Final Project
Time Series Analysis

Group: Chhavi Choudhury, Abhishek Singh, Jaime Pastor

Course: MSAN 604 - Time Series Analysis

Term: Fall 2015

I. Introduction: Problem description

This report explores the data on Canadian national bankruptcy rates from January 1987 to December 2010. The goal of this analysis is to develop a time series model to precisely forecast monthly Bankruptcy Rates in Canada. Information about the Population, Unemployment Rate and House Price Index is available for the given time period. Therefore we explore the hypothesis that these variables may help explain the variation in Bankruptcy Rates and should be included as potential covariates in the model. **SARIMA (2, 1, 0) x (0, 0, 2)** $_{[12]}$ is chosen model yields a σ^2 of 0.006278, a log-likelihood of 303.68 and an AIC of -595.36. The model obtains an RMSE of 0.001834349 on the training dataset and 0.002211695 on the validation set. This model also uses House Price Index as a covariate.

II. Methodology

A. Relationship with Covariates:

Figure 1 shows time series plots for Bankruptcy Rates, Population, Unemployment Rate and House Price Index (Y-axis) with respect to time (X-axis). We can clearly identify in Figure 1 that most variables exhibit an increasing relationship with time, except for the unemployment rate. Moreover, the relationship of the response variable with the predictor variables can be seen better in a mixed correlation plot (Figure 2).

