I. Introduction

The goal of this project is to use one or more variables to accurately predict monthly Canadian bankruptcy rates from January 2011 to December 2012. Data available for modeling includes the monthly data from January 1987 to December 2010 for Canada for the following four variables: *Unemployment Rate*, *Population*, *Housing Price Index* and *Bankruptcy Rate*.

Our approach includes an initial exploratory data analysis to get familiar with the relationship between variables to inform our model building. Next, we fit models using four different modeling approaches, both univariate and multivariate. For each approach, we tuned the appropriate parameters with the goal of maximizing prediction accuracy--in this case we measured that by minimizing root mean squared error (RMSE). RMSE measures differences between the bankruptcy rate the model predicted for the validation set to the actual bankruptcy rate given by the observed data. The validation set is a subset of data from January 2009 to December 2010 that is used to test model predictions, since this data includes the actual bankruptcy rate.

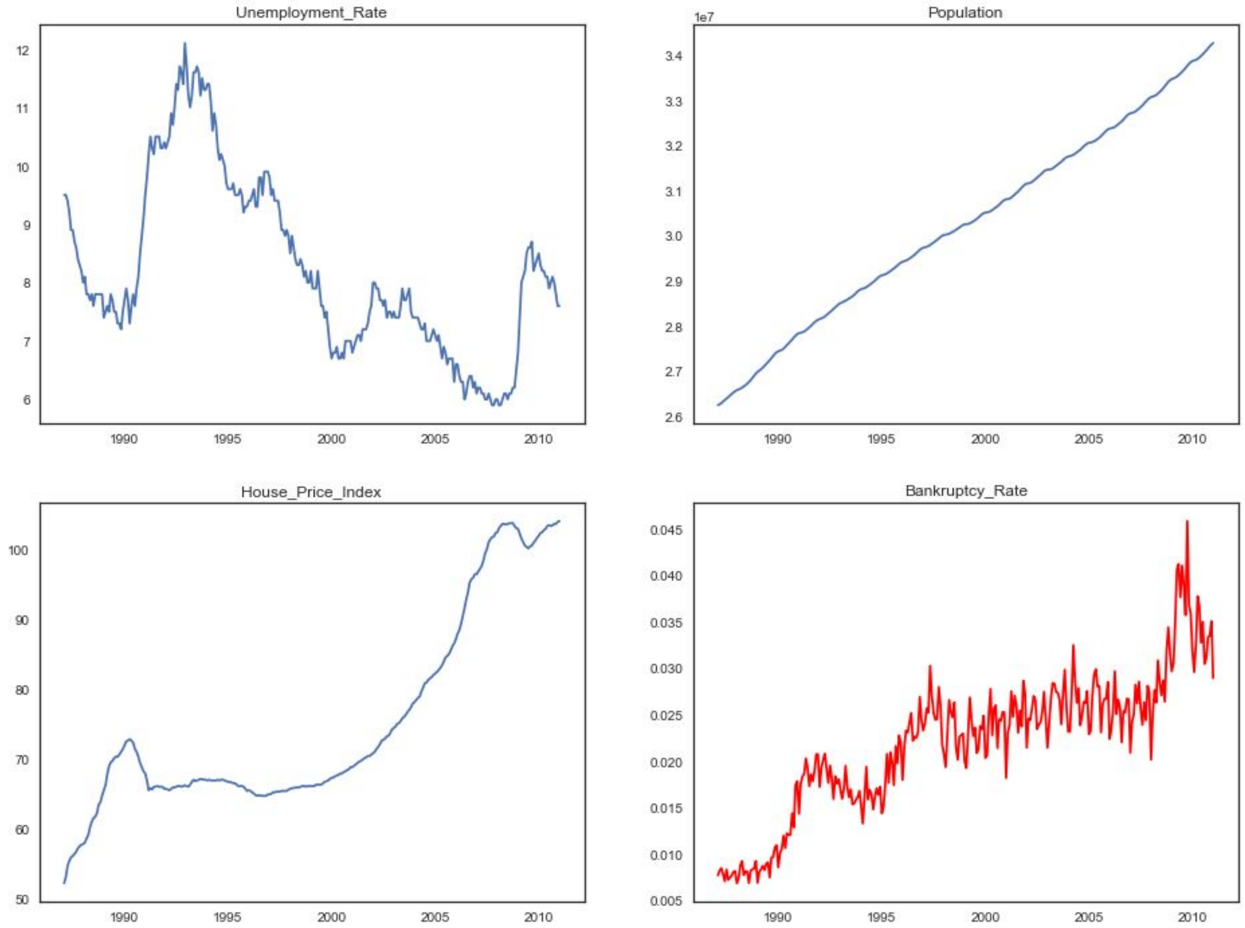
After producing optimal models for each approach, we evaluated each model's predictions and RMSE. Our final model took a combination of the best two approaches to give our final predictions for Canada bankruptcy rates between 2011-2012.

