Canadian Bankruptcy Rates Forecasting

Mengxin Qian, Vyakhya Sachdeva, Anshika Srivastava

1. Introduction

This report aims to document our efforts towards creating Canadian bankruptcy rate forecast models using time series analysis. The models are composed and validated for the period from January 1987 to December 2010. Composed models are then used to forecast bankruptcy rate in the years of 2011-2012. We also incorporated other macroeconomic trends, which can be good predictors of bankruptcy rate in our model including Unemployment Rate, Population, Housing Price Index.

2. Data

2.1 Data Summaries & Definitions

The dataset has information at a monthly granularity level starting January 1987 until December 2010, which is a sufficiently large range in time for carrying out a time series analysis. Henceforth in this report we will call this the training dataset, as our model will be "trained" using these observations.

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

Table 1: Summary table for the training dataset (January 1987 - December 2010)

We have 288 months i.e. 24 years of data for each of the 4 variables shown above. A description of the variables is as below -

1. **Bankruptcy Rate** - Response variable to be forecasted. National bankruptcy rate of Canada.
2. **Unemployment Rate** - Available covariate. Unemployment Rate of Canada.
3. **Population** - Available covariate. Number of inhabitants in Canada.
4. **House Price Index** - Available covariate. A metric that measures changes in single-family home prices across a designated market in Canada.

A quick look at the plots of each of the variables over time gives us an idea about the trends and/or seasonality in each of them [Figures 1 & 2]. Also the trend-line for House Price Index seems to follow-in very closely with the Bankruptcy Rate [Figure 3].

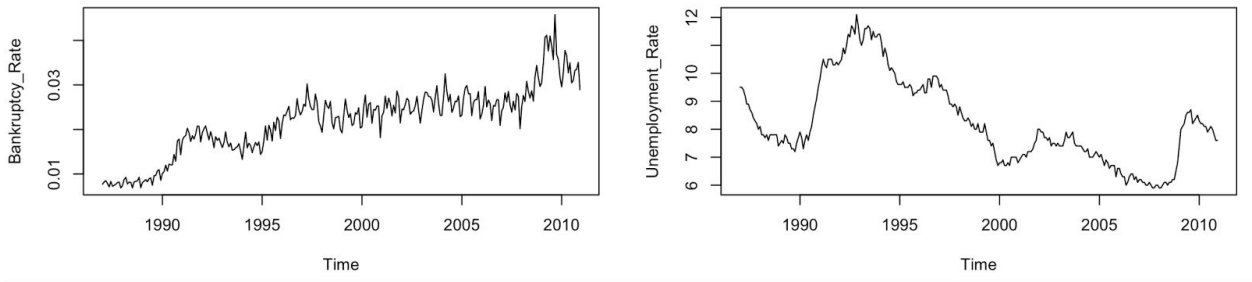


Figure 1: Bankruptcy Rate & Unemployment Rate trends over time (1987 - 2010)

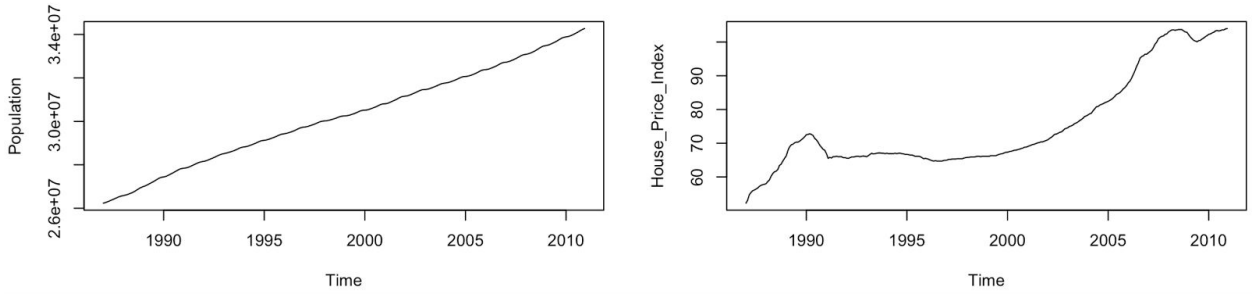


Figure 2: Population & House Price Index trends over time (1987 - 2010)

