Uncertain Times and Early Predictions of Bank Failure

Cullen F. Goenner, PhD
Professor of Economics & Finance
University of North Dakota
293 Centennial Drive Stop 8369
Grand Forks, ND 58202-8369
Email: cullen.goenner@und.edu
Web: www.ndeconomist.com

This version: June 2018

Abstract (147 words)
The recent financial crisis and economic recession has shown that bank failure in the United States, while rare is a concern during uncertain times. Understanding the magnitude of banks at risk early in a crisis is a key challenge faced by policymakers. Early warning models are quite accurate at assessing risk using rolling predictions, yet they rely on recent failures for their accuracy. Of interest here is the ability at the start of a crisis to predict future failures, when the recent past has few events to base our inferences. We use logit and survival models to show that banks' initial conditions at the start of the most recent crisis, along with model estimates from the S & L crisis, are good indicators of failures during the crisis, and that by accounting for uncertainty in our model's specification we are able to improve our model's out-of-sample predictions.

Key words: Bank failure; Cox and logistic regression; Bayesian model averaging; prediction, banking crisis.

1. Introduction

The day Lehman Brothers filed for bankruptcy, September 15 of 2008, it became quite clear to everyone that the financial sector was again in crisis and that commercial banks in the United States and elsewhere were also at a heightened risk of failure. Less clear at the time was the magnitude of this risk to banks, as the previous fifteen years had seen on average less than eight failures a year in the United States.[1] Traditional early warning models of bank failure rely on the recent pattern of previous failures to base their predictions. With few recent failures to draw from, it would seem natural to look towards a past period of crisis to guide US policymakers' predictions in late 2008 of the banks that would fail subsequent to the financial crisis. Previous research (Cole et al., 1995; Cole and Gunther, 1998) has shown that the statistical models used by the Federal Reserve were quite accurate in predicting failures during the last major banking crisis of 1985 – 1992, referred to as the Savings and Loans (S & L) crisis. These models' predictions are based on the financial conditions of banks that are captured in their call reports and reflect measures of banks' capital adequacy, asset quality, management quality, earnings, liquidity, and sensitivity to market risk (CAMELS).[2] Cole et al. (1995) and Cole and Gunther (1998) find using model estimates based on data from an earlier year during the S & L crisis that they are able to accurately predict failures out-of-sample later on during the same crisis. We consider here whether failure patterns established during the S & L crisis are also useful in predicting failures at the start of the most recent financial crisis.

---

[1] Two years (2005, 2006) had no failures. The FDIC's list of failed banks is available at https://www.fdic.gov/bank/individual/failed/banklist.html

[2] Federal bank regulators in the United States determine the health of individual banks using the revised Uniform Financial Institutions Rating System, which is also known as the CAMELS rating system. Federal Register, Vol. 61 No. 245 December 19, 1996

A number of studies (Cleary and Hebb, 2016; Cole and White, 2012; DeYoung and Torna, 2013; Jin et al., 2011; Ng and Rowchowdary, 2014) examine bank failures during the most recent financial crisis. Similar to Cole et al. (1995) and Cole and Gunther (1998), these studies use data during a particular crisis period for both their modeling and evaluation purposes. The contributions of these studies provide policymakers important insights into the factors that influenced failure during the financial crisis. Jin et al. (2011), for example, find that a bank's choice of auditor plays a role in failures during the crisis, as does their treatment of loan loss reserves as regulatory capital (Ng and Rowchowdary, 2014). DeYoung and Torna (2013) investigate whether banks' exposure to non-traditional banking activities (insurance underwriting, securitization, investment banking, and venture capital) put a bank at a higher risk of failure during the most recent crisis and find evidence that this is the case. The issue from the policymaker's perspective is the usefulness of these models' estimates is retrospective, i.e. they are useful in helping understand the conditions that contributed to failure after the crisis period examined has passed, while their relevance to failures in a future crisis is yet unclear.

This paper uses the observation made by Cole and White (2012) that the same financial conditions influencing bank failures in 2009 also affected failures during the 1980s (Cole and Gunther, 1998; Lane et al., 1986, Thomson, 1992; Whalen, 1991). The focus here though differs from Cole and White (2012) in that we use the failure experience of banks during the S & L crisis to build a prediction model, which is applied to data observed by policymakers in late 2008 for out-of-sample predictions of failures in the period 2009-2014. In a sense, we attempt to identify initial conditions that serve as common risk factors of bank failures across different crises episodes, which allow for the creation of risk scores using the estimates of past crisis episodes. If these risk scores are accurate, then policymakers in late 2008 would have had a means of assessing risks to banks early at the start of the crisis and it would suggest the most recent crisis period may help us predict risk exposure during the next.

Risk of bank failure is assessed here using both logit and survival models. Whereas others (Cole and Wu, 2010; Mayes and Stremmel, 2014) have focused on the relative accuracy of the two models, our attention is on the different information each model provides policymakers at the start of a crisis. The logit model estimated using data from the S & L crisis is able to correctly identify 97 of the 119 (82%) banks that actually fail in 2009 using year-end financial data from 2008 and a cutoff established in the earlier period. Our Cox proportional hazards model uses estimates from the S & L crisis and year-end data from only 2008 to predict risk scores and the survival experience of banks throughout the period 2009-2014. We are able to show that our model's prediction of risk of failure is as accurate later in the crisis period as it is early on, which indicates our initial conditions are good for assessing the risk to banks throughout a crisis period. Accuracy is measured here with a time dependent version of the area under the Receiver Operating Characteristic (ROC) curve (Heagerty and Zheng, 2005). Policymakers, using our Cox estimates and classification of risk scores, would identify in 2008 that 2.1% of banks were in critical condition, 6.9% were unhealthy, and 91% were healthy, which is quite comparable to the 1.6% of banks that failed within a year, the 4.1% of banks that would fail later during the crisis, and the 92% that remained healthy throughout.

Our paper further contributes to the literature, in that we demonstrate models of bank failure are subject to uncertainty as to which variables to include in the model's specification. Cole et al. (1995, p. 6) note that the Federal Reserve identifies approximately thirty financial variables as most likely to affect the probability of bank failure.[3] From this list, Cole et al.

[3] Lane et al. (1986, p. 516) though note there is little agreement by regulators on which factors are most important.

(1995) use stepwise selection to determine the subset of these variables relevant to failures during the S & L crisis. Bank failures, even during crises, are relatively rare events and in such cases where there are many potential risk factors, predictions based on a single model specification are likely sensitive to variable selection (Volinksy et al., 1997). We use techniques of Bayesian model averaging (BMA) to base our inferences on estimates that explicitly account for our uncertainty of the model's specification by averaging over the estimates from several different specifications. By accounting for model uncertainty, we improve, relative to stepwise selection, our out-of-sample predictions of the logit and Cox models.

The rest of this paper is organized as follows. In section 2 we provide discussion of the empirical logit and Cox models of bank failure. Section 3 introduces the method of Bayesian model averaging. In section 4, we discuss the data and variables, and section 5 discusses our results. Section 6 concludes.

2. Empirical models of bank failure

There are in general two approaches to modeling bank failure. The first approach models bank failure as a binary outcome, i.e. whether a bank fails in a subsequent period conditioning on a set of observables at a given point in time. These models use a cross-section of banks and estimate the probability of bank failure using either probit (Cole and Gunther, 1998; Cole et. al, 1995; Wheelock, 1992) or logistic regression (Arena, 2008; Cole and White, 2012; Ng and Rowchowdary, 2014). Cole and Gunther (1998) in their early warning model use bank financial data from year-end 1985 to predict whether a bank fails in the two year period 1986(Q2) – 1988(Q1). Their results indicate lower capital ratios, declining asset quality, lower earnings and less liquidity all play a significant role in failures during the period.

The predictions from a binary response model are probabilities, $\hat{p}_i$, and therefore one must adopt a classification rule, *c*, to distinguish between banks predicted to fail $(\hat{p}_i > c)$ from those that are not $(\hat{p}_i \le c)$. Accuracy is measured by the sensitivity and specificity of the model's classifications. Sensitivity measures the fraction of banks predicted to fail $(\hat{p}_i > c)$ among the subset of banks that actually fail during the period examined, whereas specificity measures the fraction of banks predicted not to fail $(\hat{p}_i \le c)$ among the subset that do not fail. One may also use the error rates to think about the model's classifications. A type – 1 error consists of a bank that is predicted not to fail which fails, whereas a type – 2 error consists of a bank that is predicted to fail but does not. Type-1 errors are more costly as they may result in costs from failures that could have been potentially avoided if predicted, while type – 2 errors entail diverting limited supervisory resources to a healthy bank. Mathematically the type – 1 error rate is equal to 1 – sensitivity and the type – 2 error rate equals 1 – specificity. For a given model specification a tradeoff exists between the two types of errors, i.e. reducing the cutoff to classify a failing bank reduces type – 1 errors at the expense of more type – 2 errors, whereas raising the cutoff has the opposite effect. Cole and Gunther (1998) find for a type – 2 error rate of 10% that their probit model estimated with data from 1985 has an out-of-sample type – 1 error rate of 9.8% using 1987 data and of 7.9% using 1989 data.

A plot of type-1 versus type – 2 error rates, across the range of classification cutoffs, allows for more generalizable and visual comparison of the model's classifications. Cole and Gunther (1998), for example, show that their model specification estimated with financial data from call reports has lower type – 1 errors than a specification estimated with a bank's composite

CAMELS rating over the range of type – 2 errors.[4] A closely related measure of performance is the receiver operating characteristic curve (ROC), which plots the tradeoff between the model's sensitivity relative to type – 2 errors. The area under the ROC curve (AUC) is a statistic that summarizes the model's predictive accuracy over the range of cutoffs and measures the probability a randomly selected failed bank has a higher predicted risk score than a bank that does not fail (control). A value equal to 1 indicates the model's predictions are able to completely discriminate between failed and non-failed banks, whereas a value equal to .5 indicates the predictions are no better than pure chance.

