BANK OF GREECE
EUROSYSTEM

Working Paper

Credit risk stress testing for EU15 banks: a model combination approach

George Papadopoulos
Savas Papadopoulos
Thomas Sager

203

JANUARY 2016 PAPERWORKINGPAPERWORKINGPAPERWORKINGPAPERWORK

BANK OF GREECE
Economic Analysis and Research Department – Special Studies Division
21, E. Venizelos Avenue
GR-102 50 Athens
Tel: +30210-320 3610
Fax: +30210-320 2432

www.bankofgreece.gr

Printed in Athens, Greece
at the Bank of Greece Printing Works.

ISSN 1109-6691

CREDIT RISK STRESS TESTING FOR EU15 BANKS: A MODEL COMBINATION APPROACH

George Papadopoulos
Democritus University of Thrace

Savas Papadopoulos
Bank of Greece

Thomas Sager
University of Texas

Abstract

In bank stress tests, the role of a satellite model is to tie bank-specific risk variables to macroeconomic variables that can generate stress. For valid stress tests it is important to develop a comprehensive satellite model that both preserves the sense of known economic relationships and also exhibits high predictive ability. However, it is often difficult to achieve these desiderata in a single satellite model. Multicollinearity of key macro variables and limited data may militate against inclusion of all important stress variables, thus limiting the range of stress scenarios. In order to address this problem we depart from the custom of using a single model as the "true" satellite. Instead, we generate a full space of candidate models that we then screen for reasonable candidates that remain sufficiently rich to cover a wide range of stress scenarios. We then develop composite models by combining the surviving candidate models through weighting. The result is a composite satellite model that includes all the desired macroeconomic variables, reflects the expected relationships with the dependent variable (NPL growth) and exhibits more than 20% lower RMSE compared to a commonly used benchmark model. An illustrative stress testing application shows that this approach can provide policy makers with prudent estimates of credit risk.

Keywords: Financial stability; Macroprudential policy; Non-performing loans; Forecast combination; Predictive modelling

JEL-classifications: C53; E58; G28

Acknowledgments: This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES. Investing in knowledge society through the European Social Fund.

Correspondence:
Savas Papadopoulos
Bank of Greece,
Department of Financial Stability
10250 Athens, Greece
Tel.:0030-210-3205106
Email: sapapa@bankofgreece.gr

1. Introduction

An integral part of any advanced stress testing framework is the satellite model, which maps various macroeconomic scenarios into bank-specific variables that mirror the risk under consideration. That model needs to include an adequate number of important macroeconomic variables to allow for the implementation of a wide variety of scenarios reflecting the impact of the economic environment in a comprehensive manner. For that reason special care needs to be taken to ensure that the relationships of the various macroeconomic variables with the dependent risk are appropriately captured. At a minimum, the modelled relationships should be consistent with economic theory and display high statistical significance. Another essential property of a satellite model is high predictive ability – i.e., providing reliable estimates of bank risk variables under various scenarios. The purpose of this study is to develop such a model for credit risk, assess its forecasting performance and determine its effectiveness in a stress testing application.

Past experience has shown that among the various risks that the banking sector faces, such as liquidity, market, operations, counterparty and credit risk, credit risk is the most important source of insolvency problems for banks (Buncic and Melecky, 2013). Moreover, elevated credit risk can trigger liquidity risk, with cascading consequential risks (Matz and Neu, 2006). Mutually reinforcing feedback loops can lead to a severe financial crisis (Borio, 2010). Spillover to the real economy is a real risk. If deleveraging and a credit crunch develop, very adverse effects on a society's well-being can ensue, including high rates of unemployment and severely deteriorated economic conditions. Therefore it is paramount for supervisors to have a credit risk stress testing framework in order to monitor the resilience of a financial system under possible macroeconomic shocks and assess the impact of shocks.

It is generally accepted in the literature that many macroeconomic and financial factors affect credit risk. Therefore a satellite model should be richly endowed with as many important factors as possible. There are two advantages to casting a wide net for predictive factors. The first is to minimize estimation bias due to possibly omitted variables. The second is to expand the range of scenarios to be examined in a stress testing framework, thus helping policy makers to unveil potential weaknesses and design proper corrective actions. However constructing a general model that includes all possible candidate predictors is not a trivial task. Missing data can limit the

maximum number of independent variables to be used for developing such a model. Even in the case of sufficiently long time series, near multicollinearity of related predictors may distort true relationships and force practitioners to formulate a model using a small subset from the full set of possible predictors.

The practice of using a single non-comprehensive model as if it is the "true" model might result in bias due to omitted variables and as a consequence in possible misestimation of risk. In fact while the importance of stress testing exercises is largely accepted, concerns are being raised about their ability to identify serious vulnerabilities before the onset of the financial crisis (Galati and Moessner, 2013; Haldane, 2009). In a stress testing framework the satellite model is entrusted with the task of linking macro-financial scenarios to bank-level risk parameters. Consequently and understandably, "*Financial institutions have an incentive to choose equations that imply lower provisioning needs and therefore capital requirements conditional on a scenario while conforming to the minimal requirements for economic and statistical soundness.*" (Gross and Población, 2015) It is therefore important that a model enjoy high forecasting performance conditional on a scenario.

The related literature on satellite models of credit risk displays a high degree of heterogeneity as regards the dependent risk variable modelled, methods used and level of aggregation. In a detailed survey of several major supervisory authorities' and central banks' approach to credit risk modelling, Foglia (2009) finds that the credit risk measures that are modelled may be divided into two categories defined by Cihák (2007). The first includes measures of loan portfolio performance such as non-performing loans (NPLs), loan loss provisions (LLPs) or their ratios to total loans, while the second includes measures of corporate or household sector default risk. In the main, the predictor variables are much more homogeneous. The main explanatory variables found to affect credit risk are a small set of macroeconomic indicators, including GDP growth rate, unemployment rate, inflation rate and short and long-term interest rates. The methodology used varies from simple OLS regressions to time-series and non-linear panel data techniques. In a similar study focused on Central and South Eastern European Central Banks (CSEECBs) Melecky and Podpiera (2010) find that the most common general approach for mapping macroeconomic variables to NPLs among CSEECBs is panel or time-series regressions with the same explanatory

variables mentioned in Foglia (2009), supplemented with the exchange rate and certain bank-specific predictors.

Other interesting examples from the large literature on credit risk modelling include the following: Jiménez and Saurina (2006) use annual data and find that GDP growth, real interest rates and the fourth lag of loan growth have a significant impact on Spanish banks' NPLs. Jakubík and Schmieder (2008) develop credit risk models for the Czech Republic's and Germany's corporate and household default rates. For the Czech Republic's corporate sector they find that the impact of real exchange rate and inflation is significant while for the household sector, unemployment and real interest rate affect credit risk. The respective models for Germany included nominal interest rate and GDP for the corporate sector and income and household debt to GDP for the household sector. Louzis et al. (2012) identify GDP growth, unemployment rate and lending rates as important determinants of NPL growth in Greece. Vasquez et al. (2012) using quarterly data construct a credit risk model for the Brazilian banking sector in which the previous value of NPLs, GDP growth rate and its first and second past values affect NPLs significantly. Finally Buncic and Melecky (2013) use a panel of 54 high and middle income countries and construct a macroprudential stress testing framework for credit risk. The satellite model linking macroeconomic scenarios to NPLs is estimated from annual data and includes the previous value of the dependent variable, GDP growth, inflation and the lending rate.

Despite the diversity in the aforementioned academic research and regulatory practice, all use a single equation model with a small number of statistically significant and easily interpretable explanatory variables. A reasonable assumption would be that a similar approach is followed by the banking industry to make conditional forecasts of their credit risk under baseline and adverse scenarios. Although specific information is scarce, "*in the course of the 2014 stress test and the quality assurance process led by the ECB, the documentation provided by the participating banks very clearly confirmed that virtually all institutions operate, indeed, with single equation approaches.*" (Gross and Población, 2015). One noteworthy exception comes from the European Central Bank (Henry et al., 2013; Gross and Población, 2015) where the authors model corporate distance to default (DD) for 18 EU countries using a Bayesian model averaging approach to construct scenario-conditional forecasts. Their illustrative stress test results show that even

models that may meet basic tests of economic and econometric soundness can overoptimistically underestimate risk.

In our study we depart from the use of a single model and employ various weighting schemes inspired by the forecast combination literature to link the NPL growth rate to macroeconomic variables for stress testing purposes. We focus our attention on a sample of 91 banks in EU15 countries during the period of 2006 – 2013. This period allows us to capture the behaviour of credit risk under deteriorated economic conditions. The performance of the models constructed with our approach is assessed through several goodness-of-fit measures. The results show that our models compare more than favourably to their single equation counterparts. In addition, we illustrate their predictive ability conditional on a scenario in a stress test simulation. The results are in line with Gross and Población (2015) and demonstrate that many single equation models, despite the fact that they meet economic plausibility and econometric correctness criteria, yield substantially optimistic predictions conditional on an adverse scenario, thus causing an underestimation of risk and as a consequence a false sense of security. Our model provides adequate estimation of the level of risk and provision needs. Overall our approach presents improved forecasting properties both in- and out-of-sample as well as conditional on a scenario, while retaining a clear economic meaning of the explanatory variables used. Therefore our combination approach can be a very useful tool both for policy makers and other practitioners in the field of credit risk modelling and stress testing.