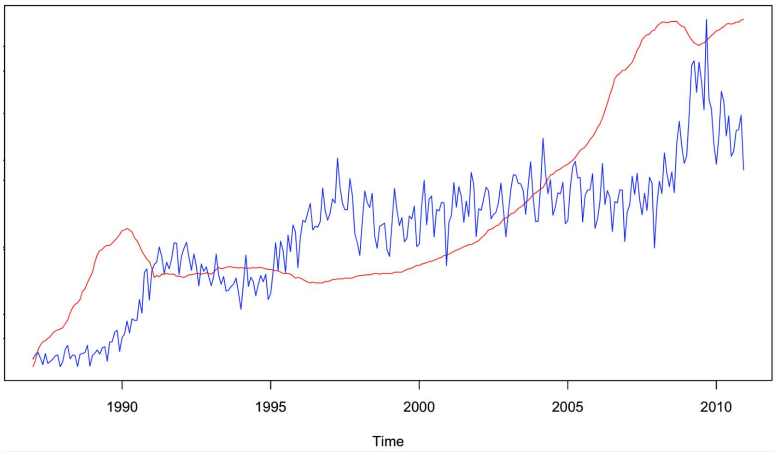


Figure 3: Overlapping trends of Bankruptcy & House Price Index

2.2 Relationships in the Data

Understanding the relationships between the target variable and available covariates is an important part of model building. Both Population and House Price Index show a strong correlation with Bankruptcy rate [Figure 4].

Figure 4: Relationship between the given macroeconomic variables

3. Methodology

In order to find the best model for forecasting the bankruptcy rates , we split the available training data in two parts. One is used to train the models in order to get the best parameter estimates and the other is used to test the accuracy of these models . We trained the model on the data of 19 years and used this model to estimate the accuracy on the data of subsequent 5 years. We assume if the model performs well on this held out sample of 5 years data, it would perform well in future forecasts too.

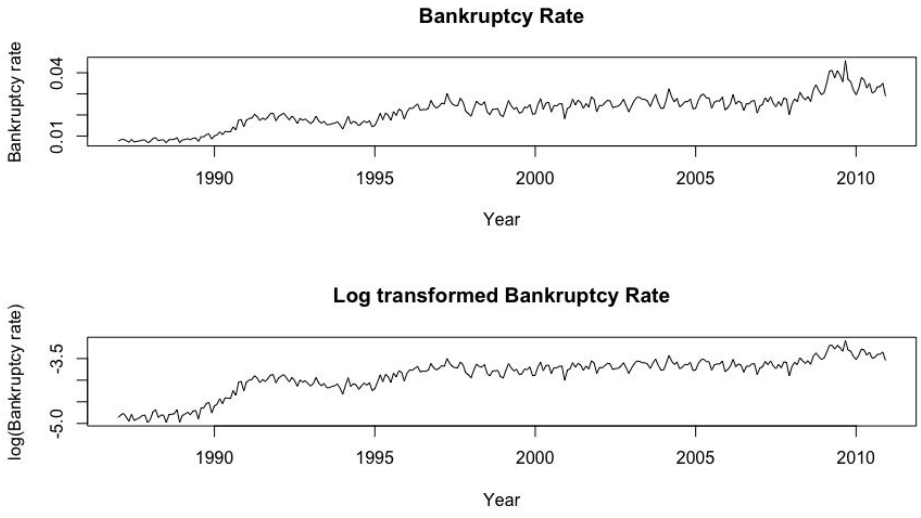


Figure 5: Bankruptcy Rate and Log transformed Bankruptcy Rate

Looking at the plot of Bankruptcy data against time, we can see observations are not random and there is some trend or pattern. Also, there are some periodic peaks which suggest seasonality effects. Since the Bankruptcy data has varying variance, that is, the variation increases with the time, it has been log transformed to get rid of varying variation [Figure 5].

3.1 Modeling Techniques

Since there is some seasonality and trend, we explored the following methods for modeling the bankruptcy data

- SARIMA
- SARIMA using covariates
- Exponential Smoothing
- VAR(Vector Autoregression)

The potential models are mainly compared on the basis of log likelihood, AIC, sigma squared. And a optimal model was selected for each of these approaches. Then, we selected the final optimal model from these four models considering other factors. We chose the model with the best predicting performance on held-out dataset as our final model. The residual diagnostics, which are the tests to ensure that the model meets the required modeling assumptions were also checked.

