Forecasting Canadian Bankruptcy Rates for 2011-2012

By: Vanessa Zheng, Nimesh Sinha, Davi Schumacher, Prince Grover

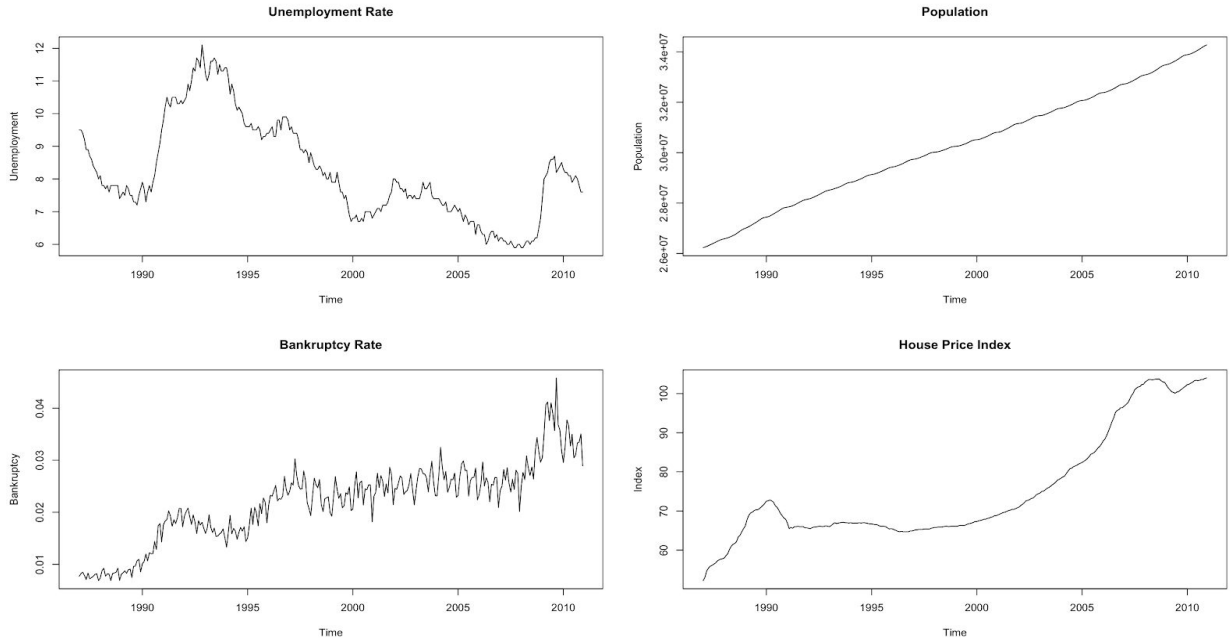
December 05, 2017

1. Introduction

The goal of this project was to build a model for forecasting Canadian bankruptcy rates for 2011 and 2012. This project used monthly data collected in Canada over the period from January 1987 to December 2010 to train time series models in preparation for forecasting. The information collected for training the model includes population, unemployment rate and housing price index, but only population and unemployment rate were used. In this report an overview of the data is presented, followed by a discussion of the modeling methods employed and their respective results. Univariate and multivariate modeling approaches are both presented along with the best models, where 'best' is defined as a model that minimizes the error between predictions and true values on our held-out validation set. Finally, our ensembled model, used for forecasting, is presented along with the actual forecast.

2. Data Overview

This project uses monthly data collected in Canada over the period from January 1987 to December 2010. The information includes measurements of population, unemployment rate, housing price index, and bankruptcy rate.



Graph 1 Time series data for forecasting bankruptcy rate in Canada (01/1987 - 12/2010)

i. Bankruptcy Rate

The bankruptcy rate generally increases from 1987 to 2010. There are a couple brief downward trends, most notably around 1992 - 1995, and there is an upward spike around 2009. The maximum bankruptcy rate never exceeds 5%. Variation in rates does increase over time, thus looking at the logarithm of the rate helps keep the variation constant.

ii. Population

The population of Canada increases linearly from around 26 million in 1987 to around 34 million in 2010. There is some slight annual seasonal trend as well. This data does not look similar enough to the bankruptcy data to help explain it, but including it did help improve some of our models.

iii. Unemployment Rate

Canada's unemployment rate has periods of ups and downs, but appears to have an overall downward trend in the period of interest, reaching a maximum of around 12% in the early 1990's and a minimum of around 6% near 2008. There were two significant spikes in unemployment, one in 1990 and the other in 2009. Bankruptcy trended upward during both of these periods as well, so even though the overall trend of the two time series are opposite, it is possible that the unemployment rate can help explain some parts of the bankruptcy rate. Specifically, large jumps in unemployment seem to be correlated with a mild increase in bankruptcy rate during the period of interest.

iv. Housing Price Index

Housing price index has similar overall trend to bankruptcy rates, generally increasing over the period of interest. Also, peaks in the index occur around 1990 and 2008, while peaks in bankruptcy occur around 1991 and 2009. The index appears to have experienced linear growth leading up to 1990, followed by no growth until 1995, and then exponential growth between 1995 and 2008. Housing price index made our models worse, and therefore was not used.

3. Methods and Results

In order to be confident in our results, the data was split into a training set and validation set. Since the data represents a time series, our validation set is taken to be the last two years of bankruptcy rates, depicted in the graphic below. Two years was chosen because we are forecasting a 2-year period and want to be confident our model can successfully predict that number of time steps. We then considered two modeling methods, Box-Jenkins and Holt-Winters.

Training data	Validation data	Test data
22 years *(1987 to 2008)*	**2 years** *(2008 to 2010)*	**2 years** *(2011 to 2012)*

i. Univariate Models

a. Box-Jenkins

The Box-Jenkins modeling approach involves modeling a time series as an autoregressive and/or moving average process, known as ARMA. An ARMA model forecasts the best fit for a time series based on that time series' past values, but is based on assumptions that our time series is stationary, meaning statistical properties like mean and correlation between observations do not depend on time. In cases where the raw time series is not stationary, we first transform it to make it stationary before modeling it. Seasonal and/or trend differencing or logarithmic transformations are some examples of common transformations. A trend and seasonally differenced time series modelled with ARMA is called a SARIMA model (Seasonal ARIMA) and is the type of model employed in this case study.