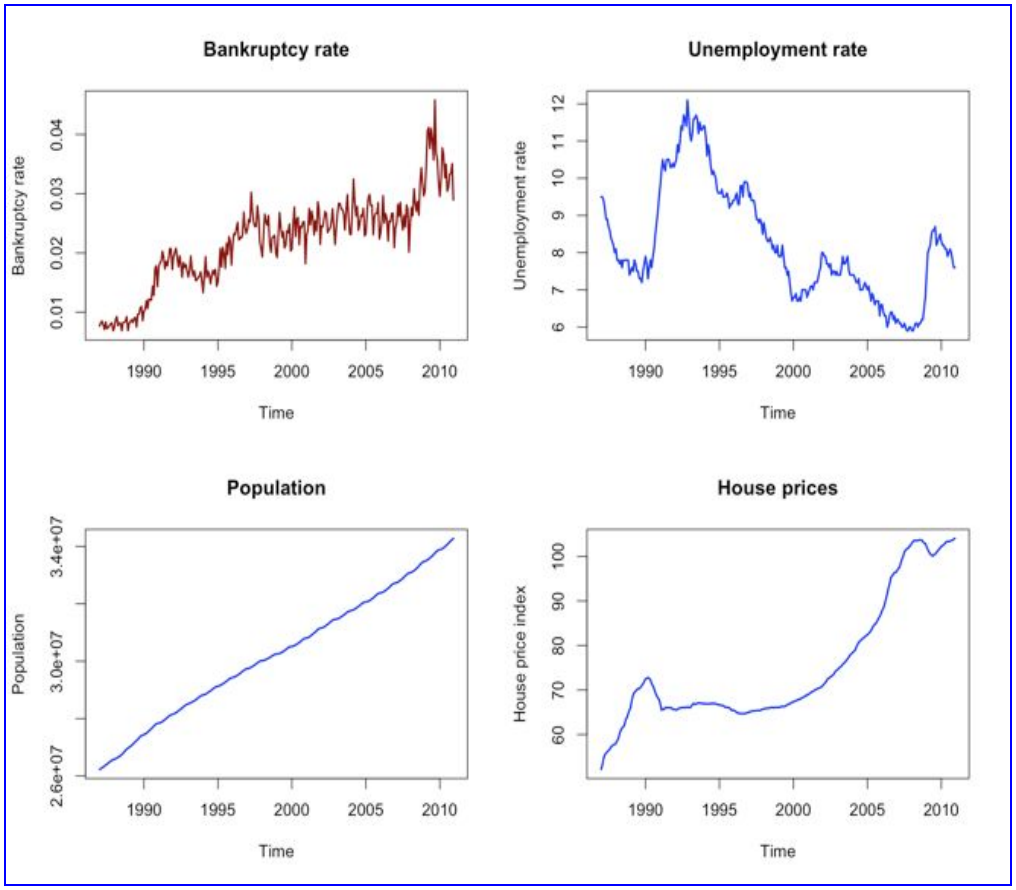

Figure 1

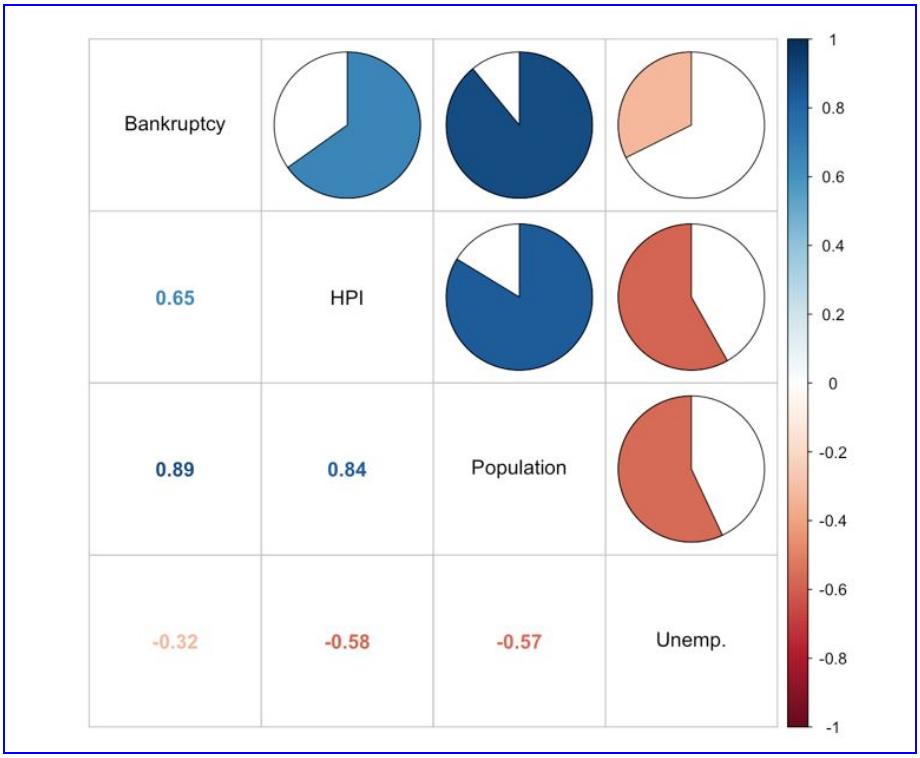

Figure 2

The correlation plot (Figure 2) shows that the *Bankruptcy Rate* has a strong correlation with the *Population* and *House Price Index*. Hence these two variables were considered while building the forecasting model for *Bankruptcy Rates*. However, we found significant evidence that the variable *Population* does not help explain the *Bankruptcy rate*, which makes intuitive sense, and consequently the final model only makes use of the *House Price Index* as a covariate.

B. Exploratory Data Analysis:

Seasonality:

The ACF plot confirms the presence of a seasonal component in the data. The periodicity of the seasonal component is determined to be 12 months (ACF plot, Figure 4). By using appropriate tests, we determine that the time series is seasonally stationary and it does not need to be seasonally differenced.

Order:

From the ACF and PACF plots, the time series shows signs of being an AR process, particularly as the ACF plot shows an exponential decay (Figure 4). However, while exploring possible models, MA components were also taken into consideration.

Stationarity:

A time series model is built for a stationary time series, but the response variable, Bankruptcy Rates, has a clear increasing trend with time. Hence ordinary differencing of the time series is done to make the time series stationary (Figure 3).

Heteroscedasticity:

A close examination of the plot for Bankruptcy Rates reveals non constant variance with time. A log transformation of the time series helps stabilize the variance and make the time series homoscedastic (Figure 3).

Mean Zero:

Time Series Analysis is performed on stationary time series which has a zero mean. The original series does not satisfy this condition. However by applying a log transformation and performing ordinary differencing, the result is a stationary time series with a mean of zero and is confirm by performing a t-test (Figure 3).

