Canadian Bankruptcy Rates Forecasting

Fiorella Tenorio, Prakhar Agrawal, Ran Huang, Nan Lin

DESCRIPTION OF THE PROBLEM

Is it possible to predict the future? If the December 21st, 2012 "end-of-the-world" and many other so-called Armageddon have taught us something is that no, it is not possible to predict the future. But it *is* possible to make an educated guess about it if you have enough historical information. We call this educated guess **forecasting**.

Now, our goal for this project is to forecast monthly bankruptcy rates for Canada, personal not for companies. Being clearer, what percent of people in the population will file for bankruptcy every month. Why this is important? Well if this rate increases a lot, this means that a lot of people will file for bankruptcy, this will have a direct and important impact in the economy, because it means that loans and debts will not get paid, this will have a direct impact for banks and the stock market, which, whether we like it or not, run every country's economy, so if they don't do well, it is bad for everyone. So, if we are caught off guard by an increase of this rate, the economy will suffer greatly, and it might take a country a lot to recover from it (remember the US market crash of 2007).

So, having the idea of what the bankruptcy rate will be in the following months is a good idea, then we can prepare better for it and make the changes necessary to prevent it or at least make the impact of it less harsh. Like we mentioned above, we need historical information, and we have information from January 1987 to December 2014. And the information we have available is:

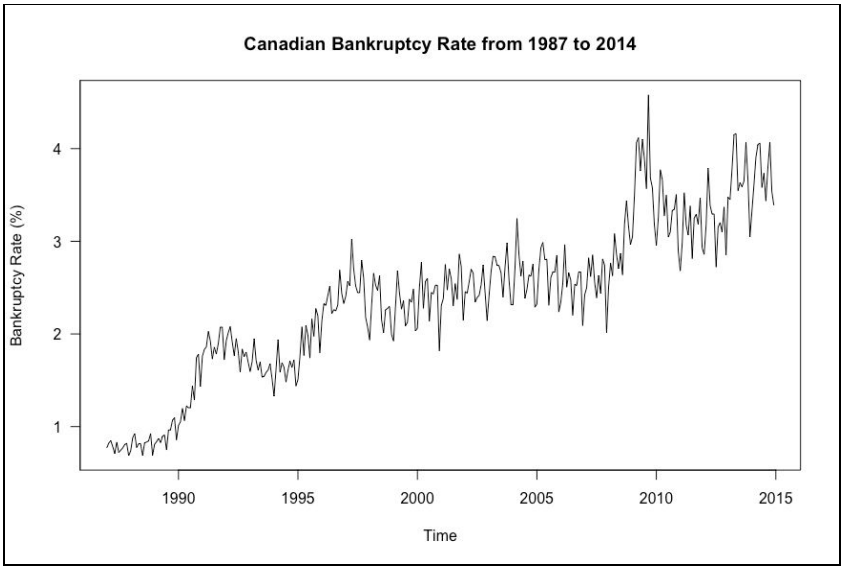
- **Unemployment Rate (%):** could the unemployment rate be related to the bankruptcy rate? This is one of the questions we will try to answer later on if knowing this can help us forecast the bankruptcy rate.
- **Population:** the increase or decrease in the population number could have an impact on the bankruptcy rate?
- **Bankruptcy Rate (%):** This is the variable we try to forecast, and we have previous values of it.
- **Housing Price Index:** could this index be related to the Canadian Bankruptcy rate? If the American market crash of 2007 thought us something is that yes, it could have a very strong impact on the rate, so we will analyze this variable and see if Canada is also impacted by it.

Now that we know what we have to work with and what we must accomplish, let's look at the data:

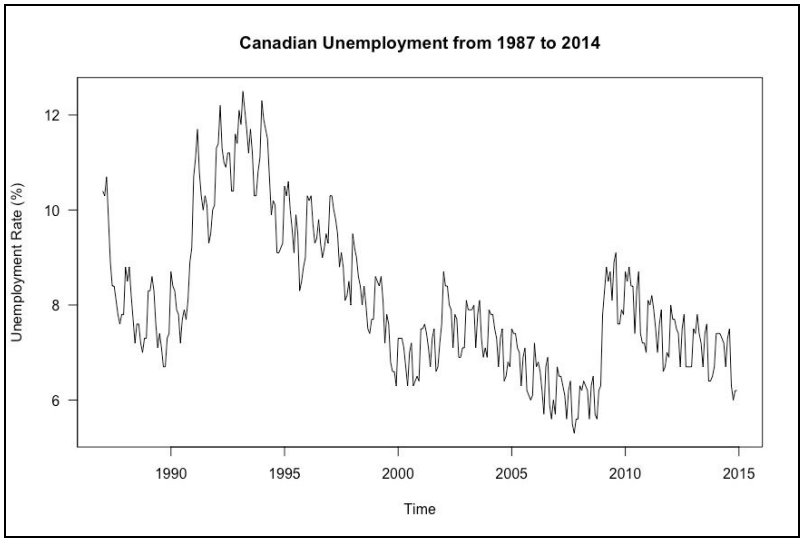
First, let's look at the bankruptcy rate: from the graph below, we can clearly see that the bankruptcy rate has been on the rise for quite a while, right from the beginning, the rate has increased a lot. In 2010, we can see a very pronounced spike, that hasn't repeated itself yet, but after it went back to its normal values, we can see that the values kept increasing. So, if we had only this, and someone asked us what we would think the bankruptcy rate would be like for 2016, what would we say? After looking at this graph, and looking at the tendency, it would be reasonable to say that the bankruptcy rate for 2016 will be greater than the one in 2015. And we can say this because we notice that the tendency is to increase. From now on, we will call the tendency: **trend**.

