Box-Jenkins Methods

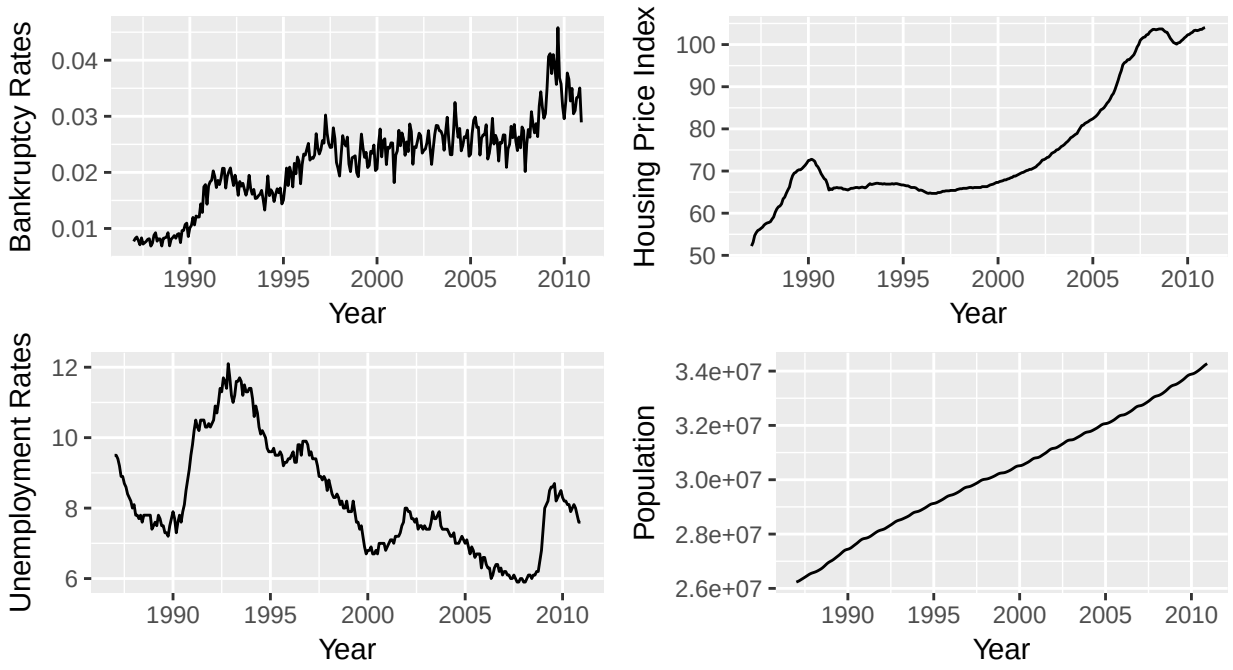
Box-Jenkin models involves statistical theory and modeling to analyze and forecast time series data. The naming standard for the various types of time series models consists of acronyms defined as the following:

- **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)

