Final Project: Time Series

Felipe Chamma, Erica JH Lee, Miao Lu

December 14th, 2015

Introduction

The goal of this project is to forecast monthly bankruptcy rates in Canada for the year 2011, with the highest possible accuracy. To do so, several time series modeling techniques were employed, and the results will be described in the upcoming sections.

1 Initial Observation of the Data

Available Data

Initial data was made available from January 1987 until December 2010, which is a wide enough range for building our prediction model. We will call this the training data, as our model will be "trained" using these observations.

Here's a summary of it:

```
> head(train)
Month Unemployment_Rate Population Bankruptcy_Rate House_Price_Index
1 11987                9.5   26232423       0.0077004              52.2
2 21987                9.5   26254410       0.0082196              53.1
3 31987                9.4   26281420       0.0084851              54.7
.  ...                 ...      ...             ...                 ...

> summary(train[-1])
 Unemployment_Rate   Population        Bankruptcy_Rate    House_Price_Index
 Min.   : 5.900    Min.   :26232423   Min.   :0.006862   Min.   : 52.20
 1st Qu.: 7.175    1st Qu.:28511929   1st Qu.:0.017277   1st Qu.: 66.00
 Median : 7.900    Median :30248741   Median :0.023127   Median : 68.30
 Mean   : 8.236    Mean   :30256218   Mean   :0.021904   Mean   : 75.22
 3rd Qu.: 9.400    3rd Qu.:32059937   3rd Qu.:0.026620   3rd Qu.: 82.25
 Max.   :12.100    Max.   :34272214   Max.   :0.045798   Max.   :104.00
```

In terms of practical usage of this data, understanding the units of the variables is crucial. For example, we can see from the summary that the response variable we are interested in, `Bankruptcy`

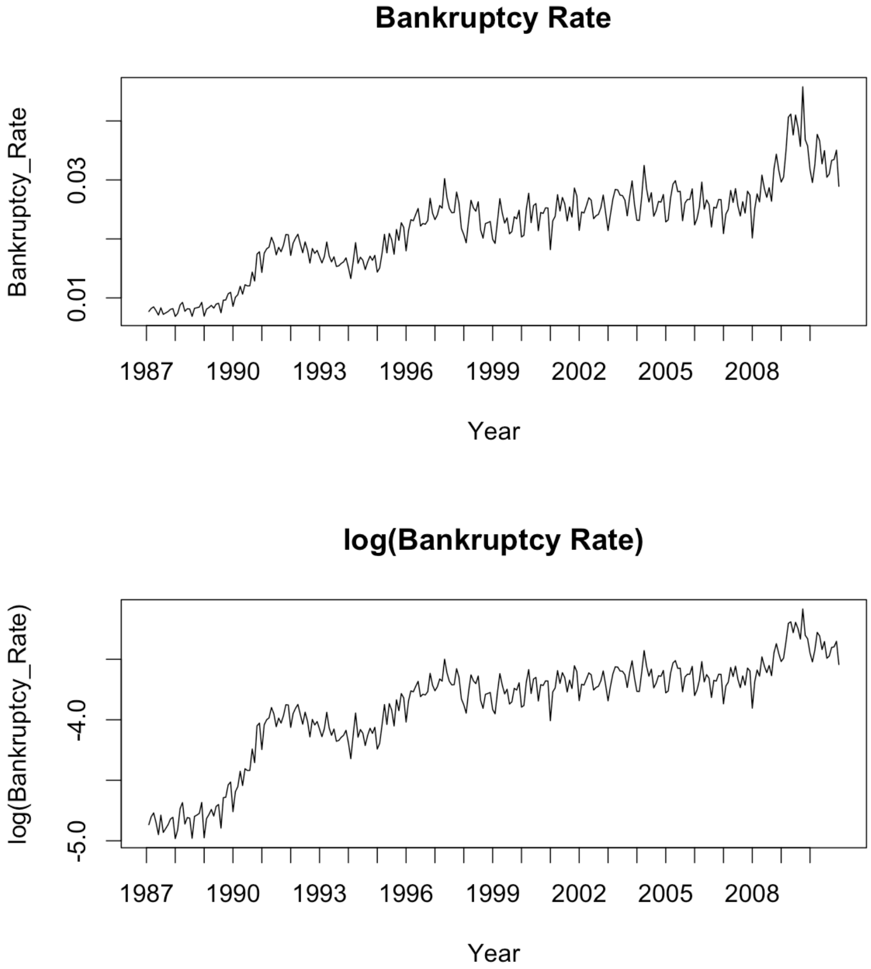
Rate, is a ratio ranging from 0.007 to 0.05.

