Time Series Final Project
Predicting Canadian Bankruptcy Rates

Paul Thompson, [redacted] and [redacted]

December 11, 2015

1 Executive Summary

Monthly Canadian bankruptcy rates from 1987 to 2010 were modeled as a time series using a SARIMA model. An optimal model was found though an exhaustive searching algorithm which fit a multitude of models and generated corresponding model assessment criteria (AIC, log likelihood and variance). From this result a handful of model were selected and were further evaluated by assessing MSPE when predicting bankruptcy rates in for 2010. An optimal model was selected and was used to forecast monthly bankruptcy rates for 2011. The predicted 2011 Canadian bankruptcy rates are as follows (see analysis for prediction intervals):

	Predicted Rate (%)
Jan	2.93
Feb	3.41
Mar	3.75
Apr	3.42
May	3.36
Jun	3.40
Jul	2.81
Aug	3.26
Sept	3.25
Oct	3.23
Nov	3.45
Dec	2.60

2 Introduction

Bankruptcy rates have steadily risen in Canada over the past 24 years, resulting in net increase of 2.1% (a nearly 3-fold change) in the percentage of Canadians who have filed for bankruptcy monthly. Bankruptcy can have many negative economic impacts, such as adverse effects on ones credit scores and interest rates, and often leaves individuals or companies in a state of limited financial capacity for several years. Various financial institutions have a large interest in predicting future bankruptcy rates.

3 Methods

We wish to model these bankruptcy rates as a time series, specifically by employing Box-Jenkins methods to determine an optimal model for this data and to forecast future bankruptcy rate values. Box-Jenkins methods apply autoregressive moving average (ARMA or ARIMA) models to find the best fit of a time series model using past values. In general an ARIMA model is used when a time series exhibits little

to no seasonal pattern (only a trend component), and a SARIMA (seasonal autoregressive integrated moving average) model is used when a time series has a clear seasonal trend which must be accounted for. These models are denoted ARIMA(p, d, q) and SARIMA(p, d, q) $(P, D, Q)_s$, where p is the order of the autoregressive model, d is the degree of differencing, and q is the order of the moving average model. In the case of a SARIMA models we also have seasonal analogs of these parameters, P, D and Q. Here P, D and Q have the same meaning as p, d and q, the only difference being that the seasonal component of the model is implemented with period s, as opposed to a period of 1 which is used in the trend component of the model.

Specifically, we will use the box-jenkins approach for forecasting as follows:
1. Plot the data.
2. Check for non-constant variation and apply box-cox transformation as needed to make variability constant.
3. Use regular and seasonal differencing if there is evidence of trend and seasonality.
4. Use the ACF and PACF of the transformed and/or differenced data to choose the parameters p, q, P, Q for a SARIMA model.
5. Fit proposed model and iterate to until an optimal model is reached based on criteria such as aic, σ^2 and log-likelihood.
6. Check residual assumptions for chosen optimal model.
7. Use model to forecast into the future.

4 Analysis

4.0.1 Model Identification

Figure 1 shows the original and the differenced log-transformed time series of the bankruptcy rates in Canada from January 1987 to December 2010.

