Forecasting Canadian Bankruptcy Rates

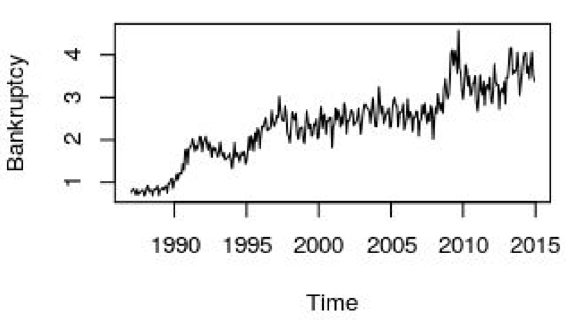
Byron Han, Louise Lai, Marwa Oussaifi, Shivee Singh

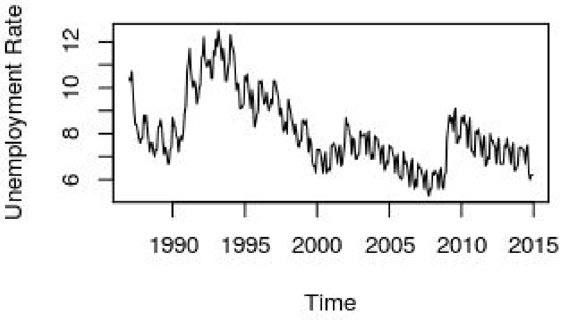
I. Objective

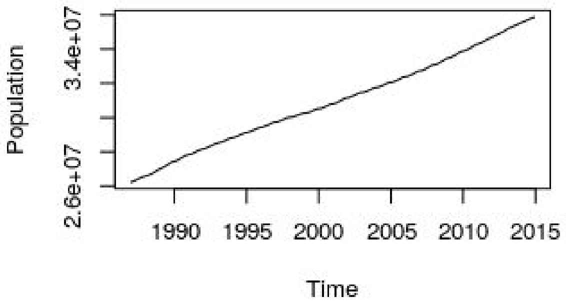
The objective of this report is to provide a forecast of monthly bankruptcy rates in Canada from January 2015 to December 2017 to justify how these predictions were generated. Four variables were available for consideration. They are:

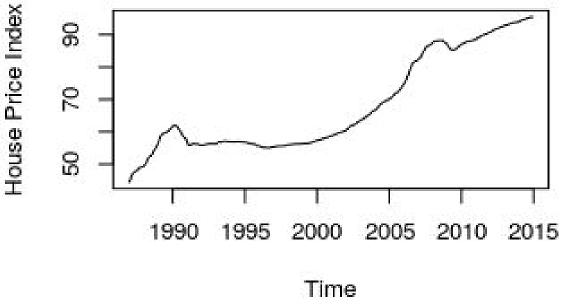
- Bankruptcy Rate (%)
- Unemployment Rate (%)
- Population
- Housing Price Index

The data are collected over 28 years - from January 1987 to December 2014 - with one entry each month.









The four charts above illustrate the four variables and their relation to one another. In the top left corner is the target variable, Bankruptcy Rate. Along with Population and House Price Index, it displays an upward trend over the years. Unemployment Rate loosely trends downwards. This suggests that Bankruptcy may be moving in sync (i.e. positively correlated) with Population and Housing Price Index, and inversely with Unemployment Rate. This report further explores these relationships below.

II. Definitions

Here are some definitions that are used throughout this report.

Trend: the overall shape of a time series

Seasonality: a repeated pattern within a certain time frame. For example, ice-cream sales fluctuate in a predictable pattern that repeats itself every 12 months.

Residuals: A residual is an actual value minus the predicted value. In this case, it is the actual bankruptcy rate minus the predicted bankruptcy rate.

$$residual = y_{actual} - y_{predicted}$$

RMSE: Residual Mean Square Error is an evaluation metric used to judge the predictive power of our model. Lower RMSE values are desirable because they indicate that the predicted bankruptcy rate is close to the actual bankruptcy rate.

$$RMSE = \sqrt{\frac{\Sigma(y_{actual} - y_{predicted})^2}{n}}$$

III. Available Methods

There are several models available for forecasting that fall into two broad categories that are explored in this report: **univariate** and **multivariate** models.