II. Exploratory Data Analysis

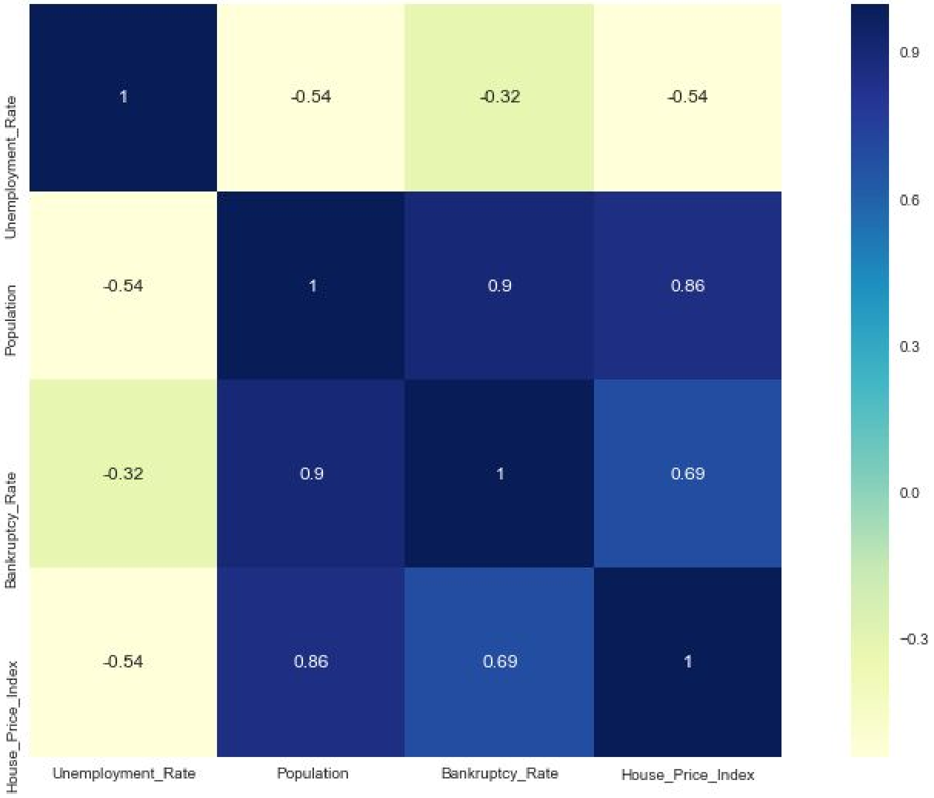
Let's look at a sample of the given data:

Unemployment Rate	Population	Bankruptcy Rate	House Price Index	Mon	Year
9.5	26232423	7.700E-03	52.2	1	1987
9.5	26254410	8.220E-03	53.1	2	1987
9.4	26281420	8.485E-03	54.7	3	1987
9.2	26313260	7.833E-03	55.4	4	1987
8.9	26346526	7.090E-03	55.9	5	1987
8.9	26379319	8.328E-03	56.1	6	1987
8.7	26412105	7.220E-03	56.4	7	1987
8.6	26446688	7.453E-03	56.7	8	1987
8.4	26482154	7.696E-03	57.2	9	1987
8.3	26516153	8.067E-03	57.5	10	1987

We can plot the four variables in the same axis to see how they evolve over time:

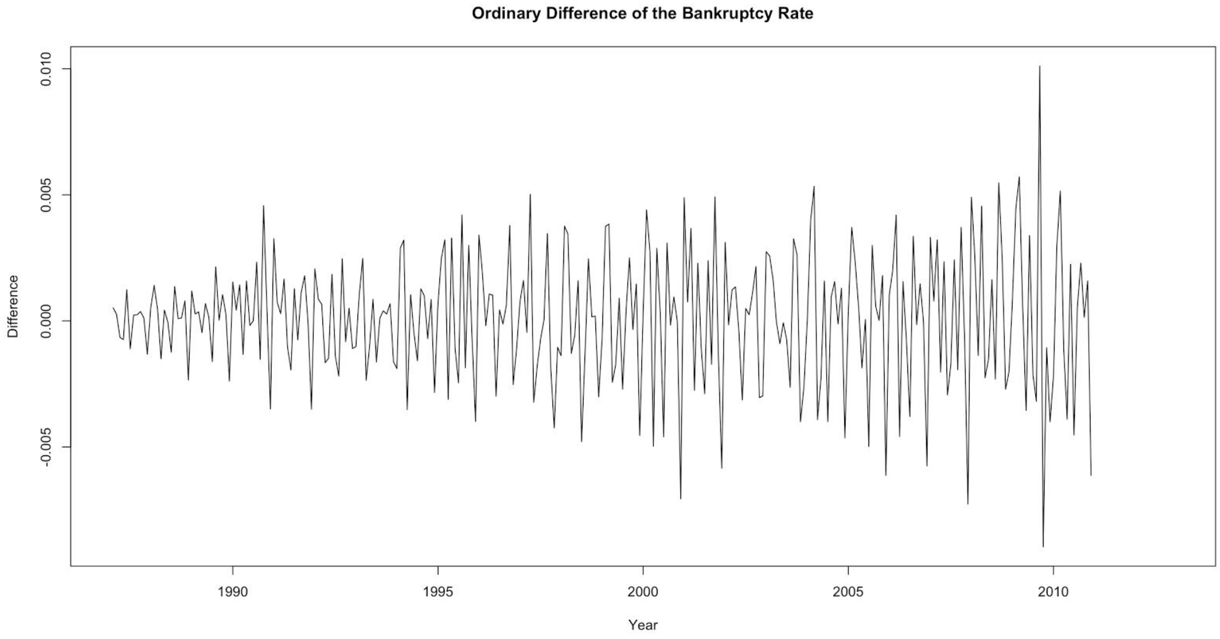


A correlation matrix can be an indicative measure on how four variables vary among each other:

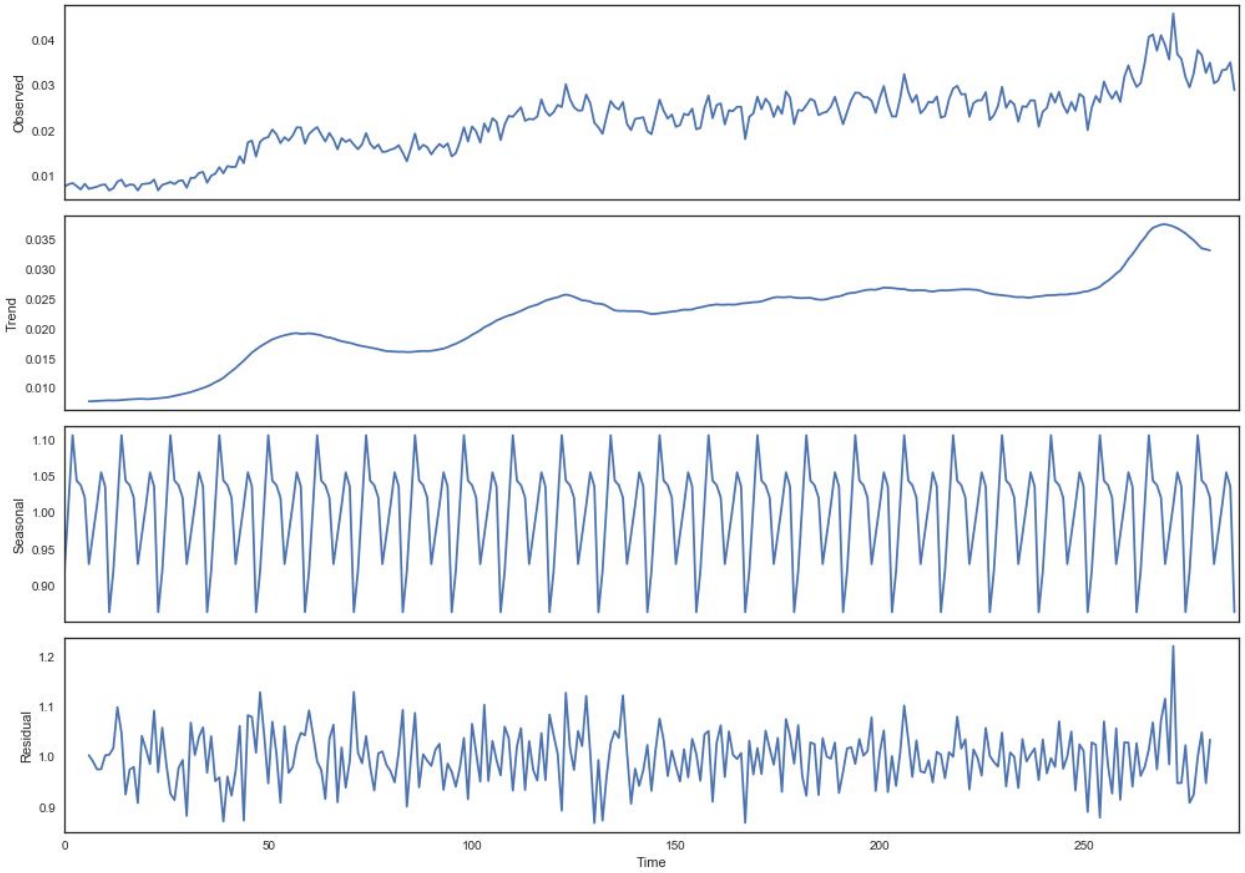


The matrix shows that *Bankruptcy Rate* is highly correlated to *Population* (a correlation coefficient of 0.9 means they almost always move in the same direction). This is counterintuitive and may be due to the fact that *Population* is constantly increasing, so *Bankruptcy Rate* is also picking up on that trend. *Bankruptcy Rate* is also highly correlated to *Housing Price Index* which leads to further exploration.

Differencing, or taking the difference between the current value and the previous value, makes a time series stationary. A stationary time series means that the properties of the time series (trend and seasonality) do not depend on the time which the series is observed.



Here we observe that the differenced bankruptcy rate does not have a constant variability, i.e. the spread seems increases overtime. This suggests a log transformation of *Bankruptcy Rate* may be beneficial for models that perform better on data with constant variance.



By using appropriate techniques, one can separate the time series into trend, seasonal and residual components. One interesting fact to note for modeling approaches is that the trend component is somewhat similar to the *Housing Price Index* time series with perhaps some lag terms.