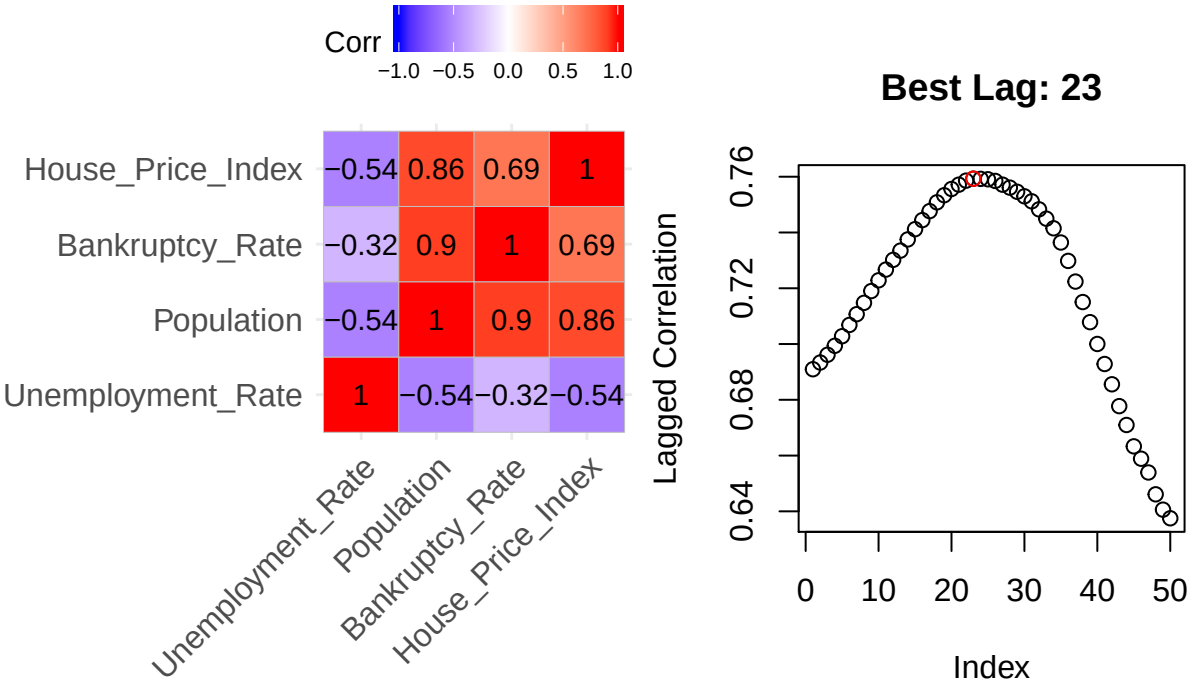
So, SARIMAX means *Seasonal Autoregressive Integrated Moving Average with Exogenous Variables*

In this part of the project we will explore ARIMA, SARIMA and SARIMAX models in order to predict bankruptcy rate. ARIMA, or in general, SARIMA models are univariate models which Wdepend on previous time series data and try to predict future values. Through this part we will check these models with their assumptions. 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.

We separated the data into a training set [1987-2008] and validation set [2009-2010], which will allow us to evaluate the predictive accuracy of our models before predicting on the unlabelled test set [2011-2012]. Here we can see time series that are provided in training data. Housing Price Index and Population seems to have a positive correlation with Bankruptcy where as Unemployment has a negative one.



There is great correlation between `House_Price_Index` and `Bankruptcy_Rate`, probably even a higher one 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.