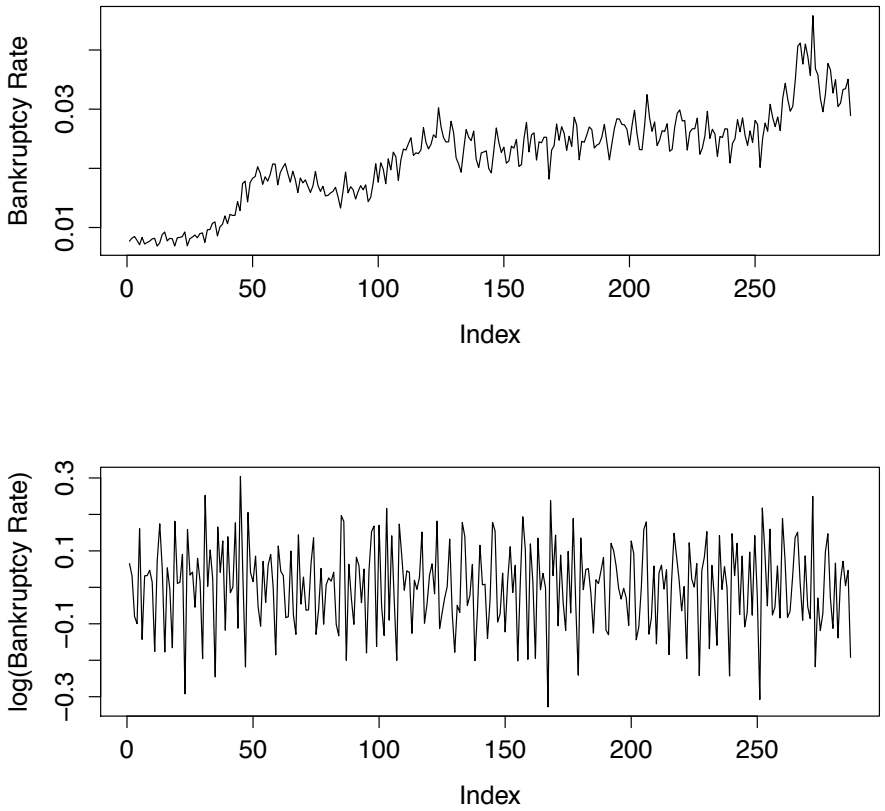

Figure 1: Plots of the original time series (top) and the differenced log time series (bottom)

There appears to be trend and possibly seasonality present in the data. Furthermore, there also appears to be heteroscedasticity present in the data. This suggests that we should difference the data, and take a log transform. After taking this transformation and differencing, we obtain the second time series shown in Figure 1 (lower plot). This time series appears to be much more homoscedastic and also stationary. This homoscedasticity is confirmed by performing a Levene Test, for which we obtain a p-value of 0.95, indicating that we do not reject the null hypothesis that the data are homoscedastic. The stationarity of the log-differenced time series is confirmed by the augmented Dickey-Fuller test. The null-hypothesis that the log differenced data isn't stationary was rejected at the $\alpha = 1\%$ level.

While we ultimately used an exhaustive search algorithm in order to choose an optimal model, we started by looking at the ACF and PACF plots. This helped guide in what ranges we searched p, q, P, Q. The plots also helped with determining if there is seasonality present.

