Forecast of Bankruptcy Rates in Canada for 2011 – 2012

Nick Levitt, Jinxin Ma, Derek Welborn

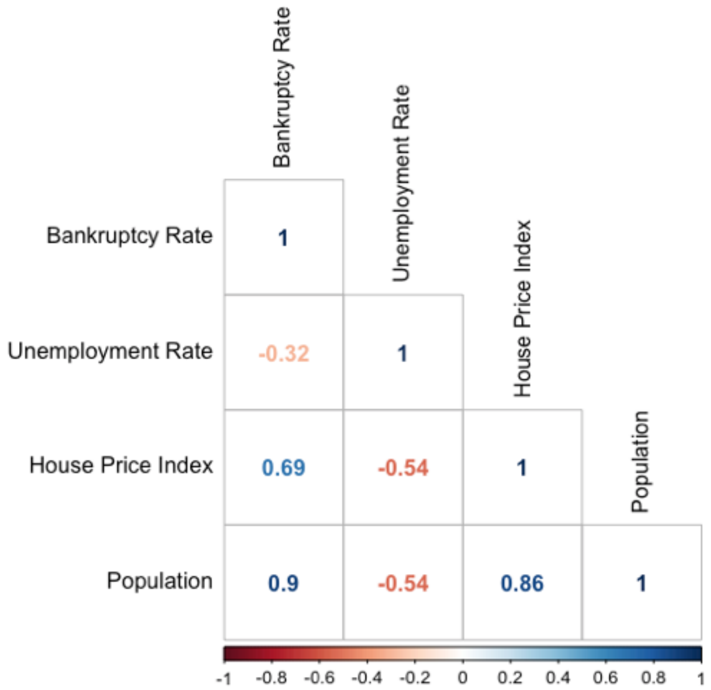
1. Introduction

This experiment sets out to build the optimal predictive model for forecasting bankruptcy rates in Canada for 2011 and 2012, using data from 1987 through 2010. Although many sources state that Canada avoided the 2008 financial crisis, Canadian bankruptcy rates still increased dramatically, and capturing such variation in a predictive model can prove tough. We show that, using external data, a strong predictive model can be built that performs well on this task.

2. Data Exploration

In the training set, we have monthly historical data for bankruptcy rates, housing price indexes, unemployment rates, and population from January 1987 to December 2010; the correlation, or mutual relationship, between these variables can be seen in Plot 1.

