deadline and efficiently use its information in their set. Yet, they could improve their forecasts should the FOMC minutes from the first quarterly meetings become available without a three-week publication lag. During their second quarterly meetings, the FOMC policy-makers accounted only for their own earlier assessment of the U.S. macroeconomy – they did not put weight on the SPF forecasts released a few weeks earlier in the same quarter. The results are robust to the use of alternative scale.

Overall, I find that directional forecasts are informative. The qualitative WES survey can produce accurate directional macroeconomic forecasts. The ROC curves analysis helps to set an association between the consensus scores and the growth rates as well as to find accurate indicators of currency crisis. The qualitative statements from monetary policy deliberations can be converted in to the GDP growth forecasts with unique information about the US economy.

Table of Contents

List of Figures

List of Tables

Chapter 1: An evaluation of the directional accuracy of the World Economic Survey U.S. macro forecasts

1.1. Introduction

Forecasting macroeconomic conditions, especially changes in the real GDP, has always been challenging, even in the short horizons. To increase accuracy of their predictions, economists often use data from the surveys of business tendencies, consumer confidence, and economic conditions. These surveys question either economic agents or professional forecasters, and thus may contain valuable information not captured by fundamentals, especially during periods of macrostructural shifts or around turning points of business cycles. For this reason, the survey data are often used as leading indicators or as explanatory variables in econometric models[1]. For example, the consumer and business tendency surveys were useful in forecasting turning points, industrial production, consumer spending, GDP growth and other macroeconomic indicators (Öller, 1990; Ludvingson, 2004; Claveria, 2007; Lahiri and Monokroussos, 2014). Croushore (2004) warns that the consumer sentiment surveys do not have real time predictive value.

An evaluation of a survey's predictive value requires rigorous procedures which depend on its type – quantitative or qualitative. Respondents of the quantitative surveys assign numerical values to the items of interest, i.e. the SPF experts forecast the rates of growth for the real GDP and other macroeconomic variables. The accuracy of such surveys is easy to assess since both the forecasts and the actual realizations are numbers[2]. In qualitative surveys, their respondents pick one of several categorical responses expressed as *verbal statements*. For example, the *World Economic Survey (WES) experts* indicate whether they expect that the overall economy in the

[1] These models can include such predictors as a balance statistic, a diffusion index, an odds ratio, or other algebraic transformation using data about the fractions of answers in each category.

[2] Common summary measures include the bias of forecast, the variance of the forecast error, the mean square forecast error, and the distribution of the forecast error. See Carnot, Koen & Tissot (2005), Ericsson (2012) and Stekler (1991) for details.

future will be “better," “the same," or “worse", or whether they assess the present economy as “good,” “satisfactory,” or “bad.”

Pesaran (1984, 1987) favored the qualitative surveys over the quantitative ones arguing that the former are less likely to be biased due to the behavioral uncertainty or sampling and measurement errors[3]. Evaluation of qualitative surveys[4] presents two problems. First, the *interpretation of words* used as responses determines the actual data to be used in the comparison[5]. For example, “the same” can be understood as a zero change in the level of the forecasted variable (zero growth rate) or as a constant growth rate (no acceleration or deceleration of growth)[6]. Second, a comparison of the qualitative forecasts with actual quantities can be made only after both are presented in a similar format. There are two methods to achieve this task.

The first method – classifying actual data in the same number of categories as the number of qualitative survey responses – requires an analyst to determine the values of actual data which mark the border between the adjacent categories. In the case of a survey with three categorical responses (i.e., “higher,” “the same,” or “lower”), this brings about the problem of finding the so-called "*indifference interval*" around a zero that includes the no change category[7]. Thus, both forecasts and realizations are classified into three groups, yielding a *3x3 contingency table* that could be tested using conventional statistics under the null hypothesis that the forecasts and realizations are independent. When the primary research interest is the direction of change, the forecasts and actuals are collapsed into a 2x2 contingency table[8]. This simplifies the problem

[3] Forecasters contributing to quantitative surveys may engage in a strategic behavior, such as herding, reputational cheap talk, radical forecasting or forecast competition (i.e. Trueman, 1994; Lamont, 2002; Ottaviani and Sorensen, 2006).

[4] I focus on the surveys with three responses, although the analysis can be extended to other cases.

[5] Sometimes, interpretation is obvious. For example, such categories as "higher," "no change" and "lower" clearly imply a comparison with a zero change in the variable of interest.

[6] When the experts judge present economic conditions as "good," "satisfactory," or "bad”, the neutral response can be understood as a near-zero growth or as a growth below its long-run trend.

[7] Theil (1961) proposed a distribution-free method to find a range of such indifference interval and to categorize observations beyond this interval as “increase” and “decrease” respectively.

[8] Most studies solved this problem adding “no change” to one of the other categories, for example, "growth" and "no-growth," or "decline" and "no decline."

as now one needs to defining an indifference interval to a determination of a single cut-off value.

The second method of evaluating qualitative surveys – their quantification and comparison to the actual realizations – requires an analyst to have data about the distribution of the survey responses between categories. There are two approaches to quantifying qualitative survey responses[9].

Anderson (1952) pioneered the first quantification method – the *regression approach.* It used a regression of the actual data on the fractions of "higher" and "lower" categories and interpreted the resulted coefficients as the upper and lower borders of the “no change” indifference interval. He also introduced the *balance statistic (B)* as a difference between the fractions of the optimistic and pessimistic responses[10]. Pesaran (1984) improved the Anderson’s regression approach to account for the asymmetry in the indifference interval letting the upper border of such interval depend on the actual level of the underlying time-series. Smith and McAleer (1995) allowed both borders of the indifference interval to vary with actual data.

Carlson and Parkin (1975) proposed second quantification method – the *probability approach.* They assumed that expectations are normally distributed and that the thresholds are constant and symmetric. Later, Lahiri and Zhao (2015) relaxed assumptions about the normality of expectations and allowed for asymmetric and time-varying indifference intervals using a hierarchical ordered probit model.

However, both methods (categorizing the actual data in 3x3 contingency table and quantification approaches) use information about the proportions of responses in each category, which is not available in some surveys. For example, the WES does not provide data on responses in each category: instead, its results are summarized as *consensus scores* ranging from 1 to 9 by construction (see section 1.2.1). The absence of an appropriate method to evaluate and interpret such scores explains their lack of use in macroeconomics and forecasting.

[9] Nardo (2003) provided a comprehensive literature review for both quantification methods.

[10] It is broadly used by surveys (i.e., by the OECD) and shown to be robust to non-responses (Seiler, 2015).

Hutson, Joutz, and Stekler (2014) were the first to assess and interpret the WES expectations for the U.S. economy. They proposed two ad-hoc approaches to produce directional forecasts[11] for the U.S. real GDP and some of its components. First, they applied a notion of the "indifference intervals" issuing forecasts "up" and "down" when the WES scores exceed the survey mean by 0.5 or 1 standard deviations. Second, they used a simple rule that the WES scores above 5.0 can be used to forecast positive economic growth.

This chapter develops *an alternative method to evaluate and interpret the qualitative survey in the absence of data about responses in each category.* Such method *combines the analysis of contingency tables* with the *ROC curves analysis* (formally defined in sec.2.4). It uses the WES data for the USA in 1989-2015 with a focus on the real GDP, consumption, and investment. It also evaluates the WES expectations for the exports, imports, trade balance, inflation rate and short-term interest rate.

The study poses several research questions: (i.) Are the WES expectations accurate in *classifying up and down movements* in the real GDP and its components, inflation rate and short-term interest rate? (ii.) What values of the WES consensus scores are optimal to produce up and down directional forecasts? (iii.) How to interpret the WES consensus scores in terms of the expected growth rates for the surveyed macroeconomic variables?

In this chapter I make several contributions to the literature. *First, I show the algebraic relation between the WES consensus scores and the balance statistics*, which implies that higher WES scores express more optimistic expectations about the present and future economy. Second, I *interpret words* in the WES questionnaire and explain implications of these interpretations for the survey users. Third, I clarify meaning of the WES answers and scores in terms of the implied growth rates using *the ROC curves analysis,* which suggests the specific values of the WES scores to produce binary forecasts (classifications) of periods when the economic variable grows

[11] Leitch and Tanner (1991) argued that for profit-maximizing firms correctly predicting the direction of change is more important than the size of the forecast error.

at least at 0, 1, 2, and 3% per annum. Then, I note that the AUC currently used in the literature on evaluation of binary forecasts and classifications (i.e. Berge and Jordà, 2011; Gorr and Schneider, 2013, and Lahiri and Wang, 2013) is not a sufficient accuracy criterion. I introduce more rigorous additional accuracy criteria to account for uncertainty in estimates of the AUC and the ROC curve itself and to distinguish the in-sample and out-of-sample predictive value. Finally, I provide motivation to use the *WES EC indicator* to produce binary directional forecasts for the overall economy.

The outline of the rest of this chapter is as follows. Section 1.2. discusses the methodology of evaluating qualitative survey forecasts. Section 1.3. describes the WES survey questions, aggregated consensus scores and the actual U.S. data used for their evaluation. Section 1.4. features empirical results and robustness checks. The chapter's conclusions are given in Section 1.5, which is followed by the references in 1.6.

1.2. Methodology of evaluating qualitative survey forecasts

In section 1.2.1 I explain how responses from any qualitative survey can be summarized as consensus scores and show their algebraic relation to the well-known Anderson's balance statistic. Section 1.2.2 outlines a conventional way to evaluate the directional accuracy of a binary forecast with contingency tables. Finally, in section 1.2.3, I suggest new application of an ROC curves analysis to the evaluation of the binary directional forecasts and discuss how it improves conventional approach.

1.2.1. Scoring the qualitative surveys: balance statistics and consensus scores

Suppose that a survey question assumes three categorical responses: 1, 2, and 3, which can be interpreted as "up/optimistic," "no change/indifferent," and "down/pessimistic respectively. Let n_1, n_2 , and n_3 be the number of answers in each category. There are several ways to aggregate this information and present it to the survey users. One way is to find the shares of responses in each category as $u = \frac{n_1}{n}$; $nc = \frac{n_2}{n}$; $d = \frac{n_3}{n}$, where n is a total number of responses

$(n = n_1 + n_2 + n_3)$ and calculate the Anderson' balance statistic (B) as the difference between the fractions of the two extreme responses:

$$B = u - d = \frac{n_1 - n_3}{n} \quad [1]$$

Another way to aggregate responses is to assign every category a value and to find a consensus score as an aggregated average. For example, let's assign values a, b, and c to the categories 1, 2, and 3 respectively. Economic interpretability requires three conditions:

(i) Scores a, b, and c are rational numbers[12];

(ii) A category with a more optimistic response is assigned a higher score:

$$a > b > c \quad [2]$$

(iii) The distances between the values assigned to categories 1, 2, and 3 are equal:

$$a - b = b - c = \Delta \quad [3]$$

Then the survey consensus score S can be found using [4]:

$$S = \frac{an_1 + bn_2 + cn_3}{n} \quad [4]$$

The derived consensus score will belong to an interval (c, a). However, some surveys (i.e. the WES) publish their results only as such aggregated consensus scores; the fractions of expert opinions in each category are not publicly available. In this case, it is helpful to find how the consensus score is related to the Anderson's balance statistic (B).

Proposition 1. When the conditions (i.)-(iii.) above hold, the survey consensus score S can be linearly transformed to obtain the balance statistic B using [5]:

$$B = \frac{S - b}{\Delta} \quad [5]$$

Proof. Eliminate the number of responses in the second category $n_2 = n - n_1 - n_3$ from the consensus score [4] and obtain: $S = \frac{an_1 + b(n - n_1 - n_3) + cn_3}{n}$. This simplifies to

[12] Using integers usually makes interpretation easier, but using fractions is not uncommon.

$$S = b + \frac{(a-b)n_1 - (b-c)n_3}{n} \qquad [6]$$

Insert [3] in [6] and simplify as $S = b + \Delta\frac{n_1 - n_3}{n}$. Use [1] to substitute B for $\frac{n_1 - n_3}{n}$ and the result in [5] follows. Thus, the aggregated average consensus score S is a linear transformation of the Anderson's balance statistic, which has a long history of applications in forecasting with survey data and therefore will possess the same statistical properties.

1.2.2. Forecasting rule and contingency tables for a given consensus score

To evaluate the directional accuracy of a consensus score as a leading indicator, an analyst should first measure the actual growth in the variable of interest z, and then construct a binary dummy variable which takes a value "1" if the success was observed, and "0" otherwise. The relative growth rates in the actual data are calculated as $g_{zq} = \frac{(z_{q+h} - z_t)}{z_q}100\%$, while the absolute changes are found as $\Delta z_q = z_{q+h} - z_q$, where q stands for the previous period and h for the forecast horizon. Success is measured as $g_{zq} \geq \bar{g}$ for the expectations about the relative growth and as ($\Delta z_q \geq \bar{\Delta}$) for the expectations about the absolute changes. The choice of the cut-off values $\bar{g}$ ($\bar{\Delta}$) depends on the research goal. For example, to evaluate whether a leading indicator accurately classifies future periods as those with positive economic growth versus recessionary periods, an analyst would consider any small and positive threshold (i.e. g=0.01% or Δ=0.01) as "success."

To assess whether a variable exceeded its long-run trend or a target value, a researcher could specify success as a long-run average observed in the past (i.e. $g = 2\%$ or $\Delta = 2$). When "success" is defined as an acceleration in the observed growth, one would set $\bar{g}$ ($\bar{\Delta}$) at the level of the past realized growth (i.e. $\bar{g} = g_{zq-h}$ or $\bar{\Delta}$= Δz_{q-h}).

The next step is to choose a forecasting rule. The formula in [5] suggests that the higher values of consensus score S indicate the larger difference between the fractions of respondents giving positive and negative answers. Thus, if the survey expectation has predictive value as a leading indicator, then the higher the score, the more likely it is to observe positive economic growth in a corresponding variable. A binary directional forecast can be produced by selection of a threshold *t* from the interval (a, c) and applying the following forecasting rule:

$$\widehat{Y_{zq}} = 1\ (the\ change\ in\ the\ variable\ z\ is\ forecasted\ as\ \text{success})\ if\ s \geq t \qquad [7]$$

$$\widehat{Y_{zq}} = 0\ (the\ change\ in\ the\ variable\ z\ is\ not\ forecasted\ as\ \text{success})\ if\ s < t$$

For any chosen threshold $t \in (c, a)$ there will be two types of correct predictions (true positives and true negatives) and two types of misclassifications (false positives and false negatives)[13]. These values are organized in a 2x2 contingency table (Table 1).

Table 1. The 2x2 contingency table for binary forecasts.

		Forecasts		Total in rows
		$\hat{Y} = 1$	$\hat{Y} = 0$	
Actuals	Y=1	True positives (TP)	False negatives (FN)	$N_{Y=1} = TP + FN$
	Y=0	False positives (FP)	True negatives (TN)	$N_{Y=0} = FP + TN$
Total in columns		$N_{\hat{Y}=1} = TP + FP$	$N_{\hat{Y}=0} = TN + FN$	$N = TP + FN + TN + FP$

A contingency table above allows us to calculate the following conditional probabilities:

$$TPR(t) = p(s \geq t | Y = 1) \qquad [8]$$

$$FPR(t) = p(s \geq t | Y = 0),$$

where TPR and FPR stand for the *true positive rate* and *false positive rate* respectively. The TPR[14] measures a probability that the consensus score S is not below the chosen threshold t, conditional on the observing the actual success (Y=1)[15]: in other words, the forecast of upward growth was correct. Similarly, the FPR[16] assesses a probability that there was an upward growth forecast while the actual growth was not observed (Y=0): in other words, the forecast of upward growth was wrong. The true negative rate (TNR) and the false negative rate (FNR) are

[13] The false positives and false negatives are often called type I and type II errors respectively.
[14] It is also known as sensitivity, recall, a hit rate, and a probability of detection.
[15] The following discussion will consider a case when "success" is defined as any positive growth.
[16] It is also known as specificity.

complementary to the two probabilities above[17]. Table 2 presents the contingency table above in terms of probabilities.

Table 2. Binary forecasts, actual changes and associated probabilities for a threshold t

Actuals	Forecasts	
	$S \geq t$	$S < t$
$g_{zq} \geq \bar{g}$	$P_1 = TPR(t) = p(s \geq t\|Y = 1)$	$1 - P_1 = FNR(t) = p(s < t\|Y = 1)$
$g_{zq} < \bar{g}$	$1 - P_2 = FPR(t) = p(s \geq t\|Y = 0)$	$P_2 = TNR(t) = p(s < t\|Y = 0)$

These probabilities are not observable, but can be estimated using the data from table 1 as follows: $TPR(t) = TP/(TP + FN)$, $FPR(t) = FP/(FP + TN)$, $FNR(t) = FN/(FN + TP)$, and $TNR(t) = TN/(TN + FP)$. Their values depend on a chosen value of threshold t used in the forecasting rule. Thus, every value of threshold will result in the corresponding contingency table and probabilities of correct (wrong) forecasts given the actual change.

1.2.3. Traditional accuracy measures and tests in evaluation binary forecasts as classifiers

The entries of each contingency table (such as table 1) are used to calculate the accuracy statistics and to test the predictive power of a consensus score S as a binary classifier[18]. *The overall accuracy ratio (ACC)* is measured as a fraction of correct predictions in the total number of observations:

$$\widehat{ACC}(t) = \frac{TP + TN}{TP + TN + FP + FN} \qquad [9]$$

The probabilities of detection the periods when Y=1 and Y=0, denoted as *PD1* and *PD0*, measure the fraction of the number of periods when the state of interest was predicted correctly to the total number of such forecasts:

$$\widehat{PD1}(t) = \frac{TP}{TP + FP} \qquad [10]$$

[17] The FNR and FPR are calculated as $FNR(t) = 1 - TPR(t) = p(g_{EWI} \leq t|Y = 1)$ and $TNR(t) = 1 - FPR(t) = p(g_{EWI} \leq t|Y = 0)$.

[18] These statistical measures depend on the given sample and chosen forecasting rule.

$$\widehat{PD0}(t) = \frac{TN}{TN + FN} \quad [11]$$

Youden (1950) offered to assess performance of binary classifiers with a J-index[19], which ranges from 0 to 1, with the higher values supporting superior predictive ability.

$$J(t) \; = \; TPR(t) - FPR(t) \; = \; TPR(t) + TNR(t) - 1 \quad [12]$$

Lahiri and Wang (2006) emphasized that in forecasting rare events (i.e., periods of GDP decline), the accuracy measure should put more weight on the accuracy of correctly identifying the less frequent state. They popularized the Heidke Skill Score (HSS) which shows the ratio of correct predictions relative to that obtained under a random forecast without skill. Proposed by Doswell, Davies-Jones, and Keller (1990), the HSS ranges from -1 to 1, with the higher values pointing to the stronger predictive power. The HSS can be rewritten using notations adopted in this chapter:

$$HSS = \frac{2 * (TP * TN - FP * FN)}{FN^2 + FP^2 + 2TP * TN + (FN + FP)(TP + TN)} \quad [13]$$

There are several tests to evaluate the predictive power of a classifier S. Fisher (1922) suggested the exact test of statistical significance (FE) of independence between binary forecasts and their realizations. It calculates the conditional probability of obtaining the values in each of the cells of the 2x2 contingency table using the hypergeometric distribution[20].

Henriksson and Merton (1981) applied the FE test to study market-timing skills of investors[21]. They argued that forecasts have value if they change the investors' prior probability distributions. This requires $P_1 \; + P_2 \; > \; 1$, where P_1 and P_2 are the unknown conditional probabilities of correct forecasts given the actual realizations (see table 2, p.8). They test a null hypothesis $H_0 : P_1 \; + P_2 \; = \; 1$ against an alternative $H_a : P_1 \; + P_2 \; > \; 1$. They reject a null if $TP \geq$

[19] The same performance measure as the J-index above is known in current literature as a Kuipers score (Granger & Pesaran, 2000, p.8-9) and as a Pierce score (Lahiri & Wang, 2013, p.185)

[20] $\binom{\mathrm{N}_{y=1}}{TP}\binom{\mathrm{N}_{y=0}}{FP} / \binom{N}{\mathrm{N}_{\hat{y}=1}} = \frac{\mathrm{N}_{y=1}!\mathrm{N}_{y=0}!\mathrm{N}_{\hat{y}=1}!\mathrm{N}_{\hat{y}=1}}{\mathrm{TP!FN!TN!FP!N!}}$

[21] They say it is a "test of independence between the market timer's forecast and whether or not the return on the market portfolio is greater than the return from riskless securities" (p. 517).

$x^*(c)$, where $x^*(c)$ is a solution to $\sum_{x=x^*}^{\overline{n_1}} (N_{y=1} x)(N_{y=0} N_{\hat{y}=1} - x)/(N N_{\hat{y}=1}) = 1 - c$, c is a chosen confidence level, and $\overline{n_1} = \min(N_{y=1,} N_{\hat{y}=1,})$.

Schnader and Stekler (1990)[22] offered an alternative test using a 2x2 contingency table. They test *Ho: forecasts are independent of the observed events* using a chi-square statistics[23]

$$\chi^2 = \sum_i \sum_j \frac{(N_{ij} - N_{pij})^2}{N_{pij}}, \qquad [14]$$

where N_{ij} and N_{pij} are the actual and predicted frequencies in the ij$^{\text{th}}$ cell.

Other accuracy measures were proposed[24]. All the measures and tests above evaluate the local power of a consensus score as a binary classifier at a given value of a threshold t. Different choices of the threshold t in a forecasting rule [7] could result in different conclusions about the classifier's predictive value. Hutson et al. (2014) proposed to find such threshold assuming that the "no change" category is bounded by 'x' standard deviations from the mean, and therefore they forecast up when $t = \mu + x\sigma$, with x=0.5 and 1.0 respectively.

This study suggests to use the ROC curves analysis in evaluating the binary predictions, including directional forecasts from the qualitative surveys. Next section discusses how it helps: (i.) to assess the global performance of a classifier over the entire range of possible thresholds; (ii.) to identify indicator's ranges for which the forecasts would be accurate out-of-sample; and (iii.) to determine optimal thresholds t^* which increase the forecast accuracy.

1.2.4. ROC curves analysis in evaluation binary forecasts as classifiers

An ROC curve was first introduced by Peterson and Birdsall (1953) as a signal detection

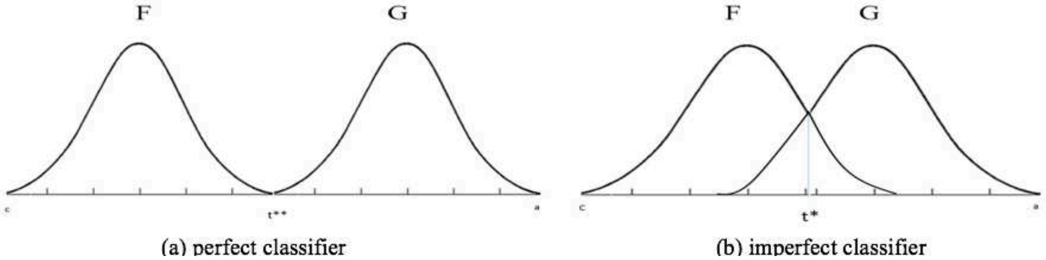
[22] They also extended Merton's method to test the value of directional forecasts for the U.S. real GNP. They suggest the forecasts have value if they can change the priors that the economists had about the probabilities of the growth rates.

[23] With the Yates' correction, the formula above becomes: $\chi 2 = \sum_i \sum_j \frac{(|N_{ij} - N_{pij}| - 0.5)^2}{N_{pij}}$

[24] Pesaran and Timmermann (1992) developed predictive failure statistics which does not require quantitative information and is distribution-free. For the 2x2 contingency table, it is equivalent to the hypothesis of independence. Lahiri and Wang (2006) also used odds ratio and odds ratio skill score.

tool, and was later adopted by other sciences. The ROC curves analysis[25] has already been applied in economics to evaluate predictors of credit defaults (Soberhart & Keenan, 2001; Blöchinger & Lieppold, 2006), to assess the ability of credit booms to predict recessionary and expansionary periods (Berge & Jordà, 2011; Shularic & Taylor, 2012), and to evaluate subjective probability forecasts of GDP declines (Lahiri & Wang, 2013).

This chapter is the first study to propose how to apply the ROC curves analysis to qualitative surveys and assess the directional accuracy of the survey consensus scores S as binary classifiers. This new methodology can be extended to evaluate value of any variable to produce binary forecasts. Section 1.2.1 motivated that S should take high values when the variable of interest is expected to grow above the given growth cut-off (Y = 1), and low values otherwise (Y = 0). The analysis assumes that scores of S when Y = 1 and Y = 0 belong to two different population distributions, F and G respectively.



(a) perfect classifier

(b) imperfect classifier

Figure 1. Examples of population distributions F and G for two different classifiers

To forecast the future direction of growth, an analyst should choose a threshold t. Ideally, she would like to find a classifier for which there exists threshold value t^{**} such that it completely separates the two distributions, observing $Y = 1$ in all the periods when $S \geq t^{**}$, and $Y = 0$ when $S < t^{**}$ (see Fig.1, panel a). Such perfect classifiers usually do not exist. Thus, an analyst's goal is to find an informative although an imperfect classifier. It would have a range of scores where the two distributions overlap (see Fig.1, panel b). Conclusions about the predictive value of a classifier will depend on the values of $TPR(t)$ and $FPR(t)$ at the chosen

[25] A reader is referred to Lahiri and Young (2013) to learn about the basics of the ROC curves analysis.

value of a threshold t. A ROC curve is a function $TPR(t) \ = \ h\,[FPR(t)]$, which maps all FPR values into respective TPR values for every value of t. Graphically, it is a curve which lies inside a [0,1]x[0,1] unit square in the Cartesian coordinate system. This is because both rates are conditional probabilities and take values from 0 to 1.

Every point on an ROC curve corresponds to a unique contingency table, which is based on the forecasting rule and a threshold t used with it. Thus, the ROC curve sums up the predictive value of a classifier at all the possible threshold values. It can be used to compare the accuracy statistics at different threshold values.

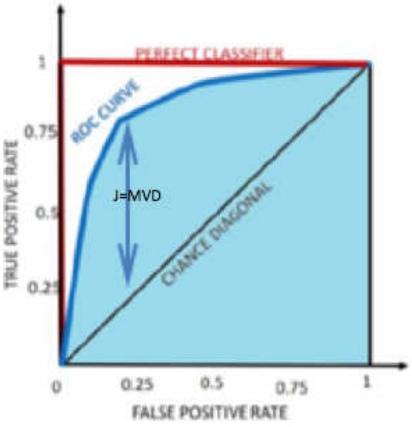


Figure 2. Sample ROC curves for two informative classifiers and a random guess

Fig. 2 above presents an ROC curve for three types of classifiers: an uninformative one (chance diagonal), an informative but imperfect classifier, and a perfect classifier. The uninformative classifier will form a 45-degree line which includes all points with $TPR(t) = FPR(t)$. It can be interpreted as an ROC curve for a random guess. An informative but imperfect classifier will result in a concave downward ROC curve above the chance diagonal including points with $TPR(t) \ > \ FPR(t)$ for any t. The higher the predictive value of a classifier S, the closer the ROC curve will be to the upper left corner of a unit square. A perfect classifier would result in the ROC curve formed by the left and upper sides of the unit square. One can note that a vertical distance indicated with an arrow in Fig. 2 equals the J-index introduced in section 1.2.3.

The statistical properties of an ROC curve are well-known[26]. The AUC[27] is found as $AUC = \int_0^1 TPR\ dFPR$ and measures a shaded area in the Fig. 2 above. It is used as a global measure of the predictive value of a classifier over the range of thresholds $t \in (c, a)$, and, therefore, it does not depend on the choice of t in a forecasting rule. The AUC can be interpreted as a probability that a classifier S will allocate a higher score to a randomly chosen observation from a population $Y = 1$ rather than from a population Y=0.

A classifier is informative in-sample if its AUC, including its 95% confidence intervals, is significantly greater than 0.5. This is tested with a null hypothesis $H_0: TPR = FPR$ against an alternative $H_a: TPR > FPR$. The higher AUC points to a higher classifying ability. This chapter applies the following grading scale: it classifies predictive value as "very high/very strong" for $AUC \in (0.9; 1]$, "high/strong" for $AUC \in (0.8; 0.9]$, "moderate" for $AUC \in (0.7; 0.8]$, "low/weak" for $AUC \in (0.6; 0.7]$, and "very low / very weak" for $AUC \in (0.5; 0.6]$.

However, we do not observe a true ROC curve which we would obtain if we could survey the entire population of macroeconomic forecasters. Instead, every quarter, the survey will have different composition of responding experts. Thus, the observed ROC curve and its AUC are only estimates for a given sample and subject to a sampling error. One can approximate a true ROC curve assessing the 95% confidence intervals around it using the maximum likelihood estimation (MLE) approach and assuming bi-normal distributions for F and G[28].

The model assumes the existence of an unobserved, continuous, latent variable that is normally distributed (perhaps after a monotonic transformation) in both populations with means μ_F and μ_G, and variances σ_F and σ_G, respectively. The *k* categories of the rating variable result

[26] See, for example, Krzanowski and Hand (2009).

[27] AUC and its standard errors are estimated either non-parametrically (using trapezoid approximation and empirical distribution of scores S), or parametrically (using MLE and assuming bi-normal distribution of scores S).

[28] Dorfman & Alf (1969) developed the MLE approach to estimate a ROC curve with confidence intervals using the rating data. Ma & Hall (1993) proposed how to obtain simultaneous confidence bands for the entire curve. Pepe (2003) and Pepe, Longton, & Janes (2009) adopted these methods for continuous data.

from partitioning the unobserved latent variable by $(k-1)$ fixed boundaries. The method fits a straight line to the empirical ROC points plotted using normal probability scales on both axes. The intercept from the fitted line $(\mu_F - \mu_G)/\sigma_F$ and its slope σ_G/σ_F measure the standardized difference between the two latent population means and the ratio of the two standard deviations. The null hypothesis that the means and variances in the two populations are equal is evaluated testing that the intercept and slope are equal zero and one respectively.[29]

I propose the following criteria to evaluate the classifier's predictive value in-sample and out-of-sample. A classifier has *in-sample predictive value* if the two conditions hold: (i.) its ROC curve is entirely above the chance diagonal, and (ii.) its AUC is significantly greater than 0.5. A classifier with in-sample predictive value has *out-of-sample predictive value* only for those ranges of the consensus score S at which its ROC curve is significant (and its confidence bands are entirely above the chance diagonal)[30].

Then I compare the accuracy and tests statistics at different cut-off values within the significant out-of-sample range. A forecasting rule with a threshold $t \leq min(S)$ will classify all the periods as "ups" and none as "not ups" and result in a point at the upper right corner of a unit square (TPR=FPR=1). The higher the threshold, the lower both the TPR and the FPR are. If the threshold $t \geq max(S)$, then all the periods are classified as "not ups" and none as "ups" and result in TPR=FPR=0. The ROC-optimal threshold $t^*$is determined by the maximum J-index[31].

In the qualitative survey of experts, the true state is neither known or verifiable as there is

[29] Let Rj for j=1,2…k indicate the categories of the rating variable, let i=1 and 2 if the subject belongs to the populations G and F respectively. Then p(i=1)=H(Zj)-H(Zj-1), where Zk=(xk-μG)/σF, H is the cumulative normal distribution, H(Z0)=0, and H(Zk)=1. Also, p(i=2)=H(bZj-a)-H(bZj-1-a), where b=σG/σF and a=(μF-μG)/σF. The parameters a, b and the (k-1) boundaries are obtained simultaneously by maximizing the log-likelihood function logL=i=12j=1krijlog{p(Rj|i)}, where rij is the number of Rj's in group i. The area under the fitted ROC curve is computed as Φ(a1+b2), where Φ is the standard normal cumulative distribution function. Point estimates for the ROC curve indexes are as follows: δ(m)=a/b, de=2a/(b+1), da=a2/(1+b2). Variances for these indexes are computed using the delta method. The δ(m) estimates (μF-μG)/σF, de estimates 2(μF-μG)/(σF-σG), and da estimates 2(μF-μG)/(σF2-σG2).

[30] We can reject the hypothesis that the consensus forecast does not have value.

[31] Maximum J-index is equivalent to the maximum vertical distance (MVD) between the ROC curve and the chance diagonal and is asymptotically equivalent to the Kolmogorov-Smirnov statistic for testing non-parametrically the equality of the two population distribution functions F and G.

no unique interpretation to the questionnaire' verbal answers. I propose to use a *reverse ROC analysis, applying the same classifier to several assumed "true state" binary indicators (i.e. when growth exceeds x%)*. This helps to find association between the optimal WES consensus scores and their implied growth rates.

1.3. World Economic Survey of the U.S. macroeconomic indicators

Section 1.3.1 explains the World Economic Survey (WES) questionnaire, summarizes its statistics, and provides some insights into its use in directional forecasting. Section 1.3.2 presents the actual US data and their transformations utilized in the evaluation of the WES directional forecasts. Section 1.3.3 compares the dynamics in actual data and the WES expectations.

1.3.1. WES questionnaire, consensus scores, and their interpretations

The WES[32] is conducted by the Center for Economic Studies Ifo Institute (CESifo), which sends its questionnaire (see Fig.A1 in Appendix A) to the volunteering experts in more than 100 countries. The experts from academia, finance, business, and economic departments of German embassies respond to the survey in the first month of each quarter, and the CESifo publishes aggregated results a month later. They form two types of expectations: (i.) future expectations about a situation by the end of the next six months (FE), and (ii.) present judgment about the situation in the current quarter (PJ).

Table 3. Summary of the WES expectations and their categorical responses[33]

Surveyed variables	Expectation type	Lead (months)	Categorical responses
Overall economy, capital expenditures, private consumption	PJ	2.5[34]	1-good, 2-satisfactory, 3-bad
	FE	8.5	1-better, 2-about the same, 3-worse
Exports, imports, inflation rate, short-term interest rate	FE	8.5	1-higher, 2-about the same, 3-lower
Trade balance	FE	8.5	1-improvement, 2-no change, 3-deterioration

[32] Stangl (2007A) and Garnitz (2015) provide a detailed survey documentation.

[33] The WES also collects future expectations for the long-term interest rate, domestic share prices, and exchange rates with three major trade partners, but the series are too short to get consistent results.

[34] The survey results are available in the middle of the quarter when it was conducted, while the first advanced estimates for a specific quarter are published only at the end of the first month next period.

Table 3 above indicates that there are three groups of responses: (i.) positive (good, better, higher, or improvement); (ii.) neutral (satisfactory, about the same, or no change); and (iii.) negative (bad, worse, lower, deterioration). However, it gives no guidance about the economic meaning of each verbal response. Hutson et al. (2014) interpreted all the positive (negative) answers as any growth above (below) zero, and all the neutral answers as a zero change for all variables in their analysis. This study agrees on the zero change benchmark for the future expectations about the exports, imports, trade balance, inflation rate and short-term interest rate. The expectations about the overall economy, consumption, and investment require different assumptions. I assume that the experts: (i) answer "better" and "good" when a variable displays a moderate to strong growth; (ii.) reply "about the same" or "satisfactory" when the economy experiences a weak growth; and (iii.) associate "worse" and "bad" with the periods of negative growth.

These assumptions are in line with professional conventions. For example, Stekler and Symington (2016) suggested how to create indexes of present and future economic outlook (see section 3.2.2 for details) and set their correspondence to nine verbal assessments of the overall economy, with three values for each of the pessimistic, neutral, and optimistic views. The WES also has three types of answers, and its consensus scores belong to one of nine brackets. This parallel between the SS index and the WES categories suggests that the survey consensus score should help its users to distinguish the periods with pessimistic economic conditions from the neutral or optimistic ones.

The consensus scores S are found as a weighted average after each type of response (1, 2, 3) is assigned a value a=9, b=5, and c=1, respectively (see [4] in sec.2.1.) and therefore belong to a range [1,9]. These scores are easier to interpret after their transformation into a balance statistic B using relation [5] from sec.2.1. Table A1 in Appendix lists all the possible pairs (B, S). One can see that S=5 when B=0, implying that there were equal shares of positive and negative responses. Scores S>5 correspond to a $B\in (0,1]$ and imply that there were 100B% more positive answers than negative ones. Scores S<5 relate to B∈[-1, 0) and suggest that there were 100B% more negative responses than positive ones.

Table 4 below summarizes statistics for the WES expectations about the U.S. economy in 1989-2015. There are 108 quarterly observations for all variables except the inflation rate[35]. Ranges $[\mu-\sigma, \mu+\sigma]$ and $[\mu-0.5\sigma, \mu+0.5\sigma]$ indicate the "no change" intervals as suggested by Hutson et. Al (2014). The upper borders are used as alternative threshold values to produce directional forecasts around 0% annual growth rate.

Table 4. Summary statistics for the WES of the U.S. economy in 1989: Q1 - 2015: Q4

WES expectation	Variable	N	M	σ	Min	max	$\mu-\sigma$	$\mu+\sigma$	$\mu-\sigma/2$	$\mu+\sigma/2$
Present judgement	Overall economy	108	5.29	2.17	1.2	9	3.12	7.46	4.21	6.38
	Capital expenditures	108	4.86	2.09	1.1	8.3	2.77	6.95	0.71	3.48
	Private consumption	108	5.32	2.04	1	8.8	3.28	7.36	0.40	3.68
Future expectations	Overall Economy	108	5.50	1.67	1.3	8.8	3.83	7.17	4.67	6.34
	Capital expenditures	108	5.44	1.55	1.9	8.5	3.89	6.99	4.67	6.22
	Private consumption	108	5.21	1.56	1.3	8.8	3.65	6.77	4.43	5.99
	Exports	108	6.31	1.35	2.6	8.4	4.96	7.66	5.64	6.99
	Imports	108	6.09	1.21	2.3	8.1	4.88	7.30	5.49	6.70
	Trade balance	108	4.88	1.43	1.6	8.4	3.45	6.31	4.17	5.60
	Inflation rate	98	5.96	0.95	3.1	8.1	5.01	6.91	5.49	6.44
	Short-term interest rate	108	6.01	1.91	1.7	9	4.1	7.92	5.06	6.97
EC indicator	Overall Economy	108	5.39	0.89	2.4	7.1	4.5	6.28	4.95	5.84

The CISifo also presents the Economic Climate (EC) indicator, which they use in its "business clock" diagram (see Appendix, Fig.A3)[36]. It is found as a simple average of the future expectations and the present judgement about the overall economy:

$$EC = 0.5\,FE + 0.5\,PJ \qquad [15]$$

Kudymova, Plenk, and Wohlrabe (2013) found that it highly correlates with the rates of economic

[35] Sample period for inflation rate is 1991: Q3 - 2015: Q4 due to a later inclusion of the question in a survey and limited availability of the real-time actual data in earlier periods.

[36] Its horizontal and vertical axes show combinations of the present judgment and future expectations about the overall economy. As economy moves through a business cycle, the combination move clockwise.

growth[37]. Intuitively, the EC combination contains information about the stage of the business cycle. For example, a "good" present economy and "better" future economic conditions signal that an economy is expanding toward its peak, while "bad" present conditions combined with "better" future expectations reveal that an economy is recovering from the trough.

Table 4 above shows that mean scores for all the expectations except the present judgment of capital expenditures and future trade balance were above 5. This means that the consensus of experts on average: (i.) evaluated present situation regarding overall economy and private consumption as "good"; (ii.) judged current situation regarding capital expenditures as "bad"; (iii) expected "better" future GDP, private consumption, and capital expenditure; (iv.) expected higher imports, exports, inflation and short-term interest rates, and lower trade balance in the short 6-months horizon. The average EC score exceeded 5, implying that the experts expected the U.S. economy was in the expansion or recovery stage in most periods. One can also note that the EC indicator has lower standard deviation than the scores for the present judgment or future expectations on the overall economy.

The analysis above allows to make the following assumptions and interpretations:

1) The WES questions with verbal answers "lower/no change/higher" or "deterioration/no change/improvement" should be used to predict whether a surveyed variable will exhibit a positive absolute growth in the next period. Questions with verbal responses "worse/the same/better," or "bad/satisfactory/good" should be used to predict whether a surveyed variable will exhibit a positive absolute growth in the next period. The WES consensus score for a positive growth forecast should exceed 5.

2) The survey user who wants to identify whether a situation is expected to be pessimistic or not, should produce such a binary forecast using the WES consensus value on the border between the negative and neutral categories. The optimal S value should be below 5.

[37] This correlation improves with time (evidence of learning effects) and does not depend on the number of responses (evidence of survey representativeness). Their analysis covered 43 countries.

3) The survey user who wants to identify whether a period is expected to be optimistic or not should produce such a binary forecast using the WES consensus value on the border between the neutral and positive situations. The assumed optimal S value is above 5.

4) The EC indicator is expected to be a more accurate classifier of the future overall economy than future expectations alone.

5) The pessimistic, neutral, and optimistic views should respectively match the periods of decline, small positive growth, and high positive growth. The numerical borders defining these growth categories depend on the past averages and/or trends. The ROC curves analysis suggests the S values which are optimal to produce directional forecasts in excess of a specific growth rate.