Logistic and probit regression models of bank failure have also been applied using panel data (Betz et al., 2014; Jin et al., 2011; Mayes and Stremel, 2014). An important aspect of bank failures is that once a bank fails, it cannot possibly fail again. The econometric issue this poses when using panel data is observations of failures from a cross-section of banks are no longer independent of each other over time, which results (Shumway, 2001) in biased and inconsistent estimates. Shumway (2001) has shown this issue can be alleviated by correcting the standard errors of the multi-period logit model to account for the lack of independence, and that the resulting specification is equivalent to a discrete time hazard model. DeYoung and Torna (2013) and Cole and Wu (2010) use this approach in their analyses of bank failure.

An alternative, which is used here, is to directly model the time to bank failure using a survival model. The dependent variable in a survival model is the time to failure, T_i, which is the difference in time (days, years, etc.) between when a bank becomes at risk of failure and when it either fails, or if it does not fail, the end of the study period. Observations in the latter case are said to be censored. Studies of bank failure (Brown and Dinc, 2005; Lane et al., 1986;

[4] CAMELS ratings assigned by regulators to banks during examinations are not made publicly available.

Mayes and Stremel, 2014; Ng and Rowchowdary, 2014; Whalen, 1991; Wheelock and Wilson, 1995, 2000) using a survival model typically specify a Cox proportional hazards model.[5] Estimates from the model allow one to specify the hazard function at time *t*, which represents the rate of failure at a point in time, conditioning on failure having yet to occur. This differs from the logit model that considers the proportion of failures within a time period. The hazard of failure $h(t,x,\beta) = h_0(t)e^{x\beta}$ is a function of two terms.[6] The baseline hazard, $h_0(t)$, characterizes how the hazard of bank failure changes relative to time at risk, while $e^{x\beta}$ is a feature of the Cox model that characterizes how the hazard of bank failure depends on the control variables. The model is said to be semi-parametric in the sense that the baseline hazard's $h_0(t)$ dependence on time is unspecified and the coefficients enter the model linearly. The effects of the covariates on survival are evaluated with a transform of the coefficients using the hazard ratio (equation 1):

$$HR(t,x_1,x_0) = \frac{h_0(t)e^{x_1\beta}}{h_0(t)e^{x_0\beta}} = \frac{e^{x_1\beta}}{e^{x_0\beta}} = e^{(x_1-x_0)\beta} \qquad (1)$$

The hazard ratio (HR) is a measure of relative risk that does not depend on time and reflects the change in a ratio of rates from a given change in covariate x's values, ($x_1 - x_0$).

Using the hazard function, one is also able to specify a survival function $S(t,x,\beta) = S_0(t)e^{x\beta}$, where the baseline survivor function $S_0(t)$ is a function of the cumulative hazard function $S_0(t) = e^{-H_0(t)}$. The survival function represents the probability of observing a survival time greater than time *t* and is mathematically related to the probability of observing a failure time less than *t*, i.e. $S(t,x,\beta) = 1 - F(t,x,\beta)$. One way to interpret the linear

[5] Arena (2008) instead uses a fully parametric Weibull regression survival model.
[6] The notation for the hazard model used here is similar to Hosmer et al. (2008).

prediction of $x\beta$ in the survival function is as a risk score, $M_i = x_i\hat{\beta}$, such that banks with similar risk scores experience similar survival rates through time. Given a risk score, M_i, one can define measures of model accuracy that are comparable to binary response models. Here we apply Heagerty and Zheng's (2005) time dependent measures of the model's incident sensitivity and dynamic specificity. [7] Incident sensitivity is referring to the sensitivity of the model's predicted risk score to discriminate between incident cases, i.e. the probability observations that fail at time t $(T_i = t)$ have a risk score greater than a user defined cutoff c $P(M_i > c \mid T_i = t)$. Dynamic specificity though refers to the fraction of banks that remain alive at time t, $(T_i > t)$, with risk scores less than or equal to the cutoff $P(M_i \leq c \mid T_i > t)$. Using the incident sensitivity and dynamic specificity measures one is then able to generate time dependent versions of the ROC curve, and calculate a time dependent measure of the AUC. The time dependent AUC(t) value allows for evaluating how the predictive performance of the model changes with time, which is important here as our interest is in developing a model that uses only initial conditions to achieve both good short-term and long-term predictive performance during a crisis period.

3. Bayesian Model Averaging

Our analysis uses Bayesian model averaging to incorporate into our inferences our uncertainty as to which dependent variables we should include in our model's specification. Rather than base inference on a single model specification, the estimates from Bayesian model averaging (BMA) use a weighted average of estimates from several specifications, with weights determined by the posterior support each receives from the data. BMA offers a theoretically appealing method of accounting for uncertainty in model specification and has been shown

[7] The time dependent ROC and AUC analyses are conducted using the R-package "risksetROC" (Heagerty and Saha-Chaudhuri, 2012).

(Raftery et. al., 1995; Volinsky et al., 1997) to have better predictive performance than other methods, such as stepwise techniques. We apply BMA to both logistic and Cox proportional hazards models and the discussion below generalizes to either type of model with differences between the two as noted.

For the purposes used here, a model specification is defined by a linear combination of variables. Both the Cox and logistic models are non-linear, though the variables enter the models in a linear fashion. The set of model specifications of interest here include each of the different linear combinations of variables that are identified as potentially relevant to predicting bank failure. Identifying *p* variables of interest implies there are then $K = 2^p$ different model specifications to consider. The weight given to each of the *K* different models' estimates are determined by the specification's posterior model probability. The posterior distribution of our parameters given the data is equal to

$$P(\beta / D) = \sum_{k=1}^{K} P(\beta / M_k, D) P(M_k / D) \tag{2}$$

The first component of equation 2, $P(\beta / M_k, D)$, represents the posterior distribution of estimates from the different model specifications. Volinsky et al. (1997) show this distribution can be approximated by applying maximum likelihood to the *K* different models in the case of logit and Cox models. The second component, $P(M_k / D)$, is the posterior model probability, i.e. weights, which represents for a given model specification the posterior likelihood the specification is the true model that generates the data. The sum of the weights across models is equal to 1. By Bayes' rule and the law of total probability the posterior model probability is

$$P(M_k / D) = \frac{P(D / M_k) P(M_k)}{\sum_{l=1}^{K} P(D / M_l) P(M_l)} \tag{3}$$

where $P(D/M_k)$ is the likelihood and $P(M_k)$ is the prior probability that model M_k is the true model. We assume a uniform prior such that each of our models under consideration is *a priori* as equally likely to be the true model as another.[8] The PMP then simplifies to become

$$P(M_k / D) = \frac{P(D / M_k)}{\sum_{l=1}^{K} P(D / M_l)} \tag{4}$$

The integrated likelihood, also referred to as the marginal likelihood, is found by integrating over parameter vector β_k

$$P(D / M_k) = \int P(D / \beta_k, M_k) P(\beta_k / M_k) d\beta_k \tag{5}$$

where β_k is a vector of parameters, $P(D/\beta_k, M_k)$ is the likelihood, and $P(\beta_k/M_k)$ is the prior density of β_k under model M_k. Volinsky et al. (1997) suggest the integral in equation 5 can be approximated using the Laplace method by using a function of Schwarz's (1978) Bayesian information criterion.

$$\begin{aligned} P(D / M_k) &\approx \exp\left(-\frac{1}{2} BIC_k^{'}\right) \\ BIC_k^{'} &= BIC_k - BIC_0 = -LRT + p_k \log(N) \end{aligned} \tag{6}$$

BIC_k and BIC_0 are the values of the Bayesian information criterion for model specification k and the null model (constant only), respectively. The difference of which is equal to the likelihood ratio statistic subtracted from the number of parameters in model k, p_k, multiplied by the natural log of the number of observations. For the logistic model the number of observations in equation 6, N, is equal to the sample size (Raftery, 1995). In the case of the Cox model, one could use the

[8] Fernandez et al. (2001) note the assumption is common when there is not strong prior information to suggest otherwise. The assumption on priors has been found by Raftery (1995) to have little impact on the posterior distribution.

number of units under observation, the total time of all units under observation, or the number of events, i.e. failures. The latter choice, which is used below is preferred by Raftery et al. (1995).

To test hypotheses under Bayesian model averaging one uses Bayes factors. A Bayes factor allows us to compare the evidence in favor of one hypothesis relative to another. Consider two hypotheses, H_0 and H_1, where we have prior beliefs as to their validity given by $P(H_0)$ and $P(H_1)$. Using Bayes rule, it can be shown that the odds of observing the null relative to the alternative are given by:

$$\frac{P(H_0 / D)}{P(H_1 / D)} = \frac{P(D / H_0)}{P(D / H_1)} \frac{P(H_1)}{P(H_0)} \qquad (7)$$

The posterior odds of the null hypothesis being true is equal to the Bayes factor multiplied by the prior odds in favor of the null. If each hypothesis is a priori equally likely, then the posterior odds of the null is simply equal to the Bayes factor. A Bayes factor is an odds ratio of probabilities and as such can be converted into the probability the null hypothesis is true.[9] A Bayes factor of 20 is therefore interpreted as the null hypothesis being 20 times more likely than the alternative, which corresponds to a 95% probability of the null being true and a 5% probability in favor of the alternative.

Bayes factors differ fundamentally in interpretation than p-values found in the Neyman-Pearson approach to statistics. A p-value measures the probability of observing an outcome in the data more extreme than what is assumed under the null hypothesis and in a sense represents $P(D / H_0)$. A p-value of .05 indicates the null hypothesis is rejected 5% of the time, when it is in fact true, and yet it does not tell us the probability the null hypothesis is true $P(H_0 / D)$. An

---

[9] The odds ratio equals = $\Omega(H_0 / D) = \frac{P(H_0 / D)}{1 - P(H_0 / D)}$ therefore $P(H_0 / D) = \frac{\Omega(H_0 / D)}{1 + \Omega(H_0 / D)}$

example by Edwards, Lindman, and Savage (1963, pp. 221-222) that follows highlights this distinction. The frequentist perspective is that by assuming the null hypothesis is true, the t-statistic from a two-tailed t-test, with many degrees of freedom will exceed 1.96 2.5% of the time. Similarly, .5% of the time the value will exceed 2.58, which implies 2% of the time the statistic will lie between 1.96 and 2.58, when the null hypothesis tested is true. It might then appear the data strongly favor the alternative if the observed test statistic lies within this interval. Consider an alternative, when the null is false the statistic lies uniformly between the values -20 and 20. In this example the value of t lies between the values of 1.96 and 2.58 with probability 1.55%. Given the alternative, the data actually favor the null.