The univariate approach relies solely on the response variable's own history to find the most appropriate model.

The multivariate approach, on the other hand, takes into consideration the impact of other variables as well and helps overcome some weaknesses of the univariate approach. Models from both categories have their own strengths, weaknesses and unique characteristics. Hence, this report explores each model, tunes the models to its best possible performance, and reports the final result at the end.

IV. Model Selection

To choose an optimal model we need to compare the performance of models on data they have not seen before. This would help us determine whether the model has generalized enough to be useful for forecasting. We applied an **80:20** split to our data, creating two sets called train and validation. The training set is used to develop models while the validation set is used to compare choose our final model.

Training Data	Validation Data	Test Data
1987-2009 (23 years)	2010-2014 (5 years)	2015-2017 (2 years)

The final model will be one that minimizes Root Mean Square Error (RMSE) is our metric on a validation set. This metric is chosen because our objective is to accurately forecast the bankruptcy rates and RMSE favors the model with closest predictions.

V. Modeling

Univariate Models

SARIMA

SARIMA models can be used to model a range of time series from simple ones which have no trend or seasonality to the complex time series with both trend and seasonality structures present.

Modeling approach

For fitting SARIMA model we first plotted the raw time series and observed it was heteroscedastic which means that the amplitudes of observations were not constant over time. SARIMA models rely on certain assumptions one of them being residuals have a constant variance or homoscedasticity. In order to overcome this problem, we applied BoxCox transformation on the raw time series with appropriate parameter lambda 0.2.

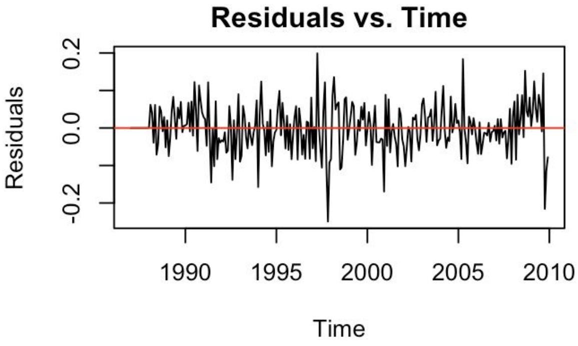
The raw time series clearly shows a trend in the upward direction. In order to fit the SARIMA model, we need to eliminate trend and any seasonality present. To eliminate trend we did an ordinary differencing of the series. Next, we used autocorrelation and partial correlation plots to find initial model parameters.

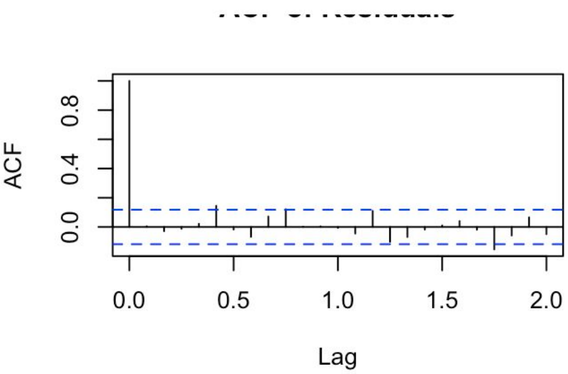
We fit several models and shortlisted handful on the basis of how well they were fitting the existing data. Then we used the shortlisted models for forecasting bankruptcy rates and compared against the validation set. The model which gave us the lowest prediction error (root mean squared error) was chosen as the optimal SARIMA model.