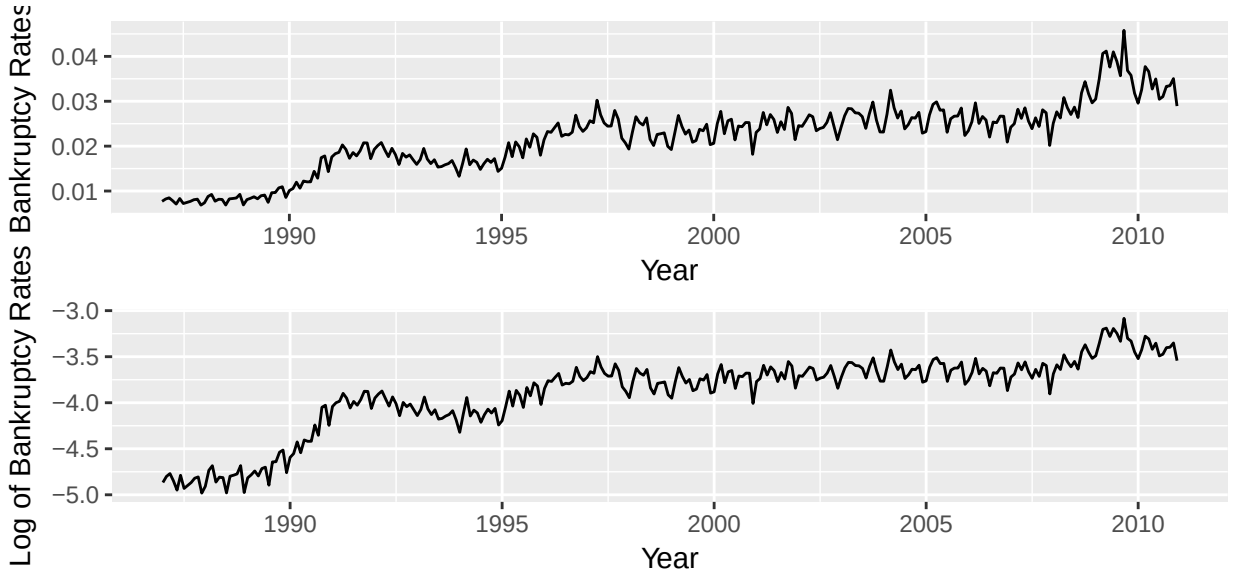


Lagged Correlation Plot h vs Correlation

We observe 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 in our SARIMAX model. Another assumptions we are making with SARIMAX is that any regressed variable during modeling is an exogenous variable, meaning that they have a uni-directional effect on dependent variable bankruptcy rate but not the other way around.

SARIMA MODEL (Univariate Bankruptcy)

Every predictive modeling task has a evaluation metric in order to assess the performance of different type of models and in order to pick the best available model that we hope to generalize to our hypothesis. Also, during these predictive modeling tasks we create a hold-out set which is also our validation set. We will use the last 2 years of our data [2009-2010] in order to assess our models with the evaluation metric root mean squared error (RMSE). 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.



Here, we observe a change in scale as we move forward in our time series, and this might be a problem during checking the constant variance of residuals (the difference between the true value and the predicted value), which is an important assumption when we fit our model with Maximum Likelihood Estimation (MLE). We generally make a transformation and look at the plot again to see if this change doesn't occur anymore; some common transformations are log, square root or in general boxcox transforms.

So we apply a log transform to our data to see if it becomes better in terms of constant variance over time. The change in variance is not as bad as before, so we will proceed our analysis with the transformed version of our time series model.

During time series modeling another important matter that one should to pay attention is stationarity of data, which means that the mean and autocovariance functions are not dependent on time t. This is important since ARMA models account for only stationary data, so we will try to decompose our time series first, then decide our parameters p, P, q, Q in order to feed our data into a SARIMA model. There are other types of models such as exponential smoothing models which takes care of seasonality and trend with the given parameters α, β and γ. But these models are subsets of general SARIMA models. So it's always better to pay good attention to SARIMA models since they can be more powerful in terms of capability of capturing many different combinations of patterns. The parameters we are trying to optimized is briefly described below:

- **p:** parameter for the order of the AR component within a month
- **q:** parameter for the order of the MA component within a month
- **P:** parameter for the order of the AR component between months
- **Q:** parameter for the order of the MA component between months

In order to make our time series stationary, we will use a well-known test called the Augmented Dickey-Fuller Test. Basically, this test will check if the unit roots lie outside the unit circle, which is an indicator of stationarity.

We are able to pass the ADF Test after performing ordinary differencing once, followed by seasonal differencing.