Bayes factors are a nice alternative to p-values, which can overstate the evidence against the null in large samples (Greene, 1997; Leamer, 1978). The issue is p-values can suggest causal relations in the data that do not exist. Raftery et al. (1997) have shown using simulated data, where the underlying relation between the data is known, BMA is better able to determine the true model's specification relative to stepwise and other single model approaches that rely on p-values. Utilizing a Bayesian approach we are left with a better understanding of the causal relationships in our data and an explicit accounting for our uncertainty in the model's specification.

We apply BMA to the logit and Cox models using the R-package BMA (Raftery et al. 2018). To increase the speed of estimation, the routine narrows down the number of models to average over by eliminating specifications that receive little support from the data. These are specifications where the odds in favor of another model specification being the true model are more than twenty to one. Excluding these specifications has little impact on our inferences given the low weight each would receive if included. The routine provides estimates of the posterior means and standard errors of the coefficients, which can be easily compared to their single equation MLE counterparts. For each variable, the routine calculates the posterior probability that the coefficient is non-zero, referred to as the posterior effect probability (PEP). Raftery (1995) considers the statistical evidence of an effect to be weak, positive, strong, and very strong according to a commonly used rule of thumb based on Bayes factors of 1, 3, 20, and 150, which corresponds to PEP values of .5, .75, .95, and .99 on the probability scale, respectively.

4. Data

The control variables used throughout this study are drawn from the Report of Condition and Income (call report) data provided by individual banks to the Federal Reserve, Federal Deposit Insurance Corporation, and the Comptroller of the Currency and are distributed by the Federal Financial Institutions Examination Council. Our sample of banks includes commercial banks, state chartered banks, and cooperative banks. We relied on theory and previous research to identify the list of financial variables most likely related to bank failure. Also guiding this choice was our desire to focus on the failure experience of banks during the S & L crisis for basing predictions of failures during the most recent financial crisis. The twenty-five variables used here include the year-end call report items considered by the Federal Reserve for use in their Financial Institutions Monitoring System (FIMS) model (Cole et al., 1995) to predict banks' risk of failure.[10] As Cole et al. (1995) note these variables were selected based on the Fed's review of the literature and their use in examination reports. Further, these measures as noted earlier were strong predictors of failures during the S & L crisis. The list of measures and

[10] The FDIC uses a similar variables in their statistical CAMELS off-site rating (SCOR) model to predict changes in CAMELS ratings (Collier et al., 2003).   The series, RCFD1406, loans past due 30-89 days and still accruing interest is included in both the government's FIMS and SCOR models. The series is not used here as it is confidential for the period 1984-1990.

their definition are included in Table 1. Each of the control variables other than banks' age and size are scaled by total assets.

[Insert Table 1 about here]

Whether a bank fails and the date of failure are identified by the FDIC's list of failed banks. At year-end 1984 there were 14,025 banks in our sample, and among these banks, 1101 would fail (7.9%) during the period 1985-1993. Table 2 reports and tests the difference in mean financial conditions and other characteristics between failed and non-failed banks. Banks that failed during the S & L crisis had characteristics that were significantly different than their counterparts in 1984. They generally speaking were younger banks that had weaker performing loans, less liquidity, a higher reliance on jumbo CDs and brokered deposits among their liabilities, and higher asset concentration in commercial real estate and C & I loans. Interestingly, failed and non-failed banks had similar levels of equity relative to total assets (8.9%), which reflects the overall weakness of the banking system subsequent to years of rising interest rates. We found no significant differences in federal funds purchased or sold, volatile liability expenses, and the shares of assets in either consumer or agriculture loans.

As of year-end 2008, there were 7445 banks in our sample, and of these banks 424 failed (5.7%) in the period 2009-2014. We find similar generalizations in the recent crisis as to the S & L crisis – failing banks in 2008 were again younger and had weaker performing loans, less liquidity, a higher reliance on jumbo CDs and brokered deposits among their liabilities, and a higher share of commercial real estate loans. Failed banks during the recent crisis also had lower shares of consumer and agriculture loans. Failed banks in the most recent crisis though were significantly less capitalized (7.9%) than their counterparts (11.5%), and were even less capitalized than banks during the S & L crisis.

A few other generalizations appear between the two crisis periods. Banks in the later crisis period were less reliant on core deposits, as their share among total assets decreased among non-failed (failed) banks from 15.8% to 10.8% (16.8% to 6.2%). The reduction in demand deposits, along with reductions in reserve requirements over time, also likely contributed to the decrease in cash held by banks. This reduction in cash though was also part of a more general trend of banks holding fewer liquid assets – securities held as a share of assets declined from 28% to 20% (15.8% to 11%) among non-failed (failed) banks. There were also notable changes to the composition of banks' loan portfolios over time. The shares of C & I, consumer, and agriculture loans decreased, whereas the shares of commercial real estate and non-commercial real estate (omitted category) increased.

[Insert Table 2 about here]

5. Results and discussion

5a. Logit models with rolling year-ahead predictions

The goal of our analysis is to examine whether the bank failure experience observed during the S & L crisis is able to predict in 2008 the failures observed during the financial crisis. Our first analysis uses logistic regression to try and identify those banks that are at imminent risk of failure at year-end 2008. Similar to the Federal Reserve's FIMS model, we use a cross-section of banks' financial data at year-end to estimate whether failure occurs in the following year. Two approaches are used to select variables for the model's specification for comparison purposes – the first uses Bayesian model averaging and the later, similar to the FIMS model specification (Cole et al., 1995) applies stepwise selection, which is determined here on the basis of Akaike's information criterion.

For each of the years 1984-1991, we estimate a rolling prediction model of bank failures in the year ahead with the two variable selection approaches. That is we use year-end data from 1984 to predict whether failure occurs in 1985, and then use 1985 year-end data to predict failures in 1986, and so on. Results (reported in an appendix) from the BMA analysis reveal there exists uncertainty as to the model's true specification with the number of specifications averaged over ranging from 4 (1984) to 56 (1985), with an average of 27 specifications.[11] Even when uncertainty is minimal, e.g. 1984, the two approaches suggest that different causal factors are important to failure. The model chosen by stepwise selection using the 1984 data includes several measures (federal funds sold, provisions for loan losses, and securities) that are not averaged over in the BMA model and includes age and cash, which receive little support from the data under BMA. Across the years, we find based on our BMA estimates evidence against there being an effect from federal funds purchased, volatile liability expense, and consumer loans as each has a posterior effect probability (PEP) less than 5%. The measures of cash, charge-offs, insider loans, and age each have PEPs less than 50%, which indicate they do not receive support for having an effect on failures.

[Insert Table 3 about here]

Our interest here though is primarily on the models' out-of-sample predictive performance. For this reason, we focus less on the potentially different causal relations implied by the two approaches, i.e. differences in the coefficient estimates and their standard errors, and are less concerned with issues of multicollinearity in our candidate regressors. Despite the observed differences in the specifications across the two approaches, the out-of-sample type – 1 versus type– 2 error rates shown in Figure 1 are quite similar for predictions throughout the S &

[11] An online appendix contains the logit model estimates from BMA and stepwise selection for each of the years 1985-1992.

L crisis period (1984-1992). To be clear, out-of-sample predictions reported in Table 4 for 1985 are determined using the model estimates from 1984, along with financial data from year-end 1985 to predict failure outcomes in 1986. For a given type – 2 error rate of 10% there is a type – 1 error rate that ranges over time between 2% and 23% for BMA and ranges between 2.5% and 21% for the stepwise model (see Table 4). BMA is shown to have a lower type – 1 error rate in three of the 8 rolling predictions and is equal in another year, relative to the stepwise model. As noted, there is little definitive difference in the predictive discrimination of the two variable selection methods across the range of type – 2 errors as shown in Figure 1 for the period of S & L crisis.

[Insert figure 1 and Table 4 about here]

An issue with rolling predictions of the type above is they rely on failures in the previous period to make predictions in the next. That is our model requires failure episodes to estimate the model, and in the years (1993-2008) following the S & L crisis there were few bank failures in any given year to update the models' estimates. We therefore examine in Figure 2 how well the prediction models' estimates from the S & L crisis years are able to predict failures in 2009 based on year-end data from 2008, i.e. early in the crisis. We find that the accuracy of the models diminishes for predictions in the recent crisis relative to predictions during the S & L crisis, which is not surprising given the predictions are based on estimates from more than seventeen years in the past, rather than the previous year. Comparison in Table 4 of out-of-sample type -1 and type – 2 errors at the start of the most recent crisis period though reveal that the estimates of BMA models outperforms stepwise regression in each of the years for a given type – 2 error rate of 10%, and in some years by a wide margin, e.g. 9% in 1986. The difference in the two models predictive discrimination is also generalizable as shown in Figure 2.

Predictions in 2008 from BMA produce lower type – 1 errors over a wide range of type-2 errors for models estimated using data from years 1984-1987, 1989, and 1991.

These results suggest BMA estimates from an earlier crisis period outperform stepwise estimates when applied to predicting whether banks will fail in a subsequent crisis period. Policymakers can use the model's estimates to create an early warning prediction. Using the estimates and a cutoff from the 1987 model, which provides the lowest type – 1 error rate for a type – 2 error rate of 10% during the S & L crisis, the model at year-end 2008 identifies 452 banks as likely to fail in the next year and correctly classifies 97 of the 119 (82%) banks that actually fail in 2009.[12] The limitation of the logit model is that it is not well suited for predicting failures through time given only a set of initial conditions. For this we turn to a survival model to account for the timing of failures during a crisis.

[Insert Figure 2 about here]

5.b. Cox model of bank failure.

We next use bank financial data from year-end 2008 and a Cox proportional hazards model to predict banks time to failure during the most recent crisis period (2009-2014). Similar to the estimation of our logit models, we use the failure experience observed during the S & L crisis to build an estimation model that will be applied out-of-sample at the start of the most recent financial crisis. The Cox model is estimated using bank financial data that is available year-end 1984 to predict this set of banks' times to failure through year-end 1993. We limit estimation of our model to the use of these "initial" conditions, which allows for our predictions to be based only on the information available to policymakers at the start of the crisis. Banks

[12] The corresponding type-2 error is equal to 5%. The error rates reported in Table 4 differ as they are based on outcomes and thus cutoffs that are not observed at year-end 2008, i.e. they are only known to policymakers at year-end 2009.

that do not fail during the period examined are said to be censored, as we do not observe their time to failure. Censoring may occur due to merger, a change in charter and hence reporting requirements, or voluntary closure. Banks are also said to be censored if they remain in the sample through the end of the period examined without failing. Banks are followed through year-end 2014, i.e. after the crisis itself to account for the often slow resolution of failures.