2. Methodology

Our approach for the development of the models draws from the forecast combination literature. At the core of this methodology is the assumption that no single model is "true." Each single model is, at best, an approximation. Models may be combined by assigning larger or smaller weights to the predictions of individual models according to their performance. The hope is that pooling the collective predictions of a set of models may result in a better prediction than any single model individually – by analogy with the well-known statistical properties that data averages enjoy over a single datum. Of course, for a weighted collective prediction to do substantially better than a single model, each model in the collective should contribute

new information to the collective. That is, the models being averaged should not be substantially the same model.

Forecast combination is closely related to model averaging and indeed some authors (Moral-Benito, 2015) consider it as a predecessor of the Frequentist Model Averaging (FMA) approach. In fact, in linear models, Hansen (2008) demonstrated that the combination of forecasts is equivalent to the forecast produced by the weighted average of the parameter estimates over the different models.

The model averaging literature is composed of two strands: The Bayesian Model Averaging (BMA) and the Frequentist Model Averaging (FMA) approaches. For the former, a very comprehensive review can be found in Hoeting et al. (1999); whereas for the latter, the works of Buckland et al. (1997), Burnham and Anderson (2002) and Claeskens and Hjort (2008) provide excellent references. In a more recent paper Moral-Benito (2015) summarizes the state of the art in both approaches.

It is well-known in the literature (Geweke and Amisano 2011; 2012) that under the BMA and FMA approaches the weight assigned to the best performing model is disproportionately large compared to the rest of the model space, essentially diminishing the contribution of other models. This is due to the fact that these methods operate under the assumption that the model space is complete, meaning that there is a "true" model and the "true" model is included in the model space (Del Negro et al. 2014).

In our study we employ three methods from the forecast combination literature: and in particular the method proposed in the seminal paper of Bates and Granger (1969), as well as the equal weights and the median forecasts, which are found to perform satisfactorily in various empirical applications (Stock and Watson 1998, 2004, 2006; Aiolfi et al. 2010; Bjørnland et al. 2012). A detailed review is provided by Timmermann (2006).

In general, if one has forecasts $f_1, f_2, \ldots, f_m$, a forecast combination is defined as the weighted sum of the individual forecasts:

$$f = \sum_{i=1}^{m} w_i \cdot f_i \,, \tag{1}$$

where w_1, w_2, …, w_m are the corresponding weights and *m* the number of forecasts, or in our case the number of models producing each forecast.

The simplest way for combining forecasts from several models is to take the average of all forecasts – that is assign equal weights to each point forecast in order to create the composite forecast. Another similarly simple way is to take the median forecast, with weight 1 on the median and 0 on all other forecasts. Despite their simplicity, these combining schemes are found to perform equally well or even better than more sophisticated combination methods in several empirical studies and simulations (Palm and Zellner 1992; Stock and Watson 2006; Timmermann 2006). Timmermann (2006) shows that equal weights are indeed optimal when the individual forecast error variances are equal and pair-wise correlations are the same. Since this may not necessarily hold in our case we also implement the weighting scheme of Bates and Granger (1969). In an early influential work Bates and Granger (1969) suggested the construction of a linear combination of forecasts using empirical weights based on out-of-sample forecast variances. The corresponding weights are:

$$w_i = \frac{\widehat{\sigma_i}^{-2}}{\sum_i^m \widehat{\sigma_i}^{-2}}, \tag{2}$$

where $\widehat{\sigma_i}$ is the out-of-sample RMSE of model *i* and *m* the number of models.

Another important finding of the literature (Granger and Jeon, 2004; Aiolfi and Timmermann, 2006; Timmermann, 2006) is that trimming the model space leads to improved performance. This is particularly evident in a situation in which very poorly performing models are combined using the equal weights scheme, as Winkler and Makridakis (1983) point out. Consequently along with the full model space combination, we will also generate combinations of the top 25% and top 50% of individual models in the model space, as ranked by their forecasting performance.

The forecasting performance of each single model, as well as of the combinations, is assessed through six standard goodness-of-fit measures (GoF).

We use three absolute GoF measures to assess the performance of the various models and their combinations directly in the same units as the variable under consideration. These GoF measures are mean absolute error (MAE), median absolute error (MdAE) and root mean squared error (RMSE). Among these measures RMSE

puts a higher penalty on large errors whereas MAE equally weights errors. Therefore large differences between the two could serve as an indication of significant variation in the magnitude of errors. Depending on whether there is a strong preference for avoiding particularly large errors or not, one can use the respective GoF as a guide. In addition MdAE can be used when robustness against possible outliers in the forecast error distribution is of importance.

We use two relative GoF measures to assess performance in percentage terms: Mean absolute percentage error (MAPE) and median absolute percentage error (MdAPE). These GoF measures present the size of the error in an intuitive way, however one should bear in mind that MAPE treats prediction errors in an asymmetric manner by potentially putting "*a heavier penalty on forecasts that exceed the actual than those that are less than the actual*" (Armstrong and Callopy, 1992), since downward errors for positive financial variables are limited to 100%, but upward errors are unlimited.

The last GoF measure is pseudo-R^2 which is estimated as the squared correlation coefficient between the actual and the predicted values (Wooldridge, 2012).[1] The corresponding formulas for each GoF measure are reported in the Appendix.

In order to get more robust results on the performance of each combination scheme we apply the method of k-fold cross-validation, setting k = 5. This procedure involves splitting the sample repeatedly into two uneven subsamples, called the training set and the validation set. The training set retains 80% of the data for model estimation. The training set estimates are then applied to the validation set, where the GoF measures are estimated in the remaining 20% of the data. The procedure is applied five times in a cyclical manner as to ensure that every element appears in the validation set once and only once. Finally the five GoF results are averaged and reported.

3. Data

The analysis is performed on a dataset covering EU15 countries using annual

[1] For OLS regression models, pseudo-R^2 is the actual R^2.

data from the period from 2006 until 2013. This period reflects the behaviour of NPLs under adverse economic conditions since the period includes the financial crisis of 2008 and the sovereign debt crisis of 2010. Thus the results not only demonstrate the feasibility and potential of the proposed methodology for stress testing purposes but can also serve as a benchmark on how NPLs could develop under a severe, real life scenario.

The dependent variable used to model credit risk is the growth rate of the stock of NPLs. The reasons behind choice of this instead of other frequently used variables such as probabilities of default (PDs) or the ratio of NPLs to total loans are two. First, information on PDs is often unavailable. However, if needed, PDs can be approximated by the formula (Hardy and Schmieder, 2013):

$$PD_t = (NPL\ ratio)_{t+1} - (NPL\ ratio)_t + \alpha \cdot (NPL\ ratio)_{t-1}\,. \qquad (3)$$

The parameter α denotes the share of loans that are written-off in period *t-1*. Hardy and Schmieder (2013) note as a rule of thumb, that in the years before a crisis NPLs are fully written off in about two years which is equivalent to an α of 0.5. After the crisis this period increases to three years, therefore parameter α can be set to 0.33. The second reason for choosing to model the growth rate of NPLs is to allow for more flexibility in a stress testing framework. By modelling the numerator of the NPL ratio one can model loans separately, apply several scenarios on them and combine the results to form the respective ratio.

Data for NPLs are collected from Bankscope database for banks that satisfy specific conditions. The sample includes commercial banks that reside in each of the EU15 countries and for which the asset side of their balance sheets exceeds 2 billion EUR as of 2010. In addition, banks are required to fulfil SSM's significance criteria (SSM, 2015). The latter condition increases bank homogeneity and ensures that the significant part of the banking sector of each country is taken into account.

However, the sample banks are subject to events that have a significant impact on NPL growth rate, albeit not directly related to macroeconomic conditions. These events include mergers and acquisitions, or even possible changes in accounting practices that lead to changes in NPLs, unrelated to variations of the general economic environment. Since we are interested in modelling the relationship of NPL growth rate with macroeconomic variables we clean the dataset by keeping only observations that

meet the following additional criteria:

1. NPLs > 0.2 billion EUR
2. -60% < NPL growth rate < 130%
3. 1% < NPL ratio < 40%.

One final condition is that we keep only individual banks that have at least four observations after the application of the previous criteria. The aforementioned criteria are considered sufficient for capturing the behaviour of NPL growth rate under stress while neutralizing the effect of events such as mergers and acquisitions without leading to a grave reduction of the original dataset. The final sample is an unbalanced dataset consisting of 91 banks and 557 observations for NPL growth rate (hereby simply referred as NPL).

The macroeconomic variables are collected from Eurostat and cover a broad part of an economy's activity including GDP (GDP), inflation (INF), unemployment measured in thousand persons (UN), long-term unemployment rate (ULT), household consumption expenditure (HHCE), net disposable income (NDI), compensation of employees (CE) and government debt to GDP (GDEBT) for each one of the EU15 countries.

Following Kalirai and Scheicher (2002) the variables used relate to a country's overall economic activity, price stability, household and government sectors. The first category includes GDP. A decline in GDP signifies a deteriorating economy which in turn can lead to a deterioration of banks' loan books due to borrowers' payment difficulties. Thus a negative relationship with NPLs is expected. The indicator related to price stability is inflation (INF). Being close to 2 percent before the crisis for EU15 countries, falling inflation indicates weakening economic conditions. In addition, declining inflation implies higher real interest rates and as a consequence is likely to result in increased loan defaults. The group of household sector indicators includes net disposable income (NDI), consumption expenditure (HHCE), compensation of employees (CE), overall unemployment (UN) and long-term unemployment rate (ULT). Higher disposable income, employee compensation and consumption relate to a positive economic environment and adequate debt servicing ability for households. Therefore these variables are expected to be inversely related to credit risk. On the contrary, increase of either of unemployment indicators indicates a deterioration of

households' repayment ability and as a result suggests a positive correlation with loan defaults. The state of government sector is represented by the variable of government debt to GDP. Several studies (Reinhart and Rogoff, 2011; Perotti, 1996; Louzis et al., 2012) have detected a positive link between rising government debt and NPLs. In particular two transmission channels have been identified: Government measures of fiscal nature such as tax increases or cuts in spending can have an impact on households' disposable income and lead to an increase in loan defaults (Perotti, 1996). In addition, weakening public finances can affect banks' credibility and give rise to liquidity problems (Reinhart and Rogoff, 2011). This in turn can result in a decrease in banks' lending and thus to refinancing problems for debtors.