The next section explains which actual data are used to evaluate the WES expectations and summarizes their statistics.

1.3.2. Actual U.S. data subject to the directional forecasting and their comparisons with the WES expectations

Assessments of the overall economy, private consumption, and capital expenditures are compared with real GDP/GNP, total personal consumption expenditures, and total investment. Future expectations about the exports, imports, and trade balance are matched with real exports, imports, and net exports of goods and services. The Real-Time Data Set for Macroeconomists (RTDSM) from the Philadelphia Fed serves as a source of actual U.S. GDP/GNP, its components, and consumer price index (CPI). Expectations about the short-run interest rate are checked with the 3-months money market rates from the IMF International Finance Statistics (IFS) Database.

Present judgments about GDP, investment, and consumption are evaluated with their three-months growth rates. Future expectations are compared with the six-months growth rates (for the real GDP and its components), or absolute changes (for the trade balance, CPI and short-term interest rate).

Table 5 below shows that the U.S. GDP, consumption, and investment grew by 0.59%

(1.21%), 0.68% (1.36%) and 0.93% (1.92) on average every three (six) months respectively[38]. Exports and imports increased respectively by 2.11% and 2.55% every six months on average causing a decline in the trade balance by 8.87 billion USD over the same period. The CPI-based inflation rate and 3-months money market interest rate declined by 0.12% and 0.15% on average every 6 months over the sample period[39].

Table 5. Summary statistics for the observed changes in actual data[40].

Variable	Notation	Sample period	N	Mean	St. Dev.	Min.	Max.
3-months growth rates, %							
GDP	G_{rgdp}	1989: Q1 – 2015: Q4	108	0.59	0.50	-1.57	1.74
Investment	G_{rinv}	1989: Q1 – 2015: Q4	108	0.93	3.16	-16.14	8.04
Consumption	G_{rcon}	1989: Q1 – 2015: Q4	108	0.68	0.50	-0.89	2.00
6-months growth rates, %							
GDP	G_{rgdp}	1989: Q1 – 2015: Q4	108	1.21	0.93	-3.17	3.12
Investment	G_{rinv}	1989: Q1 – 2015: Q4	108	1.92	5.30	-21.07	12.93
Consumption	G_{rcon}	1989: Q1 – 2015: Q4	108	1.36	0.79	-1.85	3.48
Exports	G_{rex}	1989: Q1 – 2015: Q4	108	2.11	3.48	-14.47	8.60
Imports	G_{rex}	1989: Q1 – 2015: Q4	108	2.55	3.98	-14.25	11.89
Absolute 6-months changes[41]							
Trade balance	Δtb	1989: Q1 – 2015: Q4	108	-8.87	44.72	-176.5	131.6
Inflation	Δinf	1998: Q3 – 2015: Q4	87	-0.12	1.723	-7.32	3.27
Short-term interest rate	$\Delta irsr$	1989: Q1 – 2015: Q4	108	-0.15	0.818	-2.41	1.46

For every variable (the overall economy, consumption, and investment), I construct four binary indicators Y_k which take value 1 if the observed change exceeded k=0, 1, 2 and 3% at the annualized rate[42], and 0 otherwise. Tables A2-4 in Appendix show distribution of periods into binary categories and list all periods with $Y_k = 0$ (if a variable grew below k% at the annualized rate). For example, real GDP fell in eight periods when compared to the previous quarter (1990q4-1991q1, 2001q3, 2008q3-2009q2, and 2012q4) and in six periods when compared to a

[38] The corresponding annualized growth rates are about 2.74% for consumption and 4.14% for investment.
[39] Sample means of the CPI-inflation rate and short-term interest rate are 2.13% and 3.57% respectively.
[40] Following Croushore and Stark (2001), when data are subject to revisions, I evaluate forecasts with the first vintages, using the second, third, and last available (2016: Q3) vintages in robustness check.
[41] Trade balance is measured in billion USD, inflation and interest rate are expressed as percent.
[42] This chapter uses 0.01% to exclude very small positive changes.

period six months earlier (1990q4-1991q2, 2001q3-4, and 2008q4-2009q2)[43].

The next section assesses the predictive value of the WES expectations comparing their co-movements with the actual data and analyzing the score distributions by binary states.

1.3.3. Comparisons of the dynamics in actual U.S. data and WES expectations

Fig. 3 below focuses on the GDP predictions.

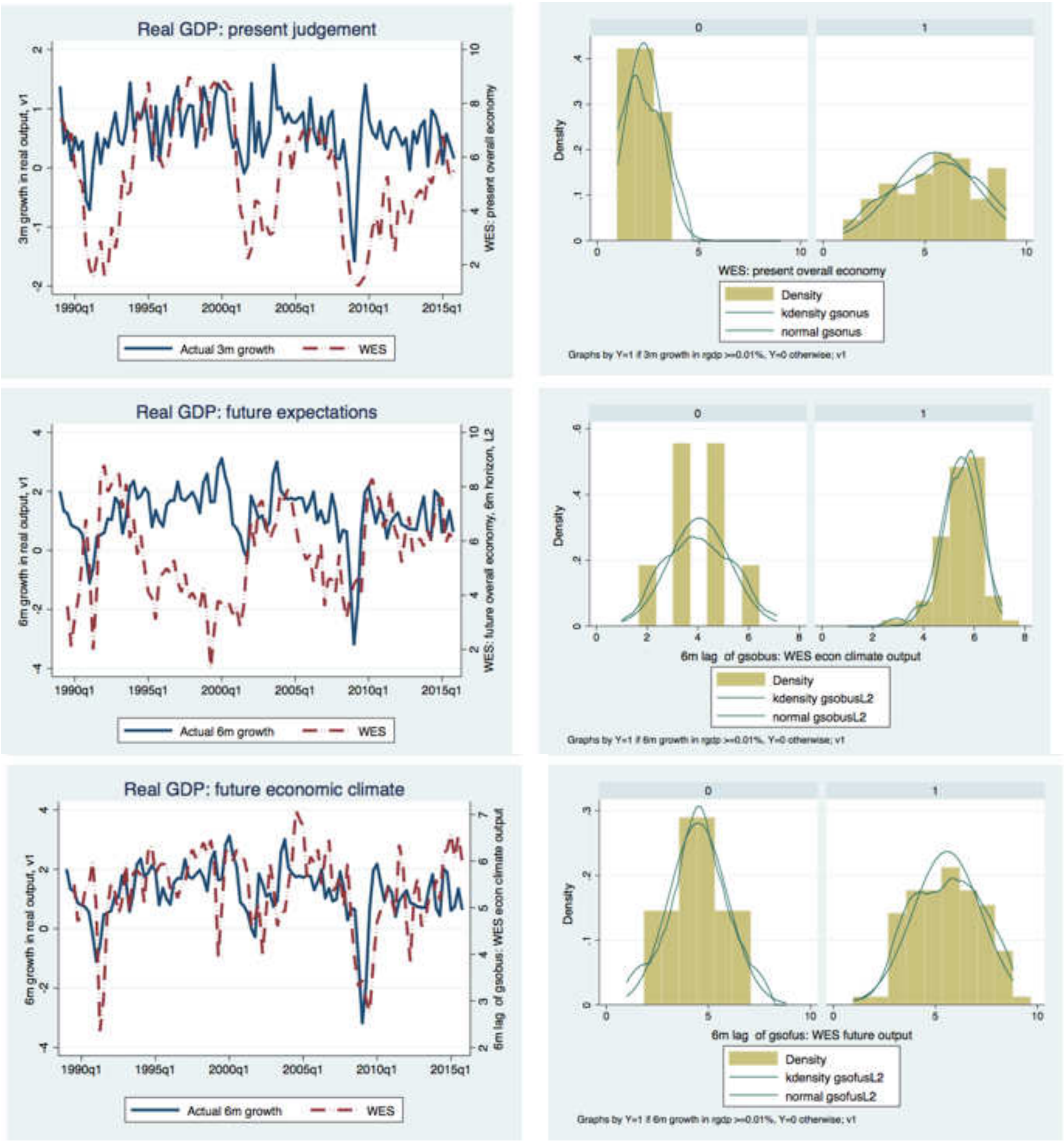


Figure 3. WES US GDP expectations: dynamics and scores distributions in Y=1 vs Y=0

[43] The NBER registered three recessions (measured as months of the peak through trough) in a sample period: 1) July 1990 – March 1991, 2) March 2001 – November 2001, and 3) December 2007 – June 2009.

The left column compares dynamics of the actual growth rates (left y-axis, solid blue line) and the respective WES consensus scores (right y-axis, red dash-dotted line). The right column shows distribution of the WES scores in periods with positive GDP growth (Y=1) and without it (Y=0) using histograms and kernel densities.

The top panel of Fig.3 assesses present judgment of the overall economy: it has the same dynamics as the three-month GDP growth rates (left panel); and its scores are always below 4 in recessionary periods (right panel) implying that it is an instructive classifier of the current quarter GDP growth. The middle and lower panels in Fig. 3 compare alternative predictors of the future GDP (h=2 quarters): future expectations of the overall economy and the EC indicator. One can see that the lagged EC indicator mimics dynamics of the 6-months GDP growth rate more accurately than the lagged future expectations of overall economy alone. The histograms in the middle and lower right panels tell a similar story: the distributions of the WES scores on the future expectations about the overall economy overlap when Y=1 and Y=0, revealing low discriminatory power; the EC indicator scores are more separated in Y=1 and Y=0 with only few misclassification cases; thus, should have high predictive value.

Fig. A5 in Appendix provides diagrams for other variables. Their analysis reveals that present judgment about consumption should have moderate-to-high predictive value, while future expectations about consumption indicate much lower discriminatory power. The WES experts had low skill in forecasting capital expenditures: the time-series show that they only confirmed what happened in the previous quarter rather than predict the actual dynamics, the histograms for Y=1 and Y=0 overlap heavily.

Similar conclusion can be drawn about the future exports, imports, trade balance and inflation rate. Expectations about the future short-run interest rate seem to co-move with the actual data well – therefore, they should have value in directional forecasting.

The next section supports the preliminary observations about the WES expectations with the empirical evidence obtained via the ROC curves analysis.

1.4. Empirical results: the ROC curves analysis and accuracy statistics

Sections 1.4.1 and 1.4.2 focus on the WES expectations about real GDP, consumption, and investment, for the current quarter and two-quarters ahead horizons. Section 1.4.3 evaluates the WES future expectations about foreign trade, inflation, and short-term interest rates.

1.4.1. Present judgement about real GDP, consumption, and investment

1.4.1.1. Present judgement: GDP

Fig. 4 below visualizes the ROC curves and their 95% confidence intervals for the WES present judgment about the U.S. overall economy.

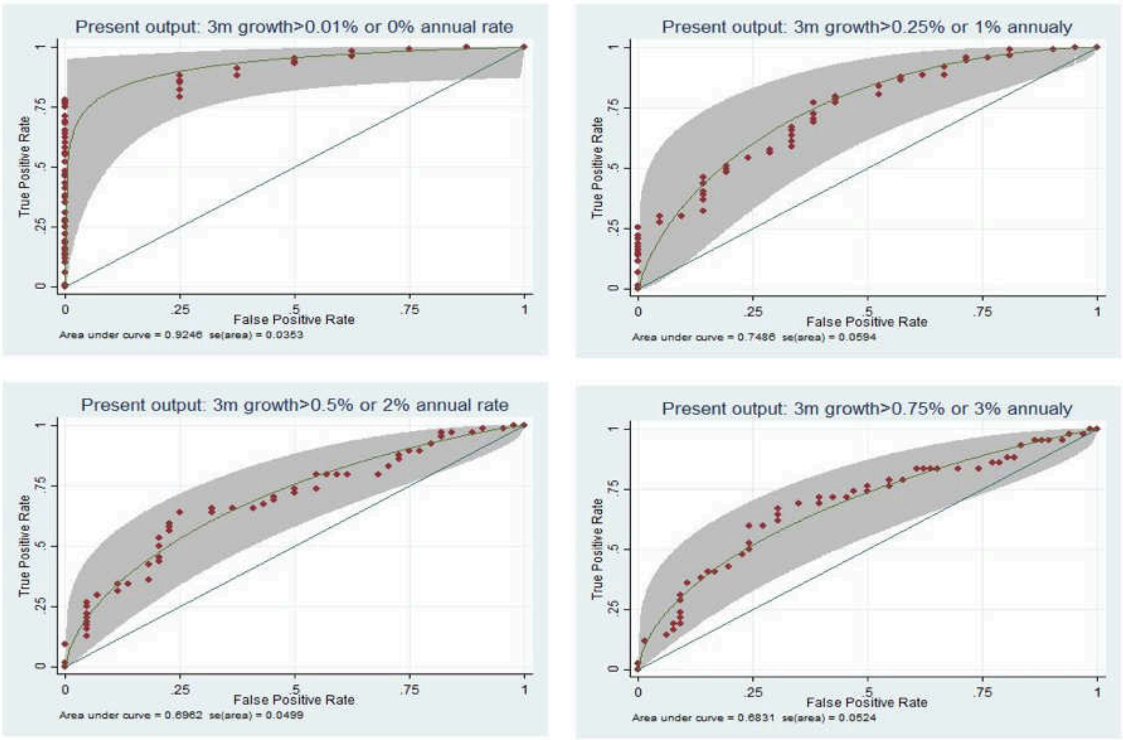


Figure 4. ROC curves for the present judgement about output

The upper left panel of Fig.4. shows that the survey has high value in predicting whether the current quarter GDP will grow above zero when compared with the previous period: the resulted ROC curve (AUC=0.92) is very close to a perfect classifier. Predictive power of the present judgment about overall economy decreases to a moderate level (AUC=0.75) when one uses it to predict whether the current quarter growth exceeds 1% at the annualized rate; its value declines further when one is interested in predicting whether the current quarter GDP growth exceeds 2% or 3% per annum (AUC=0.70 and 0.68).

Table 6 below confirms these observations presenting the AUC statistics, its standard errors and 95% confidence intervals, the ranges of the WES scores for which their respective ROC curves are significant, and the choice of optimal values to produce directional forecasts. One can see that the present judgment about output has high power in forecasting whether the real GDP growth is above zero in the current quarter: its ROC curve is significant for the consensus score values $S \in [1.4, 9]$, with the optimal S=3.6. When the present judgment about the overall economy is used to indicate whether the current GDP growth exceeds 1%, 2%, or 3% benchmark, the AUC declines to 0.75, 0.69, and 0.68 respectively.

Table 6. Present judgement about output: AUC statistics and implied WES scores

g (%)	N	AUC	Std. error	Asymptotic confidence intervals for the AUC		Optimal S value in-sample	Ranges of S for which the ROC curve is significant out-of-sample
				Lower 95%	Upper 95%		
0%	108	0.9246	0.0352	0.8555	0.9937	3.6	1.4 – 9.0
1%	108	0.7486	0.0594	0.6322	0.8650	4.3	1.5 – 6.9
2%	108	0.6962	0.0499	0.5985	0.7939	5.6	2.7 – 9.0
3%	108	0.6831	0.0524	0.5804	0.7859	6.7	3.5 – 9.0

The ROC curves are significant at broad ranges of the consensus scores. One can set a correspondence between these optimal consensus scores and the forecasted growth rates. For example, the survey user should forecast that the current quarter GDP will grow at least at 1% per annum for $S \geq 4.3$, at least at 2% for $S \geq 5.6$, and at least at 3% for $S \geq 6.7$. Comparison of the optimal WES scores for the current output with the actual annualized growth rates suggests that experts interpret GDP growth below 1% as bad, 1%-2% as satisfactory, and above 2% as good. The choice of the threshold in a forecasting rule (see [7], p.8) affects the accuracy results.

Table A5 in Appendix summarizes the entries of the 2x2 contingency table listing the number of TP, TN, FP, and FN in a column format that one would obtain if applied the ROC-suggested optimal threshold and the alternative values found as proposed by Hutson et al. (2014). It also contains the values of J-index, overall accuracy, probability of detection for Y=1 and Y=0, HSS, and Chi-square statistics. It indicates that using the ROC-optimal threshold $t^* = 3.6$, one would correctly identify all eight periods when the current quarter GDP declined and 78 out of 100 quarters with positive GDP growth. However, it would produce 22 false alarms of a GDP decline. The overall accuracy would reach 80%, with the TPR=0.78 and TNR=1. The probability of detection of the periods with Y=1 and Y=0 would reach 78% and 100% respectively. The Chi-square test with one degree of freedom rejects the null of independence at α=0.1% level.

If the survey user chose the thresholds following recommendations in Hutson et al. (2014), she could have picked t=5.0, 6.4, or 7.4. The first value assumes that the neutral consensus score corresponds to the zero change. The second and third values indicate that one should issue the forecast up when the consensus score is outside the no-change interval. Utilizing these alternative values, one would decrease the number of correctly identified periods of current quarter GDP growth but increase the number of issued false alarms when compared to the accuracy statistics a ROC-optimal $t^* = 3.6$ (see Table A5, panel A in Appendix). For example, using t=5.0, one would misclassify 16 quarters of the GDP growth as periods of decline. It would result in the 65% overall accuracy and only 62% probability of detection of the times of growth. The accuracy statistics worsen if the threshold increases further to 6.4 or 7.4. The power of the Chi-square independence test declines to 5% and 10% for t=5.0 and 6.4 respectively, losing its significance for t=7.4. The HSS is the highest for the ROC identified optimal threshold. This comparison of the accuracy measures favors the use of the ROC-implied threshold value ($t^* = 3.6$) while forecasting the current quarter directional change in GDP. This value is consistent with the interpretations above: times with the WES consensus $S \in [3.6, 5.7]$ correspond to the "satisfactory state" of the economy, while times with $S < 3.6$ predict periods of GDP decrease.

1.4.1.2. Present judgement: consumption

The ROC curves for the WES present judgment (top 4 panels of Fig. A6 in Appendix) indicate its consensus scores can foresee periods when private expenditures will increase by more than x% per annum in the current quarter. The in-sample discriminatory power is the highest (AUC=0.83) for a simple directional forecasting exercise (x=0%): it yields the ROC curve which is significant for a broad range of scores. An increase of x to 1% results in the ROC curve of a smaller area (AUC=0.73), decreasing to 0.62 and 0.61 for x=2% and 3% respectively. The higher x%, the shorter the ranges for which the present judgement about consumption has out-of-sample predictive value.

Table 7 below gives details on the ROC statistics and implied optimal values. It specifies that the scores on the present judgment about consumption chosen from a range [1.8, 9] can produce consistent out-of-sample predictions whether it will grow in the current quarter or not, with the lowest sum of type I and II errors at t*=3.5. The thresholds 5.0, 5.9, and 6.4 can be used optimally to forecast whether the consumption growth will exceed 1%, 2%, and 3% per annum, and their ROC curves are significant for broad ranges.

Table 7. Present judgement about consumption: AUC statistics and implied WES scores

g (%)	N	AUC	Std. error	Asymptotic confidence intervals for the AUC		Optimal S value in-sample	Ranges of S for which the ROC curve is significant out-of-sample
				Lower 95%	Upper 95%		
0%	108	0.8287	0.0636	0.7041	0.9533	3.5	1.8 – 9.0
1%	108	0.7245	0.0681	0.5911	0.8580	5.0	2.1 – 7.5
2%	108	0.6218	0.0537	0.5165	0.7271	5.9	5.4 – 9.0
3%	108	0.6095	0.0536	0.5044	0.7146	6.4	6.3 – 9.0

Appendix (Table A5, panel B) shows that if one used the present judgment about private consumption consensus score to predict the current quarter change in consumption expenditures, they would pick t*=3.5 as optimal threshold. This would help to correctly forecast 85 quarters when consumption increased and 7 periods when it declined[44], mislabeling 2 periods of decline and 14 quarters of growth. This t value yields the highest HSS score; the Chi-square test of independence at this threshold is significant at 0.1%. The alternative threshold values suggested by Hutson (t=3.9, 5, and 7.4) would increase the number of false alarms, and decrease the overall accuracy and probability of detection. Thus, the value t*=3.5 is indeed optimal and consistent with the interpretations above: times with the WES consensus for the present consumption

[44] These two missed periods of consumption decrease include 1989q4 and 1990q2.

S∈[3.5,5.9] indicate “satisfactory” state of the economy, while S<3.5 signals “bad” times when consumption falls.

1.4.1.3. Present judgement: investment

The ROC curves for the WES present judgment about investment (see 4 lower panels in the Fig.A6, Appendix) show some in-sample predictive value (their AUC ≈0.63-0.65), but their lower confidence borders are almost entirely below the chance diagonal implying that these expectations would not be able to discriminate the periods when the investment exhibits growth exceeding x% from those when it does not.

Table 8. Present judgement about investment: AUC statistics and implied WES scores

g (%)	N	AUC	Std. error	Asymptotic confidence intervals for the AUC		Optimal S value in-sample	Ranges of S for which the ROC curve is significant out-of-sample
				Lower 95%	Upper 95%		
0%	108	0.6298	0.0569	0.5183	0.7413	3.2	None
1%	108	0.6475	0.0540	0.5416	0.7533	4.8	3.5 – 6.5
2%	108	0.6407	0.0528	0.5372	0.7442	4.8	4.2 – 6.6
3%	108	0.6438	0.0521	0.5417	0.7459	7	3.9 – 9.0

Table 8 above shows that this survey indicator is a weak in-sample classifier of the current quarter investment growth for all x%. It has limited out-of-sample ability to recognize periods when the current quarter investment grew faster than x=1%, 2%, or 3%, but has no such skill to distinguish periods of positive and negative investment growth. Results imply that experts think of investment growth below 1% as bad, 1-3% as satisfactory, and above 3% as good.

Table A5 (panel C) in Appendix compares the in-sample accuracy statistics utilizing the ROC-implied thresholds (t*=3.2) and alternative values. At t*=3.2, the analyst would correctly classify 14 of 33 periods of investment decline and 60 of 74 quarters of its increase. It would result in the 69% overall accuracy, with 80% and 42% probability of detection of Y=1 and Y=0 respectively. Using t=3.5, one would miss 4 more periods of positive investment growth without any gain in correct predictions of the periods of investment decline. Higher thresholds (t=5, t=7) would improve in-sample identification of the periods with investment decline but worsen the number of correctly predicted times of investment growth. The Chi-square tests show that the entries of the contingency table are not independent for t>3.2 at α =5% significance level.

1.4.2. Future expectations about real GDP, consumption and investment

1.4.2.1. Future expectations: GDP

Fig.5 below indicates that the WES future expectations about the overall economy are weak in-sample predictors of the future GDP growth, but have no out-of-sample value as the lower confidence borders of the ROC curves are completely below the chance diagonal.

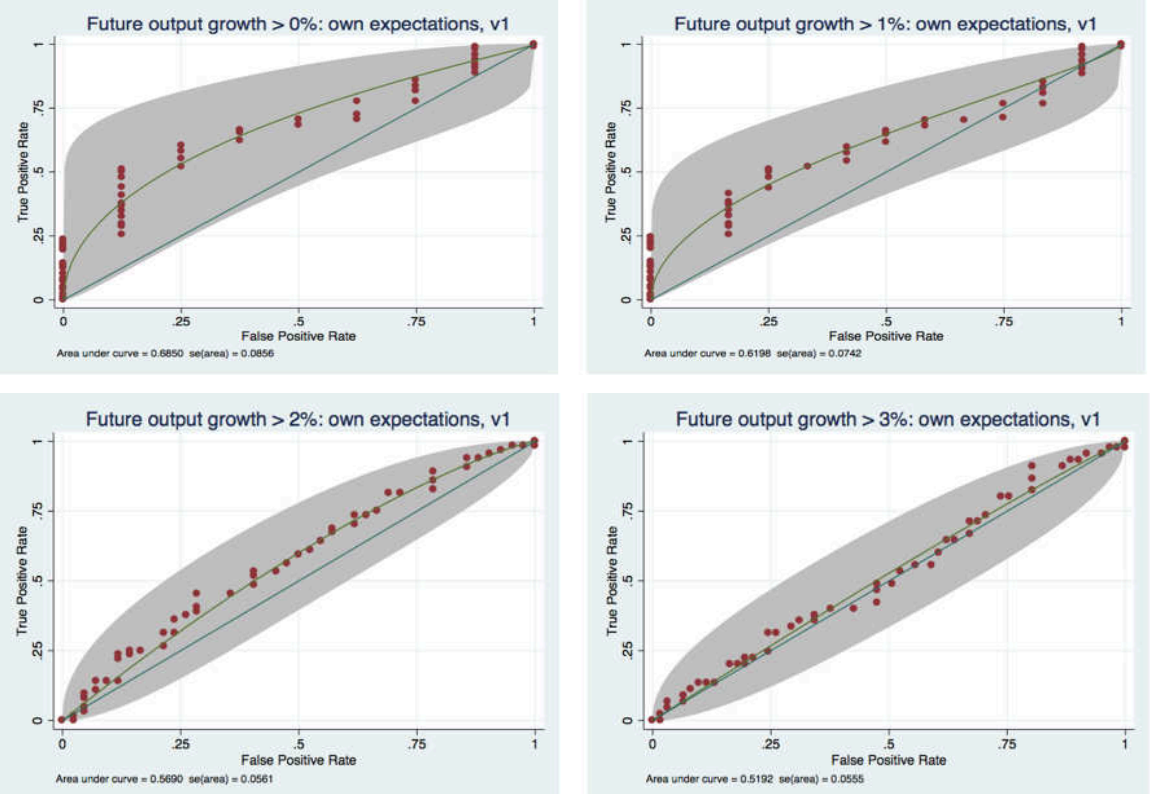


Figure 5. ROC curves for the future output growth: own expectations

Table 9 below confirms the preliminary insights from the Fig.5.

Table 9. Future GDP (own expectations): AUC statistics and implied WES scores

g (%)	N	AUC	Std. error	Asymptotic confidence intervals for the AUC		Optimal S value in-sample	Ranges of S for which the ROC curve is significant out-of-sample
				Lower 95%	Upper 95%		
0%	106	0.6850	0.0856	0.5173	0.8527	5.7	None
1%	106	0.6198	0.0742	0.4743	0.7653	5.7	None
2%	106	0.5690	0.0561	0.4592	0.6789	6.1	None
3%	106	0.5192	0.0555	0.4104	0.6281	3.7	None

If one used only the AUC criterion, she would conclude that the WES expectations about future overall economy have low value for x=0 and 1%, and very low value for x=2 and 3%. However, the AUC is significantly above 0.5 only for x=0%, indicating that this survey variable has low in-sample predictive value. Yet, it has no out-of-sample value because the ROC curve is not significant for the entire range of consensus score.

The ROC curves in Fig.6 below show that the EC indicator is a better predictor of the future GDP growth than future expectations about overall economy alone.

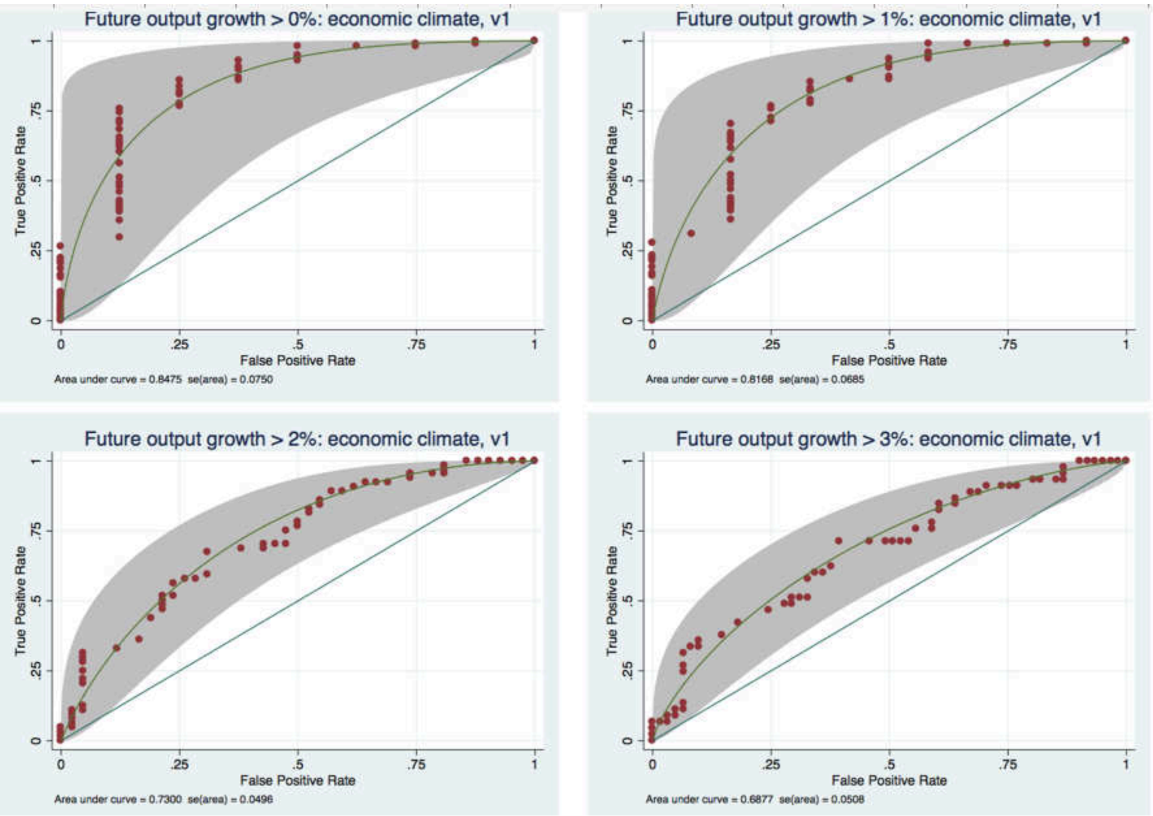


Figure 6. ROC curves for the future output growth: EC indicator

Table 10. Future GDP (EC indicator): AUC statistics and implied WES scores

g (%)	N	AUC	Std. error	Asymptotic confidence intervals for the AUC		Optimal S value in-sample	Ranges of S for which the ROC curve is significant out-of-sample
				Lower 95%	Upper 95%		
0%	106	0.8475	0.8475	0.7005	0.9945	5.05	2.75 - 6.05
1%	106	0.8168	0.0685	0.6825	0.9511	5.2	2.75 - 6.0
2%	106	0.7300	0.0496	0.6328	0.8271	5.5	2.35 - 6.6
3%	106	0.6877	0.0508	0.5882	0.7873	5.5	4.95 - 6.35

Table 10 above testifies that the EC indicator has high discriminatory power for all growth rates x% both in-sample and out-of-sample. Its predictive value is high when one aims to distinguish periods when future output growth at least at x=0 and 1% per annum and moderate for x=2% and low for x=3% per annum respectively. The optimal values of S point out that that the experts consider future 0-1% annual output growth as “the same,” and understand growth above 2-3% as a better situation.

Panel E in Table A5 in Appendix shows that if one would issue a signal about the future direction of GDP growth when the EC indicator exceeds the ROC-optimal threshold t*=5.05, she would correctly identify 7 out of 8 recessionary periods and 50 out 98 periods of GDP increase. The WES experts could not foresee recession of 1990q4. Issuing a signal of GDP growth at t=5 would result in close accuracy statistics. Using other two alternative values (t=5.8 and 6.35) would drastically decrease number of correctly identified GDP growth periods, resulting in insignificant Chi-square statistics and very low HSS value. The ROC-suggested threshold t*=5.05 is indeed optimal: it results in the highest HSS statistics and the strongest Chi-square test of independence (significant at α=0.01% level).

Comparison of panels C and D in Table A5 indicates that using the EC indicator at ROC-optimal threshold (t*=5.05) increases the overall accuracy from 54% to 76%, correctly identifying 24 more periods of GDP growth raising their probability of detection from 51 to 76% keeping the number of correctly identified recessions (7 out of 8) constant.

1.4.2.2. Future expectations: consumption

The top two panels in Fig.A7 in Appendix present the ROC curves for the WES expectations about future consumption. If one was looking only at the ROC curve being above the chance diagonal and its AUC being above 0.5, she would conclude that these expectations have low in-sample predictive value at x=0% and 1%. However, when one also accounts for the AUC standard error and compared the lower AUC confidence interval with 0.5, she would conclude that this survey expectation has no in-sample predictive value.

Table 11. Future expectations about consumption: AUC statistics and implied WES scores

g (%)	N	AUC	Std. error	Asymptotic confidence intervals for the AUC		Optimal S value in-sample	Ranges of S for which the ROC curve is significant out-of-sample
				Lower 95%	Upper 95%		
0%	106	0.6724	0.0955	0.4852	0.8596	4.2	None
1%	106	0.5830	0.1080	0.3712	0.7948	4.2	None
2%	106	0.5036	0.0603	0.3853	0.6218	6.4	None
3%	106	0.4260	0.0549	0.3184	0.5336	6.9	None

Table 11 above summarizes the ROC statistics. The optimal values suggest that the experts regard 0-1% annual consumption growth as “the same” and growth above 2-3% as a “better” situation. Panel F in Table A5 (Appendix) supports the ROC-implied threshold (t*=4.2). It accurately identifies 74 (from 100) and 4 (from 6) periods of consumption growth and decline, respectively. Thus, the survey issued 26 false alarms and missed two periods of consumption decrease (1990q4 and 1991q1). Using higher threshold values as suggested by Hutson (t=5, 6, or 6.7) would help gain 0, 1, or 2 periods of correctly classified future consumption decline, while decreasing the number of correctly identified growth periods to 55, 34, and 20 respectively. The overall accuracy and probabilities of detection of Y=1 and Y=0 are the highest at t*=4.2 (74% and 67% respectively). The Chi-square statistic is significant only at t*=4.2.

1.4.2.3. Future expectations: investment

All ROC curves for the future expectations about investment (see 4 lower panels in Fig. A7, Appendix) are above the chance diagonal suggesting some, although very low, in-sample predictive value (AUC$\approx 0.57 - 0.62$).

Table 12. Future expectations about investment: AUC statistics and implied WES scores

g (%)	N	AUC	Std. error	Asymptotic confidence intervals for the AUC		Optimal S value in-sample	Ranges of S for which the ROC curve is significant out-of-sample
				Lower 95%	Upper 95%		
0%	106	0.6202	0.0567	0.5091	0.7312	6.5	None
1%	106	0.6191	0.0558	0.5016	0.7202	6.5	None
2%	106	0.5992	0.0545	0.4923	0.7060	6.5	None
3%	106	0.5722	0.0546	0.4652	0.6791	4.8	None

However, only the lower confidence border of the AUC statistics is above 0.5 only for x=0% and 1%. Table A5 (panel G in Appendix) shows that using the ROC-implied optimal threshold (t*=6.5) the analyst would accurately predict only 26 of 76 periods when investment increased and 27 of 30 periods when it declined[45]. This yields 90% probability of detection of negative growth periods with very low probability of detection of upward periods (34%), and 50% total accuracy. The Chi-square test (with Yates correction) is significant at 5% level only for the ROC-implied threshold. However, these results hold only in a given sample. The WES for future expectation about investment is not a reliable predictor out-of-sample.

1.4.3. Future expectations about foreign trade, inflation, and short-term interest rate

Fig.A8 in Appendix presents the ROC curves and their confidence intervals for the future expectations about the exports, imports, trade balance, inflation, and short-term interest rates. The AUC for export expectations points to a low in-sample predictive value, but the confidence borders of the ROC curve are below chance diagonal suggesting that this survey indicator is not reliable out-of-sample. The other variables show both in-sample and out-of-sample predictive power (moderate for the short-term interest rates and trade balance, low for imports and inflation).

Table 13. Other future expectations: AUC statistics and implied WES scores ($g = 0\%$)

Variable	N	AUC	Std. error	Asymptotic confidence intervals for the AUC		Optimal S value in-sample	Ranges of S for which the ROC curve is significant out-of-sample
				Lower 95%	Upper 95%		
Exports	106	0.6716	0.0656	0.5431	0.8001	5.2	None
Imports	106	0.7635	0.0661	0.6340	0.8930	4.7	3.00 - 6.70
Trade balance	106	0.7107	0.0500	0.6135	0.8079	5.5	3.00 - 6.00
Inflation rate	85	0.6620	0.0553	0.5536	0.7704	5.9	4.60 - 6.40
Short-term interest rates	106	0.7800	0.0440	0.6938	0.8662	5.5	1.70 - 8.70

Table 13 above contains results of the ROC curves analysis for these variables and lists ROC-implied optimal thresholds, while Table A9 (panels H-M) in Appendix provides the accuracy statistics for these ROC-optimal thresholds and alternative values as suggested by

[45] Three missed periods of investment decrease include 1990q4, 1992q1, and 2010q4.

Hutson et al. (2014). The Chi-square test for export growth predictions (panel H) would be significant only at t*=5.2 implied by ROC analysis and at alternative t=5, but not for t=7 and 7.2. Also, t*=5.2 yields the highest HSS value, the highest Chi-square statistics, and a favorable combination of the overall accuracy (76%) and probabilities of detection of export increases and declines (85% and 42% respectively). However, these results are not reliable out-of-sample.

Panel I shows that if one predicts future import growth when the WES import expectations scores exceed the ROC- optimal threshold (t*=4.7), she would get the highest overall accuracy (0.86%) and Chi-square statistics. The results for t=5 are very close. The other two alternative thresholds (t=6.7 and 7.3) suggested by Hutson et al. (2014) yield much lower accuracy statistics and insignificant Chi-square test of independence.

Panel K in Table A5 shows that at the ROC-implied optimal threshold (t*=5) the survey future expectations about trade balance correctly predict the periods when trade balance improves (deteriorates) in 76% (60%) of cases respectively, with the overall accuracy averaging to 66%. The HSS statistics is the highest for this t. The Chi-square statistics rejects the null of the hypothesis of independence at 0.1% significance level.

Panel L shows that the optimal threshold t*=5.9 results in sending accurate signals and correct identification of future changes in inflation rate about ⅔ times. It also yields the highest HSS and Chi-square statistics.

Panel M reveals that using the ROC-optimal threshold t*=5.5, the survey expectation about the future short-term interest rate would correctly predict 61% of all changes, identifying 32 of 35 tightening periods, but only 33 of 71 quarters when interest rate declined. A higher threshold would reduce the number of missed periods when policy was loosened, but would increase a number of false alarms of the rate hike, which could harm the stock market and output.

1.5. Conclusion for chapter 1

This chapter contributes to the evaluation and interpretation of the WES expectations, which were not sufficiently covered in the current literature due to the lack of the appropriate methodology to apply to the qualitative survey forecasts lacking data about responses in each category. This research fills the gap by developing such a methodology and applying it to the U.S. data in 1989-2015 with a focus on directional forecasts for GDP, consumption, investment, foreign trade, and inflation and short-term interest rates.

I make several theoretical contributions. First, I demonstrate how to use a simple score system to express results of the qualitative surveys as numerical consensus scores and set an algebraic relation between the consensus scores and well-known balance statistics; this clarifies interpretations of the scoring system. Second, I put forward the method which combines the advantages of the ROC curves analysis with the contingency tables analysis. The former shows whether an indicator can distinguish between the two binary states and suggests the level of a threshold optimal to use in the forecasting rule. The latter helps to compare the accuracy statistics at the ROC-optimal threshold and alternative values suggested by Hutson et al. (2014).

Recent studies on the evaluation of binary forecasts already applied some elements of the ROC curves analysis, such as the AUC statistics, using the AUC>0.5 as the main criterion of an indicator's predictive value. I introduce stricter criteria for the ROC curves analysis to assess the indicators' predictive value in directional forecasting and clarify their application for an in-sample and out-of-sample evaluation. Thus, I argue that a variable has in-sample predictive value if (i.) its AUC>0.5, (ii.) lower 95% confidence bound of its AUC>0.5, and (iii.) the ROC curve lies above the chance diagonal at all possible thresholds. I also suggest that an indicator has out-of-sample predictive value only for the ranges of the threshold at which the ROC curve itself is significant (the ROC curve including its 95% confidence intervals should lie above the chance diagonal). Additionally, I show how to use the ROC curves analysis to find an optimal threshold value t* which minimizes the sum of two types of statistical errors, and demonstrate that its utilization yields the best accuracy results.

In the empirical part, this research completed the first evaluation and interpretation of the WES expectations of U.S. macroeconomic conditions through the analysis of the ROC curves.

The main findings are as follows:

1) The WES has very high directional accuracy for the GDP in the current quarter. The survey users should issue a signal about the GDP decline when the consensus score falls below 3.6. These results are consistent out-of-sample.

2) The WES future expectations about the overall economy have weak in-sample predictive value and are not reliable out-of-sample. Instead, the survey users interested in predicting future output should use the EC indicator as it has high and stable out-of-sample predictive value.

3) The WES present judgement about consumption has high predictive value, but future consumption expectations have low skill. These results are consistent out-of-sample.

4) The WES experts have weak in-sample skill in forecasting direction of change in investment regardless of the horizon which would not be reliable in the out-of-sample exercise. The same is true about future expectations for the exports.

