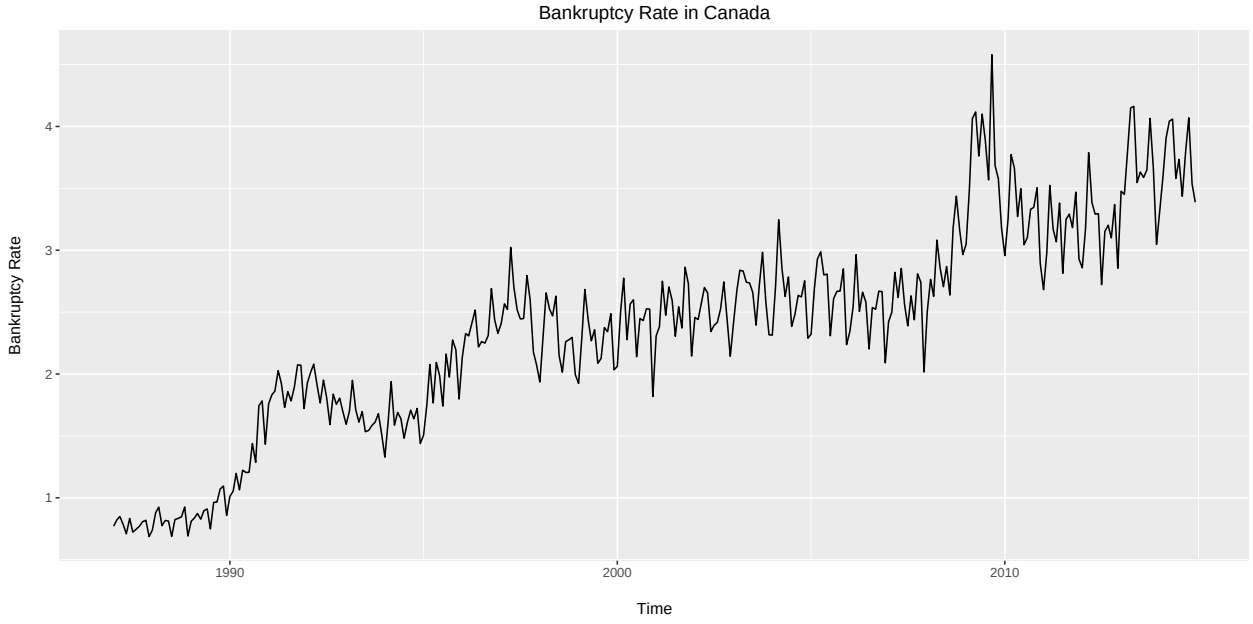
MSDS604 Final Project

December 8rd, 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 will work on contains historical bankruptcy rates, unemployment rates, population, and housing price index from January 1987 to December 2014. In this report, we will explore 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). Please see graph below for the historical monthly bankruptcy rates:



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 does not take any external information into consideration. In our case, we will only use only historical bankruptcy rates data to train our univariate models, such as `Holt-Winters` and `SARIMA` (under `Box-Jenkins` framework). `Holt-Winters` is the simplest approach because it does not rely on any statistical assumptions. Under the `Holt-Winters` approach, future observations are predicted by performing a smoothing on the previous observations in the time series. `SARIMA` is the most common type of modeling under `Box-Jenkins` framework. We will discuss in more detail 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 one treats the external information as exogenous, meaning the external variables have a uni-directional influence on the response, then a `SARIMAX` should be employed. For instance, to predict corn production, one may consider rainfall as an exogenous variable. Rainfall may have an influence on corn production, however, corn production will not influence rainfall. On the contrary, if one treats the external variables as endogenous, meaning the external variable and response have mutual influence, then a Vector Autoregression model should be employed. In our case,

with bankruptcy rates as the response variable, we have considered unemployment rates, housing price index, and population for multivariate modeling.