In the following analysis all variables are log-differenced (equivalent to growth rates in percent) unless explicitly mentioned otherwise. For the variables that are already expressed in ratios such as government debt to GDP and long-term unemployment rate their first difference is used.

The descriptive statistics of the variables reveal the adverse economic situation that many countries found themselves in and consequently the problems that borrowers and banks had to face during the study period.

Table 1 indicates that the average annual increase in banks' NPLs was nearly 20% over the eight year study period. On its face, this implies considerable and continuing deterioration in loan portfolios. Banks did experience serious problems in their loan books due to financial pressure on their borrowers. At the same time, macroeconomic variables such as GDP, income and consumption either remained mostly stagnant or even decreased, whereas government debt ratio and unemployment on the other hand presented considerable increases. However, these statistics are not differentiated by year. Further, most of these rates of change exhibit high volatility, as reflected in the standard deviations, minima, and maxima. This is an indication of the different degree of severity by which countries and banks experienced the recent economic crisis.

In Table 2, the correlation matrix of NPLs with the macroeconomic variables reveals the underlying relationships which in all cases are statistically significant and have the expected signs.

As expected, there is a negative and statistically significant relationship of NPLs

with variables such as GDP, household consumption expenditure, net disposable income and compensation of employees, the growth of which would indicate a prosperous economy. On the other hand, the relationship is positive with variables such as government debt ratio and unemployment, the growth of which signals that the economy is declining. The results are in line with Kalirai and Scheicher (2002), who present a thorough discussion about the expected relationships of macroeconomic variables with credit risk.

An important observation from Table 2 is that a few pairs of macroeconomic variables exhibit high correlation coefficients exceeding 0.9. This indicates that inclusion of the full set of predictor variables or one of these highly correlated pairs in a single model will probably give rise to multicollinearity issues. In fact, variance inflation factors (VIF) of several predictors in the full model do signal the presence of multicollinearity. The use of the proposed methodology circumvents this issue by combining sufficiently small, econometrically and economically sound models while simultaneously displaying improved performance in various GoF measures compared to the single equation counterparts.

4. Empirical results

For the development of the models and their combinations we implement a multi-stage procedure.

The first stage is the generation of the model space. Its size depends on the maximum number of regressors that can be included in a model, conditional on data availability. Specifically, the number of all possible models having at least one independent variable is 2^q-1, with q being the number of regressors. Our full sample consists of 557 observations, whereas the 5-fold cross-validation includes 445. Following the rule of thumb to have an observation-to-predictor ratio of at least ten to one in order to avoid overfitting (Harrell, 2013), we conclude that all macroeconomic variables and the dynamic term ($\Delta\ln(NPL_{t-1})$) can be used. Hence the total number of models is $m = 2^9$-1 = 511.

The next stage plays a central role in the procedure and involves the estimation of each of the 511 models. The estimation method is decided through the means of standard econometric tests. If the dynamic term is included in the regressors then the

Arellano-Bond (1991) GMM estimator is utilized. This provides a consistent estimator of the dynamic term's coefficient and is used widely in similar studies (Vasquez et al., 2012; Buncic and Melecky, 2013). In the case of static panel data models, Hausman's specification test (Hausman, 1978) is used to inform the selection between fixed or random effects estimators. All models are estimated with bank-clustered standard errors to correct for heteroskedasticity and serial correlation. Before proceeding to the stage of GoF estimation and model combination, the full model space is screened for certain desiderata. In particular, models that do not meet sign or statistical significance criteria are discarded from the model space. For the expected signs we follow economic reasoning as discussed in detail by Kalirai and Scheicher (2002). Thus, we require variables whose increase indicates deterioration of economic conditions such as GDEBT, UN and ULT to have a positive relationship with NPLs while a negative one is expected to hold for the rest. With respect to the significance criteria we demand all variables in a model to be statistically significant at 10% level, having *p*-values less than 0.1. This specific part essentially imitates the procedure an econometrician would follow to build a sound satellite model for a stress testing framework. After these conditions are applied, the size of the model space is significantly reduced and we end up with 22 models forming the *effective* model space. The diagnostic tests reported in Tables A1 and A2 indicate that the models are econometrically sound. The residuals are generally well-behaved without any significant serial correlation as suggested by the AR(2) tests and the exogeneity of the instruments used is supported by Sargan's (1958) test.

Table 3 shows the predictor variables that distinguish the 22 survivor models that constitute the effective model space. The 22 models are numbered in ascending order by their RMSE. A detailed report of the models' coefficients and their performance is given in Tables A1 to A3 in the Appendix. It is clear that models including the dynamic term largely outperform static ones. Another noticeable fact is that the most frequently appearing macroeconomic variable is GDEBT, used in 8 models. The next most frequent variables are GDP and UN in 5 models each and NDI used in 4 models. The rest appear sporadically and mostly exhibit medium or poor performance. Despite the fact that NDI is the third most frequent macroeconomic variable, the performance of the models that include it is consistently above average indicating the high explanatory power it has on NPLs. This is expected because it is

the variable which is most directly associated with the borrowers' ability to repay their loans in this group of macroeconomic variables, also reflected in the high correlation between the two variables (Table 2). Another important observation is that models with GDP also display adequate performance. This is encouraging since GDP is one of the most frequently forecasted, easy to interpret and therefore relevant variables for stress testing purposes (Hardy and Schmieder, 2013). In fact, GDP is included in virtually every satellite model for credit risk and along with UN are the two core variables used in every modern stress testing exercise (Jobst et al. 2013). Thus, the model that includes both of these variables can serve as an appropriate benchmark for comparison with the various combination schemes. However, as shown in Table 3, this model ranks 13th according to its full-sample RMSE. Consequently, the need for using a model that includes GDP and UN can lead to eventually ignoring many models that exhibit better performance.

Now we turn to the construction of the forecast combinations. We essay three weighting combinations applied to the 22 models in the effective model space. The estimated weights for the first two weighting schemes are reported in Table 4. Each of these two weighting schemes is applied to three subsets of the effective model space – respectively, all 22 models, the top 10 models, and the top 5 models, as shown by the three columns of weights under each scheme.

Obviously the weights in the equal weighting scheme are:

$$w_i = \begin{cases} \frac{1}{m^*}, i \leq m^* \\ 0, i > m^* \end{cases} \quad (4)$$

where i is the ranked model index number and m^* is the number of models used in the combination. The Equal Weights ranks by MAE differ only slightly from the Bates-Granger ranks by RMSE. In particular, the change that essentially differentiates the two approaches is that M4 is the 6th best model under the MAE order. The Bates – Granger model weights are very homogeneous, especially in the case of the trimmed subset model spaces. The relative difference of RMSE between the poorest performing model and the best one is around 33%, whereas this figure drops to 14% in the case of the 10 best models, to reach a mere 8% in the most aggressive subset trimming case (Table A3).

The third stage involves estimation of GoF measures for each single model and for model combinations. We note here that model predictions are back-transformed from log scale and all GoF measures are estimated in the levels of NPLs since that is the variable of our main focus. Bates – Granger model weights are obtained using RMSE and Equation 2, while implementation of the equal weights and median combination schemes is straightforward. The performance of the models according to their GoF is used for trimming the model space. We applied two levels of trimming. An aggressive one, discarding 75% of the models, thus keeping the top 5 (≈25%), and a milder one, keeping the 10 best models (≈50%). For the Bates – Granger model combination scheme RMSE is used for model ranking, while for the equal weights and median schemes models are ranked according to MAE. The choice of MAE is made due to its symmetrical treatment of errors both in respect to their magnitude as well as their direction. The results are robust under the use of MdAE while performance was better compared to the use of MAPE and MdAPE.

For the purpose of comparison and demonstration of its appropriateness, the full model is estimated and presented in Equation 5.

$$\begin{aligned}\Delta \ln(NPL) = {} & 0.225^{***} \cdot \Delta \ln(NPL_{t-1}) - 1.530 \cdot \Delta \ln(GDP) + 0.178 \Delta \ln(HHCE) \\ & -0.837^{***} \cdot \Delta \ln(NDI) + 1.721^{*} \cdot \Delta \ln(CE) + 3.133^{**} \cdot \Delta \ln(INF) \\ & +0.619^{***} \cdot \Delta GDEBT + 0.131 \cdot \Delta \ln(UN) + 1.221 \cdot \Delta ULT - 0.318\end{aligned} \quad (5)$$

legend: * p-value<.05; ** p-value<.01; *** p-value<.001

It is evident from Equation 5 that the estimation and use of the full model is problematic in many aspects. Half of the macroeconomic variables included are not statistically significant while three of them (HHCE, CE and INF) do not have the expected signs. Furthermore there is a serious problem of multicollinearity as indicated by the mean predictor VIF, which exceeds the empirical threshold of 4, suggested by Fox (1991).

