Canadian National Bankruptcy Rates Forecasting

Anant Agarwal, Devin Bowers, Fei Liu, Cara Qin

December 8, 2017

1. Introduction

National bankruptcy rates are of interest to multiple parties, such as insurance companies and government. Thus, it is valuable to be able to forecast these rates into the future to better inform these parties about upcoming bankruptcies. In this report, we compare multiple time series models based on monthly national bankruptcy rates, as well as monthly unemployment rate, population, and housing price index values from January 1987 through December 2010. We then use our optimal model to forecast monthly national bankruptcy rates in Canada from January 2011 through December 2012.

2. Data

Initially, we observed that trends of house price index and bankruptcy rates seem to be correlated, with bankruptcy rates leading by about an year. Population and house price index have a high positive correlation with bankruptcy rates, whereas unemployment rates appear to have a low negative correlation. We explore incorporating this correlation into our models in a later section.

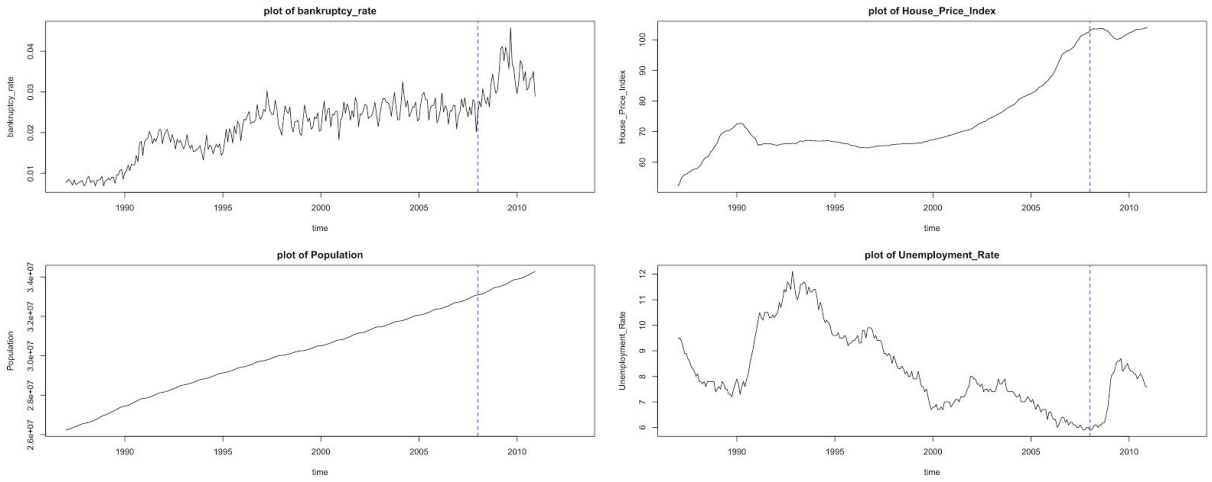


Figure 2.1 Trends of Bankruptcy rate, house price index, population, and unemployment rate

The bankruptcy rates appear to have both trend and seasonality. This is further supported by our examination of the ACF and PACF plots in Figure 2.2. We will take this into account in our model selection. In addition, we noticed that the bankruptcy rates appears to have a non-constant variance over

time. Thus, we built our models using log-transformed bankruptcy rates. This approach preserves trend and seasonality in the data, but maintains a constant variance over time.

