Time Series Analysis of
Canadian National Bankruptcy Rates

Wei Wei, Grace Zhang, Nina Hua, Anna Zeng

1. Problem Description

In Canada, bankruptcies account for insolvent corporations who cannot repay their debts to creditors and carry on with their business. The goal of this report is to forecast monthly bankruptcy rate for the years from 2015 to 2017 using time series models. The training dataset holds monthly data for 1) bankruptcy rate, 2) unemployment rate, 3) population size, 4) housing price index from 1987 to 2014. Housing Price Index (HPI) is a measurement of average price changes in repeat sales or refinancing on the same single-family houses.

2. Data

The training dataset has 4 time series, each has 336 observations from January 1987 to December 2014:

- *Bankruptcy Rate* - Response variable. National bankruptcy rate of Canada.
- *Unemployment Rate* - Unemployment Rate of Canada.
- *Population* - Number of inhabitants in Canada.
- *House Price Index* - A metric that measures changes in single-family home prices across a designated market in Canada.

Unemployment_Rate	Population	Bankruptcy_Rate	House_Price_Index
Min. : 5.300	Min. :26232423	Min. :0.6862	Min. :44.40
1st Qu.: 7.000	1st Qu.:28818164	1st Qu.:1.8088	1st Qu.:56.30
Median : 7.700	Median :30805908	Median :2.4387	Median :60.10
Mean : 8.087	Mean :30947626	Mean :2.3675	Mean :68.06
3rd Qu.: 8.925	3rd Qu.:33075674	3rd Qu.:2.8536	3rd Qu.:85.90
Max. :12.500	Max. :35872667	Max. :4.5798	Max. :95.50

Table1. Basic Statistic Summary of Data

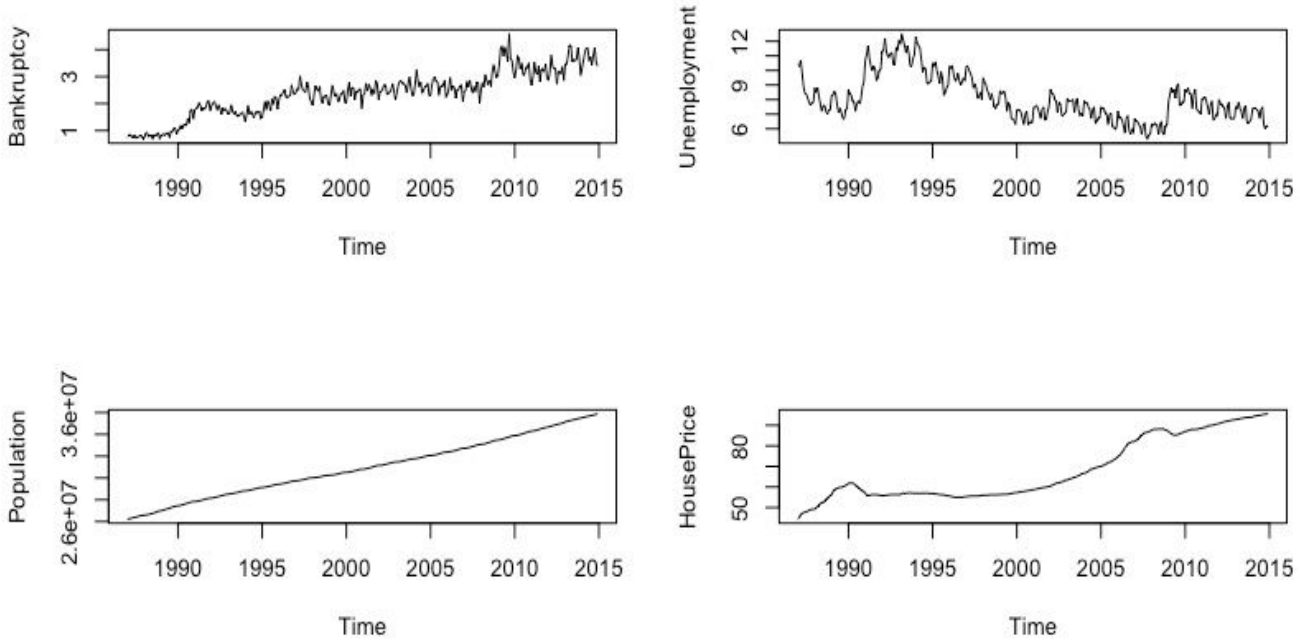

Figure 1. Time Series Trends Over Time (1987 - 2014)

