Donya Fozoonmayeh
Darren Thomas
Yuhan Wang
Nicole Kacirek

Time Series Group Project

Problem Description

Bankruptcy rate is an issue of concern for various interested parties, including banks, insurance companies, and politicians. Accurately forecasting the bankruptcy rate would provide valuable, actionable information to individuals working in these fields. To this end, our bankruptcy rate forecasting analysis was performed using two datasets. The first, `train.csv`, contains data on bankruptcy rate, unemployment rate, population, and house price index in Canada for 28 years, from January 1987 to December 2014. Our goal is to determine which information and which time series model is most effective in forecasting the bankruptcy rate. After finalizing the model, we will then test its efficacy by forecasting the bankruptcy rate in Canada in the ensuing three years (2015 to 2017) using information from the second dataset, the `test.csv` file. This file, like the 'train.csv' file, also contains data on unemployment rate, population and house price index. The goal is to do the best possible forecasting with a proper time series model or a set of models.

Available Modeling Approaches

Univariate:

The type of model we used is referred to as a time series. A time series is a collection of data points that model the relationship between a numerical value at designated time intervals. For example, the data points in this time series will be the bankruptcy rate for every month in Canada beginning in January, 1987. Visually, this can be displayed on a graph with the bankruptcy rate recorded on the y axis, and the corresponding months

indicated on the x axis. Because we are currently only looking at one variable over time, bankruptcy rate, we can define this as a univariate time series.

When modeling any time series, it is important to first understand the following terms: trend, seasonality, and serial correlation. Trend can be thought of as the general directional movement of a time series. While a time series may have spikes that vary month to month, the trend of time series could be the more general behavior of the time series. For example, the S&P 500 over a 20 year span would show a general trend upwards, with a dip in 2008 during the recession followed by another increase. Seasonality is the characteristic of a time series in which the data experiences regular and predictable fluctuations over time. An example of this would be electronic sales which spike every November and December due to holiday gifting. Lastly, serial correlation is the state of affairs when observations closer together in time tend to be more similar than observations further apart. For example, the closing price of a stock depends on the stock price the previous day. We represent correlation on an autocorrelation plot, which will be demonstrated in further detail at another time.

With these terms defined, it is also important to define the forecasting time series model as stationary or non-stationary. Most time series data is non-stationary, meaning there is some serial correlation between the variables for certain points in time. For example, if a time series demonstrates an upwards trend, a value shown at the very beginning of that time series is less likely to occur than value from the previous day. For our purposes in bankruptcy forecasting, a stationary model is required - the intrinsic trend of the measured variable, rather than the variable's absolute value, is the important factor. In order to ensure stationary data analysis, differencing is required Differencing means taking the value of every variable, and subtracting it from the previous variable. This is done until the model becomes stationary in nature. It is also important to note that it is possible to over difference, which means differencing after the model is already stationary. This increases the error terms on the model and makes

forecasting more challenging. The important thing is to difference until stationarity is met, and to then stop.

With this background information established, using the univariate modeling approach allows for the option of two different methods to establish a model: Holt-Winters and Box-Jenkins. The key difference between these two models, is that the Box-Jenkins method applies some version of an autoregressive moving average (ARMA), while Holt-Winters uses some version of exponential smoothing to find the best fit of a time series model to past values of a time series. With ARMA, the past observations are weighted equally, but with Holt-Winters, exponentially is used to assign exponentially decreasing weights over time. Different versions of both ARMA and exponential smoothing exist and choosing the best depends on the trend and seasonality of the data. If the model is stationary, ARMA and single exponential smoothing can be used. If the model exhibits trend, but no seasonality, autoregressive integrated moving average (ARIMA) and double exponential smoothing must be used. Lastly, if the model demonstrates seasonality, regardless of trend, seasonal autoregressive moving average (SARIMA) or triple exponential smoothing must be used.

With this understanding of the terms, principles, and techniques for modeling and forecasting a time series described in the univariate context, these concepts will next be discussed from the multivariate analysis perspective.