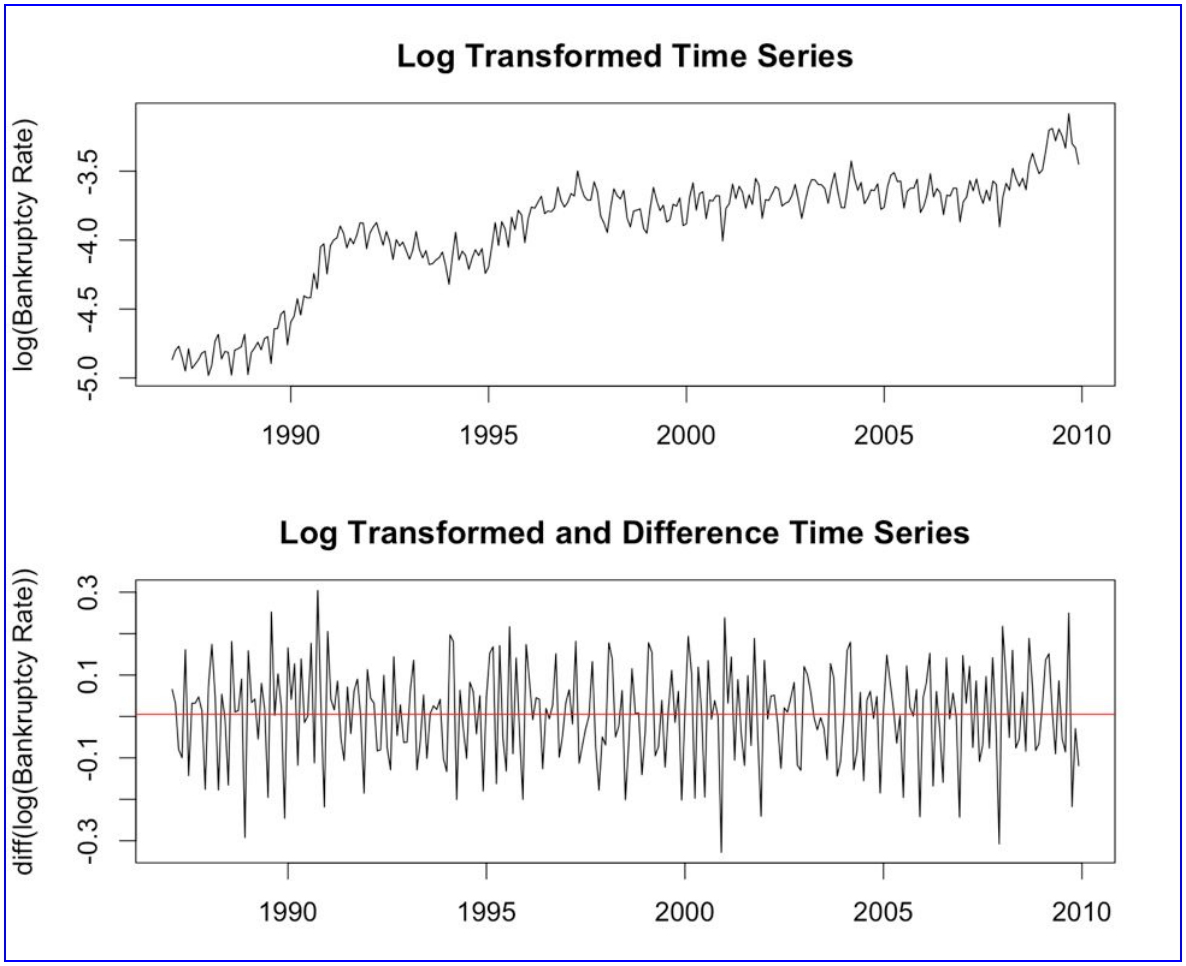

Figure 3

C. Modelling Philosophy

Data for the response variable and covariates is available for 24 years (1987 - 2010), and we aim to forecast the next 12 observations (year 2011). In order to test various models, the data for year 2010 (12 observations) is used as a validation set and excluded from the model building process. The models are built based on the data for the first 23 years (1987 - 2009). Once a model is finalized, all the available data

(24 years, 288 observation) is used to re-train the model. This changes the parameters estimates slightly but improves the forecasting ability of the model for the year 2011 (Table 1).

Table 1:

Phase 1	Training Set	Test Set
Model Building	Data for 23 years (1987 - 2009)	Data for 12 months (2009)
Forecast	Data for 24 years (1987 - 2010)	Data for 12 months (2011)

D. Model Building Process

The response time series 'Bankruptcy Rate' is log transformed to remove the non constant variance and then differenced to make the time series stationary. ACF and PACF plot for this transformed and differenced time series are studied to determine the orders for SARIMA model.

From Figure 4 it can be seen that:

p = 2 : 2 initial spikes in PACF plot.

d = 1 : Ordinary differencing of data

q = 0 : AR process as determined by Figure 6 in Appendix

P = 1 : 1 Spike at lag 12 in PACF plot

D = 0 : No need for stationary differencing

Q = 2: 1 Spike at lag 12 and 1 Spike at lag 24 in ACF plot

To select the model, the method of overfitting was used. Multiple models were explored and the values of the log-likelihood, σ^2, RMSE on the validation set as well as RMSE on the complete set were recorded (Table 2 in Appendix). The models used Least Squares method of estimation.

III. Model Selection and Validation

From the observed models, two were shortlisted:

- **Model I : SARIMA (2, 1, 0) x (1, 0, 2) $_{[12]}$**
- **Model II: SARIMA (2, 1, 0) x (0, 0, 2) $_{[12]}$**

While the first model obtains better values of log-likelihood and σ^2, the second model achieves better RMSE. The Likelihood Ratio Test was conducted to break the tie and it favored the first model over the second one (refer Appendix for details). However this may be due to an extra parameter used in the first model. Since our goal with the choice of the model is to obtain predictions as accurate as possible, **model II**, that is, **SARIMA (2, 1, 0) x (0, 0, 2) $_{[12]}$** was chosen. The residuals for the SARIMA **(p=2, d=1, q=0) (P=0, D=0, Q=2) (S=12)** model are normally distributed, which allowed us to re-train the model using the Maximum Likelihood (ML) Method. Using ML method reduces the variance of the parameter estimates, however all the parameters were found to be significant.

