MSAN 604 - Group Assignment

Andre Guimaraes Duarte, Keyang Zhang, Tim Zhou

December 13, 2016

1 Introduction

Accurately forecasting national bankruptcy rates is of interest to many parties, including banks, insurance companies, and credit lenders, for example. In this report, we used a history of monthly bankruptcy rates for Canada from January 1987 to December 2010 to forecast this data for 2011 and 2012.

The available data consists of:

- Bankruptcy Rate
- Unemployment Rate
- Population
- Housing Price Index

Figure 1 shows *time series* plots of the values in each data set. Informally, a time series is a sequence of observations that exhibits some sort of pattern over time.

Note that each of these plots seem to exhibit non-random behavior over time. This suggests that we could try and mathematically model this behavior, for the purpose of future predictions.

The goal of this report is to predict bankruptcy rates for 2011 and 2012 using the history (previous values) of these rates as well as, potentially, the other variables from 1987 to 2010.

In order to select the best model for prediction of these data, we first need to check if the model is reasonable for prediction. As such, we divide our data into a training set and a validation set, where the validation set is composed of observations from the last 3 years (2008–2010). We select the best fitted model on the training data such that its predictions are the closest to the actual values observed in the validation set. Once we have selected the best model, we re-train it using all available data and forecast for the following two years (2011 and 2012).

Figures for models that were tested for model selection, but not included in the body of this report, are included in the Appendix. R code is available upon request.

2 Model Selection and Evaluation

We first split the data into training and validation sets. We took the last 3 years of data (2008 - 2011) and used them as a validation set. Our model selection is therefore performed on a training set of bankruptcy rates from 1987 to 2007.

2.1 Model Selection

We create several models and compare them in order to select the best performing one on our validation set, which are listed below:

- Holt-Winters (Triple Exponential Smoothing)
- SARIMA
- SARIMAX
- Vector Auto-regression (VAR)