III. <u>Univariate Time Series</u>

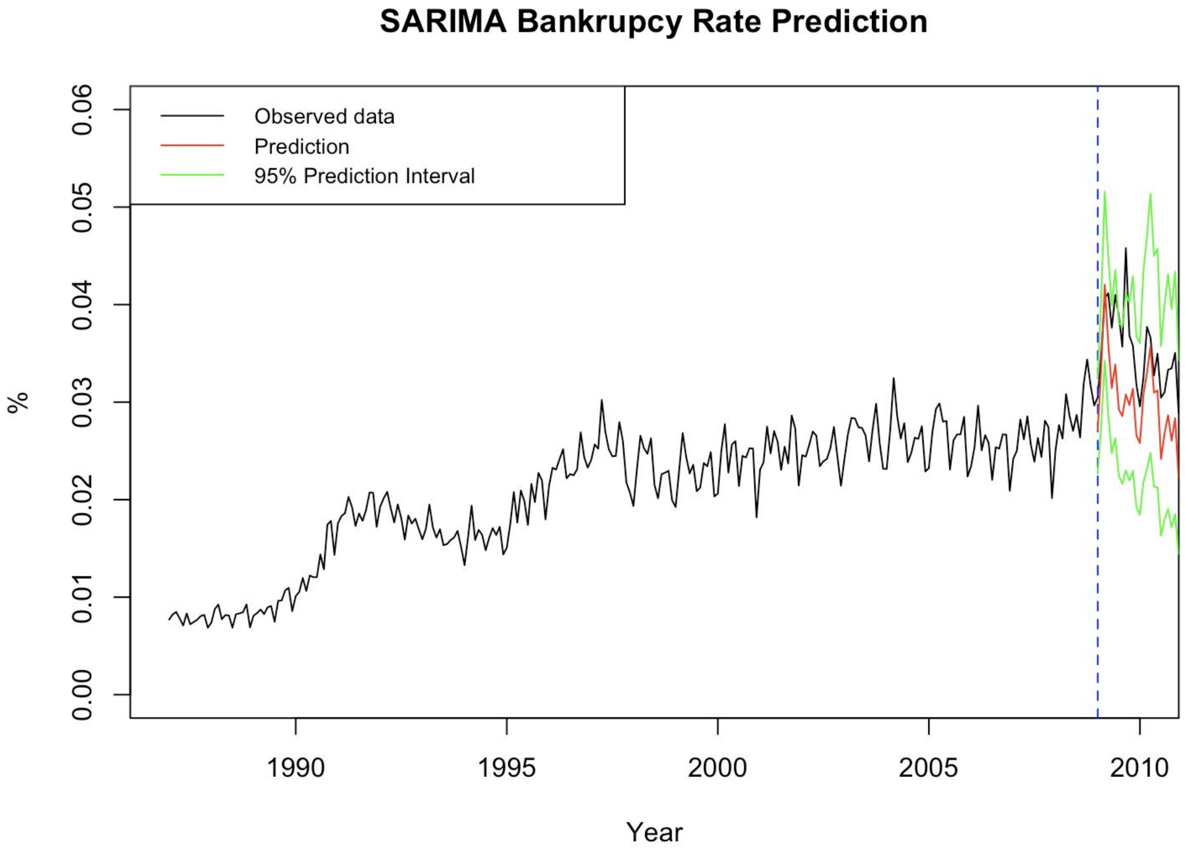
Univariate approaches to modeling this time series only consider past observations of *Bankruptcy Rate* and no other variables. We will look at two univariate time series modeling approaches: SARIMA and Exponential Smoothing.

SARIMA

The SARIMA approach is a univariate model that aims to describe autocorrelations in the data. The SARIMA model combines elements from a few different types of models. The AR part of SARIMA refers to an autoregressive model which forecasts *Bankruptcy Rate* using a linear combination of past values of the variable. The MA part of SARIMA refers to a moving average model. A moving average model uses past prediction errors in a regression-like model. The I in SARIMA means 'integrated' which refers to combining AR and MA models and differencing. The final piece of the SARIMA model is the S, which is included to model seasonal data.

The parameters of a SARIMA model are p, d, q, P, D, Q, and m. The p's refer to the order of the autoregressive part of the model, so how many lagged terms the model considered. The q's refer to the order of the moving average part of the model, so how many past forecast errors are considered. The d's refer to how many times the model must be differenced to remove trend (d) and seasonality (D). The m is the number of periods per season. Additionally, the lowercase parameters model the non-seasonal part of the model, while the uppercase parameter model the seasonal part of the model.

The optimal SARIMA model that produced the lowest RMSE on the validation set is SARIMA(2, 1, 1)(1, 2, $1)_{60}$. The forecasted predictions gave an RMSE of 0.00605. The plot below shows a visualization of the predictions. On the right side of the vertical dashed blue line is the forecasting part, where black line represents the real values of *Bankruptcy Rate*, red line represents the predicted values, and green lines represent the 95% prediction interval, which means that there is a probability of 95% for real values to be included in this interval. It is clear to see that although the predicted values have the similar trend with the real values, they are always smaller. This is probably due to the increasing bankruptcy rate after financial crisis. Since the training set we used is from January 1987 to December 2008, it may not capture that suddenly increasing trend. So the model tends to predict lower bankruptcy rates.