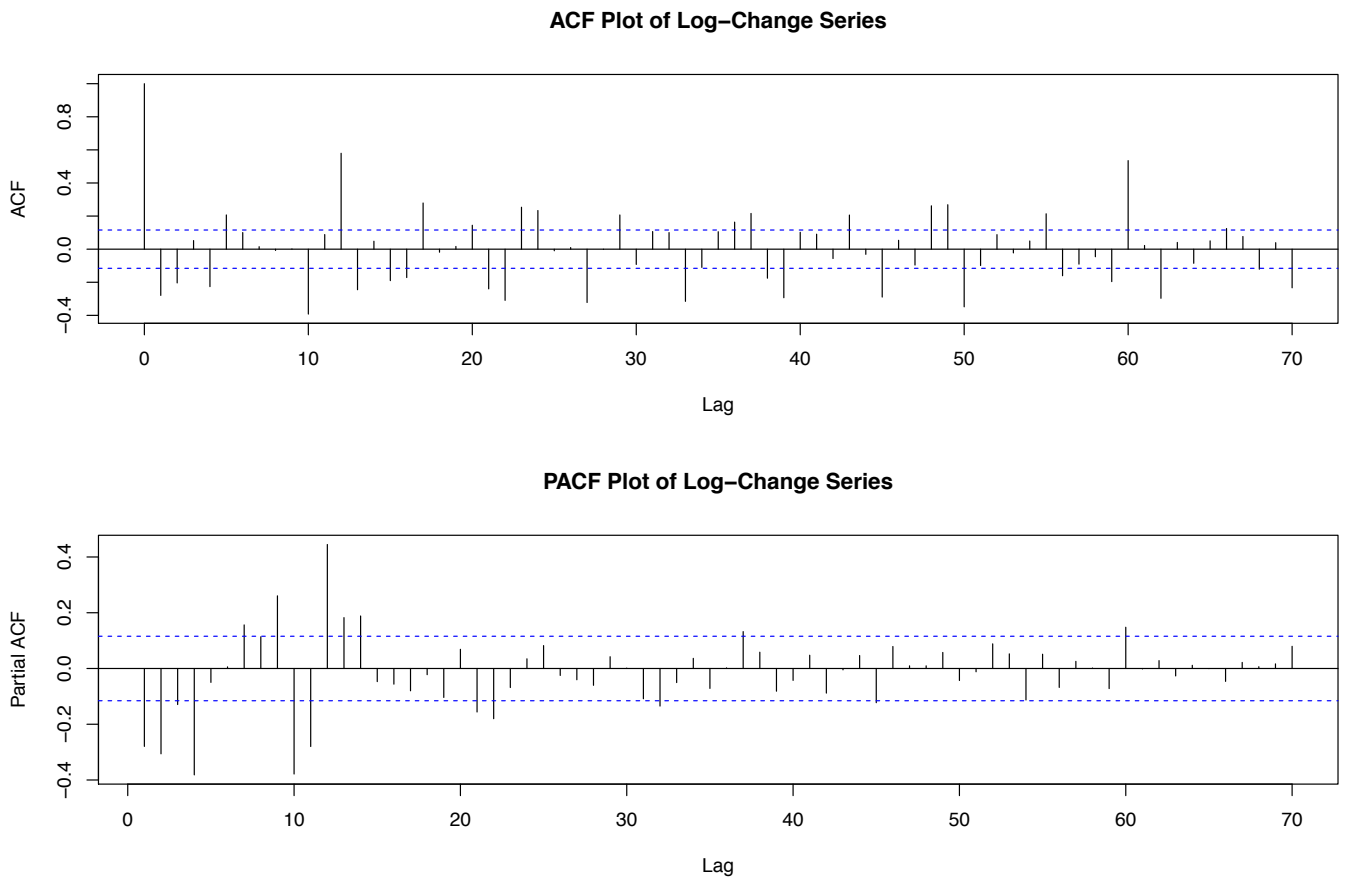

Figure 2: ACF and PACF plots for log-transformed and differenced Data

Two seasonal periods were looked at: 12 months and 60 months. The possibility of a 60 month cycle was suggested by the spikes at 60 months in both the ACF and PACF and the possibility of a 12 month cycle was suggested by the spikes at 12 months in both plots. In particular there are spikes every 12 months (up to 60 months) in the ACF which suggests that Q may be greater than one and even as large as 5.

Differencing for seasonality of 12 months actually made the ACF and PACF look worse. The inappropriateness of differencing by 12 months was confirmed by our exhaustive search, as models seasonally differenced by 12 months fared worse in terms of model selection criteria than those not seasonally differenced. Similar results were the case for 60 months.

From looking at the ACF and PACF of the log-transformed and differenced data, it appears that SARIMA model parameter p may be between 9 and 11 and parameter q looks to be 5 or 10. P (with period of 12) looks to be 1 and Q looks to be anywhere from 1 to 5.

With these observations in mind, an exhaustive search algorithm was performed to assess the fit of a wide range of SARIMA models on the difference log transformed time series. All possible SARIMA models with parameters in chosen ranges were fit and corresponding variance, log likelihood and AIC values were computed for each model (see Appendix A).

However, since our goal is to forecast a year using the data, we decided to validate that the top models in terms of aic and $sigma^2$ also did well with prediction. Therefore, we took out the months in 2010 (so only data from 1987 - 2009) and fit a wide range of SARIMA models to the remaining data. We then looked at

variance, log likelihood and AIC values to see how the top models with all the data compared to the top models with a year taken out. Though the rankings were different, many of the same models as with all the data were in the top 50 (see Appendix B).

We took the top 43 models (for the subsetted data) and used them to forecast values for 2010. For each of these forecasts a mean squared prediction error (MSPE) was computed and the models were ranked accordingly. A model that was optimal in terms of aic wasn't necessarily at the top of the list for MSPE (see Appendix C).