Another interesting thing we notice is that at the beginning, there seems to be less variation between the bankruptcy rates as we can see that at first, the spikes aren't so pronounced, they seem to be really close to each other. We call this variation between rates: **variance**. But as time goes by, we can see that the separation between rates increases as the spikes (upwards and downwards) seem to be a lot more pronounced than at the beginning. And we call this very pronounced variation between variances: **heteroscedasticity** (a big and scary word that, sadly, we must remember). The presence of this great variation of variances is something very important we will need to consider in our analysis to forecast.

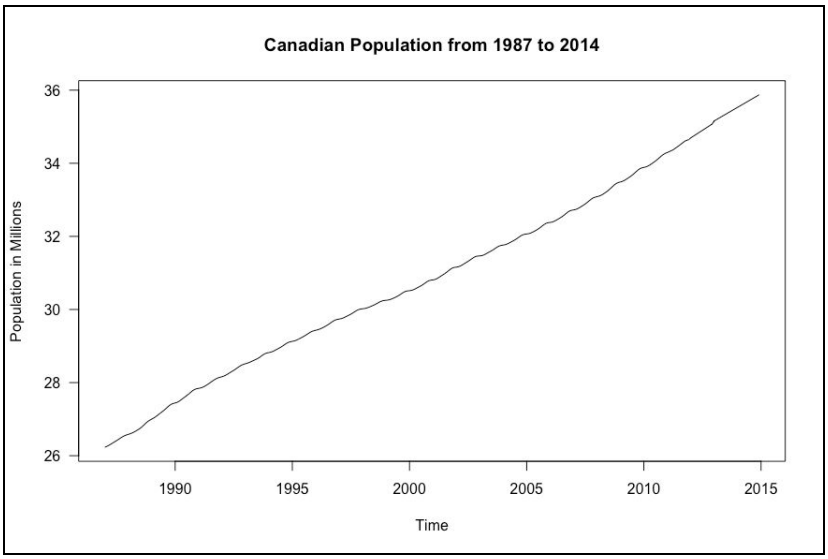
Another thing we notice, is that there seems to be a repeating pattern every certain period, we will call this **seasonality**, and we can understand better with an example about the weather, we know that every year the weather seasons repeat themselves almost at the same time, and we get almost the same temperatures, we call this repeating pattern weather seasons.



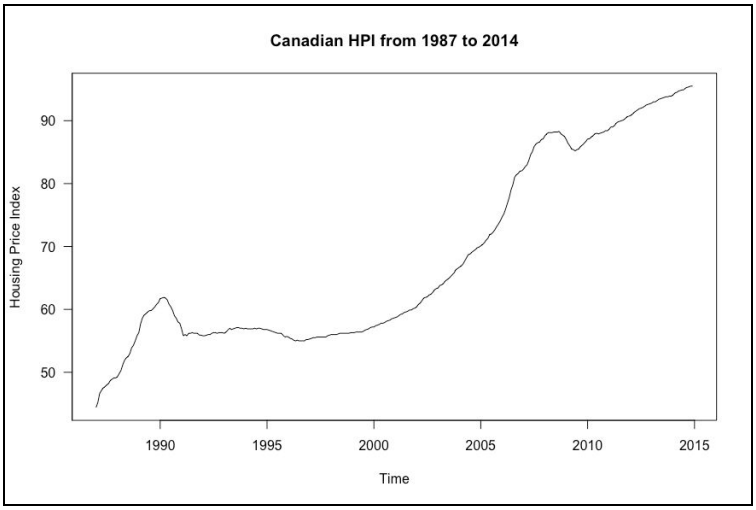
Now, let's take a look at the unemployment rate: from this graph, we can see that this rate has been decreasing almost consistently every month, we can still see some spikes every once in a while, but overall, this rate seems to be decreasing. There doesn't seem to be a very clear relationship between the bankruptcy and unemployment rate if there was I'd expect them both to increase/decrease almost identically (directly correlated), or as one increases, the other decreases (indirectly correlated). But that doesn't seem to be the case, but that's from an initial visual analysis; later, we'll perform a more sophisticated analysis to determine if a relationship between both variables exists or not.