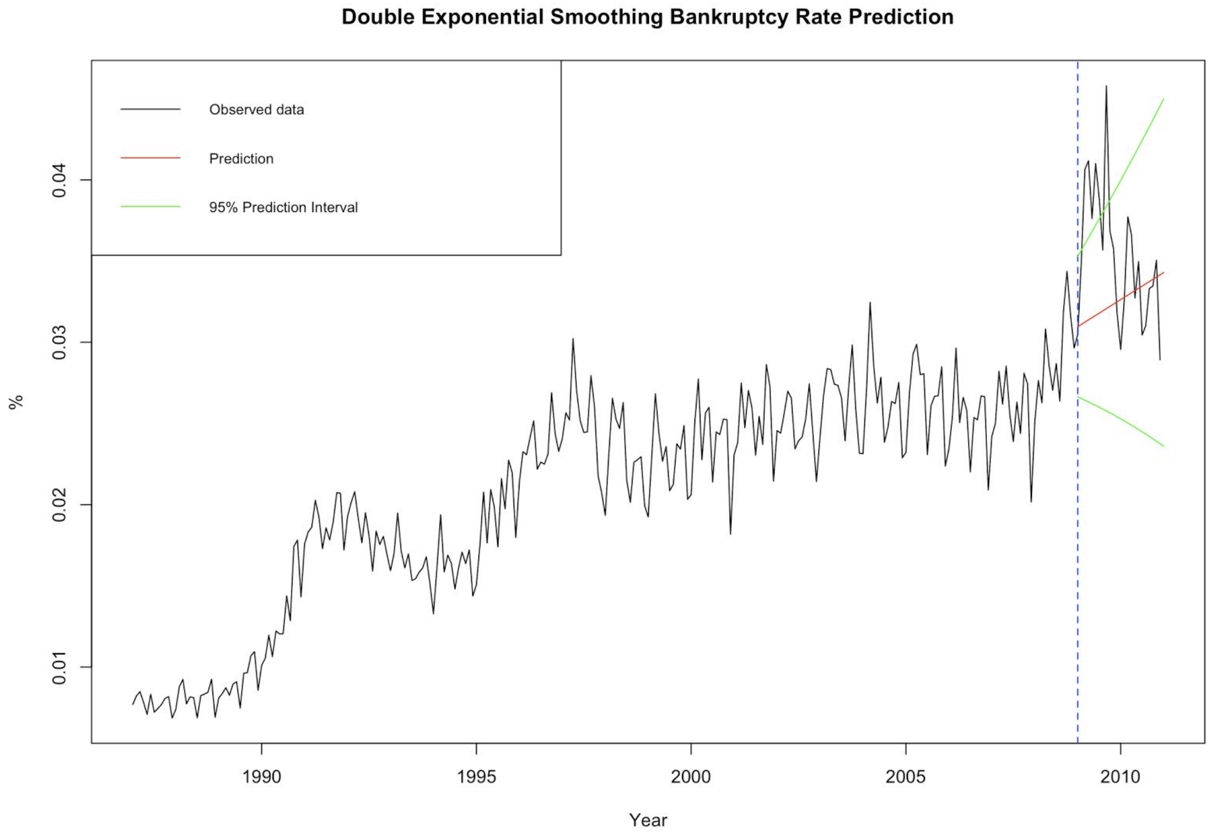
Exponential Smoothing

The exponential smoothing approach to modeling uses weighted averages of past observations, with the weights decaying exponentially as the observations get older. This approach is able to generate reliable forecasts for a variety of time series since it does not require any distributional assumptions.

The choice of exponential smoothing method depends on the components of the times series as to whether it exhibits trend, seasonality, or both. Single exponential smoothing is appropriate when the model exhibits neither trend nor seasonality, double exponential smoothing is for when a time series exhibits trend but no seasonality, while triple exponential smoothing is appropriate when a time series exhibits trend and seasonality. There are two variations to triple exponential smoothing that differ in the nature of the seasonality of the time series, additive and multiplicative. Additive triple exponential smoothing is preferred when the seasonal variations are consistent throughout the time series, while the multiplicative method is preferred when the size of the seasonal variations change across the time series.

The exponential smoothing model requires at most three parameters (alpha, beta, and gamma) which control the rates at which the weights on past observations decrease. The larger the value of the given parameter, the more weight is given to more recent observations in predicting future values. The alpha parameter controls the weight given to the level, beta controls the weights for trend, and gamma controls the weights for seasonality. Which parameters are used depends on the smoothing method.

The optimal model for exponential smoothing on the bankruptcy time series is a double exponential smoothing model with alpha= 0.335 and beta = 0.0326. The forecasted predictions gave an RMSE of 0.0054. The plot below shows a visualization of the predictions. It is clear that this is not an ideal model since the trend is increasing due to the heavy reliance on the trend of the past two years of data.