Estimates of the Cox Model using Bayesian model averaging and stepwise selection appear in Table 5. The BMA results indicate that nineteen specifications were averaged over and that the specification with the highest posterior model probability has a 34% likelihood of being the true model that generates the data. The model specification chosen by stepwise selection was not included in the set of models averaged over and includes a number of variables that did not receive support under BMA. The stepwise model includes cash, deposits, non-interest expense, dividends, and federal funds sold as a share of total assets – each measure is less than 25% likely to have an effect under BMA based on their posterior effect probability. In the case of dividends, where the posterior effect probability reported is less than 5%, we would conclude there is evidence against an effect.

[Insert Table 5 about here]

In-sample predictions are shown in figure 3 to be quite similar across the two variable selection approaches. The figure depicts the relation between type – 2 and type – 1 errors based on Heagerty and Zheng's (2005) definitions of incident sensitivity and dynamic specificity measured a year (365 days) at risk, i.e. year-end 1985. For a dynamic false positive rate (type – 2 error) of 10%, we find a corresponding type – 1 error rate of 47% (incident sensitivity is 53%) from our BMA and stepwise models. We also display in Figure 4 the area under the ROC curve over time to measure our models' accuracy. The AUC is equal to 0.82 for both BMA and stepwise models at the one year mark, which indicates banks that fail at one year at risk are 82% more likely to have a risk score higher than their counterparts that survive. We find that the two models' risk scores based on their initial financial conditions also provide good long-term predictive power as the AUC remains above 0.80 through time, which indicates banks' initial conditions are as well suited to discriminate between failures and non-failures early into a crisis as they are later on.

[Insert figures 3 and 4 about here]

Our interest here though is in the ability of our models' estimates to predict the failure experience of banks during the most recent financial crisis, i.e., out-of-sample. To determine this we apply our model estimates based on the S & L crisis and data from year-end 2008 to predict failures during the recent financial crisis. That is to say we rely only on information, model estimates and data, available to policymakers at the start of the crisis. The predictive accuracy of the BMA model is similar out-of-sample to the in-sample predictions, i.e. the AUC is greater than 0.78 during the entire time banks are at risk (Figure 4). The stepwise model performs substantially worse out-of-sample and in relation to the BMA model. A comparison of type – 1 and type – 2 errors at 365 days show for any type – 2 error that BMA has a lower type – 1 error. The AUC for the stepwise model is approximately equal to 0.66 throughout the time period examined. The results here demonstrate that the initial conditions of banks at a start of a crisis and estimation models from a previous episode are useful for failure predictions during a crisis. We further find that Bayesian model averaging improves our out-of-sample predictions.

To test the sensitivity of our prediction model to the choice of initial conditions, we re-estimated the model with year-end 1985 financial data, i.e. banks become at risk of failing starting year-end 1985. The BMA estimates (reported in an appendix) indicate 24 models were averaged over and the best specification was 25% likely to be the true model. Predictive accuracy of both the stepwise and BMA estimates are similar in-sample – a dynamic false positive rate of 10% coincides to an incident sensitivity of 55% for each model. We find out-of-sample that BMA outperforms stepwise estimation producing lower type – 1 errors (higher sensitivity) for a given type – 2 error rate. The type – 1 error of our BMA estimates is 53% as compared to 62% from stepwise selection. Our estimates further reveal the predictive accuracy is similar over time. The AUC ranges between .84 and .80 for our BMA estimates and between .76 and .75 for stepwise. It appears our Cox model's predictions from using BMA are not particularly sensitive to our choice of using estimates from the model estimated with 1984 data.

Policymakers can use the Cox model's estimates to draw inferences as to the predicted survival experiences of banks during subsequent crises. Below we compare the experiences of a representative healthy, unhealthy, and critically ill bank, where healthy banks are defined as having a risk score equal to the average value of banks that did not fail during the S & L crisis, unhealthy banks have a score equal to the average of those who failed more than a year later, and critically ill are those who failed within the first year.[13] The risk score is equal to the linear portion of the proportional hazard and is useful (Hosmer et al., 2008) in comparing the survival experiences of different representative banks. Figure 5 provides the survival experience for each of the representative risk scores. In-sample, during the S & L crisis, the BMA model's estimated risk scores classify 3.7% of the population of banks as critical, 7.8% as unhealthy, and 88.5% as healthy at year-end 1984. For this same period we observe that .7% of banks failed within a year, and another 7.1% would fail more than a year into the crisis, while 92.1% remained healthy throughout. The model provides a quite accurate overall assessment of risk to the banking

[13] Whalen (1991) uses a similar comparison of banks to describe the survivor functions in-sample. Lane (1984) also compares the survival experience of failed and non-failed banks.

system, based only on the information found in banks' conditions at the start of the crisis. If one applies the same classification criterion from 1984 to banks out-of-sample in 2008, we predict that 2.1% of banks are in critical condition, 6.9% are unhealthy, and 91% are healthy as of year-end 2008.[14] This is compared to the 1.6% of banks that failed within a year, the 4.1% of banks that would fail later during the crisis, and the 92% that remained healthy. The difference in the model's performance out-of-sample with respect to the classification of unhealthy banks, i.e. prediction of more failures than occurred, is likely a result of the unprecedented interventions taken by the government to stabilize the banking system. Therefore the model's predictions should be viewed as an "early warning" of what may occur prior to any intervention.

[Insert figure 5 about here]

5.c. Predictions for the next crisis.

Our results above suggest that when the next banking crisis strikes, policymakers should look to the failure experience of the most recent financial crisis to guide their predictions of bank failures. Table 6 reports the Bayesian model averaging estimates of the logit and Cox models applied to the most recent crisis period's data. The logit model is estimated using year-end data from 2008 to predict whether banks fail in 2009. Figure 6 displays the specifications' in-sample accuracy, relating the relationship between type – 1 and type – 2 errors. For a type – 2 error rate of 10%, the corresponding error rate is equal to 5.9% and the area under the ROC (AUC) is equal to .98. Several of the variables that receive positive support (posterior effect probability > .75) for inclusion in the model estimated using 2008 data coincide with the 1984 specification (nonaccrual loans, foreclosed real estate, and equity). A number of variables though that received very strong support for inclusion in 1984 were not averaged over in 2008 – these

[14] The risk score are estimated relative to the typical bank for each period examined with the classification cutoffs based on the risk scores and failure experience observed during the S & L crisis.

variables include loans past due 90 or more days, jumbo CDs, bank size, and the shares of C & I and agricultural loans. Receiving positive support and added to the 2008 model's specification were measures of earnings and the brokered deposits indicator.

The Cox model is applied to the most recent crisis to estimate the time to bank failure between year-end 2008 and year-end 2014 when controlling for year-end 2008 bank data. We find there is a great deal of uncertainty in the true model's specification as 113 different models are averaged over by BMA. Each of the candidate control variables, other than size and federal funds purchased, is averaged over in the 2008 BMA model. Measures that received positive support for inclusion in the 2008 model and not the 1984 model included cash, deposits, charge-offs, brokered deposits, dividends, and the share of consumer loans. The shares of C& I and agriculture loans, along with age and size, which received very strong support for inclusion in 1984 had minimal posterior effect probabilities in 2008. The in-sample accuracy of the model at 365 days is plotted in Figure 6. We find for a cumulative type - 2 error rate of 10%, an incident sensitivity of 72% (type – 1 error of 28%). The AUC(t) at 365 days is equal to .94 and remains above .87 throughout the 6 years banks are at risk. The estimates can also be used to update the risk scores of banks that are healthy, unhealthy, and in critical condition – these risk scores, measured relative to the mean, are equal to -0.226, 3.160, and 5.257, respectively. When the next crisis strikes, policymakers can use the Cox model's coefficients from Table 6, along with the most recent bank financial data to create predicted risk scores, and apply the cutoffs above to determine the range of risks to bank failure.

6. Conclusions

Prior to the recent financial crisis, one might have viewed the S & L crisis of the 1980s as a "one-off" given the otherwise rarity of bank failures in the United States following the Great

Depression of the 1930s. The bank failures that accompanied the recent financial crisis and great recession reminded us that the banking system is always at risk and therefore policymakers need to be able to understand banks' exposure to failure early into a crisis period. Fundamentals, reflected in the CAMELS acronym, are good predictors of failures. Their relative importance though may change over time, which creates uncertainty as to one's a priori choice of model specification. Here we used the failure experience of banks during the S & L crisis to build a prediction model applied to the recent financial crisis that accounted for specification uncertainty. When Bayesian model averaging is used, the accuracy of the logit model's estimates of bank failures during the S & L crisis are shown to improve the prediction of failures in 2009. Further, we show that BMA also improves the out-of-sample predictive performance of survival times with the Cox model for the period year-end 2008 – 2014.

Based on the experience during the S & L crisis, policymakers at year-end 2008 with the data available to them would have been able to use a logit model estimated from the S & L crisis correctly identify 97 of the 119 (82%) banks that actually fail in 2009 with a type – 2 error rate of 5%. The Cox model's estimates allow policymakers to think more generally about failure patterns over different periods of time. Policymakers in 2008, using our estimates, would identify that 2.1% of banks are in critical condition, 6.9% are unhealthy, and 91% are healthy, which is quite comparable, despite the massive intervention, to the 1.6% of banks that failed within a year, the 4.1% of banks that would fail later during the crisis, and the 92% that remained healthy. When the next crisis strikes, policymakers should look to recent events to base their inferences.

References

Arena, M., 2008. Bank Failures and Bank Fundamentals: A Comparative Analysis of Latin America and East Asia during the Nineties using Bank-Level Data. Journal of Banking and Finance 32, 299-310.

Betz, F., Oprica, S., Peltonen, T.A., Sarlin, P., 2014. Predicting Distress in European Banks. Journal of Banking and Finance 45, 225-241.