An important insight here is among the independent variables (predictors): population ranges from 26 million to 34 million, while unemployment rate goes from 5.90 to 8.23 - and such a magnitude difference may result in a model biased towards one of the predictors.

To account for that we will normalize all the variables to the same scale, making sure the final model is not biased due to magnitude differences.

Log and Linear Transformation

The original time series plot show some heteroscedasticity - i.e., the variance is not constant over time. A log transformation was performed to account for this variability in the variance. Plot 1 shows how the variance was normalized (showing less signs of heteroscedasticity) after the log transformation:



Bankruptcy Rate

log(Bankruptcy Rate)

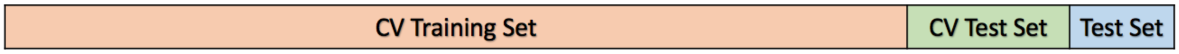
Figure 1: Log Transformation to Reduce the Heterscedasticity

Also, as mentioned in the last section, a linear transformation was performed for all other variables - using `scale()` function in `R`. This effectively centers the mean to 0, and normalize the variance to 1. When making predictions, this step will be reversed in order to show it in the original scale.

2 Model Selection Process

Subset for Cross Validation

There are different methods to select models. For this project **Box Jenkins Approach** was used to *identify* the parameters, *estimate* and *validate*. **Cross Validation** method was used to compare some of the best models.



In order to implement cross validation after selecting a few best models, 20% of the data was subsetted as a test set and the model was built on the 80% of the training set. Since the data is dependent on time, and the model will be used to predict the future, the last 20% of the data is subsetted out.

Ordinary Differencing

From the previous plot, we could clearly see that data was not stationary, and an overall upwards trend was identifiable. This consists on a violation for a proper time series modeling, and this trend must be accounted for before modeling.
An ordinary difference, which is taking the difference between an observation and its predecessor, will account for the trend and potentially make the series stationary. This result is seen in Figure 2:

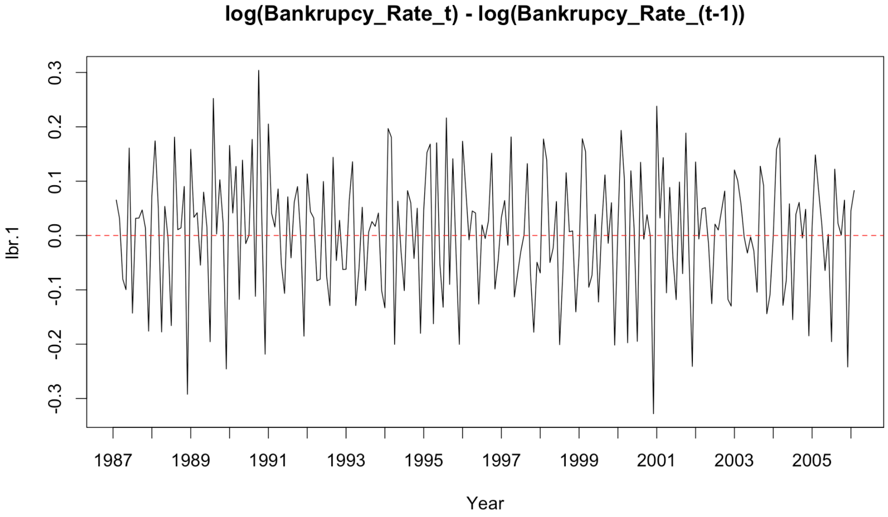


Figure 2: ACF/PACF Plot for One-Differenced Data

Visually, the upwards trend seems to have been removed. To formally confirm it, the Dickey-Fuller test was conducted and it turned out that the non-stationarity hypothesis for the data series may be rejected at a 0.01 confidence level.