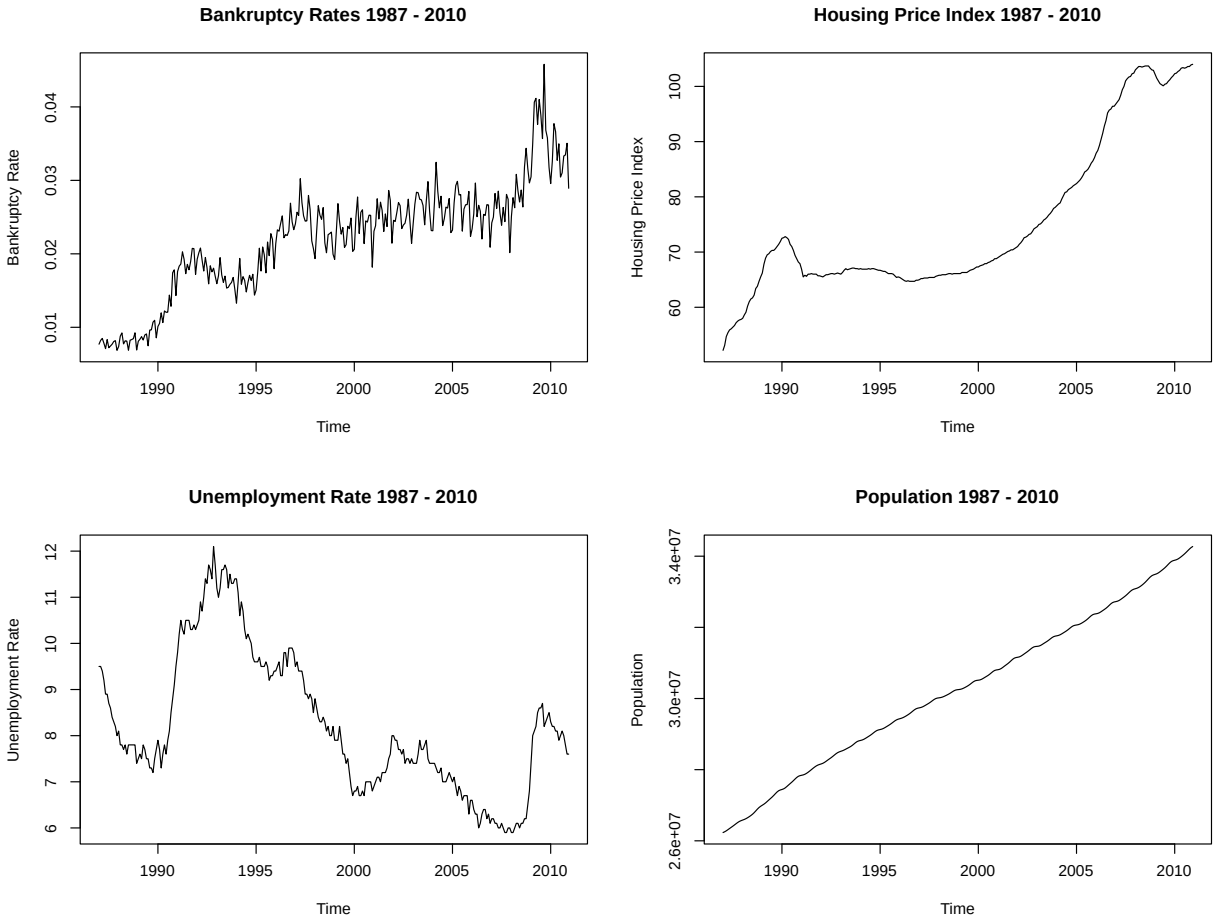

Figure 1: Time series plots of available data.

In this report, we will only explain the best performing model. The other models can be evaluated separately. In order to select the best performing model out of these, we compare the *root mean residual error* (RMSE), which is a metric that measures how closely the predictions are to the actual values in the validation set. Thus, the model with the lowest RMSE has the best forecasting power.

Table 1 shows the summarized results for RMSE and the parameters used for each model we tested. It can be seen that VAR has the lowest RMSE on our validation data. For this model, the parameter $p = 6$ means that the prediction for each observation uses a combination of the values from the previous 6 data points (in time). In addition, the log-transformed Housing Price Index data was used as an external variable to help explain the variation in bankruptcy rates (the log-transform is a mathematical process used to normalize the data).

Figure 2 shows the values predicted by our chosen model, superimposed on the actual values observed in 2008–2010, as well as the 95% prediction interval. This is an estimate of the range of values the true value of the observations will lie in, with 95% probability, given the history of data. We can see that the predictions lie close to the actual values, and that most of them fit within the prediction interval (the exceptions are addressed below). Visual representations of the predictions of other models we considered, but did not include, are shown in the Appendix.

It is pertinent to say that none of these models will do exceptionally well on this validation data. This is due to the unpredictable spike in bankruptcy rates observed after the 2008 crisis. In spite of this, the VAR (vector auto-regression) model was able to generate the best predictions on the validation set. Note that,

Table 1: Summary of results for model selection

Model	Holt-Winters	SARIMA	SARIMAX	VAR
RMSE	3.73	0.051	0.042	0.0028
Parameters	$\alpha = 0.314$ $\beta = 0.182$ $\gamma = 0.175$	$p = 4, d = 1, q = 4$ $P = 4, D = 1, Q = 1$ $s = 12$	$p = 5, d = 1, q = 3$ $P = 4, D = 1, Q = 5$ $s = 12$	$p = 6$