Brown, C.O., Dinc, I.S., 2005. The Politics of Bank Failures: Evidence from Emerging Markets. Quarterly Journal of Economics 120, 1413-1444.

Cleary, S., Hebb, G., 2016. An Efficient and Functional Model for Predicting Bank Distress: In and Out of Sample Evidence. Journal of Banking and Finance *64*, 101-111.

Cole, R. A., Cornyn, B. G., Gunther, J. W., 1995. FIMS: A New Monitoring System for Banking Organizations. Federal Reserve Bulletin, 81: 629-667.

Cole, R.A., Gunther, J.W., 1998. Predicting Bank Failures: A Comparison of On- and Off-Site Monitoring Systems. Journal of Financial Services Research 13, 103-117.

Cole, R.A., White, L.J., 2012. Deja Vu All Over Again: The Causes of U.S. Commercial Bank Failures This Time Around. Journal of Financial Services Research 42, 5-29.

Cole, R. A., Wu, Q., 2010. Is Hazard or Probit More Accurate in Predicting Financial Distress? Evidence from U.S. Bank Failures. Working paper available at SSRN.

Collier, C., 2003. The SCOR System of Off-Site Monitoring: Its Objectives, Functioning, and Performance. FDIC Banking Review 15, 17-32.

DeYoung, R., Torna, G., 2013. Nontraditional Banking Activities and Bank Failures during the Financial Crisis. Journal of Financial Intermediation 22, 397-421.

Edwards, W., Lindman, H., Savage, L. J., 1963. Bayesian Statistical Inference for Psychological Research. Psychological Review 70, 193-242.

Fernandez, C., Ley, E., Steel, M. F. J., 2001. Model Uncertainty in Cross-Country Growth Regressions. Journal of Applied Econometrics 16, 563-576.

Greene, W. H., 1997. Econometric Analysis, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Heagerty, P. J., Saha-Chaudhuri, P., 2015. Package 'risksetROC'. Available at cran.r-project.org/web/packages/risksetROC/risksetROC.pdf

Heagerty, P.J., Zheng, Y., 2005. Survival Model Predictive Accuracy and ROC curves. Biometrics 61, 92-105.

Hosmer, D. W., Lemeshow, S., May, S. 2008. Applied Survival Analysis: Regression Modeling of Time to Event Data. 2nd Edition. New York: Wiley

Jin, J.Y., Kanagaretnam, K., Lobo, G.J., 2011. Ability of Accounting and Audit Quality Variables to Predict Bank Failure during the Financial Crisis. Journal of Banking and Finance 35, 2811-2819.

Lane, W.R., Looney, S.W., Wansley, J.W., 1986. An Application of the Cox Proportional Hazards Model to Bank Failure. Journal of Banking and Finance 10, 511-531.

Leamer, E. E., 1978. Specification Searches: Ad Hoc Inference with Non-experimental Data. New York: Wiley.

Mayes, D.G., Stremmel, H., 2014. The Effectiveness of Capital Adequacy Measures in Predicting Bank Distress. SUERF Study 2014/1. Brussels: Larcier.

Ng, J., Roychowdhury, S., 2014. Do Loan Loss Reserves Behave Like Capital? Evidence from Recent Bank Failures. Review of Accounting Studies 19, 1234-1279.

Raftery, A. E., 1995. Bayesian Model Selection in Social Research, in *Sociological Methodology 1995*. (editor P. V. Marsden) , Cambridge MA.: Blackwells Publishers (111-195).

Raftery, A. E., Hoeting, J., Volinsky, C., Painter, I., Yeung, K. Y. 2018. Package "BMA". Available at https://cran.r-project.org/web/packages/BMA/BMA.pdf.

Raftery, A. E., Madigan, D., Hoeting, J. A., 1997. Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association 92, 179-191.

Raftery, A.E., Madigan, D. & Volinsky, C. T., 1995. Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with discussion). In Bayesian Statistics 5 (editors J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), Oxford: Oxford University Press (323-349).

Schwarz, G., (1978). Estimating the Dimension of a Model. The Annals of Statistics 6, 461-464.

Shumway, T., 2001. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. Journal of Business 74, 101-124.

Thomson, J.B., 1992. Modeling the Bank Regulator's Closure Option: A Two-Step Logit Regression Approach. Journal of Financial Services Research 6, 5-23.

Volinsky, C., T., Madigan, D., Raftery, A., E., Kronmal. R. A. 1997. Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke. Applied Statistics 46, 433-448.

Whalen, G., 1991. A Proportional Hazards Model of Bank Failure: An Examination of Its Usefulness as an Early Warning Tool. Economic Review 27, 21-31.

Wheelock, D.C., 1992. Deposit Insurance and Bank Failures: New Evidence from the 1920s. Economic inquiry 30, 530-543.

Wheelock, D.C., Wilson, P.W., 1995. Explaining Bank Failures: Deposit Insurance, Regulation, and Efficiency. Review of Economics and Statistics 77, 689-700.

Wheelock, D.C., Wilson, P.W., 2000. Why Do Banks Disappear? The Determinants of U.S. Bank Failures and Acquisitions. Review of Economics and Statistics 82, 127-138.

Table 1: Candidate variables for the bank failure model specification

Variable	Description	Call Report Series
Loans past due 90+ days	Loans past due 90 days or more and still accruing interest divided by assets	RCFD1407
Nonaccrual loans	Loans in nonaccrual status divided by assets	RCFD1403
Foreclosed real estate	Foreclosed real estate divided by assets	RCFD2150
Equity	Equity divided by assets	RCFD3210
Net income	Income before income taxes and discontinued operations divided by assets	RIAD4301
Securities	Available for sale and held to maturity securities divided by assets	RCFD1754 + RCFD1773
Loan loss reserves	Allowance for loan and lease losses divided by assets	RCFD3123
Jumbo CDs	CD greater than or equal to $100,000 divided by assets	RCON2604
Cash	Cash and balances due from depository institutions divided by assets	RCFD0010
Demand deposits	Total demand deposits divided by assets	RCON2210
Federal funds purchased	Federal funds purchased divided by assets	RCONB993 + RCONB995
Volatile liability expense	Interest paid on federal funds purchased and large CDS divided by assets	RIAD4190 + RIADA517
Charge-offs	Charge-offs divided by assets	RIAD4635
Brokered deposits	Indicator variable equal to 1 if the ratio of brokered deposits to total assets is greater than 1% and equal to 0 otherwise	RCON2365
Non-interest expense	Non-interest expense divided by assets	RIAD4093
Insider loans	Loans to insiders divided by assets	RIAD4093
Dividends	Dividends divided by assets	RCFD6164
Age	Age of the bank in years	RSSD9950
Size	Natural log of assets	RCFD2170
Provisions for loan losses	Provisions for loan and lease losses divided by assets	RIAD4230
C & I loans	Commercial and industrial lans divided by assets	RCFD1766
Commercial real estate	Commercial real estate loans divided by assets	RCON1480
Consumer loans	Consumer loans loans divided by assets	RCFD1975
Agriculture loans	Agriculture loans loans divided by assets	RCFD1590
Federal funds sold	Federal funds sold loans divided by assets	RCONB987 + RCONB989

Series are divided by total assets (RCFD2170) where noted. Prior to 2002 federal funds purchased (sold) were reported in series RCON2800 (RCON1350). Volatile liability expense consists the sum of (RIAD4190, RIAD4174) prior to 1997, and securities were the sum of (RCFD0390, RCFD2146) prior to 1994.

Table 2: Difference in mean characteristics of failing and non-failing banks.

	S & L Crisis		Financial Crisis	
	Failed	Non-failed	Failed	Non-failed
Loans past due 90+ days	0.0108***	0.0064	0.0039***	0.0019
Nonaccrual loans	0.0143***	0.0068	0.0547***	0.0103
Foreclosed real estate	0.0073***	0.0036	0.018***	0.0036
Equity	0.0887	0.0892	0.0789***	0.1154
Net income	-0.006***	0.0090	-0.0272***	0.0048
Securities	0.1578***	0.2842	0.1104***	0.2029
Loan loss reserves	0.0098***	0.0062	0.0195***	0.0095
Jumbo CDs	0.2182***	0.1035	0.2136***	0.1585
Cash	0.108***	0.0923	0.041***	0.0580
Demand deposits	0.1682***	0.1578	0.0624***	0.1081
Federal funds purchased	0.012	0.0136	0.0133	0.0167
Volatile liability expense	0.0915	0.0956	0.0427***	0.0363
Charge-offs	0.0123***	0.0055	0.0147***	0.0038
Brokered deposits	0.1226***	0.0248	0.8184***	0.3753
Non-interest expense	0.0396***	0.0323	0.0354	0.0329
Insider loans	0.0128***	0.0054	0.0174***	0.0151
Dividends	0.0027***	0.0037	0.0018**	0.0055
Age	37.4778***	55.8927	35.7052***	68.2487
Size	10.4049***	10.6093	12.5063***	11.9514
Provisions for loan losses	0.0144***	0.0057	0.0225***	0.0051
C & I loans	0.2112***	0.1278	0.1002	0.0976
Consumer loans	0.1305	0.1264	0.0191***	0.0455
Commercial real estate	0.0533***	0.0440	0.235***	0.1617
Agriculture loans	0.0707	0.0681	0.0106***	0.0446
Federal funds sold	0.0597	0.0598	0.0303	0.0278
Number of observations	1101	12924	424	7021

The sample of banks for each crisis period are divided between banks that either fail or do not fail during the period - S & L crisis (1985 - 1993), Financial crisis (2009-2014). The means of the controls reported are measured year-end 1984 and 2008. *, **, ***, indicate significant differences in means between failed and non-failed banks at the 10%, 5%, and 1% level, respectively

Table 3: Logit model estimates of bank failure (1984)

	Bayesian model averaging			Stepwise AIC		
	Coef	SE	PEP	Coef	SE	P-value
Constant	4.142	1.492	100	5.996	-1.683	< .001
Loans past due 90+ days	19.905	6.091	96.6	16.953	-4.973	0.001
Nonaccrual loans	17.216	3.817	100	14.124	-4.082	0.001
Foreclosed real estate	21.636	4.101	100	17.718	-4.255	< .001
Equity	-47.817	4.943	100	-43.358	-5.459	< .001
Net income	-	-	-	-	-	-
Securities	-	-	-	-4.34	-1.434	0.002
Loan loss reserves	-	-	-	-	-	-
Jumbo CDs	5.494	1.083	100	5.324	-1.029	< .001
Cash	-0.160	0.982	3.4	-6.443	-2.856	0.024
Demand deposits	-	-	-	-	-	-
Federal funds purchased	-	-	-	-	-	-
Volatile liability expense	-	-	-	-	-	-
Charge-offs	-	-	-	-	-	-
Brokered deposits	-	-	-	-	-	-
Non-interest expense	-	-	-	-	-	-
Insider loans	-	-	-	-	-	-
Dividends	-	-	-	-	-	-
Age	0.003	0.006	29.4	0.012	-0.004	0.004
Size	-0.867	0.139	100	-0.904	-0.141	< .001
Provisions for loan losses	-	-	-	7.707	-4.58	0.092
C & I loans	5.210	1.152	100	3.874	-1.115	0.001
Consumer loans	-	-	-	-	-	-
Commercial real estate	-	-	-	-	-	-
Agriculture loans	6.400	0.957	100	4.312	-0.963	< .001
Federal funds sold	-	-	-	-6.543	-2.79	0.019
Observations	14025			14025		
Models averaged over	4					

The logit model uses a cross-section of year-end bank data from 1984 to predict failures in 1985. BMA estimates reported include the posterior mean (Coef), standard deviation (SE), and effect probabilities (PEP) of the variables averaged over. The stepwise model is selected based on Akaike's information criterion (AIC).