SARIMA:

The bankruptcy data is a non-stationary time series with both seasonal and trend components. Our approach is to finitely difference the time series to remove both trend and seasonality. In this documentation the number of times the data is differenced for trend and seasonality is represented by 'd' and 'D', respectively. After differencing, we were left with a stationary time series that was able to be modeled as an ARMA process. These processes are built atop statistical assumptions, therefore it is necessary to verify the assumptions have been met in order to confidently forecast.

In order to build our ARMA model, we found:

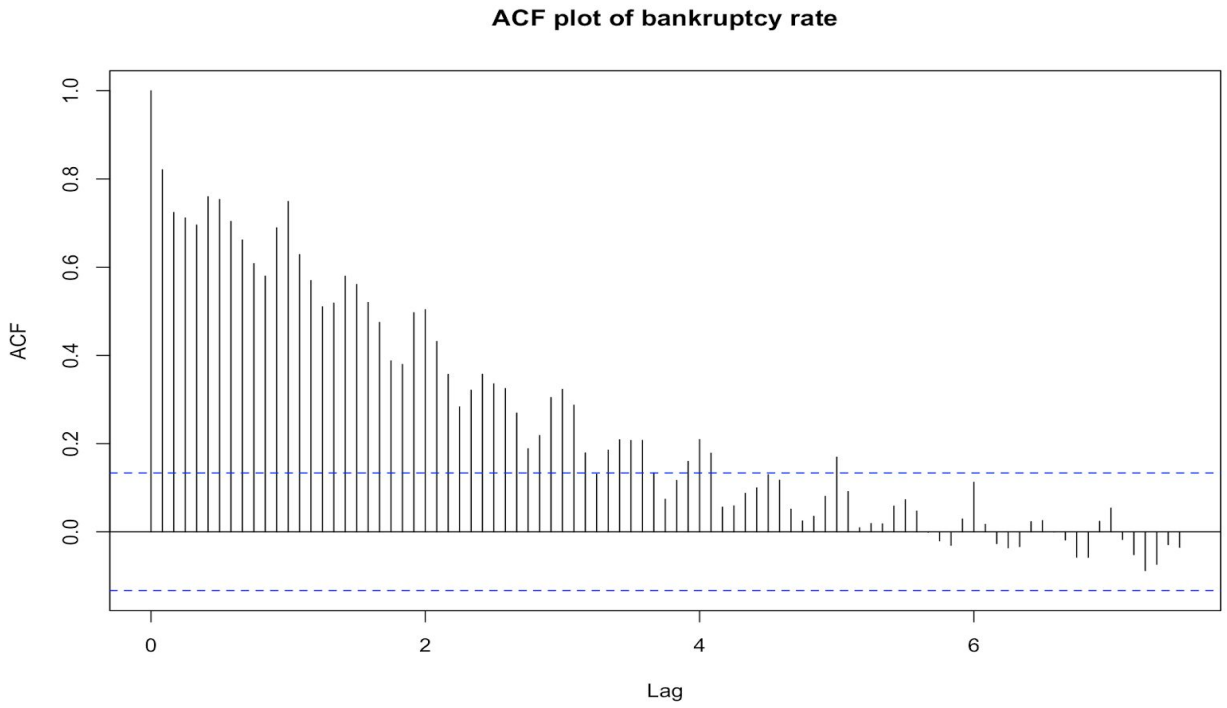
- p,q -- AR and MA orders of ***within season*** component
- P,Q -- AR and MA orders of ***between season*** components
- d,D -- Ordinary and Seasonal ***differences*** required to make time series stationary
- m -- The number of months constituting a season

Below are the modeling steps for SARIMA -

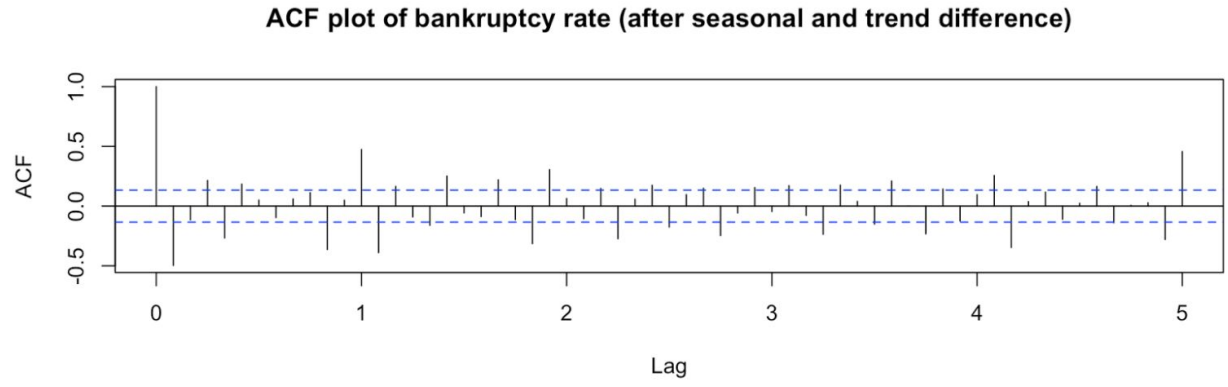
Step 1: By looking at bankruptcy rate, we decided whether or not we need a variance stabilizing transformation. We chose to apply a logarithmic transformation since variance appeared to increase with time.

Step 2: We trained our model using a subset of the data starting from year 1991 because it generated better results. This is likely because the trend of the data before and after 1991 are quite different.

Step 3: We found starting values of d and D by looking at autocorrelation (ACF) plots. Autocorrelation is the linear correlation of a signal with itself at two different points in time. The ACF of the bankruptcy data is plotted below. We chose d = 1 and D = 1.



