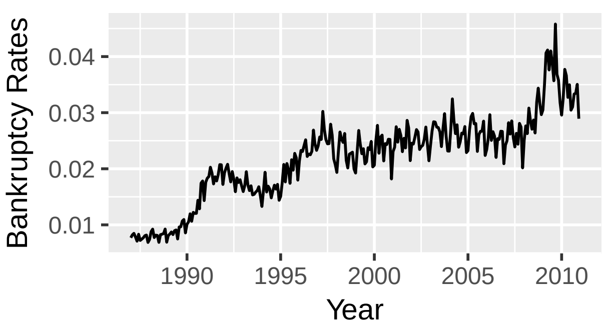
Introduction:

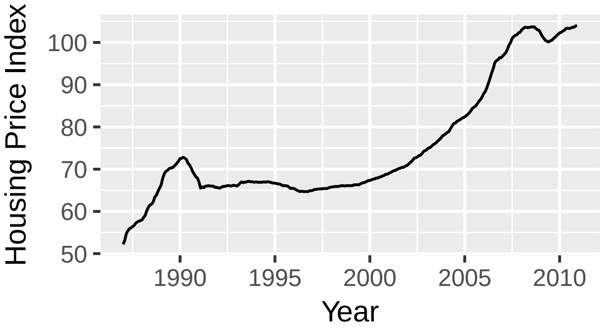
This project aims to build a predictive model to forecast monthly bankruptcy rates in Canada for the year of 2011 and 2012 with highest possible accuracy, given monthly data from January 1987 to December 2010 on bankruptcy rate, unemployment rate, population, and housing price index in Canada. then selected the optimal model and used it for forecasting.

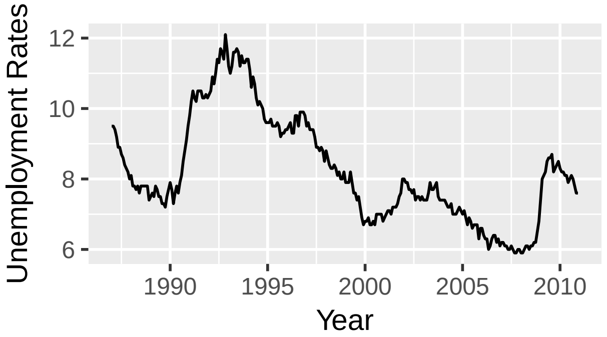
In this report, we will first explore the data, and discuss the available modeling approaches, including SARIMA, SARIMAX, Holt-Winters, and VAR. We will also explain the approach to select our best predictive model and present the forecasting results from our optimal model.

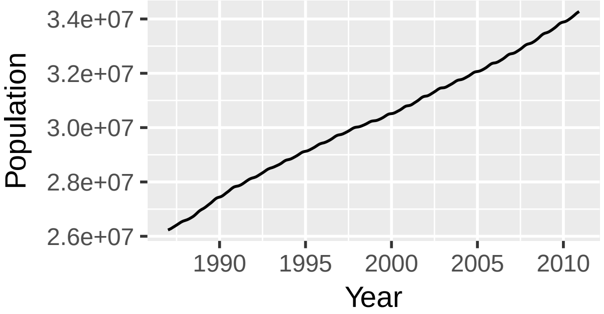
Data Overview:

The available dataset consists of monthly data from January 1987 to December 2010 on the 4 variables: bankruptcy rate, unemployment rate, population, and house price index. The four plots below correspond to each of the variable over time.