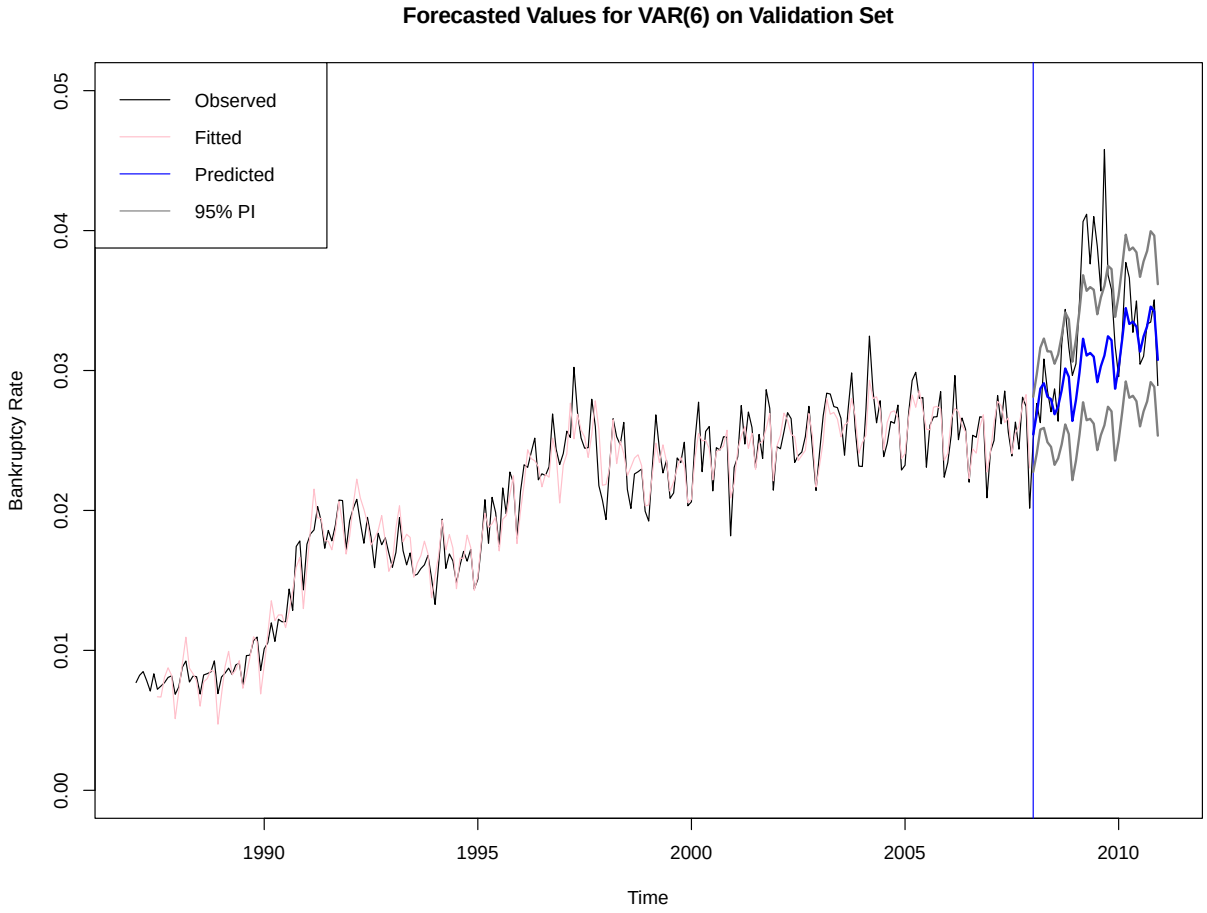

Figure 2: Comparison of predicted bankruptcy rates and actual rates for 2008 to 2010 using VAR.

disregarding the spike, the VAR model almost perfectly overlays the validation set data.

2.2 Model Evaluation

Having selected the VAR model as the best performing model, we need to verify the validity of its prediction by checking certain assumptions. In particular, we need to make sure that the residuals (i.e., the difference between the actual and fitted data) are randomly distributed around 0 over time, have constant variance, and are normally distributed. All these assumptions are shown in Figure 3, and they all pass, so we say that model is valid.