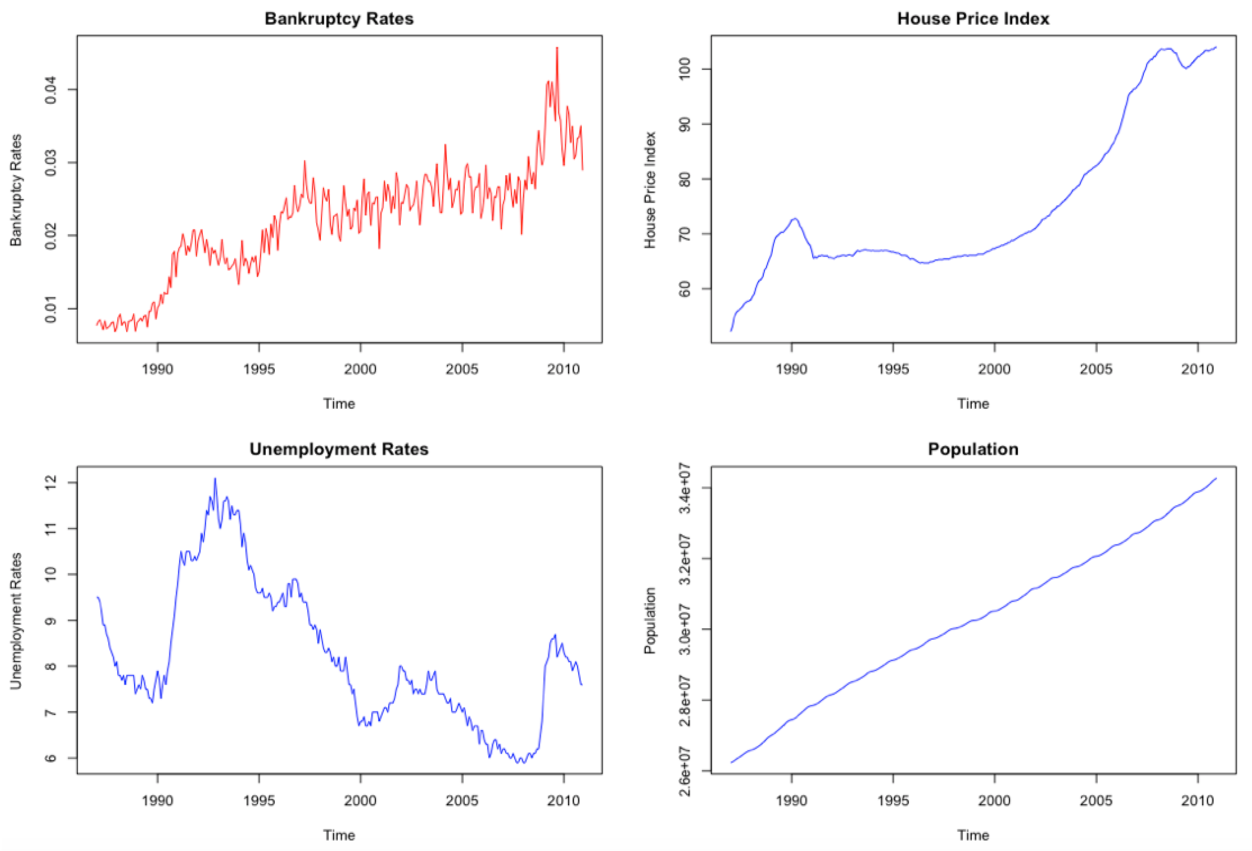
Plot 1: Correlation Plot



High correlation between variables is usually an indicator that they can be used to predict one another. Plot 1 shows that bankruptcy rates are highly correlated with population and housing price indexes.

In Plot 2, we can observe that over the years, bankruptcy rates and housing price indexes have exhibited similar trends. The paths of bankruptcy rates and unemployment rates show some similarity in the early 90's and again around 2008. Population on the other hand, displays an overall upward linear trend, which indicates it may not be helpful in predicting bankruptcy rates. In the next section, we use several modeling techniques to predict bankruptcy rates and compare the results.

Plot 2: Line Plots for All Variables over 1983 - 2010



3. Modeling

To try and accurately predict bankruptcy into the future, we implement several methods: Holt-Winters Smoothing Methods, Vector Autoregression (VAR) Models and SARIMAX models. We only have bankruptcy rates from 1987 through 2010, so to check the validity of our models we split this data in two:

- A sub-training set that has data for 1987 – 2005
- A validation set that has data for 2006 – 2010

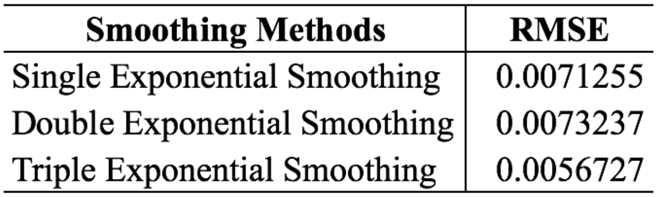
We train each model on the sub-training set, and use the validation set to determine how well the model performs. We use Root Mean Squared Error (RMSE), on the validation set as our primary metric for goodness-of-fit. RMSE is the average error we expect to make when making a prediction. Smaller RMSE values indicate better models. Finally, we train the best-performing model on the entire training set and use it to predict the bankruptcy rates for 2011 and 2012.

3.1 Holt-Winters Models

Holt Winters models use exponential smoothing to predict future values in a time series. This means that more recent observations have more weight in the prediction-making process than older observations; these types of models can be viewed as a moving window. Different Holt-Winters models have different assumptions, usually about the existence of trend and seasonality in the time series. In this case, trend can be thought of as the general direction of change over time, and seasonality can be thought of as the repeating pattern of change that occurs every 12 months.