Multivariate:

While univariate models use only the temporal observations of one variable (in this case bankruptcy), it can sometimes be helpful to also use additional variables in the model. Additional variables may add value to the model if they are 1) collected at the same frequency and for the same duration as bankruptcy and 2) are highly correlated with bankruptcy. For this project, the dataset contains three other variables - unemployment rate, population, and house price index. This data was collected at the same monthly

time intervals as bankruptcy. Therefore, so long as they are correlated with bankruptcy, they could potentially be useful to help formulate a predictive model. If these variables are not correlated with bankruptcy, their behavior does not provide any indication as to how bankruptcy changes over time and including them would not give any more information than the univariate model.

Correlation Between Bankruptcy & Other Variables

	Unemployment	**HPI**	**Population**
Bankruptcy	-0.379526	0.7804405	0.9150432

After computing the correlations (see above) between the additional variables and bankruptcy, we found population and house price index to be highly correlated with bankruptcy and unemployment rate moderately correlated with it. Thus, it might be a good idea for us to try to model bankruptcy using a multivariate model that includes one or more of these additional variables.

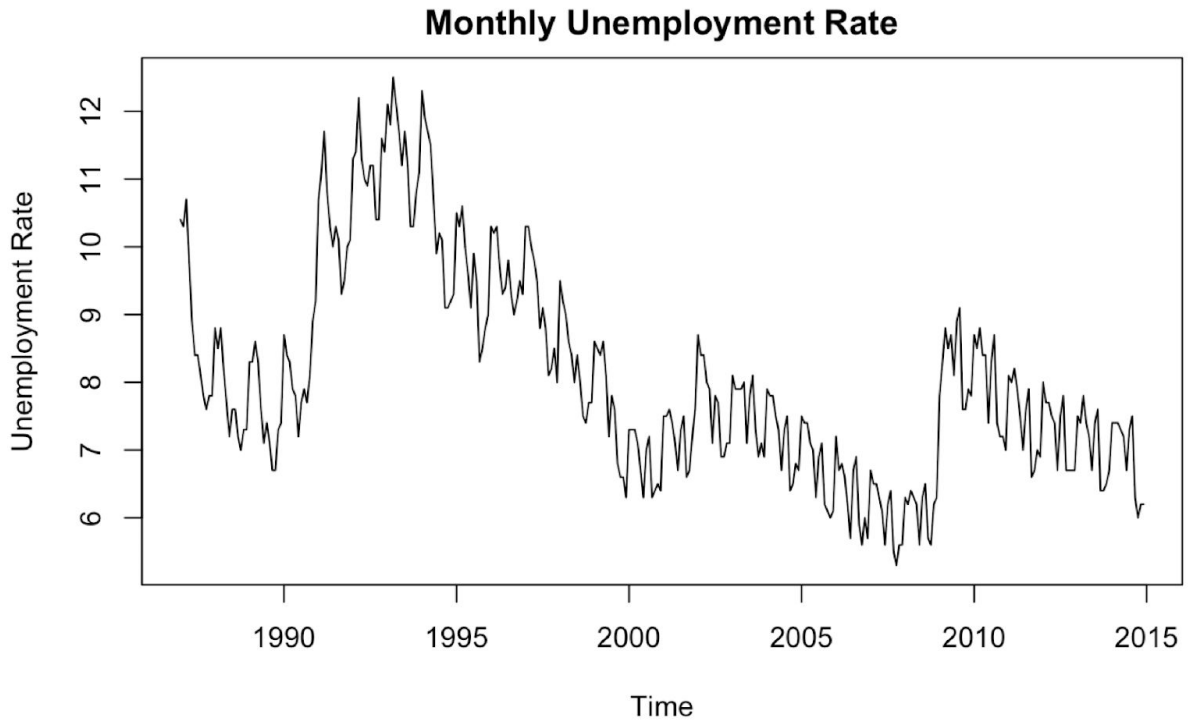
The specific model we choose depends on how we view the relationship between bankruptcy and these additional variables. We say that the relationship is exogenous if we believe that population, house price index, and unemployment rate influence the behavior of bankruptcy, but bankruptcy has no influence on them. If this were true, we would use a SARIMAX model. However, we believe bankruptcy does actually affect population, house price index, and unemployment rate - or in other words the relationship is endogenous. Therefore, a SARIMAX model would not be appropriate since it doesn't account for the influence of bankruptcy on the other variables. Instead, we should try a Vector Autoregression (VAR) model to try and account for this bidirectional relationship.