Graph 2 ACF plot of log-transformed bankruptcy rate

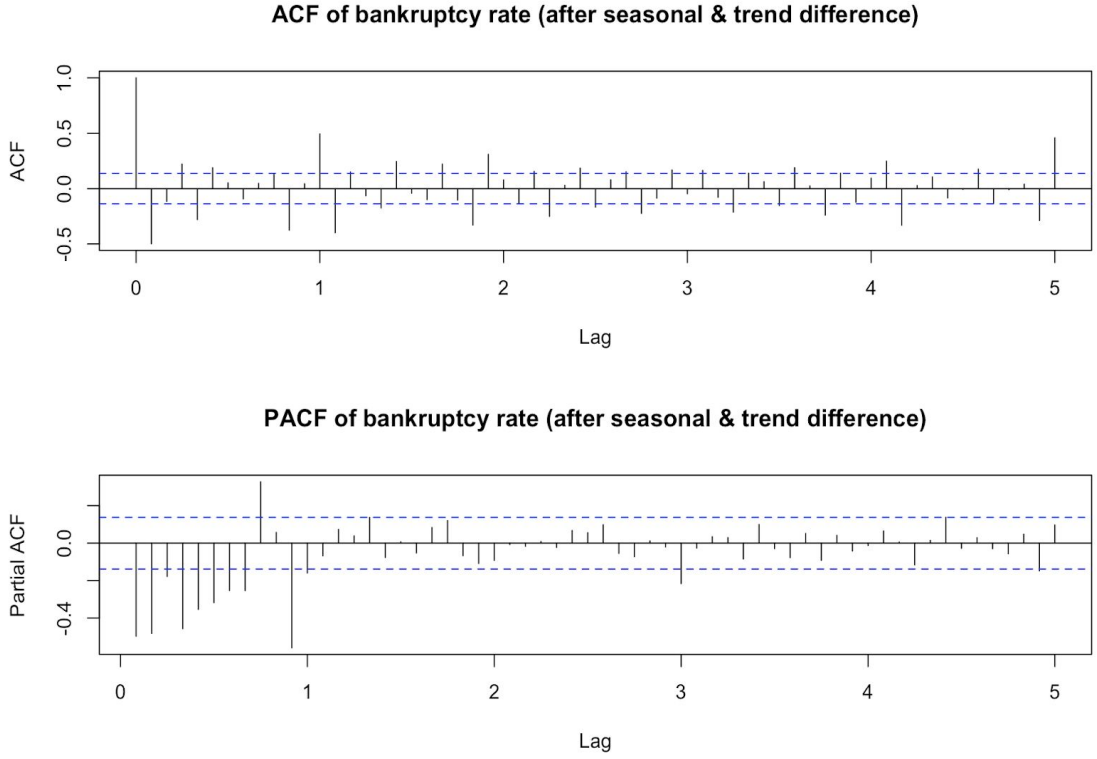


Graph 3 ACF plot of log-transformed bankruptcy rate after 2 differences

Stationarity of the time series after differencing was validated using the Augmented Dickey Fuller (ADF) unit root test. The P-value of the ADF test after differencing twice was less than our desired significance level, therefore we were confident that differencing had rendered the time series stationary.

[Null hypothesis of ADF unit root test is that our time series is non-stationary. P-value less than desired significance level makes us reject null hypothesis to argue that our time series is stationary]

Step 4: We examined ACF and partial ACF (PACF) plots of the differenced time series and found starting values of p, q, P, and Q.



Graph 4 ACF and PACF plots of log-transformed bankruptcy rate after 2 differences

Theoretically, the ARMA parameters can be deduced by observing the ACF and PACF plots, but the process is not precise. Therefore, we searched over a range of parameters to find the combination that gave the smallest prediction error, Akaike Information Criterion (AIC), and variance on our validation set.

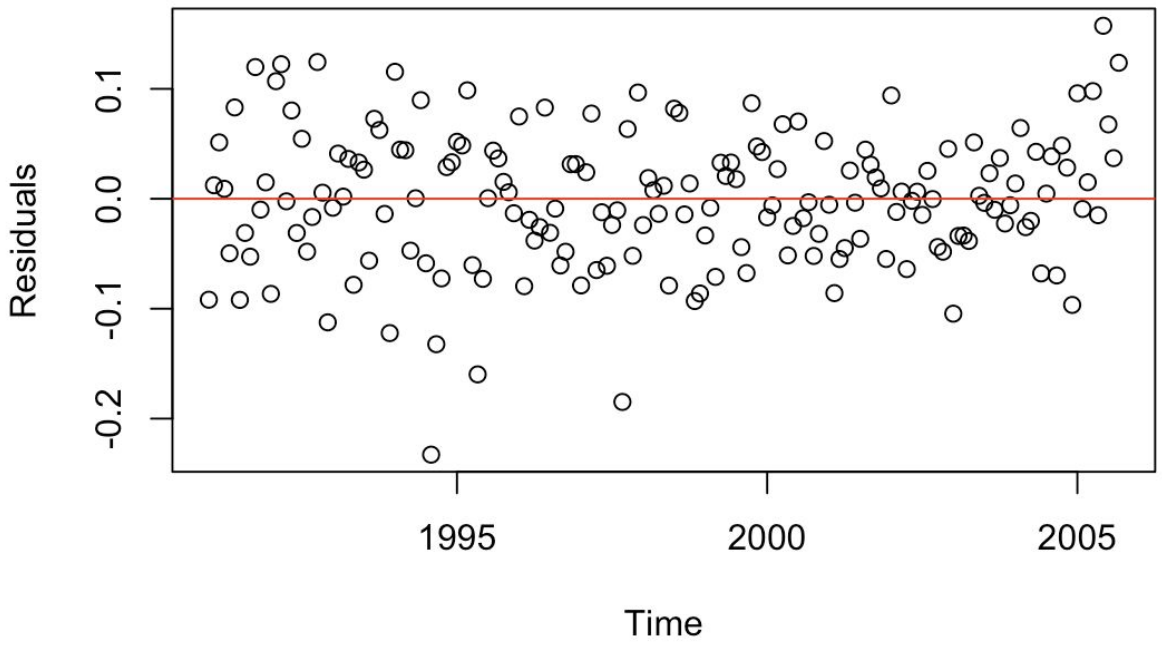
[We desire lower values of AIC, prediction error (variance) and RMSE in our model]