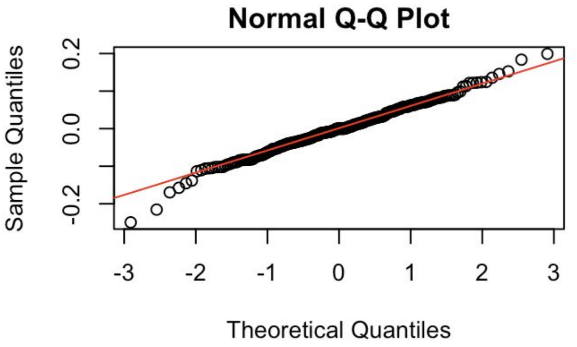
Assumptions

SARIMA models assume that error terms or residuals have expected-value of zero, are having constant variance and are uncorrelated with each other. Also since we decided to use Maximum likelihood estimation we have inherently assumed a normal distribution of data. These assumptions need to validate in order to finalize this model.

We verified all the above assumptions are being satisfied by our optimal SARIMA model. The graphs below confirm the same.







Results

Our optimal SARIMA(3,1,0)(2,1,3)[12] model with lambda 0.2 and Maximum likelihood estimation technique performed better than auto.arima results. It also satisfies the model assumptions. The RMSE we obtained on the validation set is 0.537. The limitation of this model is that it does not take into account the effect of other variables. This can be overcome by using SARIMAX approach which we explore later.

Exponential Smoothing

This technique uses a set of recursive equations which are have exponentially decreasing weights over time. The idea is to decompose the series into three parts:

- Level: an estimate of the local mean
- Trend: an estimate of the general ‘shape’ of the data
- Seasonality: an estimate of the pattern occurring within a fixed period

Modeling approach

There are three types of exponential smoothing: Single, Double and Triple. Single exponential smoothing is appropriate when the model does not display any trend or seasonality, double exponential smoothing for when the model displays trend and triple for when both trend and seasonality are present. Since our model exhibits trend and maybe seasonality, we tried both double and triple exponential smoothing.

Results

Double exponential performed better than triple exponential smoothing. The parameters were [alpha: 0.39, beta : 0.024, gamma: FALSE]. RMSE obtained with this method was 0.78 for the validation set which is higher than our optimal SARIMA model.
Exponential smoothing models are flexible since there is no assumption about the underlying distribution of data, but modeling a complex time series with just a handful of parameters makes this choice an over simplistic choice.

This report now moves on from univariate modeling to multivariate modeling.

Multivariate Models

To understand the different types of multivariate models, we must first understand the concept of response, explanatory, exogenous and endogenous variables.

Since the bankruptcy rate is the variable we are trying to predict, we call it the **response variable**, and the other variables (i.e. population, housing price index, unemployment) are called **explanatory variables**.

If we believe that an explanatory variable influences the response variable, but not vice versa then the explanatory variable is called **exogenous.** If we believe that both the explanatory variable and response variable mutually influence each other we call them **endogenous.**

SARIMAX

SARIMAX is a multivariate extension to SARIMA where ‘X’ stands for exogenous variables. It gives us more flexibility in terms of adding additional predictor variables which might influence our response variable.

Modeling approach

Therefore, it makes sense to build SARIMAX from the best performing SARIMA model. We used the optimal SARIMA model and tried different combinations of exogenous variables.

Assumptions

The assumptions of SARIMAX are same as SARIMA. We verified the assumptions related to residuals and results were satisfactory.

Results

The optimal SARIMAX model suggests that we include only the unemployment variable, the predictive power of our model is at its best. The RMSE for optimal SARIMAX is 0.4326. As we can see this has improved from Univariate methods.

Limitation of this model is that it ignores interdependence. We believe that there is a possibility that some variables might be endogenous, i.e. they influence each other, hence we shall explore VAR and VARX models.

VAR

In this model, the variables are treated symmetrically. The model accounts for the fact that variables can have an effect on each other.

Modelling approach

We choose lag to be 11 because based on the selection criteria and cross-correlation of variables. The selection criteria is an indication of the amount of predicting power increased accounting for the number of variables increased and the size of data. The cross-correlation is an indication on how much back one predicting variable affects the other. We choose this number not to overfit the data as well as having a good predicting power.