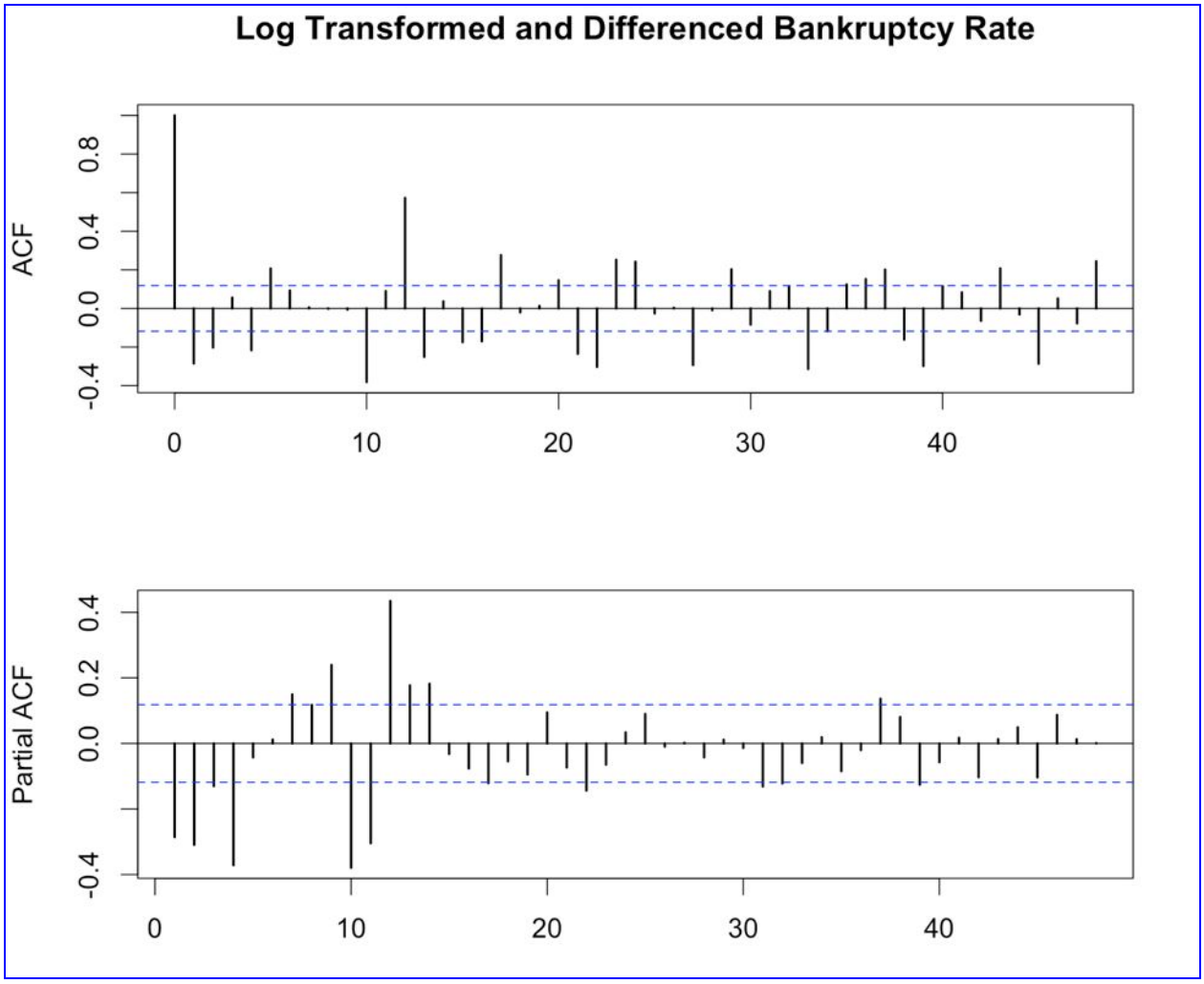

Figure 4

IV. Residual Diagnostics

1. **Zero Mean:** The Residual plot (Figure 7, Refer Appendix) suggests that the mean of the residuals is zero. However, the t-test gives a p-value of 0.04436 (Refer Appendix). Hence we have enough evidence to reject the null hypothesis and conclude that the mean of residuals is significantly different than zero.
2. **Homoscedasticity:** The Residual plot (Figure 7, Refer Appendix) suggests that the residuals are homoscedastic. The Brown-Forsythe test for homoscedasticity confirms this hypothesis. The test was conducted dividing the residuals into 4 groups. The test yielded a p-value of 0.8819 (Refer Appendix) which confirms our null hypothesis that the residuals have constant variance with time.
3. **Outliers:** Visual Inspection of Residual Plot (Figure 7, Refer Appendix) suggests there are no outliers.
4. **AutoCorrelation:** The ACF of residuals (Figure 9, Refer Appendix) shows some intermittent peaks which may be due to sampling error or because the model does not completely capture autocorrelation at various lags. A higher order model will probably overcome this issue but is avoid

to favor model parsimony. Runs Test (p value 0.6295, Refer Appendix) confirms that the residuals are random and hence have no autocorrelation.

5. **Normality:** The QQ plot (Figure 8, Refer Appendix) shows that the residuals are normally distributed. This observation is confirmed by the Shapiro-Wilk test (p-value of 0.7791, Refer Appendix), which confirms the null hypothesis that the residuals follow a normal distribution.

V. Issues

The **SARIMA (2, 1, 0) x (0, 0, 2) $_{[12]}$** model does not pass the Ljung-Box test for Autocorrelation. This can be overcome by using a higher order model. However, higher order models have insignificant parameter estimates and hence are not chosen. The residuals of this model do not have a expected value of zero. In conclusion the residuals of this model do not meet all the OLS assumptions and hence have biased parameter estimates (refer Appendix for parameter estimates).

VI. Forecasting

The **SARIMA (2, 1, 0) x (0, 0, 2) $_{[12]}$** model was trained again using the data corresponding to 24 years (1987-2010). The parameter estimates change slightly. The static window method was used for the forecasting. This model yields a σ^2 of 0.006278, a log-likelihood of 303.68 and an AIC of -595.36. The model obtains an RMSE of 0.001834349 on the training dataset and an RMSE 0.002211695 on the validation set.