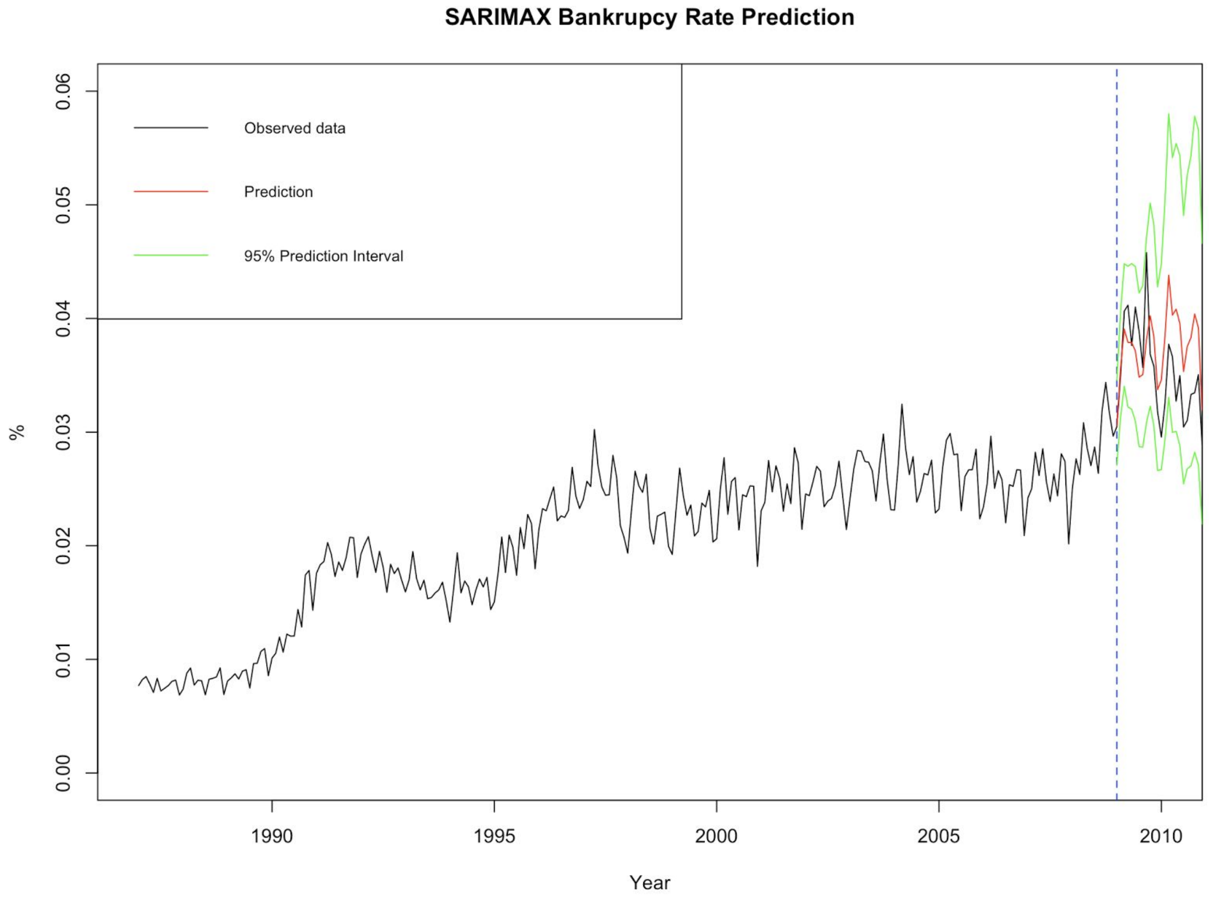
IV. Multivariate Time Series

Multivariate approaches to modeling this time series take supplementary time series variables into account to help model a response time series. In this case, we have *Unemployment Rate*, *Population*, and *Housing Price Index* at our disposal to help forecast *Bankruptcy Rate*. These supplemental variables may be one of two types that correspond to two different multivariate modeling approaches. Exogenous variables influence the response series, but the response doesn't influence them. A SARIMAX model would be most appropriate to account for this relationship. Endogenous variables influence the response and the response simultaneously influences them. In this case, a vector autoregression (VAR) model would be preferred.

SARIMAX

The SARIMAX model includes all of the model components of the SARIMA model mentioned above, with the addition of including exogenous variables. The inclusion of these variables accounts for the assumed relationship that these variables affect the response, *Bankruptcy Rate*, but *Bankruptcy Rate* does not affect these variables. The parameters of a SARIMAX model have the same interpretation as those in a SARIMA model.

When fitting combinations of different parameters and exogenous variables to SARIMAX models, the inclusion of just *Unemployment Rate* produced the smallest RMSE, equivalent to the RMSE produced with the inclusion of all three exogenous variables. Therefore, the optimal model chosen used just *Unemployment Rate* for simplicity and to reduce chance of overfitting. The optimal model, SARIMAX(2,1,3)(1,1,3)$_{12}$ + unemployment _rate produced an RMSE of 0.00451. A visualization of the prediction below show that model switches between under-predicting and over-predicting, which is indicative of a good model.