Results

After exhaustively searching for the least RMSE for all combination of predicting variables, RMSE is the lowest (0.396) when we include only unemployment as our additional predicting variables.

The winner of the VAR model suggests we should only add the unemployment rate as the additional predict variable. For example, the coefficient of lag 11 unemployment rate is significantly larger than the other lags of unemployment rate around it. This suggests the housing price index and bankruptcy influence each other most significantly after 1 year of lag.

VARX

It is a combination of VAR with exogenous variables. It is more flexible than the VAR model.

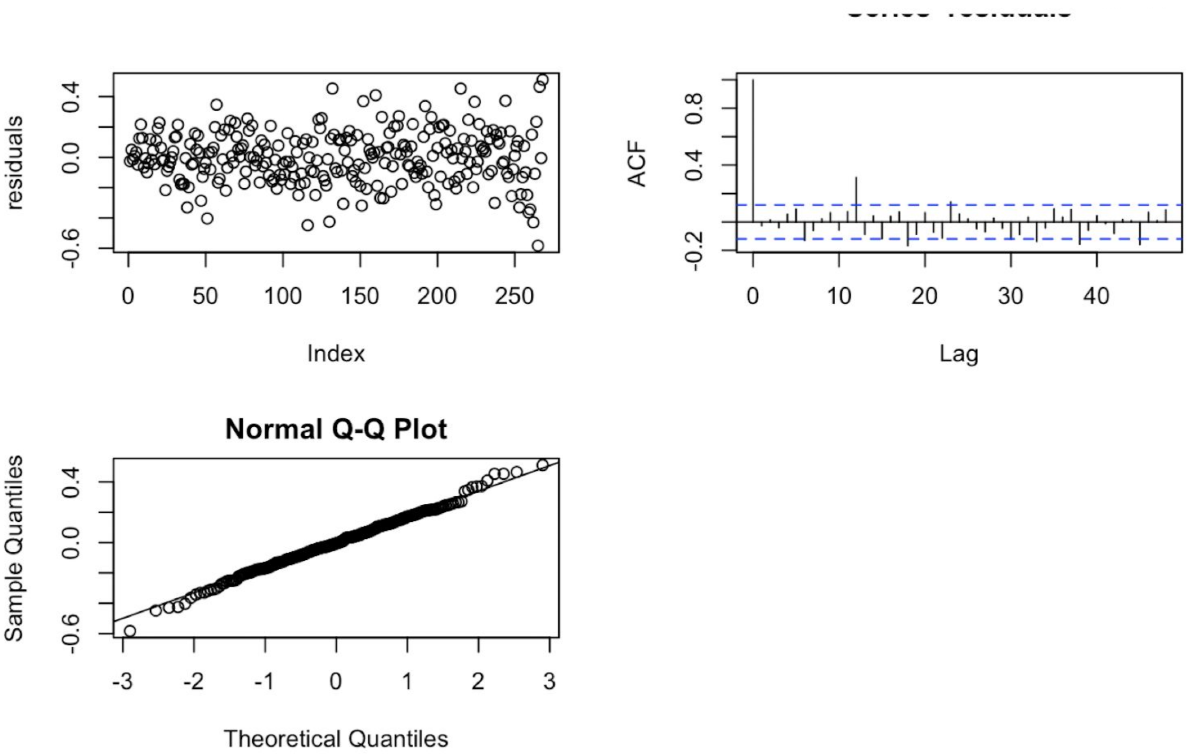
Results

We also consider if adding other variables to our best-performing VAR model will increase its predicting power. So we added the other two variables as exogenous variables.

After this, the RMSE is significantly lowered to 0.3259741.

