Tian Qi, Divya Bhargavi, Jessica Wang, Mahsa Ashabi
MSAN 604: Time Series Analysis
Professor Nathaniel Stevens

Report on Canadian Bankruptcy Rates

1. Description of Data

The goal with this time series data is to determine an appropriate modeling approach to predict Canadian bankruptcy rates. With a collection of monthly data from January 1987 to December 2014, we aim to predict the bankruptcy rates for the following 36-month period. In determining our model, there is also corresponding data on unemployment rates, population rates, and Housing Price Index that could be influential as to how we model and forecast bankruptcy rates. With various approaches described below, we will assess which model is the most appropriate fit to forecast Canadian bankruptcy rates for the January 2015 to December 2017 period.

2. Exploratory Data Analysis

The data given to us:

Month	Year	Unemployment_Rate	Population	Bankruptcy_Rate	House_Price_Index
1	1987	10.4	26232423	0.77004	44.4
2	1987	10.3	26254410	0.82196	45.2
3	1987	10.7	26281420	0.84851	46.6
4	1987	9.8	26313260	0.78326	47.1
5	1987	8.9	26346526	0.70901	47.5
6	1987	8.4	26379319	0.83285	47.7

Plotting the four variables: Unemployment rates, Population, Bankruptcy rate and Housing Price Index.

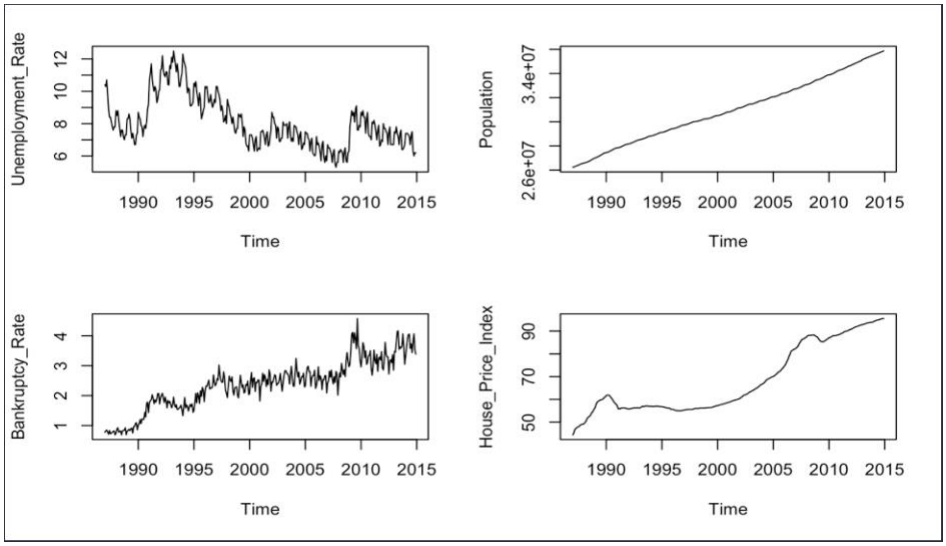

Fig. 1

Population rates and Housing Price Index indicate an increasing trend over the years, consistent with the Bankruptcy rates. However, Unemployment rate shows a decreasing trend overall.

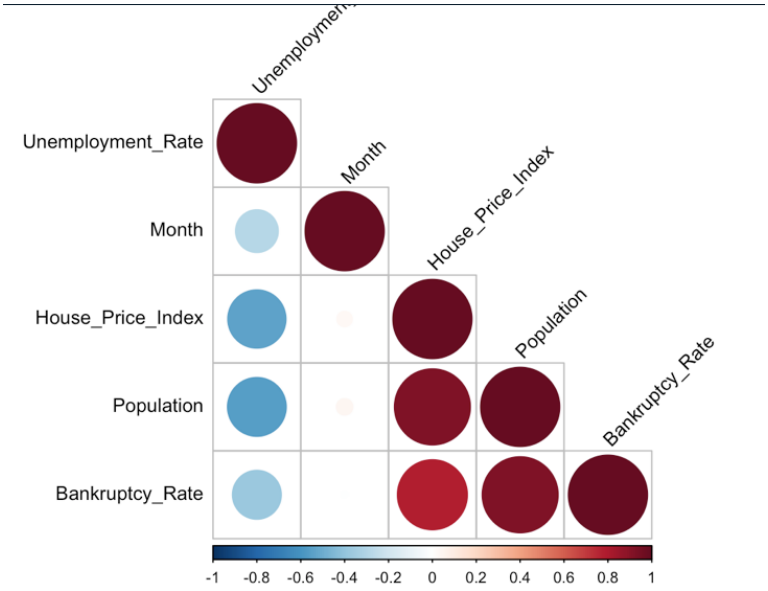

Fig. 2

From the above correlation plot we see that bankruptcy rate and House Price Index/ Population have strong positive correlation, whereas Bankruptcy rate and Unemployment rate have negative correlation.