Let's look at Population: From the graph below, we can see that the population in Canada has an increasing *trend* (let's remember this word from the beginning). This makes sense, as we can expect for every country's population to increase over time, it would be very weird to have a decreasing trend for population, or a very pronounced spike downwards (loss of population by some disease or war) or upwards (increase in births). From just a visual inspection, we could say that there isn't a very obvious relationship to the bankruptcy rate, but since the bankruptcy is calculated over the population, there might still be some relationship even if it isn't so obvious visually.



Finally, let's look at the Housing Price Index: Just like most of the previous variables, we can see an increasing trend. And we know that this index could have a direct impact in the economy of a country, so even if the relationship isn't quite obvious at first glance, we will examine it more thoroughly later to check if there is or isn't a relationship between this variable and the bankruptcy rate.



Now that we know some of the most important terms we'll use throughout the report, let's talk about the process we'll follow and how we'll do it.

We had mentioned before that our goal was to forecast the values of Canadian Bankruptcy Rate (which from now on we will call **dependent variable**) using past information we had available about that variable and some other. But how are we going to do this? We'll do this by trying to fit a mathematical formula that capture the behavior of the data, uses previous values of bankruptcy rate to predict future values of the same variable; the formula could also use some of the other variables if we find that they are related to the dependent variable and if their behavior can help explain it. This mathematical formula is formally called a **model**, and that's the name we'll use from now on.

For performance validation and model selection, we have available information from 1987 to 2014, we will split this data in train data and validation data. Based on 80/20 rule, we set 2008 as the threshold for training/test split.

Once we have our data split in these two groups, we will proceed to develop a model for the train data, and there are several different methodologies to do this (to develop a model). We will focus on 4: SARIMA, SARIMAX, Holt-Winters and VAR. In the following section, we will talk more about them and we will present the results we got from applying each methodology.

As we had mentioned before, we expect the model to be able to forecast the test data with some level of certainty (we know there will always be some error), since each model will be developed by a different methodology, we will end up choosing the model that predicts the test data with the smallest amount of error, and the way we'll measure error will be with a metric called **RMSE** (Root Mean Squared Error).

The RMSE is a metric calculated by the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Where: *Predicted:* is the value forecasted by our model, *Actual:* is the real value in our test data and *N:* is the number of observations in our test data.

As you would expect from looking at the above formula, this metric gives us an idea of how how our predicted values are, on average, from the actual values. Hence, we will use this metric as a way of assessing how good our model's predictions are.

MODEL SELECTION

We will proceed to talk about the 4 selected methods to develop the model.

1) SARIMA

This methodology is called SARIMA, which stands for Seasonal AutoRegressive Integrated Moving Average. And one of its assumptions is that data of the dependent variable must not have trend and/or seasonality, but we know from previous graphs that it does have trend and seasonality. That's what the *Integrated* component of SARIMA does, we will explain further in the following paragraphs. Once we fulfil this requirement for this methodology, we can proceed to develop an ARMA model. Now, let's break each of these words down:

S (Seasonal) - As we had mentioned in the previous section, it is possible for data to have seasonality (weather seasons example), so for this cases, we want our model to capture this seasonal component, to capture the fact that every certain period, there is a repeating pattern. The model will be able to capture the seasonal variations in the data. If the bankruptcy rates are high in February each year and low in December, then the model will take that into account.

AR (Autoregressive) - The model also considers the influence of past values on the present value. For example, the bankruptcy rate of last month, and possibly the month before that, could have an influence on the bankruptcy rate of this month. This could be a general trend across months, in the sense that each month's bankruptcy rate may depend strongly on the bankruptcy rate of last month and the month before that.We call this AutoRegressive because to predict the value of the variable we must consider its past values, but should we consider all of them? Usually, we expect that the values closest to the value we want to predict are more important than the ones further away. The number of past values we will use is p. In this case, it will be captured by an AR(p = 2) model, this means that to predict a value for bankruptcy we will use its last 2 values.

I (Integrated) - This is the component that will make sure that our data has no trend and seasonality, and we do this by *differencing*, this means taking the difference between observations. We can perform as many differences as necessary to eliminate the trend and seasonality from the data, we call a data with no trend and seasonality **stationary**. There are two types of differencing we could do:

- Ordinary: This means taking the difference between immediate observations. For example, we would take the difference between the value of March 2000 and February 2000, because they are one next to the other.
- Seasonal: This also means taking differences but, unlike before, we take the differences between observation separated m observations, where m is the period.

And our model will be developed using the results from this differences.

MA (Moving Average) - We know that the AR component captures the influence of the last values of the variable in future values, but what if the error from our predicted models also had an impact on future values? We capture this impact with the MA component, if the last q errors from our predictions influence a future value we will consider them in our model. For our model, after some analysis which we will

explain later, we concluded that the only the last 2 errors have an influence over future values of bankruptcy rate, so $q = 2$. We call that an MA($q = 2$) model.