The exclusive use of linear models allows us to express their combination of forecasts as the weighted average of the parameter estimates over the different models since in this occasion the two approaches are equivalent (Hansen, 2008). This point deserves further discussion. For example, suppose we have the top five models in one

of our weighting schemes and the weights satisfy the regularity conditions $\sum_{i=1}^{5} w_i = 1$, $0 \le w_i \le 1$. Then the five models may be written

$$\begin{array}{ccccccc} Y_i & = & \beta_{10} & +\beta_{11}x_{1i} & +\cdots+ & \beta_{19}x_{9i} & +\varepsilon_{1i} \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_i & = & \beta_{50} & +\beta_{51}x_{1i} & +\cdots+ & \beta_{59}x_{9i} & +\varepsilon_{5i} \end{array} \qquad (6)$$

The subscript *i* indexes observations. All 9 possible predictors appear on the right-hand-side, but the β coefficients may be restricted to zero in order to delete predictors from the models as required. Each model uses the same data set $(Y_i, x_{1i}, x_{2i}, ..., x_{9i})$, *i*=1, 2,…,*n*. So the only differences among the models lie in the pre-set pattern of zeroed-out β coefficients. The zeroing out of various predictors in the effective model space is a key attraction of the combination method. By this means, different models can reflect different stress-testing scenarios among the macroeconomic predictors without engaging the problematic issues (e.g., multicollinearity, coefficient signs, statistical significance) that ensue from trying to force all predictors into one satellite model. The combination model is produced by estimating (6), applying the weights, and summing:

$$\begin{array}{ccccccc} w_1\hat{Y}_i^{(1)} & = & w_1\hat{\beta}_{10} & +w_1\hat{\beta}_{11}x_{1i} & +\cdots+ & w_1\hat{\beta}_{19}x_{9i} \\ \vdots & & \vdots & \vdots & \vdots & \vdots \\ w_5Y_i^{(5)} & = & w_5\hat{\beta}_{50} & +w_5\hat{\beta}_{51}x_{1i} & +\cdots+ & w_5\hat{\beta}_{59}x_{9i} \end{array} \qquad (7)$$

$$\sum_{k=1}^{5} w_k\hat{Y}_i^{(k)} = \left(\sum_{k=1}^{5} w_k\hat{\beta}_{k0}\right) + \left(\sum_{k=1}^{5} w_k\hat{\beta}_{k1}\right)x_{1i} + \cdots + \left(\sum_{k=1}^{5} w_k\hat{\beta}_{k9}\right)x_{9i} \qquad (8)$$

The combined estimate of NPL is shown on the left-hand-side of (8). The right-hand-side of (8) shows that the coefficients of the combination model may be obtained by analogously weighting and combining the coefficients of the 5 single models that (7) comprises. The combination model (8) may be viewed as an alternative estimate of each single equation in (6). If each single equation in (6) were estimated by OLS, and OLS specifications were met, then the coefficient estimates in parentheses in (8) would not be optimal on account of the Gauss-Markov theorem. We should then expect that the RMSE of (8) would be larger than that of any single equation in (6). In fact, we find that the RMSE and other GoF measures for the combination models are generally among the best. One explanation is that (8) will

generally carry more non-zero coefficients than any of the individual equations of (6) (where zeroing restrictions apply). Thus, (8) brings additional explanatory power in the form of variables omitted from (6). In addition, the combination method may spread the risk of model misspecification that may exist in some individual models over a pool of models. Therefore, the combination method may also enjoy some robustness in its applications.

The respective combined models as well as the benchmark model (M13) along with their 5–fold cross validation GoF measures are reported in Table 5. In practice we apply the multi-stage procedure described previously (formation of the effective model space, estimation of the models, estimation of GoF measures, formation of model combinations and estimation of their performance) five times in a cyclical manner as to ensure that every element appears in the validation set once, or equivalently is included in the training set exactly four times.

For the case of the benchmark model all variables are statistically significant and have the correct signs as required by our procedure. Its low mean VIF value of 1.26 indicates that it does not suffer from multicollinearity issues. As mentioned, this model is of special interest since GDP and UN are used regularly in the context of stress testing exercises for the generation and implementation of various scenarios (Hardy and Schmieder, 2013; Jobst et al. 2013).

Several interesting findings are revealed regarding the combined models and their performance. First, the values of the simple average coefficients for all- and 10-model combinations are very similar to corresponding Bates – Granger coefficients. This is expected because the models use the same variables and the weights are similar. Second, in line with the related literature (Winkler and Makridakis, 1983; Granger and Jeon, 2004; Stock and Watson, 2004; Aiolfi and Favero, 2005; Timmermann, 2006), we find that trimming the model space improves performance. This holds for all three combination schemes and for every GoF measure used in this study. Another empirical finding, also observed and explained in related studies (Stock and Watson, 2003; Timmermann, 2006), is the fact that simple combination methods perform equally or even better than more sophisticated ones that employ differential weighting. The comparison of model combinations to the benchmark single-equation model shows that the former present significantly improved performance in every GoF measure. Specifically, all 5-model combinations perform

similarly to the 5-model median combination that has the best overall performance compared to the benchmark. It has above 20% lower MAE, MdAE and RMSE and about 3% and 4.5% lower MAPE and MdAPE respectively. The differences in pseudo-R^2 values, although in favour of the combined models, are not very large.

A more complete picture of the performance of each individual model as well as their combinations is given in Figures 1 and 2 and Table A4. The single-equation models are denoted as M1 – M22. Model combinations use the names *A* for average, *Md* for median and *BG* for Bates-Granger followed by a number which indicates the number of models used to form the respective combination.

Figs. 1 and 2 and Table A4 show that the trimmed, simple model combination schemes consistently dominate their single-equation counterparts in the model space. The 5-model median combination (Md5) ranks first in the GoF measures of MdAPE, RMSE and pseudo-R^2, second but still outperforming all individual models in MdAE and performs equally well to the second individual model in the cases of MAE and MAPE. The benchmark model including GDP and UN (M13) exhibits generally poor performance always occupying places in the lowest half of the performance range and even being the 5th worst in the GoF measures of MdAE and MdAPE. A pattern that emerges regarding the performance of the model combinations is that the 5-model combinations rank first, followed closely by their 10-model counterparts with the all-model ones lying in the middle of the performance scale for all GoF measures.

5. Stress testing application

The previous analysis is reassuring as regards to the forecasting performance of model combinations compared to their single-equation counterparts. In this part we examine the application of model combinations in a stress testing framework in order to test their operational properties.

Stress scenarios can be generated from macroeconomic models, historical events, expert judgment or a combination of these (Jones et al., 2004; Cihák, 2007; Isogai, 2009). For the needs of this illustrative application, we use the historical approach. Specifically, we select the most adverse 1% from the distribution of each macroeconomic variable per country in the period 2006 – 2013. This translates to the bottom 1% percentile of the variables whose growth is associated with a growing

economy and the top 1% percentile otherwise. This essentially captures the stressed economic environment that many countries and banks faced since the outbreak of the financial crisis of 2008 and thus poses a realistic, internally consistent and sufficiently negative scenario.

We use the estimated models to make a forecast conditional on the scenarios from Table 6 on a bank by bank basis. Next we create the forecast combinations and plot the results for 8 representative banks in Figure 3 and Figure 4. The last three years of historical data as well as the scenario conditional forecasts of the individual models and their combinations are plotted in order to get a clearer picture of their forecasting performance.

In Fig. 3 are presented the historical values of NPLs up to 2013 and the conditional forecasts in 2014 for 4 banks from the non-stressed countries while in Fig. 4 from the stressed countries. The shaded area marks the range of the forecasts conditional on the adverse scenario from the single-equation models.

In fact, the same pattern as in Figs 3 and 4 is exhibited from every bank in our sample. The conditional forecasts from the individual models display a considerable divergence in their results, ranging from very mild to aggressive, while the combination schemes appear in the middle of the forecast space as expected.

One can clearly see that the two modelling approaches can have significantly different implications for the future path of the stock of NPLs conditional on the adverse evolution of the macroeconomic variables and consequently on the provision needs of the banks and the stability of the financial system in general. Although every single-equation model meets the criterion of basic economic plausibility and is econometrically sound, there are individual models that yield very mild forecasts conditional on the adverse macroeconomic scenario. On the other hand, the forecasts of virtually all combination schemes appear in the middle of the forecast space as expected. This indicates that all combination methods employed in this analysis are just as appropriate for stress testing purposes, therefore potential users can choose the combination method that is more suitable for them on the basis of performance or even computational complexity. A comparison between the most benign individual predictions conditional on the stress scenario and those from the model combination

schemes can assist in quantifying the magnitude of the differences and their implications. The results from all 91 banks in our sample are presented in Table 7.

From Table 7 it is evident that the differences are important. On average, model combination conditional forecasts are over 3 billion EUR larger than those of their mildest counterparts. The median difference fluctuates around 800 million EUR depending on the method, which is a smaller, but still significant amount. Even the minimum values are positive, although very low at around 40 thousand EUR. The most striking observation is that the maximum difference can be as large as 24 billion EUR, a figure which could have serious implications for a bank's solvency and capital needs. The distribution of the previous differences in Figure 5 gives a clearer picture. Because of the very similar pattern exhibited by all combination methods as presented in Table 7, we plot the results for the 5-model median combination as a representative case for every combination method.

In Fig. 5 one can see the scenario-conditional forecast differences of 5-model median combination compared to the minimum obtained from the individual models. For nearly 60% of the cases these differences are positive but kept below 2.5 billion EUR. There is however a significant tail in the distribution, with differences appearing nearly uninterrupted until the very high figures of above 20 billion EUR. Therefore the proposed model combination approach can prove a useful tool for a more prudent estimation of risk.