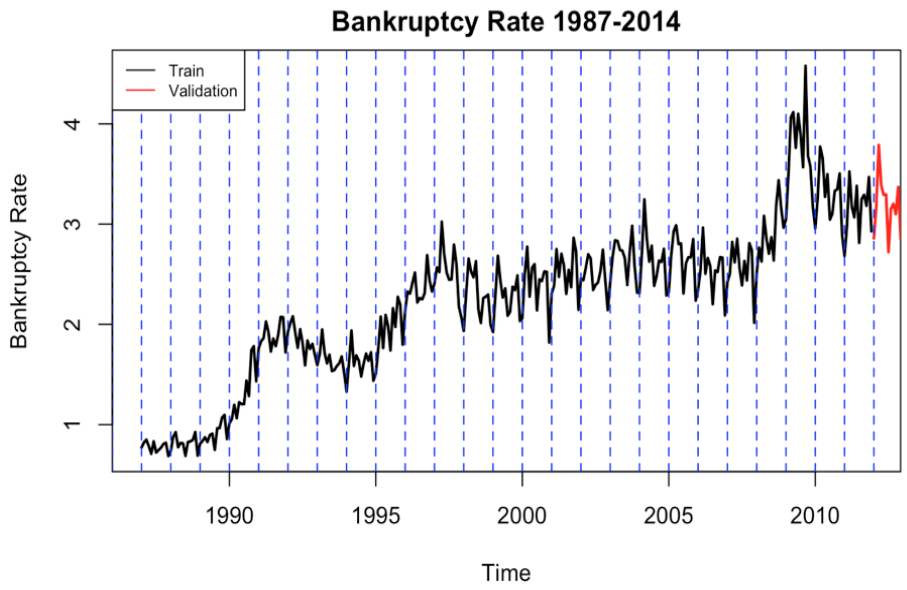

Fig. 3

Looking at the data between two blue lines, we see repeating pattern that is periodic within every year. Such repeating pattern in time series is called seasonality. There are two high peaks within a given year for most years, and more so for few years. This data might not indicate a strong seasonality. We will explore this further. Through variations of differencing, we can help stabilize the mean of a time series by removing changes in the level, and therefore eliminating (or reducing) trend and seasonal effect. Since the data shows heteroscedasticity, we do a log transform first. We split the data into rows of first 300 for training set and next 36 for validation set. Root Mean Square error (RMSE) on validation set is our criterion for model selection. The model parameters for ARIMA family of models are chosen based on the ACF/PACF plots below.

