Predicting Bankruptcy Rate in Canada

Shikhar Gupta, Qian Li, Sangyu Shen, Jake Toffler, Asmita Vikas

Goal:

The ability to forecast bankruptcy rate is a useful tool for national banks, insurance companies, credit-lenders, and more. We have been given 24 years of monthly data starting from 1987 for the **bankruptcy rate, population, unemployment rate and housing price index** in Canada. Our goal is to use this data to build a model and then forecast the monthly bankruptcy rate for the subsequent two years.

Our Process:

Time series[1] can be modeled using a number of different methods. Below is a diagram that outlines the conventional methods for time series modeling and when they are used. We have highlighted the models we deemed fit for this dataset in yellow.

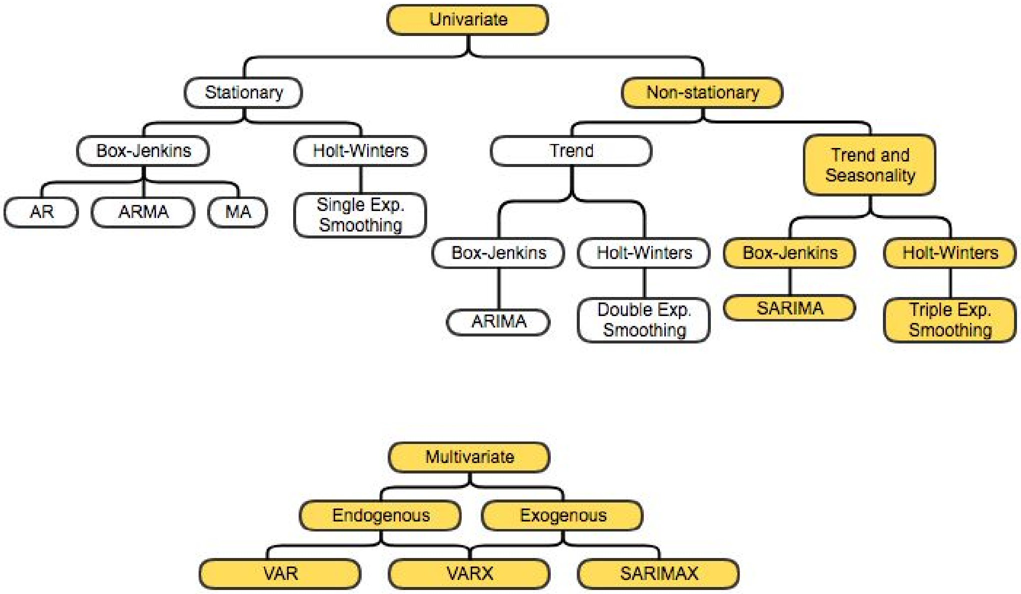


Figure 1: Decision tree of available time series modeling approaches

Choice of model depends on the data at hand. Specifically we look at the following aspects:

Univariate or Multivariate:

If our response variable (in this case bankruptcy rate) is only modeled against time, then we are in the univariate case. If there are external variables that are used to model the time series for our response variable, we are in the multivariate case. Since we have information on population, housing price index and unemployment rate, we will try both univariate and multivariate models for our analysis.

Univariate

Stationary vs Non-stationary:

Referring back to figure 1, the next question to ask is whether the data is stationary or non-stationary. We define a time series as stationary if its properties, such as mean and variance, do not change over time. If it is non-stationary, then we need to check for trend[4] and/or seasonality[5] in our data that will further narrow our choice of models.

From figure 2 it is clear that our target variable, bankruptcy rate, does not fit this definition of stationary. As time progresses, the data shows clear upward **trend**, i.e. consistent (but not necessarily constant) directional movement over time. Also from the ACF[15] plot we can conclude that bankruptcy rate is also exhibiting **annual seasonality**.