Table 4: Out-of-sample type-1 error rates for year ahead failures

Estimates Year	Prediction Year	BMA	Stepwise	Prediction Year	BMA	Stepwise
1984	1985	9.4%	7.8%	2008	29.4%	37.0%
1985	1986	7.9%	6.3%	2008	15.1%	16.8%
1986	1987	23.2%	21.1%	2008	33.6%	44.5%
1987	1988	2.0%	3.0%	2008	8.4%	15.1%
1988	1989	4.3%	3.7%	2008	10.1%	12.6%
1989	1990	2.5%	2.5%	2008	11.8%	14.3%
1990	1991	11.8%	15.5%	2008	14.3%	15.1%
1991	1992	2.4%	4.8%	2008	11.8%	13.4%

The out-of-sample type - 1 error rate (1 - sensitivity) associated with a type - 2 error rate (1 - specificity) of 10%. The logistic specifications are estimated using year-end data from the indicated estimates year, which are then applied to year end data in the prediction year to form out-of-sample predictions.

Table 5: Cox proportional hazards model estimates of bank failure - S & L Crisis

	Bayesian model averaging			Stepwise AIC		
	Coef	SE	PEP	Coef	SE	P-value
Loans past due 90+ days	10.021	2.056	100	9.279	2.057	< .001
Nonaccrual loans	7.084	1.856	100	6.696	1.861	< .001
Foreclosed real estate	10.265	1.686	100	9.739	1.718	< .001
Equity	-1.606	1.065	77.2	-1.363	0.700	0.052
Net income	-7.495	3.014	89.6	-10.930	1.319	< .001
Securities	-4.149	0.367	100	-4.683	0.372	< .001
Loan loss reserves	0.310	1.583	4.8	-	-	-
Jumbo CDs	5.296	0.259	100	5.389	0.259	< .001
Cash	-0.097	0.359	8.6	-1.150	0.473	0.015
Demand deposits	0.321	0.649	23.7	1.602	0.512	0.002
Federal funds purchased	0.021	0.199	1.9	-	-	-
Volatile liability expense	-	-	-	-	-	-
Charge-offs	-	-	-	-	-	-
Brokered deposits	-	-	-	-	-	-
Non-interest expense	-0.897	2.184	17.8	-6.257	2.182	0.004
Insider loans	4.381	0.795	100	4.476	0.835	< .001
Dividends	0.518	2.684	4.6	11.130	6.305	0.077
Age	0.005	0.001	100	0.005	0.001	< .001
Size	-0.174	0.034	100	-0.201	0.034	< .001
Provisions for loan losses	1.406	3.616	15	-	-	-
C & I loans	2.813	0.323	100	2.389	0.308	< .001
Consumer loans	-	-	-	-	-	-
Commercial real estate	2.864	0.623	100	2.239	0.615	< .001
Agriculture loans	2.640	0.352	100	2.250	0.348	< .001
Federal funds sold	-0.259	0.576	20.4	-1.487	0.497	0.003
Observations	14025			14025		
Models averaged over	19					

Cox model estimates of the time to bank failure in days between year-end 1984 and 1993. The control variables are measured using year-end bank data from 1984. BMA estimates reported include the posterior mean (Coef), standard deviation (SE), and effect probabilities (PEP) of the variables averaged over. The stepwise model is selected based on Akaike's information criterion (AIC).

Table 6: Bayesian model averaging estimates of bank failures during the financial crisis

	Logit Model			Cox Model		
	Coef	SE	PEP	Coef	SE	PEP
Constant	0.0733	0.6639	100	-	-	-
Loans past due 90+ days	-	-	-	21.912	5.198	100
Nonaccrual loans	18.8892	3.6358	100	11.566	1.051	100
Foreclosed real estate	18.0629	7.4159	93.2	11.846	2.221	100
Equity	-72.7691	6.6055	100	-34.460	2.461	100
Net income	-29.1437	8.8030	100	-19.326	5.235	100
Securities	-	-	-	-3.183	0.696	100
Loan loss reserves	-	-	-	0.858	3.125	8.8
Jumbo CDs	-	-	-	1.720	0.532	98.9
Cash	-	-	-	-3.878	1.546	95.5
Demand deposits	-	-	-	-5.001	1.304	100
Federal funds purchased	-	-	-	-	-	-
Volatile liability expense	-	-	-	1.752	2.313	40.6
Charge-offs	-10.3831	13.4806	42.5	-12.396	5.761	92
Brokered deposits	0.9700	0.4974	86.4	0.860	0.138	100
Non-interest expense	-4.4550	8.5540	23.9	-8.337	5.920	74.3
Insider loans	-	-	-	2.589	3.497	41
Dividends	-	-	-	-36.659	24.147	79.7
Age	-	-	-	-0.001	0.002	23.9
Size	-	-	-	-	-	-
Provisions for loan losses	-1.2943	5.7544	5.9	0.774	3.417	6.2
C & I loans	-	-	-	-0.017	0.158	1.8
Consumer loans	-1.3923	4.3538	11.3	-6.523	2.403	98.5
Commercial real estate	0.1827	0.7278	7.1	1.120	0.678	81.7
Agriculture loans	-	-	-	-0.015	0.214	0.8
Federal funds sold	-	-	-	1.801	1.498	65.8
Observations	7445			7445		
Models averaged over	16			113		

The logit model estimates whether banks fail in 2009 and the Cox model estimates time to bank failure in days between year-end 2008 and 2014. The control variables for both models are measured using year-end bank data from 2008. BMA estimates reported include the posterior mean (Coef), standard deviation (SE), and effect probabilities (PEP) of the variables averaged over.

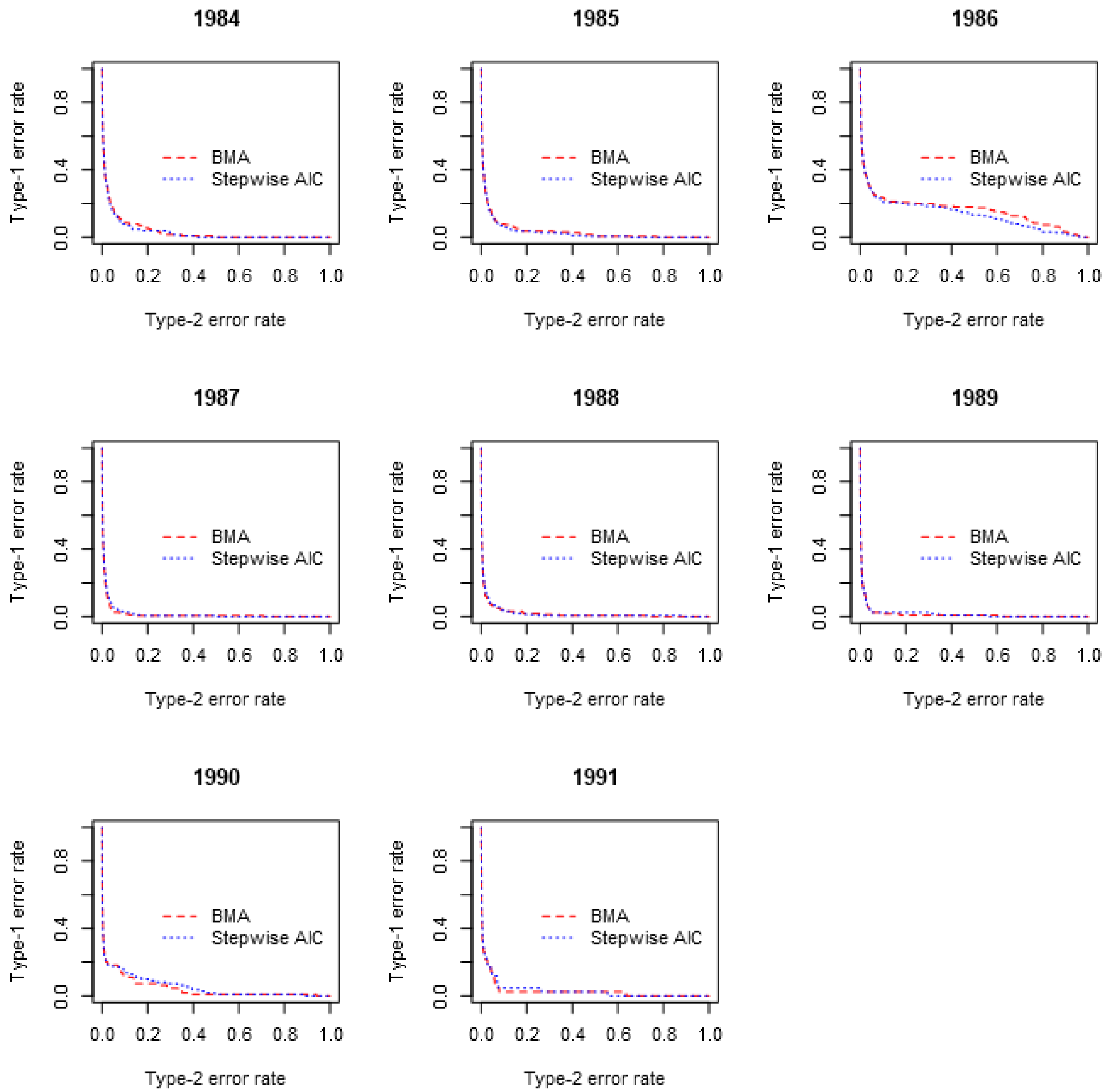

Fig. 1. The out-of-sample (year ahead) relationship between type – 2 and type – 1 errors from identifying bank failures in the year ahead using Bayesian model averaging and stepwise selection during the S & L crisis. The year indicated refers to the year-end data used to estimate the logit model of failures in the following year. These logit model estimates are then used to make out-of-sample predictions based on the subsequent years' data – e.g. (1984) refers to estimates from the specification using 1984 data to predict failures in 1985, which are then combined with year-end data from 1985 to classify out-of-sample failures in 1986.