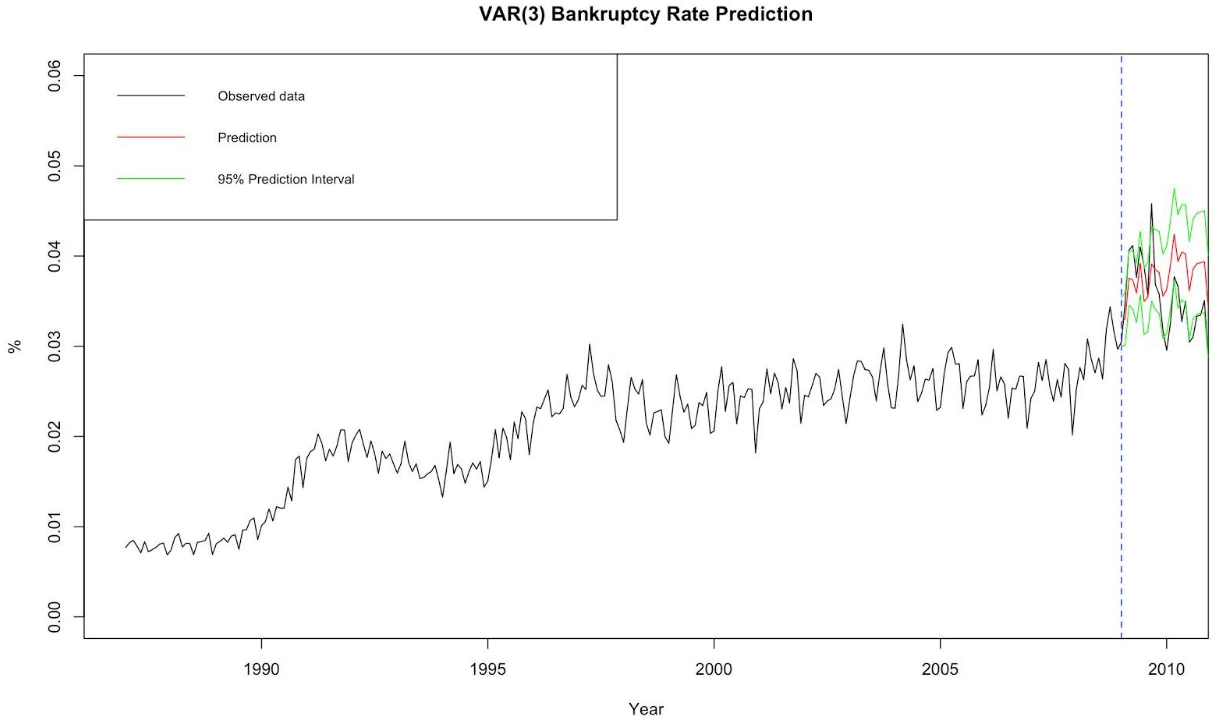


VAR

Unlike SARIMAX that treats variables as exogenous, vector autoregression (VAR) model treats all variables as endogenous, meaning they could all influence each other. Each endogenous variables is modeled similar to a regression model with the predictors come from other endogenous variables with maximum p lags, i.e. the values up to p-periods ago. The VAR model can capture quite complex patterns, however the downside is that VAR can quickly overfit the training data since the model has a large number of parameters. Therefore when tuning the parameter, one should consider using a smaller p lag even though the result might be slightly worse than another model with a larger p.

The model selection depends on three components: the different variables, the max p lag, and the amount of data to include in the training process. In addition to including all available endogenous variables (*Unemployment Rate*, *Population*, and *Housing Price Index*) we set an endogenous variable, *Seasonality*, which is an indicator for each month. To find the optimal model, each combination of the endogenous variables were tried, along with a max lag of 8 and varying training data windows. The reason for trying different training windows was that the fundamental time series trends changed in early 90's, so that the historical data that are 20 years old may not be very useful in predicting the future.

After trying 896 ways of combinations to model *Bankruptcy Rate*, with consideration for both RMSE and model complexity, the optimal model only considers *Housing Price Index* and *Bankruptcy Rate*, has a p value of 3, and uses data back to 1994. The model produced a RMSE of 0.003512. A plot of the result is show below. It is interesting to note that the upper 95% predicts the 2009 and the lower 95% predicts the 2010 *Bankruptcy Rate* extremely well. The actual prediction underestimates *Bankruptcy Rate* immediately after the financial crisis and overestimates it afterwards, but overall gives a decent prediction over the 2 years of prediction.



V. Model Choice/ Ensemble

The table below summarizes the optimal models from each approach.