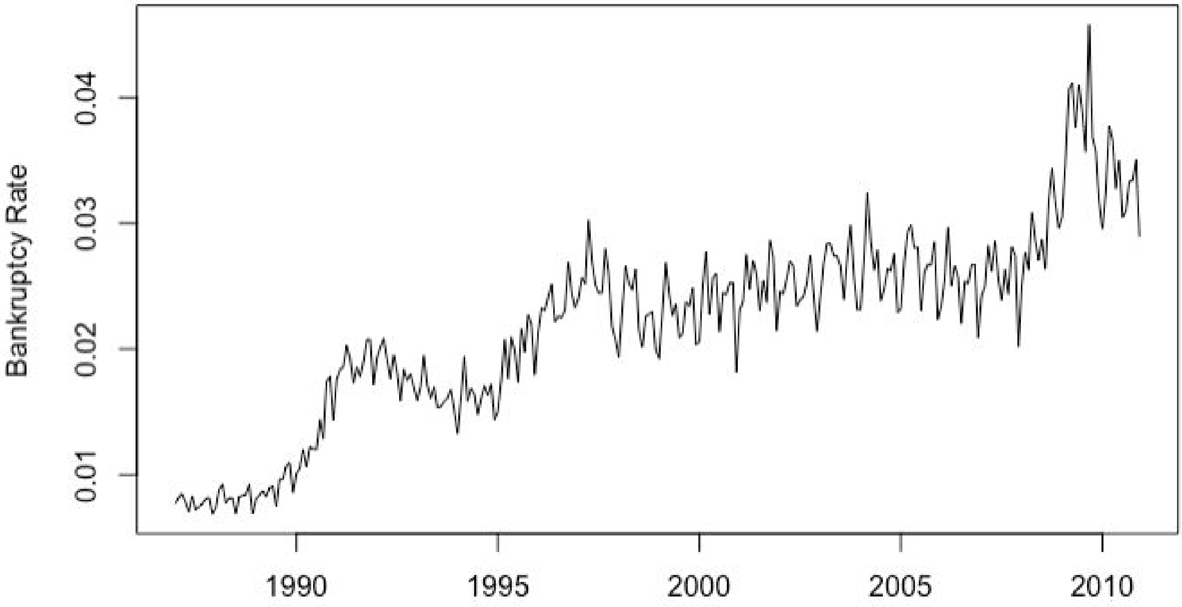


Figure 2: Plot of bankruptcy rate over time

If we are considering the univariate approach, **Box-Jenkins SARIMA models** and **Holt-Winters Triple Exponential Smoothing models** seem like suitable models to try.

Multivariate

Endogeneous vs. Exogeneous:

If we are considering a multivariate approach , we have to determine whether the external variables are exogenous[11] or endogenous[12] . External variables are considered exogenous if they influence the response variable but the response variable does not influence them. We say that a set of variables are endogenous if all variables influence one another.

In the case of exogenous variables we use a SARIMAX model, which is an extension of SARIMA with the addition of external variables. If the set of variables are endogenous we use a vector autoregressive (VAR) model, in which we try to model each variable in the system in terms of other variables. If we have both exogenous and endogenous variables, a hybrid model, which we'll refer to as a VARX model, is appropriate.