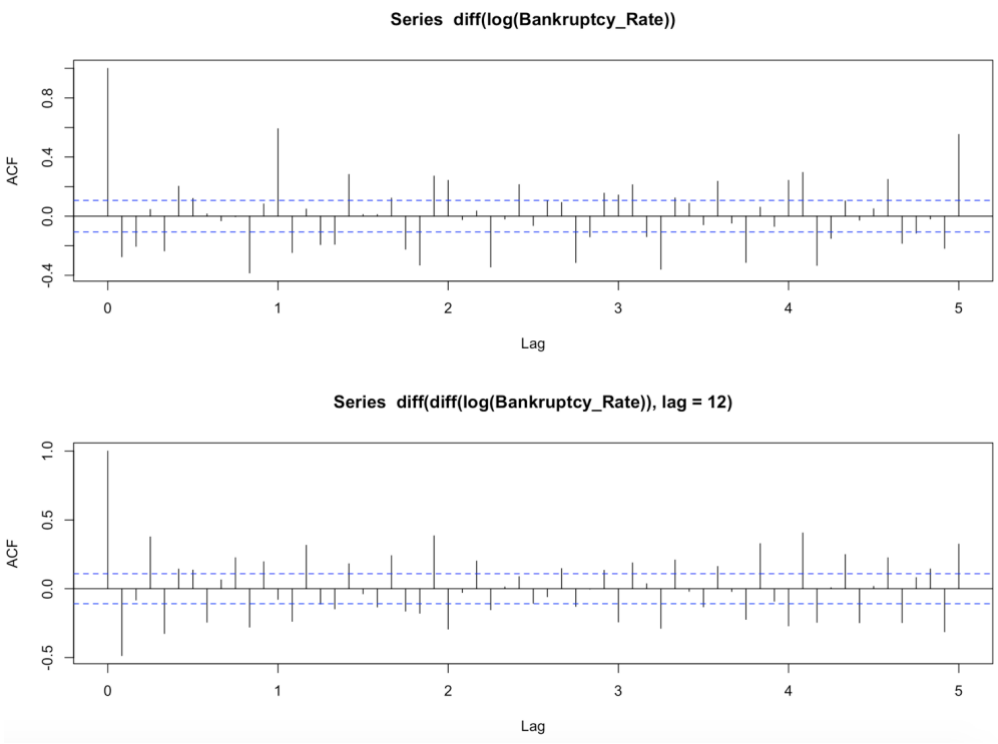

Fig. 4

Plot 1: ACF of ordinarily differenced transformed data. We see significant peaks at lags of 12,24,36 (given by 1,2,3 lags respectively in the plot). This suggests that the variables at lag 12,24,36 are strongly correlated signaling the effect of seasonality.
Plot 2: ACF of seasonally differenced data used for plot 1. Since the peaks at lag 12,24,36 are still significant, we may choose to not do seasonal differencing.

For the models described below, we chose to not include seasonal differencing due to the indication of our ACF/PACF plots and confirmation from the '*nsdiffs*' command on R.
q: First 5 lags seem only slightly significant.
Q: We can start with lags 1,2 (at 12,24) and decide on higher order. +/- 2 values from these can be tested.

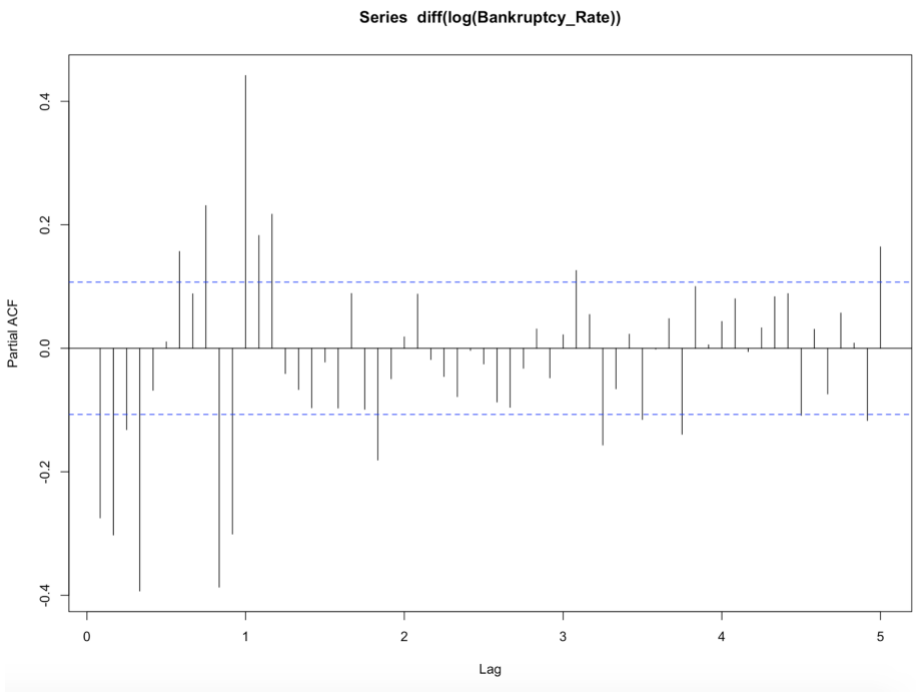

Fig. 5

For the PACF plot of the ordinarily differenced transformed data, we see that lags 1,2,3,4,7 indicate spikes that are significant. We can check the models for values p = 4-7 and decide based on log-likelihood values of the model. In testing potential orders for P, we see that the lags have significance at 1-2 lags, thus we test our SARIMA/X models with P=1-2

Next, we assess the different models available to us to fit to our data.

3. Potential Modeling Approaches

a) SARIMA: Seasonal Integrated Autoregressive Moving Average

The first model that we explored was SARIMA, which is used when we have seasonality in data for a non-stationary process. From Figure 1, it is evident that there is a directional movement that exists in the original data, so an ordinary difference is required to make the data stationary. Based on values of p,q,P,Q from the ACF/PACF plots above and setting d = 1 and D = 0 for ordinary and seasonal differencing respectively, we trained on multiple models. The model we finalized was SARIMA $(5,1,0)(2,0,1)_{12}$ [stands for $(p, d, q)(P, D, Q)_{\text{seasonality period} = 12}$] which gave lowest RMSE of 0.2554 on validation set.

In our model building, there were some higher order models that were tested which gave slightly lower RMSE but similar goodness of fit metrics. We stuck with our final choice to avoid any

overfitting of our model, which would result in capturing the noisy elements of the error rather than the data structure itself. The residual values produced in this model showed a constant variance across time, no correlation across time periods, and evidence of a normal distribution. This aids in the verification of our assumptions of the model. The final prediction of the SARIMA model and its corresponding 95% confidence interval is plotted below.