Table 1 Best parameters and corresponding scores from SARIMA

Order (p,d,q)(P,D,Q)(m = 12)	Method	Validation RMSE	AIC	Variance
(2,1,1)(2,1,2)	CSS	**0.00354**	278.3	0.003773
(2,1,1)(2,1,2)	CSS-ML	0.00364	279.9	**0.003237**
(1,1,0)(3,1,2)	CSS	**0.00343**	**276.7**	0.003831
(2,1,0)(2,1,2)	CSS	0.00354	277.7	0.003793

Step 5: Verify all residual assumptions are met. The residuals are the errors between the predicted and actual values of the validation set.

(A) Zero mean (t-test): p-value = 0.8565. Given this p-value is higher than our significance level, we are confident that the mean of the residuals is 0. This result is consistent with the plot of the residuals.



Graph 5 Residual vs. Time plot of SARIMA model

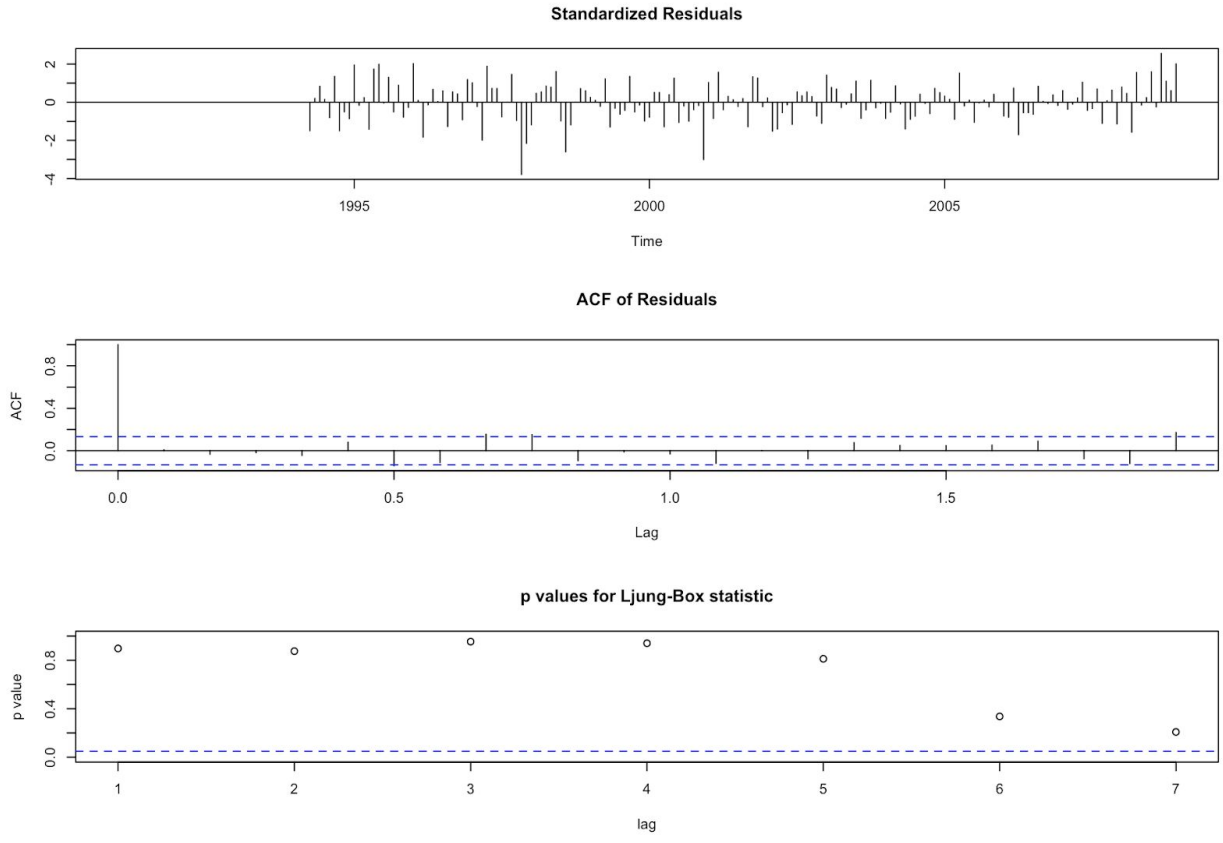
(B) Shapiro-Wilk Normality test: It tests the null hypothesis that the sample came from normally distributed population. Therefore, we desire to keep the null hypothesis.

P-value for selected model = 0.2013. Given this p-value is higher than our significance level, we are confident that the residuals are normally distributed.

(C) Bartlett test of constant variance: It is used to test if our samples come from populations with equal variances

P-value of this test on residuals = 0.03. We can accept constant variance at the 0.01 significance level. This result is consistent with the plot of the residuals.