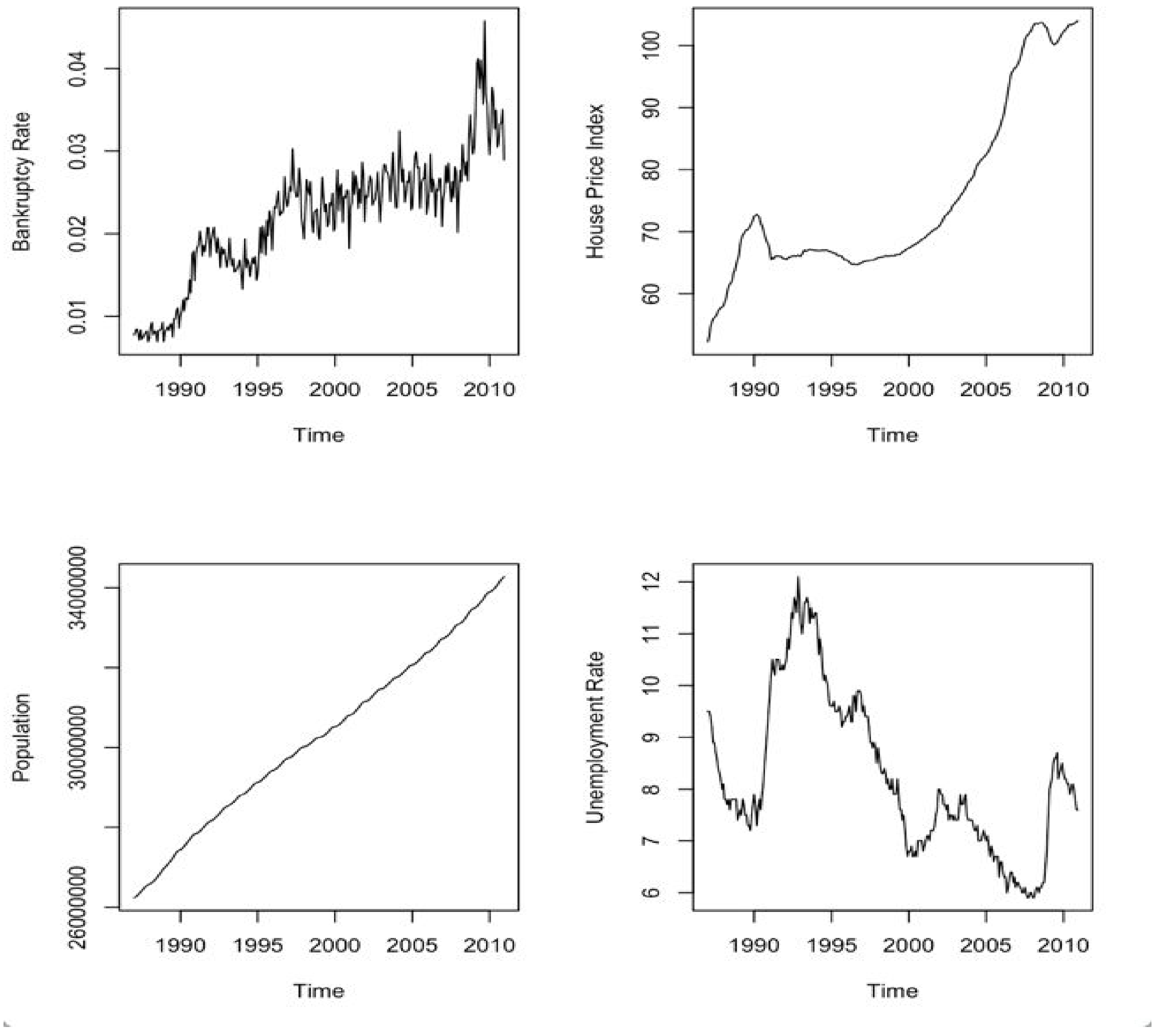


Figure 3. Plot of Bankruptcy Rate along with other influential variables

In our case the external variables are: **Unemployment Rate, Population,** and **Housing Price Index**. As we can not say with certainty whether any of them are exogenous or endogenous, we will try fitting multivariate models for both the exogenous and endogenous cases.

Modeling strategy:

Data pre-processing:

We have done some data manipulation before we begin trying to fit some of the models that we've discussed above. The purpose of these manipulations are to make the models robust in regards to random noise and outlier events in the data. This is how we pre-processed the data:

- Normalizing the bankruptcy rate using log transformation to reduce the effect of increasing amplitude of the peaks

- Subsetting the data after 1992 as the bankruptcy rate was very low before 1992. We assume that it won't help in explaining the current bankruptcy rate.

Possible models:

As discussed above, the following models seem appropriate to try for the given data:

- Triple Exponential Smoothing
  - This method predicts future data points by exponentially smoothing the weighted sum of past observations.

- SARIMA (seasonal autoregressive integrated moving average)
  - This model predicts future data points by observing past observations and errors. This method is built on the assumption that the time series after finitely differencing will be stationary.

- SARIMAX
  - This model adds one or more external predictor variables (population, unemployment rate and housing price index) to SARIMA. The extra variables help to explain the randomness and noise that cannot be explained by trend and seasonality.

- VAR
  - This model treats all variables symmetrically and captures the interaction between variables when predicting bankruptcy rate. If the variables are all dependent on one another, we expect the VAR model to outperform the SARIMAX model

Criteria for model selection:

A common framework for model selection is to divide the labeled data into training and validation sets. The idea is to train different models on the training data and then compare performance on the validation data based on some metric. For each method, we select the best model using a combination of lowest AIC and lowest RMSE of the validation set.

In order to choose a validation set as close to the test set as possible, we set apart the last two years of data from training to be the validation set, as our test set is two years of data directly following the original training set in time. The chronology chart of the data is displayed below:

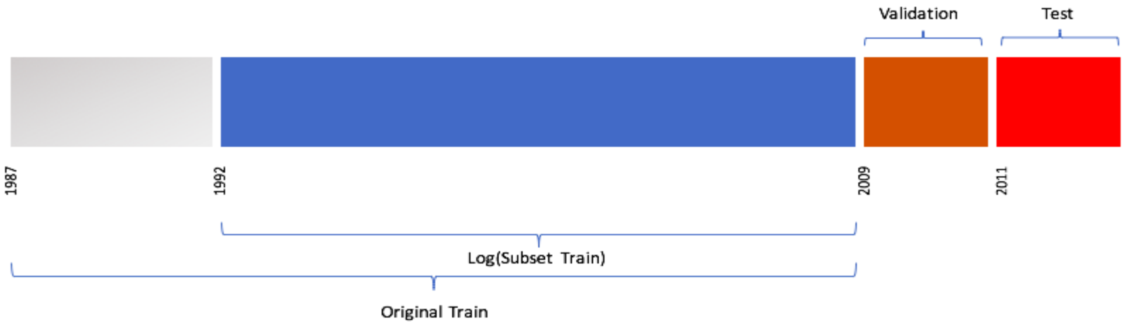
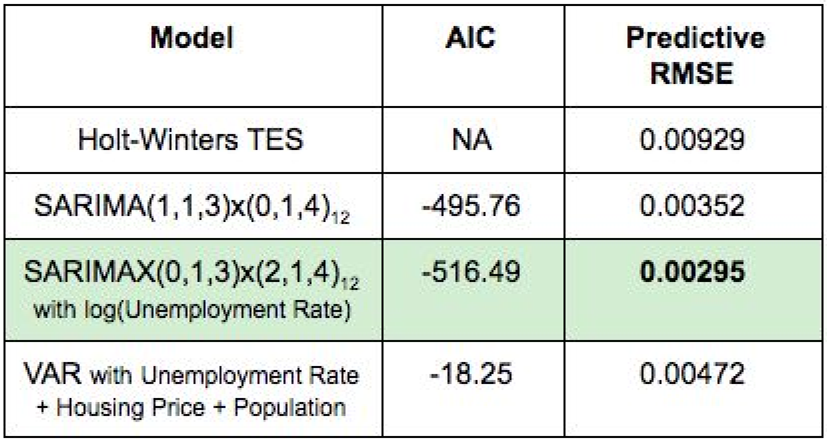


Figure 4: Subsetting the data

Model Selection:

We built a number of different models on the training set using many of the approaches discussed above. We only kept the models that passed all residual diagnostics tests. Ultimately, we found that the best model was a SARIMAX with one additional variable - unemployment rate. The parameters of SARIMAX were chosen based on ACF and PACF plots (included in the appendix) and are selected according to our goodness-of-fit metrics.

It is easy to overfit time series models by choosing very complex models. In order to avoid this, we did not choose parameters greater than p = 4 or q = 4 for our SARIMA models (Figure 7). Below is the table of best model from each approach based on AIC and predictive RMSE:



Model	AIC	Predictive RMSE
Holt-Winters TES	NA	0.00929
SARIMA(1,1,3)x(0,1,4)$_{12}$	-495.76	0.00352
SARIMAX(0,1,3)x(2,1,4)$_{12}$ with log(Unemployment Rate)	-516.49	**0.00295**
VAR with Unemployment Rate + Housing Price + Population	-18.25	0.00472

According the table, the best model is a SARIMAX(0,1,3)x(2,1,4)$_{12}$ where log(unemployment rate) is used as an exogenous variable. This model passes the model assumptions (details of the residual diagnostics tests are outlined in the appendix). As such, we can generate valid predictions and 95% prediction intervals.

Prediction:

Our final model is a SARIMAX(0,1,3)x(2,1,4)$_{12}$ with unemployment rate used as an exogenous variable. We can now use this model to make predictions on the test set to forecast[2] the bankruptcy rates for the years 2011 and 2012. Figure 5 displays the forecasted data for the bankruptcy rate with our mean forecast in blue, the 95% prediction intervals in dark grey and 80% prediction intervals in light grey.