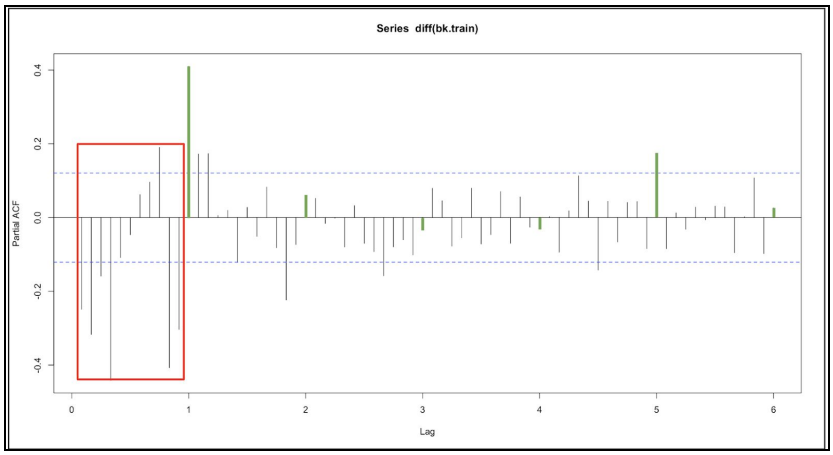
Again, as the name suggests, the SARIMA model can be used to model both the seasonal variations and the non-seasonal ones in the data. For that, we are going to need ACF (AutoCorrelation Function) and PACF (Partial AutoCorrelation Function) plots. Where AutoCorrelation means that we find the relationship between variables separated *h* lags. So, we can expect this autocorrelations to have values between -1 and 1, where the highest absolute value would mean that the observations separated *h lags* are highly related.

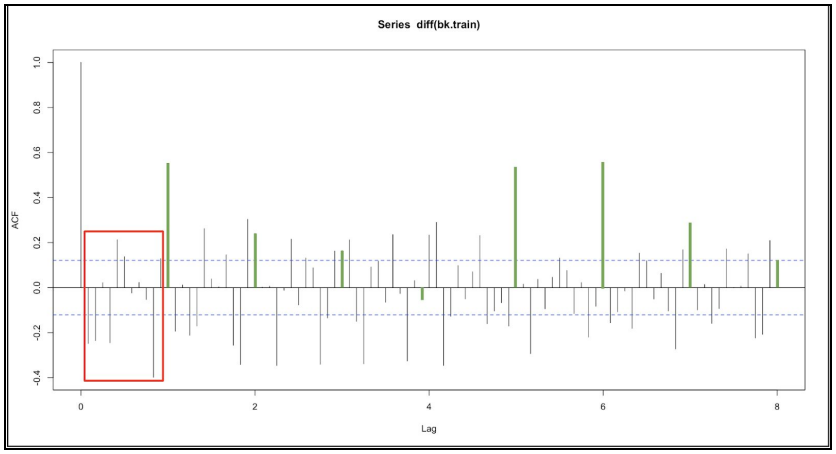
The difference between the ACF and PACF plot is that the PACF, while calculating the autocorrelation, takes into consideration the influence of intermediate values between lags. For example, if we calculate the autocorrelation between observation 1 and observation 5, it is possible that observations 2, 3 and 4 influence the value calculated, the PACF knows this and removes this effect from intermediate observations.

Hence, PACF models the relationship between the values themselves, while ACF is helpful for modelling relationship between prediction errors (as it has inadvertently taken that into account, by not excluding intermediate effects).

For our data, we'll be doing a log transform and then differencing it once (ordinally) before doing any modelling. In other words, we are modelling for the change in log of bankruptcy rate. The simple reason for that, is that it makes the model stationary, and it can then be modelled with relative ease.

Hence, let's start with making ACF and PACF plots for the differenced, log-transformed data:





From the above plots, we can discern certain patterns:

1) On the PACF plot, that focuses on the section highlighted in red. Clearly, correlation is high for lags of 1 to 4 months. This means that the current month's bankruptcy rate increase is heavily correlated with that of the last 4 months.
2) Let's now focus on the lines of the PACF plots, that are highlighted in green. These lines show us how the values are dependent across years. Clearly, correlation is high for bankruptcy rate increases 1 year apart, but not very high otherwise. This suggests that any given month's bankruptcy rate increase is dependent on how much the bankruptcy rate increased in that month, 1 year ago.
3) Similarly, the ACF plot suggests that the prediction errors for a given month depend on those from the past 5 months.
4) These errors depend on what the errors were in the corresponding months for the last 6 years.

Hence, we suspect that the model is a SARIMA(4,1,5)(1,0,6)[12].