We build three different Holt-Winters models: a single-exponential smoothing model, which assumes there is no trend and no seasonality in the time series, a double-exponential smoothing method, which assumes there is trend but no seasonality, and a triple-exponential smoothing method, which assumes there is both trend and seasonality. As can be seen in Table 1, triple-exponential smoothing performs the best.

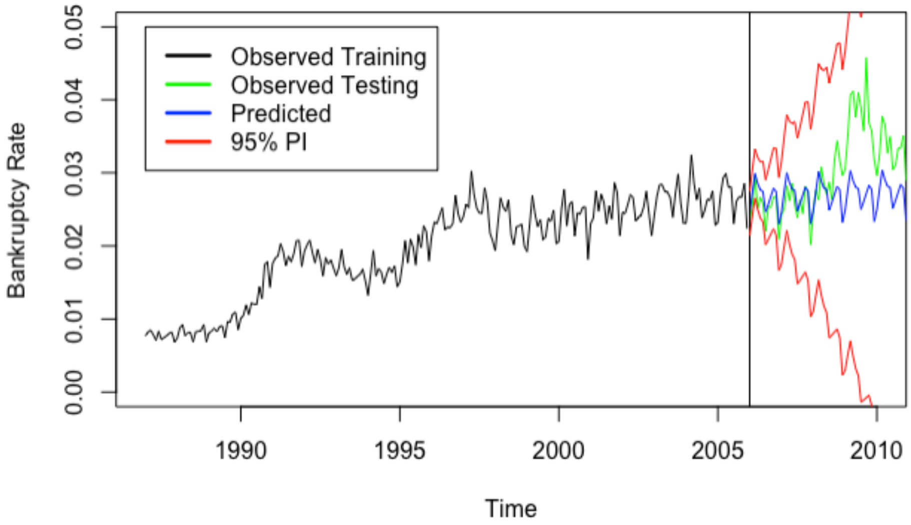
Table 1: RMSEs of Holt Winters Models



Smoothing Methods	RMSE
Single Exponential Smoothing	0.0071255
Double Exponential Smoothing	0.0073237
Triple Exponential Smoothing	0.0056727

Looking at the Plot 3 of the forecast overlaid on the observed validation data, it can be seen that the Holt-Winters model fails to capture the 2008 banking crisis. This is due to the bankruptcy rates during the crisis not adhering to the generalized trend and seasonality that had persisted previously. The red lines in Plot 3, and in future plots, are the prediction intervals. These are representative of the interval in which we are 95% confident the future values will lie. Smaller intervals are generally considered better, assuming the same level of confidence.

Plot 3: Holt Winters Forecast



Holt-Winters models use only the previous bankruptcy rates make predictions about the future. As mentioned earlier, bankruptcy rates are highly correlated with other variables, and the inclusion of these variables may increase the predictive power of our model. To test this hypothesis, we build a VAR model.

3.2 VAR Models

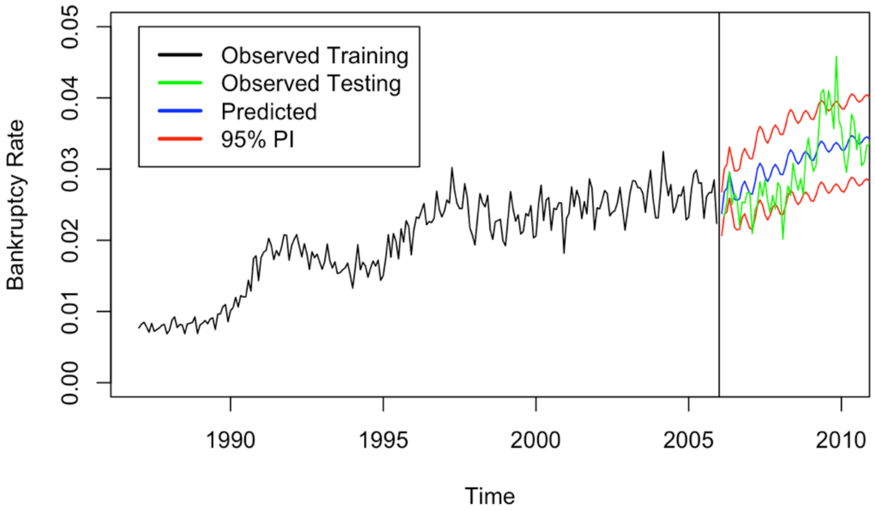
VAR models treat all the variables as endogenous. This means that all the variables, population, housing price index, unemployment and bankruptcy rates, influence each other. In a VAR model the value of a variable at any given time is a function of the its own previous values, as well as the previous values of the other endogenous variables. We build seven VAR models to predict bankruptcy rates using different combinations of the external variables, the results of which can be seen in Table 2.