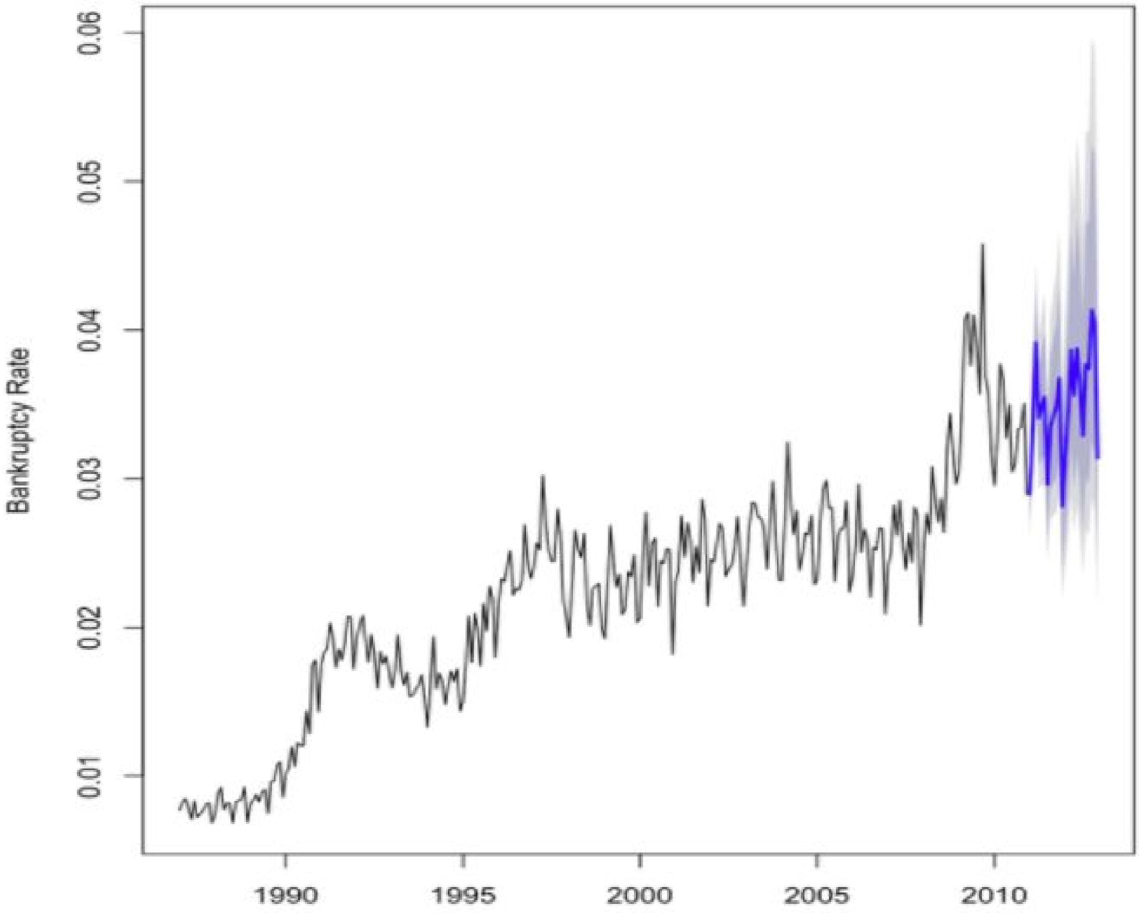


Figure 5: Forecasted bankruptcy rates

Appendix:

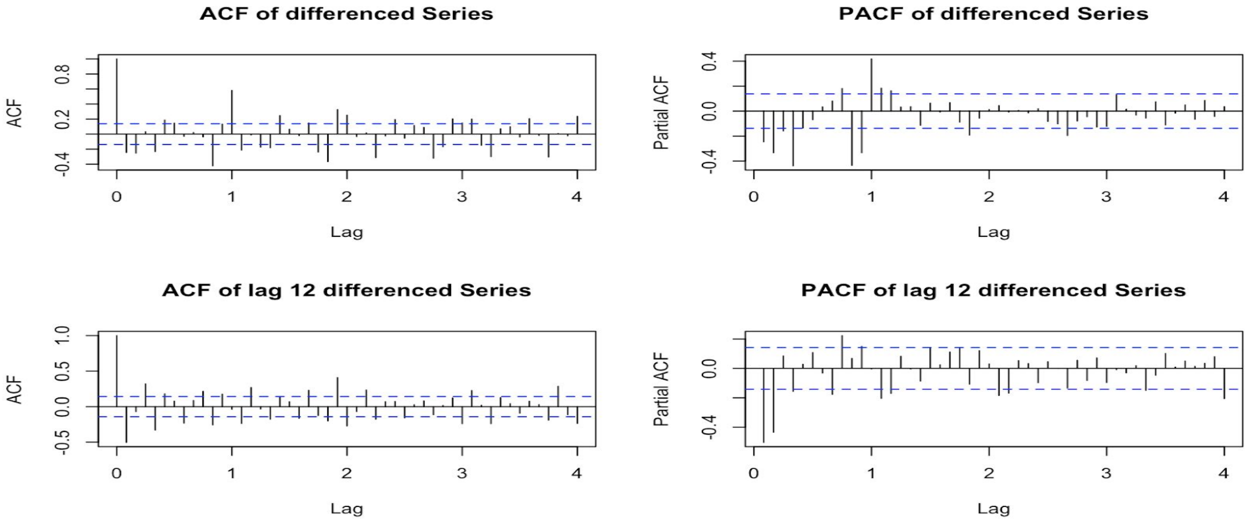


Figure 6: ACF and PACF plots of bankruptcy data

SARIMA MODEL SELECTION:

Below are all of the models that we tried with along their respective AIC[17] and RMSE[18] (the lowest AIC and RMSE in bold, and the top two optimal models are highlighted in green).

