National Bankruptcy Rates Forecasting

Chaman Preet Kaur, Alex Romriell, and Binjie Lai

May 3, 2016

1 Executive Summary

In this report we predict the bankruptcy rates in Canada for the year 2011. We have historical bankruptcy rates along with unemployment rates, population, and housing price index from the years 1987 to 2010 available for building a predictive model. We use several different time series forecasting models to do this, such as: $SARIMA$, $SARIMA\ with\ covariates$, $GARCH$, and $Holt-Winters$. This report, however, lists results obtained only from the best $SARIMA$, $SARIMA\ with\ covariates$ and $Holt-Winters$ models. Rolling window and moving window prediction methods are used to aid in forecasting. We choose the best model based on predetermined criteria. With Unemployment Rate and House Price Index as covariates, we find the best predictive model to be $SARIMA(4,1,2)(1,0,7)_1$. This report concludes that the lowest bankruptcy rates are expected to be seen during the months of January and June 2011, while the highest bankruptcy rates are expected to be seen in the months of March, April and October 2011.

2 Problem Description

Accurately forecasting national bankruptcy rates is of interest to national banks, insurance companies, credit-lenders, politicians, and so forth. The goal of this project is to precisely and accurately forecast monthly bankruptcy rates for Canada in 2011. The training set contains monthly data from January 1987 to December 2010 on the following variables:

- Bankruptcy Rate: target variable for forecasting

- Unemployment Rate: a covariate feature
- Population: a covariate feature
- Housing Price Index: a covariate feature

Table 1 contains a summary of this training set. There are 24 years (288 months) of data for each of the variables in the training set. Based on the training data, we will construct a time series model, which will then be used to precisely and accurately forecast the January 2011 - December 2011 bankruptcy rates in Canada.

Table 1: Summary Table for Training set, 1987 - 2010

Statistic	N	Mean	St. Dev.	Min	Max
Unemployment Rate	288	8.236	1.528	5.900	12.100
Population	288	30,256,218	2,199,282	26,232,423	34,272,214
Bankruptcy Rate	288	0.022	0.008	0.007	0.046
House Price Index	288	75.218	14.124	52.200	104.000

Figure 1, below, shows the correlation between the four variables. According to the figure, bankruptcy rate is highly correlated with population and is somewhat correlated with house price index. Bankruptcy rate has a smaller, and negative, correlation with unemployment rate. Variables with medium to high correlation are of interest because they can possibly be used as covariates to help accurately predict bankruptcy rates.

3 Methods

In order to find the best model to predict 2011 Canadian bankruptcy rates, we split our training data into two sub-sets. One for training models and another for testing those models. We split the data by the first 20 years (240 observations) for training and the remaining 2 years (48 observations) for testing. This allowed us to assess the limitations of the model as well as make

Figure 1: Correlation Between Variables for Training Set

comparisons between models based on how well they were able to predict on the test set.

Through iteration, the parameters for the below model types were optimized based on the training data.

- SARIMA Model
- SARIMA + Covariates Model
- Holt-Winters Model

A $GARCH$ model was not considered for forecasting. $GARCH$ models are particularly good at forecasting time series data with periods of volatility and tranquility. However, a preliminary investigation of the data indicated that there were no such periods in the data.

After finding the optimal parameters for each category of model, the test set was used to determine predictive accuracy. We used AIC, sigma squared, llk (log likelihood), and RMSE (root mean square error) as our comparison criteria. The best model among the three was chosen to forecast the bankruptcy rate from January 2011 - December 2011.

3.1 Model Selection

Each team member iterated through several different models to find the best parameter for $SARIMA$, $SARIMA\ with\ covariates$, and $Holt-Winters$ models. After comparing the optimal parameters and the fit of each model, as listed in Table 2, we chose $SARIMA(4,1,2),(1,0,7)_1 + Covariates$ as the optimal model. In this table, low values of AIC, low values of $sigma^2$, high values of llh (log-likelihood), and low values of RMSE, are preferred. AIC, $sigma^2$, and llh are not available for the Holt-Winters approach because it is only a smoothing technique for forecasting.