Apart from yielding more conservative numerical estimates, the combination approach can assist credit risk modelling practitioners in other practical ways by serving as an objective benchmark to assess a model’s ability to produce sufficiently aggressive forecasts conditional on a stress scenario. From the supervisors' point of view it could serve as a threshold which the supervised financial institutions' models should pass in a stress testing exercise as argued by Gross and Población (2015). In addition, it can inform risk management of private financial intermediaries in a more robust way about the level and potential implications of the assumed risk. Therefore, in both cases, it helps establishing a greater sense of security about the stability of the financial system and the risks associated with it.

6. Conclusions

The paper proposed an alternative approach for modelling credit risk and implementing baseline and adverse scenarios within a stress testing framework through the use of satellite models. Its performance was studied and compared to the currently used approach while a simple exemplary application demonstrated its potential. Departing from the standard way of using a single model which is often studied in the literature and used by policy makers and the industry, the analysis showed that model combination can consistently outperform its individual counterparts in terms of forecasting ability - oftentimes by a significant margin. The empirical finding that simple combination schemes such as the average and median are found to perform equally well or better than more sophisticated weighting schemes is consistent with other studies. In addition, model space trimming is found to improve the performance of every combination method. Therefore the model space should be carefully designed, either including only adequately performing models or being large enough to allow for trimming yet leaving a significant number of models for combining their forecasts.

The proposed method combines meaningful and powerful models and brings order to model space. The paper provides twenty two models as a basis or a set of influential points for model space. Stress testers could also generate a space of interesting and appropriate models for scenario analysis. For instance, the paper provides the best meaningful model for prediction purposes, the best model that includes GDP or unemployment or both, and several other useful models.

The current study also shows that the variables net disposable income and the government debt to GDP are strong predictors for NPL growth for EU15 banks. These variables could be used effectively to improve the predictive ability despite the existence of multicollinearity.

Table 1: Descriptive statistics of NPLs and macroeconomic variables

Variable, [%]	Mean	Standard deviation	Median	Min	Max
Δln(NPL)	19.250	29.520	13.970	-57.450	129.900
Δln(GDP)	0.980	4.779	1.891	-13.670	17.550
Δln(HHCE)	1.110	4.551	2.108	-13.480	16.170
Δln(NDI)	-0.139	10.690	3.170	-45.150	25.550
Δln(CE)	1.147	4.582	1.444	-13.260	13.810
Δln(INF)	2.101	1.145	2.195	-1.667	4.591
ΔGDEBT	4.830	6.378	3.499	-14.440	25.330
Δln(UN)	5.451	13.890	3.643	-17.900	60.740
ΔULT	0.370	0.950	0.200	-1.200	5.700

Table 2: The correlation matrix of NPLs, its lag and macroeconomic variables

	Δln(NPL)	$\Delta\ln(NPL_{t-1})$	Δln(GDP)	Δln(HHCE)	Δln(NDI)	Δln(CE)	Δln(INF)	ΔGDEBT	Δln(UN)	ΔULT
Δln(NPL)	1.000									
$\Delta\ln(NPL_{t-1})$	0.352*	1.000								
Δln(GDP)	-0.345*	-0.145*	1.000							
Δln(HHCE)	-0.294*	-0.165*	0.958*	1.000						
Δln(NDI)	-0.510*	-0.203*	0.690*	0.630*	1.000					
Δln(CE)	-0.240*	-0.223*	0.912*	0.927*	0.630*	1.000				
Δln(INF)	-0.116*	-0.169*	0.210*	0.209*	0.927*	0.136*	1.000			
ΔGDEBT	0.415*	0.271*	-0.503*	-0.481*	0.209*	-0.497*	-0.151*	1.000		
Δln(UN)	0.444*	0.363*	-0.581*	-0.562*	-0.481*	-0.501*	-0.207*	0.545*	1.000	
ΔULT	0.228*	0.368*	-0.328*	-0.357*	-0.562*	-0.498*	-0.199*	0.388*	0.590*	1.000

legend: *: significant at 5%

Table 3: The effective model space

Model	$\Delta\ln(NPL_{t-1})$	$\Delta\ln(GDP)$	$\Delta\ln(HHCE)$	$\Delta\ln(NDI)$	$\Delta\ln(CE)$	$\Delta\ln(INF)$	$\Delta GDEBT$	$\Delta\ln(UN)$	ΔULT
M1	□			□			□		
M2	□			□					
M3	□	□					□		
M4				□					□
M5	□	□							
M6	□						□	□	
M7	□		□						
M8				□					
M9	□							□	
M10	□						□		
M11	□					□			
M12	□								
M13		□						□	
M14							□	□	
M15		□					□		
M16								□	
M17		□							
M18			□				□		
M19			□						
M20					□	□			
M21							□		
M22					□				

Table 4: Equal and Bates-Granger model weights

	Bates-Granger Weights (models ranked by RMSE)				Equal Weights (models ranked by MAE)		
Model				Model			
	All 22	Top 10	Top 5		All 22	Top 10	Top 5
M1	0.0613	0.1153	0.2171	M1	0.0455	0.1000	0.2000
M2	0.0612	0.1151	0.2168	M2	0.0455	0.1000	0.2000
M3	0.0547	0.1029	0.1938	M3	0.0455	0.1000	0.2000
M4	0.0528	0.0994	0.1873	M5	0.0455	0.1000	0.2000
M5	0.0522	0.0982	0.1850	M6	0.0455	0.1000	0.2000
M6	0.0520	0.0979	0	M4	0.0455	0.1000	0
M7	0.0502	0.0945	0	M7	0.0455	0.1000	0
M8	0.0501	0.0943	0	M10	0.0455	0.1000	0
M9	0.0498	0.0938	0	M8	0.0455	0.1000	0
M10	0.0471	0.0886	0	M9	0.0455	0.1000	0
M11	0.0463	0	0	M11	0.0455	0	0
M12	0.0416	0	0	M12	0.0455	0	0
M13	0.0412	0	0	M14	0.0455	0	0
M14	0.0401	0	0	M15	0.0455	0	0
M15	0.0394	0	0	M13	0.0455	0	0
M16	0.0385	0	0	M18	0.0455	0	0
M17	0.0384	0	0	M17	0.0455	0	0
M18	0.0382	0	0	M16	0.0455	0	0
M19	0.0377	0	0	M19	0.0455	0	0
M20	0.0376	0	0	M21	0.0455	0	0
M21	0.0351	0	0	M20	0.0455	0	0
M22	0.0344	0	0	M22	0.0455	0	0

Table 5: Coefficients of combined and benchmark models and 5–fold cross validation GoF measures (validation set)

	Benchmark model	Average			Median			Bates-Granger		
		All	10	5	All	10	5	All	10	5
$\Delta\ln(NPL_{t-1})$		0.133	0.223	0.266	-	-	-	0.150	0.223	0.221
$\Delta\ln(GDP)$	-0.806*	-0.293	-0.280	-0.560	-	-	-	-0.292	-0.280	-0.528
$\Delta\ln(HHCE)$		-0.166	-0.131	0	-	-	-	-0.155	-0.124	0
$\Delta\ln(NDI)$		-0.200	-0.440	-0.353	-	-	-	-0.244	-0.459	-0.636
$\Delta\ln(CE)$		-0.090	0	0	-	-	-	-0.071	0	0
$\Delta\ln(INF)$		-0.232	0	0	-	-	-	-0.214	0	0
$\Delta GDEBT$		0.449	0.388	0.528	-	-	-	0.431	0.383	0.321
$\Delta\ln(UN)$	0.644***	0.145	0.105	0.085	-	-	-	0.138	0.100	0
ΔULT		0.230	0.506	0	-	-	-	0.267	0.503	0.947
Constant	16.534***	12.982	9.700	8.022	-	-	-	12.486	9.740	10.749
MAE [bn EUR]	1.841	1.665	1.526	1.486	1.653	1.526	1.456	1.628	1.525	1.498
MdAE [bn EUR]	0.499	0.416	0.398	0.374	0.440	0.418	0.386	0.422	0.415	0.394
MAPE [%]	18.506	16.711	15.745	15.512	16.643	15.768	15.501	16.458	15.765	15.566
MdAPE [%]	15.021	12.288	11.488	10.991	11.983	10.993	10.434	12.217	11.250	10.735
RMSE [bn EUR]	3.995	3.721	3.388	3.250	3.700	3.370	3.186	3.621	3.348	3.254
Pseudo-R^2 [%]	94.106	94.780	95.409	95.630	94.745	95.334	95.724	94.962	95.473	95.600