SARIMA works by removing the trend and seasonality through differencing the data and modeling the transformed data to estimate the parameters. The following model was chosen by iterating through several SARIMA models to come up with the most optimal model. The details of all such iterations is outside the scope of this report. Data is differenced once to remove the trend. Although the data exhibits seasonality, the auto correlation plot shows it does not specifically need any differencing to remove it. The transformed data series is trained on the training data and the best model was selected using the log likelihood, AIC, sigma squared metrics. SARIMA(2,1,1)(1,0,2)[12] is the optimal SARIMA model.

ARIMAX works by considering external variables in addition to SARIMA ,which influence the response series under observation. As we can be seen from the Figure 4, the population, house pricing index and bankruptcy are highly correlated with each other. Hence, we should only pick from one from population and Housing price index to avoid multicollinearity. Since Population just has a trend with no seasonality, it can not explain the seasonal effects in bankruptcy as can be well done by Housing Price Index.Therefore, we checked SARIMA model with unemployment and Housing Price Index as the covariates.The most optimal model was selected by iterating through the ARIMAX models using these covariates.ARIMAX(2,1,1)(1,0,2)[12] with House Price Index as covariate is the optimal ARIMAX model. The details of all such iterations is outside the scope of this report.

Exponential Smoothing works by assigning exponentially decreasing weights as the observations get older, that is, recent observations are given relatively higher weight in forecasting than the older

observations. Depending upon whether the data has trend or seasonality, it uses different smoothing techniques to account for them. We used Triple Exponential Smoothing as we see both trend and seasonal pattern. The optimal model was decided using the prediction accuracy on the held-out sample. Triple Exponential(alpha=0.038,beta=0.69,Gamma=1) is the best Exponential smoothing model.

VAR works by treating the other influential variables as endogenous variable, that is, they influence bankruptcy and bankruptcy influences them. VAR(8) with House Price Index and Population growth is the optimal VAR model as it had comparatively lower Predictive Error on the held-out dataset.

A GARCH model, which is generally used when data shows periods of tranquility and volatility, was also considered. However, investigation of the final model results indicated that there is not much of this kind of a variation, hence it is not used.

3.2 Model Selection

After selecting the optimal models using four time series modeling approaches respectively, we compared the Predictive Error on the held out sample of these four models. The results indicate that ARIMAX and Exponential Smoothing model have the lowest validation Predictive Error. We further compared these two models using other methods. Based on the Bankruptcy rate vs. Time plot in Figure 2, we identified that the time series exhibit some sort of variation that is not due to random chance, but rather is attributable to a certain increasing trend. This implies that smoothing techniques - Exponential Smoothing here won't give us an ideal result if there is shock values due to special events (e.g. final crisis).Besides, Figure 4 indicates that house price is a good exogenous variable for predicting. So we chose the optimal ARIMAX as our final model. The final model **ARIMAX(2,1,1)(1,0,2)[12] with House Price Index as the covariate**, meets the model assumptions, the details of which are out of the scope of the report *(Refer to Appendix 6.1 for the details)*

4. Forecasting

To build our forecasting model, we used a 'rolling window' approach along with the final model selected in Section 3.2. This ensures that for each new prediction, the most up-to-date data is used.

The predictions for 2011 and 2012, along with the 95% confidence interval are displayed in Figure 6. A snapshot of the predictions for the first 6 months are included in Table 2, and the full table for 2011 and 2012 can be found in Appendix 6.2.