Thus, trend has effectively been accounted for we may proceed with the modeling.

ACF and PACF plots 3 is showing some significant spikes outside the confidence interval for both Autocovariane and Partial-Autocovariance plots. Some exponential decay is detected as well. This signifies that the time series may be best modeled by an ARIMA (or SARIMA) model with parameters p, d, q, P, D, Q, s.

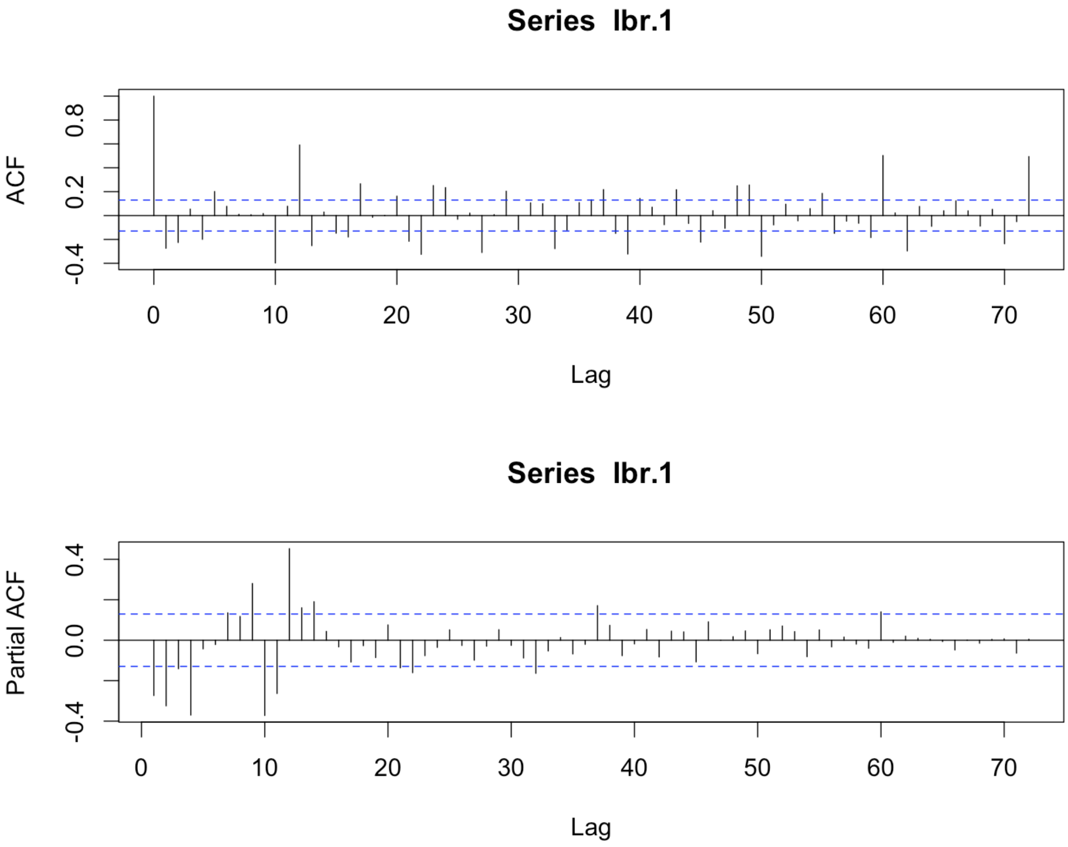


Figure 3: ACF/PACF Plot for One-Differenced Data

Seasonal Differencing

No obvious seasonality is shown in the form of sine waves in the ACF and PACF plots 3 (`nsdiffs()` function in R also suggested that no seasonal differencing is necessary, `D = 0`).

But there are still huge spikes at almost every 12 lags in the ACF plot. There is a huge spike at lag 12 in the PACF plot as well. It may indicate that seasonality exists and can be modeled well without seasonal differencing beforehand.

There is no right answer for that at this point, so we will try to fit both seasonal and non-seasonal models and see which performs best on cross validation.

Identification

When choosing the model, covariates were taken into consideration as well. It makes intuitive sense that Unemployment_Rate, Population and House_Price_Index provide more information for Bankruptcy_Rate prediction. From Pairs Plot 4, Population has a positive high correlation with Bankruptcy_Rate, while Unemployment_Rate is negatively correlated.