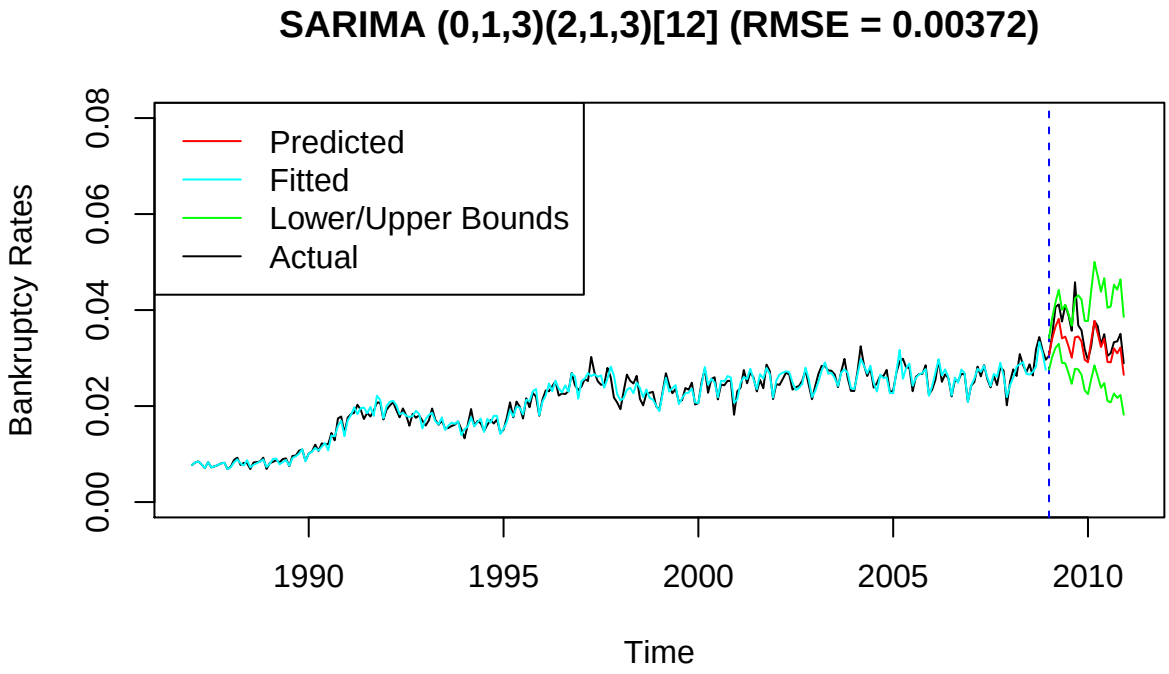
- *Ordinary differencing*: subtracting observations from the previous observation in time
- *Seasonal differencing*: subtracting observations from the observation 1 seasonal period in time prior (i.e. 12 months ago)

Note, that when performing the ADF Test, we make sure to check 4 years ahead so that we are not misled by the default parameters. After the differencing, we are now confident with 99% confidence level that the time series is indeed stationary.

To calculate the parameters of our final model, we performed a grid search where p, P, q, Q can range from 0 to 3, thus resulting in $4^4 = 256$ different SARIMA models. For each SARIMA model, we calculated the

RSME after forecasting 2 years ahead (24 data points). Then, 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 (i.e sum of $p + q + P + Q + 1$), the simpler the model will be.

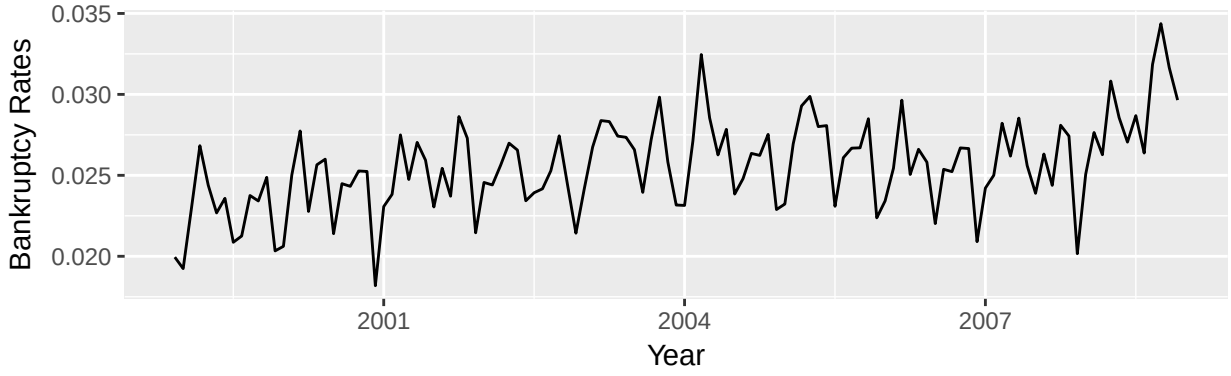
The best SARIMA model on the full data that we ended up choosing is SARIMA $(0,1,3)(2,1,3)_{12}$.