5) The WES future expectations about imports, trade balance and short-term interest rate have medium in-sample and out-of-sample predictive value. I find that a survey user should forecast that the GDP will grow in the current quarter when S>=3.6. The current quarter GDP can be forecast to increase at least at 1%, 2, and 3% per annum when its consensus score exceeds 4.3, 5.6 and 6.7 respectively. Consumption can be forecast to grow in the current quarter when its consensus score S>=3.5. The thresholds 5.0, 5.9, and 6.4 can be used optimally to predict whether the consumption growth will exceed 1%, 2%, and 3% at the annualized rate.

Comparison of the optimal WES scores with the annualized growth rates they predict in the current quarter suggests that experts interpret GDP and consumption growth below 1% as bad, 1%-2% as satisfactory, and above 2% as good. The current quarter growth for the investment can be said to be good when it exceeds 3% per annum.

The methodology proposed in this chapter can be extended to other countries and applied to other qualitative surveys (i.e. the OECD business tendency survey, EU Commission Survey, Institute of Supply Management, etc.), including those in the disciplines other than economics. Besides, the WES expectations with high predictive power can be used as instrumental variables in structural econometric models, or as predictors in other forecasting models.

1.6. References for chapter 1

Abberger, K., Frey, M., Kesina, M. & Stangl, A. (2009). Indicatoren fur die globale Konjuctur, *Ifo Schnelldienst*, 62(34/35), 32-41.

Anderson, O. (1952). The business test of the IFO-Institute for economic research. Munich, and its theoretical model. *Revue de l'Institut International de Statistique*, 20 (1), 1-17.

Ang, A., Bekaert, G., & Wei, M. (2007) Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, 54 (4), 1163-1212.

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

Blöchinger, A. & Lieppold, M. (2006). Economic benefits of powerful credit scoring. *Journal of Banking and Finance*, 30 (3), 851-873.

Carlson, J.A. & Parkin, M. (1975). Inflation Expectations. *Economica,* 42 (165), 123-138.

Carnot, N., Koen, V., & Tissot, B. (2005). *Economic Forecasting*. New York, NY: Palgrave Macmillian.

Claveria, O., Pons, E. & Ramos, R. (2007). Business and consumer expectations and macroeconomic forecasts. *International Journal of Forecasting*, 23 (1), 47-69.

Clements, M. P. (2015). Are professional macroeconomic forecasters able to do better than forecasting trends? *Journal of Money, Credit and Banking*, 47 (2-3), 349-382.

Croushore, D. (2005). Do consumer confidence indexes help forecast consumer spending in real time? *North American Journal of Economics and Finance*, 16 (3), 438-450.

Croushore, D. & Stark, T. (2001). A real-time database for macroeconomists. *Journal of Econometrics*, 105 (1), 111-130.

Cunningham, A. (1997). Quantifying Survey Data. *Bank of England Quarterly Bulletin*, 37 (3), 292-300.

Dorfman, D.D. & Alf E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. *Journal of Mathematical Psychology*, 6 (3), 487-496.

Doswell, D-J., Davies-Jones, R. & Keller, D.L. (1990). On summary measures of skill in rare events forecasting based on contingency tables. *Weather and forecasting*, 5(4), 576-585.

Ericsson, N. R. (2012). *Economics 8378.10 (Economic Forecasting): Course Lecture Notes*. Washington, DC, USA [Unpublished manuscript].

Fisher, R.A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of Royal Statistical Society*, 85 (1), 87-94.

Fisher, R.A. (1950). *Statistical Methods for Research Workers* (11th ed.). New York, NY: Hafner Publishing Company.

Garnitz, J. (2017). *Ifo World Economic Survey – Description and Information*. Munich, Germany: Ifo Institute. Retrived from https://www.cesifo-group.de/dms/ifodoc/docs/facts/survey/WES/Description_WES_2017.pdf

Gorr, W.L. & Schneider, M.J. (2013). Large-change forecast accuracy: Reanalysis of M3-competition data using receiver operating characteristic analysis. International Journal of Forecasting, 29 (2), 274-281.

Granger, C.W.J. & Pesaran, M.H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, 19 (7), 537-560.

Hamella, S. & Haupt, H. (2007). Suitability of WES data for forecasting inflation. In G. Goldrian (Ed.), *Handbook of Survey-Based Business Cycle Analysis* (99-115). Cheltenham, United Kingdom & Nortmapton, MA: Edward Elgar Publishing Limited.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143 (1), 29-36.

Haupt, H. & Waller, S. (2000). Economic analysis and short-term forecasting with qualitative data from the Economic Survey International. In K.H. Oppenlander, G. Poser, and B. Schips (Eds.), *Use of Survey Data for Industry, Research, and Economic Policy: CIRET Conference* (527-547). Avebury, United Kingdom: Aldershot.

Henriksson, R.D. & Merton, R.C. (1981). On market timing and investment performance 2: Statistical procedures for evaluating forecasting skills. *Journal of Business*, 54 (4), 513-533.

Henzel, S. & Wollmershäuser, T. (2005). Quantifying Inflation Expectations with Carlson-Parkin Method: A survey-based determination of the just noticeable difference. *Journal of Business Cycle Measurement and Analysis*, 2005 (3), 321-319.

Henzel, S. & Wollmershäuser, T. (2008). The New Keynesian Phillips curve and the role of expectations: Evidence from the CESifo World Economic Survey. *Economic Modelling*, 25 (5), 811-832.

Hutson, M., Joutz, F. & Stekler, H. (2014) Interpreting and evaluating CESIfo's World Economic Survey directional forecasts. *Economic Modeling*, 38, 6-11.

Krzanowsky, W.J. & Hand, D.J. (2009). *ROC curves for continuous data*. Boca Raton, FL: Chapman & Hall/CRC.

Kudymova, E., Plenk, J. & Wohlrabe, K (2013). Ifo World Economic Survey and the Business Cycle in selected countries. *CESIfo Forum*, 14 (4), 51-57.

Lahiri, K., & Monokroussos, G. (2013). Nowcasting US GDP: The role of ISM business surveys. *International Journal of Forecasting*, 29 (4), 644-658.

Lahiri, K. & Wang, J.G. (2013). Evaluating probability forecasts for GDP declines using alternative methodologies. *International Journal of Forecasting*, 29 (1), 175-190.

Lahiri, K. & Zhao, Y. (2015). Quantifying survey expectations: A critical review and generalization of the Carlson-Parkin method, *International Journal of Forecasting*, 31

(1), 51-62.

Lamont, O.A. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior and Organization*, 48 (3), 265-280.

Leitch, G. & Tanner, J. E. (1991). Economic Forecast Evaluation: Profits Versus the Conventional Error Measures. *The American Economic Review*, 81 (3), 580-590.

Ludvigson, S.C. (2004). Consumer confidence and consumer spending. *The Journal of Economic Perspectives*, 18 (2), 29-50.

Ma, G. & Hall, W.J. (1993). Confidence bands for the receiver operating characteristic curves. *Medical decision making*, 13 (3), 191-197.

Merton, R.C. (1981). On market timing and investment performance 1: An equilibrium theory of value for market forecasts. *Journal of Business*, 54 (3), 363-406.

Muth, J.F. (1961). Rational expectations and the theory of price movements. *Econometrica*, 29 (3), 315-335.

Nardo, M. (2003). The quantification of the qualitative data: a critical assessment. *Journal of Economic Surveys*, 17 (5), 645-668.

Business Tendency Surveys: A Handbook (2003). Paris, France: OECD Publishing. Retrived from https://www.oecd.org/std/leading-indicators/31837055.pdf.

Öller, L-E. (1990). Forecasting the business cycle using survey data. *International Journal of Forecasting*, 6 (4), 453-461.

Ottaviani, M. & Sørensen, P.N. (2006). The strategy of professional forecasting. *Journal of Financial Economics*, 81 (2), 441-466.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York, NY: Oxford University Press.

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

Pesaran, M. H. (1984). Expectations Formations and Macroeconomic Modeling. In P. Malgrange

& P-A. Muet (Eds.), *Contemporary Macroeconomic Modeling* (27-55). Oxford, United Kingdom: Basil Blackwell.

Pesaran, M. H. (1987). *The limits to rational expectations*. Oxford, United Kingdom: Basil Blackwell.

Peterson, W. W., & Birdsall, T. G. (1953). *The Theory of Signal Detectability: Part I. The General Theory. Technical Report 13*. Ann Arbor, MI: University of Michigan, Department of Electrical Engineering, Electronic Defense Group.

Schnader, M.H. & Stekler, H.O. (1990). Evaluating Predictions of Change. *The Journal of Business*, 63 (1), 99-107.

Seiler, C. (2015). On the robustness of balance statistics with respect to nonresponse. *OECD Journal of Business Cycle Measurement and Analysis*, 2014 (2), 1-18.

Schularick, M. & Taylor, A. M. (2012). Credit booms gone bust: monetary policy, leverage cycles and financial crises, 1870–2008. *American Economic Review*, 102(2), 1029–1061.

Soberhart, J. & Keenan, S. (2001). Measuring default accurately. *Credit Risk Special Report, Risk*, 14, 31-33.

Stangl, A. (2007A). World Economic Survey. In G. Goldrian (Ed.), *Handbook of Survey-Based Business Cycle Analysis* (57-65). Cheltenham, United Kingdom & Nortmapton, MA: Edward Elgar Publishing Limited.

Stangl, A. (2007B). Der Index für Konjunkturerwartungen des ZEW und des ifo Instituts, WES, sind für Deutschland identisch. *Ifo Schnelldienst*, 60 (03), 55-56.

Stekler, H.O. (1991). Macroeconomic forecast evaluation techniques. *International Journal of Forecasting*, 7 (3), 375-384.

Stekler, H.O. (1994). Are economic forecasts valuable? *Journal of Forecasting*, 13 (6), 495-505.

Stekler, H.O. & Schnader M.H. (1991). Evaluating predictions of change: an application to inflation forecasts. *Applied Financial Economics*, 1 (3), 135-137.

Stekler, H. & Symington, H. (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006-

2010. *International Journal of Forecasting*, 32 (2), 559-570.

Theil, H. (1952). On the time shape of economic macro variables and the Munich Business test. *Review of the International Statistical Institute*, 20 (2), 105-121.

Theil, H. (1961). *Economic Forecasts and Policy*, (2nd ed.). Amsterdam, Netherlands: North Holland Pub. Co.

Trueman, B. (1994). Analyst Forecasts and Herding Behavior. *The Review of Financial Studies,* 7 (1), 97-124.

Youden, W.J. (1950). Index for rating diagnostic tests. Cancer, 3 (1), 32–35.

Chapter 2: The Good, the Bad, and the Ugly…signals of currency crises: Does signal approach work in ex-ante forecasting of currency crises?

2.1. Introduction

How to construct an Early Warning System (EWS) which would forecast future currency crisis[46] episodes in an accurate and timely manner? First, one should select a list of the *Early Warning Indicators (EWIs)* – macroeconomic variables with different dynamics on the onset of currency crisis episodes, – as supported by theoretical models and empirical studies. Second, one should choose whether to use a *parametric* or a *non-parametric approach* to forecast an imminent crisis period. The former approach estimates the crisis probability using multivariate discrete choice models (i.e. $probit$ or $logit$ regressions) and assesses the predictive value of individual indicators based on their statistical significance. The latter approach monitors a list of EWIs and issues a crisis signal whenever the observed change in a variable passed a certain critical threshold.

Frankel and Rose (1996) were the first to apply parametric approach using the multivariate probit model. Kaminsky, Lizondo, and Reinhart (1998, henceforth - KLR[47]) and companion papers (Kaminsky and Reinhart, 1999; Goldstein, Kaminsky, and Reinhart, 2000) were the pioneers of the signal approach. Subsequent studies improved on either parametric or signal approaches via adding new variables, extending samples, and refining the models. For example, Milesi-Ferretti and Razin (1998) and Esquivel and Larrain (1998)[48] suggested alternative specifications for the probit model. Berg and Patillo (1999) found that the probit model was only slightly better than the KLR model, and both were more informed than a random guess[49]. Edison (2003) revisited the KLR approach

[46] According to the conventions of the international finance literature, this chapter uses terms "currency crisis," "exchange rate crisis," "balance of payment crisis," and "currency crash" interchangeably.

[47] They also provided an extensive survey of the empirical studies of the EWIs up to 1998.

[48] They found that high rates of seignorage, current account imbalances, RER misalignment, low foreign exchange reserves, negative terms of trade shocks, poor growth performance, and a measure of regional contagion have significant predictive power to explain currency crises.

[49] They note that at the 50% cut-off, the KLR model correctly called only 9% of crisis episodes while sending 44% of false alarms. At the 25% threshold, fraction of correctly called episodes increased to 41% at the cost of higher false alarm rate (63%).

and confirmed its value to identify the crisis vulnerabilities, although with such shortcomings as the high false alarm rate and inability to predict the exact timing of a crisis.

To choose which EWS is the most accurate, one should select and apply *the model superiority criteria*. Several measures were suggested in the current literature: the noise-to-signal ratio (NSR), the percentages of the correctly called crises and tranquile periods, the proportion of correctly classified periods in the total number of observations (KLR; Kaminsky and Reinhart, 1999; Edison, 2003), the quadratic probability score (QPS) and its counterparts[50] (Kaminsky, 2000, 2003, 2006), and the total misclassification error (TME) (Comelli, 2014). Candelon, Dumitrescu, and Hurlin (2012) argued that the AUC criterion from the ROC curves analysis should be preferred over a traditional QPS in the context of the EWSs. Candelon, Dumitrescu, and Hurlin (2014) and Commelli (2014) used the AUC statistics to compare different parametric specifications of the EWSs. Drehmann and Juselius (2014) applied the ROC curves analysis and the AUC criterion to evaluate the signals of the banking crises, arguing that they have advantage when one evaluates crisis signals without knowing the policy-makers utility functions. Several other authors used the AUC statistics to assess the value of the parametric EWSs of the currency crises (i.e.; Catao and Milesi-Ferretti, 2013; Caggiano et al., 2014; Frost and Saiki, 2014; Comelli, 2014).

Up to date, neither parametric nor signal approach has established its superiority in forecast accuracy. The benefit of the parametric approach is that it estimates a probability of a future crisis episode. However, it does not help to choose what probability level should be used as a threshold to predict a crisis[51]. Besides, statistical significance of a model depends on the data availability for the crisis episodes with similar features; and a model can be subject to endogeneity concern.

The advantage of the signal approach is that it does not impose parametric structure on the data and is easy to implement. However, it still needs improvement. First, it currently counts signals

[50] These include the log probability score (LPS) and the global squared bias (GSB).

[51] Ideally, such probability value would exceed 50%. However, in practice, an analyst should predict a crisis when it exceeds 20-30%.

as good if they were followed by a crisis episode in any month within *a crisis window*[52]. A preferred EWS would assess the predictive value of a signal with *a fixed forecast horizon*. Second, it marked an indicator as a relevant EWI if it produced the *NSR<1* at least for some values in a grid search. A better EWS would keep only indicators which take *consistently different values in the crisis and non-crisis episodes* (as measured by NSR<1 at the entire range of the EWIs range). Third, the published accuracy results depend on *the choice of the threshold* after reaching which an EWS issues a crisis signal. The current practice relies on the *minimum NSR* criterion and/or overall accuracy ratio. The former *minimizes the number of false alarms at the cost of missing many crisis episodes*; a preferred criterion would consider the trade-off between the false alarms and missed signals and aim to *maximize utility of the forecast user*. The latter is based on the proportion of correctly classified periods in the total number of observations and, therefore, gets *too much credit for the correct identification of tranquile times*, due to the rare nature of crisis episodes. An alternative measure would assess the EWS's ability to predict currency crisis episodes as events excluding the non-crisis episodes from the analysis. Finally, there is mixed evidence about *the out-of-sample performance* of the existing EWS.

This chapter contributes to the signal approach literature via addressing the problems stated above. It poses the following research questions: (i.) which economic indicators can accurately distinguish between the future states (a crisis vs. a non-crisis one) in h=1, 3, 6, 9, 12, 18, and 24m horizons? (ii.) How to choose a critical threshold at which an EWI should issue a crisis? (iii.) How accurately can these EWIs predict incoming currency crisis episodes as events? (iv.) What is the difference between the in-sample and out-of-sample performance of the analyzed EWS? and (v.) What are the benefits of the forecast combinations using the ad-hoc rules?

The methodology of this study builds on the novel application of the ROC curves analysis. To address the first research question, I will compare the predictive value of the indicators within

[52] For example, KLR used the 24-months window, Candalon focused on the 6-months one.

a fixed forecast horizon (h=1, 3, 6, 9, 12, 18, or 24 months). The statistical properties of the ROC curves and their use in evaluating predictive value of the binary classifiers was explained in the first chapter (see section 1.2.4 for details). I assume that each indicator (model) in contest of this chapter has an unobservable true ROC curve – the one that corresponds to the infinitely long time series, if we had such data. However, the sample is limited due to availability of macroeconomic data with monthly frequency, and the ROC curves in this study are only in-sample estimates of their true population counterparts. This chapter evaluates predictive value of the EWIs with the ROC curves analysis using the criteria established earlier in the first chapter (p.15, section 1.2.4): it focuses on the AUC statistics and its confidence intervals to assess the in-sample properties and the significance of the estimated ROC curve itself to measure the out-of-sample power.

The second research question emphasizes that the choice of a threshold value used to signal future crisis episodes affects the accuracy statistics and leads to a trade-off between the two types of errors: a missing signal and a false alarm. Following Candelon et al. (2014) and Jordà (2014), I use the ROC-optimal threshold values above which an indicator should signal an impending crisis. I also show algebraically that the ROC-optimal thresholds minimize the total misclassification cost and establish the relationship between the optimal thresholds in the traditional signal approach (as established by Kaminsky and co-authors) and those proposed here. Then, I compare the accuracy statistics at the alternative threshold values.

To answer the third question, I use a modified ROC curve so that it evaluates the accuracy of an EWS in predicting currency crises as events instead of assessing their power to produce binary classifications of future periods as the crisis or tranquil states. Stekler and Ye (2017) proposed to map the tradeoff between the signal's precision and the false alarm rate. I suggest an alternative variation of a modified ROC curve which focuses on the precision and the share of correctly called crisis episodes. The latter statistics is complementary to the false alarm rate.

To answer the fifth question and test the signal extraction EWS, I divide the data into the training and test sets. I use the training set (1970-1995) to evaluate the in-sample performance of individual EWIs, which I later apply to a test set (1996-2002) to assess the out-of-sample predictive value of the indicators chosen earlier.

Finally, I analyze four forecast combination rules. The first two rules combine information from different indicators at the same forecast horizon, sending a crisis signal when at least one (two) out of four good indicators send a signal. The other two rules combine information coming from the same indicator at several horizons (h=1, 3, 6, and 9m) and can be interpreted as results for the forecast windows as a signal is issued whenever there it is send for at least one (or two) horizons.

The methodology is discussed in Section 2.2. Section 2.3 describes the data. The empirical results are presented in Section 2.4. Section 2.5 concludes, followed by the references in 2.6.

2.2. Methodology: traditional signal approach vs. alternative

Constructing an EWS of currency crisis involves (i.) identifying dates of the crisis episodes; (ii) choice of the indicators; (iii.) choosing a forecasting rule and optimal threshold; and (iv.) evaluating the predictive value of the examined EWIs. While the first two steps are the same regardless the chosen signal approach EWS, there are some differences between the traditional and proposed alternative approaches in the last two stages.

2.2.1. Choice of Early Warning Indicators

The theoretical literature on the exchange rate crises suggests a broad set of such indicators. It groups the models of currency crises in three generations.

The “first generation” theories (i.e. Krugman,1979; Flood and Garber, 1984) believe that episodes of currency crashes stem from inconsistent macroeconomic policies (excessive fiscal and monetary expansions under the fixed exchange rate regime) which lead to a depletion of the foreign reserves and a speculative attack. The budget deficit, growing money supply and forward premium, RER overvaluation, and current account deficit usually precede such attacks.

The “second generation” model (see Obstfeld, 1996) looks at a currency crisis as an optimal choice of a policy-maker who is concerned about the recession, unemployment, or weak trade competitiveness. The weak GDP growth, RER appreciation, deterioration of terms of trade and current account deficit should signal such a crisis.

Theories of the "third generation" (i.e., Chang and Velasco, 2001) explain exchange rate crisis with problems in the banking sector, capital inflows, and financial liberalization, which are warned with increasing interest rates, high debt levels, and bank runs[53].

To limit the scope of this chapter, I examine only indicators available on a monthly frequency[54] and chosen as good predictors of the currency crisis episodes in the signal approach literature by Kaminsky and co-authors (table 14)[55].

Table 14. The Early Warning Indicators and their expected critical shock areas.

#	Indicator	Problem and critical shock
1	Deviation of the RER from the trend	Current account/ Negative
2	Growth rates of the international reserves	Capital account / Negative
3	The excess real money (M1) balances	Monetary policy/ Positive
4	Growth rates of the broad money (M2) to foreign reserves ratio	Capital account/ Positive
5	Growth rates of the exports	Current account/ Negative
6	Growth rates of the index of industrial output	Growth slowdown/ Negative
7	Growth rates of the M2 money multiplier	Overborrowing cycles/ Positive
8	Growth rates of the domestic credit to GDP ratio	Overborrowing cycles/ Positive

All the variables are converted to real terms using CPI and measured as percent values. Two variables – the deviation of the RER from the trend and the excess real M1 balances – require additional calculations[56]. The positive critical shock means that an indicator issues a crisis period when it takes very high values; i.e., very fast increase in excess M1 balances, M2/reserves, M2 multiplier, credit-to-GDP ratio and real interest rate would foresee a future crisis period. The negative critical shock means that an indicator issues a crisis signal if it takes very negative values; this is true for large declines in the RER, exports, international reserves, and industrial output. Data for each indicator are pooled across countries and grouped in 100 percentiles.

[53] When the central bank bailouts financial institutions via money creation, symptoms of the "first generation" currency crisis model will follow suit.

[54] GDP was interpolated to monthly values.

[55] The stock prices and terms of trade were excluded due to the data deficiency.

[56] To measure the deviation of the RER from the trend, I first estimate the RER using quadratic trend ($RER_n = \beta_0 + \beta_1 n + \beta_2 n^2 + e_n$), and then find the deviation between the actual and fitted RER values. The excess real M1 balances are found as the difference between the estimated demand for real M1 balances ($M1real_n = \beta_0 + \beta_1 realGDP + \beta_2 Inflation + \beta_3 n + e_n$) and their actual supply, expressed as a percentage.

2.2.2. Identification of currency crisis episodes

There are no commonly accepted currency crisis dates[57] as literature offers several ways to identify a currency crisis. For example, Kaminsky (2003, 2006) identified crisis episodes using the Exchange Rate Market Pressure Index (EMPI) [58] for each individual country[59] in-sample: a period is marked as a crisis (Y=1) if the EMPI deviates from its mean (μ_{EMP}) by more than 2.5 standard deviations (σ_{EMP}), and as a non-crisis (Y=0) otherwise. Other authors established the ad-hoc rules based on the rate of the currency devaluation or the loss in the foreign reserves. Both approaches are data-dependent: one will likely get more favorable results when the same data are used to identify the crisis episodes and to produce predictions. To provide objective analysis, I refrain from creating an own crisis ID variable, adopting the crisis dates published in Kaminsky (2003, 2006).

2.2.3. Traditional Signal Approach EWS: Kaminsky and co-authors

The traditional signal approach EWS classifies future periods as a crisis ($\hat{Y} = 1$) or non-crisis ($\hat{Y} = 0$) state based on comparison of an EWI value with a chosen threshold. The thresholds $t \in [0, 100]$ are expressed as percentiles. The percentiles are found after pooling all data across countries per each indicator, sorting them from the lowest to the highest values, and grouping into 100 percentiles. The training sample is used to determine the optimal thresholds expressed in terms of percentiles (the alternative optimality criteria will be explained later). Then an analyst finds the growth rate corresponding to the optimal percentile and uses it in the out-of-sample exercise.

When a theory suggests that a positive shock to the variable might cause a crisis, the analyst

[57] Lestano and Jacobs (2007) demonstrated that no single method could identify all the crisis dates as accepted in the IMF chronology for the Asia crisis 1997-1999.

[58] There are several alternative ways to calculate the EMPI. For example, Kaminsky and Reinhart (1999) calculated it as $EMPI_{it} = \frac{\%\Delta e_{it}}{e_{it}} - \frac{\sigma_e}{\sigma_{fxr}} \frac{\%\Delta fxr_{it}}{fxr_{it}}$, where e_{it} is a bilateral nominal exchange rate between an i-country's domestic currency and a country-issuer of the international reserve currency to which a country's currency is pegged, fxr_{it} is a stock of the country "i" foreign exchange reserves, while σ_e and σ_{fxr} are their standard deviations. Thus, the first term stands for the percentage change in the exchange rate, while the second term accounts for the negative percentage changes in the gross international reserves. Thus, the EMPI account not only for the episodes which ended up in the exchange rate adjustment, but also cases of the speculative attack which resulted in the loss of international reserves without devaluation due to the interventions of the country's central bank in a foreign exchange market.

[59] All crisis episodes are identified as single country events.

should use the following forecasting rule[60]:

$\hat{Y} = 1$ (a crisis is forecast for a chosen horizon[61]) if $g_{EWI} \geq t$ [16]

$\hat{Y} = 0$ (a crisis is not forecast for a chosen horizon) if $g_{EWI} < t$

The forecasting rule in [16] will result in two kinds of correct predictions: *the true positives*, which count the number of times when the issued signal correctly classified future period as a crisis state, and are often named the "good signals," and *the true negatives* count correctly identified non-crisis periods, which are of the least interest to the forecast users. Inevitably, such a rule will also produce two types of misclassifications: *the false positives*, which measure the number of tranquil periods misclassified as crisis ones, and are also called "false alarms" or "bad signals," and *the false negative*s, which indicate the number of missed crisis episodes when the forecasting rule failed to issue a signal about the impending crisis. These four numbers can be summarized in a contingency table such as the one presented in the first chapter (see Table 1 in section 1.2.2).

The numbers in the contingency table can be used to calculate a number of accuracy measures. Traditional signal approach focused on the assessment of the NSR. For example, KLR calculated the NSR for each indicator over all values $t \in [80, 90]$[62]:

$$NSR\ (t) = \frac{\text{fraction of tranquile periods incorrectly identified}}{\text{fraction of crisis periods correctly identified}} = \frac{\frac{FP}{FP + TN}}{\frac{TP}{TP + FN}} \qquad [17]$$

KLR pick a threshold as optimal if it minimizes the NSR. They also consider all indicators with $\min NSR < 1$ as the EWIs with strong predictive value[63]. Additionally, they evaluated the probability of a crisis conditional on a signal issued as $\frac{TP}{TP+FP}$ at t^{NSR} (this measure is known in statistics as precision).

The traditional approach outlined in this section has some drawbacks. First, it does not

[60] If the theory reveals that a variable indicates the impending crisis when it takes values from the lower tail of its distribution, the signs in the forecasting rule [16] change to the opposite.

[61] KLR used a 24-months crisis window as a forecast horizon.

[62] Alternatvely, t ∈[10, 20] perecentiles if the lower values indicate the higher probability of a crisis.

[63] See Kaminsky, Lizondo, and Reinhart (1998), p.20, for the author's definitions.

have a fixed forecast horizon: a signal is marked as good if a crisis period occurs in any of the next 24 months after its issuance. Thus, it is not conclusive about the lead time of the assessed indicators and overstates the indicator's predictive value, while understating the number of false alarms. Second, the minimum NSR<1 is a necessary but not a sufficient criterion to choose an EWI: it does not tell if a variable behaves consistently differently in crisis vs. non-crisis episodes. Third, it does not assess the out-of-sample predictive value of an EWI. Then, the percentile variable is defined for each indicator-country individually, forcing the EWS to produce equal number of crisis signals for every country. Finally, the overall accuracy measure is too optimistic as it takes too much credit for the non-crisis periods not preceded by the signal.

This chapter offers an *alternative non-parametric approach to building an EWS of the currency crisis episodes* the *ROC curves*[64]. First, it utilizes the *traditional ROC curves* to evaluate whether an indicator has *binary classification abilities to distinguish crisis periods from tranquil ones*, and then use the *modified ROC curves* to assess the *value of an indicator in forecasting crisis episodes as events*.

2.2.4. Alternative Signal Approach EWS: advantages of the ROC curves analysis

The alternative approach proposed in this chapter uses a forecasting rule similar to the one in [16], with a few notable distinctions. First, instead of focusing on the 24-months crisis window, it uses the fixed n-months ahead forecast horizon (n=1, 3, 6, 9, 12, 18, and 24), which is a stricter way to evaluate predictive power of the indicators. Second, all the indicators are transformed[65] (if necessary) so that they take higher values in crisis. Third, the percentiles for each indicator were found after pooling data for all countries unlike the KLR who used individual distribution for each country[66]. Finally, the predictive value of an indicator is assessed at the entire range of the threshold

[64] See chapter 1 for detailed references.

[65] This requires the change of the sign for the indicators with negative critical areas.

[66] KLR determined percentiles and corresponding growth rates on a country by country basis. Thus, the same percentile value will correspond to different growth rates. This method forces equal number of crisis signals for each country regardless of its fundamentals. This study finds percentile values after polling data for all countries. Therefore, it looks at the overall distribution of growth rates across countries, and searches

values $t \in [0,100]$. This is because a strong EWI should consistently take higher values in crisis periods and lower values in the non-crisis periods. Thus, it should signal better than a random guess regardless of the chosen value t.

For every value $t \in [0,100]$, the forecasting rule will issue a crisis signal when the indicator exceeds the chosen threshold. The correct predictions and misclassifications form a 2x2 contingency matrix as explained above. In the context of this chapter, every contingency table yields unique combinations of the TPR and FPR, which measure the probabilities of sending a crisis signal conditionally on the observing actual non-crisis and crisis periods, respectively:

$$FPR(t) = p(g_{EWI} \geq t | Y = 0) \qquad [18]$$

$$TPR(t) = p(g_{EWI} \geq t | Y = 1)$$[67]

Evaluating the TPR and FPR at various values of a threshold t, one can obtain an ROC curve (see Fig.2), which was discussed in detail in chapter 1. The AUC statistics here measures a probability that an indicator will take on values which are significantly higher in crisis periods than in tranquil ones.

In this study I continue to apply criteria of the in-sample and out-of-sample predictive value which were put forward in chapter 1. An indicator which meets the in-sample predictive value criteria in a training set is said to be able to classify between the crisis and non-crisis periods in sample significantly better than a random guess. Then I use a test set to assess the out-of-sample predictive value for the indicators which demonstrated in-sample predictive value in a training set.

The NSR in the traditional signal approach equals the inverse of the slope of the ROC curve. Therefore, all points on the ROC curve at which NSR<1 will lie above the chance diagonal, while the optimal threshold t^{NSR} will correspond to a point where the ROC curve has the steepest

for the extreme growth rate above which to issue a crisis signal. It yields one growth rate which forecaster will use to predict the future crisis. This is more realistic (a country with worse fundamentals is more likely to have a crisis) and simple (there is a single growth rate to use in out-of-sample test set). The conclusions about indicators are robust to the definition of the percentile variable, although the individual percentiles yielded slightly worse accuracy statistics.

[67] The FPR and TPR correspond to the nominator and denominator of the NSR presented in formula [17].

slope[68]. This observation implies that the NSR<1 criterion is a necessary but not a sufficient condition to conclude that a variable is a good EWI. The ROC curves analysis implies that the threshold t is optimal (t^{ROC}) when it maximizes the vertical distance between the ROC curve and the chance diagonal (MVD), also known as the Youden index (J)[69].

$$\mathrm{MVD(t)} = J(t) = TPR(t) - FPR(t) = TPR(t) + TNR(t) - 1 \qquad [19]$$

One can easily show that maximizing the J-index is equivalent to minimizing the sum of the type I and type II errors (TME=FPR+FNR)[70]. Note, that optimal threshold t^{ROC} implied by the ROC analysis relies on the different criteria when compared to the traditional signal approach. The choice of the optimality criteria affects the entries of the contingency table, and therefore the accuracy ratio (see formula 8 in chapter 1).

2.2.5. Choice of the optimal threshold: traditional vs. alternative approach

This section establishes the relationship between the optimal threshold in the traditional signal approach t^{NSR} and the optimal threshold in the proposed alternative signal approach t^{ROC}.

Proposition 2. Let t^{NSR} be an optimal threshold which minimizes the NSR, and t^{ROC} - an optimal threshold which maximizes the Youden index. Then the following inequalities will hold (see proof in Appendix on p. 115):

$$t^{NSR} \geq t^{ROC} \qquad [20]$$

$$TPR(t^{NSR}) \leq TPR(t^{ROC})$$

$$FPR(t^{NSR}) \leq FPR(t^{ROC})$$

Thus, we have established that the ROC-implied threshold t^{ROC} is less or equal then an optimal threshold suggested by the traditional signal approach t^{NSR}.

[68] The same point will maximize the positive likelihood ratio since $LR{+}= \frac{p(t|Y=1)}{p(t|Y=0)} = \frac{TPR(t)}{FPR(t)} = \frac{1}{NSR(t)}$.

[69] It is also equivalent to the Kolmogorov-Smirnov statistics (KS), which is used to test whether the two distributions are different. In this chapter, we test whether the values if the analyzed leading indicator belong to two different states - crisis and tranquil.

[70] First note that the FPR and FNR represent the errors of type I and II, given the null hypothesis of a non-crisis period. Then rewrite [4] using complementarity of the TPR with the FNR and the TNR and the FPR as $MVD(t) = \big(1 - FNR(t)\big) + \big(1 - FPR(t)\big) - 1 = 1 - \big(FNR(t) + FPR(t)\big) = 1 - TME(t)$

An analyst who picks a low cut-off value t will likely detect many crisis episodes but also issue many false alarms. This choice results in both high TPR and high FPR. Such a forecasting strategy can assist a forecast user who has high costs from missing a crisis but low costs from sending a false alarm. For example, an official authority would prefer to pay attention even to the small signals, because this could allow to implement the appropriate preemptive measures and regulations and to prevent the macroeconomic losses. Therefore, such authority should sacrifice high false alarm rate to minimize the missing crisis episodes.

On the opposite, private investors might lose profits if they issued many false alarms. Thus, they may prefer forecasting crisis periods using higher thresholds. This choice would result in lower TPR and lower FPR, while increasing the rate of correctly called tranquility periods (TNR) and the rate of missed crisis episodes (FNR).

Jordà[71] (2014) demonstrated that the ROC-implied threshold maximizes the forecast-user's utility function even when it is unknown. This implies that choosing a forecasting rule with a different threshold (i.e. t^{NSR}), the forecast-user will lose utility.

2.2.6. Modified ROC curves analysis: evaluating EWI's skill to forecast crises as events

When dealing with rare events such as a currency crisis, one can achieve higher accuracy statistics due to the high number of true negatives (non-crisis periods correctly predicted as such). However, the forecast users are more interested to know how accurately the signal approach predicts crisis episodes. Evaluating an EWI based on its ability to distinguish between crisis and non-crisis periods only confirms that such an indicator can be used to classify a period in two types, but it does not tell how many crisis episodes it forecasted correctly. The need to count the number of crisis episodes as events (regardless of their duration[72]) can be addressed using a modified ROC curve.

[71] He modified the ROC curve replacing the FPR with the TNR on the horizontal axis. This transformation does not change the optimal threshold because the TNR and FPR add to one (TNR=1-FPR).

[72] In this research, no crisis periods lasted longer than 1 month.

Then the following statistics are created[73]:

- True signals, as the number of crisis episodes we successfully predicted in a total number of crisis episodes;
- False signals, as the number of false signals issued when crisis episode did not occur;
- Missed signals, as the number of crisis episodes that occurred without a signal issued.

These true, false and missed signals measures are similar to the TP, FP and FN numbers discussed earlier. In addition to the true and false positive rates, calculated in the same. An analyst who wants to measure the percentage of correct signals among all crisis signals sent will calculate precision in detection of crisis periods:

$$Precision = \frac{TP}{N_{\hat{Y}=1}} \qquad [21]$$

Stekler and Ye (2017) adopted a modified ROC curve known in the statistics as the precision-recall (PR) curve. They argued that it is a proper way to evaluate a leading indicator when the frequency of the event of interest is very low. They, however, deviated from the customary PR curve used in the literature focusing on the relationship between the precision and the false alarm.

This chapter adopts a traditional use of the PR curve as a mapping of the TPR values on the horizontal axis into the precision on the vertical axis. It illustrates the trade-off[74] between the recall (TPR) and precision: to achieve a higher recall of crisis events[75], the analyst needs to choose a lower threshold t. This will issue many false alarms, often lowering precision.

It is not easy to compare the accuracy of two forecasts from the same forecasting rule at two different thresholds as one can have higher precision but lower recall, and another – lower precision but higher recall. To address this issue, I measure a harmonic mean of two values, known

[73] Note, that such classification leaves us without true negatives – because we are not interested in forecasting non-crisis periods.

[74] However, the relationship between the recall and precision is not monotonic. This is because the recall rate is monotonically decreasing in the value of a threshold t. However, there is no monotonic relationship between the value of the threshold t and the precision of a crisis signal. A higher t will increase precision if it adds more correctly identified crisis events than false alarms.

[75] Every crisis month is considered as an event.

in the machine learning literature as G-score:

$$G = \sqrt[2]{Precision * TPR} \qquad [22]$$

The higher G will imply higher overall predictive value of an indicator at chosen threshold.

2.2.7. Forecast combination rules

I analyze four forecast combinations rules using only those indicator-horizon pairs for which I found both in- and out-of-sample predictive value. Rules “At-Least-One-Indicator” (1-I) and “At-Least-Two-Indicators” (2-I) combine information derived from different indicators at the same horizon, as a given crisis can be preceded by different vulnerabilities. Rule 1-I issues a crisis signal when at least one of strong indicators issues a signal. Rule 2-I is stricter, requiring at least two indicators to issue a simultaneous signal. Rules “At-Least-One-Horizon” (1-H) and “At-Least-Two-Horizons” (2-H) combine information obtained from the same indicator at different horizons. Rule 1-H issues a crisis signal when an indicator exceeded a specified threshold at least at one horizon. Rule 2-H required an indicator to signal vulnerability at least at two horizons. Evaluating an EWI based on its ability to distinguish between crisis and non-crisis periods only confirms that such an indicator can be used to classify a period in two types.

2.3. Data on currency crisis episodes and dates

To evaluate results in the signal approach studies, mainly by Kaminsky and Reinhart (1999), this chapter replicates and extends their dataset. The monthly data are collected from the IMF IFS database, complemented with Kaminsky (2003) for missing observations. The growth rates[76] in the M2/reserves ratio[77], M2 multiplier, and domestic credit to GDP ratio, along with the excess demand for M1 balances, have positive critical areas. The growth rates of exports[78], foreign reserves, and industrial production index (IPI)[79], along with the deviation of the RER from the trend

[76] All growth rates are annual, on the month-to-month basis. KLR argued that such filtering makes data comparable across countries, ensure stationarity and well-defined moments, and remove seasonality effects.

[77] M2 was converted into USD.

[78] The value of exports is measured as a “free on board (FOB)”, in millions USD.

[79] When general IPI was not available, it was replaced with the following indexes: Brazil (seasonally adjusted IPI), Peru & Philippines (general manufacturing index), Argentina, Bolivia, Colombia, Ecuador, Venezuela

have negative critical areas. Thus, their signs are reverted.

The training set includes 76 crisis episodes in 20 countries over 1970m1-1995m12. The test set spans over 1996m1-2003m6 in 18 countries, 15 of which experienced 23 crisis episodes[80]. The unconditional probability of a BOP crisis was 1.22% in the training and 1.36% in a test set. Table A6 in Appendix list countries and the currency crisis dates.

2.4. Empirical results

2.4.1. Ability to classify periods as crisis and non-crisis ones

2.4.1.1. Evaluating the EWIs using the ROC curves analysis

Figure 7 below presents the ROC curves for each indicator across horizons. Its upper panel implies that the excess M1 balances, industrial production, domestic credit to GDP ratio, and money multiplier do not pass the in-sample value conditions for the EWI: their ROC curves are not entirely above the chance diagonal, and the AUC values are not significantly greater than 0.5. These indicators are excluded from the further analysis.