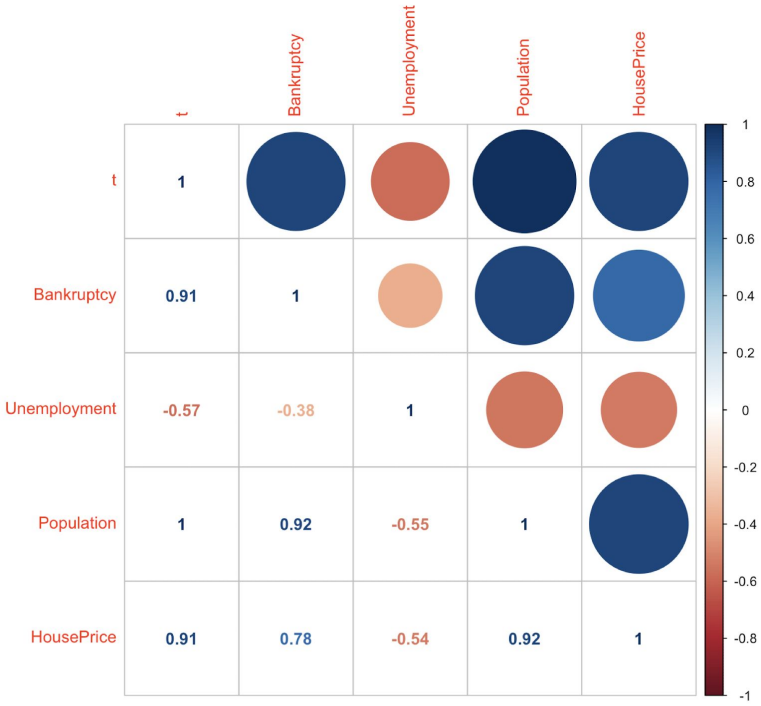

Figure 2. Cross-Correlations of Variables

3. Methodology

In order to choose the best performing model, we split the available training dataset into two parts: training and validation. Specifically, we use the training data to fit the models and use the validation data to compare different models and find the optimal model with the most appropriate parameters. Further, we use the optimal model with selected hyperparameters and pass the data from 2015 to 2017 to the model to get the forecasts.

