MSDS604 Final Project
Forecasting Canadian Bankruptcy Rate

Shulun Chen
Bowen Ma
Jinghui Zhao
Wenkun Xiao

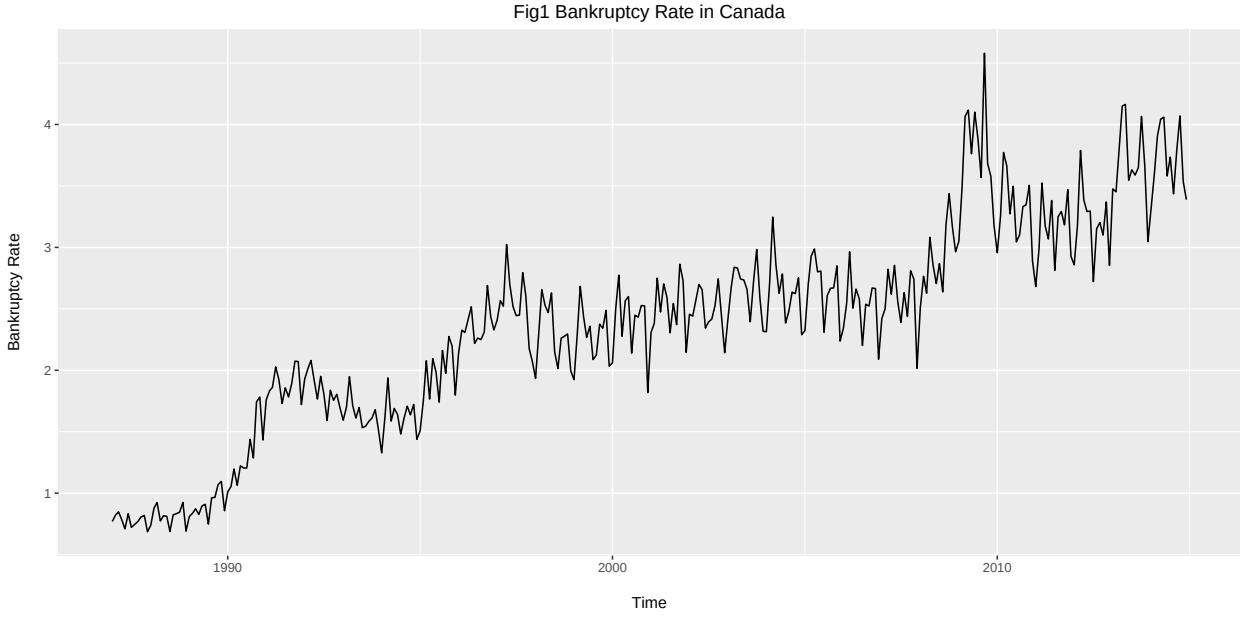
University of San Francisco

San Francisco, California

December 11, 2018

Project Description

This project aims to build a time series model to forecast monthly bankruptcy rates in Canada for the period from January 2015 to December 2017. The dataset we worked on contains historical bankruptcy rates, unemployment rates, population, and house price index from January 1987 to December 2014. In this report, we explored different time series modeling approaches, such as `Holt-Winters` , `SARIMA` , `SARIMAX` and `VAR` , to find the optimal model with the best predictive accuracy which is measured by the RMSE (Root Mean Squared Error). The historical monthly bankruptcy rates are shown below:



Modeling Methods

Numerous approaches are available for forecasting bankruptcy rates. Depending on the number of variables used for modeling, there are two main categories: univariate modeling and multivariate modeling.

A univariate modeling approach, as indicated by its name, considers only the historical data of the variable being modeled and takes no external information into consideration. In our case, we used only historical bankruptcy rates data to train our univariate models, such as `Holt-Winters` and `SARIMA` (under the `Box-Jenkins` framework). `Holt-Winters` is the simplest approach since it does not rely on any statistical assumptions. It makes predictions by performing a smoothing on the historical observations in the time series. `SARIMA` is the most common type of modeling under the `Box-Jenkins` framework and will be discussed in more details in the next section.