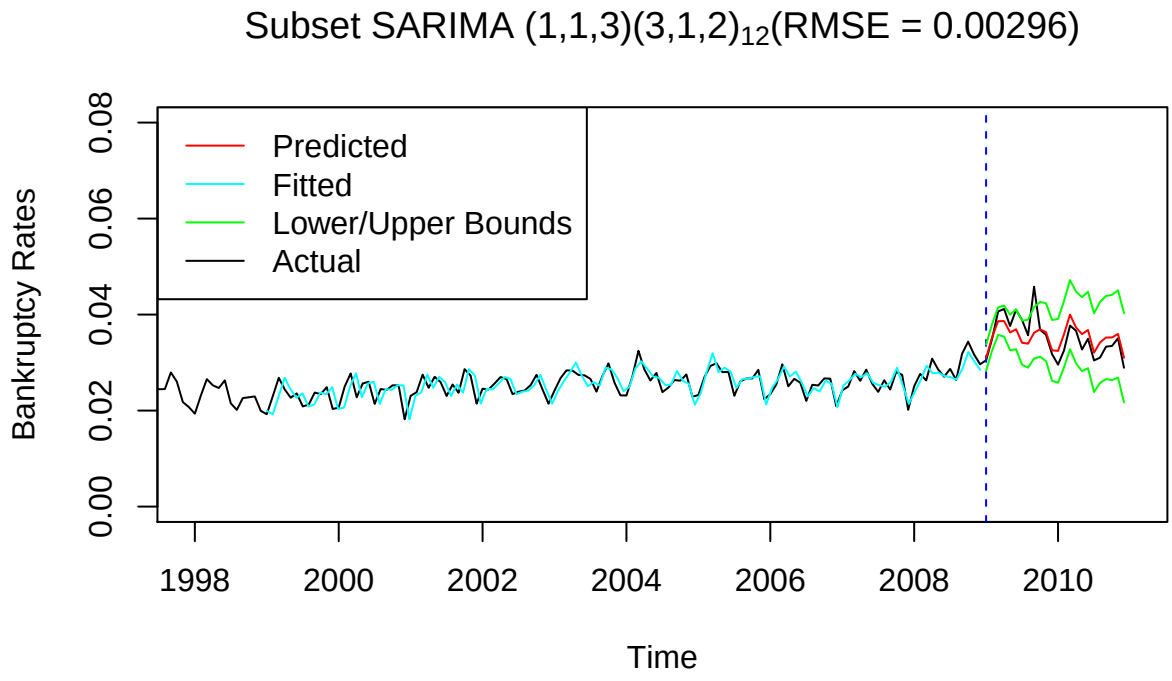
Subset SARIMA model

In this part, we will try out a hypothesis that our data in fact is not stable, in other words it has a clear and instant pattern change around after about 10 years. 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.