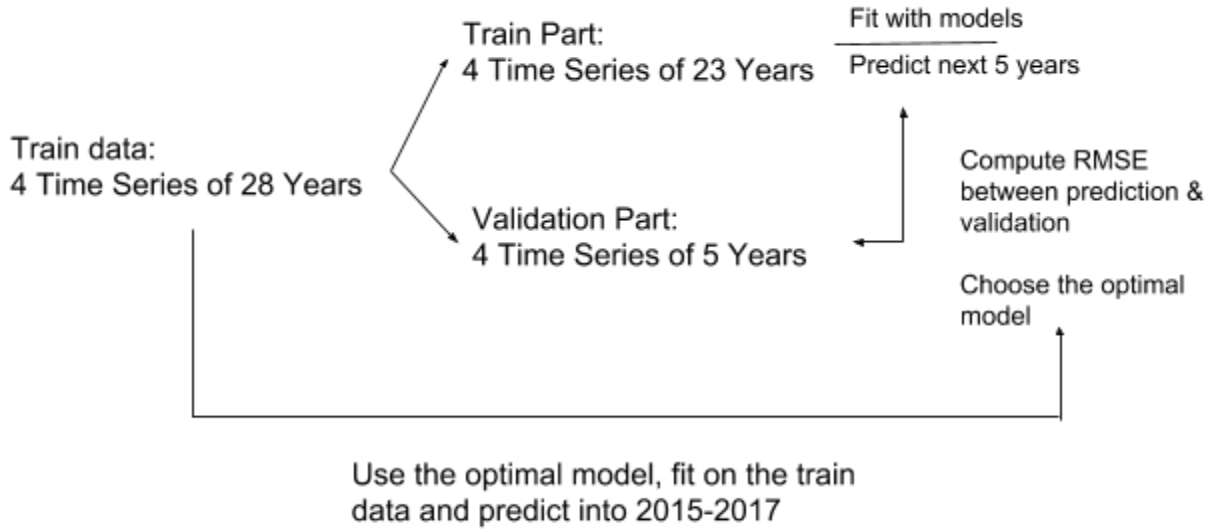

Figure 3. Methodology Flow

4. Available methods

In *Figure 1.*, we can see there are a consistent directional movement along the time (Trend) and some regular, predictable fluctuations according to some period (Seasonality). To capture both the trend and seasonality of the time series and make accurate predictions, we explored the following methods:

- SARIMA (Seasonal Autoregressive Integrated Moving Average)

The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. One shorthand notation for the model is: ARIMA(p, d, q) × (P, D, Q)S. With p = non-seasonal AR order, d = non-seasonal differencing, q = non-seasonal MA order, P = seasonal AR order, D = seasonal differencing, Q = seasonal MA order, and S = time span of repeating seasonal pattern.

- Exponential Smoothing (Holt-Winters)

Holt-Winters models use a set of recursive equations that do not require meeting distributional assumptions. There are 3 types of exponential smoothing functions. The model that is selected depends on whether or not the data exhibits trend and/or seasonality. Since bankruptcy rates display both trend and seasonality, a model that uses the Triple Exponential Smoothing method is most appropriate. The Triple Exponential Smoothing has the following notation.

$$
\begin{aligned}
s_0 &= x_0 \\
s_t &= \alpha \frac{x_t}{c_{t-L}} + (1-\alpha)(s_{t-1} + b_{t-1}) \\
b_t &= \beta(s_t - s_{t-1}) + (1-\beta) b_{t-1} \\
c_t &= \gamma \frac{x_t}{s_t} + (1-\gamma) c_{t-L} \\
F_{t+m} &= (s_t + m b_t) c_{t-L+1+(m-1) \mod L},
\end{aligned}
$$

$\mathbf{x_t}$ = response variable time series

$\mathbf{s_t}$ = smoothed value of the constant part for time t

$\mathbf{b_t}$ = the sequence of best estimates of the linear trend that are superimposed on the seasonal changes

$\mathbf{c_t}$ = the sequence of seasonal correction factors

$\mathbf{F_{t+m}}$ = an estimate of the value of x at time t+m

α = data smoothing factor; β = trend smoothing factor, and γ = seasonal change smoothing factor ; $0 < \alpha, \beta, \gamma < 1$

- VAR (Vector Autoregression)

Vector Autoregressive models are used for multivariate time series. Multivariate time series is composed of the response variable and correlated explanatory variables. These variables are assumed to be endogenous, meaning that there is a bidirectional relationship between the response and explanatory variables. A VAR model would be an appropriate

candidate model because bankruptcy rate, arguably, has a bidirectional relationship with unemployment rate, population size, and housing price index.

$$\mathbf{y_t} = \mathbf{c} + \mathbf{A_p Y_{t-p}} + \epsilon_\mathbf{t}$$

$\mathbf{y_t}$ = response variable time series

$\mathbf{c}$ = constant values

$\mathbf{A_p}$ = matrix of coefficients associated with the response time series

$\mathbf{Y_{t-p}}$ = p observations of response variable time series

$\boldsymbol{\epsilon_t}$ = error terms that are White Noise distributed with zero mean and a constant variance of σ_k^2 for k = 1, 2, 3, r

r = number of time series variables

p = number of lags

- VARX (Vector Autoregression + Explanatory Variable)

  Vector Autoregressive Process with Exogenous Variables model is similar to the VAR model, but a VAR process can be affected by other observable variables that are determined outside the system of interest instead of those of bidirectional relationships. Such variables are called exogenous variables.