On the other hand, a multivariate modeling approach considers external data. There are two common types of multivariate models: `SARIMAX`, and Vector Autoregression (`VAR`). If the external information is treated as exogenous, meaning the external variables have a uni-directional influence on the response, then a `SARIMAX` should be employed. For instance, to predict the corn production, rainfall may be considered as an exogenous variable since it has an influence on the corn production. However, corn production does not influence rainfall. On the contrary, if the external variables are treated as endogenous, meaning the external variables and response have mutual influence, then a Vector Autoregression model should be employed. In our case, with bankruptcy rates as the response variable, we considered unemployment rates, house price index, and population for multivariate modeling.

After exploring all of the above models, $\texttt{SARIMAX(2, 1, 5)} \times \texttt{(3, 0, 2)}_{12}$ was found to be the optimal model.

Justification of Modeling Approach

Data Preprocessing

(1) Train-Validation Data Split

In order to find the best model with the lowest RMSE (root mean squared error), a train-validation data split was carried out on the original `train.csv` dataset. The validation dataset was used to measure and rank the model performance based on the RMSE values. We determined the split to be at the end of year 2013 - thus the training set contains 324 data points, and the validation set contains 12 data points.

(2) Box-Cox Transformation

Box-Cox tranformation is useful in adjusting the non-constant variation in data. The bankruptcy data has shown certain degree of inflated variance over time which was mitigated by the Box-Cox transformation.

Modeling

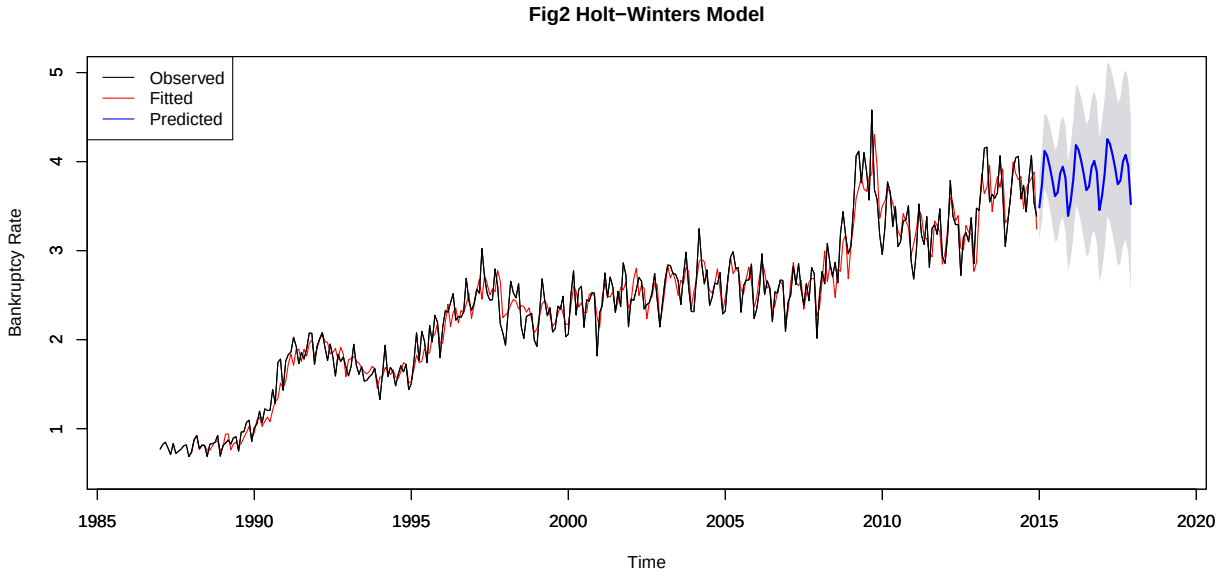
(1) Holt-Winters Model

`Holt-Winters` method, also known as exponential smoothing, is a modeling method with no distributional assumptions.