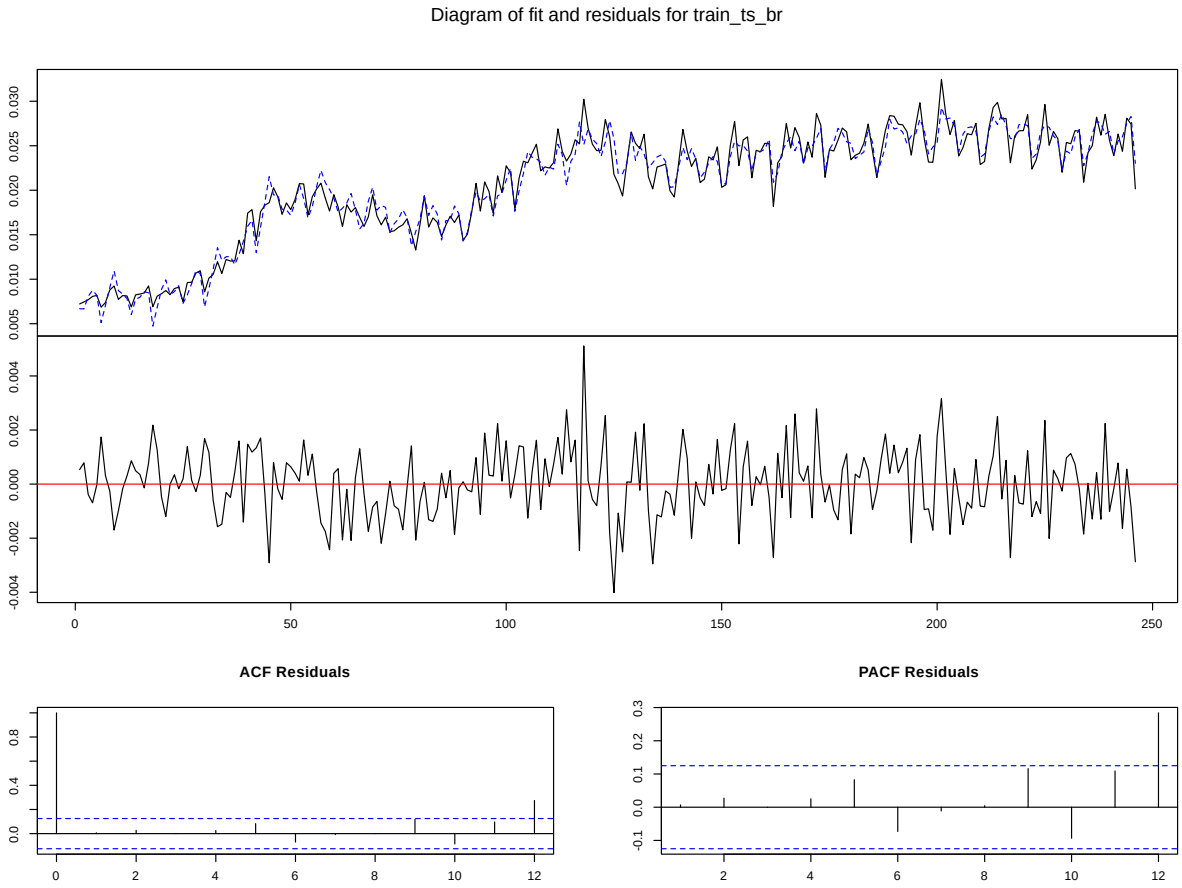

Figure 3: Residual diagnostics for VAR model.

3 Model Predictions for 2011 and 2012

Now that we have selected VAR as the best model for prediction based on the validation set, we fit this model on the entire available training data (i.e., from 1987 to 2010) and predict bankruptcy rates for the next 2 years (i.e., 2011 and 2012). We verified that the residual assumptions still hold (plots not shown in this report).

Figure 4 shows the predicted values for bankruptcy rates in 2011 and 2012, along with their 95% prediction interval. The raw values for these three elements are shown in Table A1.