Chosen Method

In order to model monthly bankruptcy rates, we first plotted monthly bankruptcy rate, unemployment rate, population and housing price index for January 1987 to December 2014.

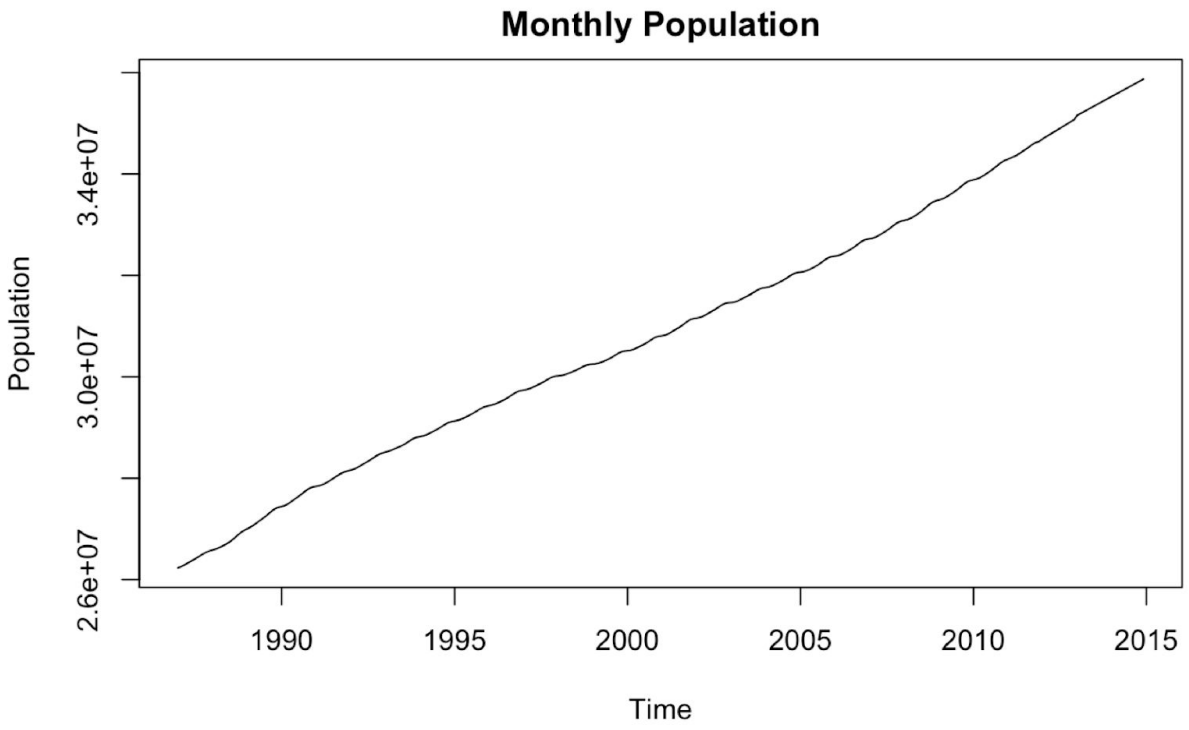
1. Unemployment Rate:

As it can be observed in the graph below, monthly unemployment rate has been variable throughout the years. The graph shows both trend and seasonality.



2. Population:

The graph below shows monthly population. As it can be seen, there is a clear trend and population has increased over the years.