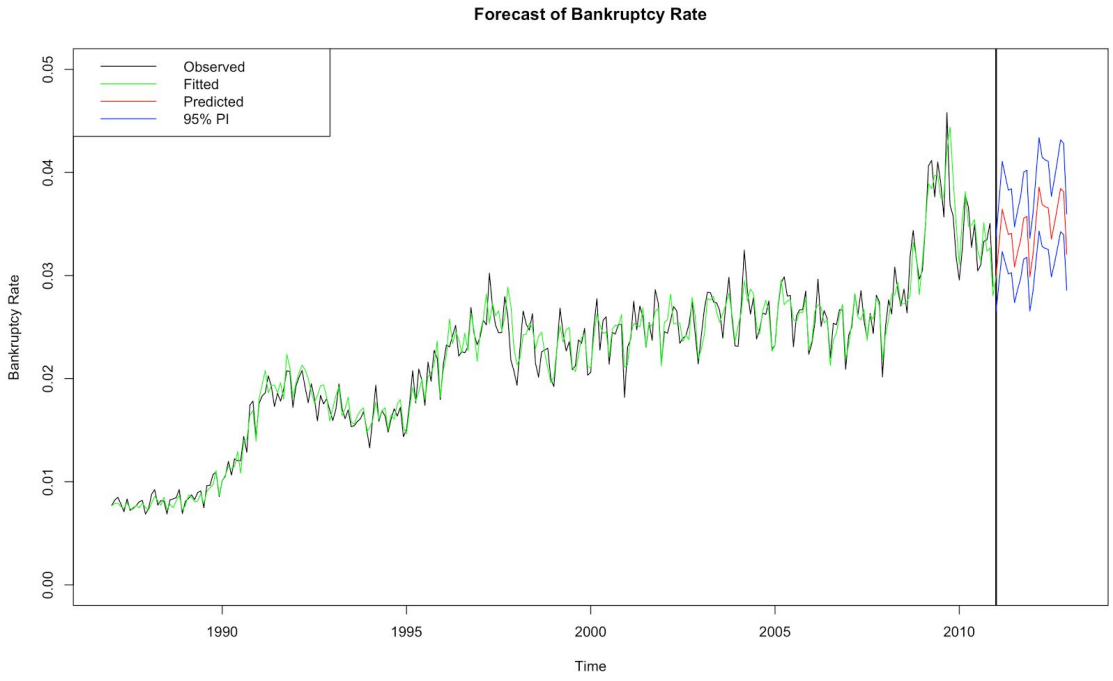


Figure 6: Forecast of Bankruptcy Rates for 2011 and 2012

	Prediction	Lower Bound (95%)	Upper Bound (95%)
January 2011	0.0299	0.0265	0.0337
February 2011	0.0331	0.0294	0.0374
March 2011	0.0364	0.0323	0.0410
April 2011	0.0351	0.0312	0.0396
May 2011	0.0339	0.0301	0.0382
June 2011	0.03409	0.0302	0.0384

