MSDS604 Final Report

Xu Lian, Zhe Yuan, and Hongdou Li

December 10, 2018

Abstract

This report aims to forecast monthly bankruptcy rates for Canada from January 2015 to December 2017. This report explores the trend and seasonality of the historical national bankruptcy rates from January 1987 to December 2014. Then, this report splits the data into a training set and testing set for the purposes of constructing models with numerous methods and implementing cross-validations. Finally, this report selects the optimal model, $SARIMAX(3,1,3)_{\times}(1,1,1)_{12}$ with information on unemployment rate, national population, and Housing Price Index, and utilizes the model to forecast 2015-17 Canada monthly bankruptcy rate.

1 Introduction

Accurately forecasting national bankruptcy rates is of interest to national banks, insurance companies, credit-lenders, politicians etc. A well-constructed model regarding this topic would provide helpful guidance in monetizing capabilities, risk mitigation or political-agenda design, ultimately bringing value to society. With this objective in mind, this report aims to provide a serviceable time series model to forecast the 2015-17 Canada monthly bankruptcy rate.

In this report, there are two types of data. The primary one is the historical national bankruptcy rates of Canada from January 1987 to December 2014. External variables, such as unemployment rate, national population, and Housing Price Index (HPI)[1], serve as the secondary complement to the bankruptcy rate when constructing more complex models.

2 Methodology

As for the process of modeling, this report features various time series modeling techniques (Fig. 1). A time series analysis characterizes the nature of the relationship among values of a given target variable at different time periods. By implementing such modeling techniques, trend and seasonality components are captured to account for the up and downs in the national bankruptcy rates of Canada.

In general, two primary types of approaches in modeling are featured in this report. Univariate approaches, like Seasonal ARIMA (SARIMA) and Triple Exponential Smoothing

[1] HPI measures the price changes of residential housing, calculated monthly by Statistics Canada.

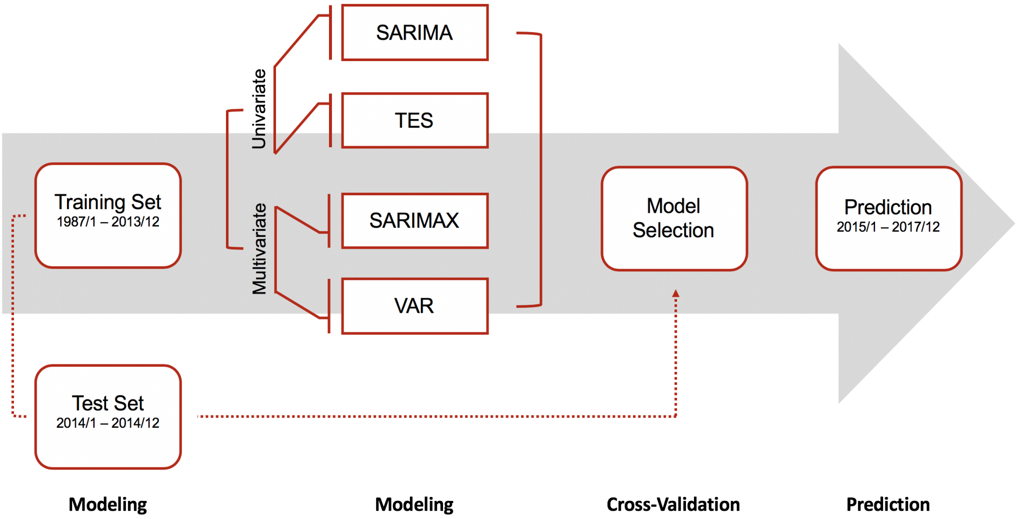

Figure 1: Time Series Modeling Flow Chart

(TES), focus on projecting the movement pattern of the bankruptcy rates in the future based on the time series of the very same variable alone.

On the other hand, the multivariate approach is utilized to investigate the relationship between bankruptcy rates and other external factors for modeling. There are two types sub-approach under this approach. A SARIMAX model is implemented under the belief that external factors influence the bankruptcy rates but the latter does not influence the former. A VAR model is based on the notion that the relationships among the bankruptcy rates and external factors are not uni-directional.

For cross-validation, multiple models are utilized to made prediction during the span of test set, which starts from January 2014 to December 2014. Our optimal model is selected based on the proximity between the predicted national bankruptcy rates generated by the models and actual values.