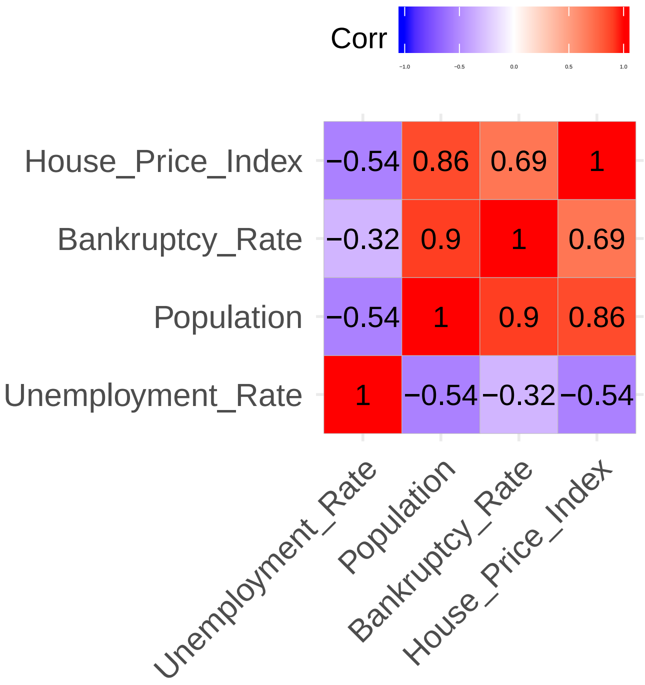


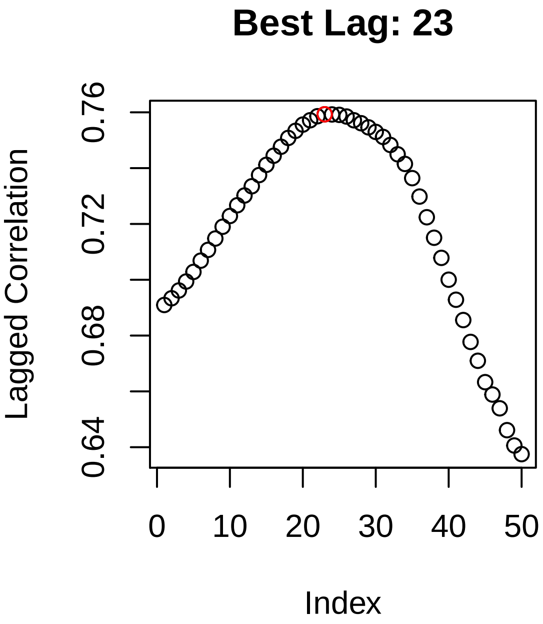






To explore the relationship between the 4 variables, we constructed a correlation matrix. We can see that bankruptcy rate is highly correlated with population and is somewhat correlated with house price index. Bankruptcy rate has a smaller, and negative, correlation with unemployment rate. Variables with medium to high correlation are of interest because they can possibly be used as covariates to help accurately predict bankruptcy rates.





More over, we ovserved from the individual time series plot that bankrupcy rate is probably has even a higher correlation with lagged values of `House_Price_Index`. So, here we plot the change in correlation between bankruptcy and housing index by different lagged values of housing index. Basically what we do here is to shift housing index to the right.

We found that highest correlation between housing index and bankruptcy happens to be with 23 lagged version of housing index. This means that there might be a pattern that bankruptcy follows from housing index after 23 months of occurence. Of course this is just a hypothetical assumptions which needs to be tried out. Hence, we can try out different lagged versions of housing index and use it as a covariate in out multivariate time series model.

Method

In order to find the best model to predict 2011-2012 Canadian bankruptcy rates, we split our training data into training and validation set. The training set consists of observations from January 1987 to December 2008, and is used for constructing the models. The last 2-year of data (48 observations) is held out to determine the predictive accuracy of the model.

We explored the following methods for modeling the bankruptcy rate:

- Box-Jenkins Methods: including ARIMA, SARIMA and SARIMAX, this approach works by removing the trend and seasonality through differencing the data and modeling the transformed data.

- Holt-Winters model: this approach works by assigning exponentially decreasing weights to older observations. Considering the trend and seasonality pattern observed in the data, we used triple exponential smoothing for modeling.

- VAR: this approach works by treating the other influential variables as endogenous variable - they influence bankruptcy rate and bankruptcy rate influences them.

For each method, the potential models are mainly compared on the basis of log likelihood, AIC and Root Mean Square Error (RMSE) on the held out data, an optimal model was selected based on its performance on the validation set. Another important point which shouldn't be forgotten is that one should be careful about not overfitting to the validation set. So, here we will also care about less complex models which will helps us avoid overfitting and as well as models that give good performance in validation set meaning that they generalize good enough to reflect the pattern on unseen data.