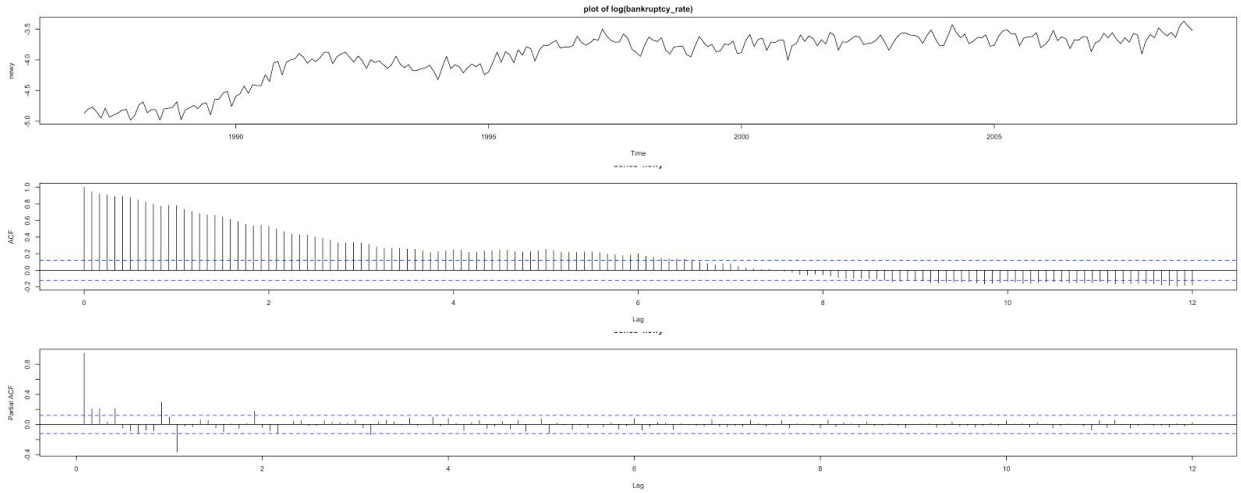


Figure 2.2 Transformed Bankruptcy rate trends, ACF, and PACF

In order to validate our model, we split the data into a training and validation set, as indicated by the blue line in Figure 2.1. We chose to use the last two years of data (January 2009 to December 2010) as our validation set since it is the same length of time as our test set.

3. Methodology

Since the bankruptcy rate exhibits both trend and seasonality, we explored the following four modeling approaches: SARIMA, SARIMAX, Vector Autoregression, and Holt-Winters. We selected one optimal model from each method and verified the validity of the model assumptions. We then selected the final optimal model from these four models based on the RMSE values and retrained the model using the complete dataset.

3.1 Modeling approaches

- *SARIMA* models predict values at a time using data values and errors at times with lags with multiple of time period S. To determine model performance, we compared the AIC, log-likelihood, and validation set RMSE for each model. Nested models with similar log-likelihood values were compared statistically using the log-likelihood ratio test. We implemented an exhaustive search, limiting the number of parameters in order to prevent overfitting during model selection process. The optimal model we selected is SARIMA(2,1,0)(3,1,3)[12] because it has lowest RMSE, performs equally well as more complex models, and satisfies all modeling assumptions.

- *SARIMAX* models are multivariate time series processes that incorporate external variables that influence response variable into the SARIMA model. From Figure 2i.1, we can see housing price index, population have similar trend with bankruptcy rate. To find our optimal model, we built on our optimal SARIMA model and iterated through all combinations of covariates. We found the optimal model to be SARIMAX(2,1,0)(3,1,3)[12] with unemployment rate, which had the lowest predictive RMSE and satisfied all modeling assumptions.

- *Vector Autoregression (VAR)* is a good choice if the unemployment rate, population, house price index, and bankruptcy rate all influence each other. The VAR model uses observations of all variables within a defined maximum lag to predict bankruptcy rate. In practice, in order to reduce model complexity, we usually only use lag values less than four. Thus, we fitted three VAR models with lags of 1, 2, and 3. After performing likelihood ratio test, we came to the conclusion that VAR(3) fits the data the best.

- The *Holt-Winters* method is useful to forecast a time series when we have no distributional assumptions. Since both trend and seasonality are present in the bankruptcy rate time series, we used a triple exponential smoothing approach with multiplicative variation. On performing a grid search over the level, trend, and seasonal smoothing parameters, we found that the optimal model with the lowest predictive RMSE has $\alpha = 0.62$, $\beta = 0.86$ and $\gamma = 0.82$.