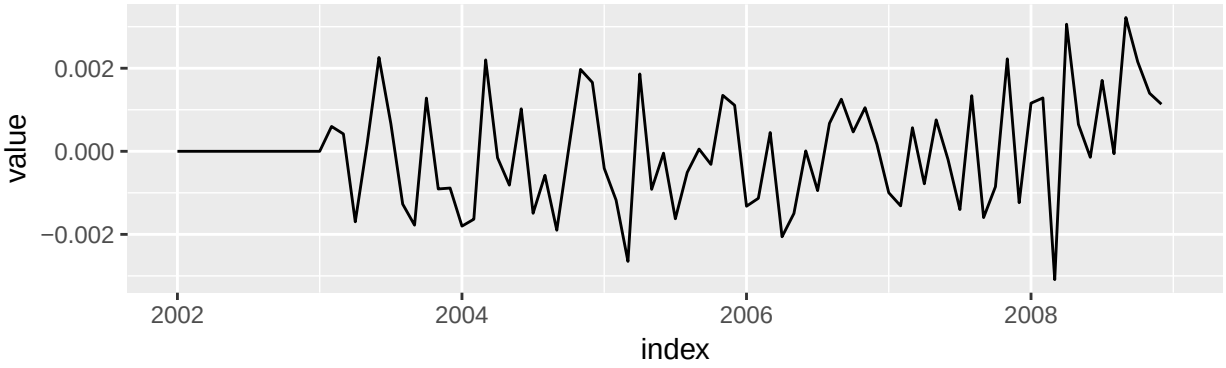
We perform another grid search and observe that best parameters are $(1, 1, 3)(3, 1, 2)_{12}$ with RMSE of ~ 0.0029 which is lower than our SARIMA model on the full data. Another important note is that since our main goal is to come up with the best predictive model we choose our optimization method as Least Squares Estimation (LSE) rather than MLE.