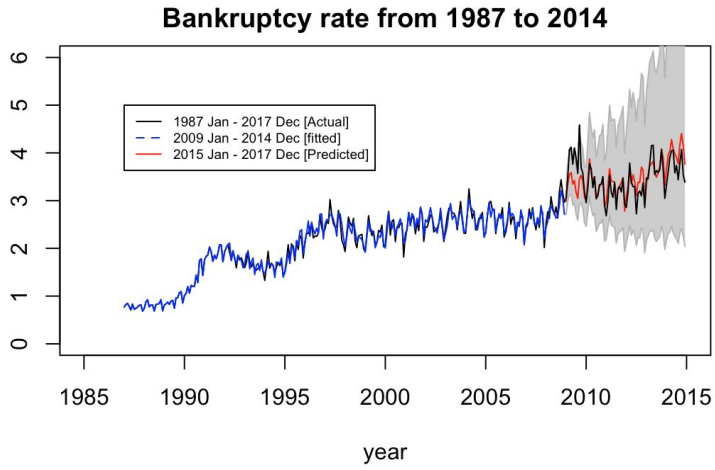
- Here the first 3 parameters (4,1,5) tell us how the "within season" time series behaves.
    - The 1 in the middle shows us how many levels of ordinary differencing we have to go through, before we fit this model.
    - Also as noted above, the values of the differenced data depend on last 4 observations and errors corresponding to it depend on last 5 errors.
- The next 4 parameters (1,0,6)[12] tell us about the seasonal pattern in the data.
    - The length of season here is 12 months, and among seasons, values depend on the last 1 month's values and errors depend on the last 6 months.

However, we suspect that the actual values of these parameters may be slightly different, and what we may be observing may be different by chance. Hence, we checked for various models with similar parameters to the one given above, and found that the following models perform best, in terms of RMSE:

Model	RMSE
SARIMA(2,1,4)(5,0,5)[12]	0.2997
SARIMA(3,1,3)(3,0,5)[12]	0.3013
SARIMA(4,1,4)(3,0,4)[12]	0.3043
SARIMA(3,1,4)(3,0,2)[12]	0.3052

The RMSE for the above purpose, was calculated by training the given model on 22 years of training data (i.e., 1987 - 2008), and then comparing its predictions on the validation data, i.e., the next 6 years (2009 - 2014), with the actual bankruptcy rates.

Going by the RMSE values, it looks like each of the above models is doing a pretty decent job at predicting the bankruptcy rate. Here's how the predictions look like for the best of these four models, i.e., SARIMA(2,1,4)(5,0,5)[12].



As can be seen in the above plot, our prediction is pretty accurate and very closely resembles the actual values. The uncertainty our prediction, however, which is shown here by the grey ribbons, is pretty high and as expected, it increases as we move farther and farther away from our training data. This is because the farther you are predicting into the future, the less information you have about current trends and hence, the more uncertain you are about your own predictions.

2) SARIMAX

Now, we know that SARIMA models are great, in the sense that they take a lot of the general characteristics of a time series, like seasonality etc., into account. However, there is only so much we can glean about the value of a variable from its past values. At some point, we also have to consider the effect other variables might have on the value of said variable.

For example, a time series based only on past values might not have been able to predict bankruptcy rates during the recession of 2009 correctly. However, these predictions could have been improved if they would have been based on not only the bankruptcy rates from the past month, but also on other factors that might affect it, such as Housing Pricing Index, unemployment rate, etc.

This is where SARIMAX comes in. It is a version of SARIMA that looks at the trend of values of not only the target variable, but also that of other variables that might be affecting it.

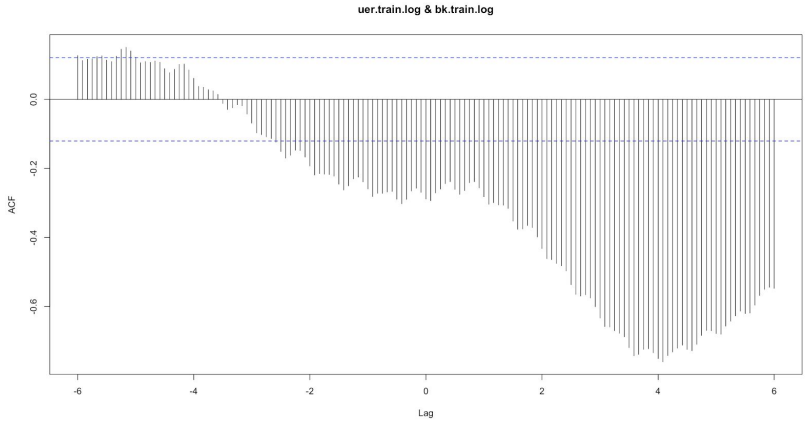
Fortunately, we had the data on 3 such variables, i.e., Unemployment Rate, Housing Price Index, Population.
We took each of the four best performing SARIMA models, and tried the following combinations of external factors, in order to see which one offers the best predictions:

1) Housing Price Index (HPI)
2) Unemployment Rate (UER)
3) Population (POP)
4) Housing Price Index (HPI) + Unemployment Rate (UER)
5) Housing Price Index (HPI) + Population (POP)

6) Population (POP) + Unemployment Rate (UER)
7) Housing Price Index (HPI) + Population (POP) + Unemployment Rate (UER)