Out of all the models, the SARIMA model that ranked consistently near the top in each of the three lists in the appendices was the one with parameters $p = 10$, $d = 1$, $q = 10$, $P = 1$, $D = 0$, $Q = 1$ and $s = 12$. Given this evidence we decided to use that model for our predictions.

We looked at using a model that included the covariates Unemployment Rate, Bankruptcy Rate, and Housing Price Index for prediction purposes. However, using these covariates with our chosen model actually produced a higher MSPE. Therefore, we decided to not use them.

4.0.2 Model Verification

Using residual diagnostics, we checked the residual assumptions to verify that the chosen is appropriate for forecasting. Figure 3 shows the residual and standardized residual plots for this model.

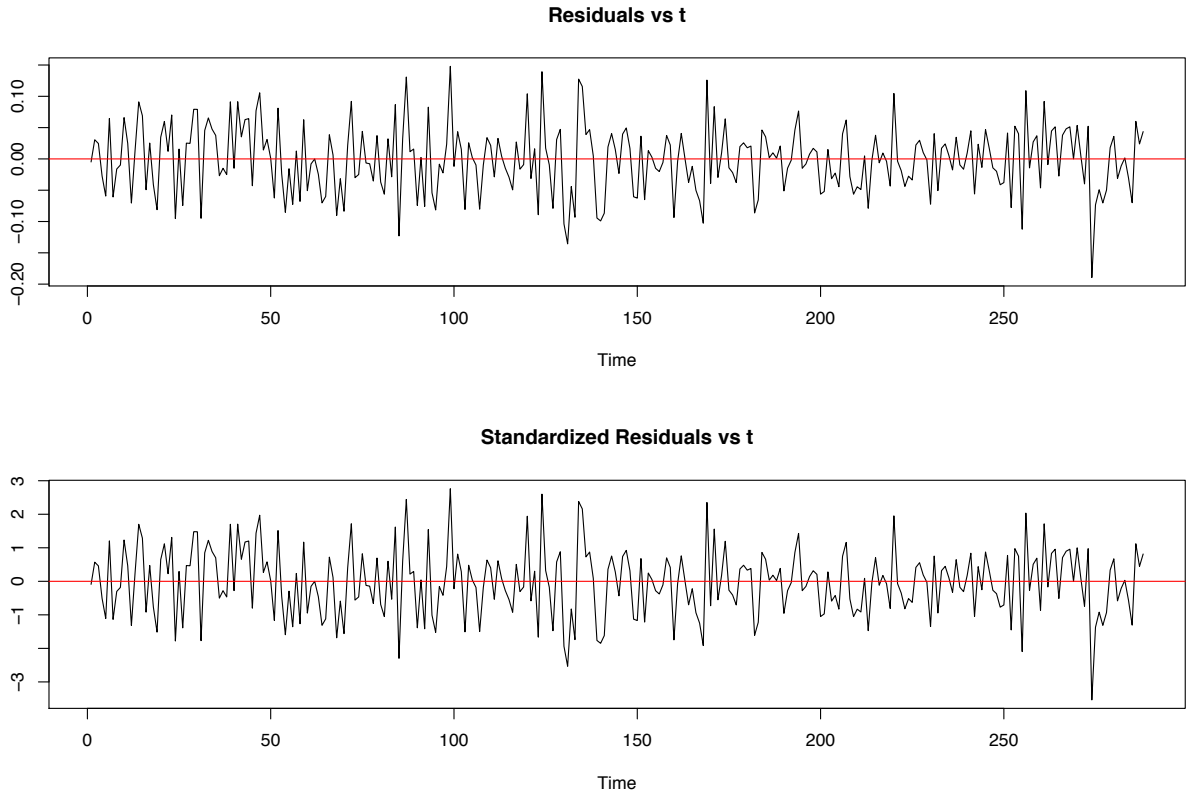

Figure 3: Residual plots for the $SARIMA$ (10,1,10)$\times$(1,0,1)$_{12}$ model

The assumption of zero mean seems to be fulfilled according to the residual plots. A t-test confirms this.

From Figure 2 we can see that the residuals appear to be somewhat homoscedastic. To verify this, we perform both Levene and Bartlett tests, which test the null hypothesis that the variance between subgroups of the residuals is equal. Both of these tests yield significant p-values on a $\alpha = 5\%$ level as can be seen in table 1.