  Though we assume all the variables are endogenous in the VAR model, intuitively, it is still possible that the bankruptcy rate doesn't influence the population directly. So we set the population as the exogenous variable and all the others as endogenous variables to see if it gives a better result.

5. Final Method

To obtain an optimal model, we fit the above-mentioned time series models and compared the goodness-of-fit metrics of these models, specifically the root mean squared error (RMSE) on the validation set and Akaike Information Criterion (AIC) on the training set. RMSE describes how concentrated the true data is around the line of best fit, i.e., the smaller the RMSE is the better. AIC focuses on the trade-off between the goodness-of-fit of the model (likelihood

function) and the simplicity of the model. The larger the likelihood and the fewer parameters, the better the model. Specifically, we prefer a smaller AIC.

Table 2 shows the best parameters for the four time series models and their goodness-of-fit metrics. We observed that VAR model with lags of 9 provides the best RMSE, following by VARX with lags of 10 and using population as an exogenous variable. However, AIC wise, the SARIMA(3,1,5)(1,0,1)[12] model actually outperforms the other models with an AIC score of -613.88.

We determined to choose the VAR model as our final "optimal" and will further justify our selection in the following paragraphs. However, we believe that the SARIMA model would also be a great option mainly because it is a simple model and it addresses both trend and seasonality in our data well.