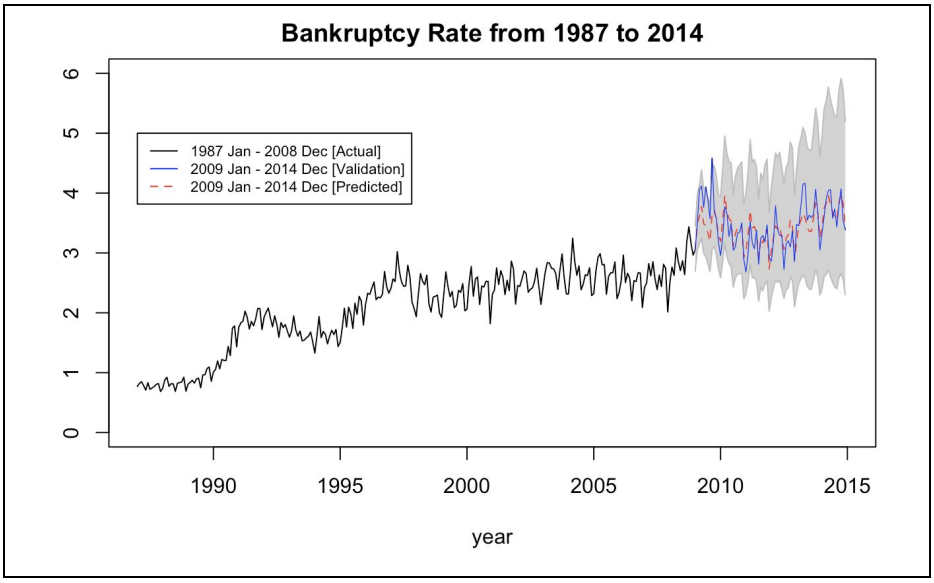
The best model to come from all these combinations, was SARIMAX(3,1,4)(3,0,2)[12](xreg = UER), i.e., SARIMA(3,1,4)(3,0,2)[12], with external regressor UER. The RMSE for this model is 0.2572, lower than the best SARIMA model gave in the last section.

Also, we looked at whether unemployment rate, with a lag of 4 years, i.e., unemployment rate from 4 years ago helps us predict the bankruptcy rate for today. This can be explained by looking at the cross-correlation function between unemployment rate and bankruptcy rate:



The strongest correlation between unemployment rate and bankruptcy rate happens at lag of 4 years, hence, it is possible that the unemployment rate from 4 years ago is dictating the bankruptcy rate this year. People who are declaring bankruptcy today, might be doing it because they might not have not been able to find work for the last 4 years.

Hence, we tried a few combinations of models with UER [lagged 4 years] as external regressor, none of which, however, improved the performance of the model. We selected SARIMAX(3,1,4)(3,0,2)[12](xreg = UER) as our final model. Here's what the predictions generated by this model look like, visually:



As you can see, our predictions for this data are now even better than what we got in through SARIMA models. This model has also predicted, at least to some extent, the increase in bankruptcy rate in 2009. This is because it also considers the value of unemployment rate when making predictions, not just the value of bankruptcy rates.

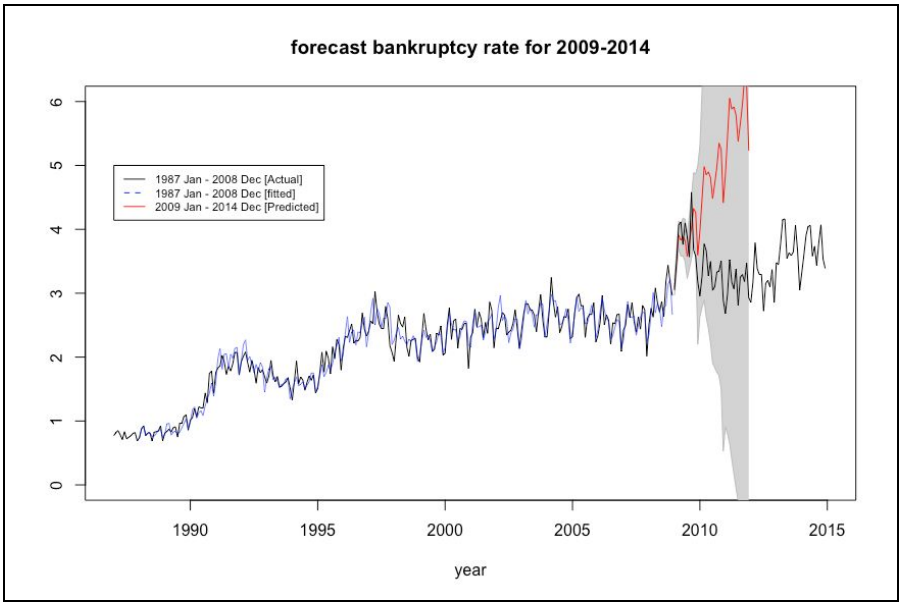
Also, it turns out that housing price index and population do not make our prediction better, and even make it worse in some cases. This suggests that the bankruptcy rates, at least in Canada, do not depend as much on these factors, instead depending on the unemployment rate in the country.