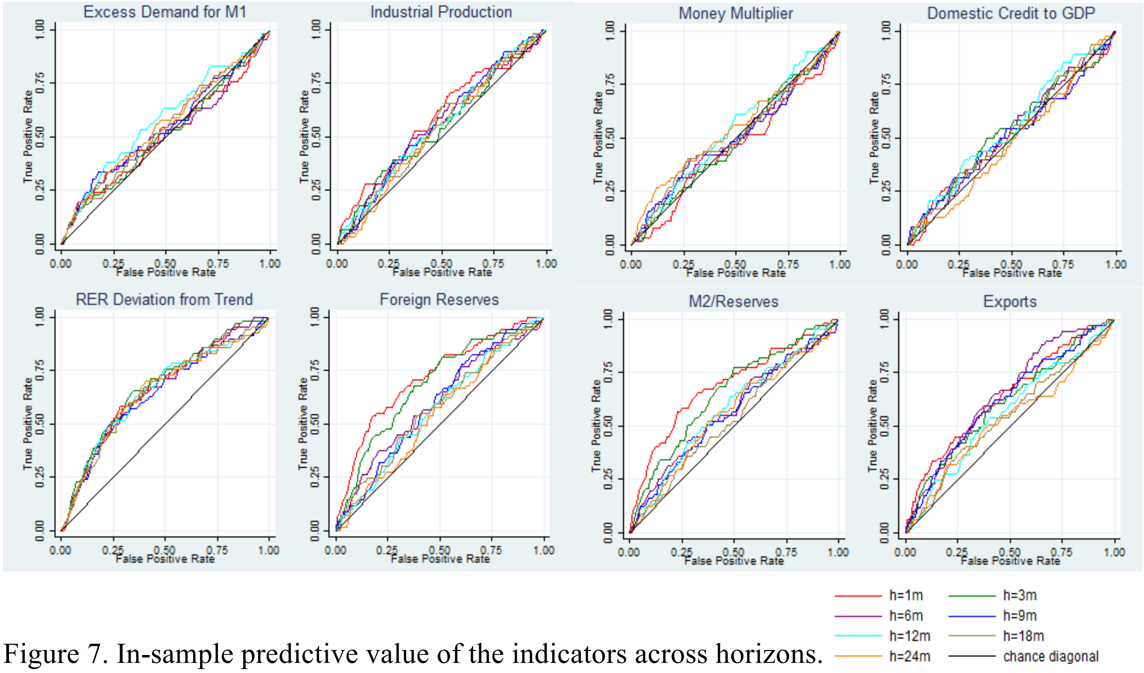


Figure 7. In-sample predictive value of the indicators across horizons.

(crude petroleum production index).

[80] Finland and Spain joined the euro. Six countries did not have a crisis identified in the test set.

The lower panel of Fig.7 shows four indicators for which the ROC curves which were entirely above the chance diagonal with the AUC values significantly above 0.5: 1) the RER overvaluation - at all horizons (AUC=0.65-0.67 with 95% confidence intervals from 0.58 to 0.73); 2) foreign reserves only at h≤18m (AUC=0.57-0.72), with significantly better results at h=1m and 3m; 3) M2/reserves at h<=12m; and 4) exports at h<=9m. Therefore, these indicator-horizons have the in-sample predictive value as they exhibited consistently different behavior in the crisis and non-crisis periods.

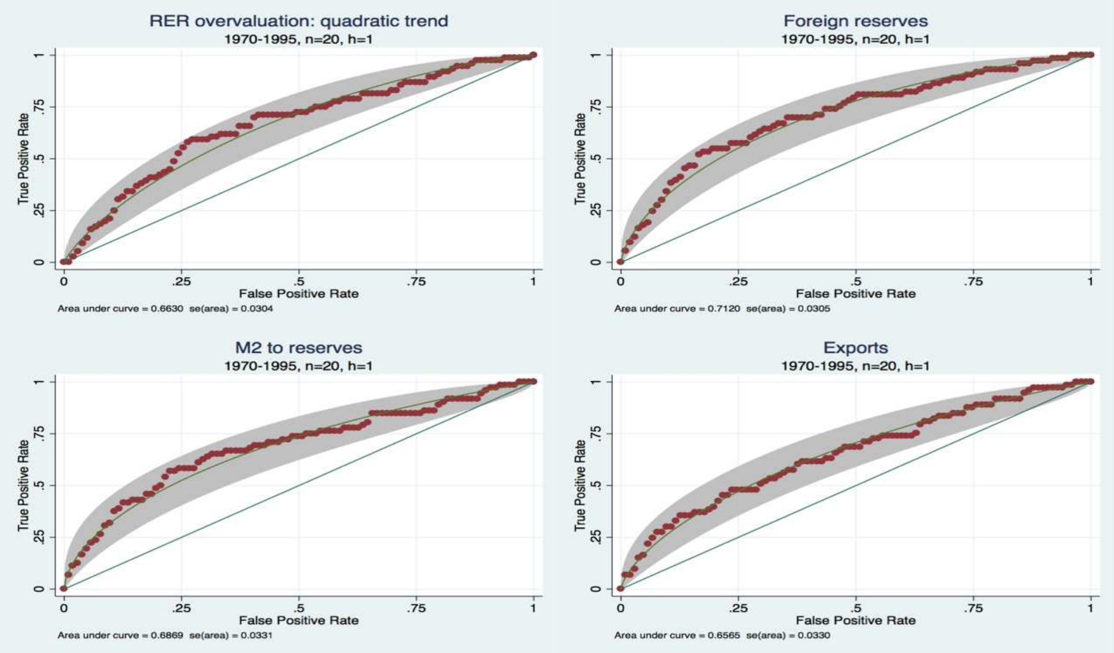


Figure 8. Out-of-sample classifying ability for indicators with in-sample value (h=1m)

Fig. 8 above shows that at h=1m forecast horizon, the ROC curves with their 95% confidence borders for RER overvaluation and decline in foreign reserves were entirely above the chance diagonal for any FPR value, suggesting their out-of-sample power to classify periods into crisis and non-crisis ones. However, for the M2/reserves ratio and exports declines the lower confidence border of the ROC curve was below the chance diagonal in the upper right corner, corresponding to the threshold values $t \in [0, 17]$ percentiles. The out-of-sample forecast ability of these indicators is analyzed via the significance of their ROC curves, presented in Fig.A9, A10, A11 of Appendix for the longer horizons.

Table 15 below provides details on the AUC statistics, lists the threshold ranges significant out-of-sample, and compares the optimal thresholds implied by the ROC and NSR criterions.

Table 15. ROC statistics for the indicators with in-sample predictive value[81]

Indicator	ROC statistics	Fixed forecast horizon, months						
		1	3	6	9	12	18	24
Overvaluation of the RER	AUC (std. error)	0.6630* (0.0304)	0.6720* (0.0310)	0.6580* (0.0302)	0.6343* (0.0331)	0.6371* (0.0347)	0.6442* (0.0335)	0.6322* (0.0353)
	t's significant out-of-sample	0-100	0-100	0-100	15-100	29-100	19-100	25-100
	Optimal threshold: t^{ROC}	73	67	55	75	64	67	57
	Optimal threshold: t^{NSR}	89	87	84	90	87	88	90
Decline in foreign reserves	AUC (std. error)	0.7120* (0.0305)	0.6726* (0.0313)	0.6059* (0.0343)	0.6092* (0.0316)	0.5759* (0.0323)	0.5788* (0.0340)	0.5395 (0.0330)
	t's significant out-of-sample	0-100	0-100	38-100	0 – 83	None	None	None
	Optimal t^{ROC}	83	64	80	53[82]	X	X	X
	Optimal t^{NSR}	90	89	88	80	X	X	X
Growth in M2/reserves ratio	AUC (std. error)	0.6869* (0.0331)	0.6572* (0.0319)	0.6004* (0.0363)	0.5824* (0.0362)	0.5741* (0.0329)	0.5588 (0.0330)	0.5379 (0.0356)
	t's significant out-of-sample	17-100	0-100	53-100	62-100	None	None	None
	Optimal t^{ROC}	78	52	73	64	X	X	X
	Optimal t^{NSR}	90	89	90	87	X	X	X
Decline in exports	AUC (std. error)	0.6565* (0.0330)	0.6174* (0.0335)	0.6335* (0.0309)	0.6048* (0.0335)	0.5606 (0.0342)	0.5396 (0.0358)	0.5184 (0.0369)
	t's significant out-of-sample	17-100	33-100	0-100	29-100	None	None	None
	Optimal t^{ROC}	77	59	57	67	X	X	X
	Optimal t^{NSR}	90	90	86	88	X	X	X

It confirms that the significance of the ROC curve is a stricter condition than the significance of the AUC statistics. For example, in the given sample, the rates of decline of foreign reserves had the AUC values significantly above 0.5 at h=12 and 18m. Thus, they had the in-sample value in classifying the crisis and tranquile periods. However, the lower confidence border of its ROC curve was entirely below the chance diagonal (see Fig. A9 in Appendix). Thus, should the exercise be repeated, decline in exports would not be able to classify two types of periods reliably.

The foreign reserves, M2/reserves, and exports were significant at the wide ranges of the thresholds for $h \leq 9$ months. The RER overvaluation was significant at all horizons, while no other

[81] These results are comparable to KLR who found that the optimal thresholds for the RER overvaluation, foreign reserves, M2/reserves ratio, and decline in exports were at 90, 85, 87, 90 and percentiles respectively (using 24-months window).

[82] This is under a restriction of t>=51 since the unrestricted ROC-t value was equal to 41.

variables had significant ROC curves at 1-year and longer horizons. This, however, is not bad news, because taking preemptive measures too early entails a risk of causing a self-fulfilling crisis and raises costs of crisis preemption. Further analysis will focus on the shorter horizons (h=1…9m ahead), as this period is sufficient to implement anti-crisis measures and eliminates long-run uncertainty about economic developments.

The ROC curves in Fig. 8 above demonstrated that excess M1 balances, industrial production, money multiplier and domestic credit to GDP did not display different behavior in crisis and non-crisis periods when evaluated at the fixed horizons. However, Kaminsky and her co-authors and followers concluded that these indicators are strong because they assessed their predictive value using the 24-months window, counting any signal in this period as a hit regardless of the horizon at which it was sent. To explain the difference in these conclusions, I propose to use *alternative convex ROC hull curves*[83] for the 24-months crisis window.

2.4.1.2. Comparison of traditional and alternative signal approaches using the ROC curves

The alternative convex ROC hull curves for the 24-months forecast window (Figure 9) are constructed as convex combination of the best TPR for each percentile value at seven fixed forecast horizons (h=1, 3, 6, 9, 12, 18, 24 months) using the training sample. Note, that these ROC curves are more conservative than the ones we would get if used the same data and a search algorithm as in Kaminsky et al. (2000). This is due to the following reasons: 1) if one created the convex hull of all 24 ROC curves, one for each h=1, 2…24m ahead, its AUC would be greater; 2) the resulted optimal thresholds t may be different; 3) Kaminsky et al. (2000) added the signals from each horizon, the convex hull takes the strongest signal for each threshold.

The convex hull ROC curves presented below in Fig.9 are sufficient to explain why the two approaches yield different results. When one combines signals from seven forecast horizons, the corresponding ROC curves lie completely above the chance diagonal hiding the fact that the

[83] In past, the convex ROC curves were used to produce a forecast randomly choosing between the two indicators. See Krzanowsky and Hand, p. 145-147.

indicators do not have strong classifying ability at the fixed horizons.

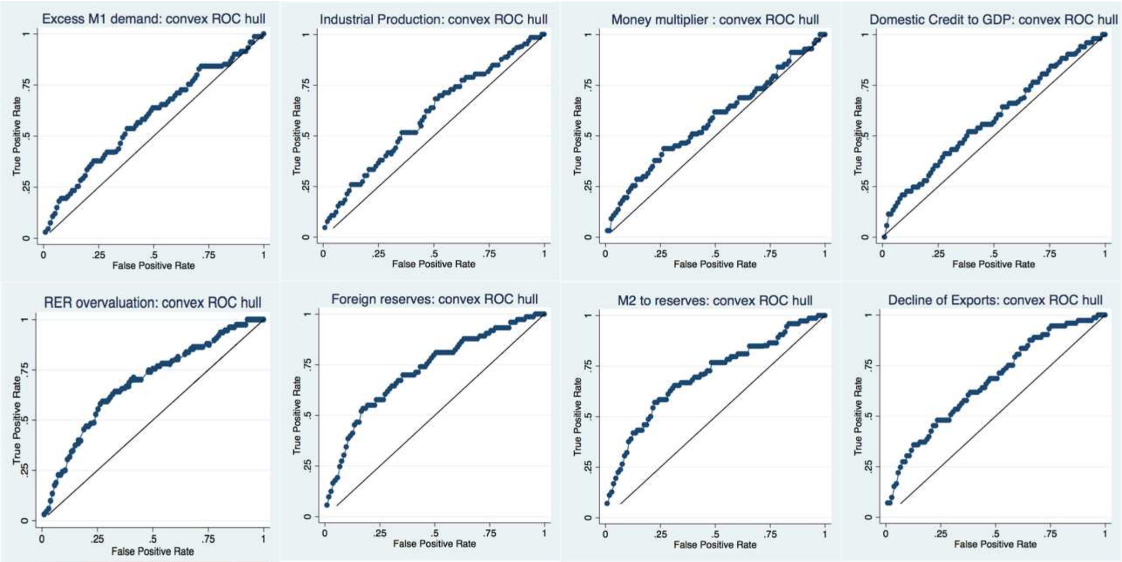


Figure 9. Convex hull ROC curves for the 24-months forecast window horizon

Table 16 below shows that the convex (over horizons) hull ROC curves for all indicators have in-sample predictive value (AUC>0.5) which means that they issued useful signals at least at one horizon during the 24-month forecast window. However, none of these indicators send a signal at h=24m ahead fixed horizon. In fact, many of them signaled only at h=1m ahead.

Table 16. AUC statistics and optimal thresholds for convex hull of each indicator

Convex hull for "any of h=1, 3, 6, 12, 18, or 24 months" horizon	AUC	Std. Err.	95% Conf. Interval		Optimal threshold and horizon		Kaminsky & Reinhart (2000)
					Max J	Min NSR	
RER deviation from the trend	0.6823	0.0541	0.5762	0.7883	73 (1m)	93 (3m)	90 (1-24m)
Foreign reserves	0.7178	0.0024	0.7131	0.7226	83 (1m)	99 (1m)	85 (1-24m)
M2 to reserves	0.6969	0.0024	0.6923	0.7016	78 (1m)	99 (1m)	87 (1-24m)
Decline of exports	0.6673	0.0027	0.6620	0.6725	77 (1m)	97 (1m)	90 (1-24m)
Excess demand for M1	0.5866	0.0027	0.5813	0.5918	63 (12m)	94 (24m)	94 (1-24m)
Industrial production	0.5976	0.0033	0.5911	0.6041	50 (1m)	99 (1m)	89 (1-24m)
Money multiplier	0.5732	0.0026	0.5681	0.5783	74 (18m)	98 (24m)	86 (1-24m)
Domestic credit to GDP	0.5784	0.0028	0.5729	0.5839	73 (12m)	98 (6m)	90 (1-24m)

A preferred method is to evaluate each indicator-horizon pair and forecast crisis episodes as events only using those indicators and threshold ranges that have out-of-sample significance.

2.4.2. In-sample ability to predict crisis episodes as events: modified ROC curves

Table 15 above indicates that the ROC and NSR criteria implied that their respective optimal thresholds are $t^{ROC} = 73$ and $t^{NSR} = 89$ percentiles. Fig. 10 below compares two contingency tables which one would obtain using these threshold values in the forecasting rule. Its left panel shows that a forecast user following the ROC-criterion would issue a signal about the crisis period when the RER reaches 73rd percentile. This would correctly identify 59% of crisis episodes (45 out of 76) and marking 27% of the tranquile periods as crisis ones, issuing 1685 false alarms. As a result, precision of the signals sent would reach only 2.6%.

t^{ROC}		Forecasts		Total
$= 73$		$\hat{Y} = 1$	$\hat{Y} = 0$	
Actuals	Y=1	TP=45	FN=31	76
	Y=0	FP=1685	TN=4459	6144
Total		1730	4490	6220

t^{NSR}		Forecasts		Total
$= 89$		$\hat{Y} = 1$	$\hat{Y} = 0$	
Actuals	Y=1	TP=23	FN=53	76
	Y=0	FP=712	TN=5432	6144
Total		735	5485	6220

Figure 10. Contingency tables for the RER overvaluation at two alternative thresholds

The right panel of Fig. 10 indicates that a forecast user following the NSR criterion would issue a signal about the crisis period when the RER reaches 89th percentile. This would correctly identify only 30% of crisis episodes (23 from 76) and mark 12% of non-crisis periods as crisis ones, issuing 735 false alarms. Precision of the signals sent would slightly increase to 3.1%, although at the cost of missing 53 crisis episodes, compared to only 31 when the ROC criterion was used.

One may be tempted to compare the overall accuracy statistics, which is higher at the NSR-implied threshold (88%) than at the ROC-implied threshold (72%). However, when the frequency of the event of interest is very low, the accuracy statistics is almost equal to the share of correctly identified traquile periods (the TNR), which reached 73% and 88% for the two alternative threshold values. A forecast user who places more cost on the false alarms could prefer to choose a threshold above the t^{ROC}, but below t^{NSR}. Table 17 below presents the entries of the contingency tables one would obtain using different threshold values and the corresponding accuracy measures in the compact column view for the RER overvaluation at h=1m fixed horizon.

Table 17. Accuracy statistics for the RER overvaluation (h=1m)[84]

T	TP	TN	FN	FP	TPR	TNR	FPR	J	NSR	ACC	Prec	G
73	45	4459	31	1685	0.59	0.73	0.27	0.32	0.46	0.72	0.026	0.12
77	37	4701	39	1443	0.49	0.77	0.23	0.25	0.48	0.76	0.025	0.11
80	32	4883	44	1261	0.42	0.79	0.21	0.22	0.49	0.79	0.025	0.10
83	30	5065	46	1079	0.39	0.82	0.18	0.22	0.44	0.82	0.027	0.10
86	26	5248	50	896	0.34	0.85	0.15	0.20	0.43	0.85	0.028	0.10
89	23	5432	53	712	0.30	0.88	0.12	0.19	0.38	0.88	0.031	0.10

One can see that a gradual increase of the threshold t used in the forecasting rule leads to a decline of the total number of crisis signals issued, which implies lower false alarm rate at the cost of lower number of correctly identified crisis episodes. This leads to an increase in the rate of correctly identified non-crisis periods (TNR), which prevail the sample, and therefore increases the accuracy ratio. The dependence of the precision, NSR, and the J-index on the t value is a non-linear.

Table 18 below presents the contingency tables and resulted accuracy statistics for the 16 combinations of 4 indicators and 4 fixed forecast horizons with the proved predictive value. Look, for example, at the deviation of the RER from the time trend. The ROC analysis and the maximum J-index imply that we would have issued a signal about the crisis in the next period whenever the RER reaches 73rd percentile. In this case, we would correctly predict 45 crisis episodes and would miss 31 crisis episodes issuing 1685 false alarms.

The minimum NSR implies that we would issue a crisis signal only when RER reaches its 95th percentile. In that case, we would correctly predict only 12 crisis episodes and would miss 64 crisis episodes issuing only 359 false alarms. However, the precision would be higher when one uses the NSR-optimal threshold (3.1%) compared to 2.6% when one uses the ROC-optimal threshold: from all the crisis signals sent, only 3.1% (2.6%) of them would be correct, and the rest 97% of signal would be false.

[84] Table A8 in Appendix presents the accuracy statistics for the RER at h=3, 6, and 9m ahead at the wide variety of thresholds bounded by the ROC and NSR optimal values.

Table 18. Accuracy statistics for each individual indicator-horizon pair (training sample)

EWI	H	t[85]	TP	TN	FN	FP	TPR	TNR	FPR	J	NSR	ACC	Prec	G
RER overvaluation	1	73	45	4459	31	1685	0.59	0.73	0.27	0.32	0.46	0.72	0.026	0.12
		89	23	5432	53	712	0.30	0.88	0.12	0.19	0.38	0.88	0.031	0.10
	3	67	48	4086	27	2019	0.64	0.67	0.33	0.31	0.52	0.67	0.023	0.12
		87	25	5300	50	805	0.33	0.87	0.13	0.20	0.40	0.86	0.030	0.10
	6	55	53	3337	21	2709	0.72	0.55	0.45	0.27	0.63	0.55	0.019	0.12
		84	27	5104	47	942	0.36	0.84	0.16	0.21	0.43	0.84	0.028	0.10
	9	75	38	4596	36	1450	0.51	0.76	0.24	0.27	0.47	0.76	0.026	0.12
		90	17	5472	57	574	0.23	0.91	0.10	0.14	0.41	0.90	0.029	0.08
Foreign reserves	1	83	39	4847	34	1034	0.53	0.82	0.18	0.36	0.33	0.82	0.036	0.14
		90	28	5252	45	629	0.38	0.89	0.11	0.28	0.28	0.89	0.043	0.13
	3	64	49	3700	24	2141	0.67	0.63	0.37	0.30	0.55	0.63	0.022	0.12
		89	24	5151	49	690	0.33	0.88	0.12	0.21	0.36	0.88	0.034	0.11
	6	80	29	4580	43	1202	0.40	0.79	0.21	0.19	0.52	0.79	0.024	0.10
		88	21	5031	51	751	0.29	0.87	0.13	0.16	0.45	0.86	0.027	0.09
	9	53	47	3029	25	2753	0.65	0.52	0.48	0.18	0.73	0.53	0.017	0.11
		80	24	4578	48	1204	0.33	0.79	0.21	0.13	0.62	0.79	0.020	0.08
IM2/reserves	1	78	41	4427	31	1293	0.57	0.77	0.23	0.34	0.40	0.77	0.031	0.13
		90	27	5108	45	612	0.38	0.89	0.11	0.27	0.29	0.89	0.042	0.13
	3	52	56	2924	17	2756	0.77	0.51	0.49	0.28	0.63	0.52	0.020	0.12
		89	21	5006	52	674	0.29	0.88	0.12	0.17	0.41	0.87	0.030	0.09
	6	73	34	4058	38	1563	0.47	0.72	0.28	0.19	0.59	0.72	0.021	0.10
		89	17	4946	55	675	0.24	0.88	0.12	0.12	0.51	0.87	0.025	0.08
	9	64	39	3559	33	2062	0.54	0.63	0.37	0.17	0.69	0.63	0.019	0.10
		87	18	4839	54	782	0.25	0.86	0.14	0.11	0.55	0.85	0.023	0.08
Exports	1	77	35	4422	38	1378	0.48	0.76	0.24	0.24	0.50	0.76	0.025	0.11
		90	22	5175	51	625	0.30	0.89	0.11	0.19	0.36	0.88	0.034	0.10
	3	59	45	3346	28	2414	0.62	0.58	0.42	0.20	0.68	0.58	0.018	0.11
		90	19	5133	54	627	0.26	0.89	0.11	0.15	0.42	0.88	0.029	0.09
	6	57	46	3184	26	2517	0.64	0.56	0.44	0.20	0.69	0.56	0.018	0.11
		86	20	4842	52	859	0.28	0.85	0.15	0.13	0.54	0.84	0.023	0.08
	9	67	39	3733	33	1968	0.54	0.65	0.35	0.20	0.64	0.65	0.019	0.10
		88	20	4950	52	751	0.28	0.87	0.13	0.15	0.47	0.86	0.026	0.09

This is a feature common to predicting a subject which occurs rarely. Low recall rate can be unpleasant for the forecasters, but should not deter them from the objective to maximize the utility function of the forecast user. And the forecast user would prefer to avoid the costs of the missed crisis events, even at the 0.5% lower precision rate. Also, note that choosing a very high threshold maximizes the number of correctly predicted non-crisis episodes, which are not useful to

[85] The 1st and 2nd rows at each indicator-horizon pair indicates the ROC (NSR) optimal thresholds.

the forecast users. The NSR-optimal threshold produced significantly lower number of correctly predicted crisis periods (17-28) than the ROC criteria (29-56). At the same time, the ROC-optimal threshold never yielded a lower G-score: it was higher than G under the NSR-optimal t for 15 out of 16 indicator-horizon pairs, and equal in one occurrence. Among all indicator-horizon pairs, one would predict the most number of crisis periods if used the M2/reserves ratio at h=3m fixed horizon. In this case the forecast would correctly identify 56 crisis periods and miss 17 crisis periods[86].

Fig. 11 below presents the modified ROC curves - the PR curves - for each indicator with good classifying properties across the selected fixed forecast horizons.

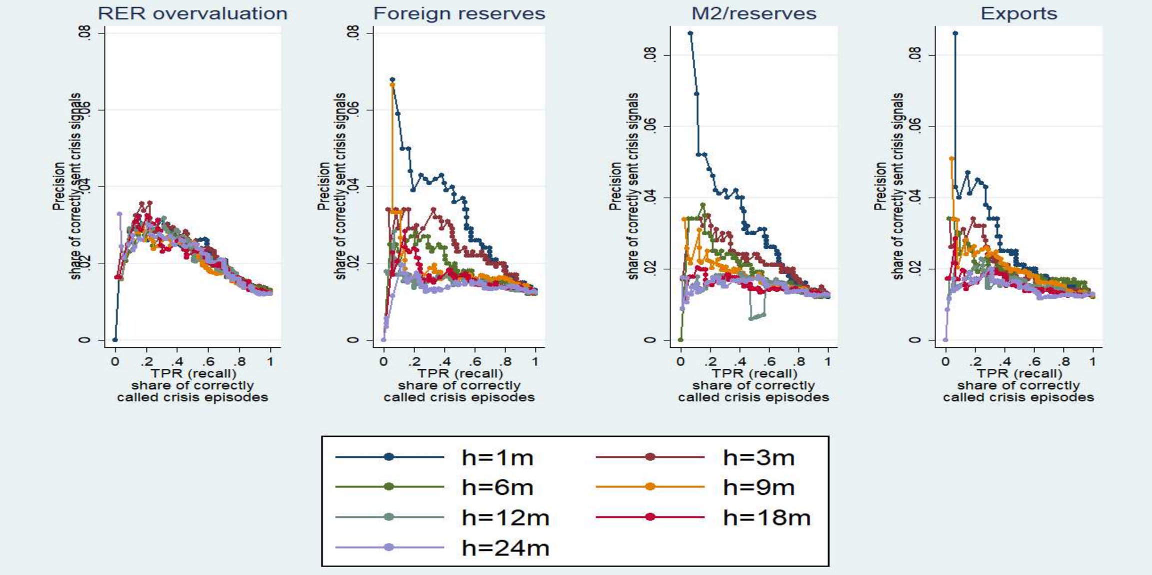


Figure 11. PR curves for each indicator with good classifying properties across horizons

Overall, these PR curves draw attention to the fact that only a small portion of all crisis signals is correctly sent. The average precision across all four indicators and horizons was around 2%, with the maximum precision of 8.6% achieved by the M2/reserves and exports at h=1m horizon. The RER overvaluation, produced almost identical PR curves across the forecast horizons. The other three indicators show that the crisis signals are more accurate at shorter horizons, with the highest accuracy 1 month before the crisis.

[86] There is no data for 3 crisis episodes for this indicator-horizon pair.

Another way to analyze the PR curves is by looking at each horizon across all 4 indicators. The focus is only at short fixed forecast horizons (h=1, 3, 6 and 9m).

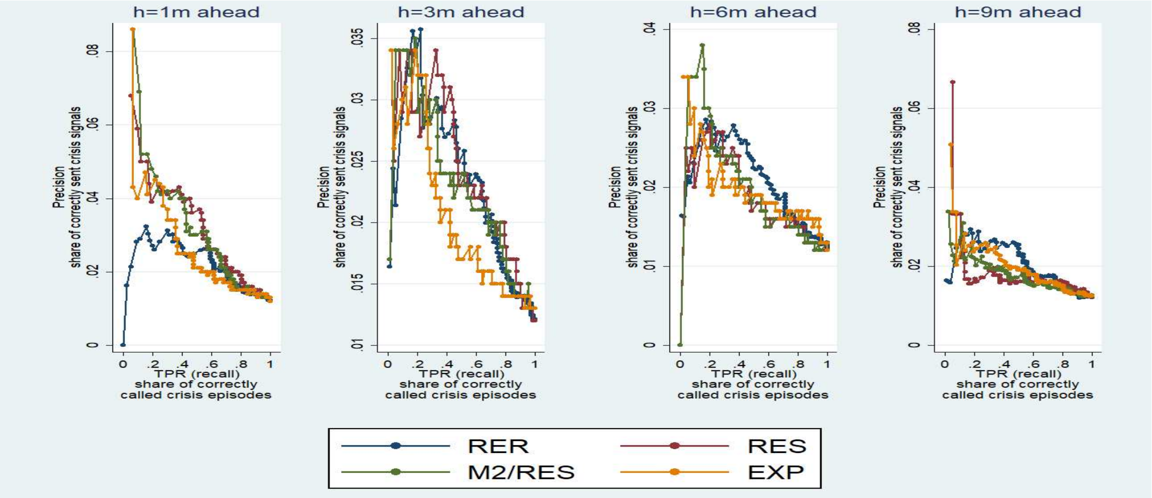


Figure 12. PR curves across indicators with good classifying properties (h=1-9m)

Figure 12 above shows that at h=1m and 3m horizons the decline in foreign reserves and increase in the M2/reserves ratio produced higher precision almost at all recall values. When forecasting a crisis event at h=6m horizon, the RER overvaluation yielded higher precision at smaller thresholds (when the TPR values are high), while the increase in M2/reserves ratio. When the crisis event is forecast 9m ahead, the decline in foreign reserves was more precise at the extreme threshold values on both ends, while the RER overvaluation – mostly in the middle range. In general, increasing the threshold raised the share of correctly called crisis signals for all indicators except the RER overvaluation. The maximum precision achieved in this sample was 8.6% at h=1m.

The in-sample precision achieved in this study seems too low. This is due to a low observed unconditional probability of the crisis events. One can show that the in-sample share of all crisis signals sent correctly is bounded between 1.22 and 55%[87]. Table A7 in the Appendix presents the formula and the resulted non-linear correspondence between the NSR and precision of correctly sent crisis signal in the given sample. For example, precision of 8.6% is achieved when the NSR

[87] One can derive the relationship between the unconditional probabilities of the crisis and non-crisis periods, the NSR, and the precision using the Bayes formula (see, i.e. Krzanovwski, p.10)

fells to 0.13. Only an indicator with the NSR<=0.12 would be able to achieve a higher precision, which would come at the cost of many missed crisis events.

2.4.3. Out-of-sample ability to predict crisis episodes as events: modified ROC curves

Table 19 below presents the accuracy statistics for the test sample (1996-2002).

Table 19. Accuracy statistics for each individual indicator-horizon pair (test sample)

EWI	h	t	TP	TN	FN	FP	TPR	TNR	FPR	J	NSR	ACC	Prec	G
RER overvaluation	1	73	21	182	2	1415	0.91	0.11	0.89	0.03	0.97	0.13	0.015	0.12
		89	18	488	5	1109	0.78	0.31	0.69	0.09	0.89	0.31	0.016	0.11
	3	67	22	142	1	1455	0.96	0.09	0.91	0.05	0.95	0.10	0.015	0.12
		87	19	440	4	1157	0.83	0.28	0.72	0.10	0.88	0.28	0.016	0.12
	6	55	22	97	1	1500	0.96	0.06	0.94	0.02	0.98	0.07	0.014	0.12
		84	19	360	4	1237	0.83	0.23	0.77	0.05	0.94	0.23	0.015	0.11
	9	75	23	200	0	1397	1.00	0.13	0.87	0.13	0.87	0.14	0.016	0.13
		90	19	492	4	1105	0.83	0.31	0.69	0.13	0.84	0.32	0.017	0.12
Foreign reserves	1	83	9	1380	14	217	0.39	0.86	0.14	0.26	0.35	0.86	0.040	0.12
		90	3	1488	20	109	0.13	0.93	0.07	0.06	0.52	0.92	0.027	0.06
	3	64	15	844	8	753	0.65	0.53	0.47	0.18	0.72	0.53	0.020	0.11
		89	3	1439	20	158	0.13	0.90	0.10	0.03	0.76	0.89	0.019	0.05
	6	80	5	1248	18	349	0.22	0.78	0.22	0.00	1.01	0.77	0.014	0.06
		88	3	1383	20	214	0.13	0.87	0.13	0.00	1.03	0.86	0.014	0.04
	9	53	12	501	11	1096	0.52	0.31	0.69	-0.16	1.32	0.32	0.011	0.08
		80	5	1203	18	394	0.22	0.75	0.25	-0.03	1.13	0.75	0.013	0.05
IM2/reserves	1	78	6	1390	17	185	0.26	0.88	0.12	0.14	0.45	0.87	0.031	0.09
		90	1	1525	22	50	0.04	0.97	0.03	0.01	0.73	0.95	0.020	0.03
	3	52	14	802	9	773	0.61	0.51	0.49	0.12	0.81	0.51	0.018	0.10
		89	1	1478	22	97	0.04	0.94	0.06	-0.02	1.42	0.93	0.010	0.02
	6	73	4	1234	19	341	0.17	0.78	0.22	-0.04	1.24	0.77	0.012	0.04
		89	0	1424	23	151	0.00	0.90	0.10	-0.10	-	0.89	0.000	0.00
	9	64	9	1028	14	547	0.39	0.65	0.35	0.04	0.89	0.65	0.016	0.08
		87	0	1363	23	212	0.00	0.87	0.13	-0.13	-	0.85	0.000	0.00
Exports	1	77	9	1116	14	481	0.39	0.70	0.30	0.09	0.77	0.69	0.018	0.08
		90	5	1389	18	208	0.22	0.87	0.13	0.09	0.60	0.86	0.023	0.07
	3	59	16	644	7	953	0.70	0.40	0.60	0.10	0.86	0.41	0.017	0.11
		90	5	1355	18	242	0.22	0.85	0.15	0.07	0.70	0.84	0.020	0.07
	6	57	17	554	6	1043	0.74	0.35	0.65	0.09	0.88	0.35	0.016	0.11
		86	5	1245	18	352	0.22	0.78	0.22	0.00	1.01	0.77	0.014	0.06
	9	67	12	742	11	855	0.52	0.46	0.54	-0.01	1.03	0.47	0.014	0.08
		88	3	1223	20	374	0.13	0.77	0.23	-0.10	1.80	0.76	0.008	0.03

With the ROC-optimal thresholds, the RER overvaluation achieved the highest accuracy result correctly predicting all 23 crisis episodes in the test sample (at h=9m fixed horizon). Exports came in second, with 17 correctly predicted crisis episodes at h=6m fixed horizon. Foreign reserves and M2/reserves correctly identified 15 and 14 crisis episodes respectively at h=3m horizon. It is interesting to note, that the accuracy results are better 3 months before the crisis than just one month ahead. Using the NSR optimality criterion, one would identify much smaller number of crisis episodes (18-19 for RER overvaluation, 3-5 for decline in foreign reserves, 0-1 for M2/reserves, and 3-5 for decline in exports). Precision is still low, from 1.6 to 3.1.%.

2.4.4. Forecast combinations

The currency crisis come in different varieties, originating from different vulnerabilities and through different propagation mechanisms. Therefore, combining information sent from different indicators at the same horizon should improve the forecast accuracy.

Table 20. Accuracy statistics for combinations of 4 indicators per horizon (training sample)

H	t[88]	TP	TN	FN	FP	TPR	TNR	FPR	J	NSR	Acc	Prec	G
Rule 1-I: “At-Least-One-Indicator”													
1	ROC	69	2463	7	3681	0.91	0.40	0.60	0.31	0.66	0.41	0.018	0.13
	NSR	57	3968	19	2176	0.75	0.65	0.35	0.40	0.47	0.65	0.026	0.14
3	ROC	74	1119	1	4986	0.99	0.18	0.82	0.17	0.83	0.19	0.015	0.12
	NSR	50	3830	25	2275	0.67	0.63	0.37	0.29	0.56	0.63	0.022	0.12
6	ROC	68	1233	6	4813	0.92	0.20	0.80	0.12	0.87	0.21	0.014	0.11
	NSR	50	3515	24	2531	0.68	0.58	0.42	0.26	0.62	0.58	0.019	0.11
9	ROC	67	1528	6	4459	0.92	0.26	0.74	0.17	0.81	0.26	0.015	0.12
	NSR	45	3500	28	2487	0.62	0.58	0.42	0.20	0.67	0.58	0.018	0.10
Rule 2-I: “At-Least-Two-Indicators”													
1	ROC	51	4426	25	1718	0.67	0.72	0.28	0.39	0.42	0.72	0.029	0.14
	NSR	30	5216	46	928	0.39	0.85	0.15	0.24	0.38	0.84	0.031	0.11
3	ROC	65	2923	10	3182	0.87	0.48	0.52	0.35	0.60	0.48	0.020	0.13
	NSR	31	5098	44	1007	0.41	0.84	0.16	0.25	0.40	0.83	0.030	0.11
6	ROC	55	3503	19	2543	0.74	0.58	0.42	0.32	0.57	0.58	0.021	0.13
	NSR	25	4971	49	1075	0.34	0.82	0.18	0.16	0.53	0.82	0.023	0.09
9	ROC	53	3092	20	2895	0.73	0.52	0.48	0.24	0.67	0.52	0.018	0.11
	NSR	24	4769	49	1218	0.33	0.80	0.20	0.13	0.62	0.79	0.019	0.08

[88] This and following tables use the thresholds optimal at ROC and NSR criteria as listed in Table 17.

Table 20 above presents results for the forecast combinations for rules 1-I and 2-I, which combine information from the four indicators with the in- and out-of-sample predictive value, all for the test sample. The top panel shows that a rule 1-I would correctly predict 67-74 crisis episodes (with a precision of the crisis signals 1.4-1.8%) when using the ROC-t values, and only 45-57 crisis episodes (with a precision of the crisis signals 1.8-2.6%). For example, if one issued a crisis signal every time when at least one indicator warned about an oncoming crisis period 1 month ahead, there would be 69 (57) correctly identified crisis episodes; and the precision would equal 1.8% (2.6%) for the ROC (NSR) t-values respectively. The ROC-threshold could achieve 99% recall and warn about 74 crisis episodes if combined information from all signals sent 3 months in advance. The NSR threshold yielded the highest recall (75%) 1 month before a crisis occurs. Rule 2-I reduced the number of false alarms, increasing the precision of a signal to 1.8–2.9% (1.9-3.1%) for the ROC (NSR) thresholds respectively. However, the number of correctly identified crisis periods reduced to 51-65 (24-31) for the ROC (NSR) optimal t values. The ROC-t yielded the G scores which were at least as good as those from the NSR-t at all horizon, with the exception of h=1m ahead.

The same crisis could be signaled during the 9m window only once (i.e. 1, 3, 6, or 9m ahead), or several times (if a signal was persistent). Table 21 below presents the accuracy of the forecasts obtained when information from the same indicator is combined at different horizons. It shows that deviation of the RER from the trend alone could correctly predict 57 (36) crisis periods if one used a rule 1-H (when a RER issued a signal at least at one of 4 forecast horizons) with the ROC (NSR) thresholds. Limiting signals to the case when an indicator issued warning at least at 2 of 4 forecast horizons and using the rule 2-I, the RER overvaluation would help predicting 53 (25) crisis periods with ROC (NSR) optimal values respectively. The precision of a crisis signal would equal 1.9 (2.3%) and 2.8% (2.6%) when the rule 1-H (2-H) was used with the ROC and NSR optimal thresholds respectively. Rule 1-H would predict the highest number of crisis periods when M2/reserves ratio used with the ROC threshold or exports with the NSR threshold. Rule 2-H would favor using exports alone, as it predicted no worse (better) with the ROC (NSR) thresholds.

Table 21. Accuracy statistics for combinations of 4 horizons per indicator (training sample)

Indicator	t	TP	TN	FN	FP	TPR	TNR	FPR	J	NSR	Acc	Prec	G
Rule 1-H: "At-Least-One-Horizon"													
RER	ROC	57	3141	19	3003	0.75	0.51	0.49	0.26	0.65	0.51	0.019	0.12
	NSR	36	4911	40	1233	0.47	0.80	0.20	0.27	0.42	0.80	0.028	0.12
Foreign reserves	ROC	59	2411	14	3480	0.81	0.41	0.59	0.22	0.73	0.41	0.017	0.12
	NSR	38	4086	35	1805	0.52	0.69	0.31	0.21	0.59	0.69	0.021	0.10
M2 to reserves	ROC	65	2110	8	3708	0.89	0.36	0.64	0.25	0.72	0.37	0.017	0.12
	NSR	39	4152	35	1666	0.53	0.71	0.29	0.24	0.54	0.71	0.023	0.11
Exports	ROC	64	1871	9	3943	0.88	0.32	0.68	0.20	0.77	0.33	0.016	0.12
	NSR	45	1871	28	1947	0.62	0.49	0.51	0.11	0.83	0.49	0.023	0.12
Rule 2-H: "At-Least-Two-Horizons"													
RER	ROC	53	3867	23	2277	0.70	0.63	0.37	0.33	0.53	0.63	0.023	0.13
	NSR	25	5211	51	933	0.33	0.85	0.15	0.18	0.46	0.84	0.026	0.09
Foreign reserves	ROC	51	3730	22	2161	0.70	0.63	0.37	0.33	0.53	0.63	0.023	0.13
	NSR	29	4878	44	1013	0.40	0.83	0.17	0.23	0.43	0.82	0.028	0.11
M2 to reserves	ROC	50	3351	23	2467	0.68	0.58	0.42	0.26	0.62	0.58	0.020	0.12
	NSR	28	4824	46	994	0.38	0.83	0.17	0.21	0.45	0.82	0.027	0.10
Exports	ROC	53	3155	20	2659	0.73	0.54	0.46	0.27	0.63	0.54	0.020	0.12
	NSR	27	2935	46	883	0.37	0.77	0.23	0.14	0.63	0.76	0.030	0.10

Overall, each indicator used individually in rule 1-H (2-H) would correctly point to 57-65 (50-53) crisis months when signals were issued with the ROC-optimal threshold values, and only 36-45 (25-28) crisis months with the NSR-optimal thresholds.