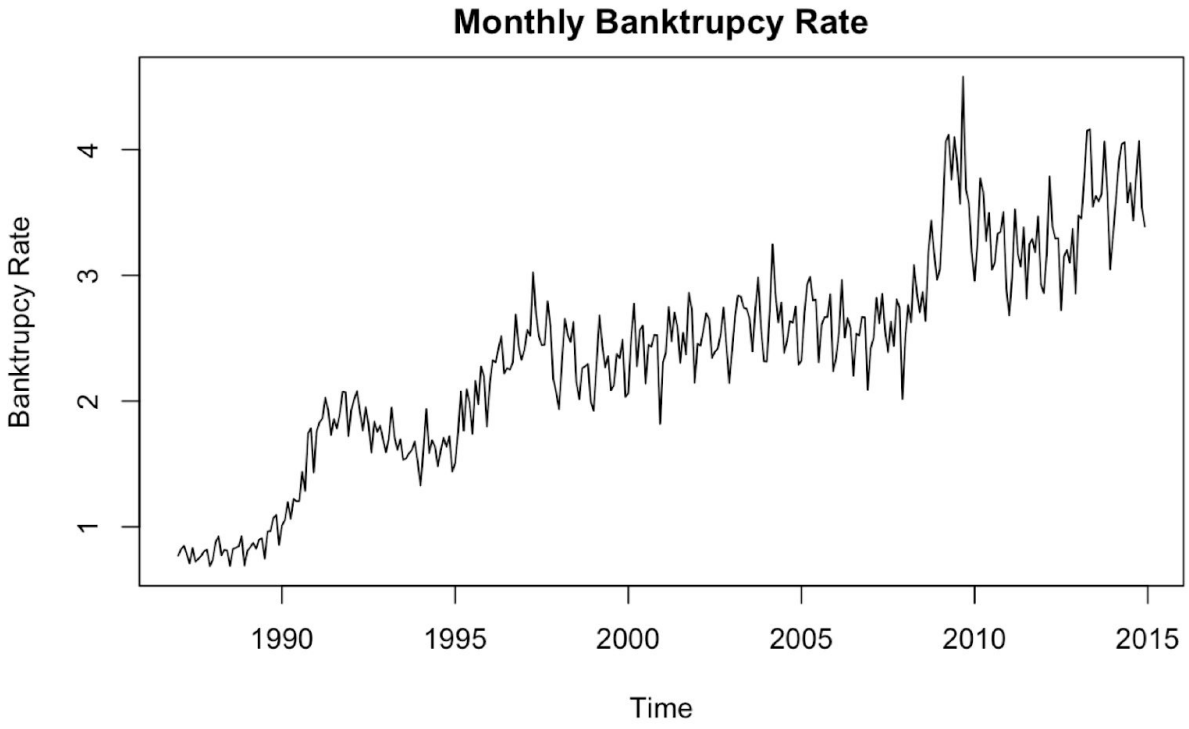


3. <u>Housing Price Index(HPI)</u>:

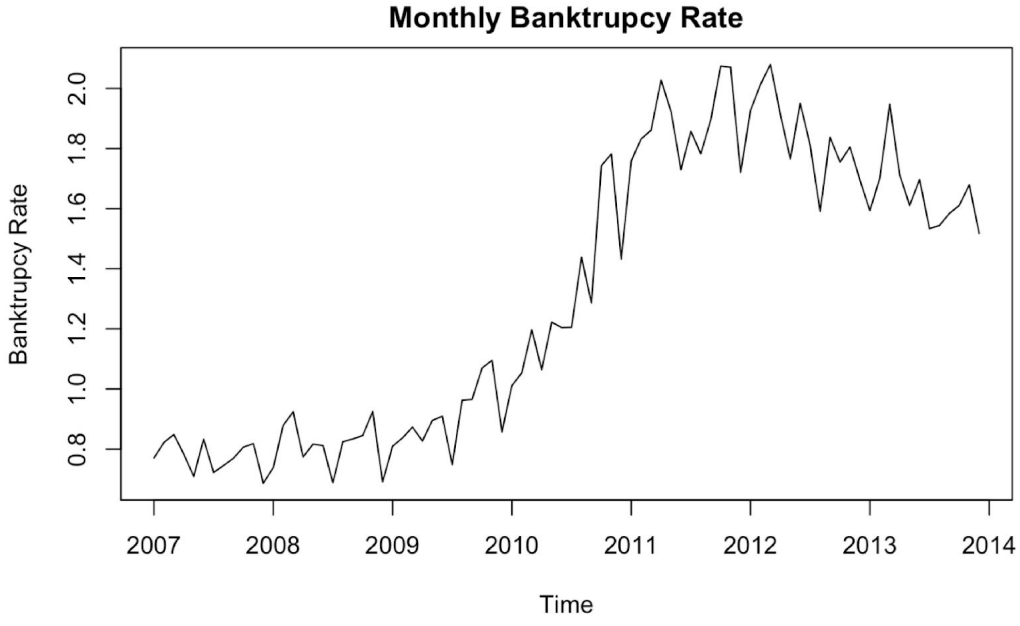
The HPI plot shows an overall increase in HPI, with two significant spikes around 1990 and 2008.