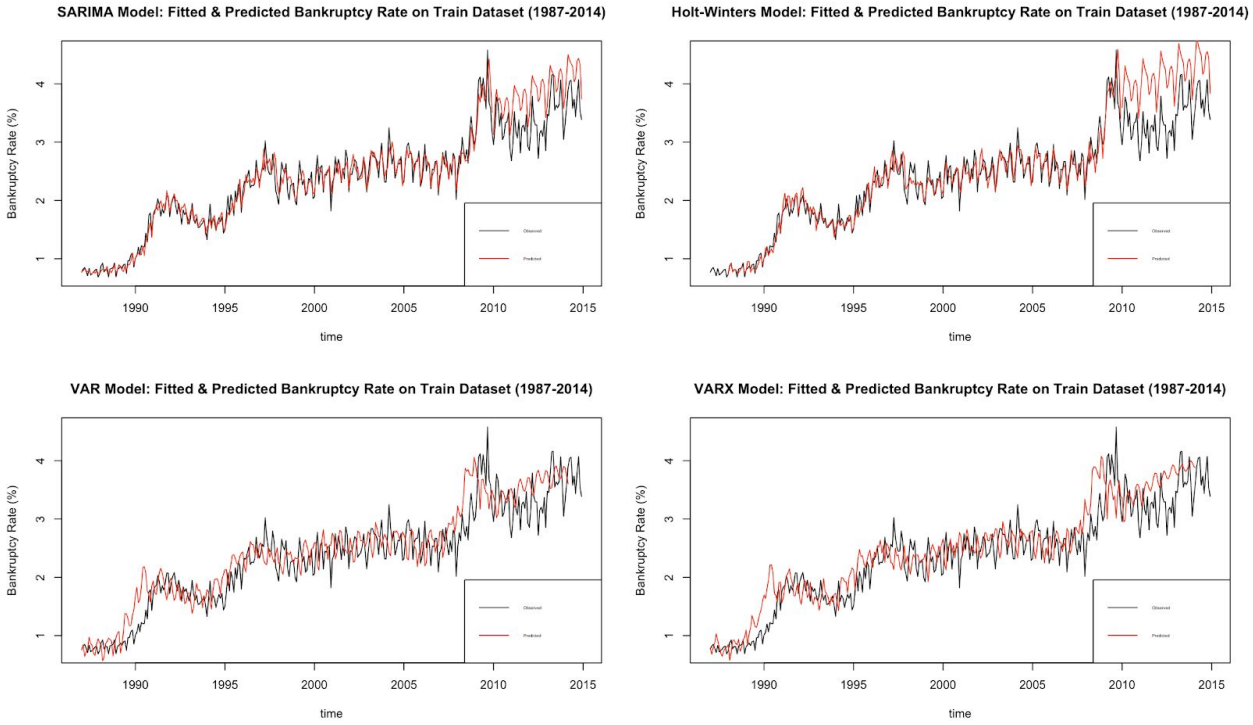

Figure 4. Fitted and Predicted Bankruptcy Rate on Train & Validation Dataset of Four Models

	Parameters	RMSE (Validation)	AIC
SARIMA	(3,1,5)(1,0,1)[12]	0.4859	**-613.88**

Holt-Winters	$\alpha = 0.3209$ $\beta = 0.1789$ $\gamma = 0.1732$ (Triple ES Additive)	0.7836	-
VAR	$p = 9$	**0.2814**	4.4820
VARX	$p = 10$ (Using population as exogenous variable)	0.3310	4.2795

Table 2. Comparison Between Four Different Models

In *Figure 2.*, we observed that the variables were correlated with others, which indicates there are possible relationships between bankruptcy rates, HPI, population size, and unemployment rates. VAR models are able to exploit the potential bidirectional relationship between those variables while also accounting for trend and seasonality. The original training data was split into a sub-training dataset and a cross-validation dataset. The sub-training dataset was composed of values from January 1987 to December 2009. The cross-validation dataset was composed of values from January 2010 to December 2014. The cross-validation dataset was used to compare the VAR models with different values of p, the number of lags, or backward shifts of the values in time. We created seventeen models but found that a lag number of nine outputted the lowest RMSE value (0.281) when we compared the predicted and actual results January 2010 to December 2014. This indicates that our predicted values for the bankruptcy rates from January 2010 to December 2014 were very similar to the actual bankruptcy rates observed.

However, there are limitations to use a VAR model. 1) Since it uses endogenous variables, models can become very complex with larger values of p in comparison to simpler models such as ARIMA/SARIMA. 2) The lag number was chosen based on a smaller subset of the training data, so the most optimal lag value could be different when the entire training data set was used in the final model. VAR models cannot be interpreted in a way where we can determine the causal relationships between variables.