Table 2: RMSEs of VAR Models

Variables	RMSE
All 3 Variables	0.003891
Excluding House Prince Index	0.004226
Excluding Population	0.004317
Excluding Unemployment Rate	0.003908
Excluding Population & Housing Price Index	0.005802
Excluding Unemployment Rate & House Prince Index	0.004005
Excluding Unemployment Rate & Population	0.003776

As can be seen, the highest performing model includes housing price indexes as the endogenous variable. Looking at Plot 4 of the forecast of this VAR model overlaid on the actual validation data it can be seen that the VAR model captures the trend of the data much better than the Holt-Winters model did. However, the VAR model still fails to capture the actual variation of the data.

Plot 4: VAR Forecast



3.3 SARIMAX Models

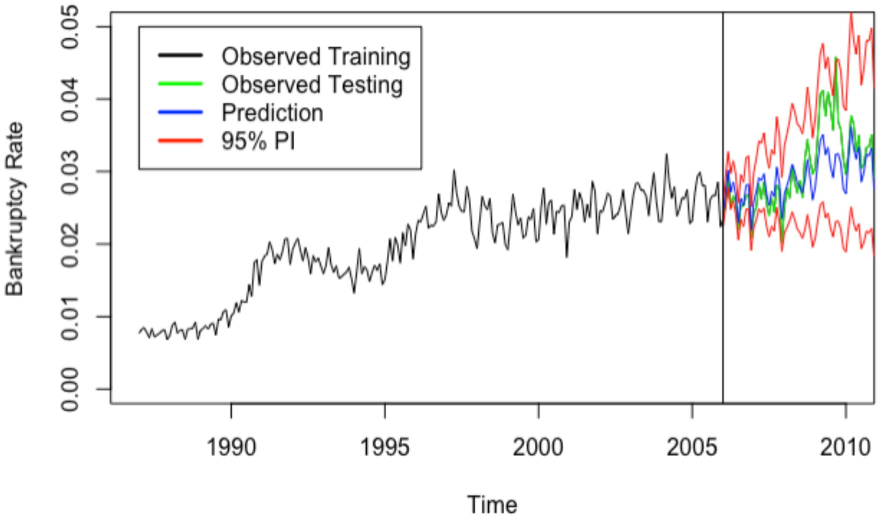
SARIMAX models can handle data with both trend and seasonality well by considering within-season and across-season variations. In our context, this means modeling the patterns in bankruptcy rates that occur every 12 months as well as occur generally over time. Like VAR models SARIMAX models incorporate external data to enhance predictive power. While the optimal VAR model included population as an endogenous variable, the low correlation between population and bankruptcy rates lead us to exclude it when building our SARIMAX model.

After testing many models, our optimal SARIMAX model is as follows:

$$\text{SARIMA } (2, 1, 2) \times (3, 1, 3)_{12} + \textit{House Price Index} + \textit{Unemployment Rate}$$

We state this model without explaining its intricacies as that is outside the scope of this report. Further details, if needed, are available upon request. The validation RMSE of this model is 0.00339, which suggests this model has stronger predictive power than both the Holt-Winters and VAR models. The predictions for the validation set are shown in Plot 5. Although the SARIMAX does not fully capture the 2008 crisis, it does captures the variation better than any of the previous models. As a result, further discussion and final predictions will use this model.