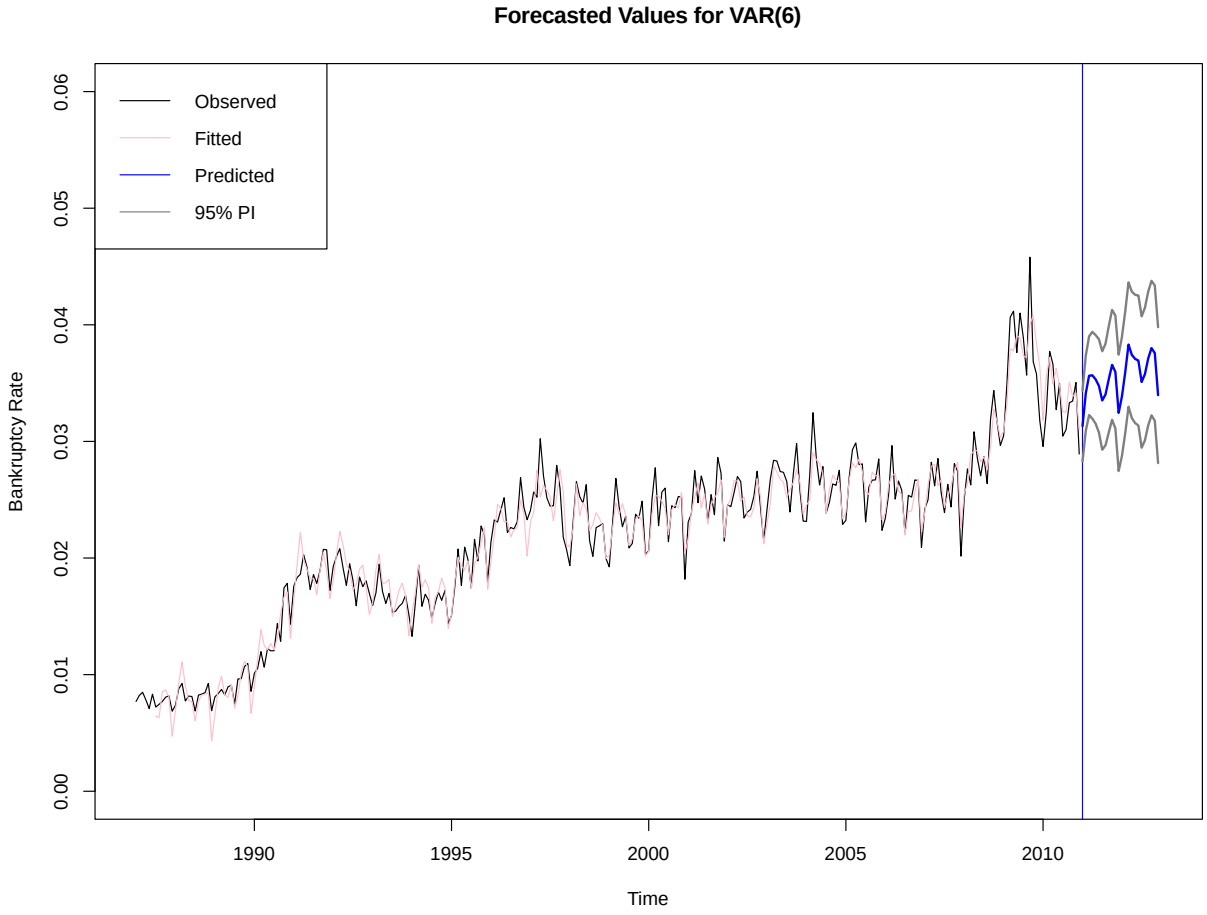

Figure 4: Predicted bankruptcy rates for 2011 and 2012 using VAR.

4 Conclusion

The goal for this project was to find and implement the best time series model in order to predict bankruptcy rates in Canada for 2011 and 2012 based on past rates and extra variables from 1987 until 2010. We first selected the best performing model based on a validation set of the last three years of available data (i.e., 2008–2010). This resulted in choosing a VAR model. Then, we fit this model on all the data and used it to predict bankruptcy rates for 2011 and 2012.

The results for this forecasting are shown in Figure 4, and the raw values in Table A1. Despite the unusual spike around 2008, we expect this model to have the best predictive performance for 2011 and 2012.

Appendix

Table A1: Summary of predictions and 95% prediction interval bounds for 2011 and 2012 using VAR.