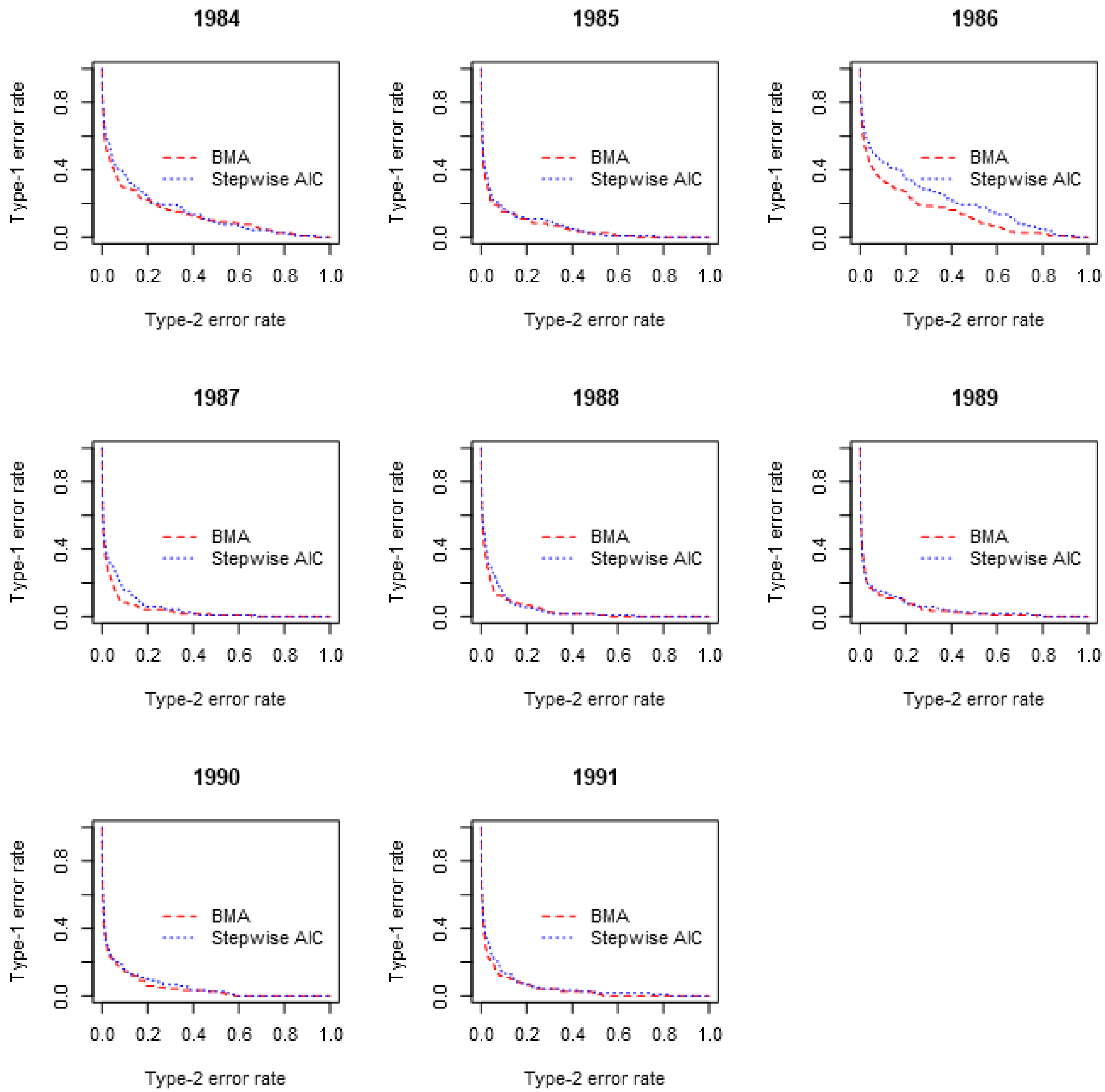

Fig. 2. The relationship between out-of-sample (2008) type – 2 and type – 1 errors from identifying bank failures in the year ahead using Bayesian model averaging and stepwise selection during the S & L crisis. The year indicated refers to the year-end data used to estimate failures in the following year of the S & L crisis. These logit model estimates are then combined with year-end 2008 data for out-of-sample prediction of failures in 2009.

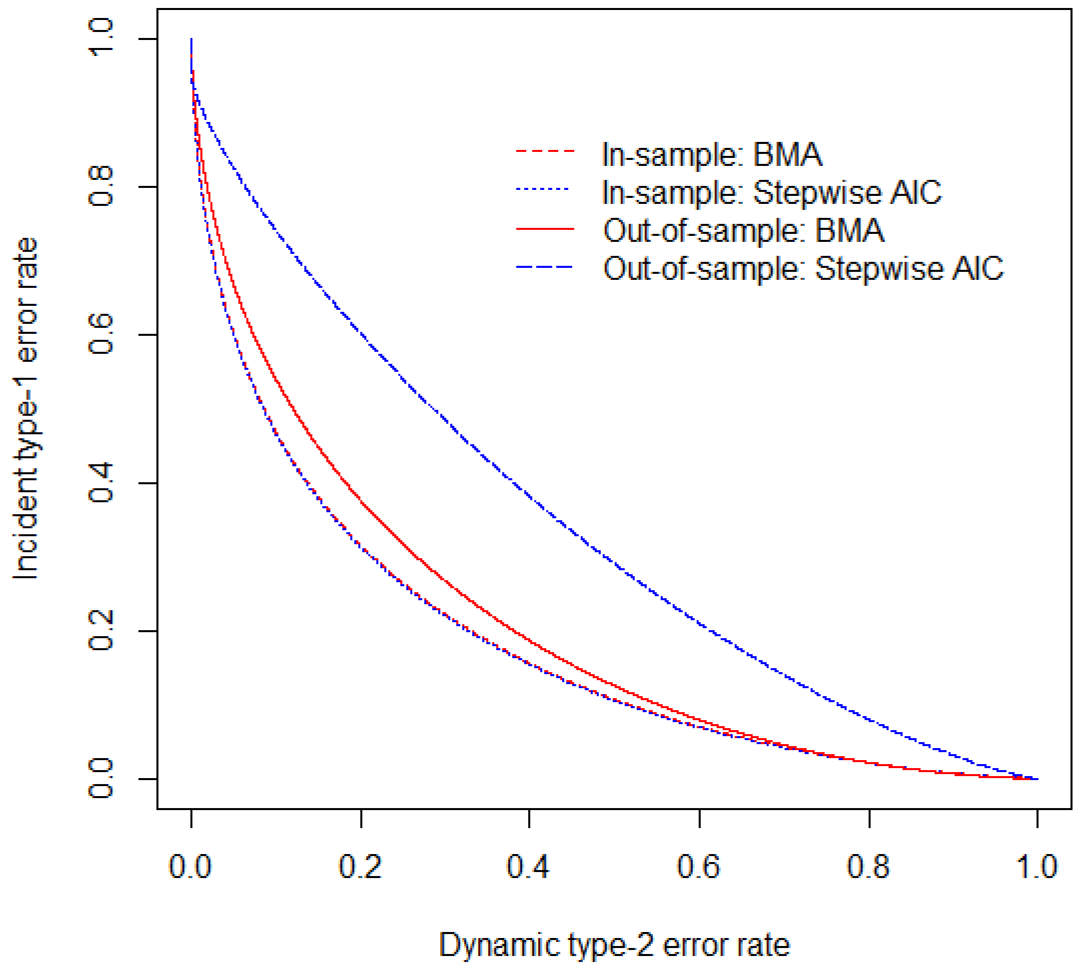

Fig. 3. The relationship at 365 days between cumulative type – 2 and dynamic type – 1 errors from a Cox model of the time to bank failure using Bayesian model averaging and stepwise selection. In-sample accuracy is measured based on the estimates of the Cox model using year-end data from 1984 and out-of-sample accuracy is measured using the in-sample estimates, combined with year-end data from 2008, to determine out-of-sample risk scores. Type – 1 errors in the survival context are defined here as equal to 1 – incident sensitivity, whereas type-2 errors are equal to 1 – dynamic specificity (Heagerty and Zheng, 2005).