Model identification was rather unclear since the spikes in ACF and PACF plots are irregular. However, iterating through some potential $(p, d, q) \times (P, D, Q), s$ and covariate combinations for ARIMA/SARIMA, 5 candidate models were generated for further consideration. Model 5 is purposefully much simpler than the rest to avoid potential over-fitting.

- m_1: (2, 1, 3)×(1, 0, 1) s = 12; co-variate: Population, House_Price_Index
- m_2: (2, 1, 3)×(1, 0, 1) s = 12; co-variate: Unemployment_Rate, Population, House_Price_Index
- m_3: (3, 1, 3)×(2, 0, 3) s = 12; co-variate: Population, House_Price_Index
- m_4: (2, 1, 1)×(1, 0, 1) s = 12; co-variate: House_Price_Index
- m_5: (2, 1, 3)×(0, 0, 0) s = 0; co-variate: Population, House_Price_Index

Estimation

`Shapiro Wilk Test` for normality suggests that the time series data is not only stationary but also has normality. Thus, `ML` (Maximum Likelihood) approach could be used to acquire the identified parameters. This would allow us to make some inferences regarding the coefficients and residuals regarding the model.

Note that for Shapiro-Wilk Test, H_0: The sample (residuals of time series in this case) does not show a significant deviation from theoretical normal distribution. Thus, we fail to reject the null hypothesis and conclude that residuals are normally distributed:

```
          Shapiro-Wilk normality test
data:  lbr.1
W = 0.99042, p-value = 0.1351
```

Validation

As we can see from Figure 4, Model 1 and Model 2 have the best performing indicators when assessing the quality of fit:

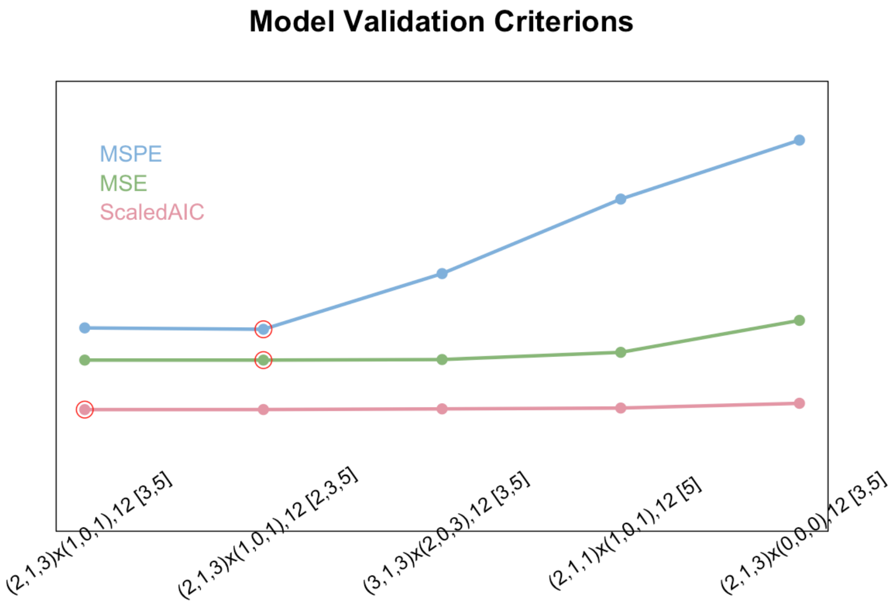


Figure 4: Selected Model for Different Criterions

Between those two, we know that Model 1 is more simple, with one less covariate. As a rule of thumb, the more austere model should be chosen in this case. Additionally, when we look at Model 2 coefficients and respective standard errors, we see that the *Unemployment_Rate* is rather insignificant. As shown in Figure 5, coefficient for `Unemployment_Rate` is not significantly different than zero. Thus, we will stick with Model 1 as be our best model for further analysis.