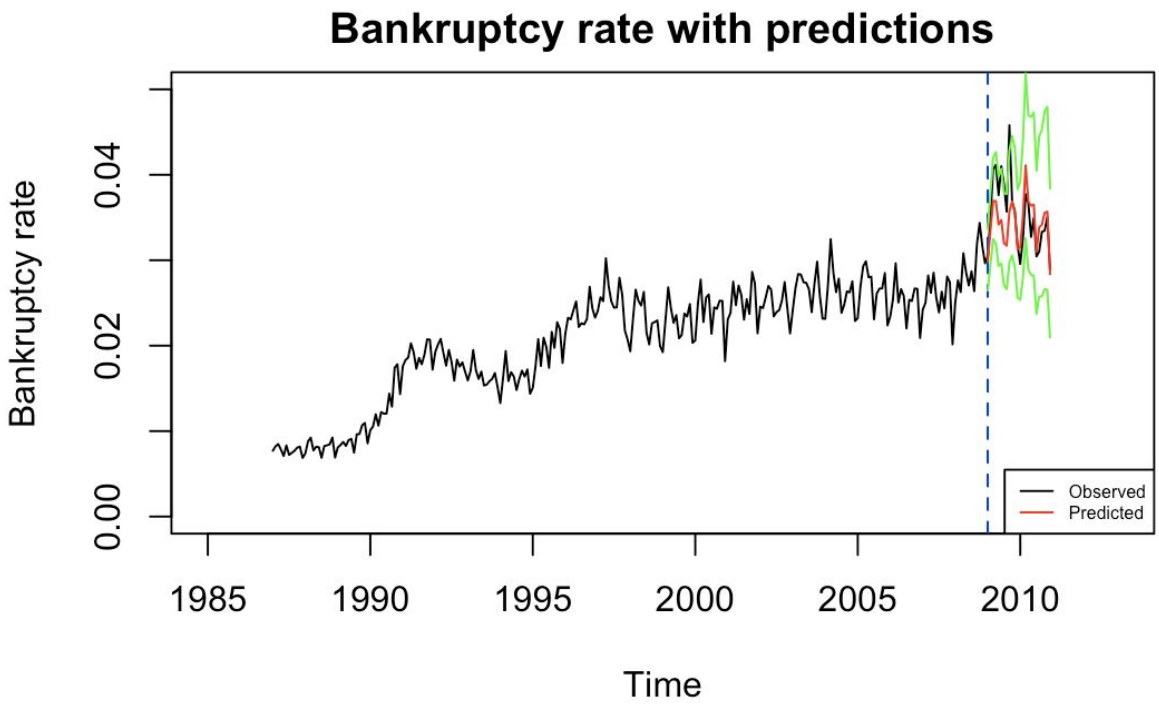
(D) Ljung-Box test for zero correlation: It is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. We desire our residuals to have zero autocorrelations for small lags.



Graph 6 Plots related to zero correlation assumption of SARIMA model

Residuals of the predicted time series appear to be uncorrelated for small lags.

Our final SARIMA model: *p = 2, q = 1, P = 2, Q = 2, d = 1, D = 1.*



Graph 7 Plots related to zero correlation assumption

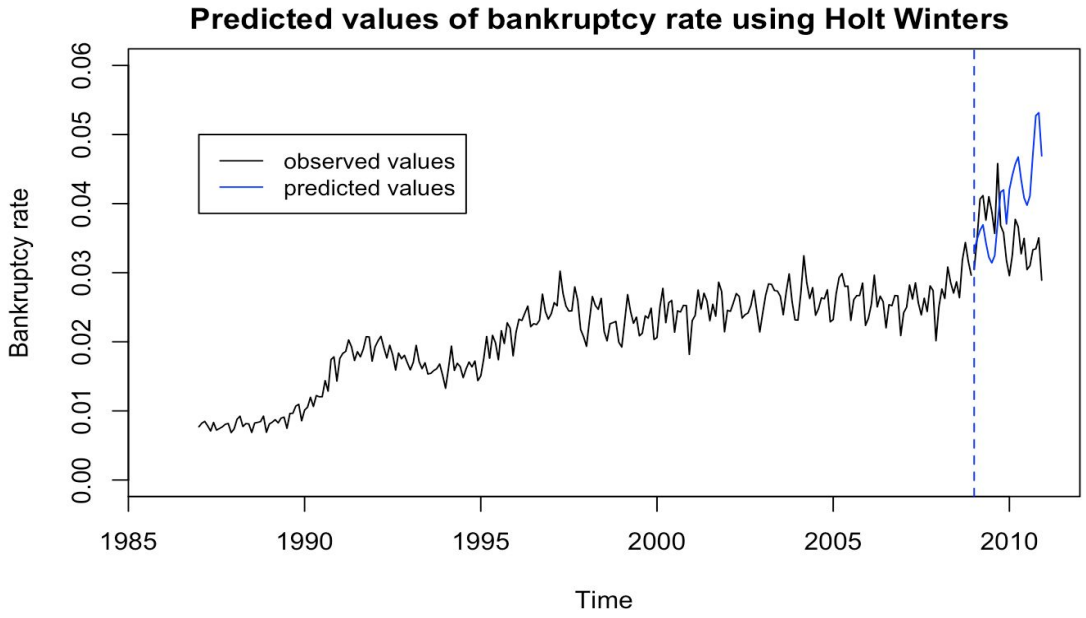
b. Holt-Winters

In exponential smoothing older data is progressively given less relative weight (importance) whereas newer data is given progressively more relative weight. There are three types of exponential smoothing:

1. Single exponential smoothing (When data has no trend and no seasonality)
2. Double exponential smoothing (When data has trend and no seasonality)
3. Triple exponential smoothing (When data has both trend and seasonality)

Since the time series of bankruptcy rate includes both trend and seasonality, we used triple exponential smoothing for modeling using Holt-Winters approach. We have used log transformed data due to its non constant variance.

The values of smoothing parameters alpha = 0.6, beta = 0.6 and gamma = 0.8 produce the lowest prediction error of 0.0049. Following is the plot for the predictions from the Holt Winters method.



Graph 8 Predicted values of bankruptcy rate using Holt Winters

We can observe that the predictions may only be based on the previous trend. In 2010, the observed values decrease, but the predictions still follow the increasing trend. The Holt winters approach has the limitation that we fail to include the effect of other variables on the bankruptcy rate.

ii. Multivariate Modeling

We also fit multivariate models to see if there were external variables that could help explain the changes in bankruptcy rates. These external variables can be either exogenic, in which case they only affect the bankruptcy rates, or endogenic, in which case they both affect and are affected by bankruptcy rates. A SARIMAX model is useful in the case of exogenic variables and vector autoregression (VAR) is useful in the case of endogenic variables.

a. SARIMAX

After trying different combinations of external variables, we decided to only use the logarithmically transformed unemployment rate since it appeared to be the only impactful external variable.