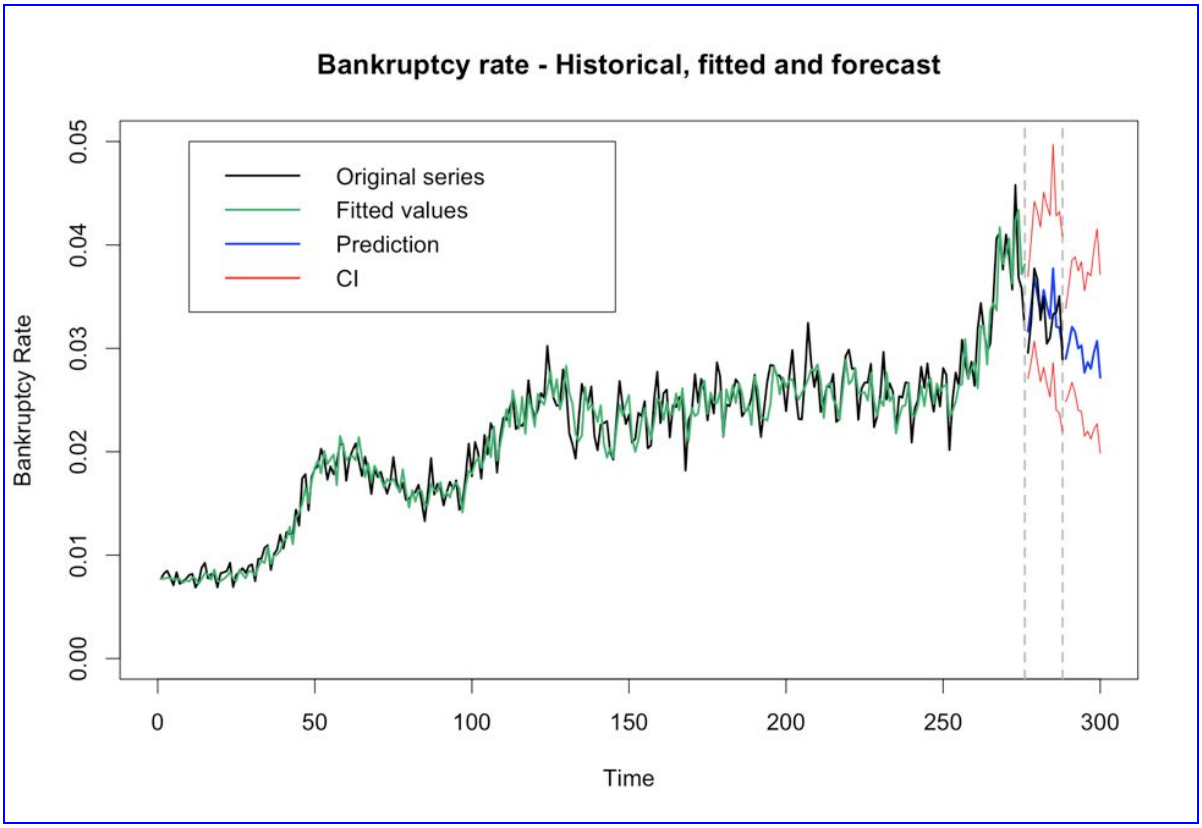

Figure 5

APPENDIX

Computations done on different orders of SARIMA process for all possible combinations of p <= 2, q <= 2, P <= 1, Q <= 2. D = 1 and d = 0

In the table below:

- **pval:** p-value for brown forsythe test
- **rmse_val:** RMSE on validation set
- **rmse_comp:** RMSE on complete data set

Table 2

	SARIMA(p,d,q)(P,D,Q)$_{[12]}$	**σ^2**	**log-lik**	**pval**	**rmse_val**	**rmse_comp**
1	0,1,0,0,0,0	0.013437666	202.3748	0.949	0.003141753	0.002559485
2	0,1,0,0,0,1	0.008832359	260.0752	0.702	0.002643649	0.002098599
3	0,1,0,0,0,2	0.008543647	264.6449	0.651	0.002461477	0.00211422
4	0,1,0,1,0,0	0.008654529	262.8719	0.169	0.002166441	0.002163395
5	0,1,0,1,0,1	0.008456669	266.0519	0.389	0.002057848	0.00212096
6	0,1,0,1,0,2	0.007255799	287.1106	0.616	0.00158323	0.001937201
7	0,1,1,0,0,0	0.010907596	231.0576	0.968	0.003044409	0.002343352
8	0,1,1,0,0,1	0.006781364	296.4087	0.686	0.003016322	0.001868752
9	0,1,1,0,0,2	0.006269668	307.1963	0.837	0.003465515	0.001825722
10	0,1,1,1,0,0	0.005761655	318.8149	0.51	0.004315057	0.00174708
11	0,1,1,1,0,1	0.005759097	318.8759	0.542	0.004205881	0.001747039
12	0,1,1,1,0,2	0.004758011	345.1317	0.793	0.00466252	0.001567429
13	0,1,2,0,0,0	0.01058479	235.1883	0.988	0.00308349	0.002291203
14	0,1,2,0,0,1	0.006738339	297.2839	0.642	0.003218614	0.001858445
15	0,1,2,0,0,2	0.006266901	307.257	0.809	0.003514757	0.001824674
16	0,1,2,1,0,0	0.005587158	323.0435	0.471	0.003717016	0.001722668
17	0,1,2,1,0,1	0.004939249	339.9914	0.182	0.004357312	0.001591163
18	0,1,2,1,0,2	0.00450594	352.6162	0.492	0.00392307	0.001536968