Model RMSE Comparisons

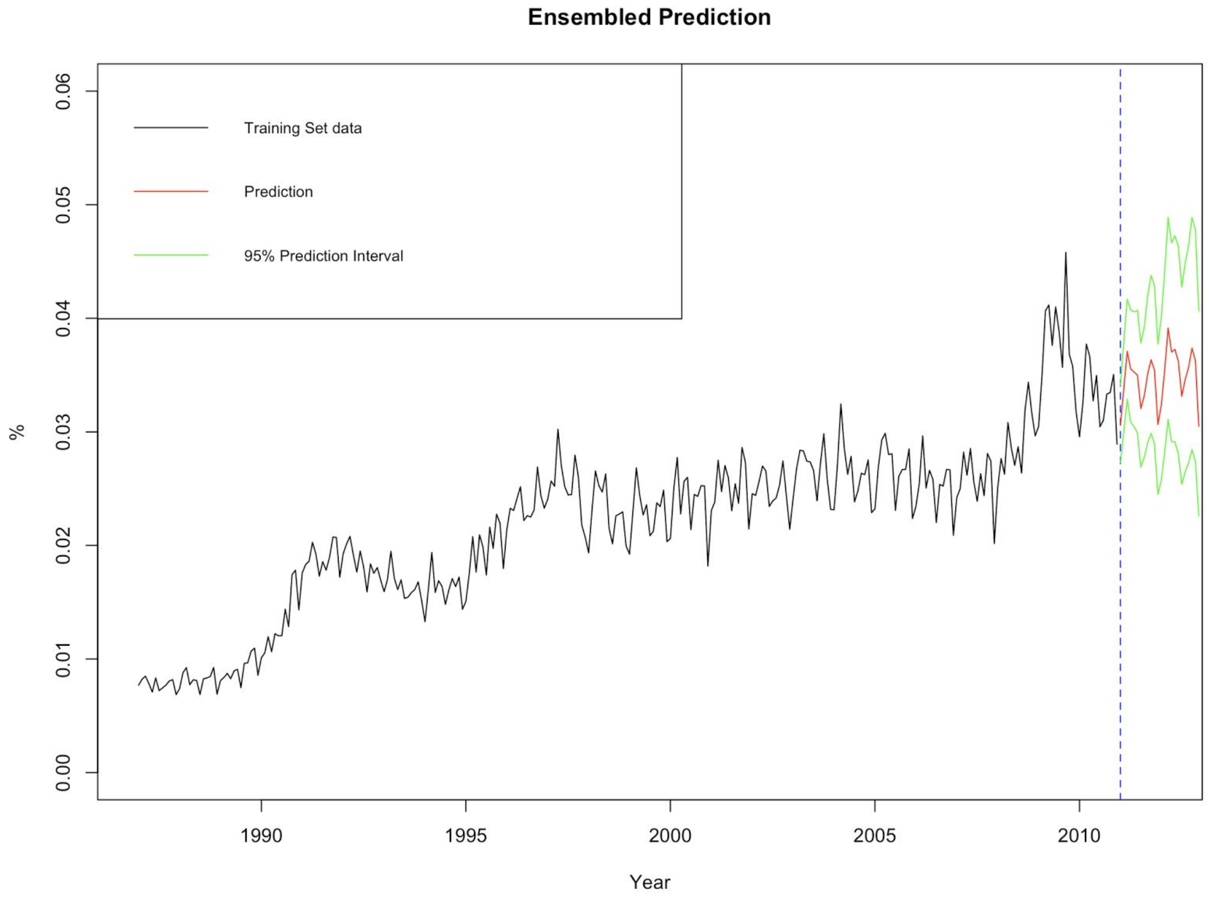
Model	Parameter	RMSE
SARIMA	SARIMA(2, 1, 1)(1, 2, 1)$_{60}$	0.006052
Exponential Smoothing	Double Exponential Smoothing: alpha= 0.335, beta = 0.0326	0.005401
SARIMAX	SARIMAX(2,1,3)(1,1,3)$_{12}$ + unemployment _rate	0.004514
VAR	VAR(3) (Bankruptcy, House Pricing Index with exogen variable of seasonality)	0.003512

By comparing the RMSE, the top two models are VAR(3) (Bankruptcy, House Pricing Index with exogen variable of seasonality) and SARIMAX(2,1,3)(1,1,3)$_{12}$ + unemployment _rate. As the data we have is limited in size, even using RMSE as the metric may cause variance and bias in prediction as potential overfitting can not be confidently identified. Therefore, we decided to ensemble the the top two models to avoid the possibility of overfitting.

Ensembling is combining the results of two separate models in attempt to obtain better predictive performance. The process of ensembling is as follows:

1. Refit the optimal VAR and SARIMAX model with all the data (including the validation set)
2. Generate mean prediction, maximum prediction, minimum possible prediction of the two models using test data
3. Average the results from the two models

Below is the visualized prediction results of the ensembled models, our final model choice. Here, the blue line represents where the last data we have for *Bankruptcy Rate* ends and predictions must be made.



VI. Conclusion

The ensemble of the VAR and SARIMAX models are the most desirable methods to predict future *Bankruptcy Rate* in terms of RMSE. From feature selections, one can say that housing price and unemployment rate have some correlation with bankruptcy rate, while population does not really affect bankruptcy rate.

Based on the forecasting diagram, one can see that the bankruptcy rate after 2011 will remain almost the same as 2009 and 2010, which is higher than the rate before financial crisis, but lower than the rate during the crisis.

Some limitations of this model include complications in the data surrounding the financial crisis of 2008. Since the training set we used is from January 1987 to December 2008, it may not capture that sudden increase after 2008 so our models tend to predict lower bankruptcy rates. However, since we do not expect such increase will happen again in the near future, we decided not to include the data from that time period in our training set in order to avoid overly predicting *Bankruptcy Rate* when financial crisis does not occur.