Plot 5: SARIMAX Forecast



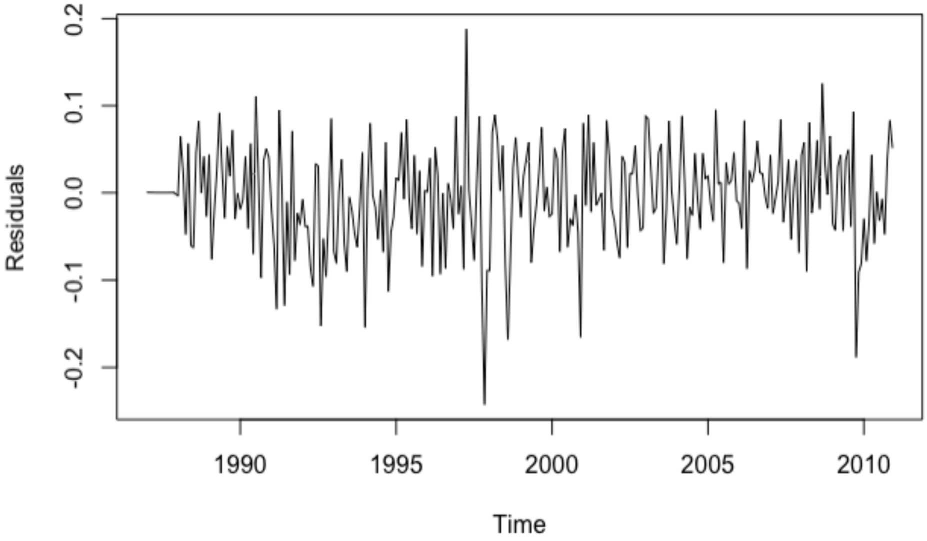
4. Residual Diagnostics

In order to build a robust model with any kind of accuracy and precision. We require the difference between the model's predictions and the true values, i.e. the residuals, to meet certain assumptions:

- The residuals do not vary with time
- The mean of the residuals is zero

Plot 6 below illustrates both assumptions holding. Other than an outlier in 1997 the sizes of the peaks and valleys in the plot remain constant. The residuals are also evenly distributed around 0.

Plot 6: Residuals vs. Time

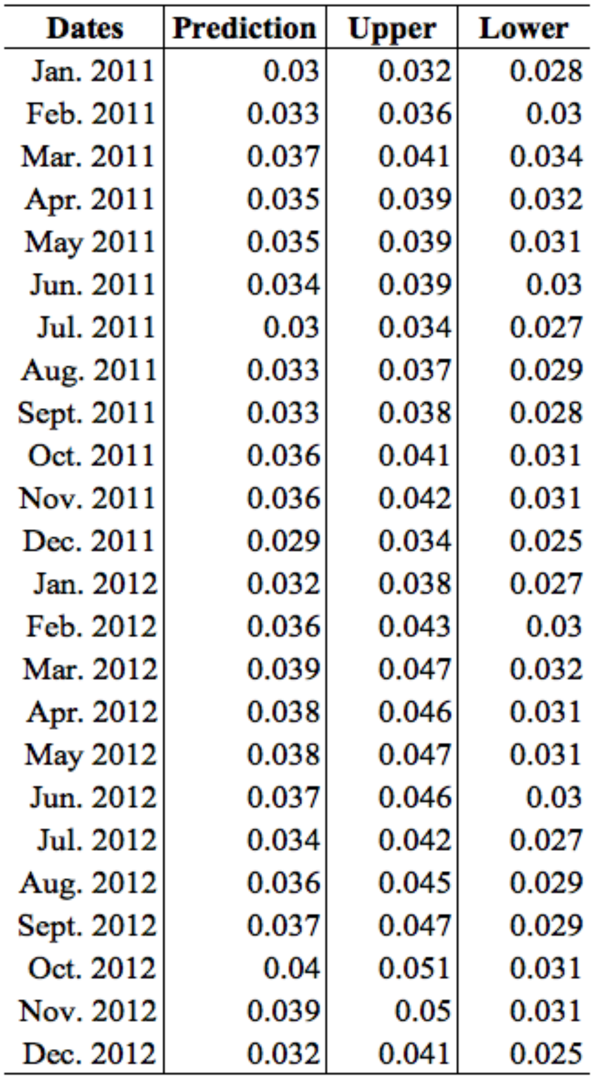


5. Prediction and Conclusion

We implemented the selected SARIMAX model to predict bankruptcy rates into 2011 and 2012. The graphical and tabular representations of the SARIMAX model's predictions can be seen in Table 3 and Plot 7, respectively.