6. Forecasting results

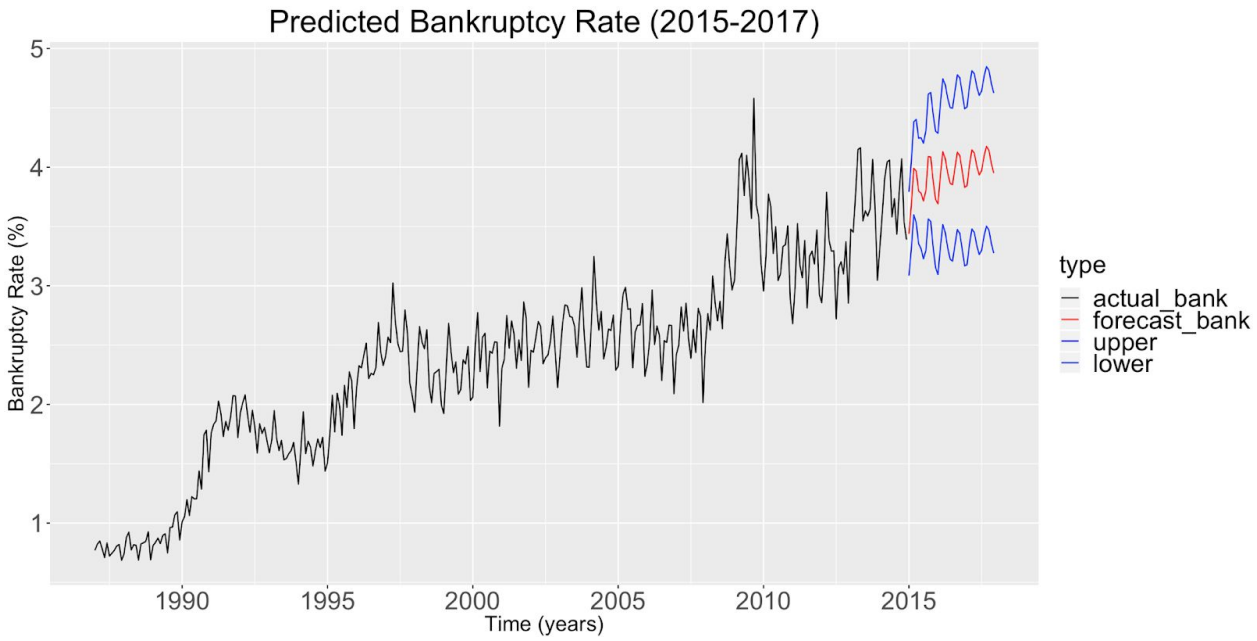

Figure 5. Forecast of Bankruptcy rates from 2015-2017

Month	Year	Forecast	Lower Bound	Upper Bound	Confidence Interval
January	2015	3.438	3.085	3.792	0.353
February	2015	3.682	3.314	4.051	0.369
March	2015	3.989	3.597	4.382	0.392
January	2016	3.691	3.096	4.286	0.595
February	2016	3.933	3.328	4.538	0.605
March	2016	4.13	3.515	4.744	0.614
January	2017	3.842	3.177	4.507	0.665
February	2017	4.018	3.352	4.684	0.666
March	2017	4.146	3.479	4.813	0.667

Table 4. Forecast, Confidence Interval Bounds, and Confidence Interval Range values

A complete table can be found in Appendix 7.1.

7. Appendix

7.1 Forecast, Confidence Interval Bounds, and Confidence Interval Range values

Month	Year	Forecast	Lower Bound	Upper Bound	Confidence Interval
January	2015	3.438	3.085	3.792	0.353
February	2015	3.682	3.314	4.051	0.369
March	2015	3.989	3.597	4.382	0.392
April	2015	3.967	3.533	4.402	0.435
May	2015	3.799	3.355	4.243	0.444
June	2015	3.78	3.312	4.247	0.467
July	2015	3.715	3.228	4.203	0.487
August	2015	3.802	3.297	4.307	0.505
September	2015	4.089	3.563	4.616	0.526
October	2015	4.086	3.542	4.629	0.544
November	2015	3.884	3.323	4.445	0.561
December	2015	3.728	3.153	4.303	0.575
January	2016	3.691	3.096	4.286	0.595
February	2016	3.933	3.328	4.538	0.605
March	2016	4.13	3.515	4.744	0.614
April	2016	4.072	3.449	4.695	0.623
May	2016	3.951	3.321	4.581	0.63
June	2016	3.864	3.225	4.502	0.639
July	2016	3.852	3.207	4.497	0.645
August	2016	3.993	3.344	4.641	0.649
September	2016	4.126	3.474	4.778	0.652
October	2016	4.097	3.441	4.753	0.656
November	2016	3.965	3.306	4.624	0.659
December	2016	3.831	3.169	4.493	0.662
January	2017	3.842	3.177	4.507	0.665
February	2017	4.018	3.352	4.684	0.666
March	2017	4.146	3.479	4.813	0.667
April	2017	4.118	3.449	4.787	0.669

May	2017	4.01	3.341	4.68	0.669
June	2017	3.934	3.264	4.605	0.671
July	2017	3.969	3.297	4.641	0.672
August	2017	4.091	3.419	4.763	0.672
September	2017	4.175	3.502	4.847	0.673
October	2017	4.143	3.47	4.817	0.673
November	2017	4.034	3.36	4.708	0.674
December	2017	3.95	3.275	4.625	0.675