legend: * p-value<.05; ** p-value<.01; *** p-value<.001

Figure 1: 5-fold CV RMSE for single-equation and combined models

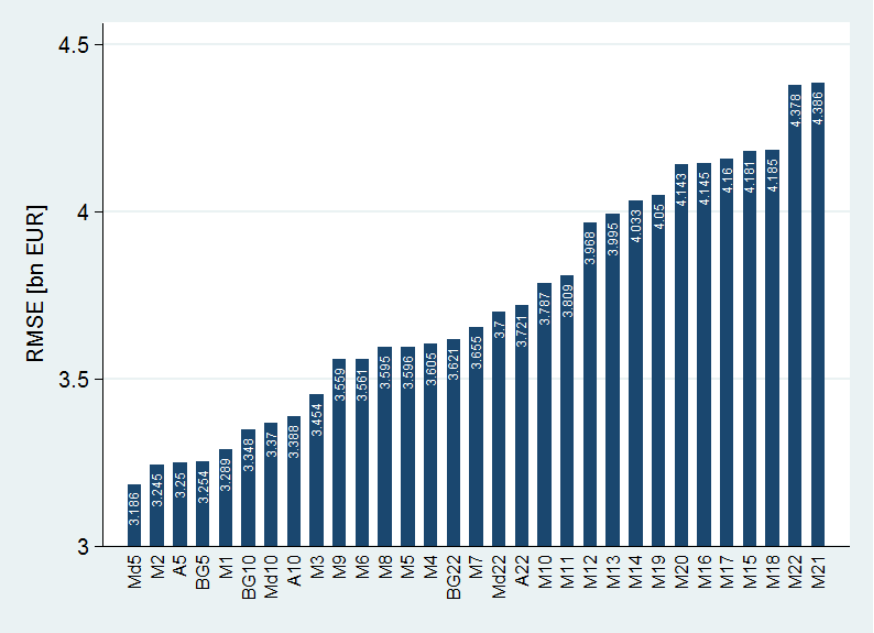


Figure 2: 5-fold CV Pseudo-R2 for single-equation and combined models

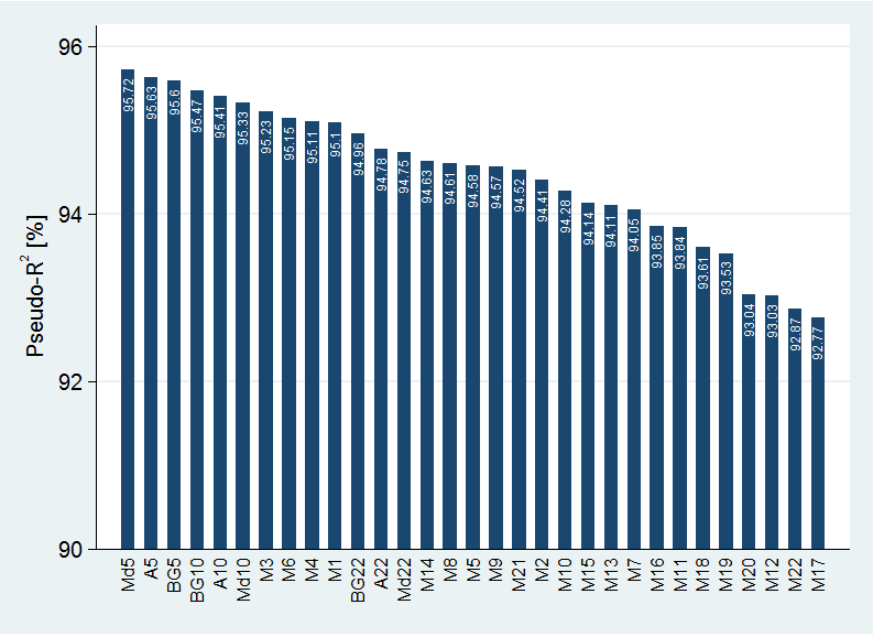


Table 6: The adverse 1% of the macroeconomic variables per country during 2006 – 2013

Country	Δln(GDP)	Δln(HHCE)	Δln(NDI)	Δln(CE)	Δln(INF)	ΔGDEBT	Δln(UN)	ΔULT
AT	-1.987	0.915	-11.780	0.931	0.400	11.210	23.050	0.100
BE	-1.522	-0.240	-12.650	0.806	-0.009	7.027	13.200	0.600
DE	-4.043	-0.498	-11.420	0.353	0.187	7.858	2.892	-0.100
DK	-4.615	-2.354	-16.660	-0.306	0.428	6.981	56.100	0.900
ES	-3.387	-4.601	-30.750	-5.800	-0.244	15.290	47.010	3.000
FI	-6.771	-1.295	-22.550	-1.140	1.624	9.042	25.070	0.600
FR	-2.889	-1.429	-14.120	0.279	0.103	10.910	21.130	0.500
GB	-13.670	-13.240	-33.330	-13.260	2.098	22.980	30.010	0.600
GR	-8.512	-9.157	-24.990	-11.250	-0.860	25.330	32.230	5.700
IE	-10.580	-13.480	-45.150	-9.375	-1.667	25.220	60.740	3.300
IT	-3.698	-2.010	-12.080	-0.683	0.738	10.120	26.370	1.300
LU	-3.882	1.178	-9.313	2.435	0.009	7.285	18.230	0.400
NL	-2.895	-2.938	-12.380	-0.388	0.925	12.060	24.630	0.600
PT	-3.759	-4.398	-26.130	-6.973	-0.909	14.900	21.480	1.600
SE	-12.900	-7.347	-16.690	-9.915	0.440	9.241	29.100	0.500

Figure 3: Scenario conditional forecast for non-stressed country banks

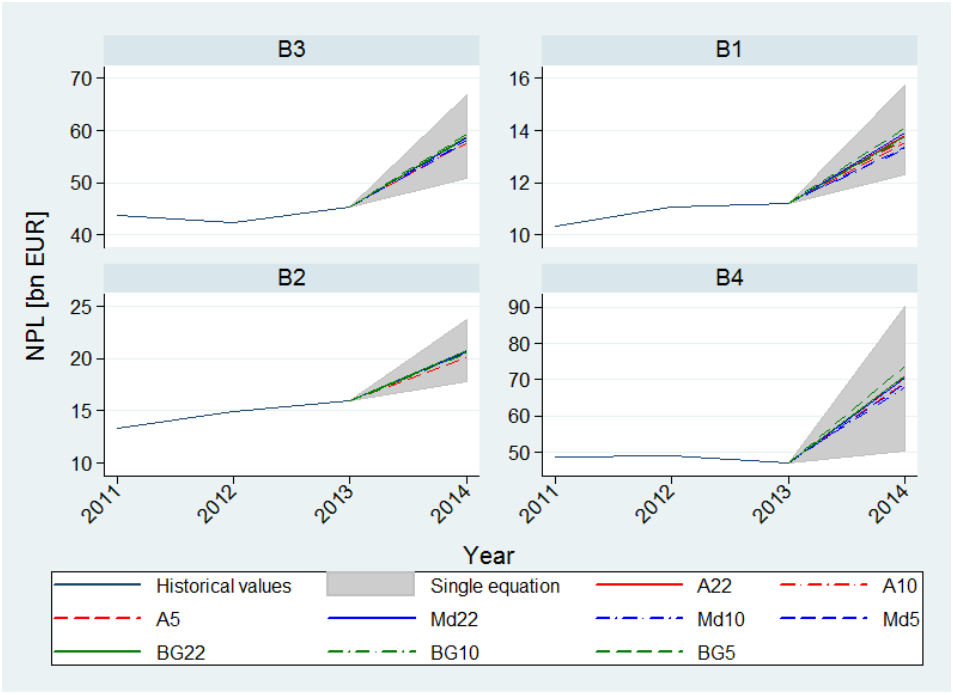


Figure 4: Scenario conditional forecast for stressed country banks

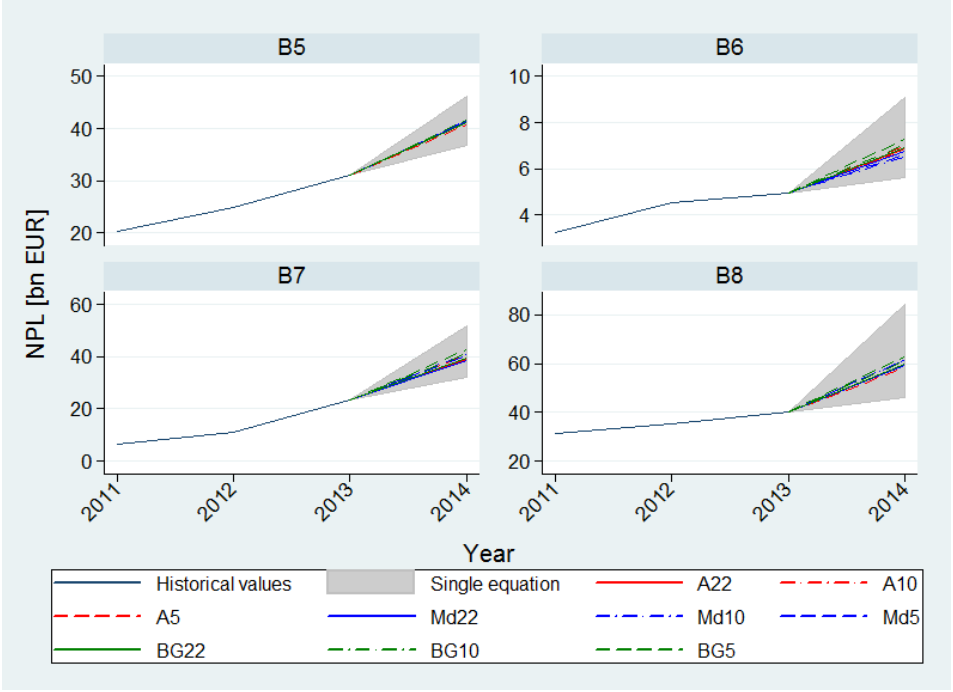
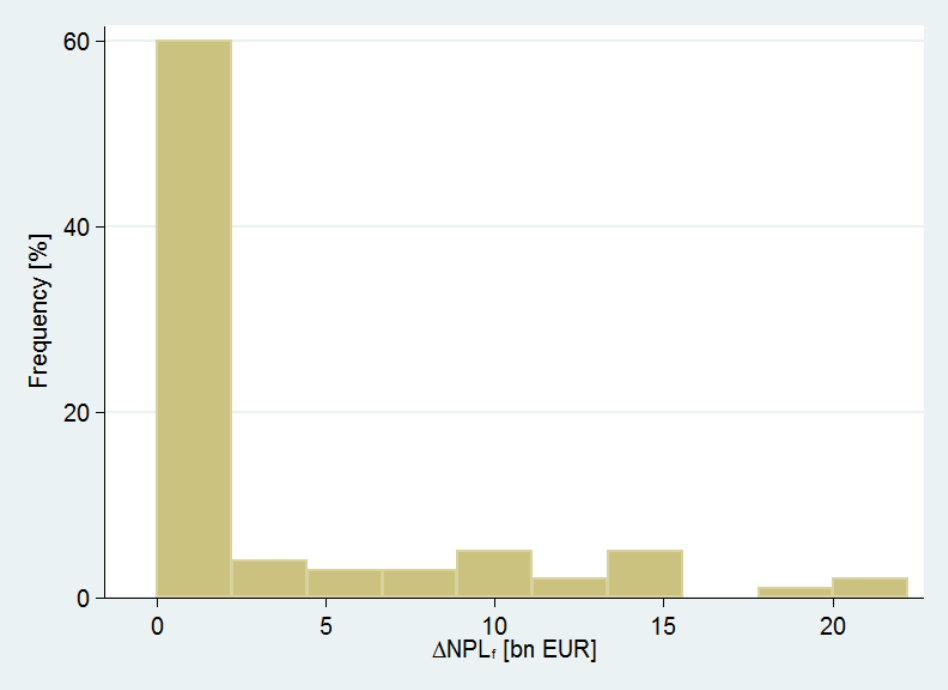