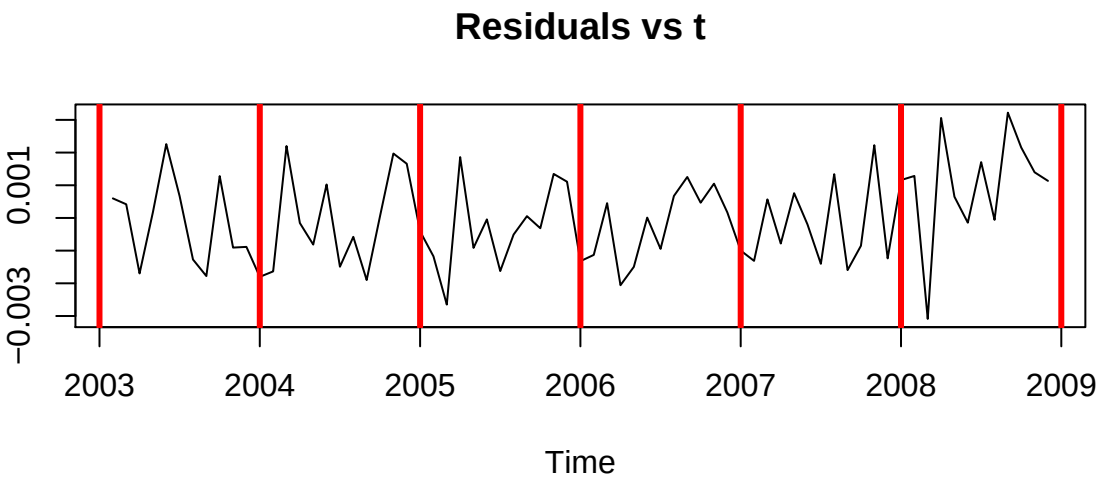
Checking Box-Jenkins Assumptions

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

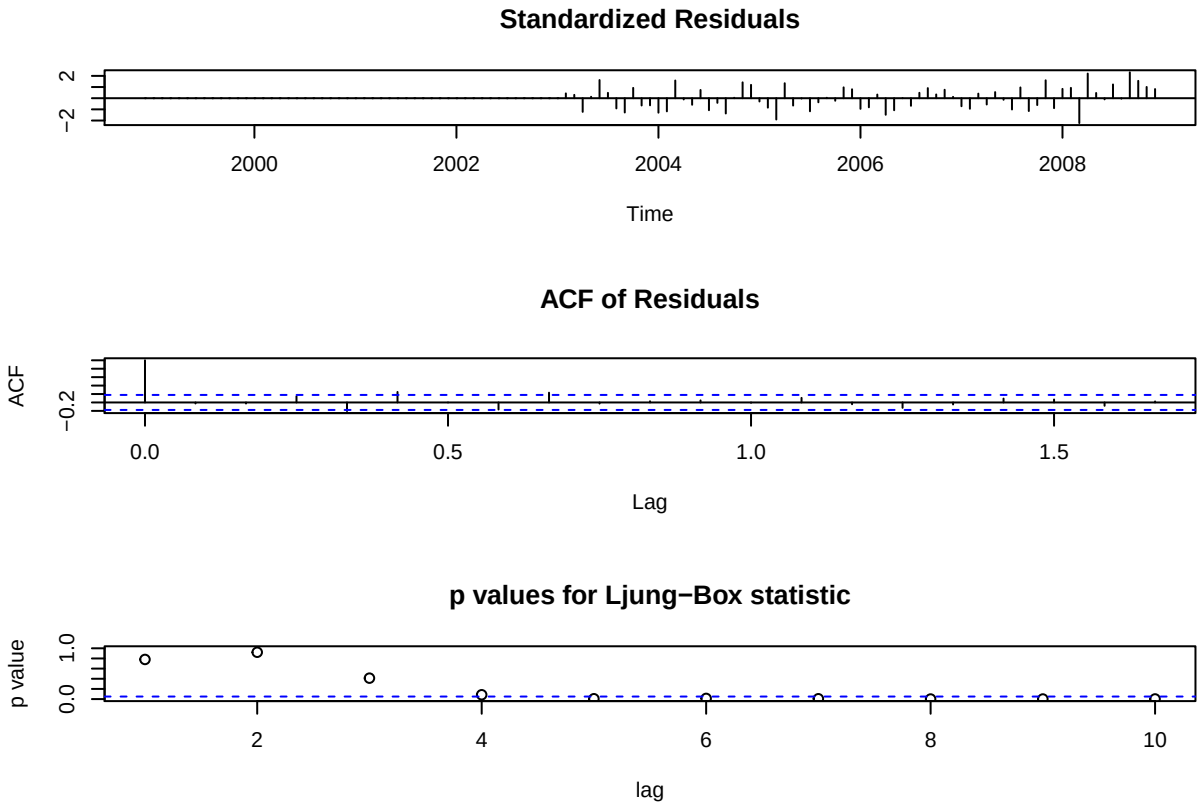
- Zero-Mean: the residuals have a mean of zero
- Homoscedasticity: the residuals have constant variance
- Zero-Correlation: the residuals are uncorrelated
- Normality: the residuals are normally distributed (no need to check for our subset model which is based on Least Squares rather than Maximum Likelihood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



We can see if the Zero-Mean assumption passes by looking at this plot. The residuals look to be more or less centered around 0, which is good.



Homoscedasticity assumption looks good because the residuals look to be fairly constant over time. The red lines simply divide the time series plot into equal groups, to see if the residuals look more or less the same within each group.



The Zero-Correlation looks to be a slight problem starting at lag 4, as we see in the plot at the very bottom. Note that we tested the three assumptions through formal tests and get the same results as indicated above. The Zero-Correlation assumption is satisified if all of the points are above the blue line.

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**. But, what exactly does do these terms mean?

Trend: A trend exists if there is a long-term increase or decrease in bankruptcy rates. As seen previously, there has been an overall decrease from 1987 to 2010. If a trend does not exist, the pattern would look more or less flat. To put it more simply, trend can be thought of as the *slope*. We see that the trend is a slow increase with an ending decrease after 2009.