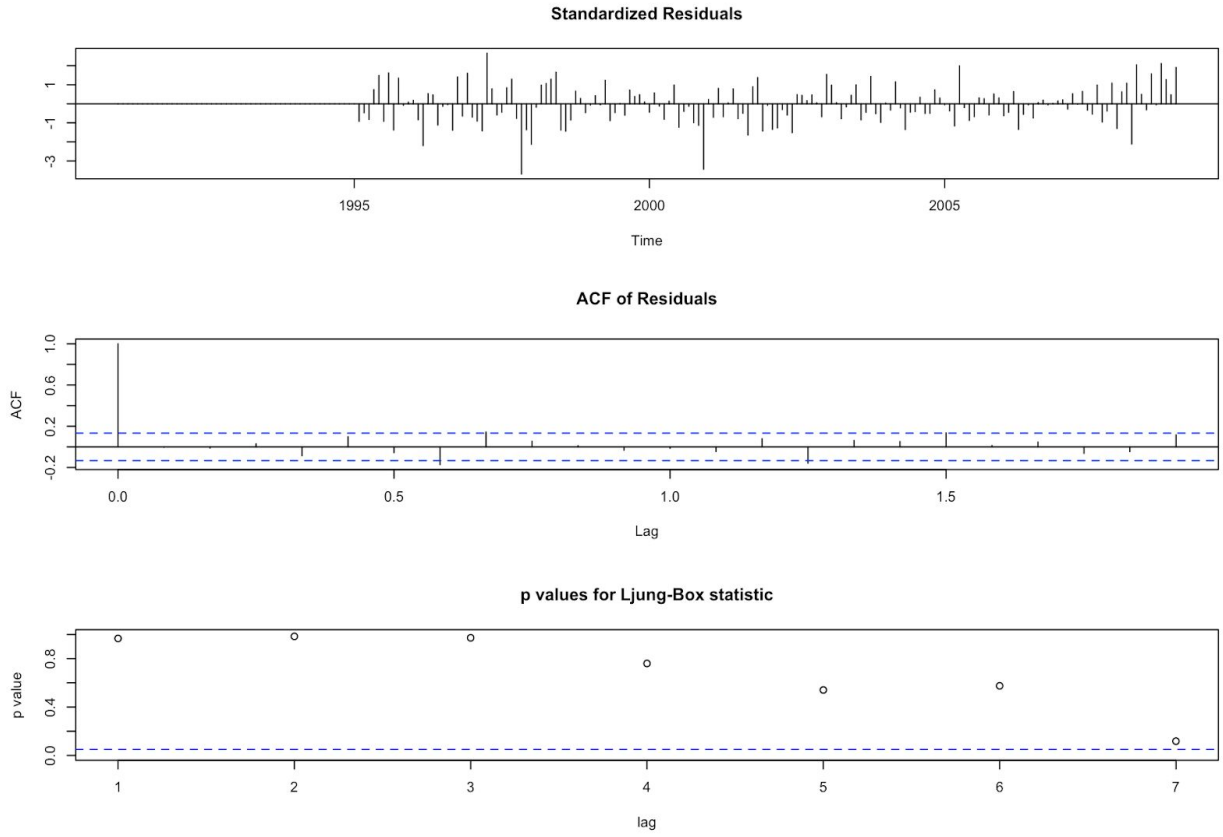
We followed all the same steps outlined in the SARIMA section above and added the unemployment data to the model. This led to the following results:

Table 2 Scores and parameters of different SARIMAX models

Order (p,d,q)(P,D,Q)(m = 12)	Method	Validation RMSE	AIC	Variance	Meet all assumptions?
(0,1,3)(3,1,2)	CSS	**0.00328**	286.3	0.00348	Yes
(2,1,3)(3,1,2)	CSS	0.00347	292.6	**0.00327**	Yes
(1,1,1)(2,1,2)	CSS	0.00333	**276.9**	0.003827	No
(1,1,3)(3,1,2)	CSS	0.00334	287.99	0.00343	Yes

Table 3 Statistical assumption tests of best SARIMAX model

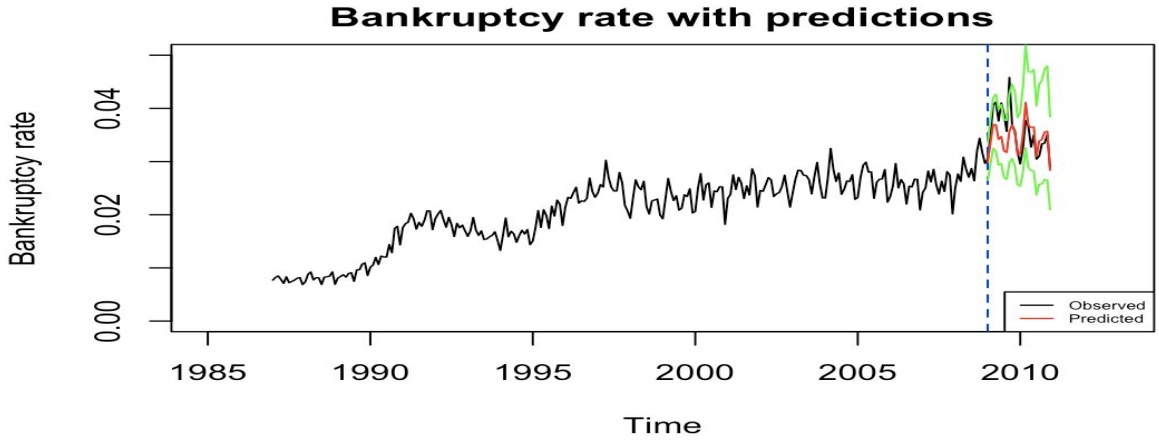
Test name	Assumption	p-value	Confidence Level	Pass or not?
T-test	0 means	0.36	95%	Pass
Shapiro-Wilk Test	Normality	0.07	95%	Pass
Levene Test	Homoscedasticity	0.24	95%	Pass
Ljung Box Test	0 correlation	Plot	95%	Pass



Graph 9 Plots related to zero correlation assumption of SARIMA model

Residuals once again appeared to be uncorrelated for small lags. All other model assumptions were satisfied as well.