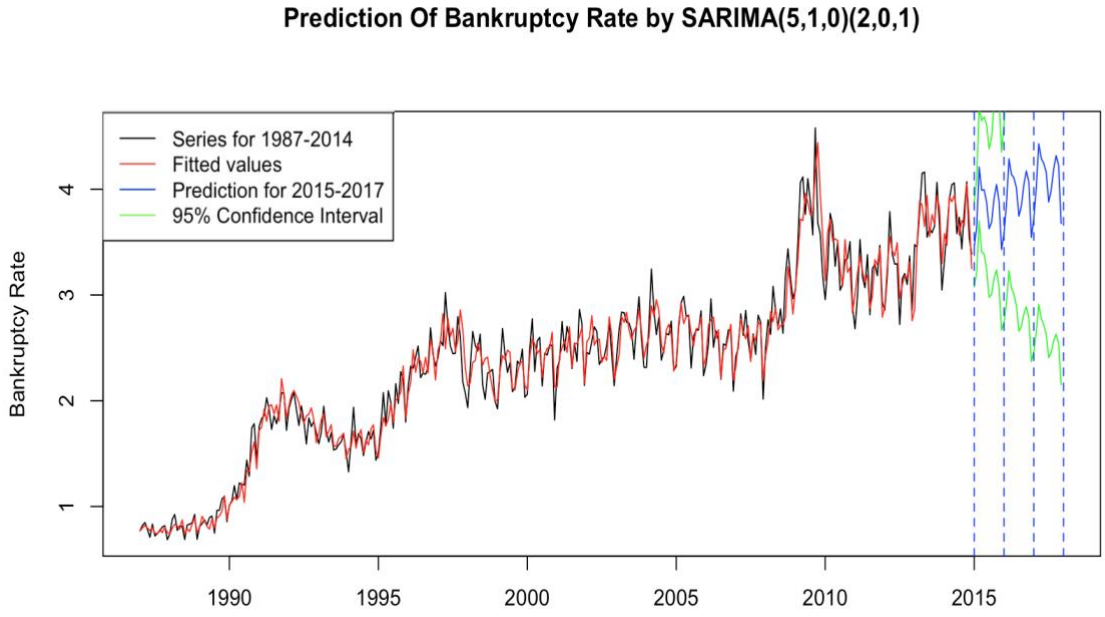

Fig. 6

b) SARIMAX: SARIMA + explanatory variables

SARIMAX model incorporates all the features of a SARIMA model but it also takes into factors in other variables that can influence the forecasting outcomes of the response variable. The parameters of the model have same interpretation as a SARIMA model. We use this type of model when the relationship between the variables and the response is unidirectional; they influence the response and not the other way. We call these variables exogenous.

From the individual plot of all variables, we see that House Price Index and Population show similar increasing trend like Bankruptcy, but Unemployment rate showing a decreasing trend. However, when we accounted for the influence of each of these and combinations of them on Bankruptcy rates, the test with Unemployment rate as exogenous variable gave the lowest RMSE of 0.2133 on the validation set with the model SARIMAX $(5,1,6)(1,0,2)_{12}$. This model used log transformations on Bankruptcy rates and Unemployment rate to get stationarity and reduce the scale of the variable respectively. We also tried the model with same lag values but with both seasonal and ordinary differencing. This gave a worse performance on all 3 goodness of fit metrics AIC, log-likelihood and variance.