Box-Jenkins Methods

Box-Jenkin models involves statistical theory and modeling to analyze and forecast time series data. The naming standard for the various types of time series models consists of acronyms defined as the following:

- **S:** *Seasonal* effects - in our case, monthly
- **AR:** *Autoregressive* is a stochastic process in which future predictions are based on a weighted sum of previous observations
- **I:** *Integrated* involves ordinary differencing, or subtracting observations from the previous observation in time, to make a time series stationary (mean, variance, autocorrelation constant over time)
- **MA:** *Moving Average* is an average over many past observations
- **X:** e*X*ogenous variables are external variables that influence the response variable but the responsen does not influence them (Example: BART ridership may be affected by weather, but weather does not depend on BART ridership)

So, SARIMAX means *Seasonal Autoregressive Integrated Moving Average with Exogenous Variables*

In this part of the project we will explore SARIMA and SARIMAX models in order to predict bankruptcy rate. SARIMA models are univariate models which depend on previous observations and try to predict future values. SARIMAX on the other hand is a multivariate time series model which has the SARIMA component but also regresses on the given multivariate data at time t.

SARIMA MODEL (Univariate Bankruptcy)

As we can see from the plot of bankruptcy time series, the fluctuation in the data seems to increase over time, so we applied a log transform to our data to stablize the variance. In order to find the best fit models, we performed a grid search. For each SARIMA model, we calculated the RSME after forecasting 2 years ahead (24 data points). Then, for the 3 best models with lowest predictive errors, we formed log-likelihood ratio tests , which is a formal test for determining whether or not one model is better than another at a specified statistically significant level. We strive our models to be as simple as possible to avoid overfitting unseen data. The lower the number of parameters, the simpler the model will be.

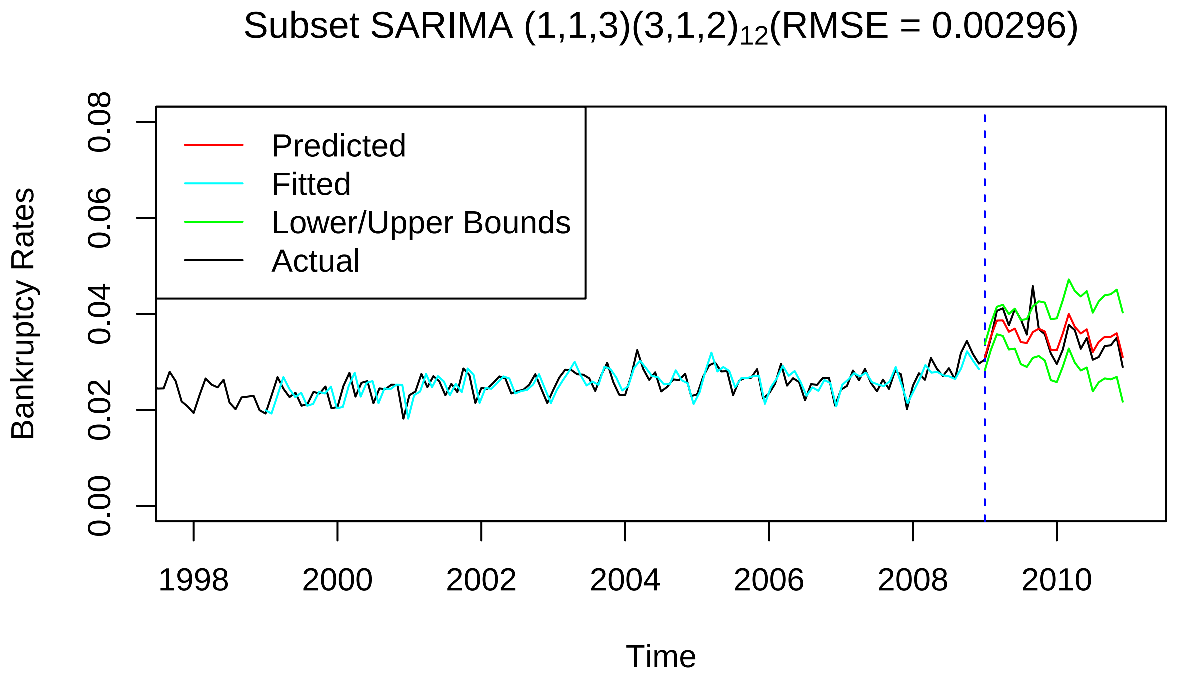
The best SARIMA model on the full data that we ended up choosing is SARIMA (0,1,3)(2,1,3)$_{12}$.