Our final SARIMAX model: *p = 0, q = 3, P = 3, Q = 2, d = 1, D = 1.*



Graph 10 Predicted values of bankruptcy rate using SARIMAX

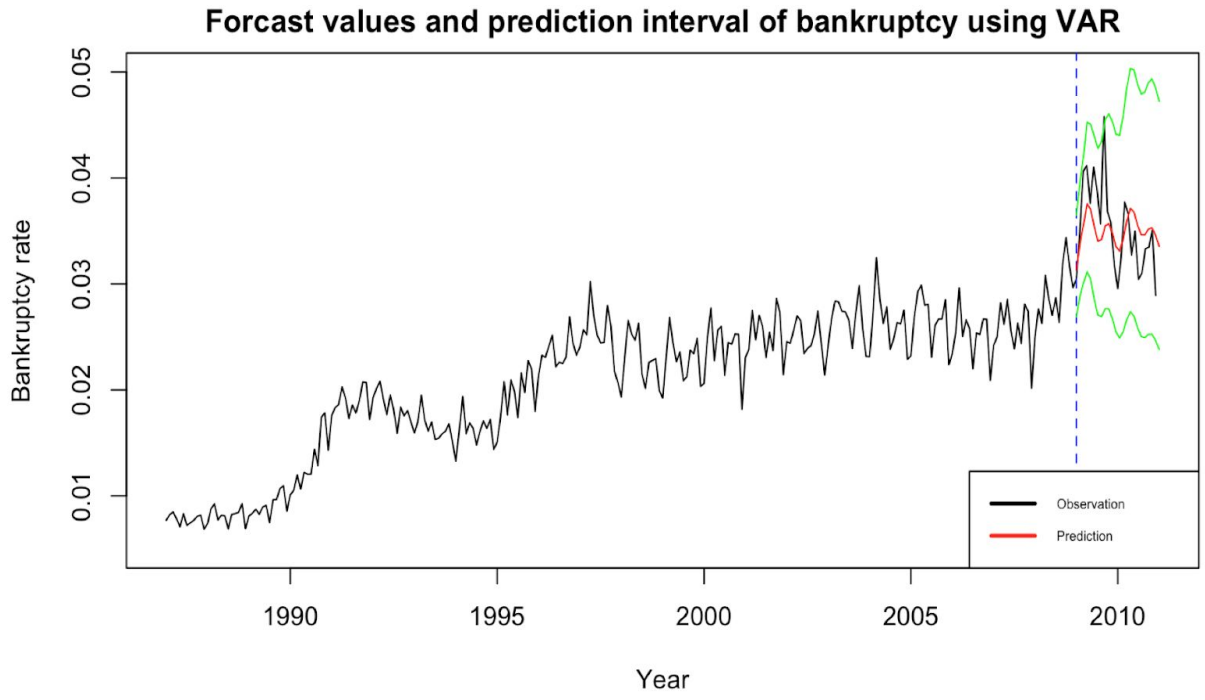
b. Vector Autoregression

Vector autoregression is a common statistical approach for modeling multiple time series that influence each other. It describes the evolution of a set of variables over the same time period, representing them as a linear function of their past values.

Since VAR treats each variable as in linear regression, we had to test all the statistical assumptions underlying linear regression for each variable, and only then we could justify the validity of our confidence intervals.

After trying different combinations of variables and hyperparameter values, the model with the smallest validation prediction error was chosen. In our chosen VAR model, bankruptcy rate and unemployment rate had been logarithmically transformed to stabilize the variance. House price index was discarded in our model because including it as an external variable hurt our model performance. We set hyperparameter p to 6.

Here, is a visualization of our VAR model. The prediction is smoother than the observed data, which during 2009 undergoes a dramatic shift not seen elsewhere in the data, making it less predictable. The trend is so abrupt that it cannot be well captured by just using historical data.



Graph 11 Forecast values and prediction interval of bankruptcy using VAR

All the assumptions corresponding to each variable hold for this model. With a confidence level of 95%, all p-values are higher than the threshold 0.05. Their values are as follows:

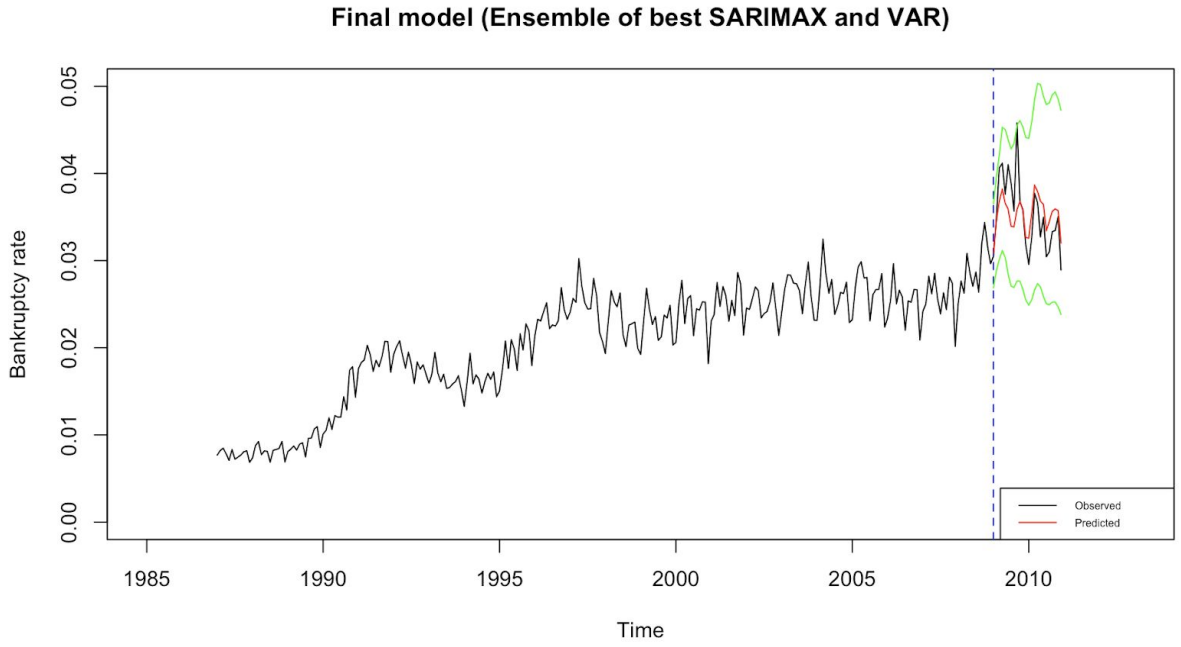
Table 4 P-values of assumption tests for each variable