Date	Prediction	Lower	Upper	Date	Prediction	Lower	Upper
Jan 2011	0.03131386	0.02827985	0.03434787	Jan 2012	0.03385525	0.02874734	0.03896316
Feb 2011	0.03413034	0.03091173	0.03734895	Feb 2012	0.03592709	0.03070760	0.04114659
Mar 2011	0.03563351	0.03225953	0.03900750	Mar 2012	0.03830768	0.03297871	0.04363666
Apr 2011	0.03567439	0.03195494	0.03939384	Apr 2012	0.03742757	0.03201075	0.04284438
May 2011	0.03531728	0.03151725	0.03911730	May 2012	0.03707839	0.03157922	0.04257756
Jun 2011	0.03476035	0.03076577	0.03875493	Jun 2012	0.03693461	0.03136239	0.04250682
Jul 2011	0.03351324	0.02929112	0.03773536	Jul 2012	0.03509538	0.02946285	0.04072791
Aug 2011	0.03402290	0.02966703	0.03837877	Aug 2012	0.03577773	0.03009068	0.04146478
Sep 2011	0.03533843	0.03078025	0.03989660	Sep 2012	0.03712203	0.03138961	0.04285445
Oct 2011	0.03656536	0.03185555	0.04127516	Oct 2012	0.03800400	0.03223431	0.04377369
Nov 2011	0.03595850	0.03111953	0.04079747	Nov 2012	0.03757131	0.03176975	0.04337286
Dec 2011	0.03244787	0.02745763	0.03743811	Dec 2012	0.03396806	0.02814142	0.03979469