Subset SARIMA model

In this part, we will try out a hypothesis that our data has two different patterns, in other words it has a clear and instant pattern change around 1998 as we can observe in the plot. To overcome this instant change, we will take the time series that is after 10 years and use this subset of data in order to come up with a better predictive model. Anyway, our main goal here is to forecast the future values of the time series as well as possible. And our hypothesis is that having this subset will provide us a more generalizable model with better performance on unseen data. One drawback is that we can only compare this method in terms of RMSE with our previous models that used full data, and cannot use a formal test such as likelihood ratio test to compare.

We will take data starting December 1998, this was determined after several experiments. We didn't apply any transforms since data seems to have a constant variance over time. We perform another grid search and observe that best parameters are (1, 1, 3)(3, 1, 2)$_{12}$ with RMSE of ~ 0.0029 which is lower than our SARIMA model on the full data. Another important note is that since our main goal is to come up with the best predictive model we choose our optimization method as Least Squares Estimation (LSE) rather than Maximun Likelihood Estimation (MLE).

The plot below shows the fit of our best SARIMA model on the training set and prediction on validation set.



Checking Box-Jenkins Assumptions

Box-Jenkins models relies on the following assumptions regarding the residuals (difference between actual - predicted) for the models to be valid.

- Zero-Mean: the residuals have a mean of zero
- Homoscedasticity: the residuals have constant variance
- Zero-Correlation: the residuals are uncorrelated
- Normality: the residuals are normally distributed (no need to check for our subset model which is based on Least Squares rather than Maximum Likelihood Estimation)

Through formal hypothesis testing, both our SARIMA models satisfied the assumptions. However, SARIMA (1,1,3)(3,1,2)$_{12}$ model trained on the subset of data is on the borderline of passing the Zero-Correlation assumption test.

SARIMAX

In this part we will use exgenous data, assuming that there is a uni-directional relationship, meaning only independent variables effect bankruptcy not the other way around. We tried lagged 23 value of housing index, since it holds the highest correlation with bankcruptcy but it doesn't seem to perform better.

Again do a grid search over combinations of candidate sarima parameters. Our final best model is SARIMAX(1,1,2)(2,1,5)$_{12}$, with a rmse of ~0.00334 by using exogenous variables; population and house_price_index. Overall, this provides a better performance than all of the non-subsetted models with this extra information. We did try SARIMAX with our subsetted time series, but it did worse than the non-subsetted model.

Holt-Winters Methods

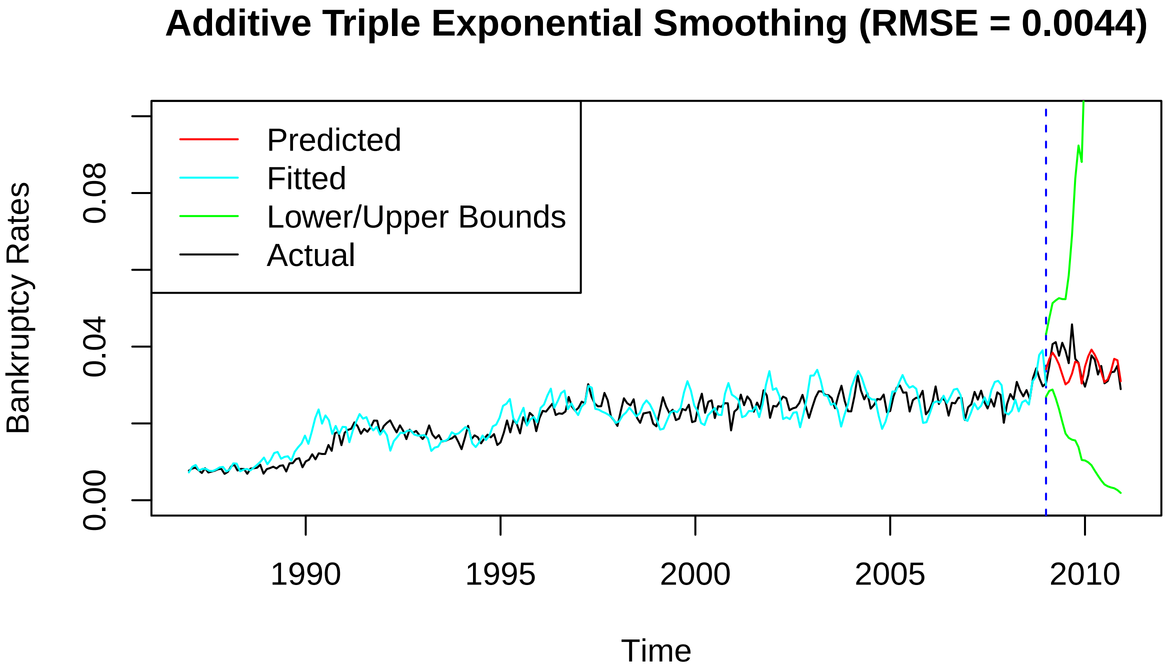
Holt-Winters Methods involves an exponentially weighted moving average. In the context of the Canadian bankruptcy rates, our model's predictions are based on averages of previously observed bankruptcy rates, with more weight on recent data. In other words, last month is a better indicator than say, 10 years ago. This makes sense because bankruptcy rates is most certainly going to change over time. Triple Exponential Smoothing is appropriate for forecasting monthly bankruptcy rates for Canada because there is both **trend** and **seasonality**.