Table 1: Homoscedasticity hypothesis test results

	Test Statistic	**p-value**
Levene	2.87	0.037
Bartlett	9.30	0.026

So there is weak evidence to suggest that the residuals do not have constant variance. This could potentially be an issue when attempting to forecast using the model. However, it is not strong heteroscedasticity so it still seems appropriate to use the model for forecasting. We checked whether ARCH/GARCH might be appropriate. However, there are no spikes in the ACF and PACF of absolute value of the residuals plots. This suggests that k = l = 0 and ARCH/GARCH isn't appropriate here.

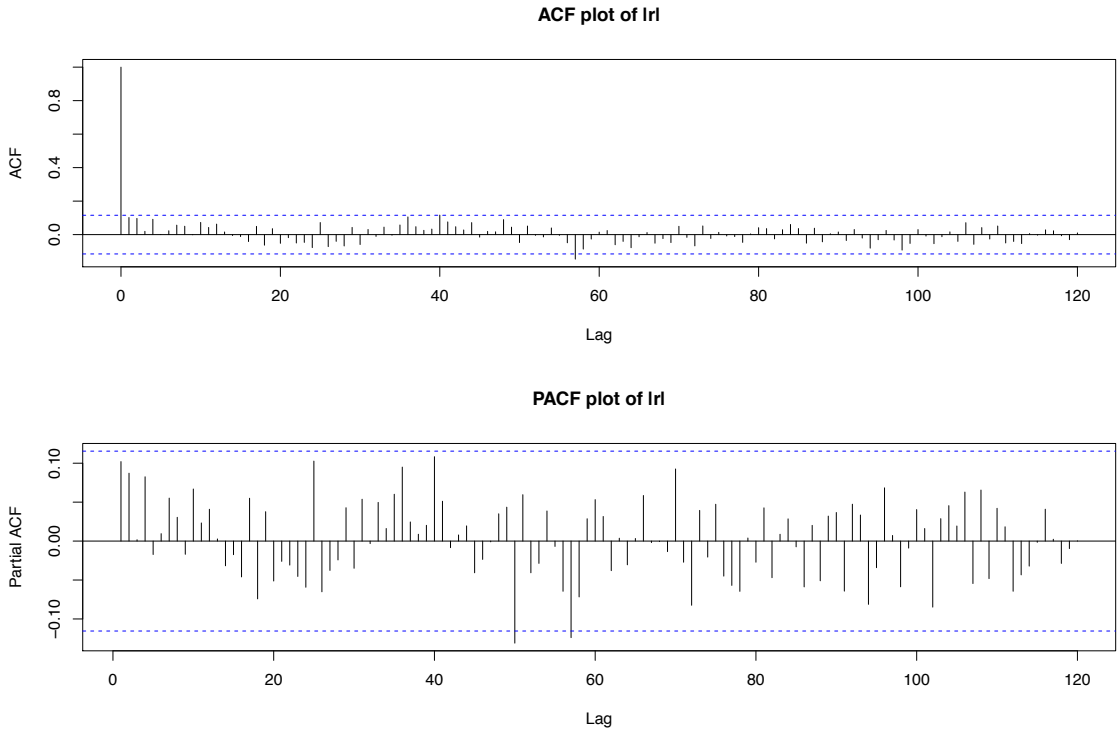

Figure 4: ACF and PACF of absolute value of residuals for $SARIMA$ (10,1,10)×(1,0,1)$_{12}$ model

Figure 5 shows a quantile-quantile plot of the residuals of the model, which compares the quantiles of the residuals to the quantiles of a standard normal distribution.