2.5. Conclusion for chapter 2

This chapter contributes to the literature on the design and evaluation of the signal approach to construct an Early Warning System (EWS) of the currency crisis episodes. It re-examines predictive value of the eight Early Earning Indicators of currency crisis which were found to have predictive value by Kaminsky and Reinhart (1999). It uses monthly data from 20 countries over a span of 26 years (1970-1995). All the indicators are calculated as percentages, and then sorted into 1-100 percentiles. These percentiles are used in the forecasting rule to predict crisis. I use the analysis of the ROC curves to test whether an indicator has distinctly different behavior in times of crisis and tranquility.

Then I employ the in-sample and out-of-sample criteria of predictive value as established in the first chapter to determine a list of indicators which take on significantly different values in two regimes (crisis vs. tranquility), and therefore can be used as classifiers to distinguish between the two states. Only the deviation of the RER from a trend, the foreign reserves, the ratio of broad money M2 to reserves, and decline in exports have demonstrated both in-sample and out-of-sample predictive value[89].

I also employed a novel way to construct the convex hull ROC curves to explain that the previous literature used more liberal criteria and therefore found more indicators had predictive value. Then, I explained how to choose an optimal threshold using the ROC-implied criteria, and how this choice differs from the minimizing noise-to-signal ratio previously used in the literature. In general, thresholds chosen in accordance with the maximum J-index in the ROC curves analysis result in the higher rate of correctly called crisis episodes.

I also employed the modified ROC curves to show the relationship between the precision of sent signals and recall of crisis episodes. Then, I analyzed the accuracy statistics to illustrate how the accuracy statistics, in particular, the tradeoff between the recall and precision depends on choice of the threshold used in the forecasting rule. Results show that although the identified EWIs do perform better than a random guess, they have very weak predictive value. In general, they identify no more 2/3 of crisis episodes, generating hundreds false alarms. Precision of the signals sent does not exceed 8.5%. It means that for every correctly sent crisis signal there are a dozen of false ones.

Finally, I exploited the benefits of forecast combinations using several ad-hoc rules and found that they help one to improve the accuracy of results, including both recall and precision.

To conclude, the alternative method to evaluate the leading indicators of currency crisis yields more conservative conclusions because it evaluates signals at the fixed forecast horizons instead of using the 24 months forecast window.

[89] The RER had in-sample and out-of-sample predictive value at all horizons, while the other three indicators were valuable only at h=6m and shorter.

2.6. References for chapter 2

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

Berg, A. & Pattillo, C. (1999b). What caused the Asian crisis: An early warning system approach. *Economics Notes*, 28(3), 285-334.

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

Bussiere, M. (2013). Balance of payment crises in emerging markets: how early were the 'early' warning signals? *Applied economics*, 45 (12), 1601-1620.

Burkart, O. and Coudert, V. (2002). Leading indicators of currency crises for emerging countries. *Emerging Markets Review*, 3 (2), 107-133.

Candelon, B., Dumitrescu, E-I. & 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.

Candelon, B., Dumitrescu, E-I. & Hurlin, C. (2012). Currency crises early warning systems: why they should be dynamic? *International Journal of Forecasting*, 30 (4), 1016-1029.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. *Biometrics*, 44 (3), 837–845.

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

Frost, J., and Saiki, A. (2014). Early Warning for Currency Crises: What is the role of financial openness? *Review of International Economics*, 22 (4), 722-743.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148 (3), 839–843.

Krzanowsky, W.J. & Hand, D.J. (2009). *ROC curves for continuous data*. Boca Raton, FL: Chapman & Hall/CRC.

Jordà, Ò. (2014). Assessing the historical role of credit: Business cycles, financial crises and the legacy of Charles S. Pierce. *International Journal of Forecasting*, 30 (3), 729-740.

Jordà, Ò., Schularick, M. & Taylor, A.M. (2011). Financial crises, credit booms and external liabilities: 140 years of lessons. *IMF Economic Review*, 59(2), 340-378.

Kamin S.B., Schindler, J. & Samuel, S. (2007). The contribution of domestic and external factors to Latin American devaluation crises: An early warning systems approach, *International Journal of Finance and Economics*, 12 (3), 317-336.

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

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

Kaminsky G.L. (1999). Currency and banking crises: the early warnings of distress. IMF working paper, 178, 1-38.

Kaminsky G. L. (2003). *Varieties of currency crises*. NBER Working chapter 10193. Cambridge, MA.: National Bureau of Economic Research.

Kaminsky G. L. (2006). Currency crises: Are they all the same? *Journal of International Money and Finance*, 25 (3), 503-527.

Lahiri, K. & Moore, G.H. (1991). *Leading economic indicators: New approaches and forecasting records*. Cambridge, United Kingdom: Cambridge University Press.

Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95 (449), 308-311.

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

Schularick, M. & Taylor, A. M. (2012). Credit booms gone bust: monetary policy, leverage cycles and financial crises, 1870–2008. *American Economic Review*, 102(2), 1029–1061.

Stekler, H.O. & Ye, T. (2017). Evaluating with a leading indicator: an application – the term spread. *Empirical Economics*, 53 (1), 183-194.

Chapter 3: Can the SPF and FOMC participants learn from each other? Using qualitative information in the FOMC minutes to elicit forecasts of the U.S. GDP growth

3.1. Introduction

There is a steady interest in assessing whether the monetary policy makers can make valuable predictions on the U.S. economy, and whether the private sector forecasters could reduce their errors if they knew the GB forecasts produced confidentially in the U.S. A seminal study by Romer and Romer (2000) was first to emphasize asymmetry of information between the FRS and private forecasters (the SPF[90] and Blue Chip Indicators), explaining it by the size of resources the FRS devotes to the process. They argued that private forecasts were irrelevant in forecasting inflation, while their output predictions were equally valuable to those in the GB. Sims (2002) found that the GB inflation and output forecasts were superior than their SPF counterparts due to the subjective nature of the GB predictions, which helped to provide an early account for the new disaggregated data and unusual events. Gamber and Smith (2009) argued that the advantage of the GB inflation forecasts declined with time. Faust and Wright (2009) compared the GB forecasts to the time-series models and found that the GB inflation forecasts were more informed at all horizons, while its output predictions were superior only for the current quarter.

The accuracy of the GB forecasts and their five-year publication lag motivated economists to seek ways to elicit them from the monetary policy makers' deliberations, such as the Federal Reserve Beige Book[91] and the FOMC minutes[92], which become available to the public in a timelier manner. Balke and Peterson (2002) developed an ad-hoc procedure to quantify the qualitative information in the Beige Book and forecast U.S. macroeconomic conditions. Goldfarb, Stekler and David (2005) pioneered a simple and replicable method to quantify qualitative statements about

[90] The SPF was started by the American Statistical Association and the NBER in 1968 and have been run by the Philadelphia Fed since 1990.
[91] The Beige Book is published about two weeks ahead of the FOMC meetings.
[92] Today the FOMC minutes are released three weeks after the meeting. Danker and Luecker (2005) explain how the content and release dates of the FOMC minutes have developed over time.

the U.S. macro conditions, which they grouped in nine categories and scored from -1 (most pessimistic views) to +1 (most optimistic views) using a step of 1/4, with 0 indicating the neutral mindset. Lindquist and Stekler (2012) improved on their descriptions of economic conditions in each category and used this method to score the Beige Book' qualitative statements. Stekler and Symington (2016) were the first to apply the updated procedure to the texts in the FOMC minutes from 2006-2010 and presented an abbreviated list of the key words corresponding to each type of economic conditions. They scored the current and future U.S. economic outlook and calibrated them to the real GDP growth rates using the GB and SPF forecasts. Ericsson (2016) reinterpreted these calibrations as elicited GB forecasts.

This chapter contributes to the literature by extending the Stekler and Symington (2016) indexes for the U.S. current and future outlooks (denoted as $SS0_t$ and $SS1_t$ respectively) in both directions to cover a full sample of 1986-2016. It also provides a rigorous comparison of their calibrations' predictive value with the SPF forecasts of the U.S. GDP growth rates in real-time. Additionally, it contributes to the discussion on the asymmetric information between the FRS and private forecasters raised by Romer and Romer (2000) and others. While confirming the FRS' forecasting advantage, this chapter validates a simple and efficient method to elicit valuable information contained in the GB from the FOMC minutes just three weeks after the meeting adjourned, almost five years earlier than the GB publication date. This will allow private forecasters to improve the quality of their output forecasts at all horizons via better assessment of initial conditions in their econometric models. Overall, this chapter once again confirms the value of the subjective and expert forecasting methods.

The study poses the following research questions:

1. Do the $SS0_t$ and $SS1_t$ indexes for the current and future outlooks suggested by Stekler and Symington (2016) remain well-calibrated with the U.S. economy even after being interpolated into past (1986-2005) and extrapolated into future (2011-2016)?

2. Are the calibrated FMI forecasts unbiased and rational?

3. Do the FOMC minutes contain unique information which is not accounted by the SPF forecasts of the U.S. real GDP growth rates?

4. Do the FOMC policy-makers improve their earlier forecasts during the second quarterly meetings?

5. Do the private forecasters and policy-makers learn from each other? This question is two-fold: (i.) do the SPF forecasts released in quarter t improve the FOMC views of the current and future trends formed during the second meeting of the same quarter? (ii.) do the FOMC minutes published ahead of the SPF deadline improve the SPF predictions?

I assume that the commercial sector is well-versed in academic research on the accuracy of the GB forecasts and the methods to elicit them from publicly available monetary policy deliberations. Thus, the SPF participants should use all the information available to them, including that from already published FOMC minutes. I also assume that the SPF forecasters could improve their predictions if they had access to the minutes of the FOMC meeting immediately before the SPF deadline. As the FOMC minutes are now published with a three-week lag, the FRS' forecasting advantage should prevail at the short horizons, including the current and next quarter.

The outline of the remaining sections is as follows. Section 3.2 explains the data and timeline, while section 3.3 specifies the methodology. Empirical results are presented in section 3.4. Section 3.5 summarizes findings in this chapter and highlights how they stand out from the previous research. The references follow the chapter in section 3.6.

3.2. Timeline, data, and hypotheses

3.2.1. Timeline of the FOMC meetings and SPF forecasts

FOMC meets eight times per year, twice every quarter. The $\mathrm{FOMC_t}^{\mathrm{a}}$ and $\mathrm{FOMC_t}^{\mathrm{b}}$ denote the first and second meeting in period t respectively. For example, the consecutive FOMC meetings in 2004 are denoted as $\mathrm{FOMC_{04q1}}^{\mathrm{a}}$, $\mathrm{FOMC_{04q1}}^{\mathrm{b}}$, $\mathrm{FOMC_{04q2}}^{\mathrm{a}}$ …, and $\mathrm{FOMC_{04q4}}^{\mathrm{b}}$. The SPF forecasts are produced once every quarter: their deadline and release dates fall between the two

FOMC meetings in the same quarter. Table 22 below summarizes the typical cycle of the FOMC meetings and publication of their minutes with the dates of the SPF forecasts production and release using year 2004 as an example.

Table 22. Typical timeline of the FOMC and the SPF forecasts with an example.

Dates	Event description	2004 as an example			
		2004-q1	2004-q2	2004-q3	2004-q4
$\mathrm{FOMC}_t{}^{a}$- meeting	Actual day of the 1st FOMC meeting in the period t	01.27.04	05.04.04	08.10.04	11.10.04
$\mathrm{FOMC}_{t-1}{}^{b}$- release	Release of the FOMC minutes from the previous meeting	01.29.04	05.06.04	08.12.04	11.11.04
SPF_t- deadline	Deadline to submit SPF forecasts for the current quarter and up to 4 quarters ahead	02.14.04	05.14.04	08.13.04	11.13.04
SPF_t- release	Day when the SPF forecasts were released	02.23.04	05.24.04	08.20.04	11.22.04
$\mathrm{FOMC}_t{}^{b}$- meeting	Actual day of the 2nd FOMC meeting in the period t	03.16.04	06.29.04	09.11.04	12.14.04
$\mathrm{FOMC}_t{}^{a}$- release	Release of the FOMC minutes from the previous meeting	03.18.04	07.01.04	09.23.04	12.16.04

Table above suggests that, the SPF forecasts are usually issued after the first quarterly FOMC meeting took place, but before its minutes are published. Therefore, producing their forecasts, the SPF participants had access only to the minutes from the second FOMC meeting in the previous quarter. The FOMC members know the SPF forecasts issued in the same period only before their second quarterly meetings.

3.2.2. Data, word choice, scoring and calibrating the FOMC minutes

The FOMC minutes are available at the Federal Reserve Board web-site. Their forecasts are qualitative and need a special quantification procedure. The current and future economic outlook indexes are based on the deliberations on the present (current quarter) and future (one quarter ahead) U.S. economic outlooks, which are usually presented in the paragraphs beginning as “The information reviewed at the meeting suggested that…” and “In their discussion of the economic situation and outlook, meeting participants”. Reading these paragraphs, I focus on the predominance of the recurring words. For example, the most optimistic outlook would use words as “strong”, “robust”, “substantial”, “considerable”, “upbeat”, “brisk”, “surge” and “buoyant”. It would be marked as strong expansion with SS=1, and the corresponding FMI is calibrated as 4% GDP growth.

Table 23. Criteria for scoring and calibrating deliberations on the economic outlook

Outlook	Score (SS)	Condition diagnosed or forecast	Assessment	Recurring words in the minutes to score	Calibrations to annual real GDP growth rate, %	
					Stekler and Symington (2016)	Alternative
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Optimistic	+1	The economy is strong or will expand very strongly	Strong growth	Strong, robust, substantial, considerable, upbeat, brisk, surge, buoyant	>=4.0	>=4.5
	+3/4	The economy is growing normally or will definitely continue to grow	Normal growth	Normal, solid, steady, close to potential, continued to rise	3.4	3.8
	+1/2	The economy is growing at a “modest” pace or will do well barring unforeseen circumstances	Modest growth	Modest, moderate, sustainable	2.8	3
	+1/4	There is some risk of recession or downturn <30%, or the economy will grow but slower than usual	Slow growth	Slow, gradual, subdued, muted, subpar, bottoming out	2.1	2.3
Neutral	0	It is unclear where the economy is or where it will go because the signs are mixed	Unclear	Unclear, mixed, uneven, uncertain	1.5	1.6
	-1/4	The economy is visibly slowing, decelerating, or there is quite a bit of risk of recession, >30% but <60%	Decelerating growth	Decelerating, stabilizing, outgoing adjustment leveling out, flattening	0.9	0.8
	-1/2	The economy is sluggish, barely growing, or there is >60 risk of recession.	Continued weakness	Continued weakness, sluggish, slack, below potential, flat	0.3	0.1
	-3/4	The economy is declining, will contract, or there are mild recession conditions	Decline	Declining, deteriorating, mild downturn	-0.4	-0.7
	-1	Recession conditions are here or imminent and it is worse than any recession in recent history	Recession	Recession, contraction, sharp and widespread /appreciable / substantial decline	<=-1	<=-1.4

Table 23 above summarizes criteria for scoring and calibrating the outlook of the FOMC minutes' qualitative forecasts. Columns 1 through 6 encompasses Table A2 in Goldfarb et al. (2005), Table 1 in Lindquist and Stekler (2012), and Tables 3 and 4 in Stekler and Symington (2016), extending their list of recurring words. The index values SS=-1, -3/4, -1/2 imply the pessimistic state (recession, decline, and continued weakness). Decelerating growth, unclear, and slow growth are associated with SS=-¼, 0, and ¼ and characterize the neutral state. The modest, normal, and strong growth correspond to SS=1/2, 3/4, and 1, respectively.

Stekler and Symington (2016) proposed that economic outlook indexes from column 2 can be mapped into the annualized real GDP growth rates using the calibrations presented in column 6. To check the robustness of results based on the Stekler and Symington calibrations I considered an alternative mapping of SS scores into FMI calibrations (see column 7). These alternative calibrations were obtained by regressing the actual real-time GDP growth rates on the current economic outlook[93]. I obtained rgdp3=1.565+2.968SS0a and used the fitted values corresponding to each SS value as the alternative FMI calibrations. These alternative FMI calibrations fall within the range [-1.4%, +4.5%].[94] I used these calibrations to check the sensitivity of the results provided in this paper. Qualitatively, the results obtained with the alternative calibrations were the same as those obtained using the Stekler and Symington mapping of SS scores into FMI calibrations. Therefore, for the sake of brevity, only the results for the Stekler and Symington calibrations are reported.

The study covers 124 quarters over 1986-2016. To extend the Stekler and Symington (2016) current and future economic outlook scores (denoted as SS0 and SS1 respectively), I've read and quantified the texts of the FOMC minutes for all meetings in 1986-2005 and 2011-2016, and used the full (combined) score series to calibrate the real GDP annual growth forecasts

[93] This regression used data for 1986-2007.

[94] Coefficients are estimated at 1% significance level. Results are not sensitive to the isomorphic transformations of the economic outlook index. For example, adding 1 to SS index for each category will result in the linear relation rgdp3=-1.403+2.968(SS0a+1), which results in the same alternative FMI scale as in column 7.

$FMI_{t+h|t}$ for h=0 and h=1 horizons.

To align the economic outlook scores and their calibrations with the quarterly data frequency, I use the upper-scripts “a” and “b” to denote the series corresponding to the first and second FOMC meeting in every period t. Calibrations of these scores to the U.S. GDP growth rates from the first ($FMI^{a}_{t+h|t}$) and second ($FMI^{b}_{t+h|t}$) FOMC meetings constitute the main subject of this research. The “final” (third) real-time estimates[95] of the actual U.S. GDP growth and their SPF forecasts are collected by the Philadelphia Fed RTDSM.

3.3. Methodology

3.3.1. Research hypotheses

The timeline discussed above suggests the following hypotheses:

H1. The SS0 and SS1 indexes for the current and future outlooks are well calibrated with the U.S. economy.

H2. The calibrated $FMI^{a}_{t+h|t}$ and $FMI^{b}_{t+h|t}$forecasts are rational (unbiased and efficient).

H3. The competing GDP growth forecasts produced by the monetary policy-makers ($FMI^{a}_{t+h|t}$) and private sector ($\mathrm{SPF}_{\mathrm{t+h|t}}$) both contain unique information not accounted for by their rival: neither of the two forecasts encompasses another.

H4. The policy-makers are consistent: their FMI forecasts from the second meetings encompass their own FMI forecasts from the first meetings.

H5. The SPF forecasts are more efficient than those made by monetary policy-makers: they use all the information available to them, including the FOMC minutes from previous quarter published before the SPF deadline, but the FOMC does not pay attention to the SPF forecasts released before their second quarterly meetings. In other words, the $\mathrm{SPF}_{\mathrm{t+h|t}}$ are strongly efficient ($FMI^{b}_{t|t-1}$ could

[95] In general, literature uses the third real-time estimates to evaluate forecasts. However, Romer and Romer (2000) used the second real-time estimate, while Sims (2002) used historical data. The robustness check conducted using the second real-time estimate and historical data yielded the same conclusions.

not improve them), but $FMI_{t+h|t}^{b}$ are not (SPF$_{\mathrm{t+h|t}}$ could improve on $FMI_{t+h|t}^{b}$).

3.3.2. Evaluating calibration of the SS economic outlook indexes with actual economy

To answer the first research question and test hypothesis H1, I assess the calibration of the current and future economic outlook indexes SS0 and SS1 with the U.S. economy using the following tools: (1) mean comparison with the actual data and SPF forecasts; (2) graphing the movements in the SS indexes and the actual US economy; (3) graphing the movements in the calibrated FMI forecasts compared with the actual economy and SPF forecasts; (4) evaluation of the FMI forecast error, defined as $e_{t+h|t} = A_{t+h} - F_{t+h|\,t}$, where A_{t+h} and $F_{t+h|\,t}$ are the actual and forecast values for period $(t + h)$ conditionally on the information available in period t.[96]

3.3.2.1. Rationality tests

To answer the second research question and test hypothesis H2, I use the forecast rationality test[97] using the Mincer and Zarnowitz (1969) regression[98]:

$$A_{t+h} = \beta_o + \beta_1 F_{t+h|t} + e_{t+h|t} \qquad [23]$$

The forecasts are said to be rational if they are unbiased and weakly efficient. Rejection of a joint null hypothesis Ho: $\beta_o = 0, \beta_1 = 1$ implies that forecasts are biased and/or weakly inefficient, and could be improved if they were scaled by β_1 and intercept-corrected by β_o. Failure to reject Ho: $\beta_o = 0, \beta_1 = 1$ with β_o significantly different from zero provides evidence of the conditionally biased forecasts. After establishing the rationality of both competing forecasts, I turn attention to comparison of their predictive ability via encompassing tests.

3.3.2.2. Encompassing tests for competing forecasts

[96] The forecast accuracy measures include: the mean and standard deviation of the forecast error: $\mu(e_{t+h|t})$ and $\sigma(e_{t+h|t})$; forecast bias or mean absolute error $MAE = \frac{1}{N}\sum_{t=1}^{N}\left|e_{t|t-h}\right|$; root mean square forecast error $RMSFE = \sqrt{\frac{1}{N}\sum_{t=1}^{N} {e_{t|t-h}}^{2}}$; and the mean absolute percentage error $MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|p_{t|t-h}\right|$, where the percentage forecast error $p_{t|t-h} = 100\frac{e_{t|t-h}}{A_t}$.

[97] Tests of single coefficient restrictions use t-statistics and normal distribution, while tests of the joint linear restrictions use the Wald statistics and F distribution.

[98] The Newey-West procedure to estimate the Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors was applied in all regressions when the AR(1) error terms could not be rejected.

To answer the third and fourth research questions, I employ the Chong and Hendry (1986) regression, which evaluate the contributions of the considered forecasts to the prediction of the actual value:

$$A_{t+h} = \beta_1 F1_{t+h|t} + \beta_2 F2_{t+h|t} + e_{t+h} \qquad [24],$$

where $F1_{t+h|t}$ and $F2_{t+h|t}$ are two competing forecasts.

The encompassing test involves two null hypotheses: (i) Ho: $\beta_1 = 1, \beta_2 = 0$; and (ii) Ho: $\beta_1 = 0, \beta_2 = 1$. A rejection of the first null and a failure to reject the second null give the evidence that forecast F2 encompasses F1, and vice versa. A joint rejection of both nulls above reveals that both F1 and F2 are relevant. Additional restrictions, i.e. Ho: $\beta_1 = \beta_2 = 0.5$, test the size of each forecast contributions to the prediction of the actual value. Using $FMI^a_{t+h|t}$ as $F1_{t+h|t}$ and $\mathrm{SPF}_{\mathrm{t+h|t}}$ (and $FMI^b_{t+h|t}$) as $F2_{t+h|t}$, I test hypotheses H4 (H5) respectively.

3.3.2.3. Orthogonality tests for strong information efficiency

To answer the last research question and test hypothesis H5, I employ the orthogonality test. It assesses whether the forecasts used all the available information, and therefore are strongly efficient. The test uses regression of the actual value in the forecast and a variable z_t, which contains a relevant piece of data from the forecasters' information set:

$$A_{t+h} = \beta_1 F_{t+h|t} + \delta z_t + e_{t+h} \qquad [25]$$

Rejection of the null hypothesis Ho: δ =0 provides evidence that the forecasts have not used all the information available to them. To test whether the FMI forecasts from the 2nd FOMC meetings account for the SPF forecasts released in the same quarter, I use [25] with $F_{t+h|t} = FMI^b_{t+h|t}$, $z_t = \mathrm{SPF}_{\mathrm{t+h|t}}$, and h=0, 1 quarters ahead. To test whether the SPF forecasts account for the information available from the recently published FOMC minutes, I use [25] with $F_{t+h|t} = \mathrm{SPF}_{\mathrm{t|t}}$ (current quarter SPF forecasts) and $z_t = FMI^b_{t|t-1}$ (next quarter FMI from the 2nd meeting a quarter earlier).

3.4. Results

3.4.1. Evaluating SS calibrations and their forecast accuracy

First, I compare dynamics in the SS outlook scores and actual GDP growth. The shadowed areas in the top panel in the Fig.13 below show the actual GDP growth (measured on the left vertical axis, as annualized percent rate). The solid red and dashed green lines indicate the SS scores from the first and second meetings respectively (measured on the right vertical axis as indexes). The current economic outlook scores and the actual GDP growth exhibit very similar dynamics. However, movements in the SS index are not as pronounced as those in the actual GDP during the times of very high growth or very severe recession due to its limited range [-1; 1].

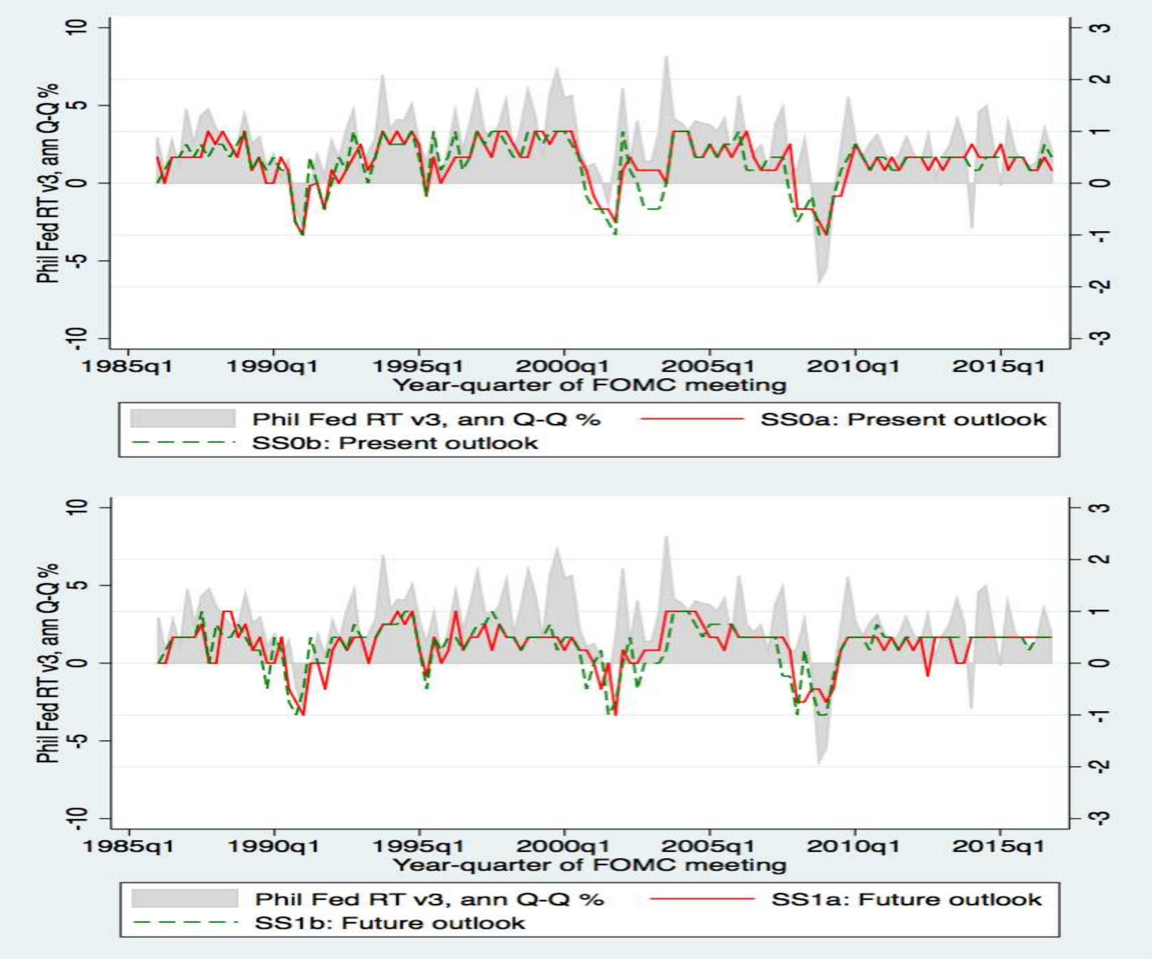


Figure 13. Present and future economic outlook scores and actual GDP growth (1986-2016)

The lower panel of Fig.13 above draws a similar picture as it compares the next quarter economic outlook indexes with the actual GDP growth. Comparing the scales on the left and right vertical axes, one can see that the calibrations suggested by Stekler and Symington (2016) and summarized above in Table 23 are visually valid for an extended sample of 1986-2016. For a more accurate comparison, let's look at the FMI forecasts obtained with these calibrations, compared with the corresponding SPF forecasts and the actual GDP growth data in Figure 14 below, focusing on the forecasts for h=0 and h=1 quarters ahead respectively. In both graphs, the solid red and dashed green lines indicate $\mathrm{FMI}^{a}_{t+h|t}$ and $\mathrm{FMI}^{b}_{t+h|t}$, while the long-dashed black line denotes the rival forecasts ($\mathrm{SPF}_{t+h|t}$), and the shadowed gray area refers to the actual GDP growth.

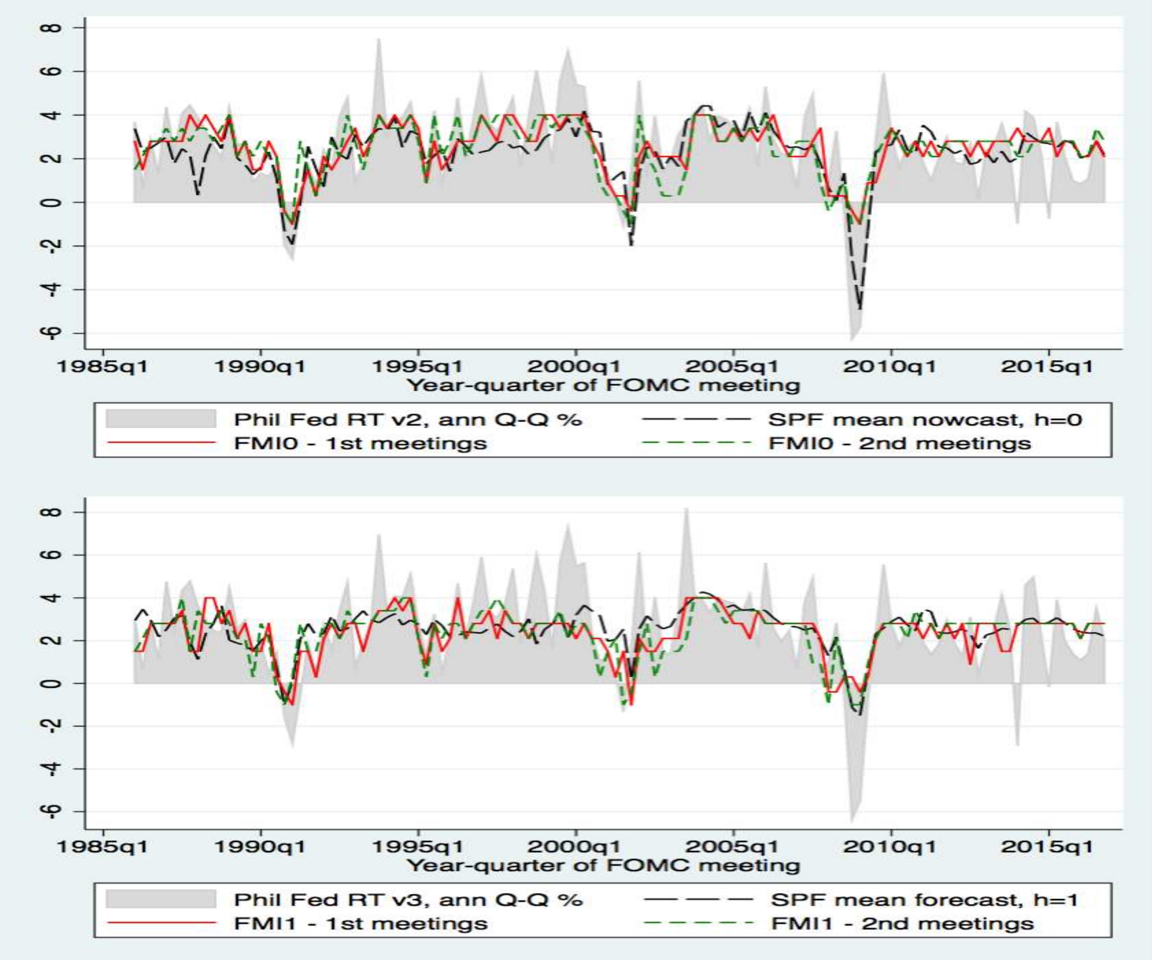


Figure 14. Dynamics in actual and forecast GDP growth rates (annualized %) for h=0 and h=1 (1986-2016)

Figure 14 above compares the actual and forecast GDP growth rates for h=0 (top panel) and h=1 (lower panel). It confirms visually that FMI forecasts are closely calibrated to the actual GDP growth, and are not far from the quantitative SPF forecasts. In some periods, notably in late 80th and late 90th, the FMI forecasts seemed to do better than their SPF counterparts. The SPF forecasts more accurately recognized recession periods (h=0), which can be explained by the fact that the FMI calibrations are truncated by their range. Yet, the SPF forecasts, although not truncated in either direction, were very conservative about expansions, predicting growth above 4% only in 5 periods. Table 24 below compares means and standard deviations for the SS scores, FMI and SPF forecasts, and actual data.

Table 24. Summary statistics for the actual GDP growth rates, SPF forecasts, SS outlook indexes and FMI elicitcasts (1986-2016)

	$rgdp_o$	$rgdp_1$	SPF_0	SPF_1	SS_0^a	SS_1^a	FMI_0^a	FMI_1^a	SS_o^b	SS_1^b	FMI_0^b	FMI_1^b
Mean	2.54	2.54	2.28	2.52	0.39	0.33	2.49	2.33	0.38	0.34	2.47	2.38
St. dev.	2.20	2.21	1.37	0.93	0.45	0.43	1.15	1.08	0.50	0.45	1.24	1.14

Table 24 indicates that the SS current and future economic outlook indexes from the first and second meetings had very close mean values. Their respective FMI calibrations averaged 2.49% (2.47%) and 2.33% (2.38%) for the current (oncoming) quarter. On average, actual real-time GDP grew at 2.54% per annum. Thus, the FMI now-casts were very close to the actual data while the next quarter forecasts were more conservative. The SPF forecasts, on the opposite, were more conservative in their current quarter estimates while overly optimistic about the next quarter predictions.

Another way to check calibrations of the SS indexes is to analyze their correlations with the actual data and SPF forecasts.

Table 25. Correlations between the actual data and forecasts (1986-2016)

Horizon	$(fmi^a_{h;}rgdp)$	$(fmi^b_{h;}rgdp)$	$(fmi^a_{h;}spf_h)$	$(fmi^b_{h;}spf_h)$	$(spf_h; rgdp)$
h=0	0.62	0.67	0.76	0.69	0.65
h=1	0.44	0.43	0.68	0.68	0.47

The correlation between the FMI calibrations of the current quarter GDP growth with the actual data was 0.62 and 0.67 for the 1st and 2nd FOMC meetings respectively. For comparison, their corresponding correlations with the SPF forecasts for h=0 were 0.76 and 0.69. This indicates reasonably good correlation, very close to the one between the SPF forecasts and the actual data for the same horizon.

The correlations between the FMI calibrations of the next quarter GDP growth with the actual data were 0.44 and 0.43 for the 1st and 2nd FOMC meetings. While low, it is close to the correlation between the SPF next quarter forecasts and actual realizations. The h=1 FMI and SPF forecasts were strongly correlated, reaching 0.68 for both FOMC meetings.

Analysis of the forecast errors gives the last piece of evidence on the calibration of the SS indexes: the current quarter FMI calibrations had lower mean forecast error and mean absolute error than their SPF counterparts. The standard error of the current quarter FMI forecasts equaled 1.745 and 1.646 for the 1st and 2nd meetings respectively, bounding the SPF forecasts standard error (1.685) for h=0. For h=1 quarter, the SPF forecast error terms have lower mean and standard deviation, but higher MAE.

Table 26. Measures of forecast accuracy (1986-2016)

	FMI_0^a	FMI_0^b	SPF_0	FMI_1^a	FMI_1^b	SPF_1
$\mu(e_{t\|t-h})$	0.052	0.075	0.227	0.216	0.171	0.022
$\sigma(e_{t\|t-h})$	1.745	1.646	1.685	1.987	2.003	1.958
MAE	1.240	1.207	1.320	1.484	1.382	1.496

The alternative FMI calibrations result in slightly negative forecast errors in the first quarterly meetings; and very close to zero for the second quarterly meetings. The standard deviations and mean absolute errors are very robust to these new calibrations.

Finally, Figure 15 below compares the plots for the FMI and SPF forecast errors for the current quarter (top panel) and the next quarter (bottom panel). The error terms from the FMI and SPF forecasts show similar dynamics. The SPF forecasts had higher error term than FMI predictions before 2000, while the FMIs had their biggest mistakes in periods of very high or very low growth, i.e. around 2008 crisis, which was due to the lower bound of the FMI calibrations limited by -4%. Figure 15 does not indicate any bias or systematic errors.

To sum up, the results in this section indicate that the method proposed by Stekler and Symington (2016) is not restricted to their sample (2006-2010) and produces very good calibrations of the U.S. output growth forecasts for the current and next quarters ahead. Following Ericsson (2016), these calibrations can be reinterpreted as elicitcasts of the GB forecasts. These elicitcasts become available to commercial forecasters just three weeks after the FOMC meetings. Their contribution to the forecasts of the actual US output growth will be analyzed in section 3.4.3.

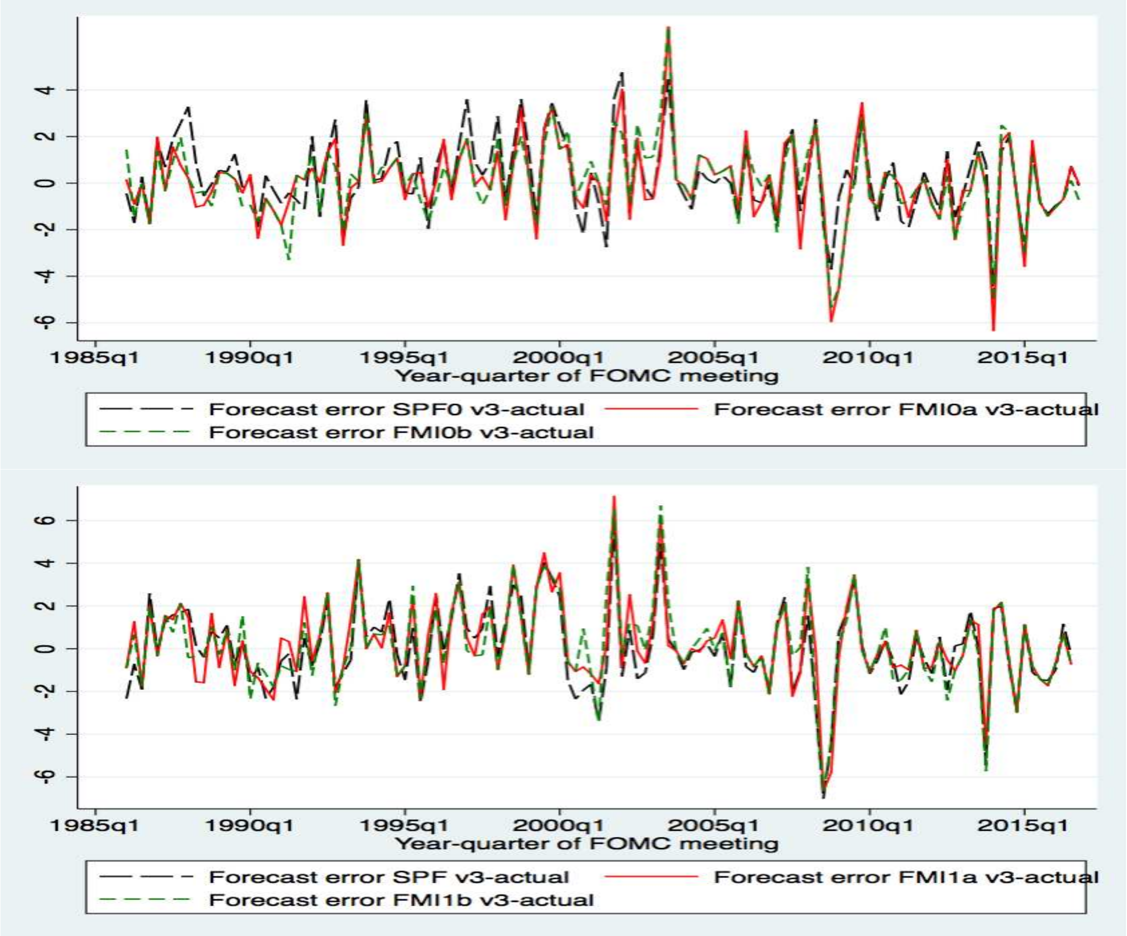


Figure 15. SPF and FMI forecast errors, h=0 and h=1 (1986-2016)

The next section tests the rationality of the SPF forecasts and the elicitcasts.

3.4.2. Rationality tests for the FMI and SPF forecasts

In the previous section I demonstrated that the FMI elicitcasts are well calibrated with the US economy even when a sample was extended to cover 1986-2016. This and following sections will use data from the full sample (1986-2016) to test the hypotheses posed above in 3.3.1.