4. <u>Bankruptcy Rate:</u>

The graph for bankruptcy rate shows a trend for an overall increase.



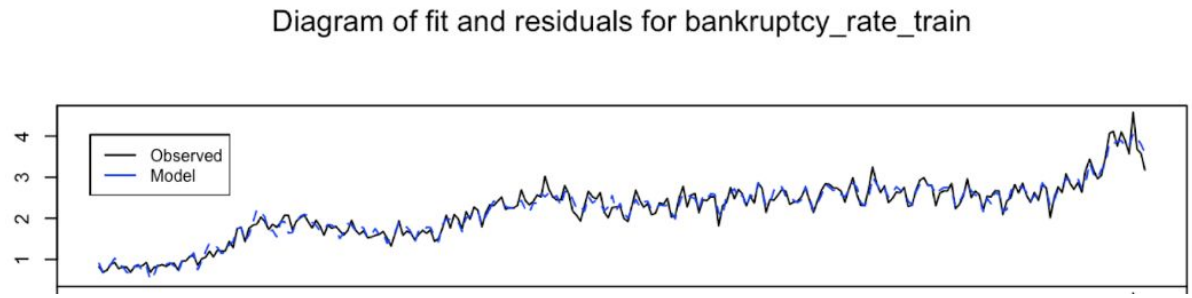
Zooming in on some of the periods, we can see that there is also a seasonal component for bankruptcy as there are consistent and regular fluctuations between each year.



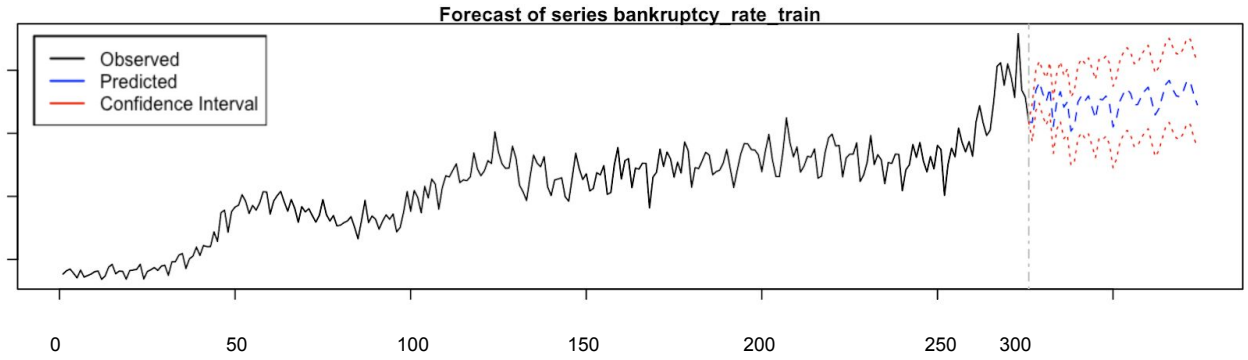
It can be observed from the graphs that there is a clear relationship between bankruptcy and HPI. We also thought that population and unemployment rate might influence bankruptcy rate. Therefore, in order to take into account these relationships, we decided to use a VAR model. Because we think that the relationship can be bidirectional, we avoided using a SARIMAX model.