Table 2: Models-Comparison

ID	Model	AIC	$Sigma^2$	llh	RMSE
1	SARIMA	-622.784	0.00336	324.392	0.00568
2	SARIMA + Covariates	-535.032	0.00502	285.516	0.00478
3	Holt-Winters	—	—	—	0.00221

From Table 2, we see that the $SARIMA + Covariates$ model performed the best in terms of the AIC while $SARIMA$ performed well in terms of $sigma^2$ and log-likelihood. The Holt-Winters approach had the lowest RMSE. In order to ensure a fair selection of the optimal model, a rolling window and moving window approach was used to fit the $SARIMA$, $SARIMA + Covariates$ models as well as the Holt-Winters model. A moving window prediction was found to provide more accurate results. The graphs shown in Figure 2, Figure 3, and Figure 4 depict the fit obtained for each of these approaches.

It can be seen from Figures 2 - 4 that the $Moving\ Window\ SARIMA + Covariate$ model, Figure 3, fits the training data the best. It is able to adjust the most for the spike in bankruptcy rates seen in 2009. The RMSPE (root mean squared predicted error) values obtained for each model is presented in Table 3. Smaller numbers in RMSPE are desired. This represents the difference between the actual value and predicted value for the held-out test set. The results in Table 3 confirm what is seen visually.