After exploring all of the above models, `SARIMAX` 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 its RMSE value. 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 on the entire dataset.

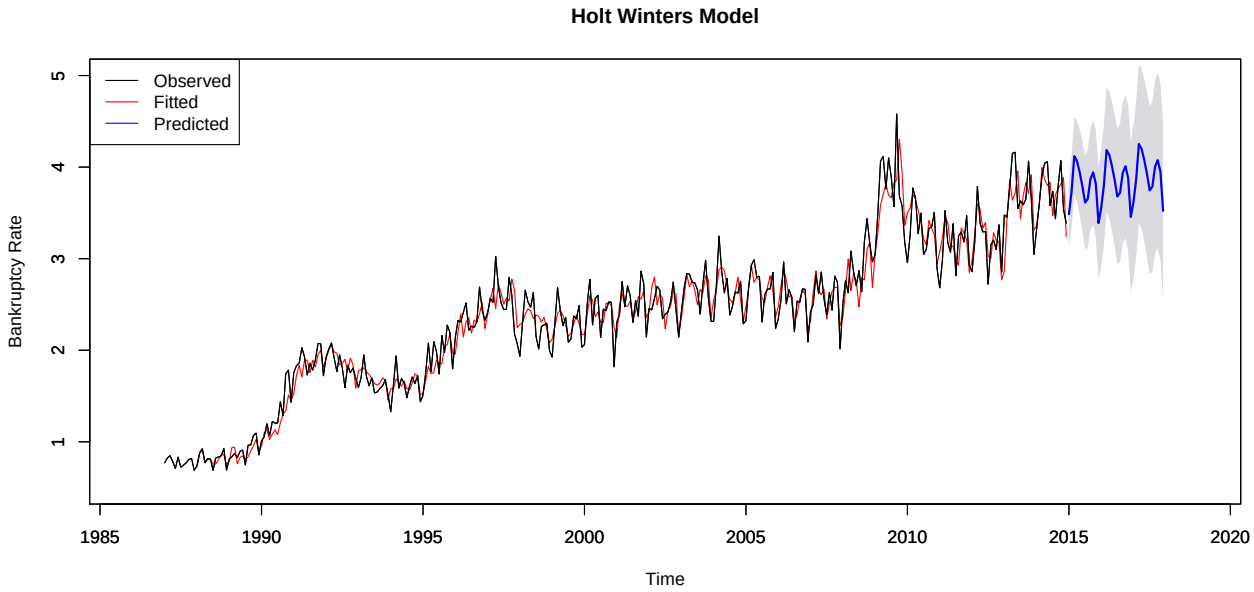
Data Modeling

(1) Holt-Winters Model (Exponential Smoothing)

Triple Exponential Smoothing, also known as the `Holt-Winters` method, is one of the many methods that can be used to forecast data points in a time series, provided that the series is "seasonal", i.e. repetitive over some period. The objective is to predict $y_n + h$ given the observed history $\{ y_1, y_2, \ldots, y_n\}$ of the time series.

Exponential smoothing is a method of time series modeling by which we model its level, trend, and seasonality components by exponential equations. Since our data exhibits both trend and seasonal components, we employed triple exponential smoothing to model and forecast the data. We modeled the seasonal component either additively or multiplicatively, depending on how the variation in our data changes over time. Since the variation does appear to inflate over time (in a non-linear fashion), we selected multiplicative seasonality for modeling.

The parameters α, β, and γ, which represent the model's sensitivity to its level, trend, and seasonality, were found using a grid search that gave the lowest RMSE value.