Exponential smoothing uses exponential equations to model level, trend, and seasonality components of the time series. Since our data exhibits both trend and seasonal components, we employed triple exponential smoothing to model and forecast the data. The seasonal component could be modeled either additively or multiplicatively depending on the variance. Since the variance in data appears to be non-constant over time, we selected multiplicative seasonality for our model.

The parameters α, β, and γ, which represent the model's sensitivity to level, trend, and seasonality, were found to be $\alpha = 0.3802$, $\beta = 0.0013$, $\gamma = 0.1906$ for the optimal `Holt-Winters` model.

The fitted and forecast results of the `Holt-Winters` model are shown below:



(2) SARIMA Model

Seasonal Autoregressive Integrated Moving Average, or `SARIMA` is widely used for modeling univariate non-stationary time series that exhibits seasonality and/or trend. The idea of `SARIMA` modeling is to transform the non-stationary time series into a new stationary time series, such that the stationary time series can be modeled by an `ARMA` model (the basic time series model). This transformation can be achieved by taking finite times of seasonal differencing and trend (ordinary) differencing.

`SARIMA` model is composed of a within-season time series, which can be modeled by `ARMA(p, q)`, and a between-season time series, which can be modeled by `ARMA(P, Q)`.

Since the data exhibits both trend and seasonality, we considered both components. For the trend component, p represents trend autoregression order, d represents ordinary (trend) difference order, and q represents trend moving average order. For the seasonal component, P represents seasonal autoregressive order, D represents seasonal difference order, Q represents seasonal moving average order, and m represents the number of time steps for a single seasonal period. These are all captured in the `SARIMA (p,d,q)`$\times$`(P, D, Q)`$_m$ model.

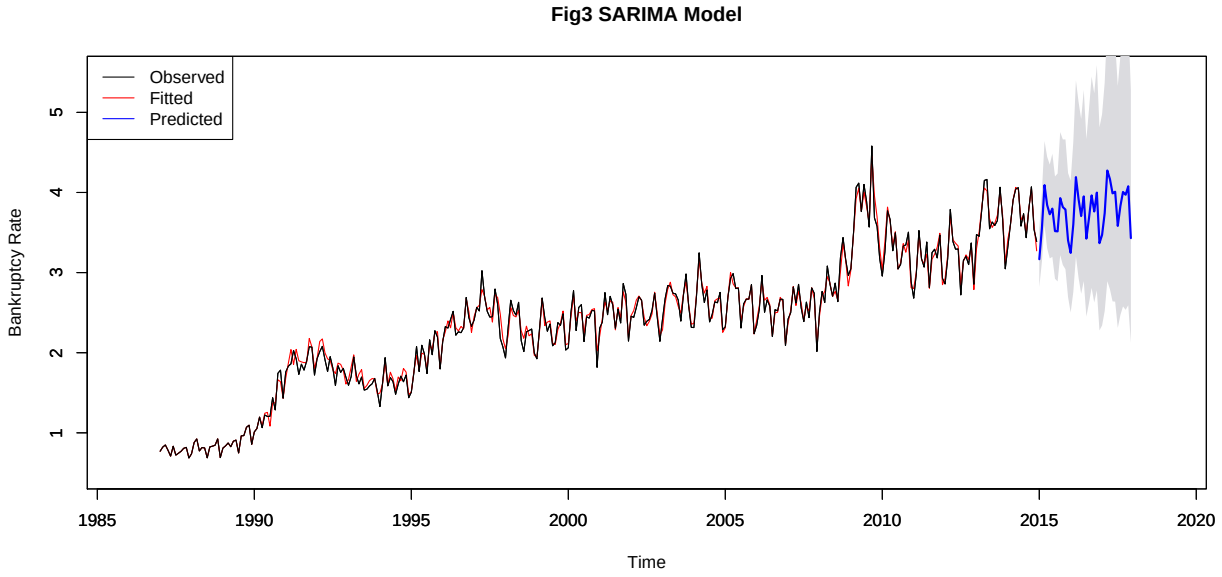
The Box-Jenkins Framework for SARIMA modeling process was followed:

- Applied Box-Cox transformation to fix the non-constant variance;
- Took ordinary difference once to eliminate trend (we have confirmed the differencing orders by applying R functions `ndiffs` and `nsdiffs`);
- Identified that $p \leq 2$, $q \leq 5$, $P \leq 3$, $Q \leq 8$ from ACF and PACF plots of the differenced time series;
- Fitted proposed models with orders within above range and iterated to find optimum;
- Checked residual plots to ensure that the model assumptions were valid;
- Forecast into the future.