Table 2: Forecast of Bankruptcy Rates for first 6 months of 2011

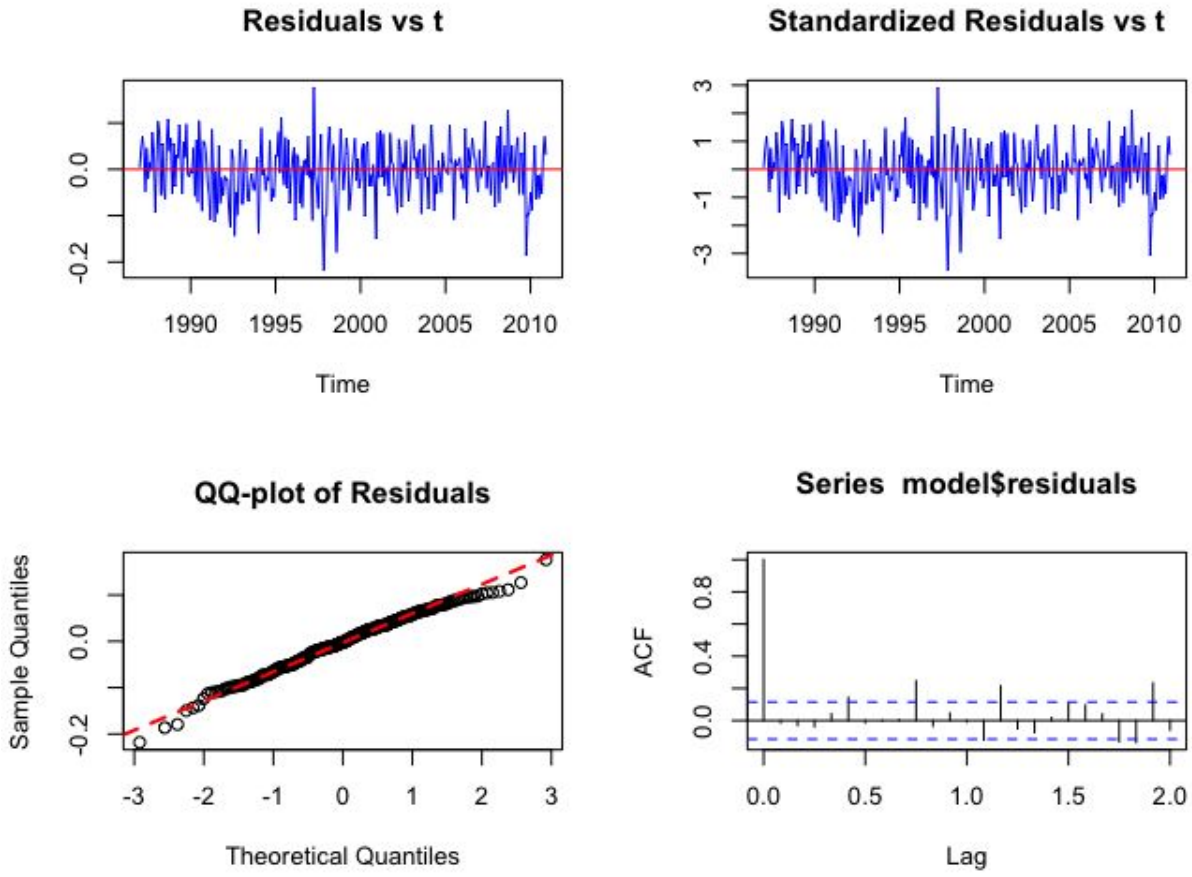
5. Conclusion

Our final model predicts bankruptcy rates using the past values of Bankruptcy Rate, and House Price Index as covariate. House price index has a positive influence on the bankruptcy rate, which means when house prices increase, bankruptcy rate will increase. It aligns with our intuition that an increase in house price index would lead to an increase in high real estate prices, which would in-turn contribute to higher bankruptcy rates.

Our model is simple, and ensures that the dependency of bankruptcy rates on its past values does not go too far back in time. However, we acknowledge that there are certain limitations of our model. It does not include other important macroeconomic variables that may improve the prediction. Also, we can consider using Ensemble Methods, which take weighted values of predictions from different models to give better accuracies.

6. Appendix

6.1 Residual Diagnosis -ARIMAX (2,1,1)(1,0,2)[12] with House Price Index



Informal Diagnostics:
Zero-Mean- The residual plot shows randomness around 0 hence the mean of residual appears to be 0.
Homoscedasticity- The residual plot shows constant variance except for some unusual high spikes. The number of such spikes seem less than the 5% of the data so might not be significant hence the residual can be said to be homoscedastic
Zero-Correlation- ACF plot shows no correlation between the residuals.
Normality-QQ norm plot appears to be Normal.

Formal Diagnostics:
Zero-Mean- One Sample t tests gives a p-value of 0.36 hence we fail to reject the null hypothesis that residuals have a mean zero. Hence the residuals have an expected value of 0.

Homoscedasticity- Barlett's test gives a p-value of 0.16 for for 4 group split and a p value of 0.23 for a 3 group split, hence we fail to reject the null hypothesis that residuals are homoscedastic.The assumption of Homoscedasticity is met.
Zero-Correlation- Run test gives a p value of 0.63 hence we fail to reject the null hypothesis that residuals are Uncorrelated.The assumption of Uncorrelatedness is met.
Normality-Shapiro Wilk test gives a p value of 0.06 clearly failing to reject the null hypothesis that residuals are normally distributed. The residuals meet the normality distribution assumption.

6.2 Full Forecasts

Month	Prediction	Lower Bound(95%)	Upper Bound(95%)
Jan 2011	0.0300	0.0266	0.0338
Feb 2011	0.0332	0.0294	0.0374
Mar 2011	0.0364	0.0323	0.0411
Apr 2011	0.0352	0.0312	0.0397
May 2011	0.0340	0.0302	0.0383
Jun 2011	0.0341	0.0303	0.0384
Jul 2011	0.0308	0.0274	0.0347
Aug 2011	0.0322	0.0286	0.0363
Sep 2011	0.0335	0.0297	0.0377
Oct 2011	0.0356	0.0316	0.0400
Nov 2011	0.0357	0.0318	0.0402
Dec 2011	0.0299	0.0266	0.0336
Jan 2012	0.0320	0.0284	0.0360
Feb 2012	0.0353	0.0314	0.0397
Mar 2012	0.0386	0.0343	0.0434
Apr 2012	0.0369	0.0328	0.0415
May 2012	0.0367	0.0326	0.0412
Jun 2012	0.0366	0.0325	0.0411
Jul 2012	0.0335	0.0299	0.0377
Aug 2012	0.0350	0.0312	0.0394
Sep 2012	0.0367	0.0327	0.0412
Oct 2012	0.0384	0.0342	0.0432
Nov 2012	0.0381	0.0340	0.0428
Dec 2012	0.0321	0.0286	0.0360