In addition to the above models, we attempted ensembling similar models together, but no additional improvement in the validation RMSE was found. It is also worth noting that adding all three covariates fits the training data well, but does not produce a good prediction on held-out validation data, indicating overfitting on the training set.

3.2 Model selection

After selecting the optimal model from each modeling approach, we compared the predicted RMSE on validation data. SARIMA and SARIMAX model have the two lowest predictive RMSE values (0.0038 vs. 0.0029 respectively as shown in Table 3.1). We further compared these two models using likelihood ratio test, which suggested that the SARIMAX model fits the data better than the SARIMA model with p-value 0.06. Finally, we examined the residual assumptions to ensure our selected model is valid, results show our model satisfied all the modeling assumptions (residual diagnostic tests are in Appendix 6.1). Therefore, our final optimal forecasting model is SARIMAX(2,1,0)(3,1,3)[12] with unemployment rate as exogenous variable.

Table 3.1 Model Comparison

Model	RMSE
SARIMA(2,1,0)(3,1,3)[12]	0.0038
SARIMAX(2,1,0)(3,1,3)[12] Unemployment Rate	0.0029
VAR(3)	0.0057
Holt-Winters	0.0039

4. Forecasting

The following are our predictions for 2011 and 2012 bankruptcy rates along with 95% prediction interval using the best model are displayed in Figure 4.1. The full prediction result table can be found in Appendix 6.2.