We searched all combinations of the models within the range and obtained the top 10 models with the least RMSE. Since prediction accuracy was used as the metric, we picked the model `SARIMA(2,1,5)`$\times$`(3,0,2)`$_{12}$ with the lowest RMSE from the table below. In addition, the assumptions of zero-mean, constant variance, and uncorrelatedness were all satisfied.

	p	q	P	Q	loglik	sigma2	rmse
300	2	5	3	2	362.3195	0.0056923	0.1531865
310	2	5	3	3	363.6797	0.0056627	0.1561818
191	1	1	2	1	317.9173	0.0076629	0.1562966
272	1	2	3	0	303.5353	0.0080647	0.1564372
130	2	5	1	3	373.5860	0.0057373	0.1566100
110	2	5	1	1	367.3456	0.0059254	0.1568215
275	1	5	3	0	321.2991	0.0072937	0.1572141
210	2	5	2	2	371.5211	0.0055855	0.1579558
230	2	5	2	4	372.4029	0.0055909	0.1582082
220	2	5	2	3	371.5261	0.0056032	0.1583980

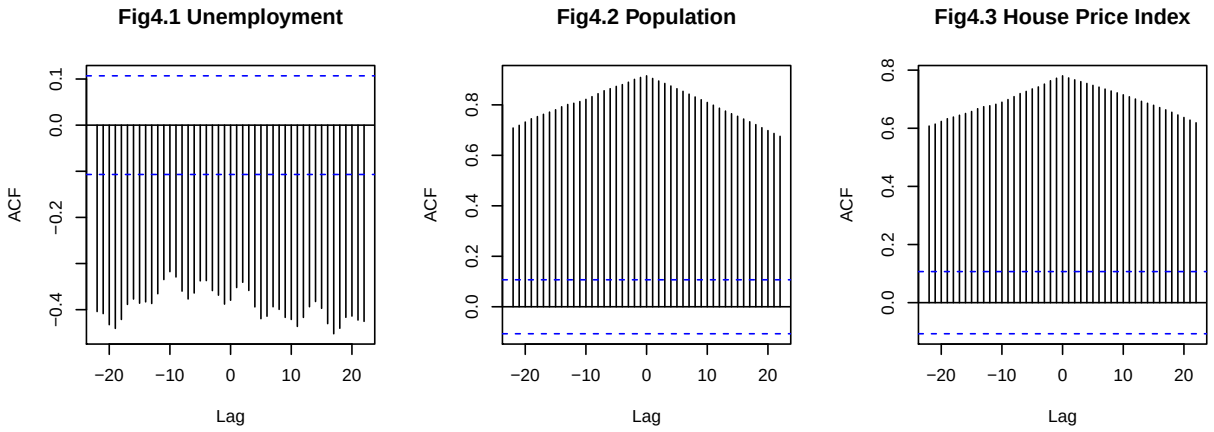
The fitted and forecast results of the `SARIMA` model are shown below:



(3) SARIMAX Model

A `SARIMAX` model is a `SARIMA` model with explanatory variables. `SARIMAX` model is a popular method for modeling multivariate time series. We considered multivariate time series when there exists other variables that are highly correlated with the response variable.