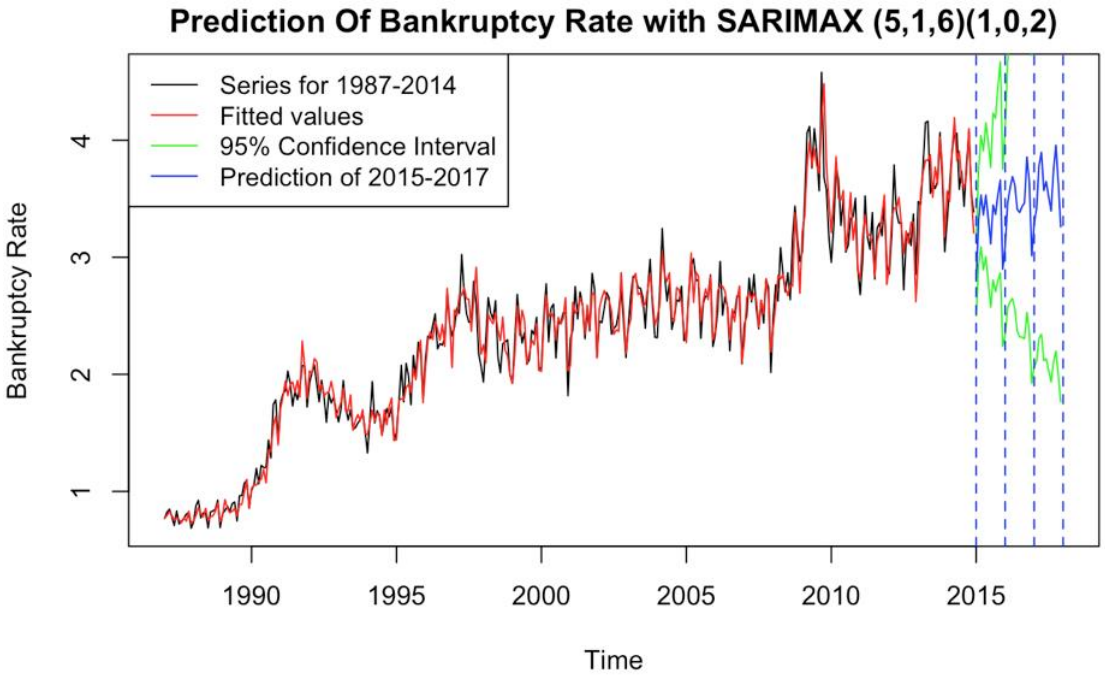

Fig. 7

c) Holt-Winters Method: Exponential Smoothing

The Holt-Winters method uses exponential smoothing to forecast into future. This is done by taking the exponentially weighted average of previous observations using a set of recursive equations with the observed data. In doing so, the recent observations are typically weighted more and there is no strict assumptions about the distributions. When plotting the data, there is a clear trend and seasonality that indicates the need for a triple exponential smoothing(TES) model. This version of exponential smoothing is able to capture all three factors of level, trend, and seasonality in the data. In addition, heteroscedasticity that is present indicates our need to choose a multiplicative version of smoothing to deal with the dependence on variability.

With TES, we estimate three parameters for this model: alpha for level, beta for trend and gamma for seasonality. If these parameters are close to 1, more weight has been given to recent observations, demonstrate that the model has less smoothing. If the parameter is close to 0, less weight has been given to recent observations, which demonstrates more smoothing. In practice, alpha = 0.2 is a commonly chosen standard but a more fitting value of the parameters can be determined from the data. Through the model iteration, the parameters chosen by RMSE on the validation data are as follows: alpha = 0.23, beta = 0.16 and gamma = 0.11. The forecasted prediction produced an RMSE of 0.229 on the validation data. The plot below indicates a visualization of these predictions with the TES model, with wider prediction intervals than the other models tested.