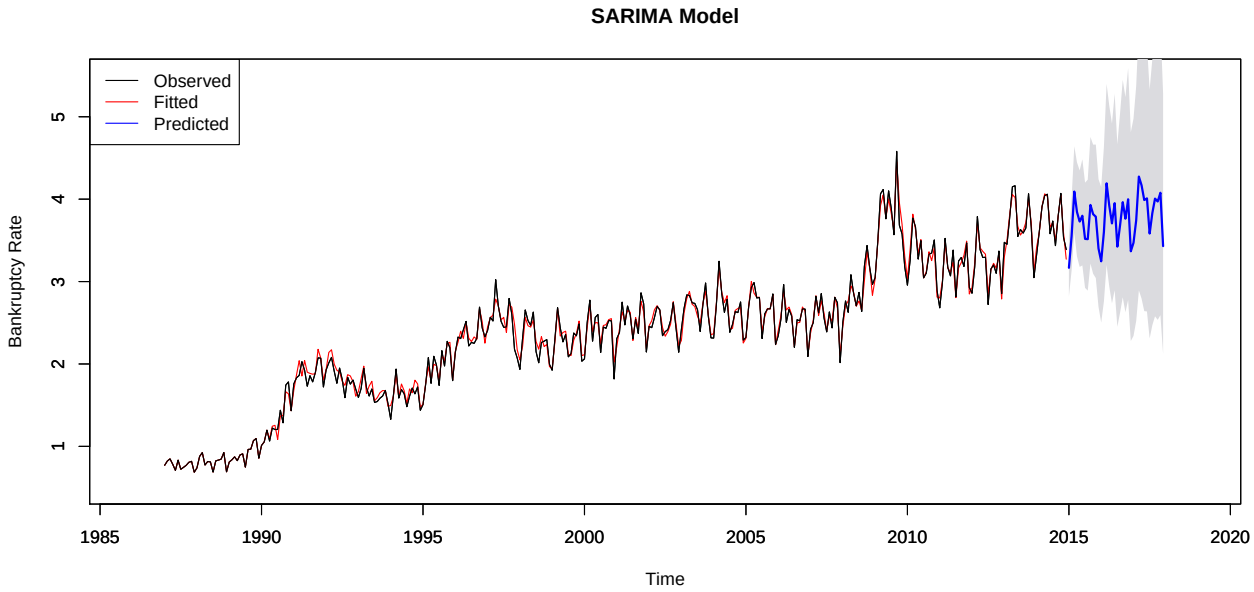
(2) SARIMA Model

Seasonal Autoregressive Integrated Moving Average, or `SARIMA` is widely used for modeling univariate time series with seasonality with or without trend. One can think of `SARIMA` model from two dimensions, first a within-season time series which can be modeled by `ARIMA(p, q)`, then between-season time series.

Since the data exhibits both trend and seasonality, we considered both trend and seasonality components. With regard to the trend component: p represents trend autoregression order, d represents ordinary (trend) difference order, and q represents trend moving average order. With regard to the seasonal component: P represents seasonal autoregressive order, D represents seasonal difference order, Q represents seasonal moving average order, 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.

Now the question is how to determine the orders for a SARIMA model. First we used `ndiffs` and `nsdiffs` to determine the ordinary differencing and seasonal differencing order. Then we found the maximum orders, p, q, P and Q. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of the transformed data were examined and iterated to find out the optimal model. `auto.arima`, a less computationally expensive method, was also employed to optimize the model. Then we compared other models with this the model produced by `auto.arima` and chose the best model in the end.