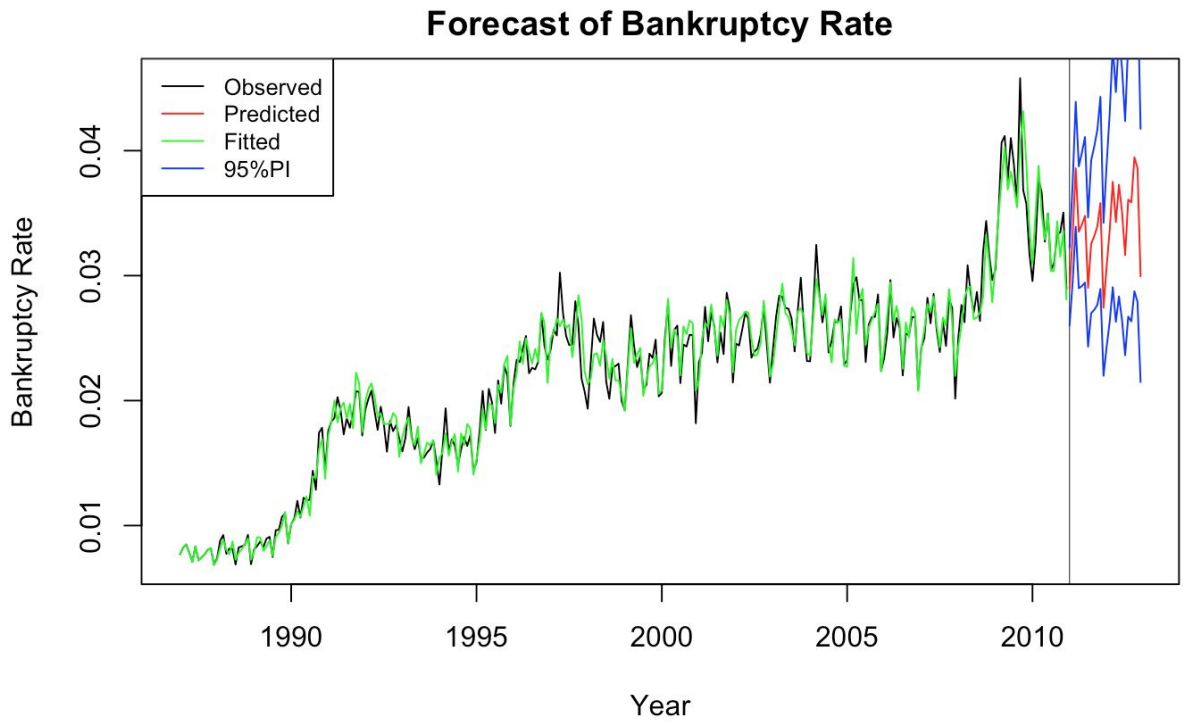


Figure 4.1 Forecast of Bankruptcy Rate for 2011 and 2012

5. Conclusion

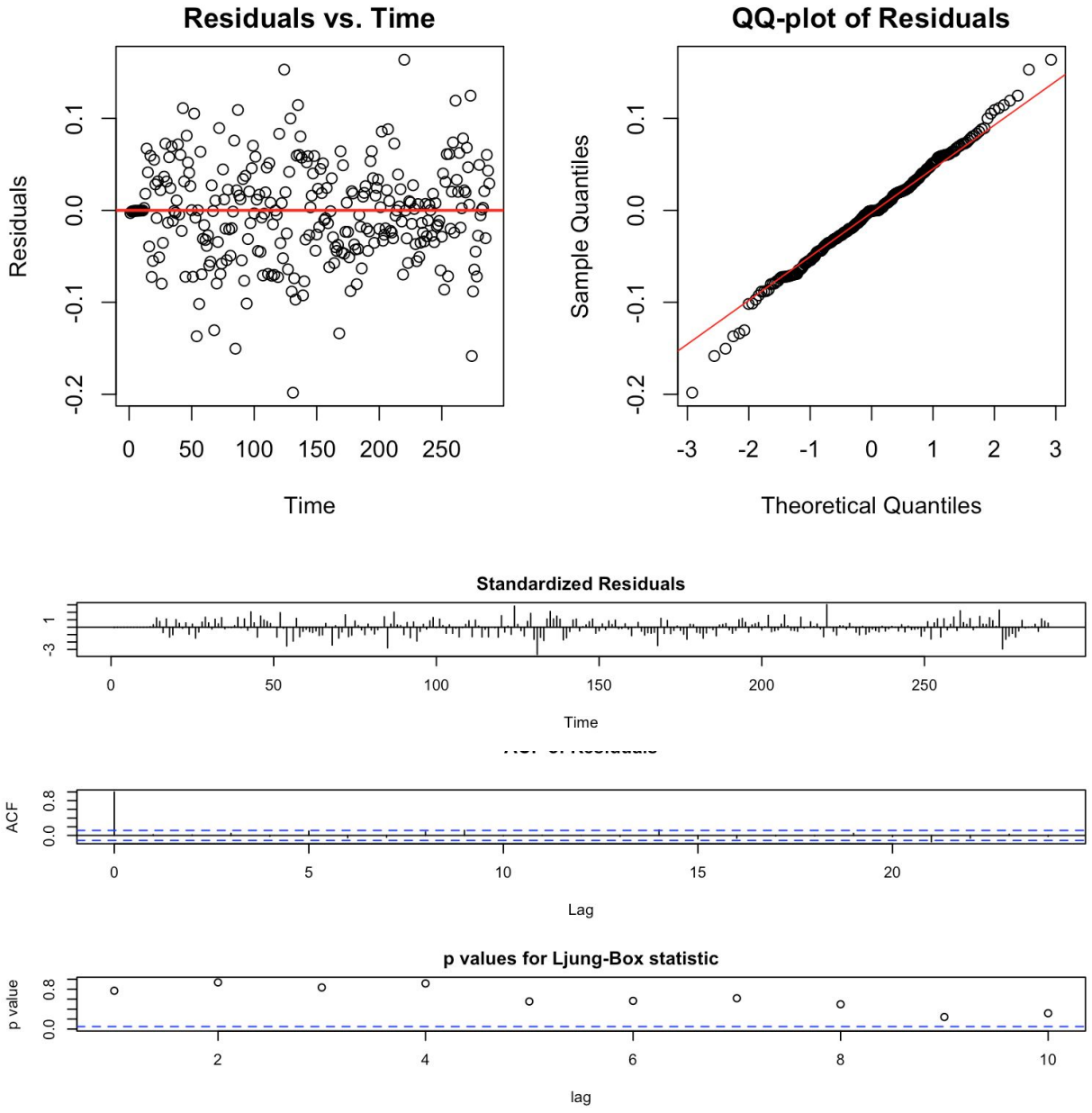
After comparing the four different models, we found our optimal model to be the SARIMAX(2,1,0)(3,1,3)[12] model with unemployment rate as the exogenous variable, based on this model having the smallest predicted RMSE on the held-out validation dataset. In addition, figure 2.1 shows unemployment rate aligns the best with bankruptcy rate spikes appears after 2008, that indicates unemployment rate is an appropriate exogenous variable to be included in the forecasting model.