Variable name	Zero-Mean	Zero-correlation	Homoscedasticity	Normality
Unemployment rate	0.9804	0.9168	0.3222	0.4850
Population	0.9800	0.3038	0.8907	0.7067
Bankruptcy rate	0.9995	0.9308	0.0790	0.1980

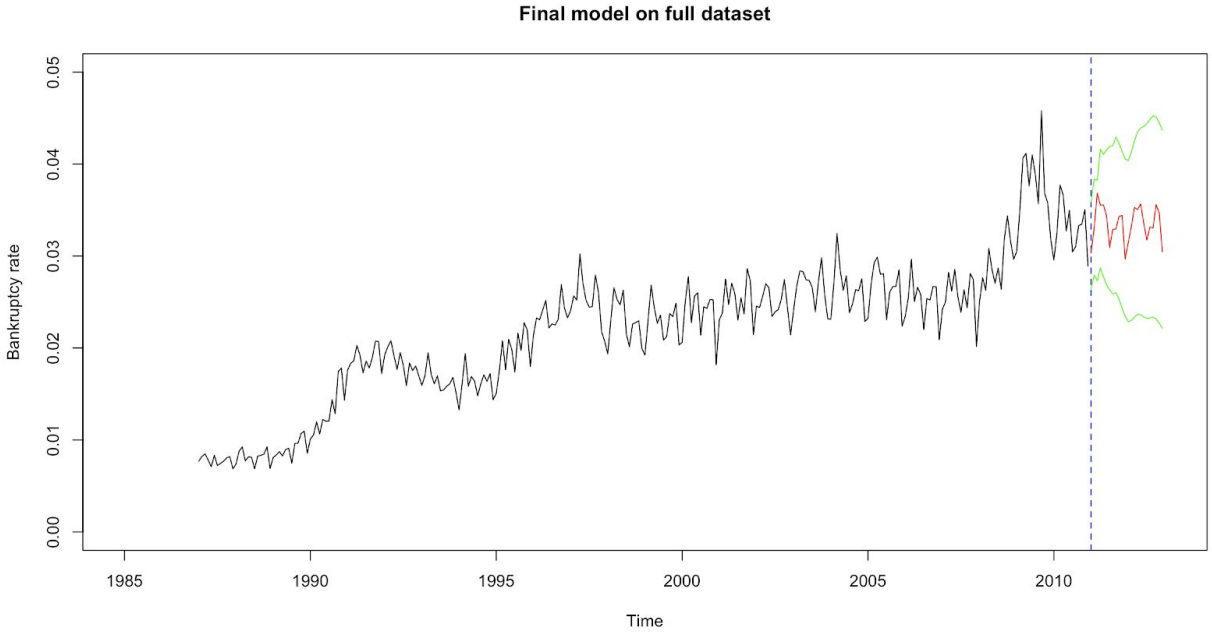
There are also a few limitations that come with VAR. Firstly, as mentioned above, every variable is assumed to influence every other variable in the system, which makes direct interpretation of the estimated VAR coefficients very difficult. Another limitation is that, VAR assumes all variables are linearly related, therefore missing potential nonlinear relationships between variables.

4. Conclusion

The goal of this project was to forecast Canadian bankruptcy rates for 2011 and 2012. Univariate SARIMA and Holt-Winters methods and multivariate SARIMAX and VAR methods were tested, but Holt-Winters was rejected after predicting poorly. The final model was an average of the SARIMAX and VAR models. This was chosen because it performed effectively as well as the individual models in terms of predictive accuracy, but also is more robust to new data since SARIMAX and VAR are not highly correlated and therefore provide diverse predictions. SARIMA was left out of the ensemble because it was highly correlated with SARIMAX. The widest prediction interval was used so as to remain more conservative. All model assumptions were tested again on the full dataset and passed. Our final model and forecasted time series is presented below.



Graph 12 Final forecast values and prediction intervals on validation set



Graph 13 Final forecast values and prediction intervals

5. Appendix

Table 1: Final forecast values for Canadian bankruptcy rate

Month	Predicted
2011/01	0.0303362
2011/02	0.0330493
2011/03	0.0368382
2011/04	0.0355263
2011/05	0.0355619
2011/06	0.0343551
2011/07	0.0309335
2011/08	0.0329125
2011/09	0.0329245
2011/10	0.0342876
2011/11	0.0344156
2011/12	0.0296764
2012/01	0.0314572
2012/02	0.0332060
2012/03	0.0352756
2012/04	0.0350621
2012/05	0.0356567
2012/06	0.0335783
2012/07	0.0317499
2012/08	0.0331478
2012/09	0.0330514
2012/10	0.0355786
2012/11	0.0346959
2012/12	0.0304683