Assumptions

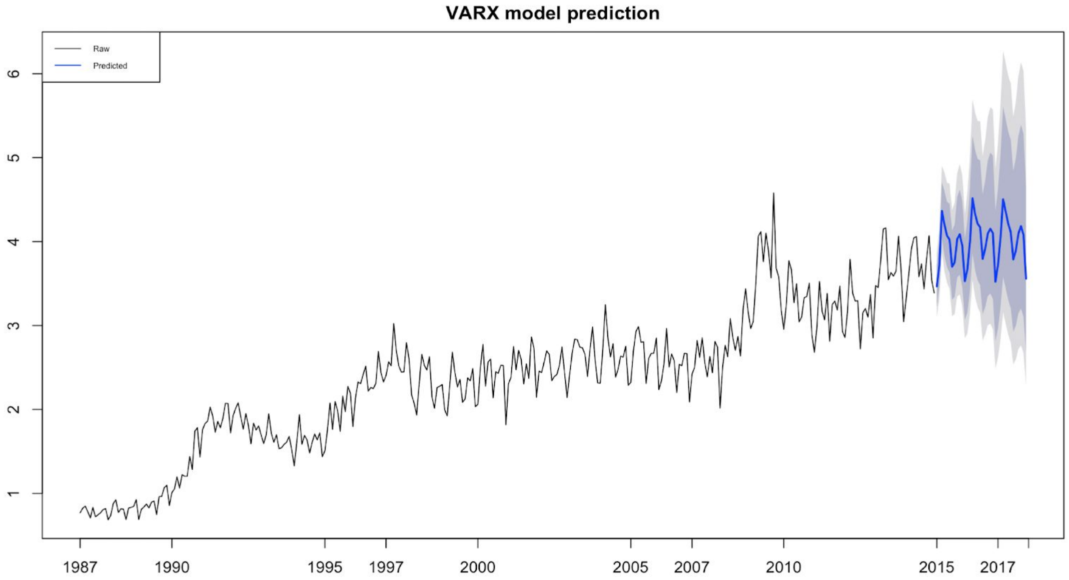
VAR and VARX model are built on similar assumptions as the SARIMA model
The optimal VARX model satisfies all the model assumptions. This was verified graphically using below plots residual vs time, autocorrelation and Q-Q plot.


residuals
0.4
0.0
-0.6
0
50
100
150
200
250
Index
ACF
0.8
0.4
-0.2
0
10
20
30
40
Lag
Normal Q-Q Plot
Sample Quantiles
0.4
0.0
-0.6
-3
-2
-1
0
1
2
3
Theoretical Quantiles

IV. Summary of results

Model	RMSE	Parameters
Univariate		
SARIMA	0.537	ARIMA(3,1,0)(2,1,3)[12]
Exponential Smoothing	0.789	alpha 0.48, beta=0.02, gamma=FALSE
Multivariate		
SARIMAX (using only unemployment rate)	0.432	ARIMA(3,1,0)(2,1,3)[12]
VAR	0.396	P=11 bankruptcy, unemployment
VARX	0.326	P=11 Endogenous: bankruptcy, unemployment Exogenous: housing, population

The above table summarizes all the models and their RMSE. We see that VARX has the lowest RMSE, hence we will use this model to generate our prediction for Canadian bankruptcy rates for data from 2015 to 2017.



V. Evaluation and Limitations

The lag order in VARX model is limited by the data we have as well as the extent we prevent overfitting. Therefore, we are only modeling the linear correlation within lags of 1 year. If unemployment or other variables affect bankruptcy in lag longer than 12, we will miss it. Moreover, VARX is a linear model so it is limited in that sense.

For VARX we need to first predict the exogenous variables in order to predict our response variable. Hence we are relying on methods to give us accurate predictions of other exogenous variables which might be difficult to obtain.

VI. Conclusion

In conclusion, the model with the best performance is the VARX model with bankruptcy and unemployment as endogenous variables and housing and population as exogenous variables.

Intuitively this makes sense, as the model is indicating that unemployment can affect Canadian bankruptcy rates and vice versa, while the relationship between housing and population are only unidirectional. Unemployment tends to have a domino effect, leading to a drop in consumer spending and pushing some businesses into bankruptcy to more layoffs, thus increasing unemployment. Thus, it is logical that treating unemployment as an endogenous variable resulting in the best model.