Trend: A trend exists if there is a long-term increase or decrease in bankruptcy rates.

Seasonality: Seasonality is when the time series exhibits similar behavior at regular intervals, or *seasons*. In our scenario, bankruptcy rates are recorded every month, and therefore the period of a season is 12 months (1 year).

Furthermore, we choose to use an *additive* method because the size of the peaks are roughly the same throughout the time series. There are three parameters to be estimated: α, β, and γ. These parameters range from 0 to 1 inclusive. To decide the optimal values of the parameters, we used an iterative approach. For each of the iterations, we calculated the smallest RSME on the validation set, and decided to use the parameters with the lowest RSME on these data.

Our best model for Holt-Winters consists of $\alpha = 0.25, \beta = 0.65, \gamma = 0.35$.



Although this was our best Holt-Winters model, the prediction intervals are quite large. This means that although our point estimates are accurate, we do not have high confidence of or results. One advantage of Holt-Winters is that it does not depend on any distribution assumptions. For interpretability purposes, this method is fairly easy to understand because it just involves exponential smoothing over and over. A disadvantage of this model is that it is heavily dependent on the most recent data in the training set. Overall, in terms of RMSE, this Holt-Winter models are competitive with standard SARIMA and SARIMAX models, but the subsetted SARIMA model performs even better.

VAR - Vector Autoregression

Vector Autoregressive models are used for multivariate time series. The model assumes that the variables are endogenous - they influence each other, so in this model each variable is a linear function of past lags of itself and past lags of the other variables.

Since we have observed high correlation between bankrupcy rate and population and house price index, and we believe the influence between bankrupcy rate and both variables are bidirectional, we will fit a VAR model. We also observed some seasonality in the time series, so we added month as a exogenous variable. After iteration through different lag values, the best model with VAR method is VAR(4) model, with population and house price index as covariate, and month as exogenous variable. The RMSPE on the validation set is 0.004019.