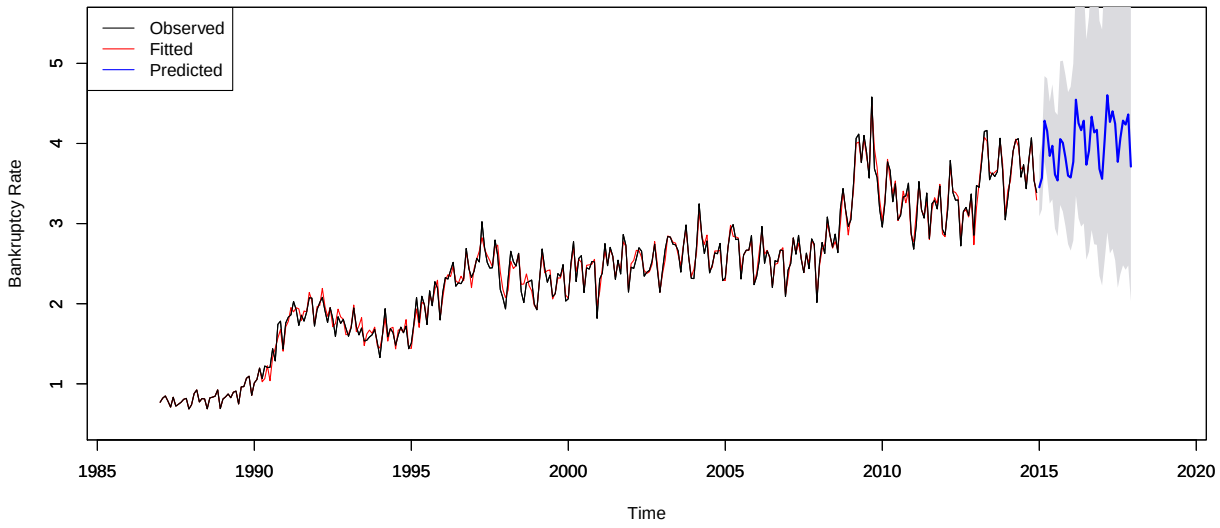
In our case, bankruptcy rates were highly correlated with population and house price index and was negatively correlated with unemployment rates (according to CCF plot below). By considering a `SARIMAX` model, we hoped to improve the predictive accuracy based on the previously built `SARIMA` model.



We tried all combinations of SARIMAX model and found that the best SARIMAX model contains explanatory variable unemployment rate. This model gives the lowest RMSE (0.1523) so far, which is lower than the RMSE we obtained from the `SARIMA` model. In addition, the assumptions of zero-mean, constant variance, and uncorrelatedness are all satisfied.

The fitted and forecast results of the `SARIMAX` model are shown below:

Fig5 SARIMAX Model



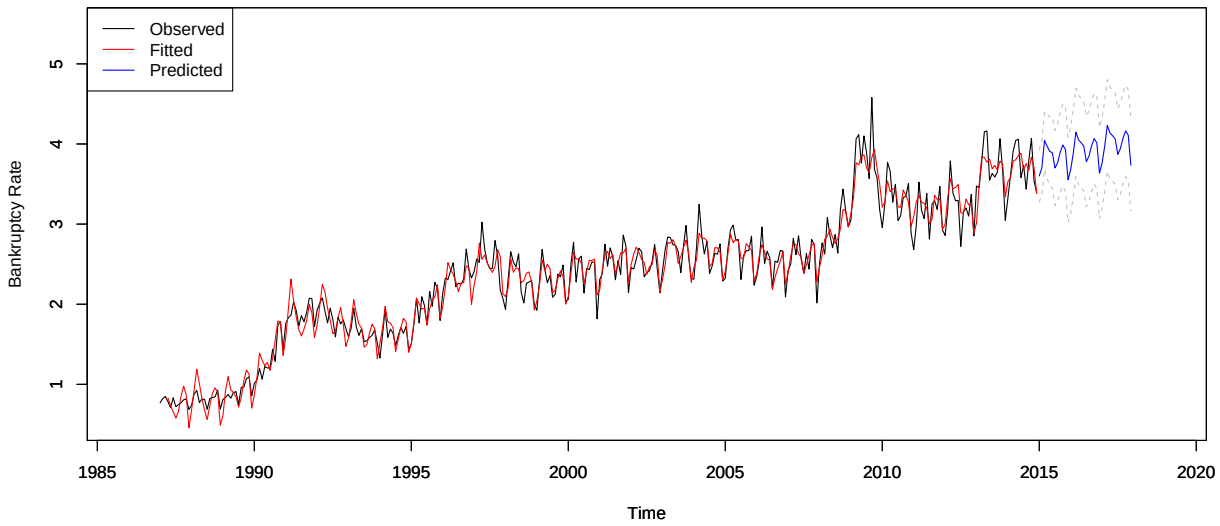
(4) VAR Model

Vector Autoregression (`VAR(p)`) model, an extension of the univariate autoregression model to multivariate time series data, is a system of equations whose variables are treated as endogeneous. The model consists of r equations, one for each variable, that are each autoregressions of order p.

We chose this model to account for the relationships between explanatory variables. For example, house price index and population might be influencing each other, since a larger population boosts the house price, on the other hand, a higher house price reduces the population in an area. Their mutual influence could be fully accounted for by the `VAR` model.

In our case, p was chosen to be 3 based on the RMSE values. The fitted and forecast results of the `VAR` model are shown below:

Fig6 VAR Model



Forecasting Results

We chose the final model to be SARIMAX(2, 1, 5)×(3, 0, 2)$_{12}$ and it was plotted in Fig5. It has the lowest RMSE value of 0.1523 on our test set. The model assumptions, including zero-mean, constant variance, and uncorrelatedness, are all satisfied. The best models out of all categories are shown in the table below:

Models	RMSE
SARIMAX(2, 1, 5)×(3, 0, 2)$_{12}$	0.152270
SARIMA(2, 1, 5)×(3, 0, 2)$_{12}$	0.153187
VAR(3)	0.187653
Holt-Winters	0.190232

The forecasting results for 2015-2017 are shown in the table below:

	Prediction	Lower Bound	Upper Bound
Jan 2015	3.453249	3.090858	3.847102
Feb 2015	3.569987	3.176269	3.999788
Mar 2015	4.283855	3.776529	4.841449
Apr 2015	4.157311	3.573641	4.810839
May 2015	3.845208	3.249869	4.520116
Jun 2015	3.971005	3.302359	4.738009
Jul 2015	3.611288	2.937339	4.396920
Aug 2015	3.538384	2.843188	4.355949
Sep 2015	4.054666	3.232982	5.026379
Oct 2015	4.008055	3.150162	5.033139
Nov 2015	3.824395	2.969242	4.855208
Dec 2015	3.598352	2.744653	4.640519
Jan 2016	3.575325	2.663007	4.708025
Feb 2016	3.772009	2.788271	4.999955
Mar 2016	4.546585	3.349665	6.044154
Apr 2016	4.252659	3.067787	5.757326
May 2016	4.166086	2.963274	5.708709
Jun 2016	4.284242	3.002624	5.944995
Jul 2016	3.734843	2.549566	5.299129
Aug 2016	3.908336	2.648022	5.580331
Sep 2016	4.334300	2.919869	6.218191
Oct 2016	4.138183	2.742452	6.018279
Nov 2016	4.170400	2.735171	6.117582
Dec 2016	3.682833	2.347533	5.530085
Jan 2017	3.558725	2.228445	5.421406
Feb 2017	4.051634	2.544867	6.157065
Mar 2017	4.602563	2.891133	6.993847
Apr 2017	4.269070	2.627944	6.593743
May 2017	4.401777	2.686643	6.845285
Jun 2017	4.245172	2.542890	6.701153
Jul 2017	3.770544	2.199656	6.077357
Aug 2017	4.068884	2.368994	6.568435
Sep 2017	4.285359	2.480297	6.950033
Oct 2017	4.236018	2.424004	6.931233
Nov 2017	4.362681	2.477143	7.181740
Dec 2017	3.711829	2.032265	6.283839