Table 7: Descriptive statistics of differences between model combination and minimum individual models' conditional forecasts

	Average			Median			Bates-Granger		
	All	10	5	All	10	5	All	10	5
Mean [bn EUR]	3.306	3.375	3.107	3.241	3.195	3.448	3.309	3.407	3.860
Median [bn EUR]	0.789	0.846	0.696	0.792	0.679	0.688	0.776	0.855	0.950
Min. [bn EUR]	0.038	0.034	0.032	0.037	0.030	0.030	0.037	0.034	0.040
Max.[bn EUR]	21.093	21.380	20.683	21.506	20.676	22.247	21.048	21.592	24.742

Figure 5: Conditional forecast difference distribution between 5-model median combination and minimum individual model



Appendix

A1. Goodness-of-fit measure definitions

Equations A1 to A6 define the goodness-of-fit measures used in the study to estimate each models' performance.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{A1}$$

$$MdAE = median(|y_i - \hat{y}_i|) \tag{A2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{A3}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{A4}$$

$$MdAPE = median\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \tag{A5}$$

$$pseudo - R^2 = \left(corr(y_i, \hat{y}_i)\right)^2 \tag{A6}$$

A2. Regression estimation results

In Tables A1 and A2 are presented the estimated models forming the effective model space accompanied by standard statistical tests that demonstrate their econometric validity as well as their sound economic interpretation. The order in which they are presented is the same as in Table 3, ranked according to their full-sample RMSE with the best performing (lowest RMSE) first.

Table A1: M1 – M10 regression estimation results

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
$\Delta\ln(NPL_{t-1})$	0.231***	0.251***	0.270***		0.347***	0.233***	0.347***		0.294***	0.257***
	(0.050)	(0.050)	(0.053)		(0.057)	(0.048)	(0.058)		(0.050)	(0.061)
$\Delta\ln(GDP)$			-1.189*		-1.609**					
			(0.517)		(0.560)					
$\Delta\ln(HHCE)$							-1.309*			
							(0.626)			
$\Delta\ln(NDI)$	-0.789***	-0.978***		-1.347***				-1.287***		
	(0.164)	(0.155)		(0.109)				(0.116)		
$\Delta\ln(CE)$										
$\Delta\ln(INF)$										
$\Delta GDEBT$	0.546*		1.043***			1.052***				1.238***
	(0.238)		(0.241)			(0.237)				(0.336)
$\Delta\ln(UN)$						0.423***			0.625***	
						(0.125)			(0.143)	
ΔULT				5.057**						
				(1.529)						
Constant	8.693***	11.566***	6.045**	17.196***	10.606***	3.198	10.457***	19.075***	6.176***	3.991
	(2.018)	(1.471)	(1.922)	(1.242)	(1.348)	(1.735)	(1.478)	(0.016)	(1.503)	(2.262)
AR(2) (p-value)	0.752	0.769	0.600		0.616	0.932	0.826		0.757	0.651
Sargan test (p-value)	0.148	0.212	0.202		0.107	0.173	0.055		0.122	0.047

legend: * p-value<.05; ** p-value<.01; *** p-value<.001

Robust standard errors in parentheses

Table A2: M11 – M22 regression estimation results

	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22
$\Delta\ln(NPL_{t-1})$	0.347*** (0.060)	0.361*** (0.062)										
$\Delta\ln(GDP)$			-0.806* (0.337)		-1.060*** (0.229)		-1.776*** (0.250)					
$\Delta\ln(HHCE)$								-0.795** (0.238)	-1.557*** (0.244)			
$\Delta\ln(NDI)$												
$\Delta\ln(CE)$										-0.915** (0.296)		-1.057*** (0.292)
$\Delta\ln(INF)$	-2.491* (1.117)									-2.617* (1.180)		
$\Delta GDEBT$				1.140*** (0.249)	1.291*** (0.218)			1.650*** (0.221)			1.922*** (0.200)	
$\Delta\ln(UN)$			0.644*** (0.159)	0.659*** (0.125)		0.828*** (0.119)						
ΔULT												
Constant	14.237*** (2.850)	8.497*** (1.446)	16.534*** (1.131)	10.151*** (1.398)	14.056*** (1.128)	14.737*** (0.649)	20.994*** (0.245)	12.168*** (1.752)	20.983*** (0.271)	25.803*** (2.480)	9.969*** (1.627)	20.467*** (0.335)
AR(2) (p-value)	0.431	0.615										
Sargan test (p-value)	0.109	0.054										

legend: * p-value<.05; ** p-value<.01; *** p-value<.001

Robust standard errors in parentheses

In Table A3 are reported the full-sample GoF measures for each individual model in the effective model space. Ranking according to their RMSE is used to define their names throughout the paper.

Table A3: Individual model full-sample GoF measures

Model	MAE [bn EUR]	MdAE [bn EUR]	MAPE [%]	MdAPE [%]	RMSE [bn EUR]	R^2 [%]
M1	1.441	0.386	15.286	10.774	3.326	95.198
M2	1.453	0.444	15.607	10.808	3.328	95.076
M3	1.494	0.429	15.770	10.581	3.520	94.755
M4	1.584	0.449	16.620	11.724	3.581	95.017
M5	1.542	0.432	16.623	11.313	3.603	94.175
M6	1.543	0.422	15.806	10.980	3.609	94.780
M7	1.589	0.452	17.018	11.656	3.672	93.995
M8	1.621	0.449	16.971	11.972	3.677	94.517
M9	1.629	0.421	16.755	11.484	3.687	94.207
M10	1.614	0.409	16.281	11.348	3.792	94.137
M11	1.728	0.507	18.174	12.912	3.824	93.390
M12	1.776	0.471	18.113	13.466	4.037	92.981
M13	1.812	0.476	18.324	14.314	4.054	93.939
M14	1.800	0.446	17.962	14.559	4.109	94.599
M15	1.805	0.482	17.902	13.779	4.146	94.069
M16	1.866	0.482	18.696	15.266	4.193	93.734
M17	1.865	0.502	18.830	14.726	4.199	93.052
M18	1.833	0.467	18.124	14.561	4.212	94.149
M19	1.891	0.514	19.218	15.286	4.237	92.966
M20	1.919	0.479	19.597	16.276	4.243	92.817
M21	1.916	0.475	18.425	14.753	4.395	93.915
M22	1.978	0.497	19.585	15.217	4.438	92.619

In Table A4 are reported the 5-fold cross validation GoF measures in the validation set, for each of the 22 single-equation models.

Table A4: Single-equation model 5-fold cross validation GoF measures (validation set)

Model	MAE [bn EUR]	MdAE [bn EUR]	MAPE [%]	MdAPE [%]	RMSE [bn EUR]	R^2 [%]
M1	1.456	0.387	15.381	10.812	3.289	95.097
M2	1.453	0.407	15.465	11.039	3.245	94.410
M3	1.553	0.450	15.841	11.494	3.454	95.226
M4	1.597	0.456	16.713	12.432	3.605	95.105
M5	1.538	0.457	16.667	11.757	3.596	94.577
M6	1.574	0.425	16.057	11.312	3.561	95.148
M7	1.609	0.450	16.871	11.655	3.655	94.053
M8	1.634	0.460	17.091	12.621	3.595	94.613
M9	1.662	0.426	16.577	12.721	3.559	94.567
M10	1.644	0.451	16.372	11.533	3.787	94.280
M11	1.716	0.498	18.225	13.947	3.809	93.844
M12	1.790	0.499	18.085	13.962	3.968	93.033
M13	1.841	0.499	18.506	15.021	3.995	94.106
M14	1.821	0.488	18.082	14.756	4.033	94.632
M15	1.810	0.469	18.016	14.042	4.181	94.137
M16	1.868	0.489	18.679	15.207	4.145	93.853
M17	1.868	0.509	18.845	14.808	4.160	92.770
M18	1.844	0.508	18.237	14.343	4.185	93.610
M19	1.907	0.509	19.372	15.515	4.050	93.534
M20	1.915	0.492	19.751	15.946	4.143	93.045
M21	1.921	0.508	18.486	14.288	4.386	94.524
M22	1.977	0.493	19.591	15.935	4.378	92.873

References

Aiolfi, M., Capistrán, C., & Timmermann, A. G. (2010). Forecast combinations. CREATES research paper(2010-21).