19	1,1,0,0,0,0	0.012356625	213.9068	0.812	0.00283056	0.002472171
20	1,1,0,0,0,1	0.007832439	276.5956	0.796	0.002053537	0.001975336
21	1,1,0,0,0,2	0.00733704	285.5796	0.972	0.002055463	0.001956539
22	1,1,0,1,0,0	0.006621125	299.6968	0.214	0.002787364	0.001895536
23	1,1,0,1,0,1	0.006620347	299.7129	0.221	0.002721052	0.001895116
24	1,1,0,1,0,2	0.005453537	326.3719	0.768	0.003001408	0.00168978
25	1,1,1,0,0,0	0.010720008	233.4429	0.983	0.003022816	0.002307515
26	1,1,1,0,0,1	0.0068202	295.6236	0.537	0.002958632	0.001860053
27	1,1,1,0,0,2	0.006319613	306.1053	0.817	0.003348921	0.001823862
28	1,1,1,1,0,0	0.005593719	322.8822	0.361	0.004082877	0.001729628
29	1,1,1,1,0,1	0.005580287	323.2127	0.292	0.004425779	0.001725128
30	1,1,1,1,0,2	0.004666541	347.8008	0.812	0.004393114	0.001559227
31	1,1,2,0,0,0	0.01019177	240.3909	0.952	0.002926305	0.002264221
32	1,1,2,0,0,1	0.006453404	303.2247	0.714	0.002912824	0.001824595
33	1,1,2,0,0,2	0.005996541	313.3206	0.77	0.003308375	0.001786181
34	1,1,2,1,0,0	0.005519247	324.7251	0.325	0.003611353	0.001716904
35	1,1,2,1,0,1	0.005061218	336.6373	0.304	0.004277636	0.001605268
36	1,1,2,1,0,2	0.004572634	350.5959	0.806	0.003722959	0.001544574
37	2,1,0,0,0,0	0.011201283	227.4044	0.85	0.002674174	0.002374428
38	2,1,0,0,0,1	0.007070849	290.6609	0.682	0.002030943	0.001885172
39	2,1,0,0,0,2	0.006464612	302.9861	0.866	0.00224855	0.001834349
40	2,1,0,1,0,0	0.005230814	332.1053	0.334	0.003143343	0.001658397
41	2,1,0,1,0,1	0.004660551	347.9774	0.135	0.004822815	0.001531868
42	2,1,0,1,0,2	0.004303992	358.9211	0.625	0.003788918	0.001488823
43	2,1,1,0,0,0	0.010479779	236.5592	0.982	0.003156789	0.002281338
44	2,1,1,0,0,1	0.006768464	296.6706	0.577	0.002774219	0.001851898

45	2,1,1,0,0,2	0.006286295	306.8321	0.875	0.003087436	0.001815905
46	2,1,1,1,0,0	0.005148215	334.2939	0.266	0.00282313	0.001641534
47	2,1,1,1,0,1	0.00455334	351.1774	0.226	0.004402947	0.001522195
48	2,1,1,1,0,2	0.004221453	361.5836	0.63	0.003385647	0.001475341
49	2,1,2,0,0,0	0.010255357	239.5357	0.925	0.002920284	0.002266194
50	2,1,2,0,0,1	0.006665697	298.7743	0.53	0.002930002	0.001832066
51	2,1,2,0,0,2	0.006140093	310.0678	0.632	0.003301537	0.001781988
52	2,1,2,1,0,0	0.00504205	337.159	0.363	0.003544023	0.001632812
53	2,1,2,1,0,1	0.004553156	351.1829	0.227	0.004425871	0.00152213
54	2,1,2,1,0,2	0.004193092	362.5105	0.794	0.003761104	0.001473194