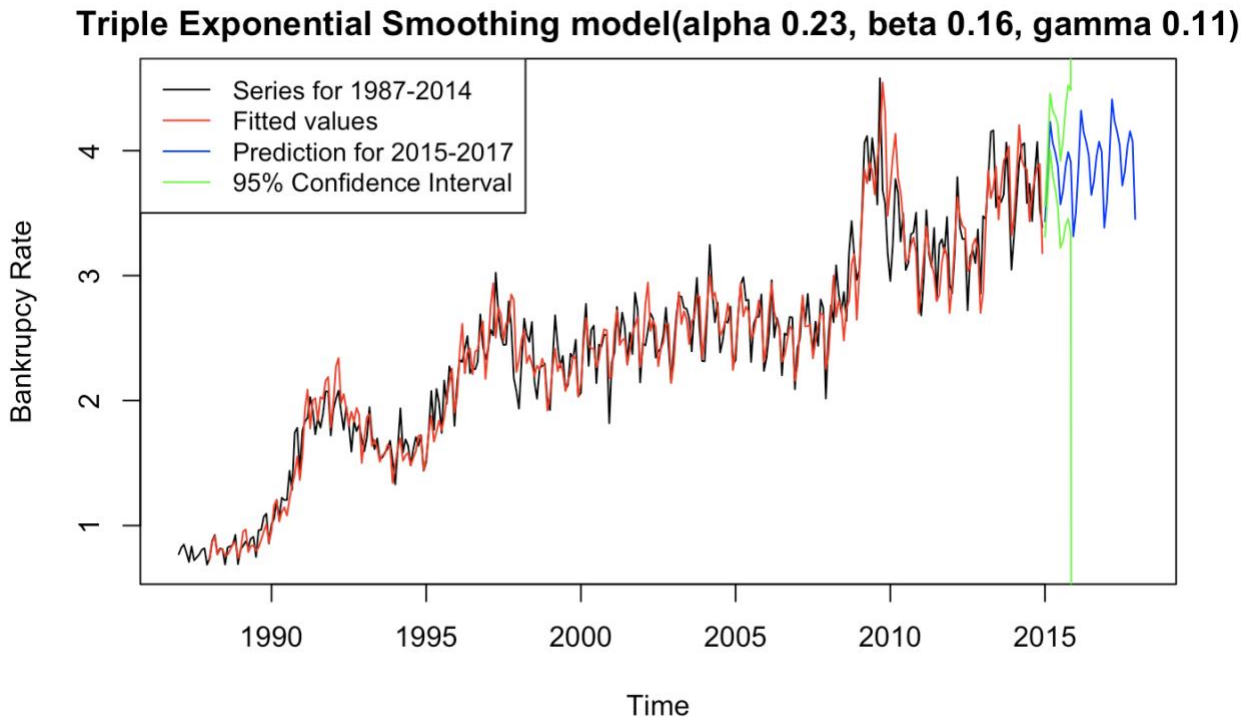

Fig. 8

d) VAR: Vector Autoregression

Instead of treating all variables as exogenous like *SARIMAX*, the vector auto-regression, abbreviated as *VAR*, treats all variables as endogenous, meaning they all influence each other. The structure of the model is that each variable is a linear function of other variables with maximum *p* lags, and past *p* lags of itself. The VAR model is one of the most flexible models for multivariate time series. It excels at capturing the dynamic behavior of economic and financial time series data because these two variables are more likely to be endogenous. However, VAR can easily over-fit a training data set because it utilizes too many parameters. Therefore, when building the models to create forecasts, it is helpful to consider a smaller p value and a small subset of existing variables we have.

When building the VAR model, there are three parts to consider: endogenous variables, the time lag *p*, and the length of past training data. Besides the housing price index, unemployment rate, and population, we also set a seasonality variable to indicate each month. The two possible training windows are all previous data because it may be beneficial in exploring any underlying relationships, or the data from the past five years, which allows us to see the bankruptcy rate had significant variation in recent years, indicated by two distinct spikes. This is compared to where it remained roughly horizontal from 1993-2007, demonstrating that these data might not useful for predicting the future. For the lag *p* parameter, I set the maximum value of p to be 3 since a higher value would result in overfitting. In order to find the best *p*, length of the past training data, and set of variables,

we tried numerous combinations to see which gave the lowest RMSE values on validation data, indicating the best predictive accuracy of the data. The final result is a VAR (3) model with all endogenous variables previously mentioned, resulting in an RMSE value of 0.249.

In order to predict the forecasted values for the bankruptcy rate, the model will be rebuilt with the parameters from the previous steps and trained on the overall data we have, via the training and validation data in the last step. A plot of our prediction is below:

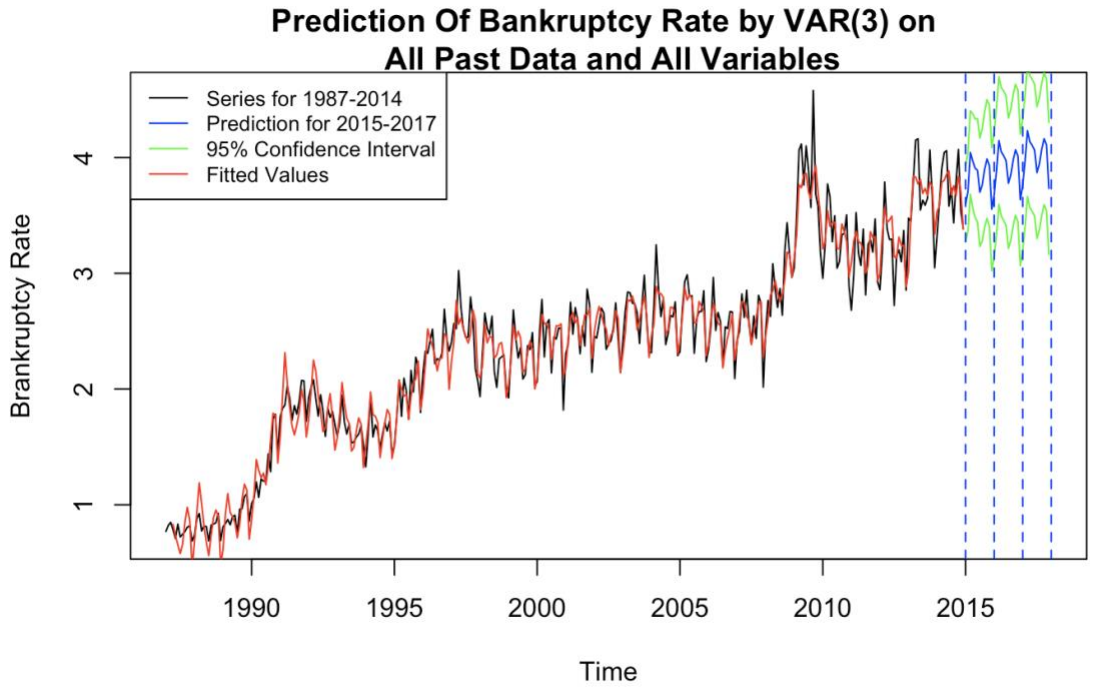

Fig. 9

It is apparent that the model produces a prediction with a slightly increasing trend. Therefore, it is fitting due to the fact that the bankruptcy rate had a large variation in the most recent 10 years period. In addition, the 95% confidence interval was not particularly large, which provides a suitable range for the forecast.

4. Optimal Model Selection & Forecasting Discussion

The following is the summary of best models including their parameters and RMSE on validation set.

Table 1: Summary of Models

Model	Parameters	RMSE
SARIMA	SARIMA(5,1,0)(2,0,1)12 lambda=0.21	0.2550
SARIMAX	SARIMAX(5,1,6)(1,0,2)12 +Unemployment Rate	0.2133
Exponential Smoothing	TES alpha=0.23 beta=0.16 gamma=0.11	0.2290
VAR	VAR(3) (4 variables + Seasonality)	0.2490

After assessing the various goodness-of-fit metrics, RMSE values for predictive accuracy and residual assumptions of our 4 potential approaches, the model selected as most appropriate for forecast our data is an ensemble model that averages the prediction of models.

Since all four models gave a relatively small value for RMSE, the final forecasting result is the average of four models, meaning we take the average result from the 4 models as the final prediction. By doing so, we can avoid overfitting and can produce a better predictive performance overall. Below, we include our final forecast for 2015-2018 plotted with the original data set.

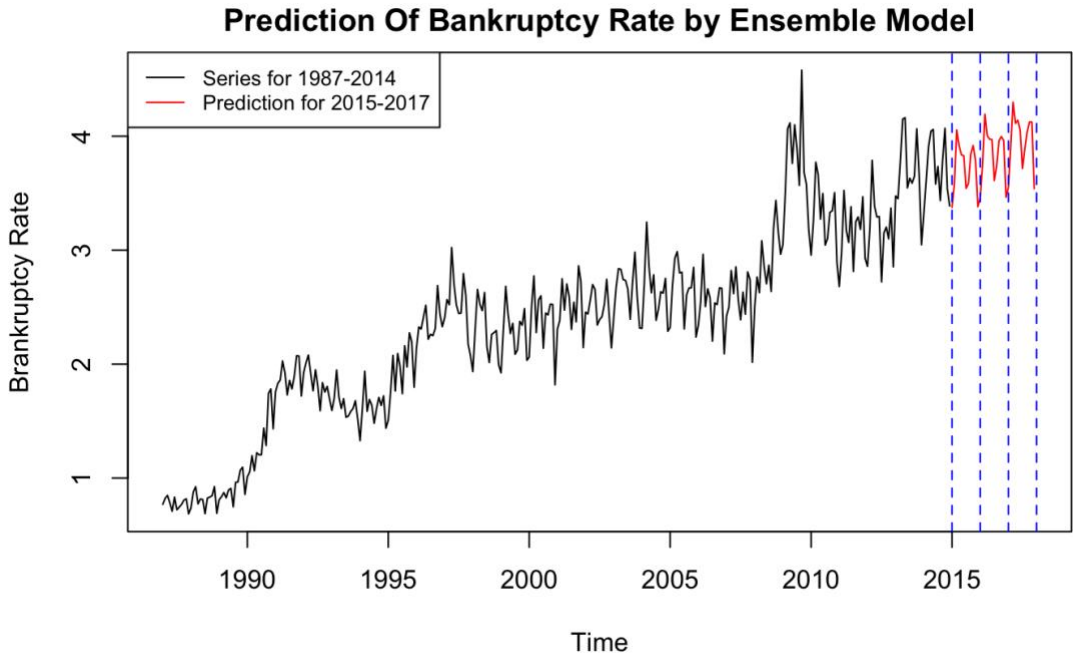

Fig. 10

Appendix: Time Series Vocabulary

Transformation: From the Bankruptcy plot we see above, the variance of the peaks within every year lacks uniformity. The initial few years have insignificant peaks and later we see prominent ones. Thus, we move forward with a log transformation to balance the data. Below, the log-transformed graph shows lower variance than the original plot.