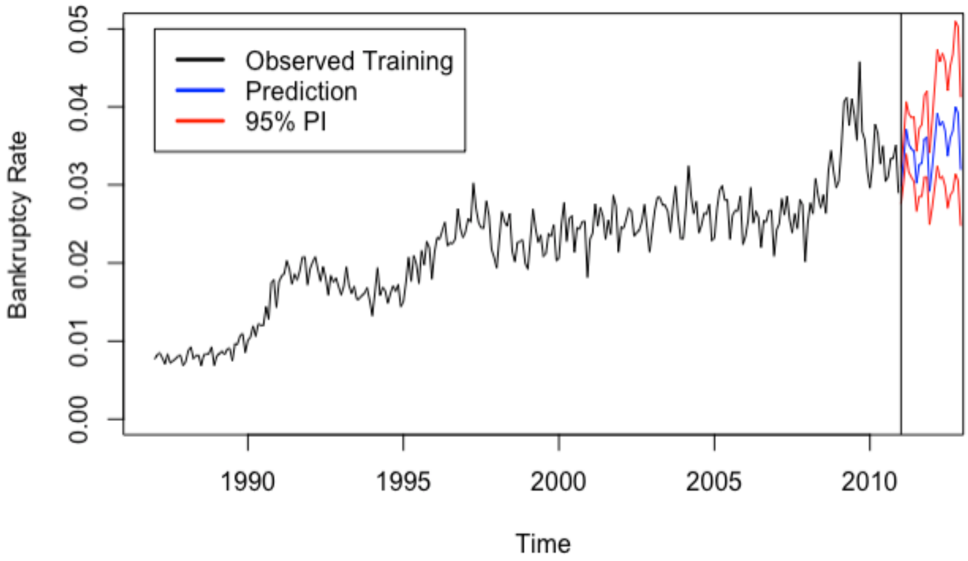
In summary, we used Holt-Winters Methods, VAR models, and SARIMAX models to fit the bankruptcy rates. We split the training set into a sub-training set and a validation set. Each model was fitted on the sub-training set, and then tested on the validation set. Among all models, the SARIMAX model that incorporated house price index and unemployment rate had the best performance.

Table 3: Predictions with 95% Prediction Intervals



Dates	Prediction	Upper	Lower
Jan. 2011	0.03	0.032	0.028
Feb. 2011	0.033	0.036	0.03
Mar. 2011	0.037	0.041	0.034
Apr. 2011	0.035	0.039	0.032
May 2011	0.035	0.039	0.031
Jun. 2011	0.034	0.039	0.03
Jul. 2011	0.03	0.034	0.027
Aug. 2011	0.033	0.037	0.029
Sept. 2011	0.033	0.038	0.028
Oct. 2011	0.036	0.041	0.031
Nov. 2011	0.036	0.042	0.031
Dec. 2011	0.029	0.034	0.025
Jan. 2012	0.032	0.038	0.027
Feb. 2012	0.036	0.043	0.03
Mar. 2012	0.039	0.047	0.032
Apr. 2012	0.038	0.046	0.031
May 2012	0.038	0.047	0.031
Jun. 2012	0.037	0.046	0.03
Jul. 2012	0.034	0.042	0.027
Aug. 2012	0.036	0.045	0.029
Sept. 2012	0.037	0.047	0.029
Oct. 2012	0.04	0.051	0.031
Nov. 2012	0.039	0.05	0.031
Dec. 2012	0.032	0.041	0.025

Plot 7: Forecast on Test Set from SARIMAX


Observed Training
Prediction
95% PI
Bankruptcy Rate
Time