Aiolfi, M., & Favero, C. A. (2005). Model uncertainty, thick modelling and the predictability of stock returns. Journal of Forecasting, 24(4), 233-254.

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. Journal of Econometrics, 135(1), 31-53.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The review of economic studies, 58(2), 277-297.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. International journal of forecasting, 8(1), 69-80.

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

Bjørnland, H. C., Gerdrup, K., Jore, A. S., Smith, C., & Thorsrud, L. A. (2012). Does Forecast Combination Improve Norges Bank Inflation Forecasts?*. Oxford Bulletin of Economics and Statistics, 74(2), 163-179.

Borio, C. (2010). Ten propositions about liquidity crises. CESifo Economic Studies, 56(1), 70-95.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: an integral part of inference. Biometrics, 603-618.

Buncic, D., & Melecky, M. (2013). Macroprudential stress testing of credit risk: A practical approach for policy makers. Journal of Financial Stability, 9(3), 347-370.

Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-theoretic approach: Springer Science & Business Media.

Cihák, M. (2007). Introduction to applied stress testing. IMF Working Papers(7-59), 1-74.

Claeskens, G., & Hjort, N. L. (2008). Model selection and model averaging (Vol. 330): Cambridge University Press Cambridge.

Del Negro, M., Hasegawa, R. B., & Schorfheide, F. (2014). Dynamic prediction pools: an investigation of financial frictions and forecasting performance: National Bureau of Economic Research.

Foglia, A. (2009). Stress Testing Credit Risk: A Survey of Authorities' Approaches. International Journal of Central Banking, 5(3), 9-45.

Fox, J. (1991). Regression diagnostics: An introduction (Vol. 79): Sage.

Galati, G., & Moessner, R. (2013). Macroprudential policy–a literature review. Journal of Economic Surveys, 27(5), 846-878.

Geweke, J., & Amisano, G. (2011). Optimal prediction pools. Journal of Econometrics, 164(1), 130-141.

Geweke, J., & Amisano, G. (2012). Prediction with misspecified models. The American Economic Review, 102(3), 482-486.

Granger, C. W., & Jeon, Y. (2004). Thick modeling. Economic Modelling, 21(2), 323-343.

Gross, M., & Población, J. (2015). A False Sense of Security in Applying Handpicked Equations for Stress Test Purposes. ECB Working Paper no. 1845.
Haldane, A. (2009). Why banks failed the stress test. BIS Review, 18, 2009.
Hansen, B. E. (2008). Least-squares forecast averaging. Journal of Econometrics, 146(2), 342-350.
Hardy, D. C., & Schmieder, C. (2013). Rules of thumb for bank solvency stress testing. IMF Working Papers(13-232).
Harrell, F. E. (2013). Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis: Springer Science & Business Media.
Hausman, J. A. (1978). Specification tests in econometrics. Econometrica: Journal of the Econometric Society, 1251-1271.
Henry, J., Kok Sorensen, C., Amzallag, A., Baudino, P., Cabral, I., Grodzicki, M., Leber, M. (2013). A macro stress testing framework for assessing systemic risks in the banking sector. ECB Occasional Paper no. 152.
Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. Statistical science, 382-401.
Isogai, T. (2009). Scenario design and calibration. Stress-testing the Banking System: Methodologies and Applications, M. Quagliariello (ed.), Cambridge University Press.
Jakubík, P., & Schmieder, C. (2008). Stress testing credit risk: is the Czech Republic different from Germany? Czech National Bank, Working Papers(9).
Jiménez, G., & Saurina, J. (2006). Credit Cycles, Credit Risk, and Prudential Regulation. International Journal of Central Banking, 2(2).
Jobst, A., Ong, L., & Schmieder, C. (2013). A Framework for Macroprudential Bank Solvency Stress Testing: Application to S-25 and Other G-20 Country FSAPs. IMF Working Papers(13-68).
Jones, M. T., Hilbers, P. L. C., & Slack, G. L. (2004). Stress Testing Financial Systems: What to Do When the Governor Calls (Vol. 4): International Monetary Fund.
Kalirai, H., & Scheicher, M. (2002). Macroeconomic stress testing: preliminary evidence for Austria. Financial Stability Report(3), 58-74.
Louzis, D. P., Vouldis, A. T., & Metaxas, V. L. (2012). Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios. Journal of Banking & Finance, 36(4), 1012-1027.
Matz, L., & Neu, P. (2006). Liquidity risk measurement and management: a practitioner's guide to global best practices (Vol. 408): John Wiley & Sons.
Melecky, M., & Podpiera, A. M. (2010). Macroprudential stress-testing practices of central banks in central and south eastern Europe: an overview and challenges ahead. World Bank Policy Research Working Paper Series.
Moral-Benito, E. (2015). Model averaging in economics: an overview. Journal of Economic Surveys, 29(1), 46-75.
Palm, F. C., & Zellner, A. (1992). To combine or not to combine? Issues of combining forecasts. Journal of Forecasting, 11(8), 687-701.
Perotti, R. (1996). Fiscal consolidation in Europe: Composition matters. American Economic Review, 86(2), 105-110.
Reinhart, C. M., & Rogoff, K. S. (2011). From Financial Crash to Debt Crisis. American

Economic Review, 101(5), 1676-1706.
Sargan, J. D. (1958). The Estimation of Economic Relationships using Instrumental Variables. Econometrica, 26(3), 393-415.
Single Supervisory Mechanism (2015). The list of significant supervised entities and the list of less significant institutions.
Stock, J. H., & Watson, M. W. (1998). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series: National Bureau of Economic Research.
Stock, J. H., & Watson, M. W. (2003). Forecasting Output and Inflation: The Role of Asset Prices. Journal of Economic Literature, 41, 788-829.
Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. Journal of Forecasting, 23(6), 405-430.
Stock, J. H., & Watson, M. W. (2006). Forecasting with many predictors. Handbook of economic forecasting, 1, 515-554.
Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger & A. Timmermann (Eds.), Handbook of economic forecasting (Vol. 1, pp. 135-196): Elsevier B.V.
Vazquez, F., Tabak, B. M., & Souto, M. (2012). A macro stress test model of credit risk for the Brazilian banking sector. Journal of Financial Stability, 8(2), 69-83.
Winkler, R. L., & Makridakis, S. (1983). The combination of forecasts. Journal of the Royal Statistical Society. Series A (General), 150-157.
Wooldridge, J. (2012). Introductory econometrics: A modern approach: Cengage Learning.

BANK OF GREECE WORKING PAPERS

185. Adam, A., and T., Moutos, "Industry-Level Labour Demand Elasticities Across the Eurozone: Will There Be Any Gain After the Pain of Internal Devaluation?" July, 2014.

186. Tagkalakis, O.A., "Fiscal Policy, Net Exports, and the Sectoral Composition of Output in Greece", September 2014.

187. Hondroyiannis, G. and D., Papaoikonomou, "When Does it Pay To Tax? Evidence from State-Dependent Fiscal Multipliers in the Euro Area", October 2014.

188. Charalambakis, C. E., "On Corporate Financial Distress Prediction: What Can we Learn From Private Firms in a Small Open Economy?, November 2014.

189. Pagratis, S., E., Karakatsani and E. Louri, "Bank Leverage and Return on Equity Targeting: Intrinsic Procyclicality of Short-Term Choices", November 2014.

190. Evgenidis, A. and C., Siriopoulos, "What are the International Channels Through Which a US Policy Shock is Transmitted to the World Economies? Evidence from a Time Varying Favar, January 2015.

191. Louzis, D. P., and A.T., Vouldis, "Profitability in the Greek Banking System: a Dual Investigation of Net Interest and Non-Interest Income", February 2015.

192. Papaspyrou, S.T, "EMU 2.0 - Drawing Lessons From the Crisis - a New Framework For Stability and Growth", March 2014.

193. Litina, A and T, Palivos, "Corruption and Tax Evasion: Reflections on Greek Tragedy", June 2015.

194. Balfoussia, H. and H.D. Gibson, "Financial Conditions and Economic Activity: The Potential Impact of the Targeted Longer-Term Refinancing Operations (TLTROS)", July 2015.

195. Louzis, P. D., "Steady-State Priors and Bayesian Variable Selection in VAR Forecasting", July 2015.

196. Zografakis, S. and A., Sarris, "The Distributional Consequences of the Stabilization and Adjustment Policies in Greece During the Crisis, with the Use of A Multisectoral Computable General Equilibrium Model", August 2015.

197. Papageorgiou, D. and E. Vourvachaki, "The Macroeconomic Impact of Structural Reforms in Product and Labour Markets: Trade-Offs and Complementarities", October 2015.

198. Louri, H., and P. M. Migiakis, "Determinants of Euro-Area Bank Lending Margins: Financial Fragmentation and ECB Policies", October 2015.

199. Gibson, D. H, S.G. Hall, and G. S. Tavlas, "The effectiveness of the ECB's asset purchase programs of 2009 to 2012", November 2015.

200. Balfoussia, H and D. Malliaropulos, "Credit-less recoveries: the role of investment-savings imbalances", November 2015.

201. Kalyvitis, S., “Who Exports High-Quality Products? Some Empirical Regularities From Greek Exporting Firms”, December 2015.

202. Papadopoulos, S., P. Stavroulias and T. Sager, “Systemic Early Warning Systems for EU15 Based on the 2008 Crisis”, January 2016.