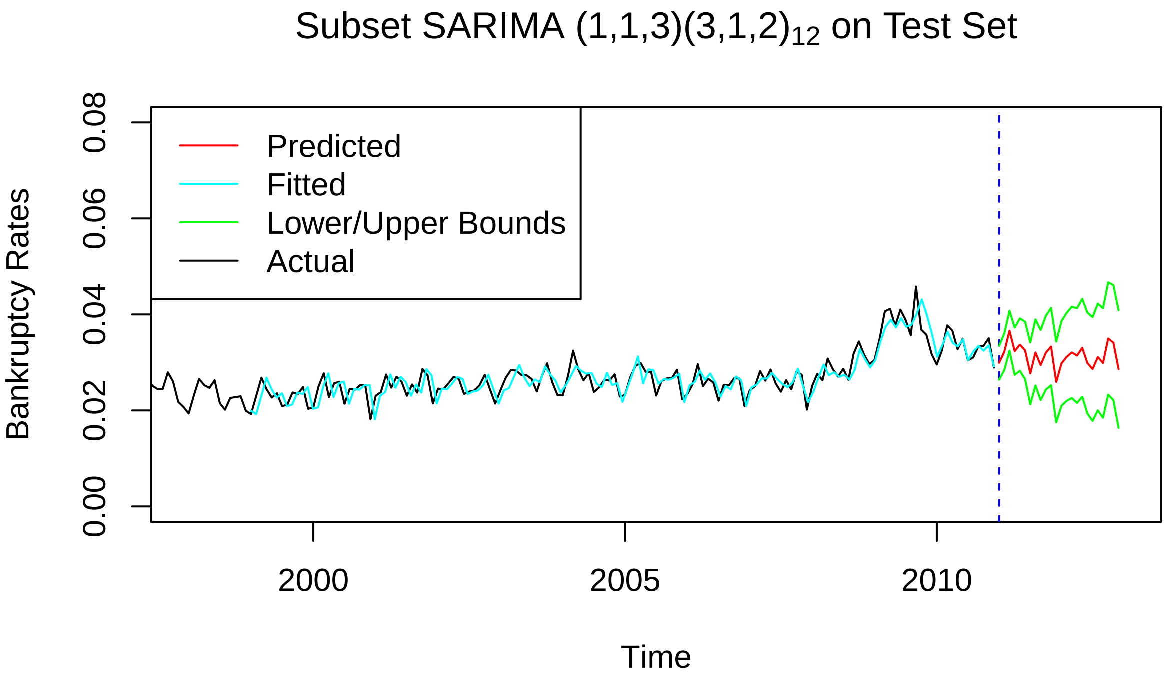
Conclusion

Comparison of RMSE on validation set for our models

Model	Root Mean Squared Error
Subset SARIMA (1,1,3)(3,1,2)$_{12}$	0.00296
SARIMAX(1,1,2)(2,1,5)$_{12}$	0.00334
Additive TES ($\alpha = 0.25, \beta = 0.65, \gamma = 0.35$)	0.0044
SARIMA (0,1,3)(2,1,3)$_{12}$	0.00372
VAR(4)	0.004019

In the end, we choose to select our best model based on the Root Mean Squared Error. In this case, it is our subsetted model SARIMA (0,1,3)(2,1,3)$_{12}$, where we only chose to use points from January 1999 to December 2008. This makes sense practically because the time series behaved differently prior to the year 1999. The assumptions needed for the model are fairly met, though the Zero-Correlation assumption is on the borderline, which isn't too bad.

Here are the predictions intervals of our final model SARIMA (0,1,3)(2,1,3)$_{12}$ on the unlabelled test set [2011-2012].



Predictions from January 2011 to December 2012

Point Forecast	Lower Bound	Upper Bound	Month
0.0299374	0.0264078	0.0334670	Jan 2011
0.0322459	0.0284316	0.0360602	Feb 2011
0.0365993	0.0324435	0.0407550	Mar 2011
0.0323550	0.0274241	0.0372859	Apr 2011
0.0337145	0.0282551	0.0391739	May 2011
0.0325095	0.0265299	0.0384891	Jun 2011
0.0277137	0.0212667	0.0341606	Jul 2011
0.0320847	0.0251989	0.0389705	Aug 2011
0.0294523	0.0221550	0.0367497	Sep 2011
0.0320374	0.0243503	0.0397246	Oct 2011
0.0333055	0.0252475	0.0413636	Nov 2011
0.0259192	0.0175065	0.0343319	Dec 2011
0.0297942	0.0209753	0.0386132	Jan 2012
0.0311612	0.0219919	0.0403305	Feb 2012
0.0320925	0.0225822	0.0416027	Mar 2012
0.0314290	0.0215727	0.0412854	Apr 2012
0.0330580	0.0228725	0.0432436	May 2012
0.0298720	0.0193660	0.0403780	Jun 2012
0.0286283	0.0178118	0.0394449	Jul 2012
0.0311412	0.0200227	0.0422596	Aug 2012
0.0298949	0.0184825	0.0413073	Sep 2012
0.0349903	0.0232913	0.0466893	Oct 2012
0.0341286	0.0221499	0.0461073	Nov 2012
0.0286110	0.0163590	0.0408630	Dec 2012