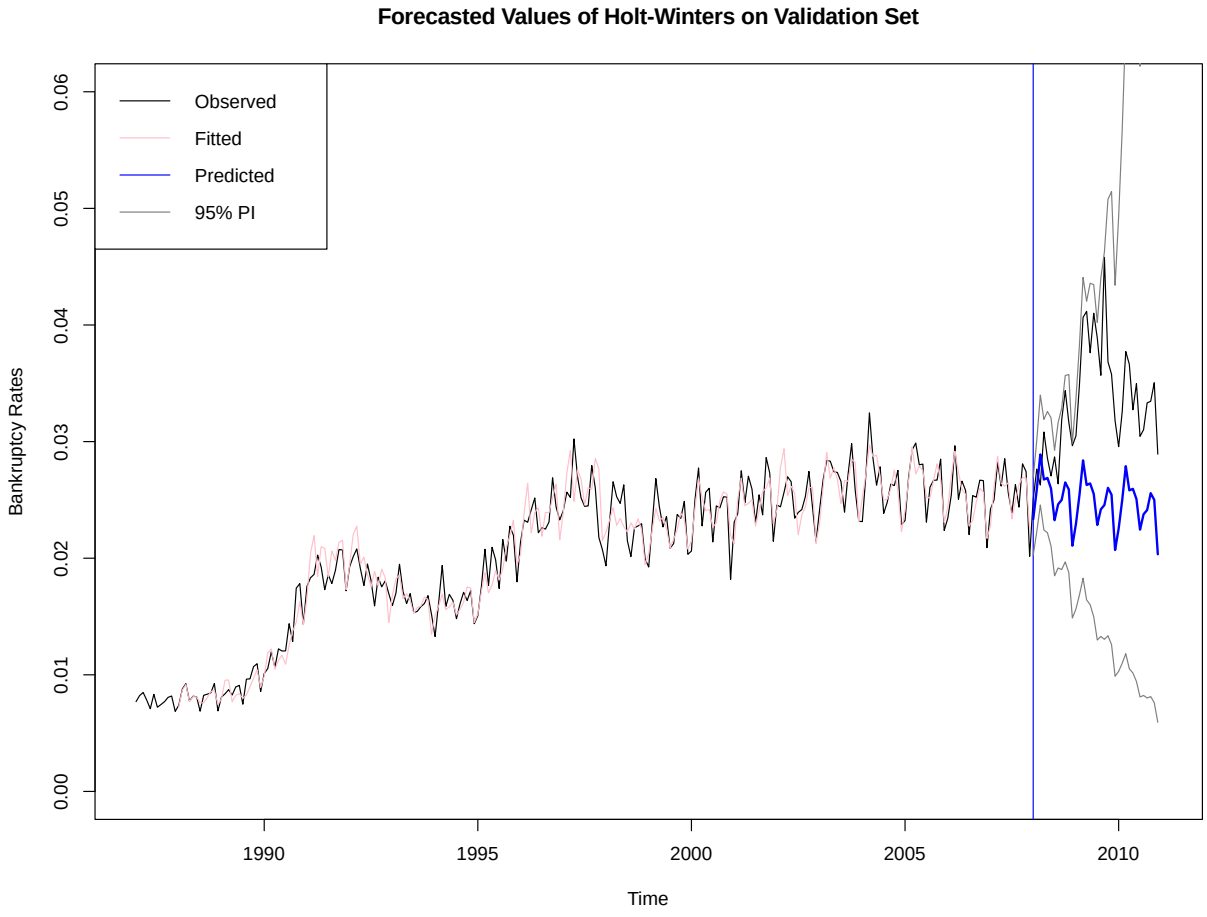

Figure A1: Comparison of predicted bankruptcy rates and actual rates for 2008 to 2010 using Holt-Winters.

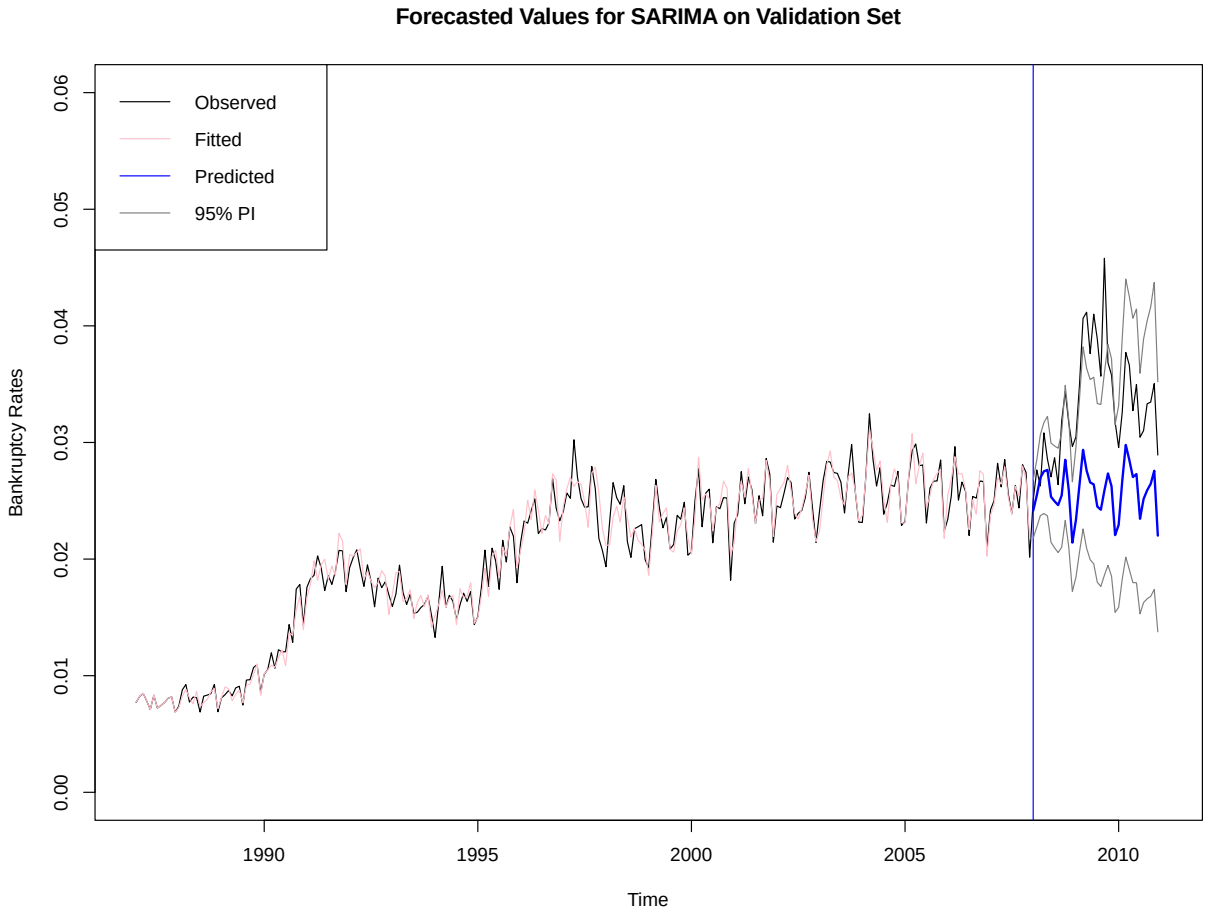

Figure A2: Comparison of predicted bankruptcy rates and actual rates for 2008 to 2010 using SARIMA.