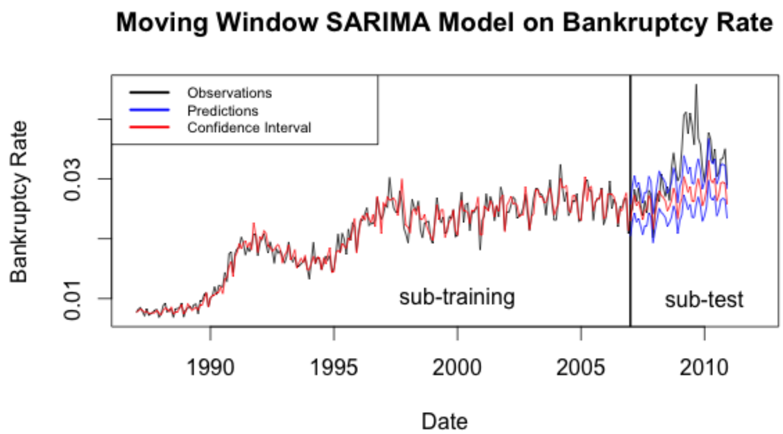

Figure 2: SARIMA

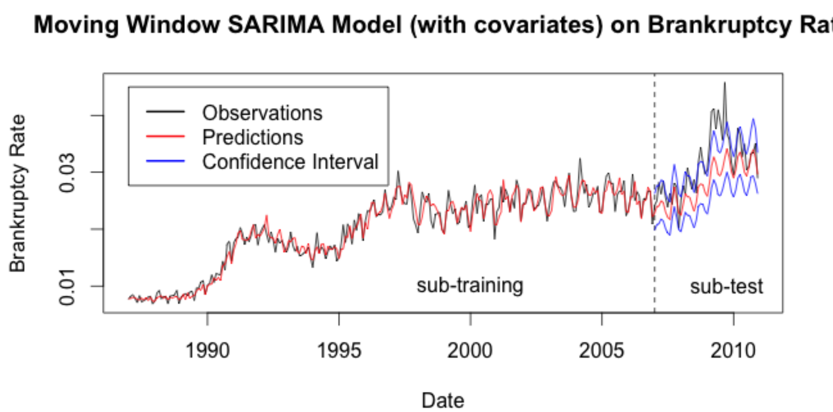

Figure 3: Moving window SARIMA with covariate information

The covariates used to produce this model were Unemployment Rate and House Price Index. As an aside, during model selection and calculation of RMSPE, there was one model in the $SARIMA + Covariates$ category that generated an RMSPE value lower than the best model reported here. This model was the $SARIMA(4,1,2),(1,0,7)_1$ with *rolling window* predictions and with *all* the covariates. This model gave an RMSPE value of 0.00448. However, the $SARIMA(4,1,2),(1,0,7)_1$ with *moving window* predictions, and only Unemployment Rate and House Price Index as covariates, gave a better fit in terms of the finer (tiny) ups and downs in the data. The RMSPE value of this model was comparable at 0.00482. Because this difference is quite small we opted for the model that better adjusts with the monthly nuances. Our final and

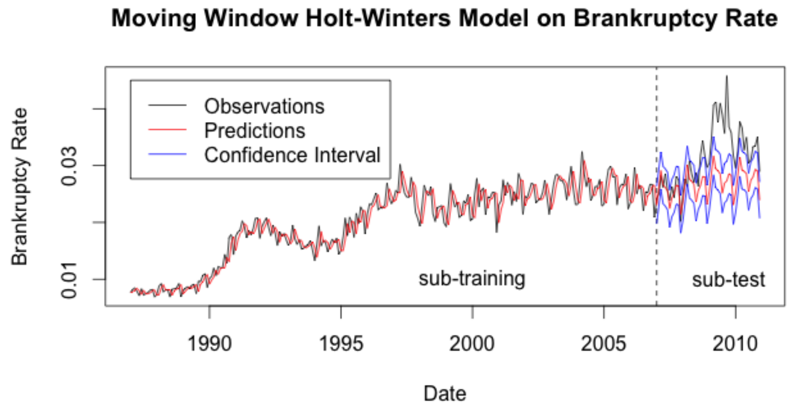

Figure 4: Moving Window Holt-Winter Model on Bankruptcy Rate

optimal model of the $SARIMA+Covariates$ class is the $SARIMA(4,1,2)(1,0,7)_1$ model. These numerical values represent the best parameters for this model. This model, when accompanied with the covariates: Unemployment Rate and House Price Index, provided the best prediction on our test set, and will thus provide the best predictions for 2011 Bankruptcy Rates in Canada.

Table 3: RMSPE-Comparison

ID	Model	Method	RMSPE
1	SARIMA	Moving window	0.00671
2	SARIMA + Covariates	Moving Window	0.00482
3	Holt-Winters	Moving Window	0.00621

3.2 Model Diagnostics and Limitation

Using residual diagnostics, we checked the assumptions of the residuals for our final model, $SARIMA(4,1,2)(1,0,7)_1$. See Figure 5 and Figure 6 for reference. The residuals passed all the tests, including normality test, randomness test, and homoscedasticity test. The model we chose satisfied all the assumptions, which means that we can confidently make predictions with this model and provide corresponding confidence intervals for those predictions.