Model	**Original unemployment rate**		**Log(unemployment rate)**	
	AIC	**Predictive RMSE**	**AIC**	**Predictive RMSE**
SARIMA(2,1,3)x(2,1,2)$_{12}$	**-524.56**	.00345	**-525.31**	0.00353
SARIMA(2,1,4)x(2,1,2)$_{12}$	-522.77	.00347	-523.52	0.00357
SARIMA(2,1,3)x(1,1,3)$_{12}$	-522.28	.00354	-523.06	.003618
SARIMA(0,1,1)x(3,1,3)$_{12}$	-517.62	.00350	-518.04	.003233
SARIMA(0,1,3)x(2,1,3)$_{12}$	-517.95	**.00314**	-518.46	**.002297**
SARIMA(0,1,3)x(2,1,4)$_{12}$	-515.96	**.00313**	-516.49	**.002295**
SARIMA(0,1,1)x(3,1,2)$_{12}$	-505.28	.00339	-505.86	.003361
SARIMA(0,1,1)x(2,1,2)$_{12}$	-500.76	.00335	-501.65	.003321

Model Diagnostics:

Below are the tests to see if the optimal SARIMAX(0,1,3)x(2,1,4)$_{12}$ model meets all of the necessary assumptions.

Zero-Mean

We will examine the plot of residuals vs. time and conduct a t-test to formally check if they are mean-zero.

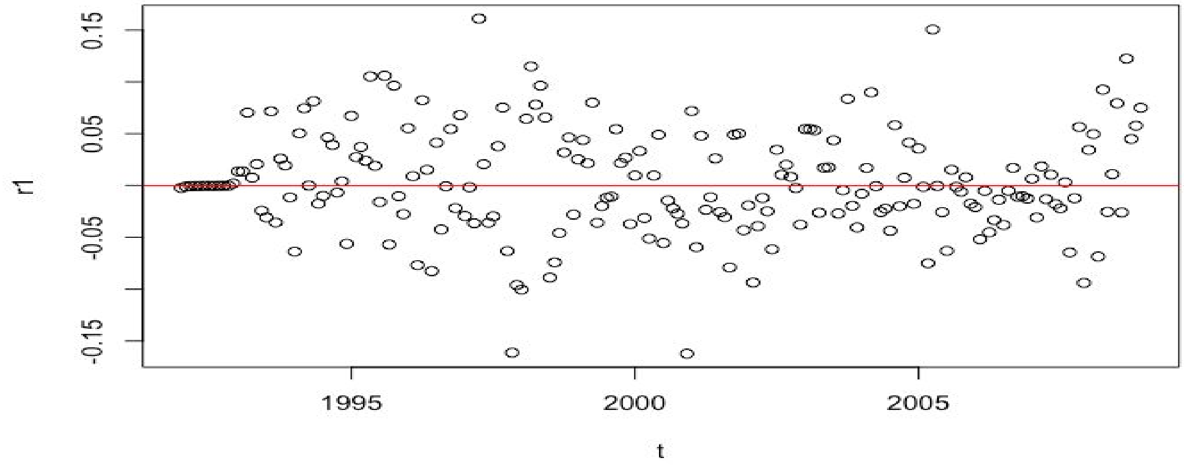


Figure 7. Residual vs Time plot

- From the plot it appears that the residuals are mean-zero
- T-test suggests that mean of residual is 0 at significance level of .05 (p-value: 0.31)

Normality

For normality we examine the Q-Q plot of residuals and formally conduct a Shapiro-Wilk test.