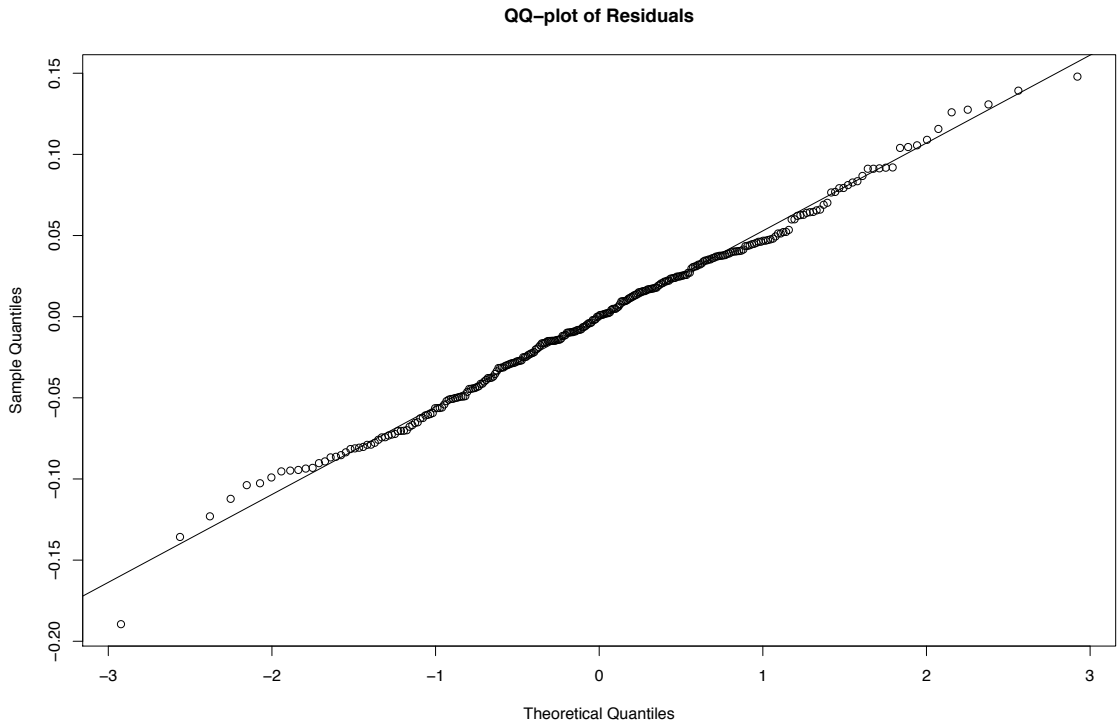

Figure 5: QQ-Plot of residuals for the $SARIMA$ $(10,1,10)\times(1,0,1)_{12}$ model

We can see that the residuals match up with the standard normal quantiles quite nicely, which suggests that the residuals are normally distributed. To formally test the null hypothesis that the residuals are normal we perform a Shapiro-Wilk test, the results of which are summarized in Table 2.

Table 2: Normality hypothesis test results

	Test Statistic	**p-value**
Shapiro-Wilk	0.99	0.61

Since the p-value is insignificant there is no evidence to suggest that the residuals are not normally distributed. Figure 6 shows a Ljung-Box plot and an ACF plot which looks for autocorrelation within the residuals. There isn't evidence that there is autocorrelation remaining after fitting the model.