Table 27 below provides evidence on testing the second hypothesis in this study. Regardless of horizon, the calibrated FMI forecasts from both 1st and 2nd FOMC meetings and their SPF counterparts are rational (unbiased and weakly efficient) at both horizons of interest. All these forecasts contain important information about the U.S. GDP growth in the current quarter and one quarter ahead. The estimates of slope coefficient are all positive and close to one with intercepts not significantly different from zero.

Table 27. Mincer and Zarnowitz (1969) rationality tests (1986-2016)

	h=0			h=1		
	FMI_0^a	FMI_0^b	SPF_0	FMI_1^a	FMI_1^b	SPF_1
β_o[99]	-0.415 (0.586)	-0.400 (0.573)	0.193 (0.381)	0.439 (0.736)	0.552 (0.736)	-0.255 (0.823)
β_1	1.187*** (0.205)	1.192*** (0.119)	1.035*** (0.141)	0.904*** (0.265)	0.839*** (0.260)	1.110*** (0.289)
RMSE	1.738	1.636	1.692	1.992	2.002	1.964
Adj. R-sq.	0.377	0.449	0.411	0.189	0.180	0.212
Ho[100]: $\beta_o = 0, \beta_1 = 1$	0.66 (0.517)	1.21 (0.302)	1.65 (0.196)	0.67 (0.513)	0.39 (0.675)	0.11 (0.896)

The next step in this analysis is to assess whether one of the considered forecasts is superior over its rival using the forecast encompassing tests.

3.4.3. Tests of forecast encompassing for FMI_h^a and SPF_h

This section sheds light on the third hypothesis detecting whether only one of the rival forecasts is relevant to predict the actual value and, therefore, encompasses another.

Table 28. Chong and Hendry (1986) encompassing tests for FMI_h^a and SPF_h (1986-2016)

	h=0	h=1
F1=FMI_h^a	0.446** (0.242)	0.446* (0.203)
F2=SPF_h	0.663*** (0.254)	0.616*** (0.205)
RMSE	1.644	1.933
Adj. R-sq.	0.761	0.670
Ho: $\beta_{F1} = 1, \beta_{F2} = 0$	F(2,122)=4.06 (0.020*)	F(2,121)=4.67 (0.011*)
Ho: $\beta_{F1} = 0, \beta_{F2} = 1$	F(2,122)=4.00 (0.021*)	F(2,121)=2.72 (0.069**)
Ho: $\beta_{F1} = \beta_{F2} = 0.5$	F(2,122)=1.87 (0.159)	F(2,121)=0.63 (0.534)

[99] Numbers in brackets next to coefficient estimates show their standard errors. Symbols *, ** and *** indicate statistical significance at 5%, 10% and 1% level respectively.

[100] The first number indicates the F statistics, the number in brackets next to it indicates the p-value of $F(q,k) > F_{critical}$, q=2, k=n-q-h.

Results indicate that regardless of a horizon, both FMI_0^a and SPF_0 forecasts contain unique information which helps to predict the actual growth in real GDP: we reject the null that either one or another forecast is not significant. Also, I fail to reject that the two forecasts make equal contributions to prediction of the actual value – this confirms the observation by Romer and Romer (2000) which they, however, did not test formally.

Results in Table 28 indicate that both the FOMC members and SPF forecasters know some unique information about the path of the U.S. output which is not known to their rival. The FMI calibrations retain the GB forecasts' informational advantage which was previously explained in the literature by the FRS forecasting capacity, including its better understanding of the available disaggregated data and plentiful resources devoted to the process. This conclusion is in line with Romer and Romer (2000) and Gamber and Smith (2009). However, neither of them verified whether the FRS' information set is superior over the private forecasters with the encompassing tests. Results in this chapter indicate the SPF forecasters also possess unique information not available to the FOMC members – this was not previously known in the literature. The mastery of the SPF forecasts might come from the averaging results of different models which were likely used by the survey participants and their level of expertise. Therefore, results in this section speak in favor of forecasting methods using both qualitative and judgmental forecast methods and expert surveys' techniques.

3.4.4. Tests of forecast encompassing for FMI_h^a and FMI_h^b

Another pair of the encompassing tests helps find evidence on hypothesis H4.

Table 29. Chong and Hendry (1986) encompassing tests for FMI_h^a and FMI_h^b (1986-2016)

	h=0	h=1
F1=FMI_h^a	0.235 (0.204)	0.591* (0.218)
F2=FMI_h^b	0.837*** (0.203)	0.479* (0.216)
Adj. R-sq.	0.763	0.659
RMSE	1.637	1.965
Ho: $\beta_{F1} = 1, \beta_{F2} = 0$	F(2,122)=8.95 (0.000***)	F(2,121)=3.98 (0.021**)
Ho: $\beta_{F1} = 0, \beta_{F2} = 1$	F(2,122)=1.37 (0.259)	F(2,121)=2.94 (0.056*)
Ho: $\beta_{F1} = \beta_{F2} = 0.5$	F(2,122)=2.09 (0.129)	F(2,121)=0.75 (0.475)

For both forecast horizons, we can reject a null that FMI_h^b is irrelevant, but fail to reject that FMI_h^a is irrelevant (both at 5% significance level). Thus, FMI_h^b encompasses FMI_h^a, which confirms the hypothesis H4: that the policy-makers are consistent in their forecasts and use all the information available to them at the first quarterly FOMC meeting to make statements about the economic outlook during the second FOMC meeting.

Finally, the next section employs the orthogonality tests of strong efficiency to collect the statistical evidence on the fifth hypothesis in this study.

3.4.5. Tests of forecast strong efficiency (orthogonality)

This section supports hypothesis H5 by testing whether the FMI and SPF forecasts used all information available to them from their rivals.

Table 30. Strong efficiency tests for SPF_0, FMI_0^b, and FMI_0^b (1986-2016)

	F=$SPF_{0,t}$; Z=$FMI_{1,t-1}^b$	F=FMI_0^b; Z=SPF_0	F=FMI_1^b; Z=SPF_1
β_1	1.296 (0.208***)	0.617*** (0.194)	0.398* (0.185)
δ	-0.208 (0.212)	0.502* (0.215)	0.651*** (0.188)
Adj. R-sq.	0.748	0.783	0.669
RMSE	1.691	1.566	1.936

The second column in Table 30 suggests that making their current quarter GDP growth predictions, the SPF forecasters fully used qualitative forecasts contained in the FOMC minutes available to them before the forecast deadline. This conclusion proves that professional forecasters were influenced by Romer and Romer (2000) and others who pointed to the FRS forecasting superiority, and, therefore, paid due attention to the deliberations in the FOMC minutes. The third and fourth columns in Table 30 indicate that the FOMC forecasts made during the second meeting each quarter could be improved with the SPF forecasts released before it in the same quarter. This conclusion is consistent with one of the results in Section 3.4.3, which highlighted the fact that the SPF forecasters also possess unique information about the path of the U.S. output.

However, existing academic studies had overstated superiority of the FRS information set. It seems that the U.S. monetary policy makers do not put high weights on the SPF forecasts in their forecasting process. This may be due to concerns about the private forecasters strategic behavior[101].

3.5. Conclusion for chapter 3

Public availability of the FOMC minutes, official views they contain, and the ability to elicit the GB forecasts from the meeting delibarations make them a valuable reference to infer the current and future economic outlook in the USA.

This chapter contributes to the literature on the economic forecasting in several ways. First, it extended the Stekler and Symington (2016) $SS0_t$ and $SS1_t$ indexes for the U.S. current and future outlooks, adding 26 years of bi-quarterly observations to cover a full sample of 1986-2016. The analysis demonstrated that the constructed indexes are well calibrated with the U.S. real GDP growth rates even after they are extended in both directions. Thus, the method proposed by Stekler and Symington (2016) is not restricted to their sample (2006-2010) and produces very good calibrations of the U.S. output growth forecasts for the current and next quarters ahead. Following Ericsson (2016), these calibrations are reinterpreted as elicitcasts of the GB forecasts; statistical tests found them to be unbiased and weakly efficient, and therefore, rational.

This research also contributes to the discussion on the asymmetric information between the FRS and private forecasters raised by Romer and Romer (2000) and others. This chapter rigorously demonstrated that both the FOMC minutes and SPF forecasts contain unique information which is not accounted for by their rival: neither of the forecast encompasses another. This is the first time when contributions of the FRS and SPF forecasts to the actual US output growth predictions were tested using encompassing techniques. The analysis suggested that one should put equal weights on the FMI elicitcasts and SPF forecasts – this result is similar to the one published in Romer and Romer (2000) when they analyzed the GB and SPF forecasts. Unlike Gamber and Smith (2009), who found that the GB inflation forecasts' advantage declined in time, I found that the SS indexes and the FMI elicitcasts had stable performance even for a significantly extended sample.

[101] Forecasters may engage in a strategic behavior, such as herding, reputational cheap talk, radical forecasting or forecast competition (i.e. Trueman, 1994; Lamont, 2002; Ottaviani and Sorensen, 2006).

Another contribution this chapter made to the literature was to show that unlike the SPF forecasters, who fully used qualitative forecasts available to them in the already published FOMC minutes, the FOMC members, in their turn, did not pay due attention to the SPF forecasts published between their first and second meetings. The FRS policy-makers accounted only for their assessment of the U.S. economic conditions expressed in the first quarterly meetings: the FOMC forecasts made during the second meeting each quarter could be improved with the SPF forecasts released earlier in the same quarter. Results are robust to the use of alternative FMI calibrations.

This chapter demonstrated that the method developed by Stekler and Symington (2016) gives a informational gain: while the GB forecasts are available to the private sector only after a five-year publication embargo, the FMI elicitcasts constructed here using simple textual analysis technique can be obtained as soon as just three weeks after the FOMC meeting. This finding will allow private forecasters to quickly and easily improve their assessment of initial conditions used in the econometrics models, and therefore the quality of their output forecasts at all horizons. Yet, the question whether removal of such publication embargo on the FOMC minutes is desirable remains unanswered. I believe that the current publication policy is optimal. Overall, results in this chapter favor of the qualitative forecasting techniques, including the subjective and expert methods.

3.6. References for chapter 3

Balke, N. & Petersen D. (2002). How well does the Beige Book reflect economic activity? Evaluating qualitative information quantitatively. *Journal of Money, Credit and Banking*, 23(1), 114-136.

Chong, Y. & D. Hendry. (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, 53 (4), 671-690.

Danker, D. J. & Luecker, M.M. (2005). Background in FOMC Meeting Minutes. *Federal Reserve Bulletin*, 91 (2), 175-179.

Ericsson, N. R. (2016). Eliciting GDP forecasts from the FOMC minutes around the financial crisis. *International Journal of Forecasting*, 32 (2), 571-583.

Faust, J. & Wright, J.H. (2009). Comparing Greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business and Economic Statistics*, 27 (4), 468-479.

Gamber, E.N. & J. Smith, J.K. (2009). Are the Fed's inflation forecasts still superior to the private sector's? *Journal of Macroeconomics*, 31(2), 240-251.

Goldfarb, R.S., Stekler, H. O. & David, J. (2005). Methodological issues in forecasting: Insights from the egregious business forecast errors of late 1930, *Journal of Economic Methodology*, 12 (4), 517-542.

Joutz, F. & Stekler, H. (2000). An evaluation of the predictions of the Federal Reserve. *International Journal of Forecasting*, 16 (1), 17-38.

Lamont, O.A. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior and Organization*, 48 (3), 265-280.

Mincer, J. & Zarnowitz, V. (1969). The evaluation of economic forecasts. In J. Mincer (Ed.), *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance* (3-46). New York, National Bureau of Economic Research.

Newey, W.K. & West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703-708.

Ottaviani, M. & Sørensen, P.N. (2006). The strategy of professional forecasting. *Journal of Financial Economics*, 81 (2), 441-466.

Romer, C. D. & Romer, D.H. (2000). Federal Reserve information and the behavior of interest rates. *American Economic Review*, 90(3), 429-457.

Sims, C. (2002). The role of models and probabilities in the monetary policy process. *Brookings Chapters on Economic Activity*, 2002 (2), 1-40.

Stekler, H. & Symington, H. (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006-2010. *International Journal of Forecasting*, 32 (2), 559-570.

Trueman, B. (1994). Analyst Forecasts and Herding Behavior. *The Review of Financial Studies,* 7 (1), 97-124.

References

Abberger, K., Frey, M., Kesina, M. & Stangl, A. (2009). Indicatoren fur die globale Konjuctur, *Ifo Schnelldienst*, 62(34/35), 32-41.

Anderson, O. (1952). The business test of the IFO-Institute for economic research. Munich, and its theoretical model. *Revue de l'Institut International de Statistique*, 20 (1), 1-17.

Ang, A., Bekaert, G., & Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, 54 (4), 1163-1212.

Balke, N. & Petersen D. (2002). How well does the Beige Book reflect economic activity? Evaluating qualitative information quantitatively. *Journal of Money, Credit and Banking*, 23(1), 114-136.

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

Berg, A & Pattillo, C. (1999b). What caused the Asian crisis: An early warning system approach. *Economics Notes*, 28(3), 285-334.

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

Blöchinger, A. & Lieppold, M. (2006). Economic benefits of powerful credit scoring. *Journal of Banking and Finance*, 30 (3), 851-873.

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

Bussiere, M. (2013). Balance of payment crises in emerging markets: how early were the 'early' warning signals? *Applied economics*, 45 (12), 1601-1620.

Burkart, O. and Coudert, V. (2002). Leading indicators of currency crises for emerging countries. *Emerging Markets Review*, 3 (2), 107-133.

Candelon, B., Dumitrescu, E-I. & 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.

Candelon, B., Dumitrescu, E-I. & Hurlin, C. (2012). Currency crises early warning systems: why they should be dynamic? *International Journal of Forecasting*, 30 (4), 1016-1029.

Carlson, J.A. & Parkin, M. (1975). Inflation Expectations. *Economica,* 42 (165), 123-138.

Carnot, N., Koen, V., & Tissot, B. (2005). *Economic Forecasting*. New York, NY: Palgrave Macmillian.

Chong, Y. & D. Hendry. (1986). Econometric evaluation of linear macro-economic models. *Review of Economic Studies*, 53 (4), 671-690.

Claveria, O., Pons, E. & Ramos, R. (2007). Business and consumer expectations and macroeconomic forecasts. *International Journal of Forecasting*, 23 (1), 47-69.

Clements, M. P. (2015). Are professional macroeconomic forecasters able to do better than forecasting trends? *Journal of Money, Credit and Banking*, 47 (2-3), 349-382.

Croushore, D. (2005). Do consumer confidence indexes help forecast consumer spending in real time? *North American Journal of Economics and Finance*, 16 (3), 438-450.

Croushore, D. & Stark, T. (2001). A real-time database for macroeconomists. *Journal of Econometrics*, 105 (1), 111-130.

Cunningham, A. (1997). Quantifying Survey Data. *Bank of England Quarterly Bulletin*, 37 (3), 292-300.

Danker, D. J. & Luecker, M.M. (2005). Background in FOMC Meeting Minutes. *Federal Reserve Bulletin*, 91 (2), 175-179.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating curves: a nonparametric approach. *Biometrics*, 44 (3), 837–845.

Dorfman, D.D. & Alf E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. *Journal of*

Mathematical Psychology, 6 (3), 487-496.

Doswell, D-J., Davies-Jones, R. & Keller, D.L. (1990). On summary measures of skill in rare events forecasting based on contingency tables. *Weather and forecasting*, 5(4), 576 – 585.

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

Ericsson, N. R. (2012). *Economics 8378.10 (Economic Forecasting): Course Lecture Notes*. Washington, DC, USA [Unpublished manuscript].

Ericsson, N. R. (2016). Eliciting GDP forecasts from the FOMC minutes around the financial crisis. *International Journal of Forecasting*, 32 (2), 571-583.

Faust, J. & Wright, J.H. (2009). Comparing Greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business and Economic Statistics*, 27 (4), 468-479.

Fisher, R.A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of Royal Statistical Society*, 85 (1), 87-94.

Fisher, R.A. (1950). *Statistical Methods for Research Workers* (11th ed.). New York, NY: Hafner Publishing Company.

Frost, J., and Saiki, A. (2014). Early Warning for Currency Crises: What is the role of financial openness? *Review of International Economics*, 22 (4), 722-743.

Gamber, E.N. & J. Smith, J.K. (2009). Are the Fed's inflation forecasts still superior to the private sector's? *Journal of Macroeconomics*, 31(2), 240-251.

Garnitz, J. (2017). *Ifo World Economic Survey – Description and Information*. Munich, Germany: Ifo Institute. Retrived from https://www.cesifo-group.de/dms/ifodoc/docs/facts/survey/WES/Description_WES_2017.pdf

Goldfarb, R.S., Stekler, H. O. & David, J. (2005). Methodological issues in forecasting: Insights from the egregious business forecast errors of late 1930, *Journal of Economic Methodology*, 12 (4), 517-542.

Gorr, W.L. & Schneider, M.J. (2013). Large-change forecast accuracy: Reanalysis of M3-competition data using receiver operating characteristic analysis. International Journal of Forecasting, 29 (2), 274-281.

Granger, C.W.J. & Pesaran, M.H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, 19 (7), 537-560.

Hamella, S. & Haupt, H. (2007). Suitability of WES data for forecasting inflation. In G. Goldrian (Ed.), *Handbook of Survey-Based Business Cycle Analysis* (99-115). Cheltenham, United Kingdom & Nortmapton, MA: Edward Elgar Publishing Limited.

Haupt, H. & Waller, S. (2000). Economic analysis and short-term forecasting with qualitative data from the Economic Survey International. In K.H. Oppenlander, G. Poser, and B. Schips (Eds.), *Use of Survey Data for Industry, Research, and Economic Policy: CIRET Conference* (527-547). Avebury, United Kingdom: Aldershot.

Henriksson, R.D. & Merton, R.C. (1981). On market timing and investment performance 2: Statistical procedures for evaluating forecasting skills. *Journal of Business*, 54 (4), 513-533.

Henzel, S. & Wollmershäuser, T. (2005). Quantifying Inflation Expectations with Carlson-Parkin Method: A survey-based determination of the just noticeable difference. *Journal of Business Cycle Measurement and Analysis*, 2005 (3), 321-319.

Henzel, S. & Wollmershäuser, T. (2008). The New Keynesian Phillips curve and the role of expectations: Evidence from the CESifo World Economic Survey. *Economic Modelling*, 25 (5), 811-832.

Hutson, M., Joutz, F. & Stekler, H. (2014) Interpreting and evaluating CESIfo's World Economic Survey directional forecasts. *Economic Modeling*, 38, 6-11.

Jordà, Ò. (2014). Assessing the historical role of credit: Business cycles, financial crises and the legacy of Charles S. Pierce. *International Journal of Forecasting*, 30 (3), 729-740.

Jordà, Ò., Schularick, M. & Taylor, A.M. (2011). Financial crises, credit booms and external liabilities: 140 years of lessons. *IMF Economic Review*, 59(2), 340-378.

Joutz, F. & Stekler, H. (2000). An evaluation of the predictions of the Federal Reserve. *International Journal of Forecasting*, 16 (1), 17-38.

Kamin S.B., Schindler, J. & Samuel, S. (2007). The contribution of domestic and external factors to Latin American devaluation crises: An early warning systems approach, *International Journal of Finance and Economics*, 12 (3), 317-336.

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

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

Kaminsky G.L. (1999). Currency and banking crises: the early warnings of distress. IMF working paper, 178, 1-38.

Kaminsky G. L. (2003). *Varieties of currency crises*. NBER Working chapter 10193. Cambridge, MA.: National Bureau of Economic Research.

Kaminsky G. L. (2006). Currency crises: Are they all the same? *Journal of International Money and Finance*, 25 (3), 503-527.

Krzanowsky, W.J. & Hand, D.J. (2009). *ROC curves for continuous data*. Boca Raton, FL: Chapman & Hall/CRC.

Kudymova, E., Plenk, J. & Wohlrabe, K (2013). Ifo World Economic Survey and the Business Cycle in selected countries. *CESIfo Forum*, 14 (4), 51-57.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143 (1), 29-36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148 (3), 839–843.

Lahiri, K., & Monokroussos, G. (2013). Nowcasting US GDP: The role of ISM business surveys. *International Journal of Forecasting*, 29 (4), 644-658.

Lahiri, K. & Moore, G.H. (1991). *Leading economic indicators: New approaches and forecasting records*. Cambridge, United Kingdom: Cambridge University Press.

Lahiri, K. & Wang, J.G. (2013). Evaluating probability forecasts for GDP declines using alternative methodologies. *International Journal of Forecasting*, 29 (1), 175-190.

Lahiri, K. & Zhao, Y. (2015). Quantifying survey expectations: A critical review and generalization of the Carlson-Parkin method, *International Journal of Forecasting*, 31 (1), 51-62.

Lamont, O.A. (2002). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior and Organization*, 48 (3), 265-280.

Leitch, G. & Tanner, J. E. (1991). Economic Forecast Evaluation: Profits Versus the Conventional Error Measures. *The American Economic Review*, 81 (3), 580-590.

Ludvigson, S.C. (2004). Consumer confidence and consumer spending. *The Journal of Economic Perspectives*, 18 (2), 29-50.

Ma, G. & Hall, W.J. (1993). Confidence bands for the receiver operating characteristic curves. *Medical decision making*, 13 (3), 191-197.

Merton, R.C. (1981). On market timing and investment performance 1: An equilibrium theory of value for market forecasts. *Journal of Business*, 54 (3), 363-406.

Mincer, J. & Zarnowitz, V. (1969). The evaluation of economic forecasts. In J. Mincer (Ed.), *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance* (3-46). New York, National Bureau of Economic Research.

Muth, J.F. (1961). Rational expectations and the theory of price movements. *Econometrica*, 29 (3), 315-335.

Nardo, M. (2003). The quantification of the qualitative data: a critical assessment. *Journal of*

Economic Surveys, 17 (5), 645-668.

Newey, W.K. & West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703-708.

Business Tendency Surveys: A Handbook (2003). Paris, France: OECD Publishing. Retrived from https://www.oecd.org/std/leading-indicators/31837055.pdf.

Öller, L-E. (1990). Forecasting the business cycle using survey data. *International Journal of Forecasting*, 6 (4), 453-461.

Ottaviani, M. & Sørensen, P.N. (2006). The strategy of professional forecasting. *Journal of Financial Economics*, 81 (2), 441-466.

Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95 (449), 308-311.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York, NY: Oxford University Press.

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

Pesaran, M. H. (1984). Expectations Formations and Macroeconomic Modeling. In P. Malgrange & P-A. Muet (Eds.), *Contemporary Macroeconomic Modeling* (27-55). Oxford, United Kingdom: Basil Blackwell.

Pesaran, M. H. (1987). *The limits to rational expectations*. Oxford, United Kingdom: Basil Blackwell.

Peterson, W. W., & Birdsall, T. G. (1953). *The Theory of Signal Detectability: Part I. The General Theory. Technical Report 13*. Ann Arbor, MI: University of Michigan, Department of Electrical Engineering, Electronic Defense Group.

Romer, C. D. & Romer, D.H. (2000). Federal Reserve information and the behavior of interest rates. *American Economic Review*, 90(3), 429-457.

Schnader, M.H. & Stekler, H.O. (1990). Evaluating Predictions of Change. *The Journal of*

Business, 63 (1), 99-107.

Seiler, C. (2015). On the robustness of balance statistics with respect to nonresponse. *OECD Journal of Business Cycle Measurement and Analysis*, 2014 (2), 1-18.

Schularick, M. & Taylor, A. M. (2012). Credit booms gone bust: monetary policy, leverage cycles and financial crises, 1870–2008. *American Economic Review*, 102(2), 1029–1061.

Sims, C. (2002). The role of models and probabilities in the monetary policy process. *Brookings Chapters on Economic Activity*, 2002 (2), 1-40.

Soberhart, J. & Keenan, S. (2001). Measuring default accurately. *Credit Risk Special Report, Risk*, 14, 31-33.

Stangl, A. (2007A). World Economic Survey. In G. Goldrian (Ed.), *Handbook of Survey-Based Business Cycle Analysis* (57-65). Cheltenham, United Kingdom & Nortmapton, MA: Edward Elgar Publishing Limited.

Stangl, A. (2007B). Der Index für Konjunkturerwartungen des ZEW und des ifo Instituts, WES, sind für Deutschland identisch. *Ifo Schnelldienst*, 60 (03), 55-56.

Stekler, H.O. (1994). Are economic forecasts valuable? *Journal of Forecasting*, 13 (6), 495-505.

Stekler, H.O. (1991). Macroeconomic forecast evaluation techniques. *International Journal of Forecasting*, 7 (3), 375-384.

Stekler, H.O. & Schnader M.H. (1991). Evaluating predictions of change: an application to inflation forecasts. *Applied Financial Economics*, 1 (3), 135-137.

Stekler, H.O. & Symington, H. (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006-2010. *International Journal of Forecasting*, 32 (2), 559-570.

Stekler, H.O. & Ye, T. (2017). Evaluating with a leading indicator: an application – the term spread. *Empirical Economics*, 53 (1), 183-194.

Theil, H. (1952). On the time shape of economic macro variables and the Munich Business test. *Review of the International Statistical Institute*, 20 (2), 105-121.

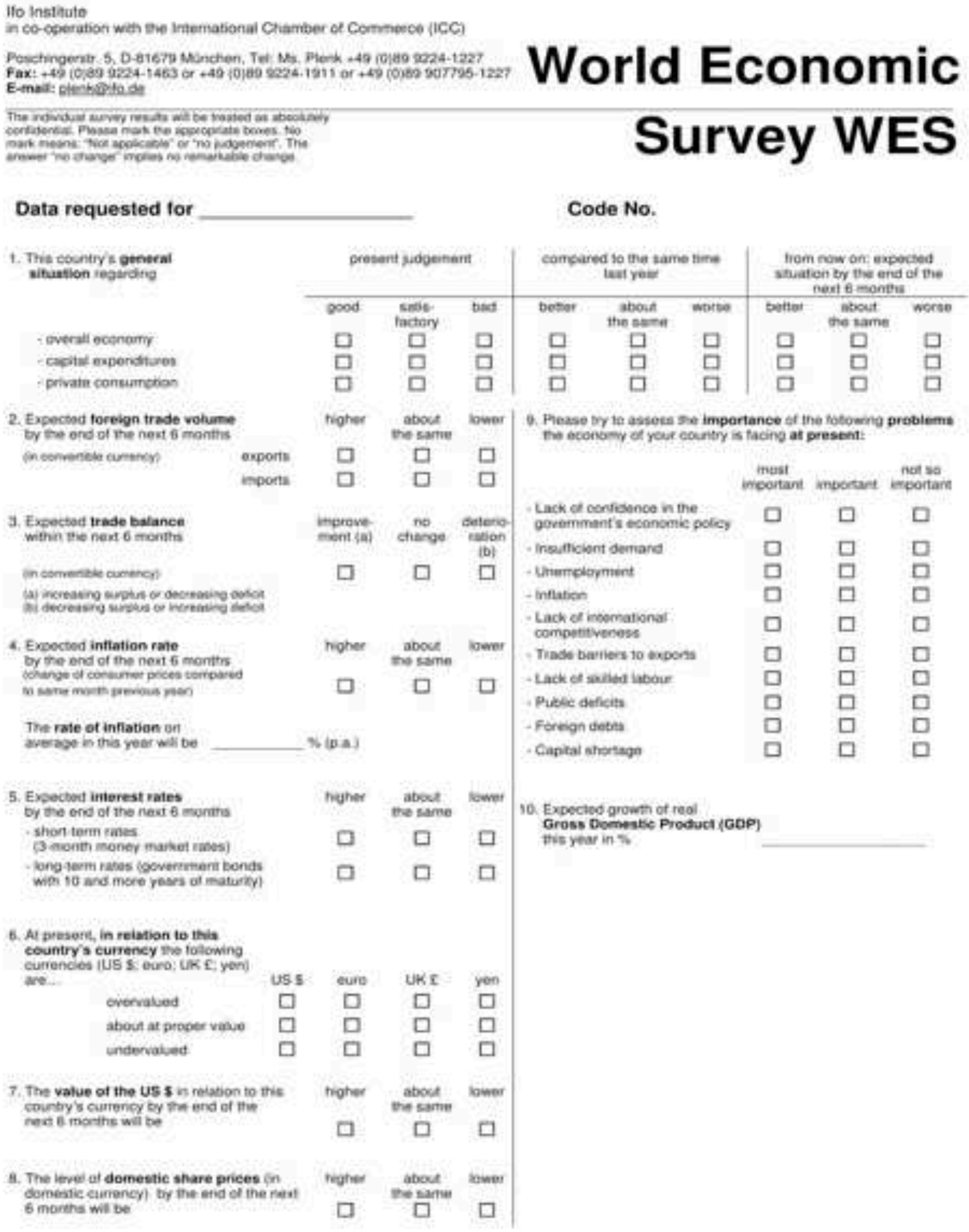
Theil, H. (1961). *Economic Forecasts and Policy*, (2nd ed.). Amsterdam, Netherlands: North

Holland Pub. Co.

Trueman, B. (1994). Analyst Forecasts and Herding Behavior. *The Review of Financial Studies,* 7 (1), 97-124.

Youden, W.J. (1950). Index for rating diagnostic tests. Cancer, 3 (1), 32–35.

Figure A1. WES Questionnaire Sample



Ifo Institute
in co-operation with the International Chamber of Commerce (ICC)

Poschingerstr. 5, D-81679 München, Tel: Ms. Plenk +49 (0)89 9224-1227
Fax: +49 (0)89 9224-1463 or +49 (0)89 9224-1911 or +49 (0)89 907795-1227
E-mail: plenk@ifo.de

World Economic Survey WES

The individual survey results will be treated as absolutely confidential. Please mark the appropriate boxes. No mark means: "Not applicable" or "no judgement". The answer "no change" implies no remarkable change.

Data requested for ____________________ **Code No.**

1. This country's **general situation** regarding	present judgement			compared to the same time last year			from now on: expected situation by the end of the next 6 months		
	good	satis-factory	bad	better	about the same	worse	better	about the same	worse
- overall economy	□	□	□	□	□	□	□	□	□
- capital expenditures	□	□	□	□	□	□	□	□	□
- private consumption	□	□	□	□	□	□	□	□	□

2. Expected **foreign trade volume** by the end of the next 6 months		higher	about the same	lower
(in convertible currency)	exports	□	□	□
	imports	□	□	□

3. Expected **trade balance** within the next 6 months	improve-ment (a)	no change	deterio-ration (b)
(in convertible currency)	□	□	□

(a) increasing surplus or decreasing deficit
(b) decreasing surplus or increasing deficit

4. Expected **inflation rate** by the end of the next 6 months (change of consumer prices compared to same month previous year)	higher	about the same	lower
	□	□	□

The **rate of inflation** on average in this year will be __________ % (p.a.)

5. Expected **interest rates** by the end of the next 6 months	higher	about the same	lower
- short-term rates (3-month money market rates)	□	□	□
- long-term rates (government bonds with 10 and more years of maturity)	□	□	□

6. At present, **in relation to this country's currency** the following currencies (US $; euro; UK £; yen) are...

	US $	euro	UK £	yen
overvalued	□	□	□	□
about at proper value	□	□	□	□
undervalued	□	□	□	□

7. The **value of the US $** in relation to this country's currency by the end of the next 6 months will be	higher	about the same	lower
	□	□	□

8. The level of **domestic share prices** (in domestic currency) by the end of the next 6 months will be	higher	about the same	lower
	□	□	□

9. Please try to assess the **importance** of the following **problems** the economy of your country is facing **at present**:

	most important	important	not so important
- Lack of confidence in the government's economic policy	□	□	□
- Insufficient demand	□	□	□
- Unemployment	□	□	□
- Inflation	□	□	□
- Lack of international competitiveness	□	□	□
- Trade barriers to exports	□	□	□
- Lack of skilled labour	□	□	□
- Public deficits	□	□	□
- Foreign debts	□	□	□
- Capital shortage	□	□	□

10. Expected growth of real **Gross Domestic Product (GDP)** this year in % ____________________

Figure A2. WES - Country coverage with average number of participants (1990-2014)

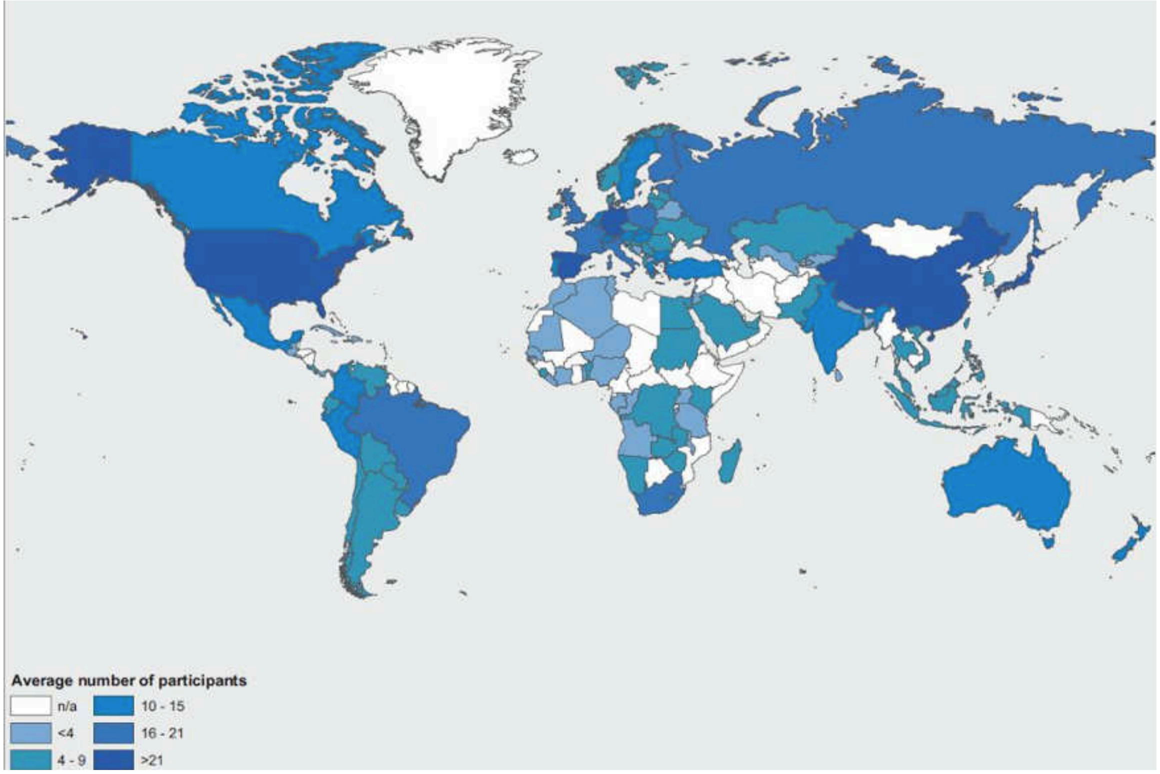
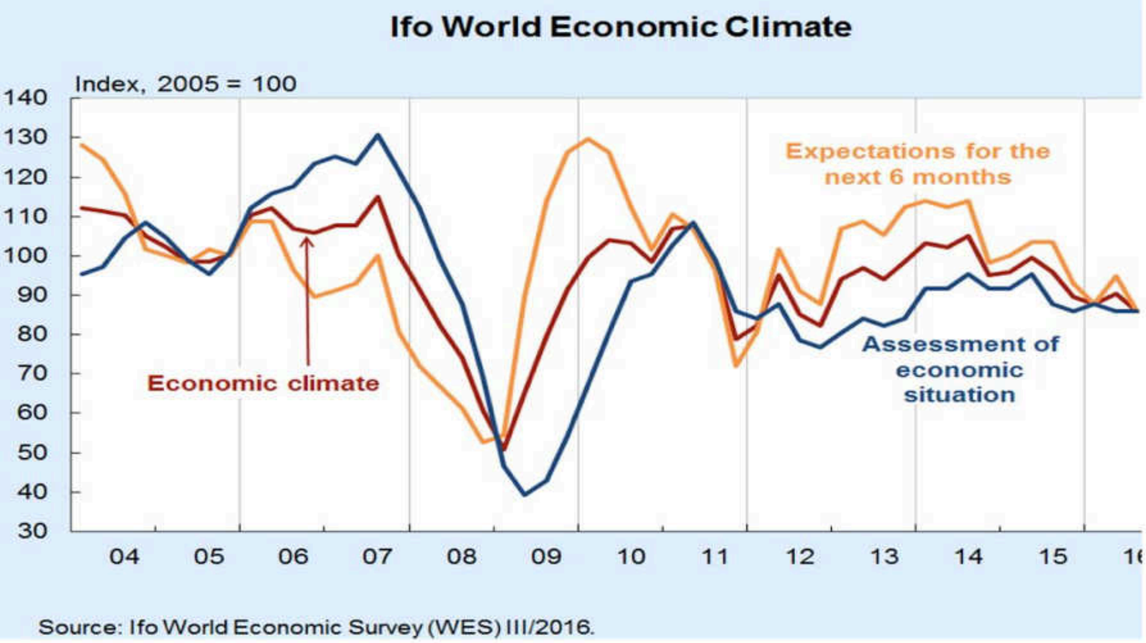
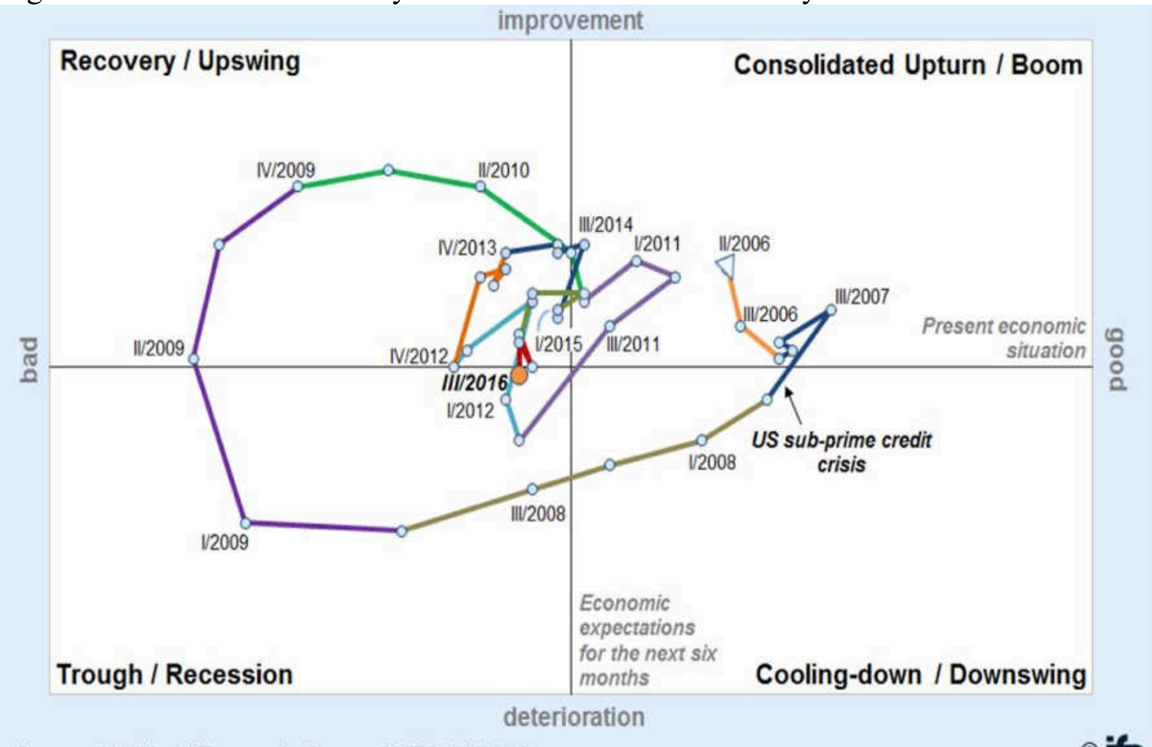


Figure A3. CESifo World EC Indicator



Source: Ifo World Economic Survey (WES) III/2016.

Figure A4. CESifo Business Cycle Clock for a World Economy in 2006-2016



Source: Ifo World Economic Survey (WES) III/2016.