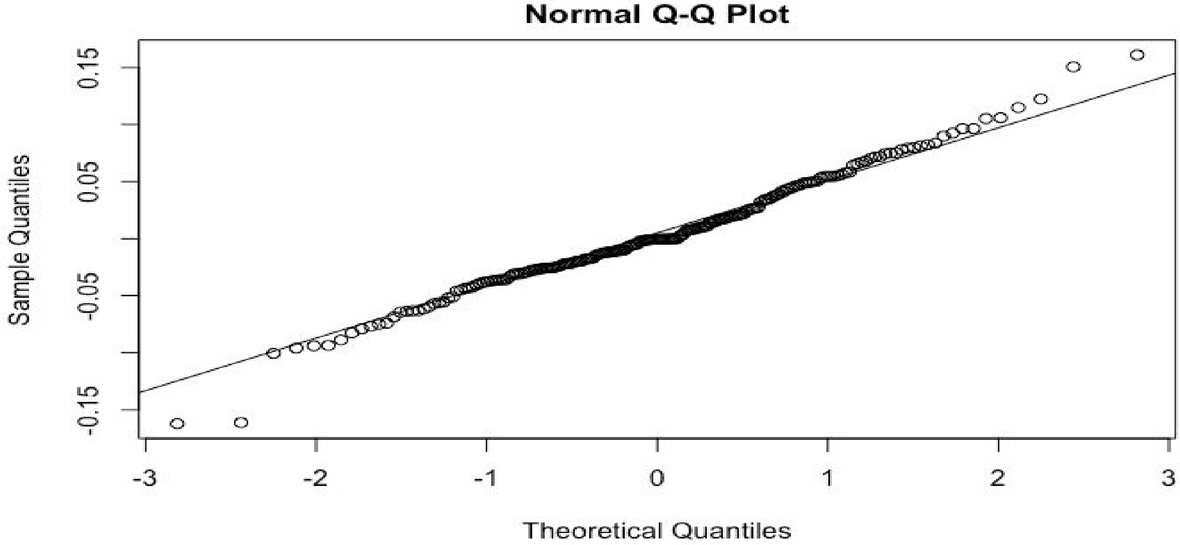


Figure 8: Q-Q plot of residuals

- The Q-Q plot appears to suggest normality except for some outliers at the tails.

- The Shapiro-Wilk test suggests borderline non-normality (p-value: .04824) but given the Q-Q plot we will assume the residuals are normally distributed.

Homoscedasticity

For homoscedasticity we will examine a plot of residuals vs. time and formally conduct Bartlett's test.

- Looking again at figure 7, it appears that the data is homoscedastic.
- Bartlett's test gives us a p-value of 0.09 which is above the significance level of 0.05. Hence we cannot reject the null hypothesis of homoscedasticity.

Zero-correlation

For zero-autocorrelation we will examine the ACF plot of residuals and conduct a Ljung-Box test.