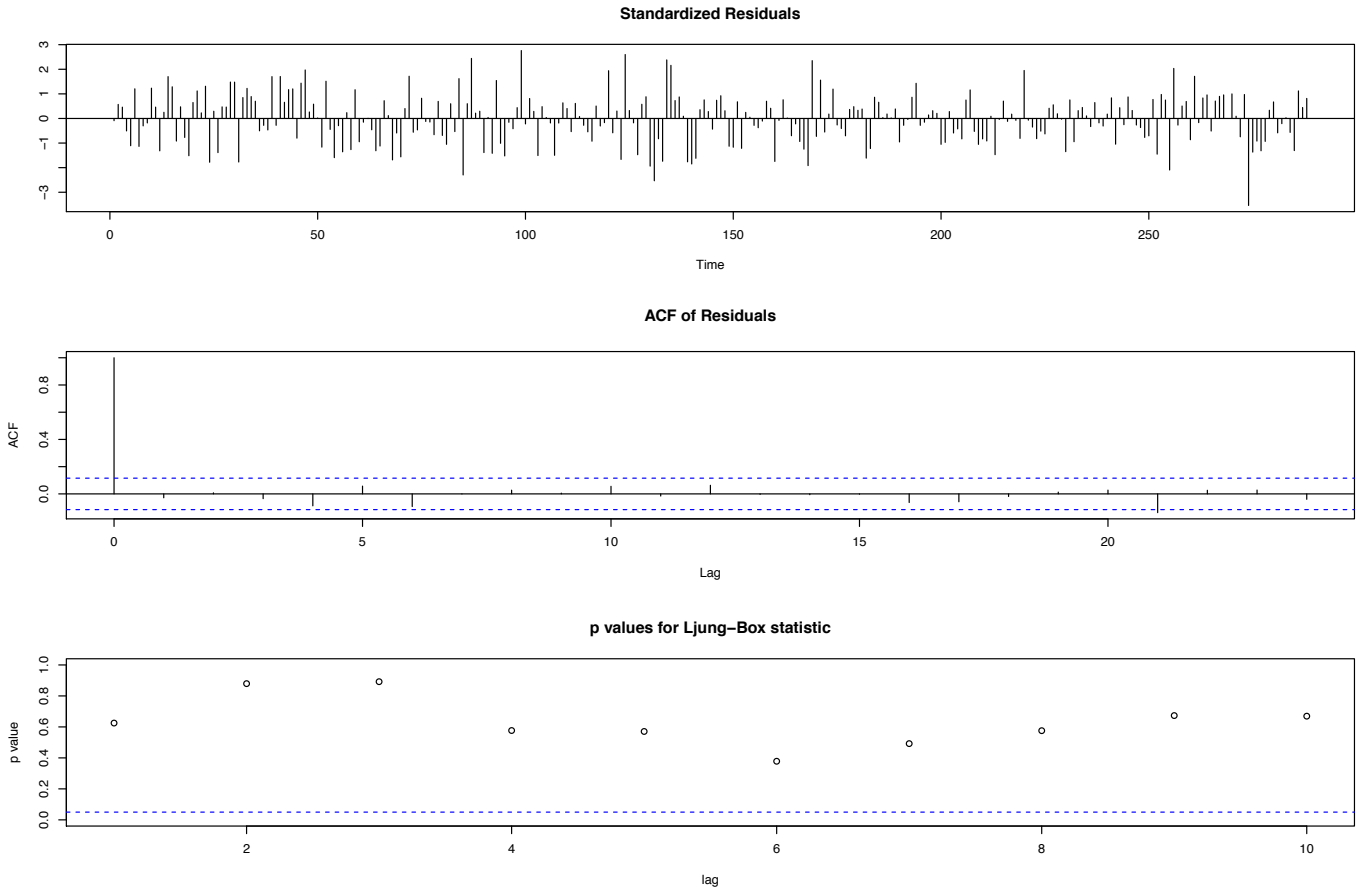

Figure 6: Residual vs. Time Plot, Residual ACF plot, and Ljung-Box Statistic Plot

All of the residual assumptions except for homoscedasticity are fulfilled. However, the heteroscedasticity isn't strong enough to make the chosen model seem inappropriate for forecasting.

4.1 Forecasting Results

Figure 7 shows the 2011 bankruptcy rate predictions (green) as well as 95% confidence limits (red) using the model, and Table 3 displays this information numerically.