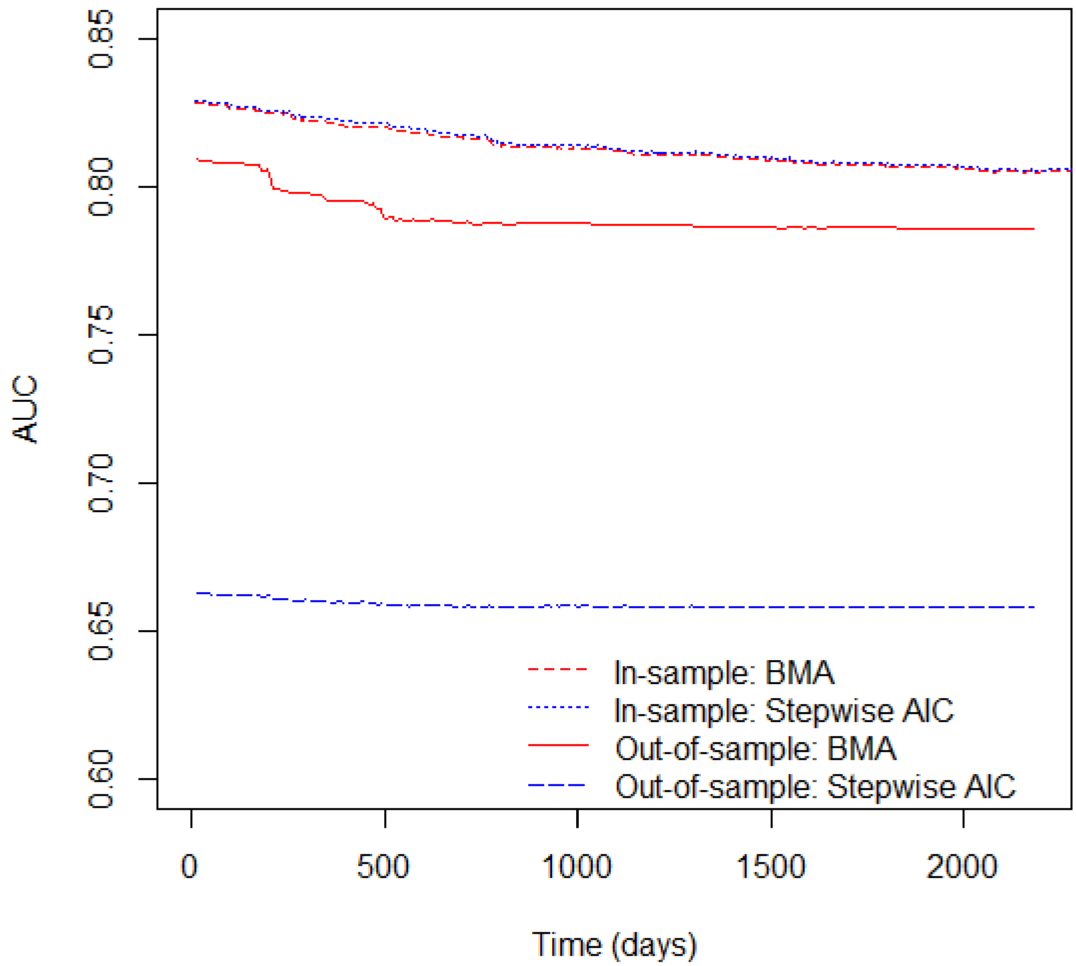

Fig. 4. Plot of the area under the incident/dynamic ROC curve (AUC) at various times for the Cox model estimated using Bayesian model averaging and stepwise selection. The in-sample AUC(t) is measured using year-end data from 1984 and out-of-sample accuracy is measured using the in-sample estimates, combined with year-end data from 2008, to determine out-of-sample risk scores, and the corresponding out-of-sample AUC.

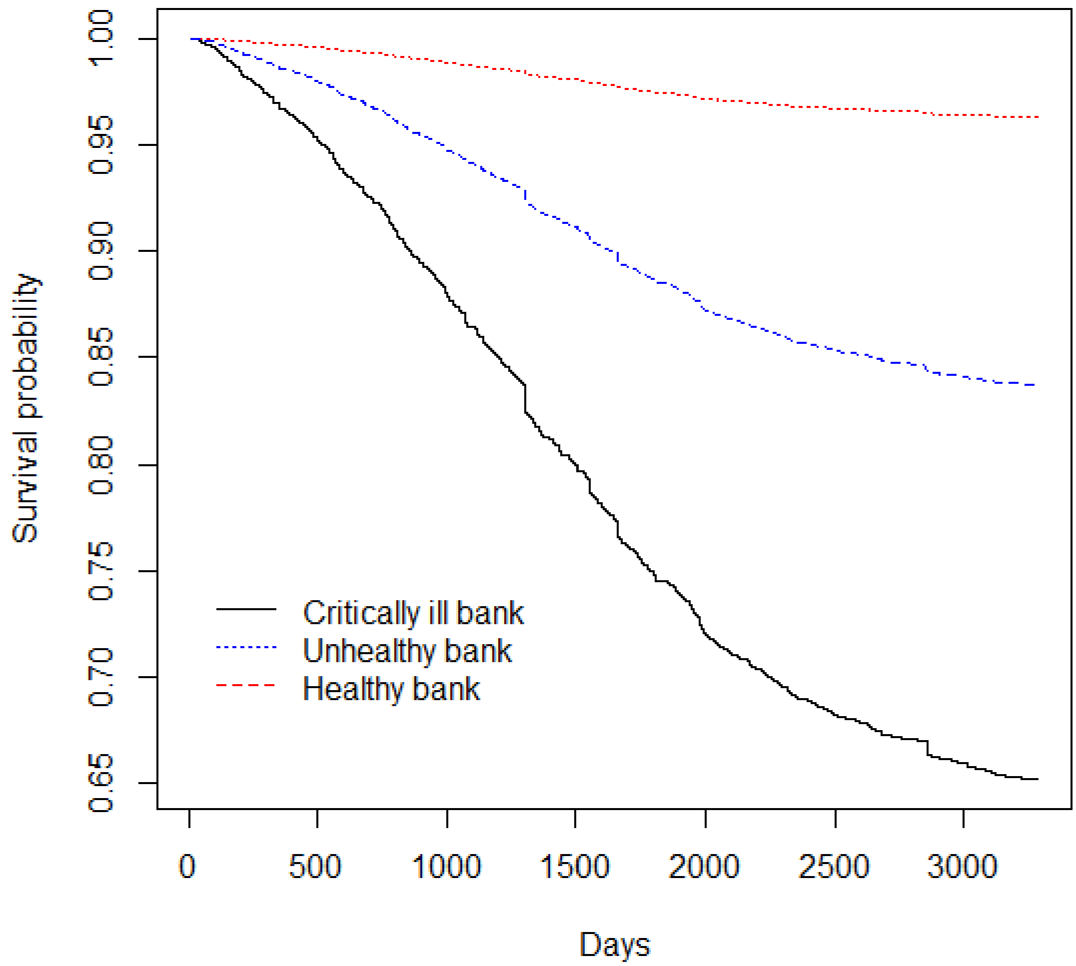

Fig. 5. Estimated survival profiles of healthy, unhealthy, and critically ill banks. Estimates are based on the survival experience of banks during the S & L crisis based on their conditions at year-end 1984. Healthy banks are defined as having risk scores equal to the mean score of banks that did not fail, whereas unhealthy banks' scores are equal to the mean risk score of banks that failed after a year, and critical banks have scores that fail within a year at risk.

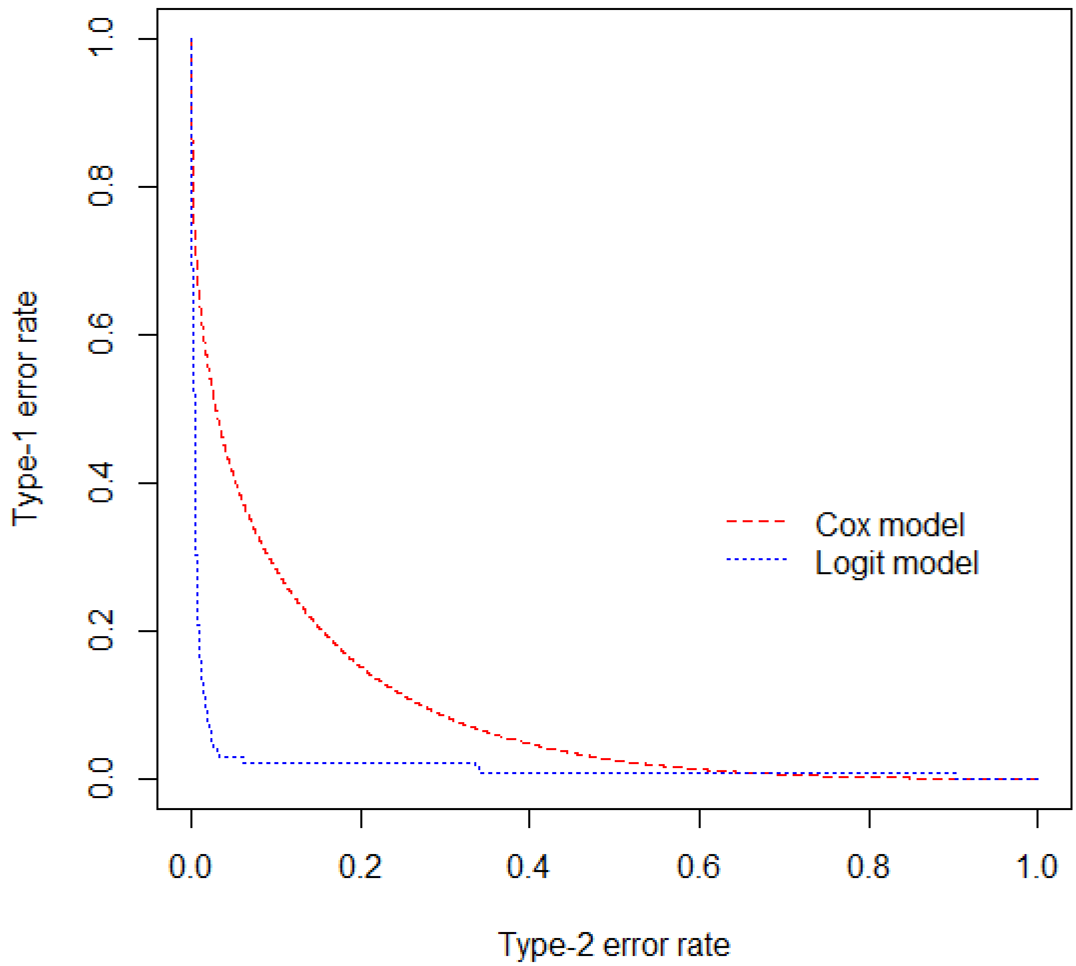

Fig. 6. The relationship between in-sample error rates of the logit and Cox models estimated using Bayesian model averaging during the financial crisis. The logit model estimates failures in 2009, controlling for 2008 year-end data. The Cox model estimates the time to failure in the period year end period 2008-2014 using only year-end data from 2008. Type – 1 errors in the survival context are defined here as equal to 1 – incident sensitivity, whereas type-2 errors are equal to 1 – dynamic specificity (Heagerty and Zheng, 2005).