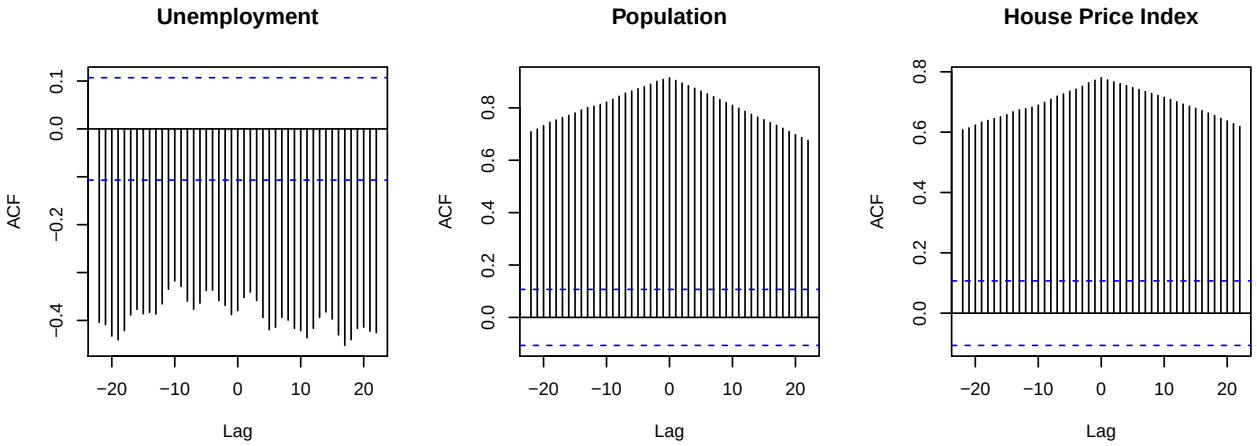
	p	q	P	Q	loglik	sigma2	rmse
300	2	5	3	2	362.3195	0.0056923	0.1531865
310	2	5	3	3	363.6797	0.0056627	0.1561818
191	1	1	2	1	317.9173	0.0076629	0.1562966
272	1	2	3	0	303.5353	0.0080647	0.1564372
130	2	5	1	3	373.5860	0.0057373	0.1566100
110	2	5	1	1	367.3456	0.0059254	0.1568215
275	1	5	3	0	321.2991	0.0072937	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

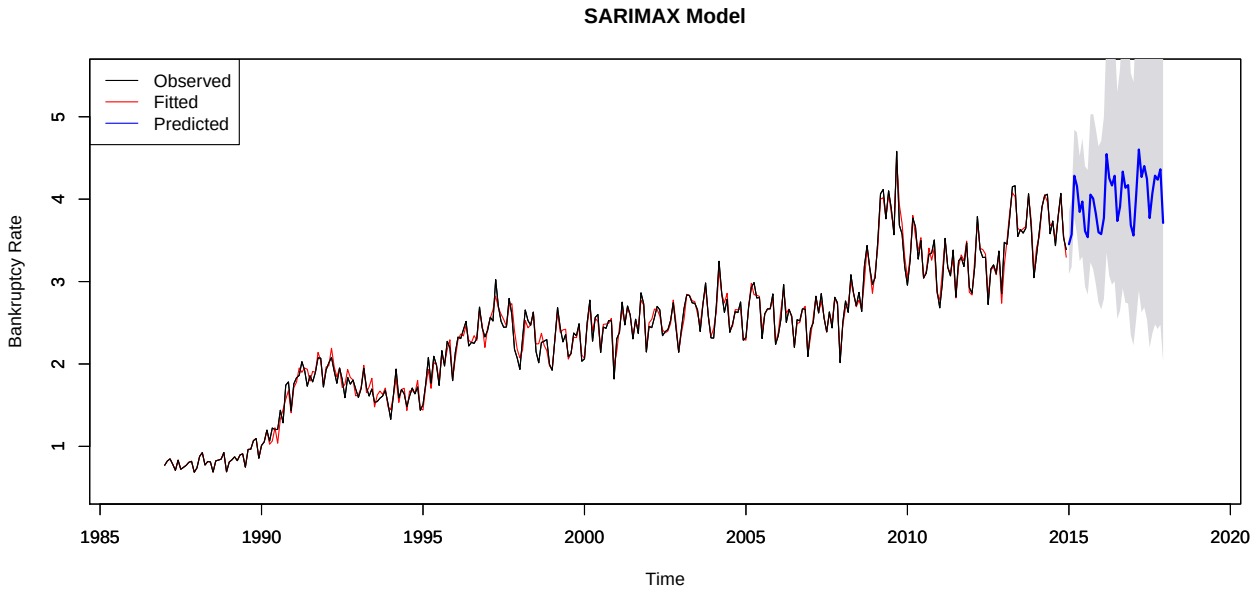


(3) SARIMAX Model

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 addition, the data on other variables have to be collected at the same frequency and for the same duration as our response variable.