m_1: (2, 1, 3)×(1, 0, 1) s = 12; co-variate: Population, House_Price_Index

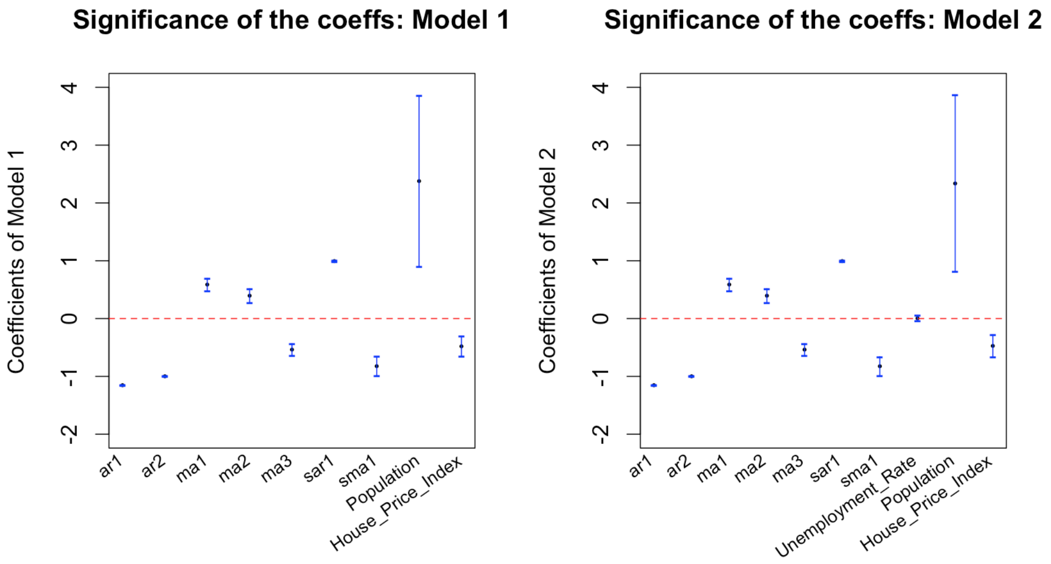


Figure 5: Estimated Coefficients and 95% C.I.

Residual Diagnostics

Residual diagnostic plots in figure 6 show that there are some significant autocovariance between the residuals (ACF: Residuals), and also some partial-autocovariance that is outside the confidence interval (PACF: Residuals).

However, none of the other identified models could completely remove the autocovariances among the residuals. Besides, the chosen model passes the tests for normality, homoscedasticity assumptions on the residuals. Those tests results are available at Table 2 as a reference. Thus, for further prediction of the future, we will proceed with the selected model.

Test	Statistic	p-value	Conclusion
t.test	-1.286	0.200	$\mathbb{E}[e_i] \approx 0$
adf.test	-5.945	0.010	$\{e_i\} \sim$ Stationary
Levene.test	1.976	0.098	$\mathrm{Var}[e_i] = \mathrm{c}$ (constant)
Shapiro.Wilk.test	0.993	0.242	$\{e_i\} \sim$ Norm