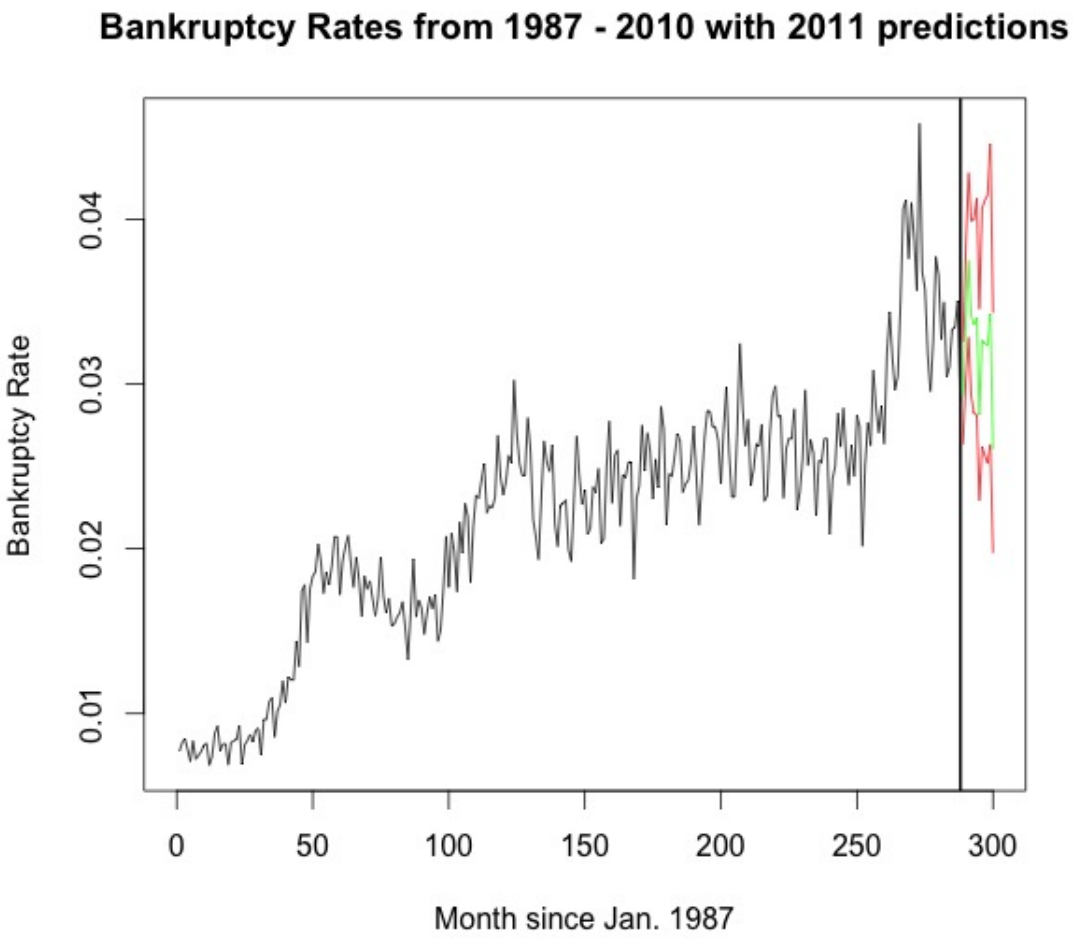

Figure 7: Plot of Canadian bankruptcy rate predictions for 2011

Table 3: 2011 Canadian Bankruptcy rate predictions

	Prediction	**95% Lower Bound**	**95% Upper Bound**
Jan	2.93	2.63	3.26
Feb	3.41	3.02	3.84
Mar	3.75	3.28	4.28
Apr	3.42	2.93	4.00
May	3.36	2.82	4.00
Jun	3.40	2.81	4.13
Jul	2.81	2.29	3.46
Aug	3.26	2.62	4.07
Sept	3.25	2.57	4.15
Oct	3.23	2.52	4.46
Nov	3.45	2.63	3.43
Dec	2.60	1.98	3.43

So the model predicts that Canadian bankruptcy rates will ultimately decrease over the course of 2011, however we can see that the variance is increasing as we forecast further out so we are less confident in

these predictions.

5 Conclusion and Recommendations

Through this investigation a SARIMA time series model was trained using 24 years of monthly Canadian bankruptcy rate data. The model was then used to forecast monthly bankruptcy rates for the 2011 year, the results of which suggest that the Canadian bankruptcy rate is expected to oscillate above and below 3% over the course of 2011. Financial institutions who use the Value at Risk measure at the 5% level for economic capital allocation may be particularly interested in the 95% upper bound we included. Bankruptcy rates in any one month are predicted to be as high as 4.46% in 2011. We suggest that any organizations with a financial interest in future Canadian bankruptcy rates plan accordingly.

6 Sources

(1) "A Fresh Start." The Economist. The Economist Newspaper, 14 Mar. 2015.

(2) Brockwell, Peter J., and Richard A. Davis. Introduction to Time Series and Forecasting. New York: Springer, 2002.