We started by dividing our dataset into two parts - one part for training our model and one for testing it. The training data is from January 1987 to December 2009 and the test data is from January 2010 to December 2014. Doing this train-test split will be useful for evaluating how good our model is at predicting future values.

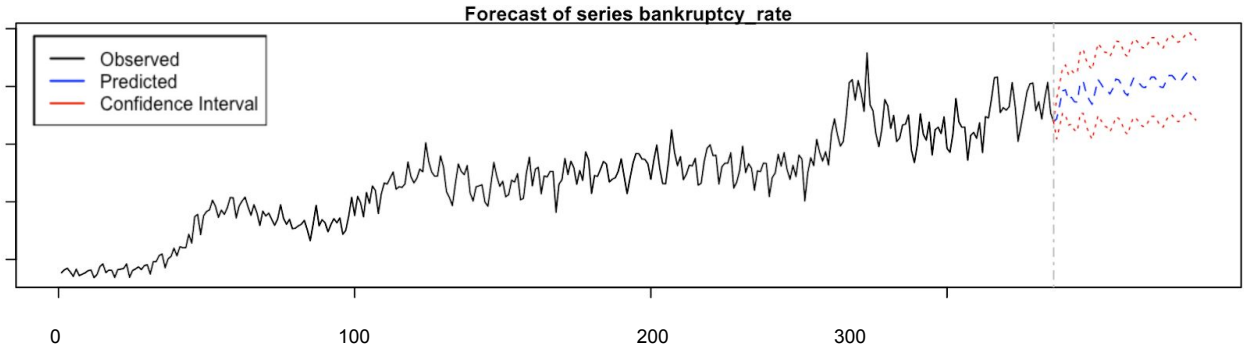
We selected the optimal parameter p for our VAR model using AIC as our goodness-of-fit metric. AIC is an estimator of how well the model fits the data while penalizing overly complex models. When comparing AICs, the smaller the better. The value of p that minimized the AIC was p=10, therefore VAR(10) is our optimal model. Below a graph of the model and the actual observations can be seen.



We then looked at the forecast of the model for the following five years: January 2009 to December 2014. A graph of the forecast of the model can be seen below:



We calculated the RMSE (root mean squared error) for our model by taking the square root of the squared difference between the observed values and the predicted values. A small RMSE indicates that the fitted values are close to the observed values, and that the model is doing a good job at forecasting. The RMSE of our model is 2.31. Considering that having an RMSE close to zero suggests that the predicted values are close to the observed values, we decided to pick this model to forecast the bankruptcy rate for 2015-2017.



One limitation to our final model is it assumes that the relationship between all the various is strictly endogenous. Given more time, we could have also explored a VARX model that accounts for both endogenous and exogenous variables. For example, it may be reasonable to presume unemployment rate and house price index are endogenous, while population is exogenous. However, based on our forecast and the evaluation metrics outlined, this prediction appears to be well suited for this time series.

Forecasting Results

Year.Month	Forecasted Bankruptcy Rate
2015.1	3.436381
2015.2	3.664566
2015.3	3.924636
2015.4	3.940424
2015.5	3.766818
2015.6	3.812675
2015.7	3.734753
2015.8	3.729439
2015.9	4.074070
2015.10	4.085480
2015.11	3.884437
2015.12	3.769002
2016.1	3.687864
2016.2	3.901591
2016.3	4.118680
2016.4	4.043425
2016.5	3.959534
2016.6	3.937971
2016.7	3.867620
2016.8	3.964096
2016.9	4.127475
2016.10	4.110607
2016.11	4.023133
2016.12	3.889011

2017.1	3.842000
2017.2	4.011735
2017.3	4.142800
2017.4	4.114115
2017.5	4.048495
2017.6	3.989217
2017.7	3.981134
2017.8	4.075215
2017.9	4.161459
2017.10	4.162317
2017.11	4.092829
2017.12	3.990638