3) Holt-Winters Method

This method also called "Exponential Smoothing", is a technique used to detect significant changes in data by considering the most recent data. Keeping this in mind, this method is useful for making short-term forecasts, because it gives more "weight" (importance) to the most recent data. That's why it is only useful for the short-term forecast because if we use it for large data, at some point, we will forecast using forecasted values, this will add a lot of uncertainty to our predictions.

Even knowing this, we decided to try this method with our data, given that we were using a large window to test our predictions, we expected it not to give us a good result. Now, there are different ways to implement Holt-Winters methodology, and the one we chose for our project was "Triple Exponential Smoothing", and that's because, as we explained right at the beginning, the bankruptcy rate has trend and seasonality, so this method is appropriate for those situations. Also, since we have heteroscedasticity, we decided to use the multiplicative approach.

The picture below shows the results that we got for this method. The black line are the real values, and the blue line is the values that we forecasted. We can see that for the forecasted values for the test data, we get very different results, we can see that our prediction (forecast) is not accurate, it is very different from the real values. So, we can say that this method does not work for this data and we should continue to explore for a better method.

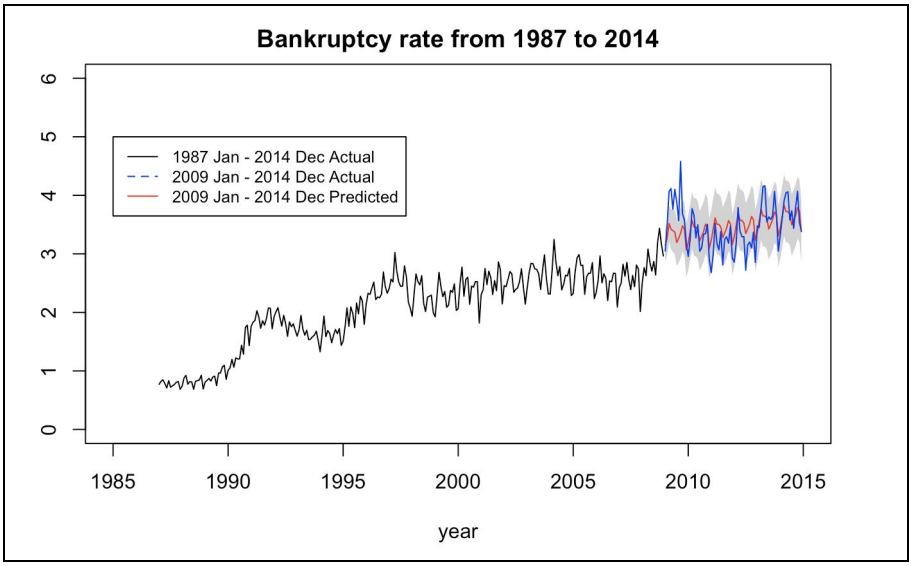


4) Vector Autoregression

In SARIMAX model, there is only one response variable, which is influenced by other variables but not is not influencing others. But in some cases, variables are influencing each other. For example, if we want to forecast daily number crimes committed in San Francisco, we may also need daily data on a number of SFPD officers on patrol when modeling. However, we should note that not only does the number of officers on patrol influence the number of crime, it is also influenced by the crime. Under this situation, we would call daily number of crimes as "response", and daily number of SFPD officers on patrol as an "endogenous variable". So, in our case, if we believe our response, the bankruptcy rate, is influenced by those variables as well as influencing them, then we should treat all other variables as endogenous variables.

What VAR does is to make all variables enter the model in the same way. Each variable will be the response once, having an equation explaining its evolution based on its past values, the past values of other variables, and an error term. A VAR model consists of these equations. In our case, we put all those variables into VAR model, including bankruptcy rate, which is our target, as well as other variables which are population, unemployment rate, and housing price index. Once we get the equations, we would only use the equation of which the response variable is the bankruptcy rate to forecast the future bankruptcy rate.

Note that there are many different VAR models we can choose because there are many options we can choose. First, we can choose the variables we want to put into the model. We do not necessarily put every variable into the model. Second, we can choose the lag terms. As we stated before, in each equation, the response is correlated to the lag terms of itself and also lag terms of other variables. So choosing an appropriate lag is also important. Third, whether specify the seasonality of the series or not also influence the results a lot. For example, we believe there is a seasonal effect in bankrupt series as well as other series where the period is 12. Using our metric RMSE on the validation set, we pick VAR with four variables, lag = 3 and season = 12 as the best model since this combination returns the lowest RMSE.



Final model

After comparing the RMSE on our validation data, we found that SARIMAX with external factor unemployment rate outperforms all the other models.

Hence, we are choosing SARIMAX as our final model for forecasting. Why SARIMAX will outperform SARIMA has already been explained earlier, but the fact that it outperforms the other three methodologies too, is probably not that surprising, given the nature of our data. Why?