Eventually, our predictions of national bankruptcy rates of Canada are obtained via our optimal models.

2.1 Overview of the Time Series

The initial step of the model construction process is to visualize how Canadas national bankruptcy rates have altered over the time span from January 1987 to December 2013. This procedure helps to recognize the existence of trend and seasonality in the time series of bankruptcy rates, serving as the filter to narrow down the range of modeling approaches utilized in later steps.

As shown in Fig. 2(a), it is clear that there is a consistent directional movement in the time series of the bankruptcy rates. In addition, there are cyclical trends along the overall direction, indicating the existence of seasonality. The trend and seasonality components are also confirmed in the ACF plot on the right. For the first 48 lags, the bars over the dotted

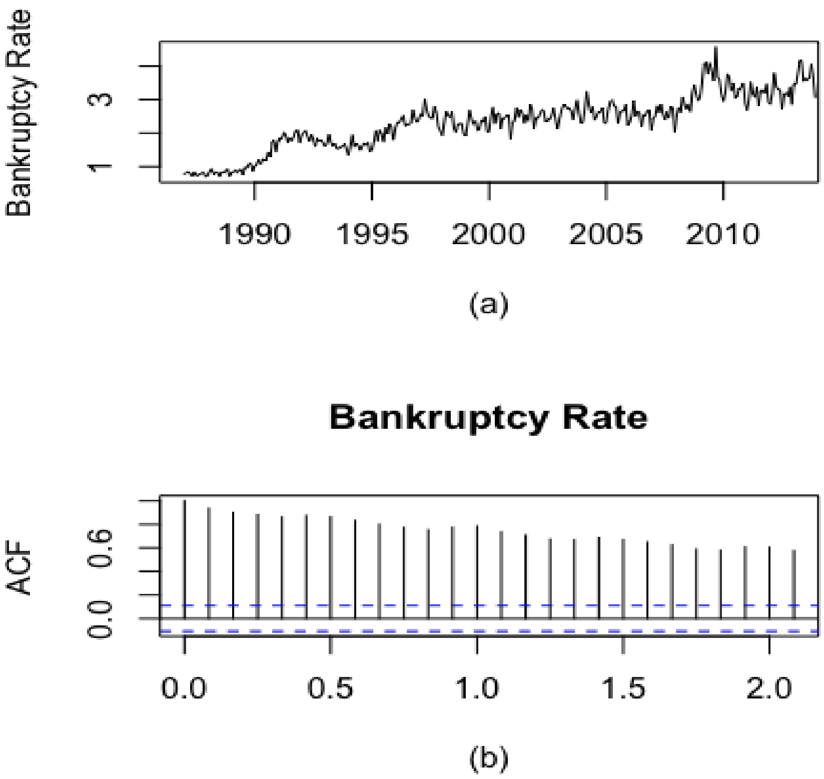

Figure 2: Bankruptcy Rate Trend and ACF Plot

lines illustrate the significance of the trend, while relatively high spikes at every 12 lags imply the strong seasonality. Hence, modeling techniques with capabilities to capture both the trend and seasonality components of the movement in bankruptcy rates are chosen for implementation in this study.

2.2 Univariate Model

Univariates models were built to forecast future values of bankruptcy rates solely based on the knowledge of past values of y. Under this scope, Seasonal Autoregressive Integrated Moving Average (SARIMA) and Triple Exponential Smoothing (TES) are implemented.

2.2.1 Integrated Seasonal Autoregressive Moving Average (SARIMA) model

The first model we tried is a Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The Autoregressive parts of the SARIMA model indicate that the value variable is regressed on its own previous values. The Moving Average parts indicate that the regression error is a linear combination of error terms in the past. Meanwhile, the integrated part corresponds to the differencing steps. In addition, the model also takes into accounts of the seasonality effects.

A $SARIMA(3,1,3)_{\times}(1,1,1)_{12}$ model was established based on the information we get from Autocorrelation Function(ACF) and Partial Autocorrelation Function(PACF) plots. In addition, to ensure the models adequacy, a series of diagnostic tests are used to examine whether a model satisfies certain assumption. For example, a SARIMA model needs to meet certain assumptions regarding constant variance, zero mean, and uncorrelatedness before generating meaningful predictions.