Ljung-Box test also shows no significance ($\alpha = 0.05$) for all lags < 11

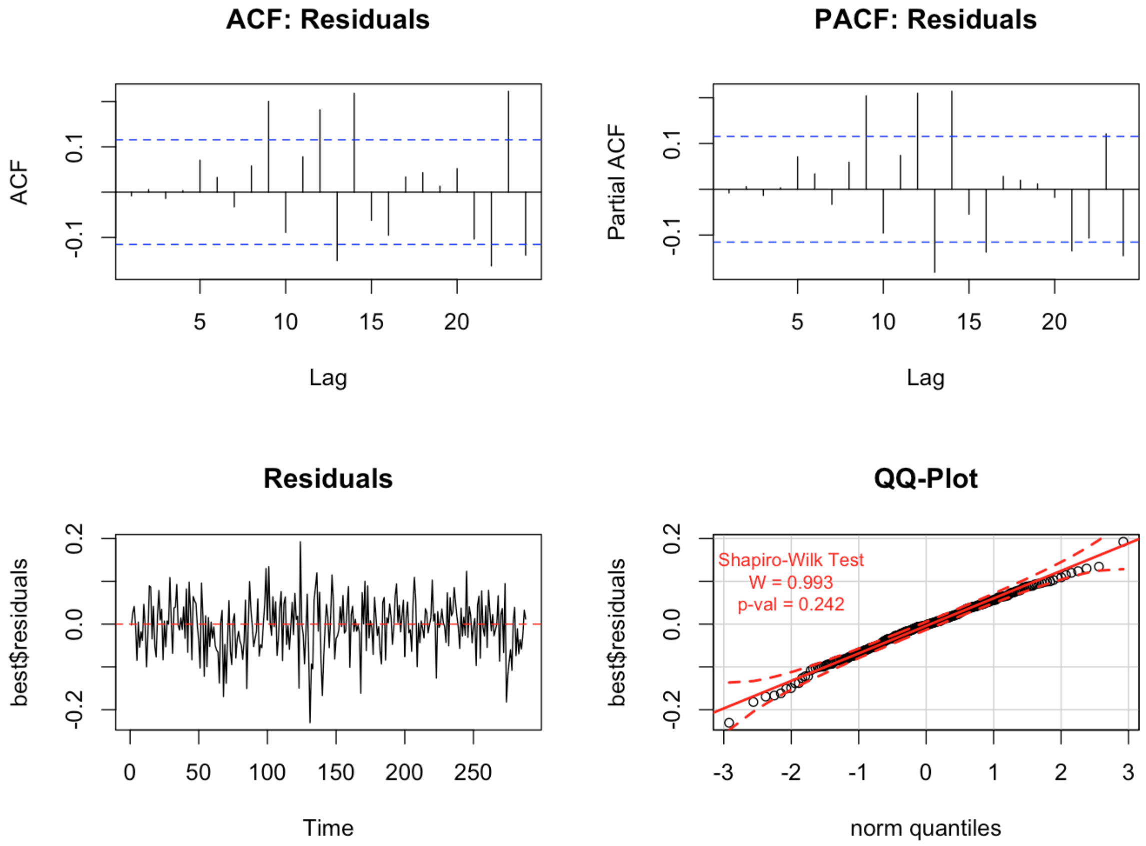


Figure 6: Residual Diagnostics

Potential ARCH/GARCH Model

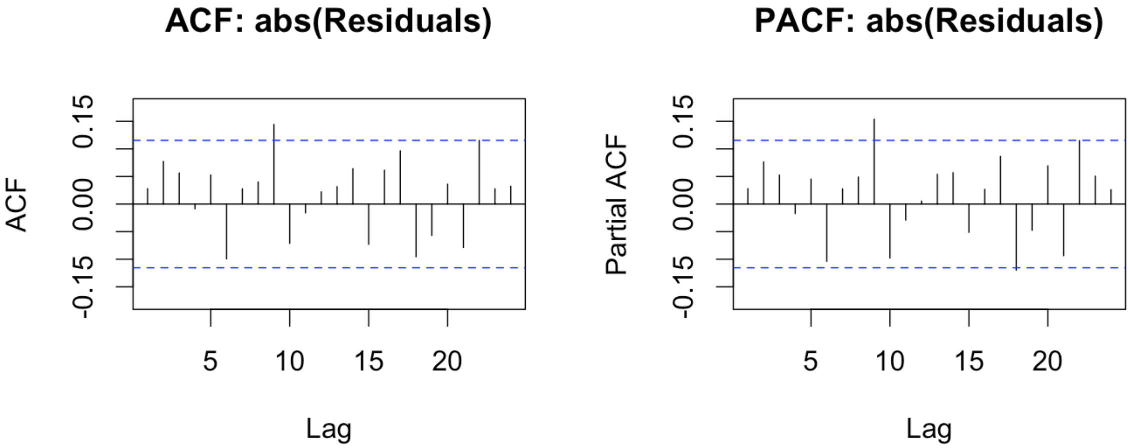


Figure 7: Absolute Value Residual Diagnostics

From Residuals Plot 6 there is no evidence of clusters with high volatility. Even for the ACF(abs(Residuals)) and PACF(abs(Residuals)) plots 7, there is no decay that is detected. There is a higher spike at 9, but nothing suggests there is a volatility trend behind it. Thus, there is no necessity to fit the residuals with ARCH/GARCH model right now, and ARIMA/SARIMA model will be good enough to capture all the information available.