There are some limitations to our current model. Our current model has been trained on the unemployment rate time series data observed from January 1987 through December 2010. If there were to be any unexpected changes in the observed unemployment rate in future years that are not accounted by the previous data, our model would not be able to accurately forecast the bankruptcy rates. An alternative method would be to forecast the unemployment rates as well, but then the error in our forecasted employment rates would also make our predictions of bankruptcy rate worse.

In the future, we could explore more advanced time series models, as well as a machine learning approach, such as decision trees or deep learning, as these have been shown to perform just as well as the models explored in this report.

6. Appendix

6.1 Residual Diagnosis - SARIMAX(2, 1, 0)(3, 1, 3)[12] with unemployment rate



- Residual Zero-mean: Residuals against time plot shows residuals are randomly scattered around horizontal line y=0, which informally confirms zero-mean. The one sample t-test fails to reject the

null hypothesis at 95% confidence level (P-value 0.42), suggesting that the residuals have zero-mean.

- Residual Homoscedasticity: Residuals against time plot shows residuals are randomly scattered around horizontal line y=0 except for a few points, which correspond to the unpredictable high spikes in the raw data. Overall the plot informally shows constant variance. The Bartlett-test using three groups suggested that the homoscedasticity assumption is satisfied (P-value 0.45).
- Residual zero-correlation: ACF plot shows no significant correlation between residuals.The Lyung-Box test suggested that the residuals are uncorrelated (P-value 0.57) .
- Residual normality: QQ-plot shows residuals are aligned with normal quantile well except for a few points which correspond to the unpredictable high spikes in the raw data. Overall the plot informally shows Residual are normal distribution. Shapiro-Wilk suggests that the residuals are distributed normally (P-value 0.17).

6.2 Full Forecasts

Month	Prediction	Lower Bound(95%)	Upper Bound(95%)
2011-01	0.0290	0.0260	0.0323
2011-02	0.0332	0.0295	0.0373
2011-03	0.0386	0.0339	0.0439
2011-04	0.0335	0.0290	0.0388
2011-05	0.0341	0.0292	0.0398
2011-06	0.0348	0.0294	0.0411
2011-07	0.0290	0.0243	0.0347
2011-08	0.0325	0.0270	0.0392
2011-09	0.0332	0.0273	0.0403
2011-10	0.0339	0.0277	0.0416
2011-11	0.0358	0.0289	0.0443
2011-12	0.0274	0.0220	0.0342
2012-01	0.0307	0.0243	0.0388
2012-02	0.0336	0.0263	0.0430
2012-03	0.0375	0.0291	0.0484
2012-04	0.0343	0.0263	0.0447
2012-05	0.0373	0.0283	0.0490
2012-06	0.0351	0.0264	0.0466
2012-07	0.0316	0.0236	0.0424
2012-08	0.0361	0.0267	0.0487
2012-09	0.0359	0.0263	0.0488
2012-10	0.0395	0.0287	0.0542
2012-11	0.0386	0.0279	0.0534
2012-12	0.0299	0.0215	0.0417