© ifo

Table A1. The WES consensus scores and the corresponding Anderson's Balance Statistics

WES	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
ABS	-1	-0.975	-0.95	-0.925	-0.9	-0.875	-0.85	-0.825	-0.8	-0.775
WES	2	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9
ABS	-0.75	-0.725	-0.7	-0.675	-0.65	-0.625	-0.6	-0.575	-0.55	-0.525
WES	3	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
ABS	-0.5	-0.475	-0.45	-0.425	-0.4	-0.375	-0.35	-0.325	-0.3	-0.275
WES	4	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9
ABS	-0.25	-0.225	-0.2	-0.175	-0.15	-0.125	-0.1	-0.075	-0.05	-0.025
WES	5	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9
ABS	0	0.025	0.05	0.075	0.1	0.125	0.15	0.175	0.2	0.225
WES	6	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9
ABS	0.25	0.275	0.3	0.325	0.35	0.375	0.4	0.425	0.45	0.475
WES	7	7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9
ABS	0.5	0.525	0.55	0.575	0.6	0.625	0.65	0.675	0.7	0.725
WES	8	8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9
ABS	0.75	0.775	0.8	0.825	0.85	0.875	0.9	0.925	0.95	0.975
WES	9									
ABS	1									

Table A2. Distribution of periods in binary categories: 6-month growth rates, 1st vintage

g	N	$N_{y=1}$	$N_{y=0}$	List of periods with Y=0
Real output				
0%	106	98	8	1990q4-1991q2, 2001q3-4, 2008q4-2009q2
1%	106	94	12	1991q3, 2008q1, 2011q2, 2014q2, and all the above
2%	106	64	42	1989q4-1990q3, 1991q4, 1992q1, 1993q2, 1995q2, 1996q1, 2001q1-2, 2003q1-2, 2007q1-2, 2008q2-3, 2009q3, 2010q3, 2011q3, 2012q2-2013q2, 2014q1, 2015q1-2, 2015q4, and all the above
3%	106	45	61	1989q3, 1992q2-3, 1993q3, 1995q3-4, 1998q3, 2002q3-4, 2005q4, 2006q3-4, 2007q4, 2010q4, 2011q1, 2011q4, 2012q1,
Consumption				
0%	106	100	6	1990q4-1991q1, 2008q3-2009q2
1%	106	98	8	1990q2, 1991q4, and all the above
2%	106	77	29	1990q1, 1990q3, 1991q2, 1992q3, 1995q4, 1996q3-4, 2001q3, 2003q1, 2008q1-2, 2009q3, 2010q2, 2011q2-4, 2012q2-4, 2013q3,
3%	106	47	59	1989q4, 1992q1-2, 1993q2, 1994q2-3, 1995q2, 1996q1, 2001q1-2, 2002q2-4, 2003q2, 2004q2, 2005q4, 2006q3, 2007q4, 2009q4, 2010q1, 2010q3, 2012q1, 2013q2, 2013q4, 2014q3, 2015q2.
Investment				
0%	106	76	30	1990q1, 1990q3-1991q2, 1992q1, 2000q4-2002q1, 2003q2, 2005q3, 2006q4-2007q2, 2007q4, 2008q1-2009q3, 2010q4, 2011q1, 2014q1,
1%	106	74	32	2006q3, 2012q3, and all the above
2%	106	68	38	1990q2, 1996q1, 1998q3, 2002q4, 2003q1, 2005q2, and all the above
3%	106	64	42	1993q3, 1995q4, 2012q4, 2015q1, and all the above
Exports				
0%	106	75	31	1989q3, 1990q2, 1991q1, 1992q2, 1993q1, 1993q3, 1994q1, 1995q1, 1998q1-3, 1999q1, 2001q1, 200q4, 2001q1-4, 2002q4, 2003q1-2, 2004q4, 2007q1, 2008q4-2009q2, 2012q3-4, 2014q1,
Imports				
0%	106	81	25	1990q1, 1990q4, 1991q1, 1991q4, 1992q1, 1993q1, 2001q1-4, 2003q1, 2005q2-3, 2006q4, 2007q2, 2008q2–2009q2, 2010q4,

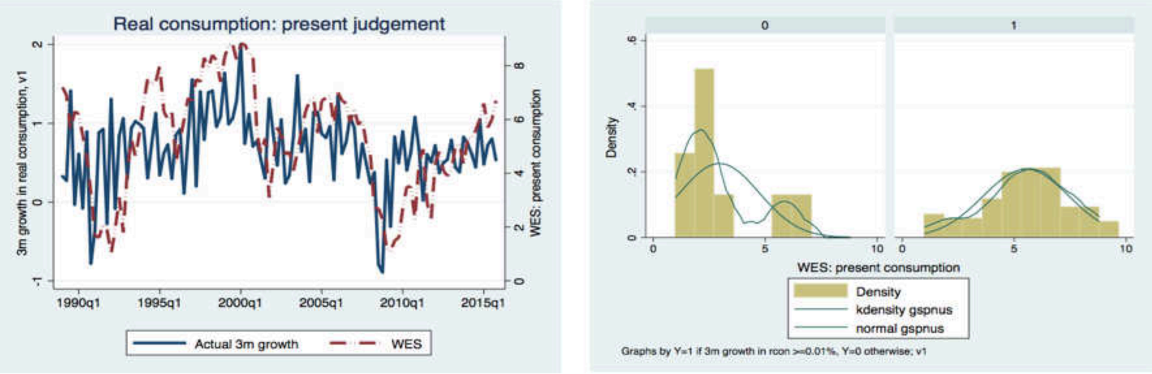
Table A3. Distribution of periods in binary categories: absolute change, 1st vintage

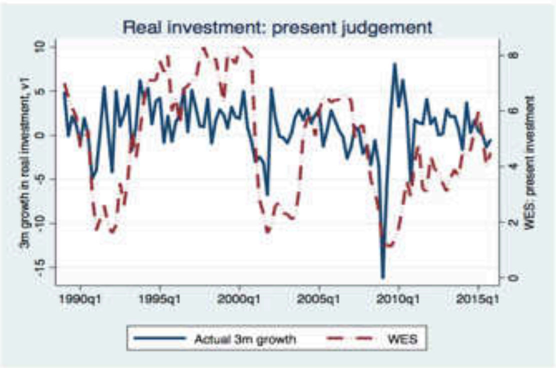
Variable	N	$N_{y=1}$	$N_{y=0}$	List of periods with Y=0
Y=1 if 6-month absolute change is greater or equal 0.01				
Trade balance	106	41	65	1989q3-4, 1990q3, 1991q2-3, 1992q2-1995q3, 1996q2-3, 1997q2-2000q4, 2001q4-2003q3, 2004q1-2005q1, 2005q4-2006q1, 2006q3, 2009q4-2010q3, 2012q1-2, 2013q1-2, 2014q2, 2014q4-2015q2, 2015q4
Inflation	87	46	41	1994q2-4, 1996q1-2, 1997q2-1999q1, 2001q3-2002q4, 2004q1, 2006q4-2007q2, 2009q1-4, 2011q1, 2012q2-2013q2, 2013q4-2014q1, 2014q3, 2015q1-4
Short term interest rate	106	35	71	1989q3-1993q2, 1993q4, 1995q3-1996q3, 1997q1, 1997q4-1998q2, 1998q4-1999q2, 2001q1-2004q1, 2007q1-2010q1, 2010q4-2011q4, 2012q4, 2013q1-

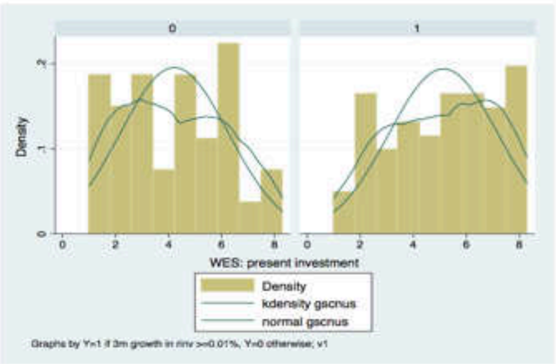
Table A4. Distribution of periods in binary categories: 3-month growth rates, 1st vintage

g	N	$N_{y=1}$	$N_{y=0}$	List of periods with Y=0
				Real output
0%	108	100	8	1990q4, 1991q1, 2001q3, 2008q3-2009q2, 2012q4
1%	108	87	21	1989q4, 1991q2, 1991q4, 1995q2, 1995q4, 2001q2, 2001q4, 2002q4, 2007q4, 2008q1, 2014q1, 2015q1, 2015q4, and all the
2%	108	64	44	1989q2, 1990q2-3, 1992q1-2, 1993q1-2, 1998q2, 2000q4, 2001q1, 2002q2, 2003q1, 2005q4, 2006q3, 2007q1, 2008q2, 2010q3, 2011q1-2, 2012q2-3, 2013q2, 2015q3, and all the above
3%	108	42	66	1989q3, 1990q1, 1991q3, 1992q3, 1993q3, 1994q1, 1995q1, 1996q1, 1996q3, 1997q2, 1999q2, 2000q3, 2003q2, 2006q2, 2010q2, 2011q3-4, 2012q1, 2013q1, 2013q3, 2014q4, 2015q2, and all the
				Consumption
0%	108	99	9	1989q4, 1990q2, 1990q4, 1991q1, 1991q4, 1992q2, 2008q3-4, 2009q2
1%	108	94	14	1996q3, 1997q2, 2002q4, 2008q1, 2011q2, and all the above
2%	108	72	36	1989q1, 1989q2, 1993q1, 1994q2, 1995q1, 1995q4, 2001q3, 2002q2, 2003q1, 2004q2, 2005q4, 2007q2, 2007q4, 2008q2, 2009q4, 2010q2, 2012q2-3, 2013q2-3, 2014q3, 2015q1, and all the above
3%	108	51	57	1990q1, 1994q3, 1995q2-3, 2000q2, 2000q4, 2001q2, 2003q4, 2006q2, 2007q3, 2009q1, 2010q3, 2011q1, 2011q3-2012q1, 2012q4, 2014q1-2, 2015q2, 2015q4, and all the above
				Investment
0%	108	75	33	1989q2, 1990q1, 1990q3-1991q1, 1992q1, 1993q2, 1995q2, 1995q4, 1998q2, 2000q4, 2001q1-4, 2002q3-4, 2003q1, 2005q2, 2006q3-2007q1, 2007q4, 2008q1-2009q2, 2010q4, 2014q1,
1%	108	71	37	2003q2, 2012q3-4, 2015q2, and all the above
2%	108	66	42	1991q2, 1996q4, 2007q3, 2014q3, 2015q1, and all the above
3%	108	62	46	1991q4, 2005q3, 2006q2, 2013q4, and all the above

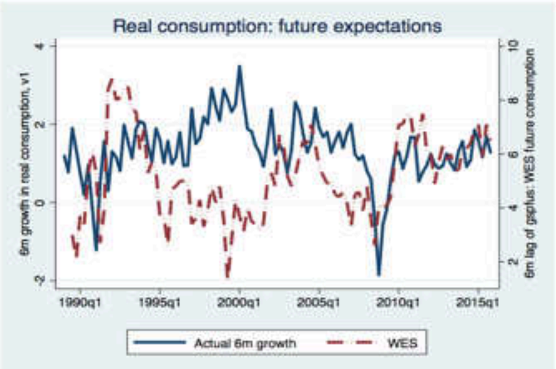
Figure A5. WES expectations: dynamics and scores distributions in Y=1 vs Y=0

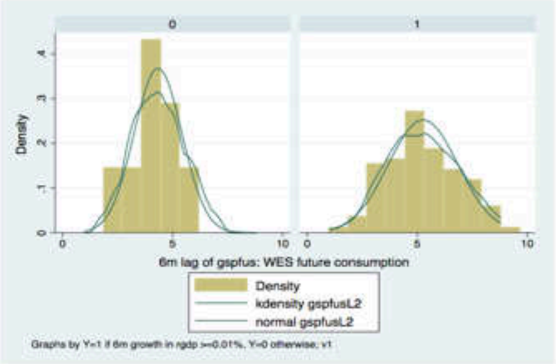



Real investment: present judgement

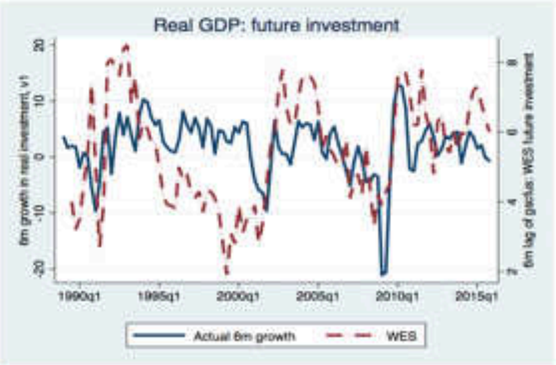


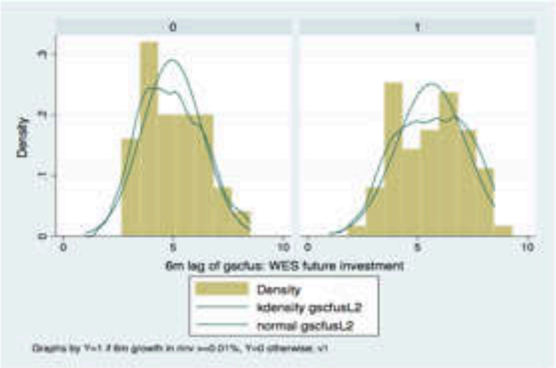
Graphs by Y=1 if 3m growth in rinv >=0.01%, Y=0 otherwise; v1


Real consumption: future expectations

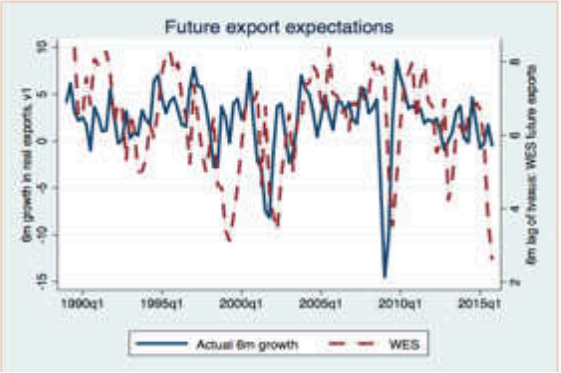


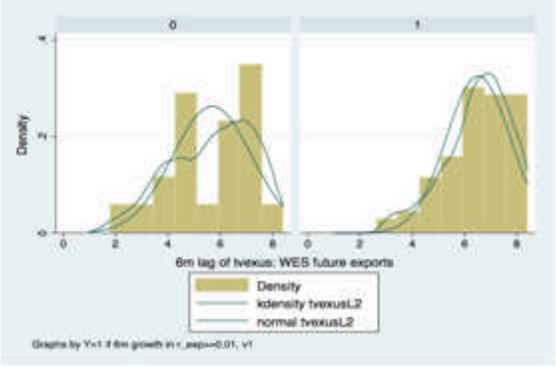
Graphs by Y=1 if 6m growth in rgdp >=0.01%, Y=0 otherwise; v1


Real GDP: future investment



Graphs by Y=1 if 6m growth in rinv >=0.01%, Y=0 otherwise; v1


Future export expectations



Graphs by Y=1 if 6m growth in r_exp>=0.01, v1

Figure A5 (cont.). WES expectations: dynamics and scores distributions in Y=1 vs Y=0

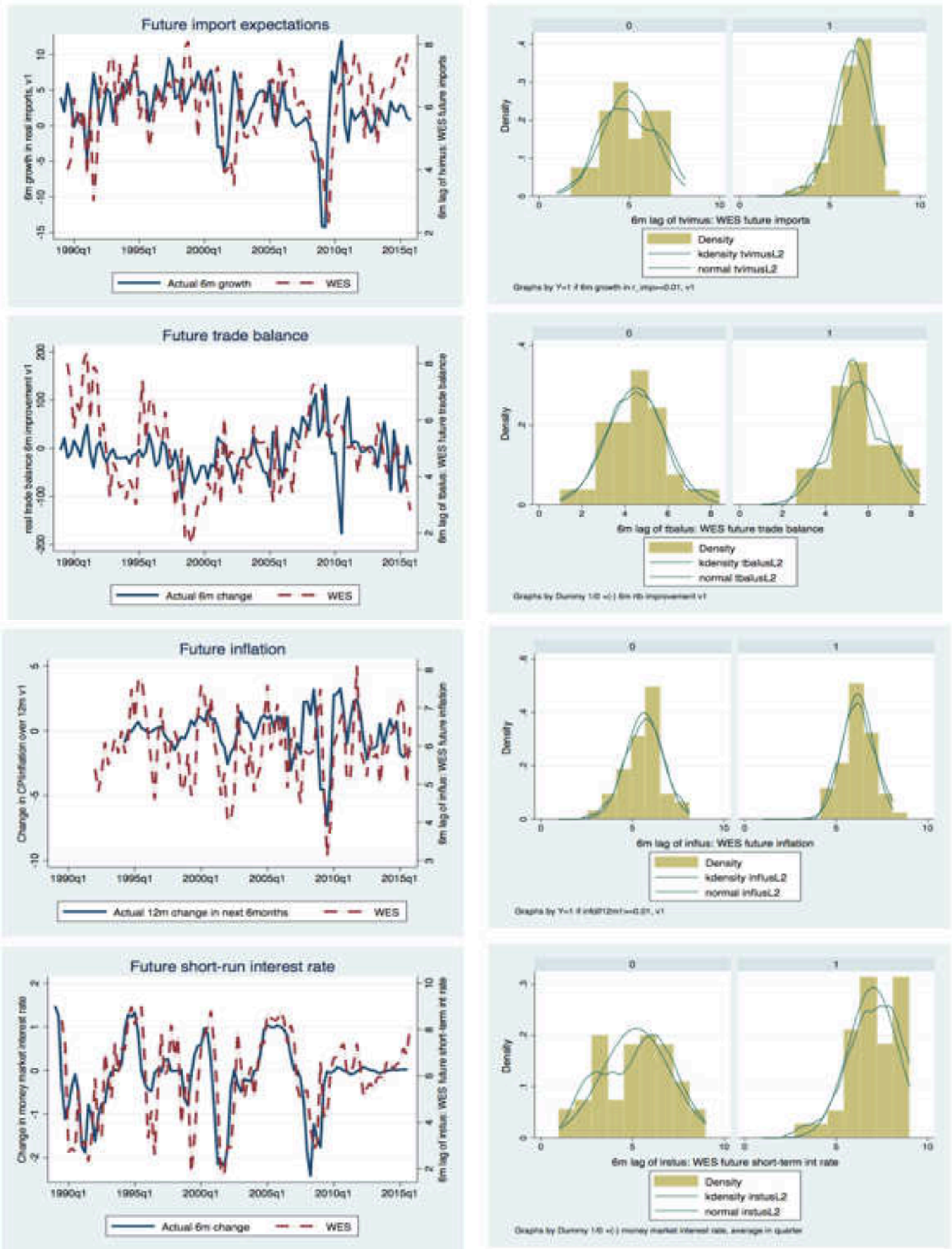


Figure A6. ROC curves for the present judgement about consumption and investment

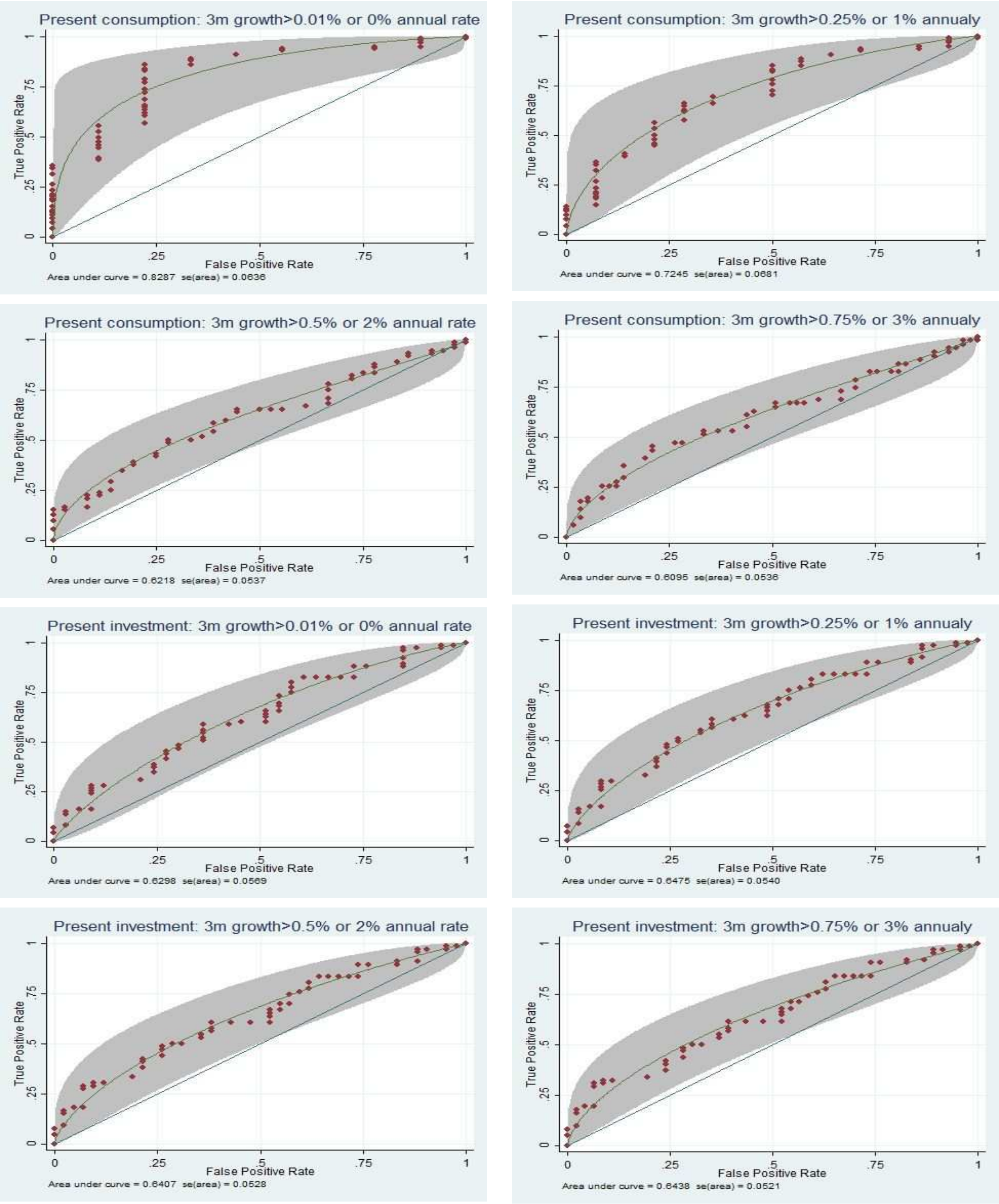


Figure A7. ROC curves for the future expectations about consumption and investment

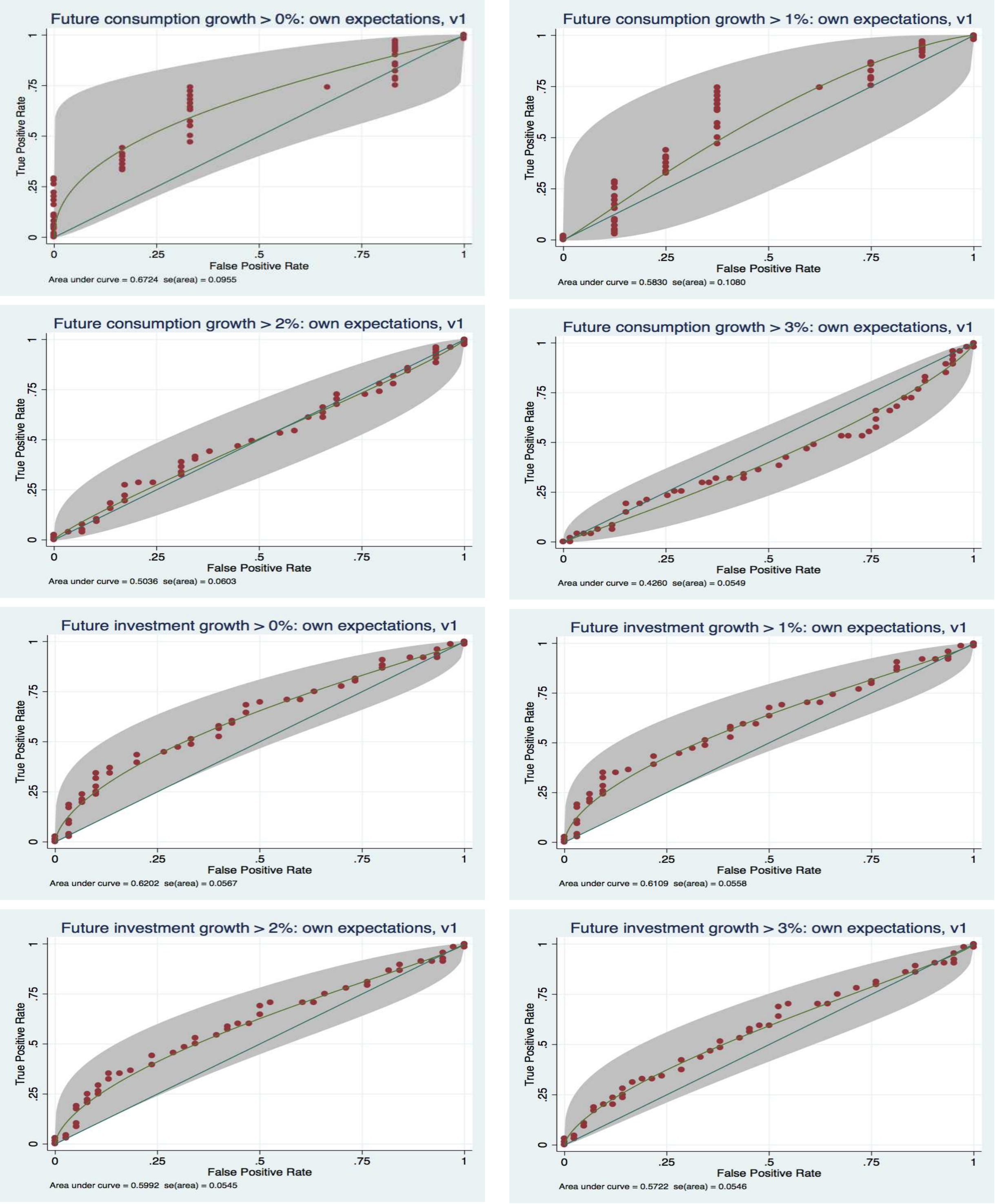


Figure A8. ROC curves for the future expectations about trade, inflation, and interest

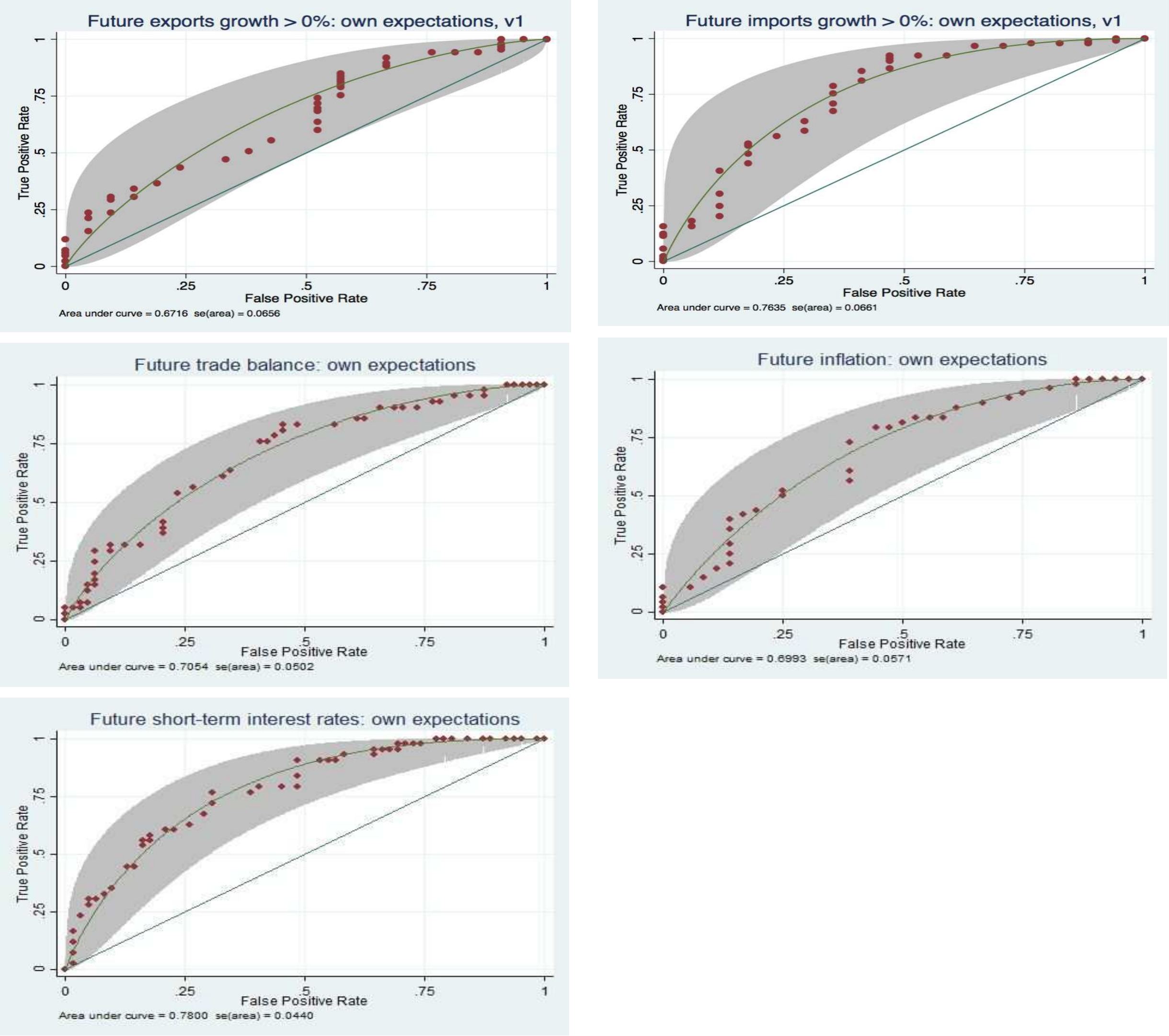


Table A5. Accuracy statistics for the alternative ways to define a threshold[102]

t>=	J	TP	TN	FP	FN	TPR	TNR	FPR	ACC	PD1	PD2	HSS	Chi2	Chi2Y
A. Present judgment about overall economy: own expectations (N1=100, N0=8)														
3.6*	**0.78**	**78**	**8**	**22**	**0**	**0.78**	**1.00**	**0.00**	**0.80**	**0.78**	**1.00**	**0.53**	**22.46**	**18.74**
5	0.62	62	8	38	0	0.62	1.00	0.00	0.65	0.62	1.00	0.24	11.65	9.25
6.4	0.38	38	8	62	0	0.38	1.00	0.00	0.43	0.38	1.00	0.09	4.69	3.17
7.4	0.19	19	8	81	0	0.19	1.00	0.00	0.25	0.19	1.00	0.03	1.84	0.77
B. Present judgment about consumption: own expectations (N1=99, N0=9)														
3.5*	**0.64**	**85**	**7**	**14**	**2**	**0.86**	**0.78**	**0.22**	**0.85**	**0.86**	**0.78**	**0.68**	**21.33**	**17.46**
3.9	0.62	83	7	16	2	0.84	0.78	0.22	0.83	0.84	0.78	0.58	18.69	15.19
5	0.42	64	7	35	2	0.65	0.78	0.22	0.66	0.65	0.78	0.20	6.25	4.59
7.4	0.18	18	9	81	0	0.18	1.00	0.00	0.25	0.18	1.00	0.04	1.96	0.87
C. Present judgment about investment: own expectations (N1=75, N0=33)														
3.2*	**0.22**	**60**	**14**	**15**	**19**	**0.80**	**0.42**	**0.58**	**0.69**	**0.80**	**0.42**	**0.36**	**5.87**	**4.78**
3.5	0.17	56	14	19	19	0.75	0.42	0.58	0.65	0.75	0.42	0.25	3.16	2.40
5	0.20	42	21	33	12	0.56	0.64	0.36	0.58	0.56	0.64	0.24	3.54	2.79
7	0.16	19	30	56	3	0.25	0.91	0.09	0.45	0.25	0.91	0.13	3.73	2.79
D. Future overall economy: own expectations (N1=98, N0=8)														
5	0.25	61	5	37	3	0.62	0.625	0.375	0.62	0.62	0.63	0.10	1.89	1.00
5.7*	**0.39**	**50**	**7**	**48**	**1**	**0.51**	**0.875**	**0.125**	**0.54**	**0.51**	**0.88**	**0.12**	**4.40**	**2.99**
6.3	0.24	36	7	62	1	0.37	0.875	0.125	0.41	0.37	0.88	0.06	1.91	0.99
7.2	0.21	21	8	77	0	0.21	1.00	0.00	0.27	0.21	1.00	0.04	2.14	1.00
E. Future overall economy: EC indicator (N1=98, N0=8)														
5	0.515	75	6	23	2	0.77	0.75	0.250	0.76	0.77	0.75	0.32	9.88	7.46
5.05*	**0.630**	**74**	**7**	**24**	**1**	**0.76**	**0.875**	**0.125**	**0.76**	**0.76**	**0.88**	**0.38**	**14.19**	**11.31**
5.8	0.283	40	7	58	1	0.41	0.875	0.125	0.44	0.41	0.88	0.07	2.50	1.45
6.35	0.102	10	8	88	0	0.10	1.000	0.000	0.17	0.10	1.00	0.02	0.90	0.10
F. Future consumption: own expectations (N1=100, N0=6)														
4.2*	**0.407**	**74**	**4**	**26**	**2**	**0.740**	**0.667**	**0.333**	**0.74**	**0.74**	**0.67**	**0.17**	**4.61**	**2.83**
5	0.217	55	4	45	2	0.550	0.667	0.333	0.56	0.55	0.67	0.05	1.07	0.38
6	0.173	34	5	66	1	0.340	0.833	0.167	0.37	0.34	0.83	0.03	0.77	0.19
6.7	0.200	20	6	80	0	0.200	1.000	0.000	0.25	0.20	1.00	0.03	1.48	0.46
G. Future investment: own expectations (N1=76, N0=30)														
5	0.178	49	16	27	14	0.645	0.533	0.467	0.61	0.65	0.53	0.23	2.83	2.14
6.2	0.195	30	24	46	6	0.395	0.800	0.200	0.51	0.40	0.80	0.18	3.64	2.82
6.5*	**0.242**	**26**	**27**	**50**	**3**	**0.342**	**0.900**	**0.100**	**0.50**	**0.34**	**0.90**	**0.21**	**6.34**	**5.18**
7	0.150	19	27	57	3	0.250	0.900	0.100	0.43	0.25	0.90	0.11	2.94	2.10
H. Future expectations about exports (N1=75, N0=31)														
5	0.22	66	10	9	2	0.88	0.33	0.67	0.77	0.88	0.32	0.34	6.12	4.82
5.2*	**0.28**	**64**	**13**	**11**	**1**	**0.85**	**0.43**	**0.57**	**0.76**	**0.85**	**0.42**	**0.47**	**9.31**	**7.82**
7	0.17	27	25	48	6	0.36	0.81	0.19	0.45	0.36	0.81	0.15	2.83	2.11
7.7	0.16	16	30	59	1	0.21	0.95	0.05	0.36	0.21	0.97	0.13	5.34	4.08
I. Future expectations about imports (N1=81, N0=25)														
4.7*	**0.45**	**75**	**13**	**6**	**1**	**0.92**	**0.53**	**0.47**	**0.86**	**0.93**	**0.52**	**1.02**	**25.8**	**22.8**
5	0.43	73	13	8	1	0.90	0.53	0.47	0.84	0.90	0.52	0.88	21.3	18.7
6.7	0.26	35	21	46	4	0.44	0.82	0.18	0.50	0.43	0.84	0.22	6.08	4.97
7.3	0.10	13	24	68	1	0.16	0.94	0.06	0.28	0.16	0.96	0.07	2.42	1.48

[102] Hereafter a star (*) sign identifies the optimal threshold implied by the ROC curves analysis.

Table A5 (continued). Accuracy statistics for the alternative ways to define a threshold

t>=	J	TP	TN	FP	FN	TPR	TNR	FPR	ACC	PD1	PD2	HSS	Chi2	Chi2Y
K. Future expectations about trade balance (N1=41, N0=65)														
5*	**0.36**	**31**	**39**	**10**	**26**	**0.76**	**0.60**	**0.40**	**0.66**	**0.76**	**0.60**	**0.58**	**12.83**	**11.43**
5.6	0.21	17	52	24	13	0.41	0.80	0.20	0.65	0.41	0.80	0.35	5.71	4.70
6.3	0.23	12	61	29	4	0.29	0.94	0.06	0.69	0.29	0.94	0.38	10.48	8.76
L. Future expectations about inflation (N1=46, N0=41)														
5	0.15	42	10	4	31	0.91	0.24	0.76	0.63	0.91	0.24	0.21	3.95	2.88
5.9*	**0.26**	**31**	**24**	**15**	**17**	**0.67**	**0.59**	**0.41**	**0.64**	**0.67**	**0.59**	**0.43**	**5.89**	**4.89**
6.4	0.16	18	32	28	9	0.38	0.78	0.22	0.55	0.39	0.78	0.24	2.99	2.24
6.9	0.10	10	36	36	5	0.22	0.88	0.12	0.50	0.22	0.88	0.11	1.38	0.80
M. Future expectations about short-term interest rate (N1=35, N0=71)														
5	0.27	33	23	2	48	0.94	0.32	0.68	0.53	0.58	0.38	-0.04	9.26	7.84
5.5*	**0.38**	**32**	**33**	**3**	**38**	**0.91**	**0.46**	**0.54**	**0.61**	**0.63**	**0.43**	**0.09**	**15.02**	**13.38**
7	0.43	22	57	13	14	0.63	0.80	0.20	0.75	0.76	0.52	0.37	19.45	17.58
7.9	0.30	13	66	22	5	0.37	0.93	0.07	0.75	0.84	0.55	0.36	15.06	13.00

Proof of proposition 2:

1. Since t^{NSR} minimizes NSR, it implies that $\frac{FPR(t^{NSR})}{TPR(t^{NSR})} \leq \frac{FPR(t^{ROC})}{TPR(t^{ROC})}$. Let $k \leq 1$ be a coefficient, and rewrite the previous condition as

$$\frac{FPR(t^{NSR})}{TPR(t^{NSR})} = k\frac{FPR(t^{ROC})}{TPR(t^{ROC})} \qquad [7]$$

2. Since t^{ROC} maximizes the J index, it implies that

$$J^{ROC} = TPR(t^{ROC}) - FPR(t^{ROC}) \geq TPR(t^{NSR}) - FPR(t^{NSR}) = J^{NSR} \qquad [8]$$

3. Use [10] to rewrite $FPR(t^{NSR})$ and plug in [11] to obtain

$$J^{NSR} = \frac{TPR(t^{NSR})}{TPR(t^{ROC})}[TPR(t^{ROC}) - k\, FPR(t^{ROC})] \geq \frac{TPR(t^{NSR})}{TPR(t^{ROC})}J^{ROC} \qquad [9]$$

since $[TPR(t^{ROC}) - k\, FPR(t^{ROC})] \geq [TPR(t^{ROC}) - FPR(t^{ROC})] = J^{ROC}$

4. Both [11] and [12] may hold simultaneously if and only if $\frac{TPR(t^{NSR})}{TPR(t^{ROC})} \leq 1$, which implies

$$TPR(t^{NSR}) \leq TPR(t^{ROC}) \qquad [10]$$

5. Both the true and false positive rates are decreasing in t: $\frac{dTPR(t)}{dt} < 0$; $\frac{dFPR(t)}{dt} < 0$[103]. Therefore, $t^{NSR} \geq t^{ROC}$.

[103] See, for example, Krzanovski and Hand (2009) for the reference.