1) Holt-Winters suffers from the same drawbacks as SARIMA does - it doesn't account for external variables, when, in our case, unemployment rate clearly appears to be influencing bankruptcy rate.
2) VAR is helpful if the variables in question influence each other, i.e., are endogenous. Since that is not the case here (bankruptcy is influenced by unemployment, but unemployment is not influenced as much by bankruptcy), consideration of such effects only adds noise to the data and makes the model worse.

Hence, we move ahead with SARIMAX[(3,1,4)*(3,0,2), xreg = uer], as it is the best model for our data.

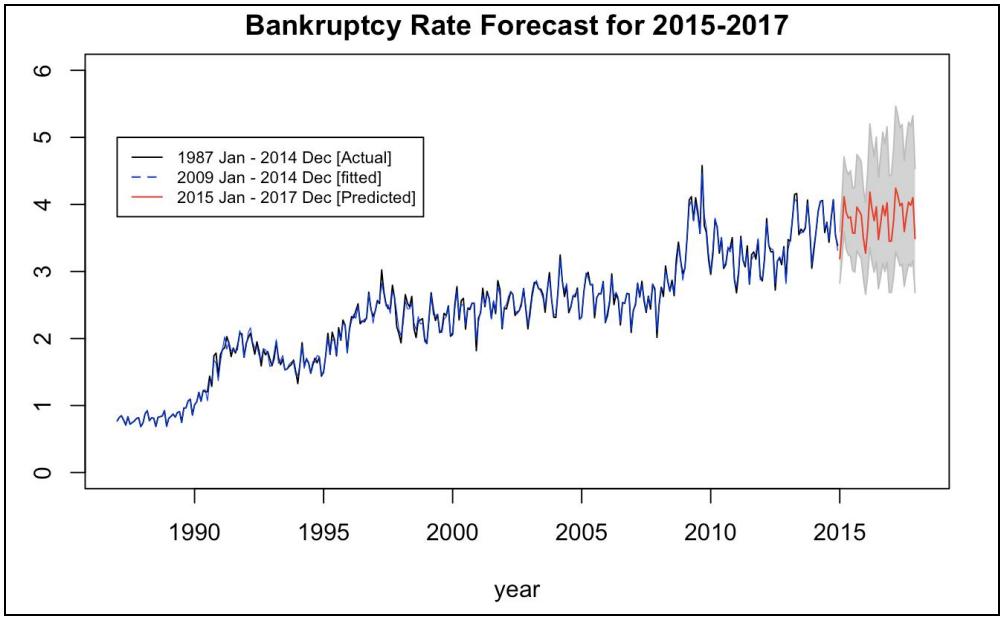
FORECASTING

1) Description

So far, we built our model on the training set from train.csv and assess the accuracy on the validation set from train.csv, which means our model is based on the data from 1987 to 2008 when selecting the optimal model. Once we have the optimal model, we need to build our model on the whole train.csv data and forecast bankruptcy rate from 2015 to 2017.

Table 1 Bankruptcy Rate Forecast for 2015-2017

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2015	3.188882	3.595684	4.113544	3.882103	3.797381	3.813474	3.574234	3.568759	3.960443	3.904386	3.843666	3.494487
2016	3.272762	3.638865	4.183524	3.928864	3.756102	3.964098	3.476172	3.718556	3.983535	3.830848	4.025643	3.448172
2017	3.454344	3.743357	4.241575	4.131406	3.97938	4.018088	3.595073	3.83967	4.034842	3.984844	4.100613	3.490025



The prediction for 2015 to 2017, along with a 95% confidence interval is displayed in the figure above. These intervals are the range of values between which we are confident that the true bankruptcy rate would be. The full table for 2015 to 2017 forecast can also be found above.

In order to forecast with a SARIMAX model, future values of the external variable, unemployment rate are needed. In other words, the prediction interval shown in the plot does not include the uncertainty of predicting unemployment rate values. Predictions of bankruptcy were calculated using the observed unemployment rate for 2005-2017. Taking this into account, generally speaking, Canadian bankruptcy rate would witness a slight uptrend with seasonal period 12 for the coming two years.

CONCLUSION

The final model uses the past value of Bankruptcy Rate of 1987-2014 and unemployment rate as a covariate. The unemployment rate has a positive influence on the bankruptcy rate, which means when the unemployment rate increase, the bankruptcy rate will increase. It aligns with our intuition that an increase in the unemployment rate would lead to an decrease in people income, which would, in turn, contribute to bankruptcy rates.

Our model is simple with high interpretability. But there are still limitations of our models. It does not include other important macroeconomic variables that might improve our forecast. Further, we can consider using Ensemble Methods, which can aggregate prediction values from multiple time series models to give a better prediction.

So in practice, if we want to forecast the bankruptcy rate in Canada, what we should do is gathering past bankruptcy rate data, past unemployment rate data, as well as future unemployment rate data. The future unemployment rate data can be gained by other source. Also, we acknowledge that this model may not be suitable for other countries since there are so many differences between countries.