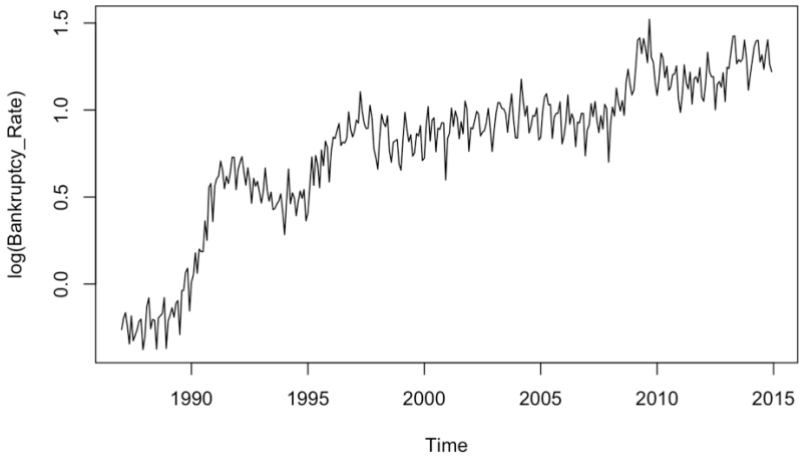

Fig. 11

Stationarity: A stationary time series is one whose properties do not depend on the time at which the series is observed. A time series variable that doesn't exhibit this invariance of time is said to be non-stationary. Thus, time series data with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the data at different times.

Ordinary differencing: Computing differencing between consecutive terms. For a variable Y, it is $Y_t - Y_{t-1}$. This is denoted by 'd'.

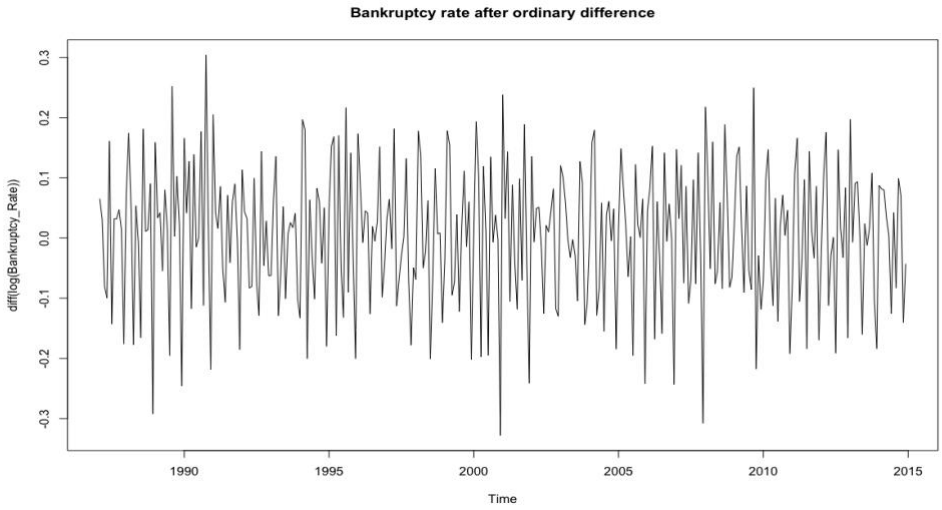

Fig. 12

We see that after a log transformation and ordinary differencing of the data, there are no obvious trends. The graph shows stationarity.

Seasonal differencing: The **seasonal difference** of a time series is the series of changes from one *season* to the next. For monthly data, in which there are 12 periods in a season, the seasonal difference of Y at period t is $Y_t - Y_{t-12}$. We use this to remove the correlation between occurring over the seasonal period. This is denoted by D, in which we posit that every observation within a season is a time series.

ACF (Autocorrelation Function) plots: We can calculate the correlation for time series observations with observations from previous time steps, called lags. The ACF is a way to measure the linear relationship between an observation at time *t* and the observations at previous times. Since the correlation of the time series observations is calculated with values of the same series at previous times, this is called a serial correlation, or an autocorrelation. We use an ACF plot to determine what is the last q/Q lags to consider for the forecast error to predict future observations. The lag considered for between-season time series is denoted by Q and within-season time series is given by q.

PACF (Partial Autocorrelation Function) plots: Partial autocorrelation is the amount of correlation between a variable and a lag of itself that is not explained by correlations at all intermediary lags. We use PACF plots to determine what is the last p/P lags to consider for previous observations of Y when forecasting. The lag considered for between-season time series is denoted by P and within-season time series is given by p.

AR: (Auto-regressive): An **autoregressive model** is when a value from a time series is regressed on previous values from that same time series. for example, y_t on y_{t-1}.

MA (Moving average): A **moving average** model is when a time series is regressed on past errors.

SARIMA: includes AR, MA and seasonality component.