Figure A9. ROC curves for EWI with the out-of-sample value (h=3, 6, 9 m)

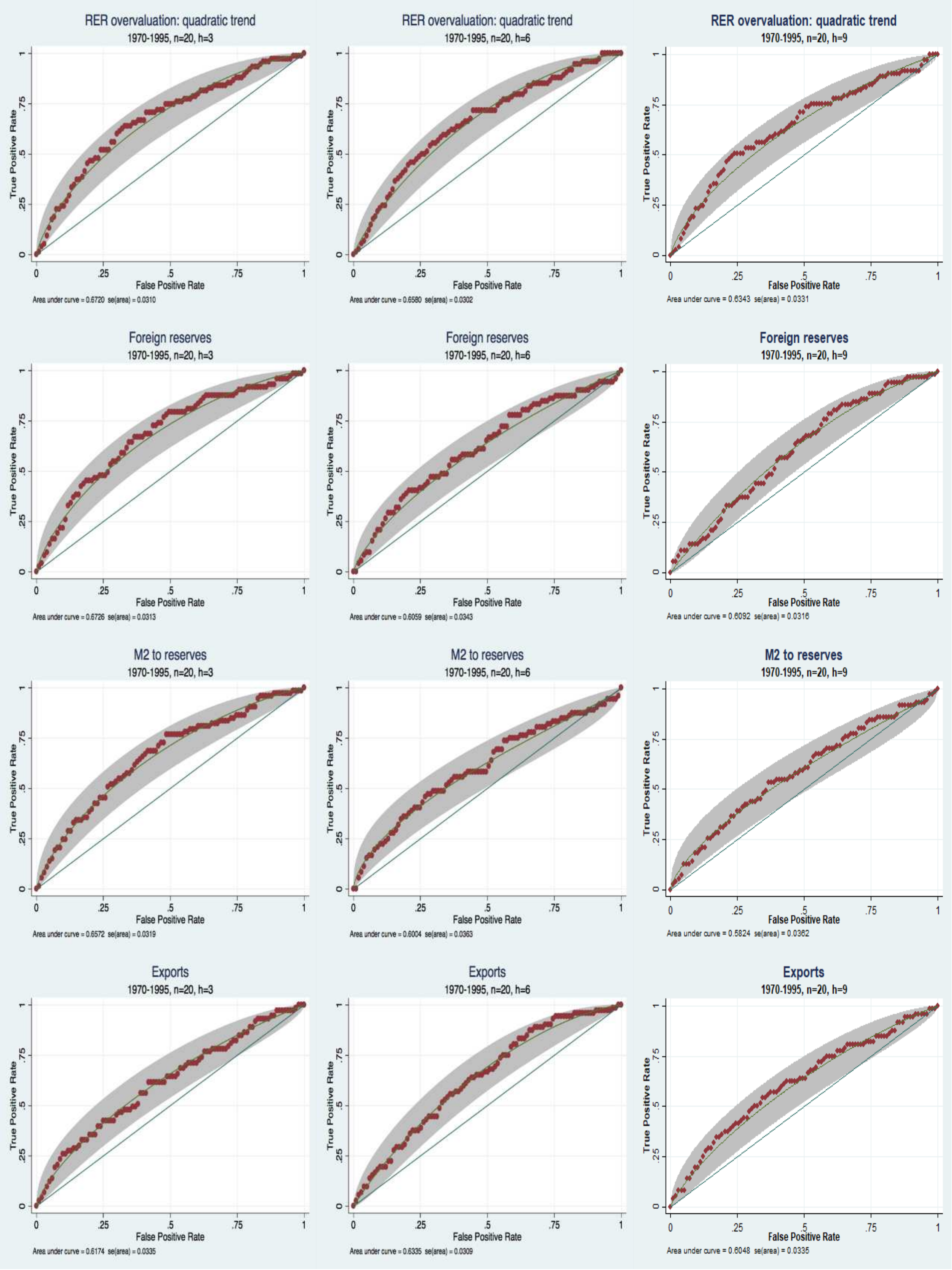


Figure A10. ROC curves for EWI with the out-of-sample value (h=12, 18, 24 m)

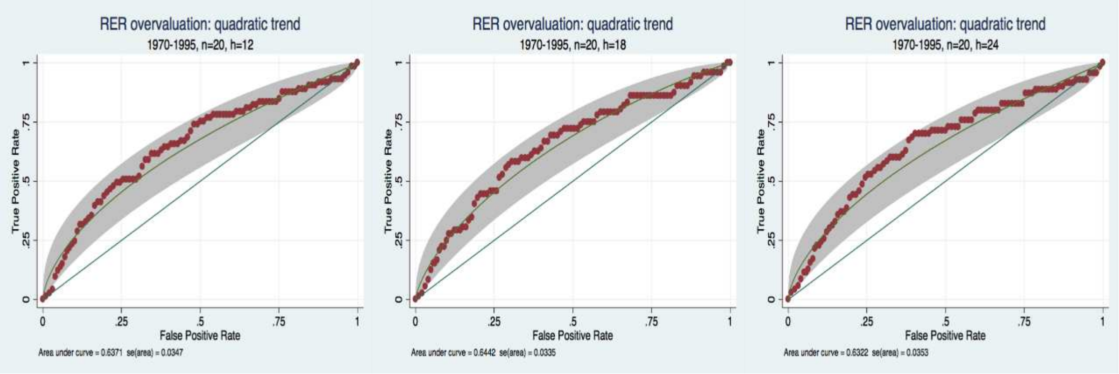


Figure A11. Indicators with in-sample value not significant out-of-sample

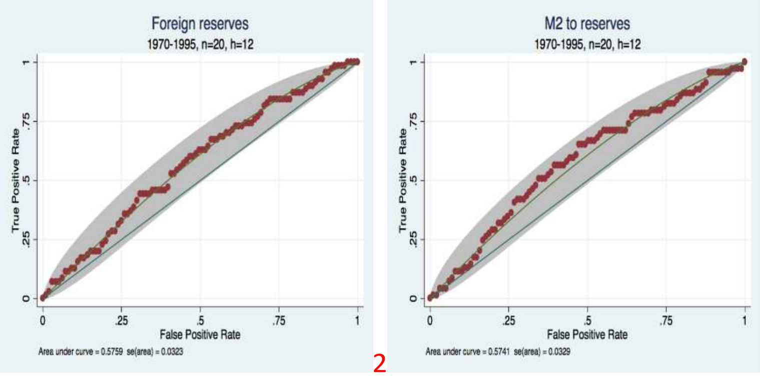


Figure A12. Precision-recall curves for R deviation at h=12, 18, and 24m.

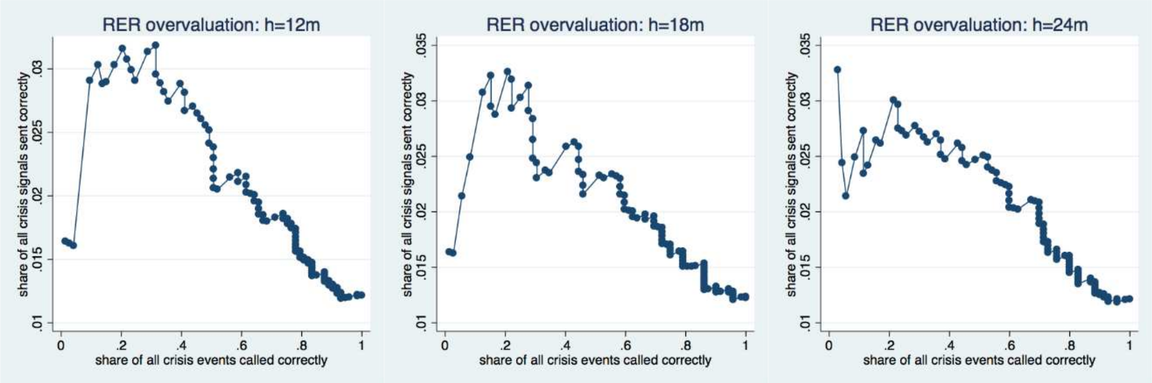


Table A6. Dates of identified currency crisis episodes in the training and test sets

Country	Training set	Test set
Argentina	1970m6, 1975m6, 1981m2, 1982m7, 1986m9, 1989m4, 1990m2	2002m1
Bolivia	1982m11, 1983m11, 1985m9	None
Brazil	1983m2, 1986m11, 1989m7, 1990m11, 1991m10	1999m1
Chile	1971m12, 1972m8, 1973m10, 1974m12, 1976m1, 1982m8, 1984m9	None
Colombia	1983m3, 1985m2	1997m9, 1998m9, 1999m8, 2002m7
Denmark	1971m5, 1973m6, 1979m11, 1993m8	2003m6
Finland	1973m6, 1982m10, 1991m11, 1992m9	Dropped the sample as a EMU member
Indonesia	1978m11, 1983m4, 1986m9	1997m12, 1998m6
Israel	1974m11, 1977m11, 1983m10, 1984m7	None
Malaysia	1975m7	1997m8, 1998m6
Mexico	1976m9, 1982m2, 1982m12, 1994m12	None
Norway	1973m6, 1978m2, 1986m5, 1992m12	1998m1, 1999m7, 2000m11, 2003m2
Peru	1976m6, 1987m10	None
Philippines	1970m2, 1983m10, 1984m6, 1986m2	1997m12
Spain	1976m2, 1977m7, 1982m12, 1992m9, 1993m5	Dropped the sample as a EMU member
Sweden	1977m8, 1981m9, 1982m10, 1992m11	None
Thailand	1978m11, 1981m7, 1984m11	1997m7, 1998m6, 1999m9, 2000m7
Turkey	1970m8, 1980m1, 1994m3	2001m2
Uruguay	1971m12, 1982m10	2002m7
Venezuela	1984m2, 1986m12, 1989m3, 1994m5, 1995m12	2002m2
Total	76	23

Table A7. Relationship between the NSR and precision in the given training sample[104]

NSR	Precision
0.01	0.55
0.02	0.38
0.03	0.29
0.04	0.24
0.05-0.06	0.20-0.17
0.07-0.11	0.15-0.10
0.12-0.15	0.09-0.08
0.16-0.19	0.07-0.06
0.2-0.26	0.05
0.27-0.34	0.04
0.35-0.48	0.03
0.49-0.81	0.02
0.82-1.0	0.012-0.014

[104] Using the Bayes theorem and notations accepted in this chapter, we can express precision as following:

$$Precision = \frac{TPR * p(Y=1)}{TPR * p(Y=1) + FPR * p(Y=0)}$$

After a simple algebraic transformation, one can rewrite this as:

$$Precision = 1\colon \left[1 + \frac{FPR * p(Y=0)}{TPR * p(Y=1)}\right] = \frac{1}{1 + NSR\frac{p(Y=0)}{p(Y=1)}}$$

In the given training sample, we had $p(Y=1) = 1.22\%$ and $p(Y=0) = 98.78\%$ respectively. These numbers were used to calculate the correspondence between the NSR (with 0.01 step) and the resulted precision, which are grouped in the table A3 above.

Table A8. Accuracy statistics for the number of indicator-horizon pairs

T	TP	TN	FN	FP	TPR	TNR	FPR	J	NSR	AC	Precision
RER overvaluation, h=3m											
67	48	4086	27	2019	0.64	0.67	0.33	0.31	0.52	0.67	0.023
68	47	4147	28	1958	0.63	0.68	0.32	0.31	0.51	0.68	0.023
69	46	4208	29	1897	0.61	0.69	0.31	0.30	0.51	0.69	0.024
70	45	4268	30	1837	0.60	0.70	0.30	0.30	0.50	0.70	0.024
71	42	4326	33	1779	0.56	0.71	0.29	0.27	0.52	0.71	0.023
72	42	4389	33	1716	0.56	0.72	0.28	0.28	0.50	0.72	0.024
73	39	4448	36	1657	0.52	0.73	0.27	0.25	0.52	0.73	0.023
74	39	4511	36	1594	0.52	0.74	0.26	0.26	0.50	0.74	0.024
75	39	4573	36	1532	0.52	0.75	0.25	0.27	0.48	0.75	0.025
76	39	4635	36	1470	0.52	0.76	0.24	0.28	0.46	0.76	0.026
77	36	4695	39	1410	0.48	0.77	0.23	0.25	0.48	0.77	0.025
78	36	4757	39	1348	0.48	0.78	0.22	0.26	0.46	0.78	0.026
79	35	4818	40	1287	0.47	0.79	0.21	0.26	0.45	0.79	0.026
80	35	4879	40	1226	0.47	0.80	0.20	0.27	0.43	0.80	0.028
81	34	4939	41	1166	0.45	0.81	0.19	0.26	0.42	0.80	0.028
82	31	4998	44	1107	0.41	0.82	0.18	0.23	0.44	0.81	0.027
83	29	5057	46	1048	0.39	0.83	0.17	0.22	0.44	0.82	0.027
84	28	5119	47	986	0.37	0.84	0.16	0.21	0.43	0.83	0.028
85	28	5181	47	924	0.37	0.85	0.15	0.22	0.41	0.84	0.029
86	26	5240	49	865	0.35	0.86	0.14	0.21	0.41	0.85	0.029
87	25	5300	50	805	0.33	0.87	0.13	0.20	0.40	0.86	0.030
RER overvaluation, h=6m											
55	53	3337	21	2709	0.72	0.55	0.45	0.27	0.63	0.55	0.019
56	50	3396	24	2650	0.68	0.56	0.44	0.24	0.65	0.56	0.019
57	49	3458	25	2588	0.66	0.57	0.43	0.23	0.65	0.57	0.019
58	49	3520	25	2526	0.66	0.58	0.42	0.24	0.63	0.58	0.019
59	48	3582	26	2464	0.65	0.59	0.41	0.24	0.63	0.59	0.019
60	47	3643	27	2403	0.64	0.60	0.40	0.24	0.63	0.60	0.019
61	47	3705	27	2341	0.64	0.61	0.39	0.25	0.61	0.61	0.020
62	46	3767	28	2279	0.62	0.62	0.38	0.24	0.61	0.62	0.020
63	46	3829	28	2217	0.62	0.63	0.37	0.25	0.59	0.63	0.020
64	45	3891	29	2155	0.61	0.64	0.36	0.25	0.59	0.64	0.020
65	44	3952	30	2094	0.59	0.65	0.35	0.25	0.58	0.65	0.021
66	44	4014	30	2032	0.59	0.66	0.34	0.26	0.57	0.66	0.021
67	43	4076	31	1970	0.58	0.67	0.33	0.26	0.56	0.67	0.021
68	42	4136	32	1910	0.57	0.68	0.32	0.25	0.56	0.68	0.022
69	41	4196	33	1850	0.55	0.69	0.31	0.25	0.55	0.69	0.022
70	41	4257	33	1789	0.55	0.70	0.30	0.26	0.53	0.70	0.022
71	40	4317	34	1729	0.54	0.71	0.29	0.25	0.53	0.71	0.023
72	38	4377	36	1669	0.51	0.72	0.28	0.24	0.54	0.72	0.022
73	37	4437	37	1609	0.50	0.73	0.27	0.23	0.53	0.73	0.022
74	37	4500	37	1546	0.50	0.74	0.26	0.24	0.51	0.74	0.023
75	36	4561	38	1485	0.49	0.75	0.25	0.24	0.50	0.75	0.024
76	35	4622	39	1424	0.47	0.76	0.24	0.24	0.50	0.76	0.024

77	34	4684	40	1362	0.46	0.77	0.23	0.23	0.49	0.77	0.024
78	34	4745	40	1301	0.46	0.78	0.22	0.24	0.47	0.78	0.025
79	33	4806	41	1240	0.45	0.79	0.21	0.24	0.46	0.79	0.026
80	31	4864	43	1182	0.42	0.80	0.20	0.22	0.47	0.80	0.026
81	30	4924	44	1122	0.41	0.81	0.19	0.22	0.46	0.81	0.026
82	29	4983	45	1063	0.39	0.82	0.18	0.22	0.45	0.82	0.027
83	28	5043	46	1003	0.38	0.83	0.17	0.21	0.44	0.83	0.027
84	27	5104	47	942	0.36	0.84	0.16	0.21	0.43	0.84	0.028
RER overvaluation, h=9m											
75	38	4596	36	1450	0.51	0.76	0.24	0.27	0.47	0.76	0.026
76	36	4657	38	1389	0.49	0.77	0.23	0.26	0.47	0.77	0.025
77	35	4720	39	1326	0.48	0.78	0.22	0.26	0.46	0.78	0.026
78	34	4779	40	1267	0.47	0.79	0.21	0.26	0.45	0.79	0.026
79	31	4839	43	1207	0.42	0.80	0.20	0.23	0.47	0.80	0.025
80	30	4899	44	1147	0.41	0.81	0.19	0.22	0.46	0.81	0.025
81	29	4957	45	1089	0.40	0.82	0.18	0.22	0.45	0.81	0.026
82	26	5015	48	1031	0.36	0.83	0.17	0.19	0.48	0.82	0.025
83	26	5076	48	970	0.36	0.84	0.16	0.20	0.45	0.83	0.026
84	25	5137	49	909	0.34	0.85	0.15	0.19	0.44	0.84	0.027
85	23	5196	51	850	0.32	0.86	0.14	0.17	0.45	0.85	0.026
86	20	5250	54	796	0.27	0.87	0.13	0.14	0.48	0.86	0.025
87	18	5308	56	738	0.25	0.88	0.12	0.12	0.50	0.87	0.024
88	18	5363	56	683	0.25	0.89	0.11	0.13	0.46	0.88	0.026
89	17	5416	57	630	0.23	0.90	0.10	0.13	0.45	0.89	0.026
90	17	5472	57	574	0.23	0.91	0.10	0.14	0.41	0.90	0.029
Foreign reserves, h=1m											
83	39	4847	34	1034	0.53	0.82	0.18	0.36	0.33	0.82	0.036
84	38	4905	35	976	0.52	0.83	0.17	0.35	0.32	0.83	0.037
85	34	4961	39	920	0.47	0.84	0.16	0.31	0.34	0.84	0.036
86	34	5020	39	861	0.47	0.85	0.15	0.32	0.31	0.85	0.038
87	33	5079	40	802	0.45	0.86	0.14	0.32	0.30	0.86	0.040
88	30	5135	43	746	0.41	0.87	0.13	0.28	0.31	0.87	0.039
89	29	5194	44	687	0.40	0.88	0.12	0.28	0.29	0.88	0.041
90	28	5252	45	629	0.38	0.89	0.11	0.28	0.28	0.89	0.043
Foreign reserves, h=3m											
64	49	3700	24	2141	0.67	0.63	0.37	0.30	0.55	0.63	0.022
65	47	3757	26	2084	0.64	0.64	0.36	0.29	0.55	0.64	0.022
66	47	3817	26	2024	0.64	0.65	0.35	0.30	0.54	0.65	0.023
67	45	3873	28	1968	0.62	0.66	0.34	0.28	0.55	0.66	0.022
68	43	3931	30	1910	0.59	0.67	0.33	0.26	0.56	0.67	0.022
69	43	3989	30	1852	0.59	0.68	0.32	0.27	0.54	0.68	0.023
70	41	4045	32	1796	0.56	0.69	0.31	0.25	0.55	0.69	0.022
71	40	4103	33	1738	0.55	0.70	0.30	0.25	0.54	0.70	0.022
72	40	4163	33	1678	0.55	0.71	0.29	0.26	0.52	0.71	0.023
73	39	4222	34	1619	0.53	0.72	0.28	0.26	0.52	0.72	0.024
74	36	4279	37	1562	0.49	0.73	0.27	0.23	0.54	0.73	0.023
75	35	4336	38	1505	0.48	0.74	0.26	0.22	0.54	0.74	0.023
76	35	4395	38	1446	0.48	0.75	0.25	0.23	0.52	0.75	0.024

77	35	4455	38	1386	0.48	0.76	0.24	0.24	0.49	0.76	0.025
78	34	4512	39	1329	0.47	0.77	0.23	0.24	0.49	0.77	0.025
79	34	4572	39	1269	0.47	0.78	0.22	0.25	0.47	0.78	0.026
80	33	4631	40	1210	0.45	0.79	0.21	0.24	0.46	0.79	0.027
81	33	4691	40	1150	0.45	0.80	0.20	0.26	0.44	0.80	0.028
82	33	4750	40	1091	0.45	0.81	0.19	0.27	0.41	0.81	0.029
83	32	4809	41	1032	0.44	0.82	0.18	0.26	0.40	0.82	0.030
84	31	4866	42	975	0.42	0.83	0.17	0.26	0.39	0.83	0.031
85	28	4919	45	922	0.38	0.84	0.16	0.23	0.41	0.84	0.029
86	28	4976	45	865	0.38	0.85	0.15	0.24	0.39	0.85	0.031
87	27	5035	46	806	0.37	0.86	0.14	0.23	0.37	0.86	0.032
88	25	5092	48	749	0.34	0.87	0.13	0.21	0.37	0.87	0.032
89	24	5151	49	690	0.33	0.88	0.12	0.21	0.36	0.88	0.034
Foreign reserves, h=6m											
80	29	4580	43	1202	0.40	0.79	0.21	0.19	0.52	0.79	0.024
81	28	4637	44	1145	0.39	0.80	0.20	0.19	0.51	0.80	0.024
82	27	4694	45	1088	0.38	0.81	0.19	0.19	0.50	0.81	0.024
83	26	4752	46	1030	0.36	0.82	0.18	0.18	0.49	0.82	0.025
84	23	4806	49	976	0.32	0.83	0.17	0.15	0.53	0.82	0.023
85	23	4861	49	921	0.32	0.84	0.16	0.16	0.50	0.83	0.024
86	21	4912	51	870	0.29	0.85	0.15	0.14	0.52	0.84	0.024
87	21	4972	51	810	0.29	0.86	0.14	0.15	0.48	0.85	0.025
88	21	5031	51	751	0.29	0.87	0.13	0.16	0.45	0.86	0.027
Foreign reserves, h=9m											
53	47	3029	25	2753	0.65	0.52	0.48	0.18	0.73	0.53	0.017
54	46	3084	26	2698	0.64	0.53	0.47	0.17	0.73	0.53	0.017
55	43	3136	29	2646	0.60	0.54	0.46	0.14	0.77	0.54	0.016
56	42	3194	30	2588	0.58	0.55	0.45	0.14	0.77	0.55	0.016
57	41	3252	31	2530	0.57	0.56	0.44	0.13	0.77	0.56	0.016
58	41	3310	31	2472	0.57	0.57	0.43	0.14	0.75	0.57	0.016
59	41	3366	31	2416	0.57	0.58	0.42	0.15	0.73	0.58	0.017
60	41	3424	31	2358	0.57	0.59	0.41	0.16	0.72	0.59	0.017
61	40	3480	32	2302	0.56	0.60	0.40	0.16	0.72	0.60	0.017
62	37	3533	35	2249	0.51	0.61	0.39	0.13	0.76	0.61	0.016
63	35	3590	37	2192	0.49	0.62	0.38	0.11	0.78	0.62	0.016
64	35	3648	37	2134	0.49	0.63	0.37	0.12	0.76	0.63	0.016
65	34	3706	38	2076	0.47	0.64	0.36	0.11	0.76	0.64	0.016
66	32	3763	40	2019	0.44	0.65	0.35	0.10	0.79	0.65	0.016
67	32	3821	40	1961	0.44	0.66	0.34	0.11	0.76	0.66	0.016
68	32	3879	40	1903	0.44	0.67	0.33	0.12	0.74	0.67	0.017
69	32	3936	40	1846	0.44	0.68	0.32	0.13	0.72	0.68	0.017
70	30	3991	42	1791	0.42	0.69	0.31	0.11	0.74	0.69	0.016
71	29	4049	43	1733	0.40	0.70	0.30	0.10	0.74	0.70	0.016
72	27	4106	45	1676	0.38	0.71	0.29	0.09	0.77	0.71	0.016
73	27	4167	45	1615	0.38	0.72	0.28	0.10	0.74	0.72	0.016
74	27	4225	45	1557	0.38	0.73	0.27	0.11	0.72	0.73	0.017
75	27	4283	45	1499	0.38	0.74	0.26	0.12	0.69	0.74	0.018
76	26	4341	46	1441	0.36	0.75	0.25	0.11	0.69	0.75	0.018

77	25	4400	47	1382	0.35	0.76	0.24	0.11	0.69	0.76	0.018
78	24	4457	48	1325	0.33	0.77	0.23	0.10	0.69	0.77	0.018
79	24	4518	48	1264	0.33	0.78	0.22	0.11	0.66	0.78	0.019
80	24	4578	48	1204	0.33	0.79	0.21	0.13	0.62	0.79	0.020
M2/reserves, h=1m											
78	41	4427	31	1293	0.57	0.77	0.23	0.34	0.40	0.77	0.031
79	39	4483	33	1237	0.54	0.78	0.22	0.33	0.40	0.78	0.031
80	36	4538	36	1182	0.50	0.79	0.21	0.29	0.41	0.79	0.030
81	35	4594	37	1126	0.49	0.80	0.20	0.29	0.41	0.80	0.030
82	33	4650	39	1070	0.46	0.81	0.19	0.27	0.41	0.81	0.030
83	33	4708	39	1012	0.46	0.82	0.18	0.28	0.39	0.82	0.032
84	31	4764	41	956	0.43	0.83	0.17	0.26	0.39	0.83	0.031
85	31	4823	41	897	0.43	0.84	0.16	0.27	0.36	0.84	0.033
86	31	4881	41	839	0.43	0.85	0.15	0.28	0.34	0.85	0.036
87	30	4937	42	783	0.42	0.86	0.14	0.28	0.33	0.86	0.037
88	30	4995	42	725	0.42	0.87	0.13	0.29	0.30	0.87	0.040
89	28	5051	44	669	0.39	0.88	0.12	0.27	0.30	0.88	0.040
90	27	5108	45	612	0.38	0.89	0.11	0.27	0.29	0.89	0.042
M2/reserves, h=3m											
52	56	2924	17	2756	0.77	0.51	0.49	0.28	0.63	0.52	0.020
53	53	2978	20	2702	0.73	0.52	0.48	0.25	0.66	0.53	0.019
54	53	3034	20	2646	0.73	0.53	0.47	0.26	0.64	0.54	0.020
55	52	3090	21	2590	0.71	0.54	0.46	0.26	0.64	0.55	0.020
56	50	3144	23	2536	0.68	0.55	0.45	0.24	0.65	0.56	0.019
57	50	3200	23	2480	0.68	0.56	0.44	0.25	0.64	0.56	0.020
58	50	3257	23	2423	0.68	0.57	0.43	0.26	0.62	0.57	0.020
59	50	3314	23	2366	0.68	0.58	0.42	0.27	0.61	0.58	0.021
60	49	3371	24	2309	0.67	0.59	0.41	0.26	0.61	0.59	0.021
61	48	3428	25	2252	0.66	0.60	0.40	0.26	0.60	0.60	0.021
62	47	3485	26	2195	0.64	0.61	0.39	0.26	0.60	0.61	0.021
63	46	3541	27	2139	0.63	0.62	0.38	0.25	0.60	0.62	0.021
64	45	3596	28	2084	0.62	0.63	0.37	0.25	0.60	0.63	0.021
65	43	3652	30	2028	0.59	0.64	0.36	0.23	0.61	0.64	0.021
66	42	3708	31	1972	0.58	0.65	0.35	0.23	0.60	0.65	0.021
67	42	3764	31	1916	0.58	0.66	0.34	0.24	0.59	0.66	0.021
68	41	3822	32	1858	0.56	0.67	0.33	0.23	0.58	0.67	0.022
69	40	3879	33	1801	0.55	0.68	0.32	0.23	0.58	0.68	0.022
70	40	3935	33	1745	0.55	0.69	0.31	0.24	0.56	0.69	0.022
71	39	3992	34	1688	0.53	0.70	0.30	0.24	0.56	0.70	0.023
72	38	4049	35	1631	0.52	0.71	0.29	0.23	0.55	0.71	0.023
73	38	4103	35	1577	0.52	0.72	0.28	0.24	0.53	0.72	0.024
74	37	4160	36	1520	0.51	0.73	0.27	0.24	0.53	0.73	0.024
75	33	4213	40	1467	0.45	0.74	0.26	0.19	0.57	0.74	0.022
76	33	4271	40	1409	0.45	0.75	0.25	0.20	0.55	0.75	0.023
77	33	4330	40	1350	0.45	0.76	0.24	0.21	0.53	0.76	0.024
78	31	4384	42	1296	0.42	0.77	0.23	0.20	0.54	0.77	0.023
79	31	4442	42	1238	0.42	0.78	0.22	0.21	0.51	0.78	0.024
80	29	4498	44	1182	0.40	0.79	0.21	0.19	0.52	0.79	0.024

81	28	4554	45	1126	0.38	0.80	0.20	0.19	0.52	0.80	0.024
82	26	4609	47	1071	0.36	0.81	0.19	0.17	0.53	0.81	0.024
83	26	4667	47	1013	0.36	0.82	0.18	0.18	0.50	0.82	0.025
84	25	4724	48	956	0.34	0.83	0.17	0.17	0.49	0.83	0.025
85	25	4782	48	898	0.34	0.84	0.16	0.18	0.46	0.84	0.027
86	25	4840	48	840	0.34	0.85	0.15	0.19	0.43	0.85	0.029
87	24	4894	49	786	0.33	0.86	0.14	0.19	0.42	0.85	0.030
88	21	4949	52	731	0.29	0.87	0.13	0.16	0.45	0.86	0.028
89	21	5006	52	674	0.29	0.88	0.12	0.17	0.41	0.87	0.030
M2/reserves, h=6m											
73	34	4058	38	1563	0.47	0.72	0.28	0.19	0.59	0.72	0.021
74	33	4115	39	1506	0.46	0.73	0.27	0.19	0.58	0.73	0.021
75	31	4168	41	1453	0.43	0.74	0.26	0.17	0.60	0.74	0.021
76	29	4224	43	1397	0.40	0.75	0.25	0.15	0.62	0.75	0.020
77	29	4282	43	1339	0.40	0.76	0.24	0.16	0.59	0.76	0.021
78	29	4338	43	1283	0.40	0.77	0.23	0.17	0.57	0.77	0.022
79	28	4395	44	1226	0.39	0.78	0.22	0.17	0.56	0.78	0.022
80	27	4452	45	1169	0.38	0.79	0.21	0.17	0.55	0.79	0.023
81	26	4507	46	1114	0.36	0.80	0.20	0.16	0.55	0.80	0.023
82	26	4563	46	1058	0.36	0.81	0.19	0.17	0.52	0.81	0.024
83	25	4618	47	1003	0.35	0.82	0.18	0.17	0.51	0.82	0.024
84	23	4673	49	948	0.32	0.83	0.17	0.15	0.53	0.82	0.024
85	21	4728	51	893	0.29	0.84	0.16	0.13	0.54	0.83	0.023
86	20	4783	52	838	0.28	0.85	0.15	0.13	0.54	0.84	0.023
87	20	4836	52	785	0.28	0.86	0.14	0.14	0.50	0.85	0.025
88	18	4891	54	730	0.25	0.87	0.13	0.12	0.52	0.86	0.024
89	17	4946	55	675	0.24	0.88	0.12	0.12	0.51	0.87	0.025
M2/reserves, h=9m											
64	39	3559	33	2062	0.54	0.63	0.37	0.17	0.69	0.63	0.019
65	35	3614	37	2007	0.49	0.64	0.36	0.14	0.72	0.64	0.017
66	34	3669	38	1952	0.48	0.65	0.35	0.13	0.72	0.65	0.017
67	32	3724	40	1897	0.45	0.66	0.34	0.11	0.75	0.66	0.017
68	32	3782	40	1839	0.45	0.67	0.33	0.12	0.73	0.67	0.017
69	31	3836	41	1785	0.44	0.68	0.32	0.12	0.73	0.68	0.017
70	31	3891	41	1730	0.44	0.69	0.31	0.13	0.70	0.69	0.018
71	31	3947	41	1674	0.44	0.70	0.30	0.14	0.68	0.70	0.018
72	30	4004	42	1617	0.42	0.71	0.29	0.13	0.68	0.71	0.018
73	30	4057	42	1564	0.42	0.72	0.28	0.14	0.66	0.72	0.019
74	29	4114	43	1507	0.41	0.73	0.27	0.14	0.66	0.73	0.019
75	28	4169	44	1452	0.39	0.74	0.26	0.14	0.66	0.74	0.019
76	28	4228	44	1393	0.39	0.75	0.25	0.15	0.63	0.75	0.020
77	26	4284	46	1337	0.37	0.76	0.24	0.13	0.65	0.76	0.019
78	26	4341	46	1280	0.37	0.77	0.23	0.14	0.62	0.77	0.020
79	24	4396	48	1225	0.34	0.78	0.22	0.12	0.64	0.78	0.019
80	23	4454	49	1167	0.32	0.79	0.21	0.12	0.64	0.79	0.019
81	22	4509	50	1112	0.31	0.80	0.20	0.11	0.64	0.80	0.019
82	22	4566	50	1055	0.31	0.81	0.19	0.12	0.61	0.81	0.020
83	20	4619	52	1002	0.28	0.82	0.18	0.10	0.63	0.82	0.020

84	20	4677	52	944	0.28	0.83	0.17	0.11	0.60	0.83	0.021
85	19	4733	53	888	0.27	0.84	0.16	0.11	0.59	0.83	0.021
86	18	4786	54	835	0.25	0.85	0.15	0.11	0.59	0.84	0.021
87	18	4839	54	782	0.25	0.86	0.14	0.11	0.55	0.85	0.023
Exports, h=1m											
77	35	4422	38	1378	0.48	0.76	0.24	0.24	0.50	0.76	0.025
78	33	4479	40	1321	0.45	0.77	0.23	0.22	0.50	0.77	0.024
79	33	4538	40	1262	0.45	0.78	0.22	0.23	0.48	0.78	0.025
80	31	4595	42	1205	0.42	0.79	0.21	0.22	0.49	0.79	0.025
81	29	4652	44	1148	0.40	0.80	0.20	0.20	0.50	0.80	0.025
82	28	4710	45	1090	0.38	0.81	0.19	0.20	0.49	0.81	0.025
83	27	4768	46	1032	0.37	0.82	0.18	0.19	0.48	0.82	0.025
84	27	4827	46	973	0.37	0.83	0.17	0.20	0.45	0.83	0.027
85	27	4886	46	914	0.37	0.84	0.16	0.21	0.43	0.84	0.029
86	26	4944	47	856	0.36	0.85	0.15	0.21	0.41	0.85	0.029
87	26	5002	47	798	0.36	0.86	0.14	0.22	0.39	0.86	0.032
88	26	5061	47	739	0.36	0.87	0.13	0.23	0.36	0.87	0.034
89	24	5118	49	682	0.33	0.88	0.12	0.21	0.36	0.88	0.034
90	22	5175	51	625	0.30	0.89	0.11	0.19	0.36	0.88	0.034
Exports, h=3m											
59	45	3346	28	2414	0.62	0.58	0.42	0.20	0.68	0.58	0.018
60	41	3401	32	2359	0.56	0.59	0.41	0.15	0.73	0.59	0.017
61	41	3458	32	2302	0.56	0.60	0.40	0.16	0.71	0.60	0.017
62	41	3517	32	2243	0.56	0.61	0.39	0.17	0.69	0.61	0.018
63	37	3571	36	2189	0.51	0.62	0.38	0.13	0.75	0.62	0.017
64	36	3629	37	2131	0.49	0.63	0.37	0.12	0.75	0.63	0.017
65	36	3688	37	2072	0.49	0.64	0.36	0.13	0.73	0.64	0.017
66	35	3745	38	2015	0.48	0.65	0.35	0.13	0.73	0.65	0.017
67	35	3803	38	1957	0.48	0.66	0.34	0.14	0.71	0.66	0.018
68	35	3860	38	1900	0.48	0.67	0.33	0.15	0.69	0.67	0.018
69	34	3918	39	1842	0.47	0.68	0.32	0.15	0.69	0.68	0.018
70	34	3976	39	1784	0.47	0.69	0.31	0.16	0.66	0.69	0.019
71	33	4033	40	1727	0.45	0.70	0.30	0.15	0.66	0.70	0.019
72	31	4089	42	1671	0.42	0.71	0.29	0.13	0.68	0.71	0.018
73	31	4146	42	1614	0.42	0.72	0.28	0.14	0.66	0.72	0.019
74	31	4204	42	1556	0.42	0.73	0.27	0.15	0.64	0.73	0.020
75	31	4262	42	1498	0.42	0.74	0.26	0.16	0.61	0.74	0.020
76	31	4321	42	1439	0.42	0.75	0.25	0.17	0.59	0.75	0.021
77	29	4378	44	1382	0.40	0.76	0.24	0.16	0.60	0.76	0.021
78	29	4437	44	1323	0.40	0.77	0.23	0.17	0.58	0.77	0.021
79	26	4493	47	1267	0.36	0.78	0.22	0.14	0.62	0.77	0.020
80	26	4552	47	1208	0.36	0.79	0.21	0.15	0.59	0.78	0.021
81	26	4611	47	1149	0.36	0.80	0.20	0.16	0.56	0.80	0.022
82	24	4668	49	1092	0.33	0.81	0.19	0.14	0.58	0.80	0.022
83	24	4727	49	1033	0.33	0.82	0.18	0.15	0.55	0.81	0.023
84	24	4786	49	974	0.33	0.83	0.17	0.16	0.51	0.82	0.024
85	22	4843	51	917	0.30	0.84	0.16	0.14	0.53	0.83	0.023
86	21	4901	52	859	0.29	0.85	0.15	0.14	0.52	0.84	0.024

87	21	4958	52	802	0.29	0.86	0.14	0.15	0.48	0.85	0.026
88	20	5016	53	744	0.27	0.87	0.13	0.14	0.47	0.86	0.026
89	20	5075	53	685	0.27	0.88	0.12	0.16	0.43	0.87	0.028
90	19	5133	54	627	0.26	0.89	0.11	0.15	0.42	0.88	0.029
Exports, h=6m											
57	46	3184	26	2517	0.64	0.56	0.44	0.20	0.69	0.56	0.018
58	45	3242	27	2459	0.63	0.57	0.43	0.19	0.69	0.57	0.018
59	44	3298	28	2403	0.61	0.58	0.42	0.19	0.69	0.58	0.018
60	43	3354	29	2347	0.60	0.59	0.41	0.19	0.69	0.59	0.018
61	42	3408	30	2293	0.58	0.60	0.40	0.18	0.69	0.60	0.018
62	41	3465	31	2236	0.57	0.61	0.39	0.18	0.69	0.61	0.018
63	41	3523	31	2178	0.57	0.62	0.38	0.19	0.67	0.62	0.018
64	40	3581	32	2120	0.56	0.63	0.37	0.18	0.67	0.63	0.019
65	40	3640	32	2061	0.56	0.64	0.36	0.19	0.65	0.64	0.019
66	39	3696	33	2005	0.54	0.65	0.35	0.19	0.65	0.65	0.019
67	38	3752	34	1949	0.53	0.66	0.34	0.19	0.65	0.66	0.019
68	37	3808	35	1893	0.51	0.67	0.33	0.18	0.65	0.67	0.019
69	35	3864	37	1837	0.49	0.68	0.32	0.16	0.66	0.68	0.019
70	32	3918	40	1783	0.44	0.69	0.31	0.13	0.70	0.68	0.018
71	32	3975	40	1726	0.44	0.70	0.30	0.14	0.68	0.69	0.018
72	32	4033	40	1668	0.44	0.71	0.29	0.15	0.66	0.70	0.019
73	32	4089	40	1612	0.44	0.72	0.28	0.16	0.64	0.71	0.019
74	31	4146	41	1555	0.43	0.73	0.27	0.16	0.63	0.72	0.020
75	30	4203	42	1498	0.42	0.74	0.26	0.15	0.63	0.73	0.020
76	28	4260	44	1441	0.39	0.75	0.25	0.14	0.65	0.74	0.019
77	27	4318	45	1383	0.38	0.76	0.24	0.13	0.65	0.75	0.019
78	27	4377	45	1324	0.38	0.77	0.23	0.14	0.62	0.76	0.020
79	27	4436	45	1265	0.38	0.78	0.22	0.15	0.59	0.77	0.021
80	26	4494	46	1207	0.36	0.79	0.21	0.15	0.59	0.78	0.021
81	24	4551	48	1150	0.33	0.80	0.20	0.13	0.61	0.79	0.020
82	22	4608	50	1093	0.31	0.81	0.19	0.11	0.63	0.80	0.020
83	21	4666	51	1035	0.29	0.82	0.18	0.11	0.62	0.81	0.020
84	21	4725	51	976	0.29	0.83	0.17	0.12	0.59	0.82	0.021
85	21	4784	51	917	0.29	0.84	0.16	0.13	0.55	0.83	0.022
86	20	4842	52	859	0.28	0.85	0.15	0.13	0.54	0.84	0.023
Exports, h=9m											
67	39	3733	33	1968	0.54	0.65	0.35	0.20	0.64	0.65	0.019
68	37	3789	35	1912	0.51	0.66	0.34	0.18	0.65	0.66	0.019
69	36	3846	36	1855	0.50	0.67	0.33	0.17	0.65	0.67	0.019
70	36	3904	36	1797	0.50	0.68	0.32	0.18	0.63	0.68	0.020
71	35	3960	37	1741	0.49	0.69	0.31	0.18	0.63	0.69	0.020
72	34	4018	38	1683	0.47	0.70	0.30	0.18	0.63	0.70	0.020
73	32	4073	40	1628	0.44	0.71	0.29	0.16	0.64	0.71	0.019
74	32	4132	40	1569	0.44	0.72	0.28	0.17	0.62	0.72	0.020
75	31	4189	41	1512	0.43	0.73	0.27	0.17	0.62	0.73	0.020
76	30	4248	42	1453	0.42	0.75	0.25	0.16	0.61	0.74	0.020
77	30	4307	42	1394	0.42	0.76	0.24	0.17	0.59	0.75	0.021
78	29	4366	43	1335	0.40	0.77	0.23	0.17	0.58	0.76	0.021

79	28	4425	44	1276	0.39	0.78	0.22	0.17	0.58	0.77	0.021
80	27	4483	45	1218	0.38	0.79	0.21	0.16	0.57	0.78	0.022
81	27	4543	45	1158	0.38	0.80	0.20	0.17	0.54	0.79	0.023
82	26	4601	46	1100	0.36	0.81	0.19	0.17	0.53	0.80	0.023
83	25	4659	47	1042	0.35	0.82	0.18	0.16	0.53	0.81	0.023
84	25	4719	47	982	0.35	0.83	0.17	0.17	0.50	0.82	0.025
85	23	4776	49	925	0.32	0.84	0.16	0.16	0.51	0.83	0.024
86	21	4834	51	867	0.29	0.85	0.15	0.14	0.52	0.84	0.024
87	21	4891	51	810	0.29	0.86	0.14	0.15	0.49	0.85	0.025
88	20	4950	52	751	0.28	0.87	0.13	0.15	0.47	0.86	0.026