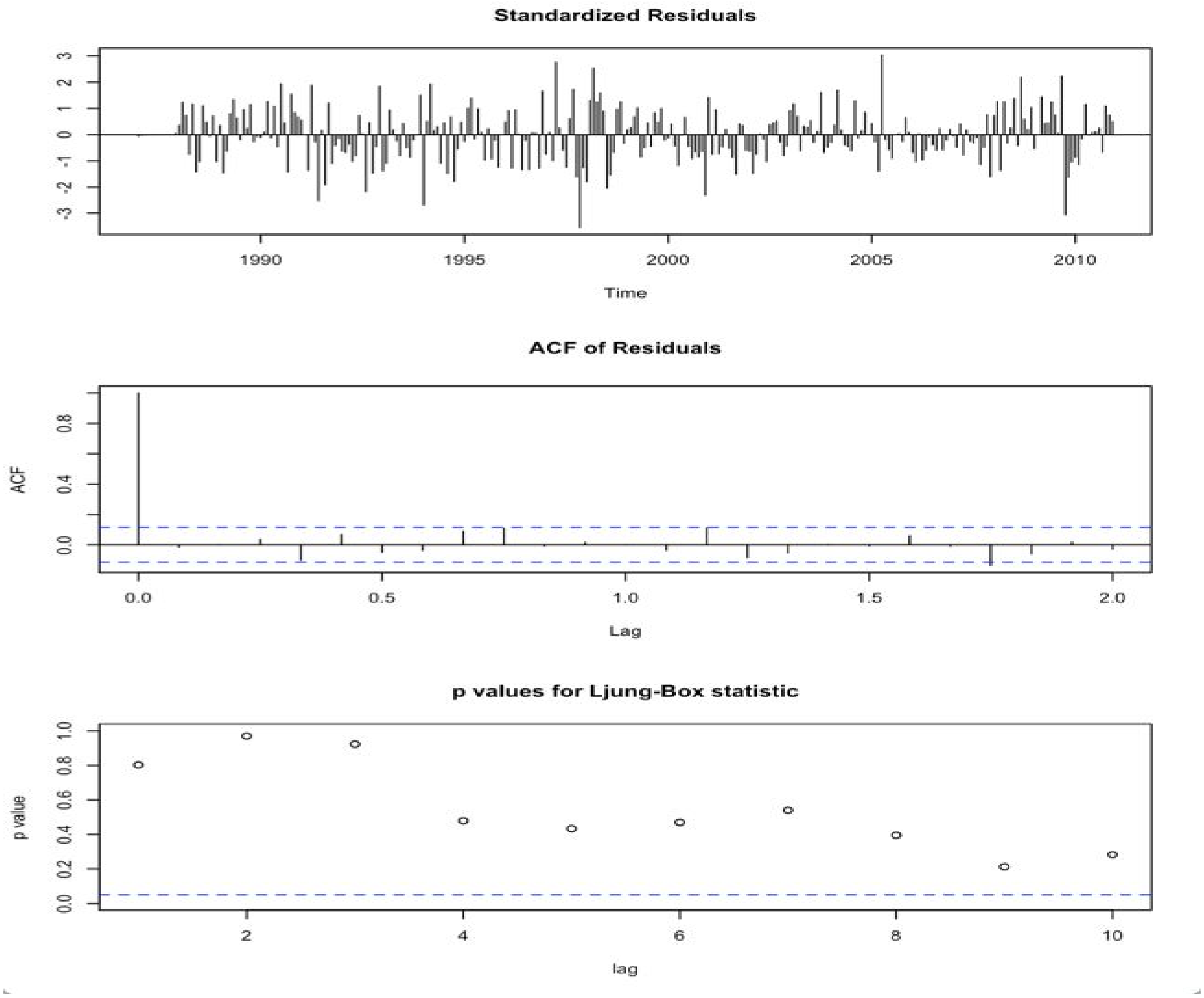


Figure 9. Plots from Ljung-Box test

- Ljung-Box test suggest zero-autocorrelation which is also evident in ACF plot with almost no significant spikes.

Hence, the model satisfies all assumptions.

Important Definitions and Test Descriptions:

Here are some key definitions that would go a long way in understanding our data and its analysis:

1. Time series: A collection of data points corresponding to temporal measurements of some measurable quantity
2. Forecasting: Use of a model to predict values of a response variable given history of previously observed data points.
3. Stationarity: A time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Generally, there are two types of stationarity: strong and weak. For this report, when we say stationary, we are referring to weak stationarity
4. Trend: Consistent directional movement in a time series
5. Seasonality: The regular and predictive fluctuations according to some period. Seasonality is quantified by m, where data points that are m periods apart are related to one enother
6. MA models: Moving-average (MA) models specify that the output variable depends linearly on the current and various past values of a stochastic (imperfectly predictable) term. (source Wiki) MA models have an order of q, which signifies how many prior observations are needed to predict future values
7. AR models: Autoregressive (AR) models are stochastic processes used in statistical calculations in which future values are estimated based on a weighted sum of past values. (source Investopedia) AR models have an order of p, which signifies how many prior observations are needed to predict future values
8. ARMA modesl: Autoregressive moving average (ARMA) models are built from a combination of AR(p) and MA(q) models
9. SARIMA models: A SARIMA model is a combination of AR and MA models that are used when there is seasonality in the time-series
10. Exogenous: External variables are considered exogenous if they influence the response variable but are not influenced by the response variable
11. Endogenous: External variables are endogenous if they not only influence the response, but also are influenced by the response
12. SARIMAX model: The SARIMAX models are SARIMA models with explanatory variables. SARIMAX is appropriate when dealing with exogenous variables
13. VAR model: Vector autoregression (VAR) models are ones where all variables are treated symmetrically. VAR is appropriate when dealing with endogenous variables
14. Single Exponential Smoothing: A Holt-Winters set of recursive equations that is used when a time series exhibits neither trend nor seasonality
15. Double Exponential Smoothing: A Holt-Winters set of recursive equations that is used when a time series exhibits trend but not seasonality
16. Triple Exponential Smoothing: A Holt-Winters set of recursive equations that is used when a time series exhibits both trend and seasonality
17. ACF: Autocorrelation between different lags of time-series
18. PACF: Partial autocorrelation between different lags, with effects of correlation between intermediate lags being accounted for.
19. AIC: The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. The smaller the AIC, the better the model.

AIC = -2(log-likelihood) + 2(p+q+1)

20. RMSE: The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values.
21. LRT test: The likelihood ratio test is a statistical test used for comparing the goodness of fit of two statistical models, one of which (the null model) is a special case of the other (the alternative model).

$$\mathbf{LRT} = -2\log_e\left(\frac{\mathcal{L}_s(\hat{\theta})}{\mathcal{L}_g(\hat{\theta})}\right)$$

22. Bartlett's test: To formally test for heteroscedasticity of residuals. High p-value indicates that the model has constant variance. This is sensitive to non-normality of residuals
23. Levene's test: A more robust alternative to Bartlett's test, where high p-value indicates that the model has constant variance.
24. Ljung-Box test: This is used to test whether the ACF is equal to zero for a variety of lags.
25. Shapiro-Wilk test: This test is used to check if the residuals are normally distributed. High p-values indicate normality.