Below are the coefficients for the chosen model m_1:(2, 1, 3)×(1, 0, 1), s = 12 with covariates as Population and House_Price_Index:

	ar1	ar2	ma1	ma2	ma3	sar1	sma1	Population	HPI
Coefficients	-1.02	-0.494	0.266	-0.154	0.114	0.997	-0.868	1.907	-0.402

3 Prediction

There are a few different ways to use the model to predict into the future, where the basic idea is to adjust the model coefficients by iterating through the training data. Some of the approaches include Fixed window (n-steps ahead), Rolling window, and Moving window. For this project, Rolling window method will be applied.

Rolling Window

The idea of Rolling window is to add each monthly Bankruptcy_Rate predicted back to the original data set and predict for the following one. An expanding data set is used for next prediction. It does 1-step ahead prediction in this case. It also fully uses all the Bankruptcy_Rate information available. Generally speaking, using more relevant information captures more signals and 1-step ahead gives a higher accuracy.

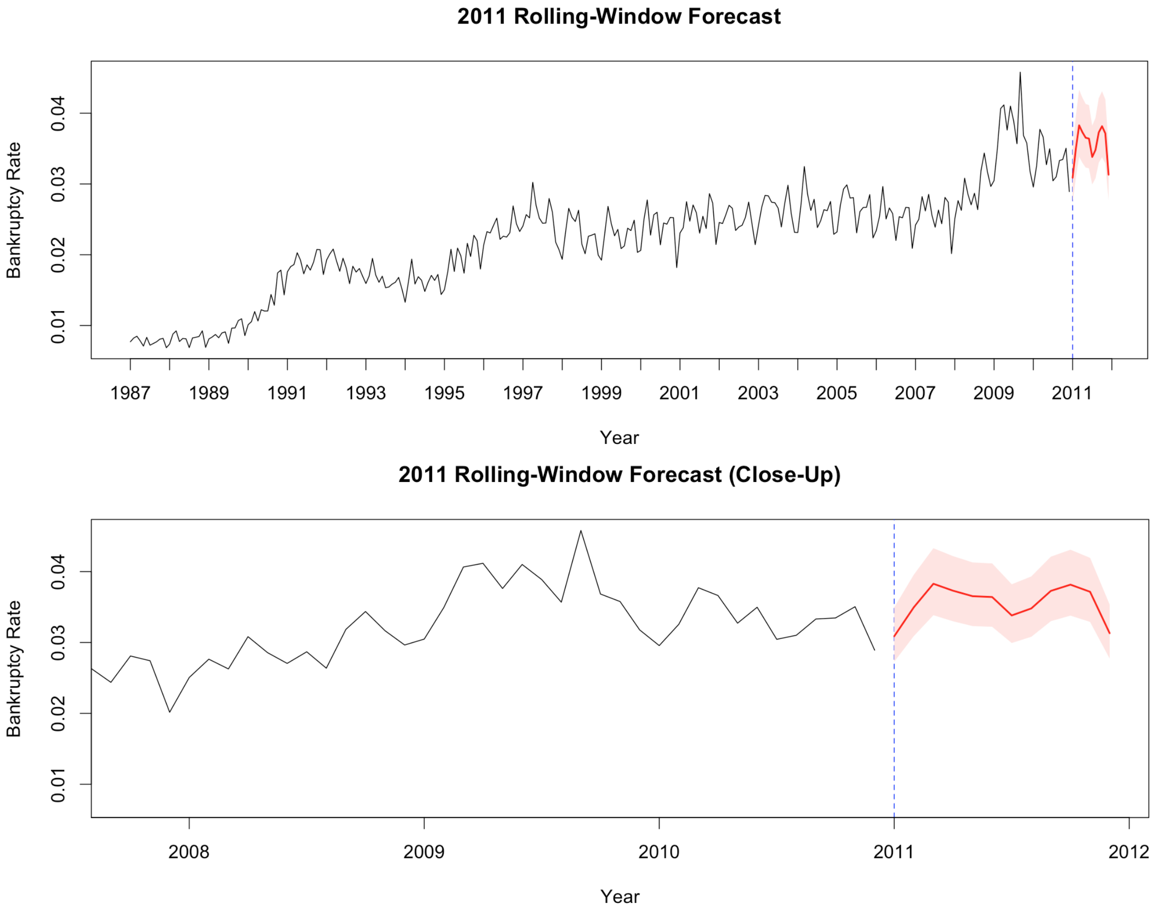


Figure 8: Prediction of 2011

```
             fitted      lower      upper
Jan-2011 0.03088119 0.02729115 0.03494347
Feb-2011 0.03496809 0.03091011 0.03955882
Mar-2011 0.03828502 0.03384986 0.04330130