2.2.2 Triple Exponential Smoothing model

Another univariate model tried is a Triple Exponential Smoothing (TES) model Because both treand and seasonality in bankruptcy rate data. A TES model is composed with four equations. A Level Equation with a weighted average of seasonally adjusted observation and a non-seasonal forecast on it. A Trend Equation with a weighted average of current trend component and historical prediction of that same component. A Seasonal Equation with a weighted average between the current seasonal index and the seasonal index of the same period in the previous season and a forecast function combining all three equations.

2.3 Multivariate Model

From a socio-economic perspective, bankruptcy rates are more likely related to external factors than not. In this study, factors such as HPI, population, and unemployment rates are selected for model optimization. As shown in Fig. 3, it is reasonable to speculate that correlations between bankruptcy rates and the other three variables exist.

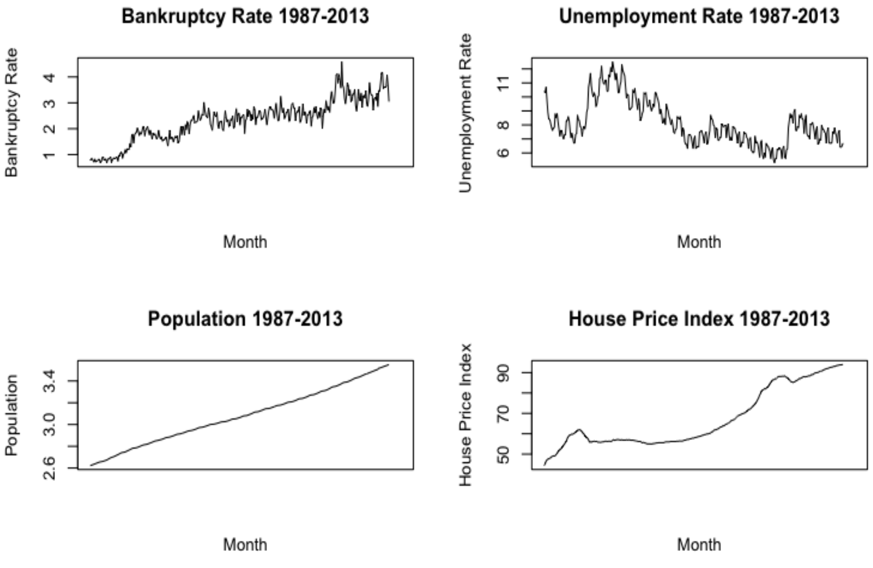

Figure 3: Plot of all variables

For the construction of multivariate models, two sub-approaches are implemented. A SARIMAX model is implemented under the belief that external factors influence the bankruptcy rates but the latter does not influence the former. A Vector Autoregression (VAR) model is based on the notion that the relationships among the bankruptcy rates and external factors are not uni-directional.

2.3.1 SARIMAX

As previously discussed in Fig. 3, correlation between bankruptcy and the other three variables exists to some extent. Such a notion is verified by calculating the correlation matrix, which shows significant correlation between bankruptcy, population and house price index.

A SARIMAX model is built under the assumption that all these three factors are exogenous, which means they influence the response and not the other way around. To find the optimal SARIMA model, except for RMSE, other good-of-fit metrics are also used to serve as criteria to determine the optimal model. Table X shows an example of such a process, with the best SARIMAX model obtaining the lowest RSME score and highest log-likelihood.

Table 1: SARIMAX RMSE comparison table

Features	RMSE	Log likelihood
Population + HPI + Unemployment	0.1420	143.69
Unemployment	0.1426	122.92
Population + Unemployment	0.1442	133.08
HPI + Unemployment	0.1457	127.65
HPI	0.1481	124.81
Population	0.1625	102.79
Population + HPI	0.1792	118.74

2.3.2 VAR

Owing to the nature that the social economy is a large and complex system, the interaction between various factors could be complicated as well. Below are CCF plots which show cross correlation between bankruptcy and the other three factors.

The results of CCF (Fig. 4)show that both the cross correlation between population and bank and the cross correlation between house and bank have a peak around 0, which indicating that there is no obvious cross correlation between these two pairs. However, the cross correlation plot between unemployment and bank has a peak around lag 4, which means the unemployment lags the bankruptcy for a period of 4 months.