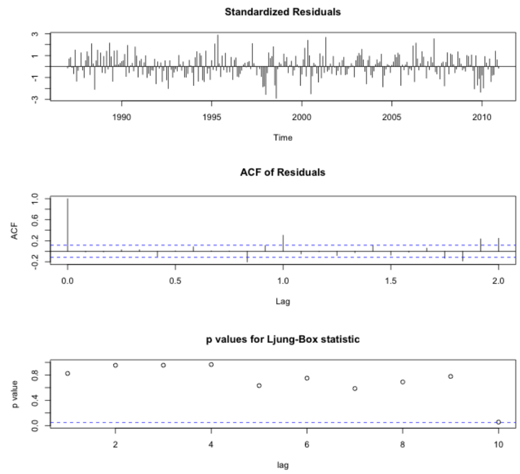

Figure 5: Residuals, ACF, and Ljung-Box Plots

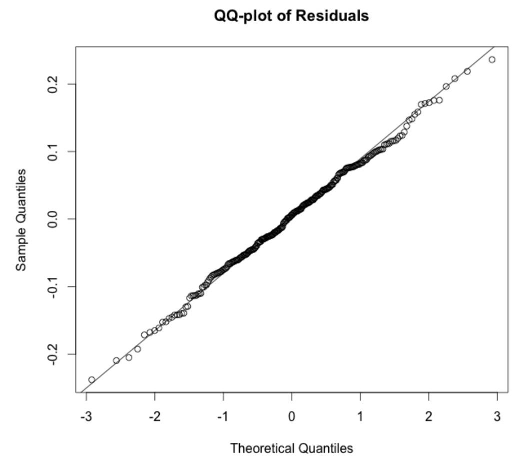

Figure 6: QQplot for Residuals

This model does, however, have a few limitations. While compiling the final model, a some warnings pertaining to possible non-convergence at certain points were given. However, the predicted values from this model gave the best root mean square predicted error. We suspect that if there are any more significant spikes in bankruptcy rates in the future, we will have to re-tune the parameters for this model.

4 Forecasting

In order to make forecasts into 2011, we expanded the window width to the length of the whole training set and implemented the final $SARIMA + Covariates$ model with the 2011 covariate information to obtain the monthly bankruptcy rates from January 2011 to December 2011. Table 4 displays the predictions with its corresponding upper and lower confidence interval for each month. Figure 7 is a visual representation of this forecast.

5 Conclusion

We conclude that according to our prediction model, the months of January and June are expected to have the lowest bankruptcy rates in 2011, whereas the months of March, April, and October are expected to have the highest bankruptcy rates. In our opinion, since January is the

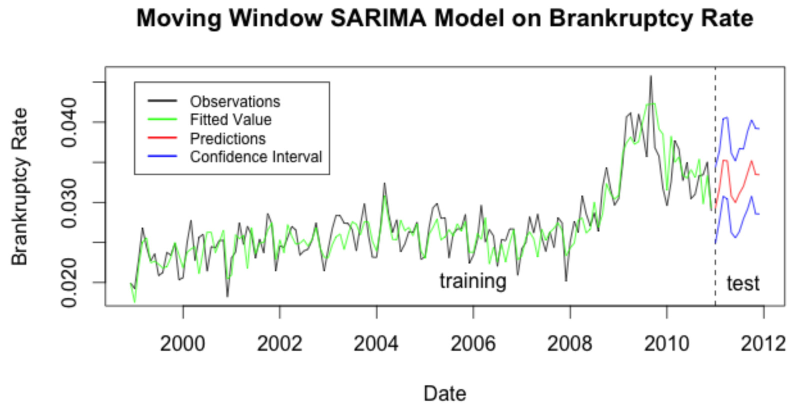

Figure 7: Moving Window SARIMA Model (with covariates) on Bankruptcy Rate

Table 4: Forecating Result

df	Forecast	Upper Interval	Lower Interval	Date
1	0.02925690	0.03437360	0.02490184	01/01/2011
2	0.03181410	0.03646543	0.02775607	02/01/2011
3	0.03526236	0.04041522	0.03076649	03/01/2011
4	0.03515061	0.04062120	0.03041677	04/01/2011
5	0.03080364	0.03614961	0.02624825	05/01/2011
6	0.02997545	0.03515452	0.02555937	06/01/2011
7	0.03110887	0.03671838	0.02635632	07/01/2011
8	0.03201328	0.03666493	0.02795178	08/01/2011
9	0.03366551	0.03886439	0.02916209	09/01/2011
10	0.03521253	0.04030760	0.03076150	10/01/2011
11	0.03348762	0.03923511	0.02858208	11/01/2011
12	0.03345200	0.03919103	0.02855337	12/01/2011

end of a fiscal year and June is mid year review time, most people get salary hikes during this time and hence this could be a contributing factor towards lower bankruptcy rates during these months.