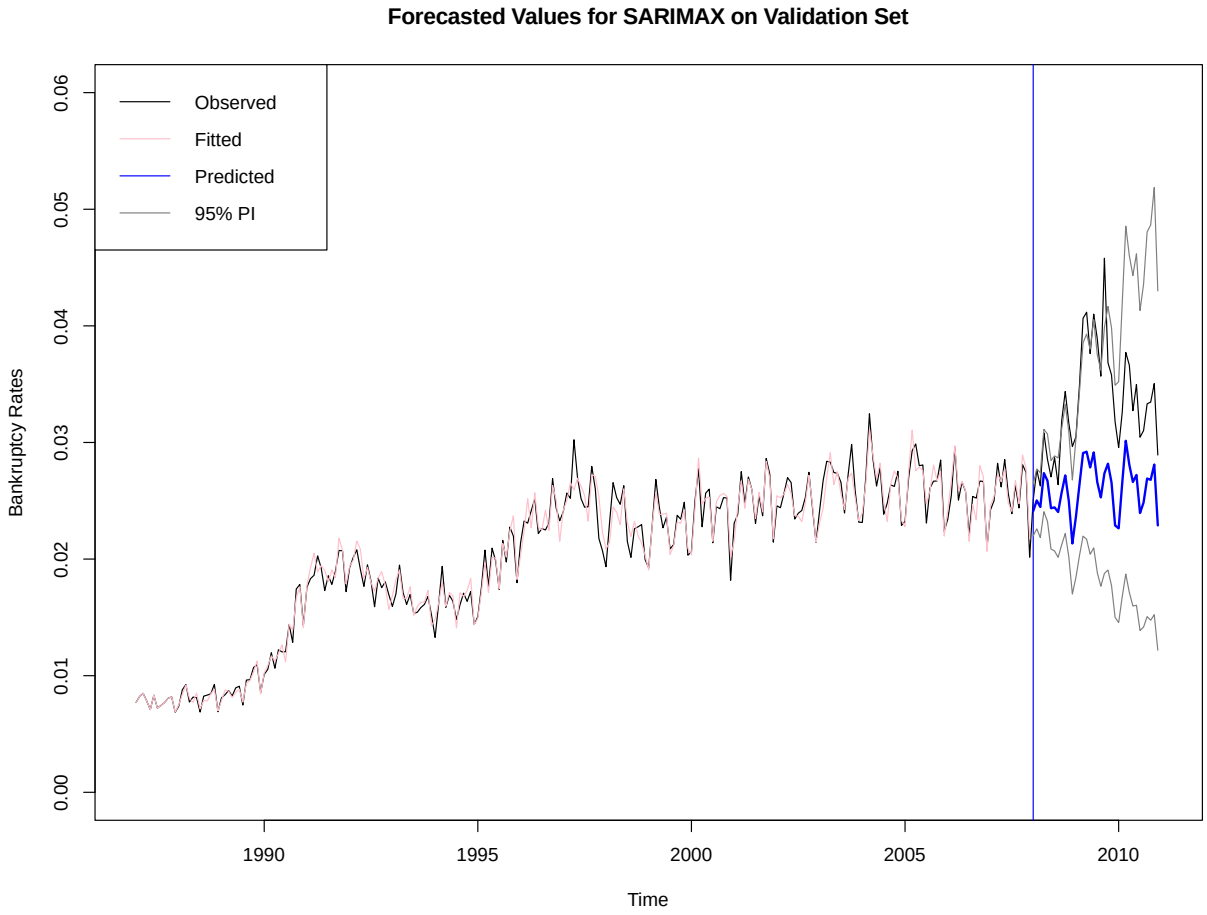

Figure A3: Comparison of predicted bankruptcy rates and actual rates for 2008 to 2010 using SARIMAX.