Hence, it is fair to speculate that these variables are endogenous, which means they influence the response and the response influences them simultaneously. Therefore, Vector Autoregression (VAR) model is introduced to account for this kind of relationship. Iteration in order p was implemented to check the model performance.

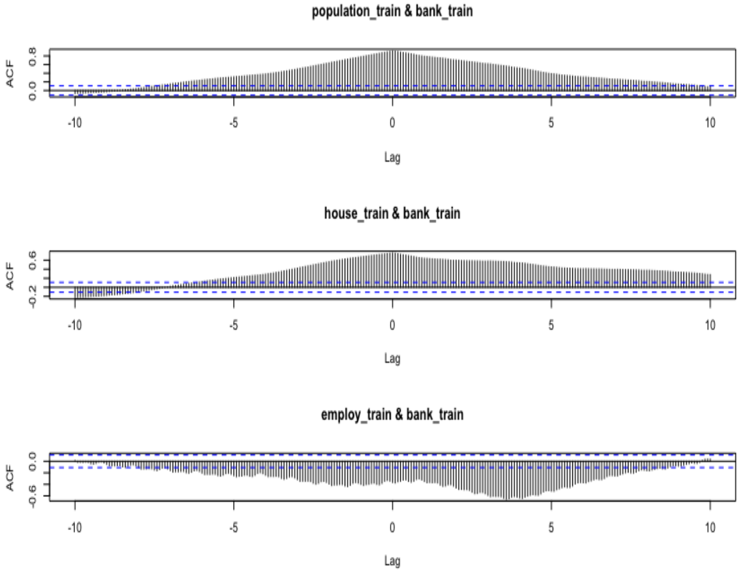

Figure 4: Cross Correlation Plot

Table 2: SARIMAX RMSE comparison table

Order	RMSE
p=1	0.2680
p=2	0.2515
p=3	0.1734
p=4	0.2640

2.4 Model Selection

In the context of quantitative studies, the definition of accuracy is determined by how close the error between predicted values and actual values are. In this study, RMSE is used to determine which model is optimal.

Table 3: RMSE comparison table

Models	RMSE
SARIMA(3,1,3)(1,1,1)[12]	0.1805
TES	0.1826
SARIMAX(3,1,3)(1,1,1)[12]	0.1420
VAR(p=3)	0.1733

As shown in Table 2, the $SARIMAX(3,1,3)_{\times}(1,1,1)_{12}$ model has the lowest RMSE value, with HPI, population, and unemployment rates as external variables. This particular model is generated by least square estimation, satisfying all the assumptions related to zero mean, constant variance, and uncorrelation.

3 Prediction

The bankruptcy rates of the years 2015 to 2017 is predicted by the selected SARIMAX(3,1,3)(1,1,1)[12] model and the given data. The Fig. 5 below shows the Canadian monthly bankruptcy rate has a consistently stable trend in the future 3 years. The prediction also contains 95% prediction interval, which means there is a 95% probability that the bankruptcy rate is believed to lie in.

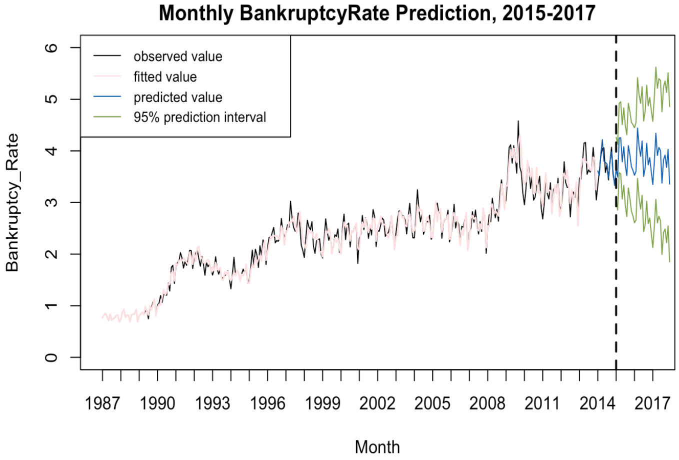

Figure 5: Prediction

References

[1] N. Stevens *Time Series Analysis: Course notes.*

[2] P.J. Brockwell, R.A. DAVIS *Introduction to Time Series and Forecasting*, (Springer Texts in Statistics.).

[3] R. Hyndman *Forecasting: Principles and Practice*, (OTexts).