ACF and PACF Plot of Original Time Series

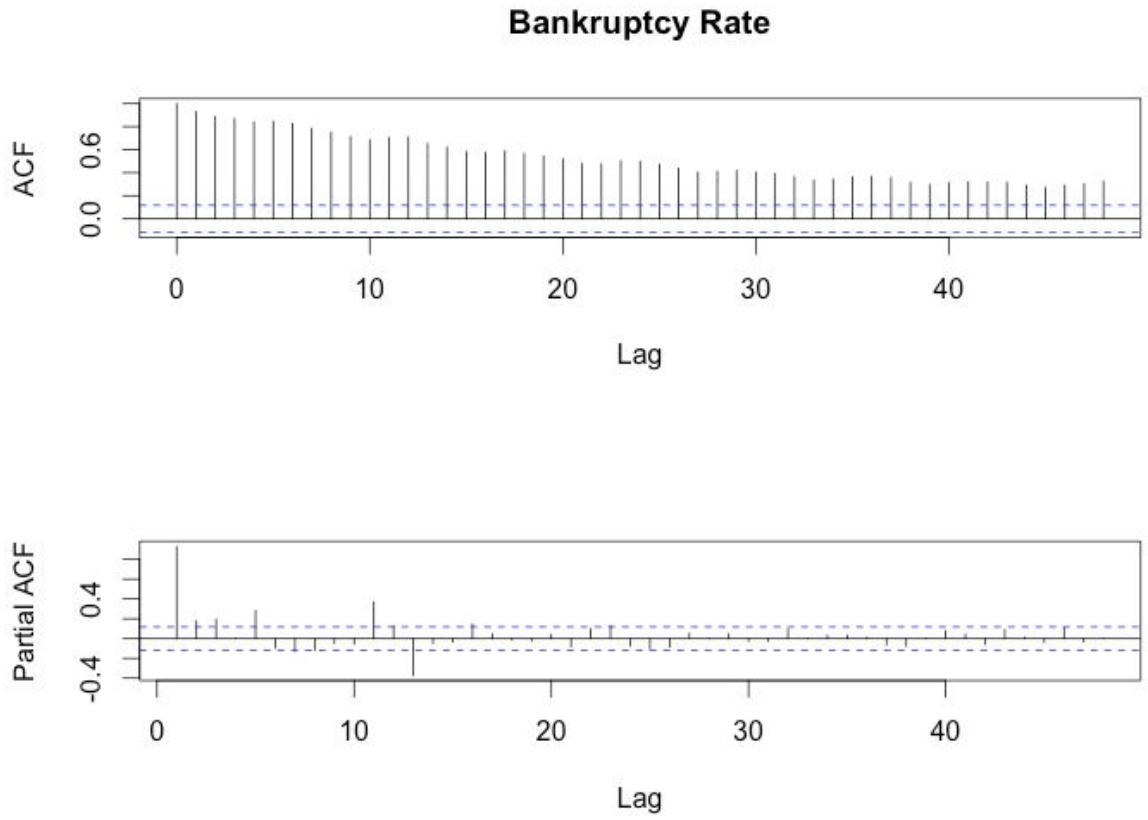

Figure 6

CHOSEN MODEL

SARIMA (2,1,0) x (0,0,2)$_{[12]}$ Model Parameters on Training Set (Yr 1987-2009)

```
Call:
arima(x = log(Bankruptcy_Rate_train), order = c(2, 1, 0), seasonal = list(order = c(0,
    0, 2), period = 12), xreg = data.frame(House_Price_Index_train), method = "ML")

Coefficients:
          ar1      ar2    sma1    sma2  House_Price_Index_train
      -0.5346  -0.3461  0.7049  0.2713                  -0.0280
s.e.   0.0581   0.0566  0.0672  0.0586                   0.0078

sigma^2 estimated as 0.006278:  log likelihood = 303.68,  aic = -595.36
```

NEXT BEST MODEL

SARIMA (2,1,0) x (1,0,2)$_{[12]}$ Model Parameters on Training Set (Yr 1987-2009)

```
Call:
arima(x = log(Bankruptcy_Rate_train), order = c(2, 1, 0), seasonal = list(order = c(1,
    0, 2), period = 12), xreg = data.frame(House_Price_Index_train), method = "CSS")

Coefficients:
          ar1      ar2    sar1     sma1     sma2  House_Price_Index_train
      -0.7779  -0.4754  0.9931  -0.5040  -0.3137                  -0.0296
s.e.   0.0558   0.0549  0.0105   0.0644   0.0648                   0.0052

sigma^2 estimated as 0.004304:  part log likelihood = 358.92
```

Residual Diagnostics:

Residual Plot

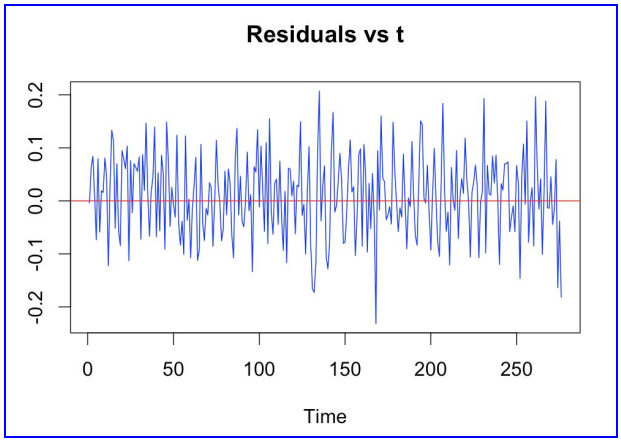

Figure 7

QQ Plot

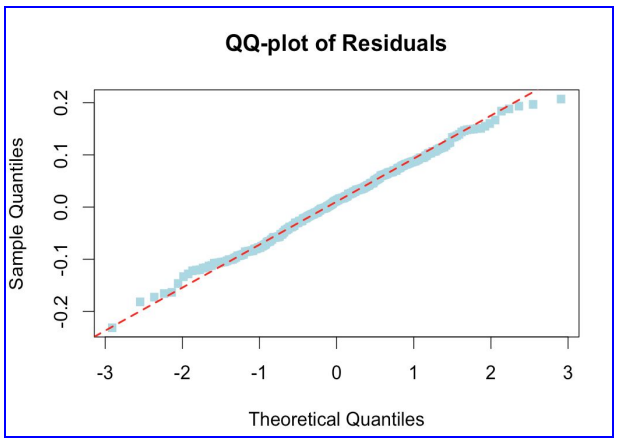

Figure 8

Ljung Box plots: AutoCorrelation

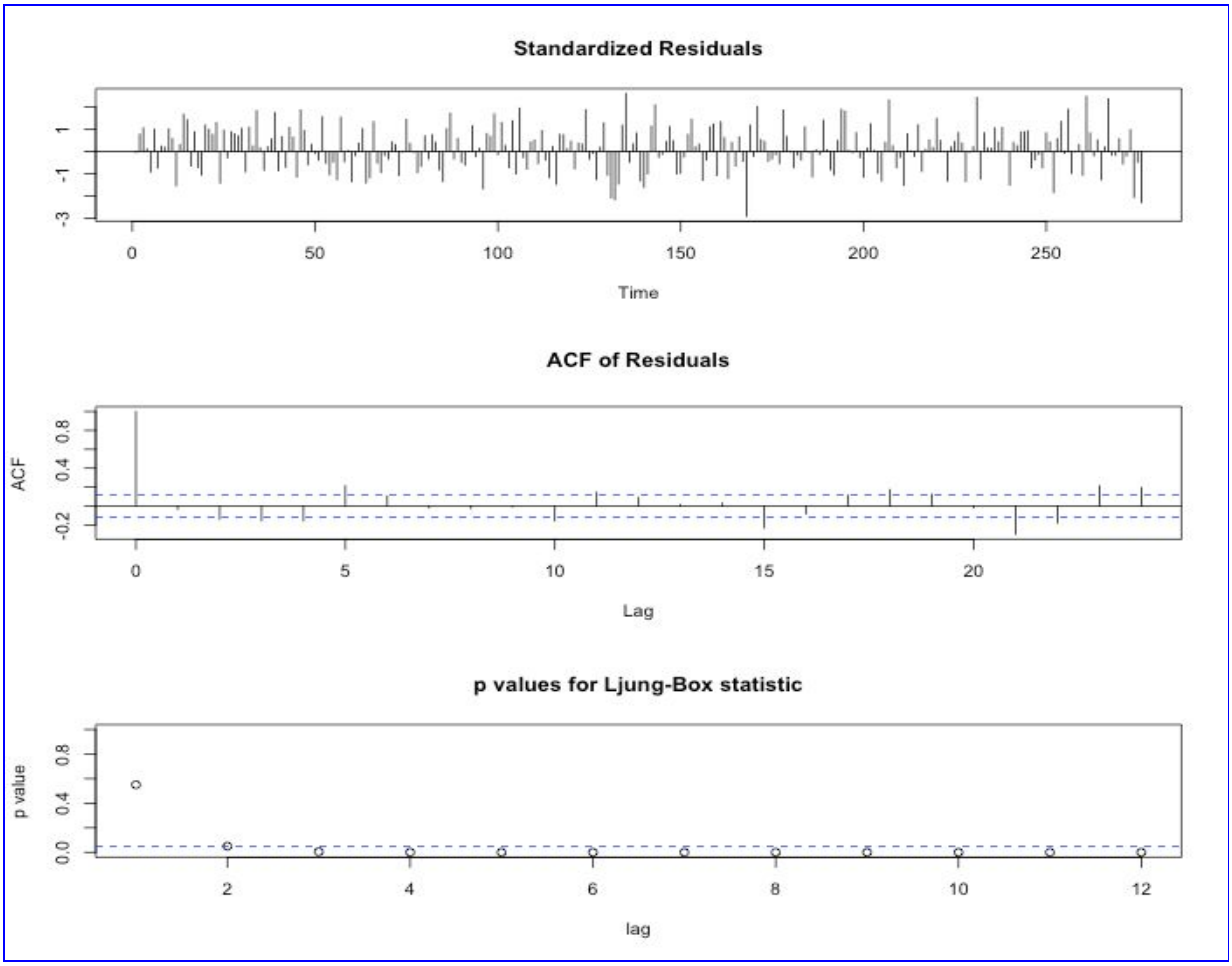

Figure 9

FORMAL TESTS:

One Sample t-test: Testing Mean Zero Assumption

```
        One Sample t-test

data:  e
t = 2.0199, df = 275, p-value = 0.04436
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.0002427474 0.0188829578
sample estimates:
  mean of x
0.009562853
```

Test for Homoscedasticity

```
        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  e
Test Statistic = 0.22085, p-value = 0.8819
```

Runs Test - Two sided For Randomness

```
        Runs Test - Two sided

data:  e
Standardized Runs Statistic = 0.48242, p-value = 0.6295
```

Shapiro-Wilk normality test

```
        Shapiro-Wilk normality test

data:  e
W = 0.99636, p-value = 0.7791
```

Ljung Box test:

Our model residuals don't fully qualify the requirement of the Ljung Box test. This is evident from the plot on the previous page.

Model Trained on full dataset:

```
Call:
arima(x = log(c(Bankruptcy_Rate_train, val$Bankruptcy_Rate)), order = c(2, 1,
    0), seasonal = list(order = c(0, 0, 2), period = 12), xreg = data.frame(c(House_Pr
ice_Index_train,
    val$House_Price_Index)), method = "ML")

Coefficients:
          ar1      ar2    sma1    sma2
      -0.5213  -0.3389  0.7095  0.2569
s.e.   0.0574   0.0557  0.0621  0.0558
      c.House_Price_Index_train..val.House_Price_Index.
                                                -0.0284
s.e.                                             0.0075

sigma^2 estimated as 0.006253:  log likelihood = 317.66,  aic = -623.33
```

Likelihood Ratio Test

```
Test Statistic: 111.87 P-value: 0
```

Bibliography

Domain specific:

- Nelson, John P. - "*Consumer Bankruptcies and the Bankruptcy Reform Act: A Time-Series Intervention Analysis, 1960–1997".* Journal of Financial Services Research

Time Series:

- Brockwell, Peter and Davis, Richard - "*Introduction to Time Series and Forecasting"*