Apr-2011 0.03730873 0.03299422 0.04218743
May-2011 0.03653256 0.03231508 0.04130047
Jun-2011 0.03641183 0.03221560 0.04115465
Jul-2011 0.03381739 0.02992682 0.03821375
Aug-2011 0.03480465 0.03080724 0.03932074
Sep-2011 0.03728328 0.03300852 0.04211164
Oct-2011 0.03815853 0.03379086 0.04309074
Nov-2011 0.03715153 0.03290640 0.04194431
Dec-2011 0.03130846 0.02773707 0.03533969
```

4 Conclusion

Making predictions about future values is a challenging matter, as we'll always be surrounded by the uncertainties of the unexpected. Good predictions have the right balance between confidence intervals that are not too wide that would include any naive guessing predictions nor too short that would fall apart after the first few observations.

In that sense, we believe our model seems good enough when looking at trends of past data, with reasonably wide confidence intervals. We can tell with a high confidence that Bankruptcy Rates in Canada will start 2011 going up a little bit, as it usually does on the first quarter of the year, and then move a little bit downwards to the end of the second quarter. That pattern is also observed for the second half of the year.

Appendix

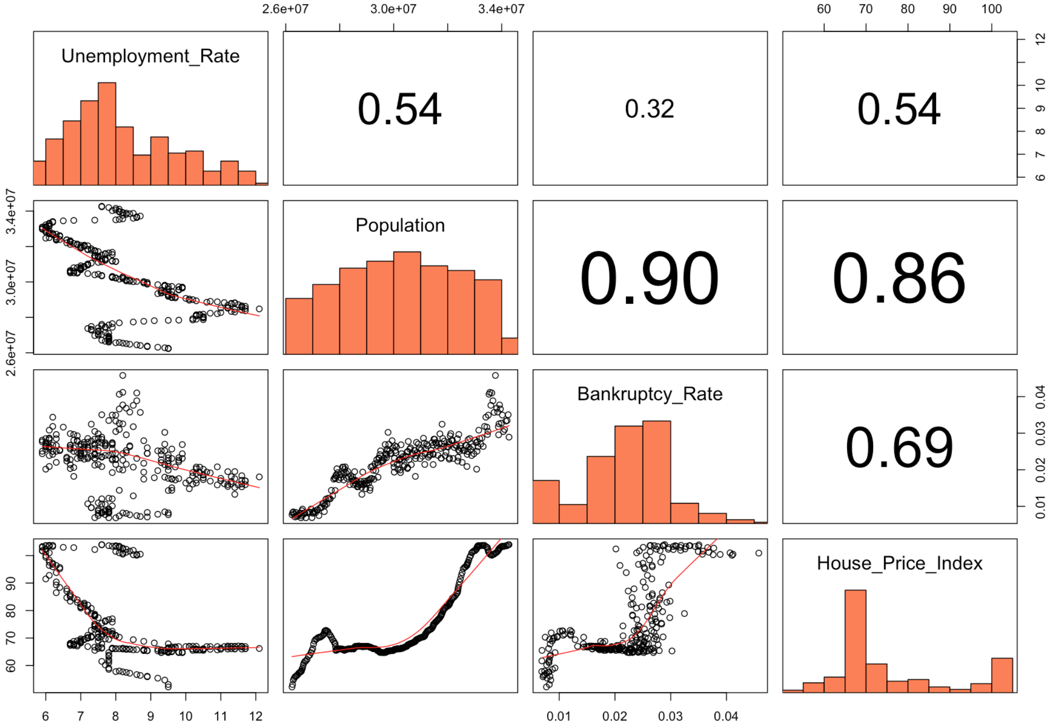


Figure 9: Pairs Plot: Relationship between Given variables