In our case, bankruptcy rates are highly correlated with population and housing price index (according to CCF plot) and is negatively correlated with unemployment rates. By considering a `SARIMAX` model, we hoped that such a model would provide more accurate forecasts than other univariate models.



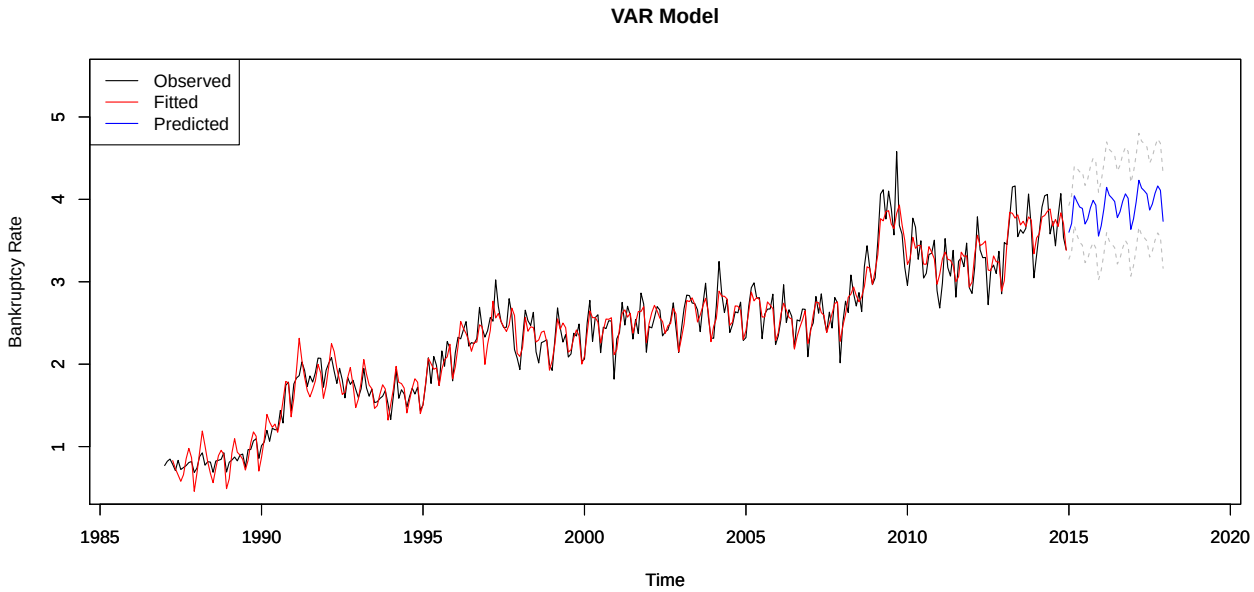


(4) VAR Model

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 have chosen this model to account for the relationships between predictor variables. For example, house price index and population might be influencing each other, since a larger population will boost the housing price, on the other hand, a high housing price will reduce the population in an area. Their endogeneous relationship was accounted for by the `VAR` model.

In our case, p was chosen to be 3 based on the RMSE values.



Forecasting Results

We have chosen the final model to be SARIMAX(2, 1, 5)×(3, 0, 2)$_{12}$. It has the lowest RMSE value of 0.1523 on our test set. The best models out of all categories are shown in the table below:

Models	RMSE
SARIMAX(2, 1, 5)×(3, 0, 2)$_{12}$	0.1522703
SARIMA(2, 1, 5)×(3, 0, 2)$_{12}$	0.1531865
VAR(1)	0.1876525
Holt-Winters	0.1902321

	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