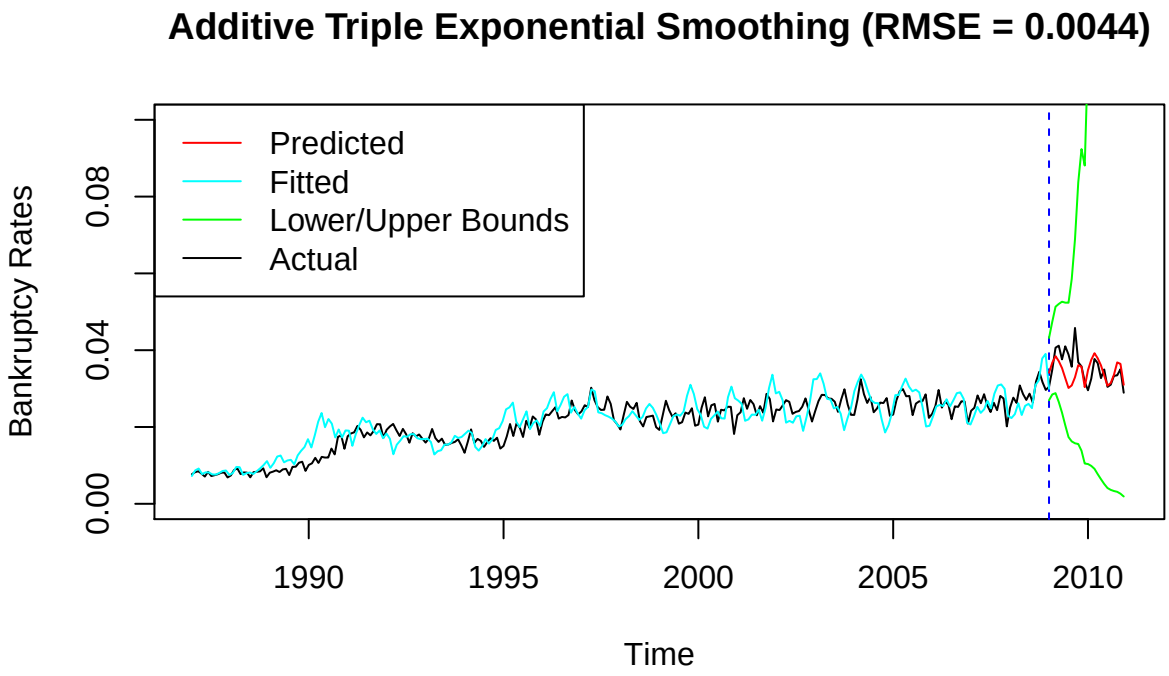
Seasonality: Seasonality is when the time series exhibits similar behavior at regular intervals, or *seasons*. Seasons can be quarterly, monthly, day of the week, etc. In our scenario, bankruptcy rates are recorded every month, and therefore the period of a season is 12 months (1 year).

Furthermore, because we choose to use an *additive* method because the size of the peaks are roughly the same throughout the time series. The overall concept behind triple exponential smoothing is to apply exponential smoothing and incorporating the level, trend, and seasonal components. While the trend is the *slope*, the level can be treated as the *intercept*.

Next, the amount of smoothing to be done needs to be calculated. Every time series behaves differently and thus require different set of smoothing parameters. There are three smoothing parameters and are the following: level(α), trend(β), and seasonality(γ). These parameters range from 0 to 1 inclusive, where values close to 0 represent *extreme* smoothing and values closer to 1 represent *no* smoothing. To decide the optimal values of α, β, and γ, we used an iterative approach. We tried values from $0.05, 0.10, ..., 0.95, 1$ for each of α, β, and γ so a total of $20^3 = 8,000$ combinations. 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.

Best Holt-Winters Model

Our best model for Holt-Winters consists of $\alpha = 0.25, \beta = 0.65, \gamma = 0.35$. Because $\beta = 0.65$ (higher means less smoothing) *level* and *seasonality* is the most important when it comes to prediction. These parameters are based on what gives us the lowest RMSE.



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.

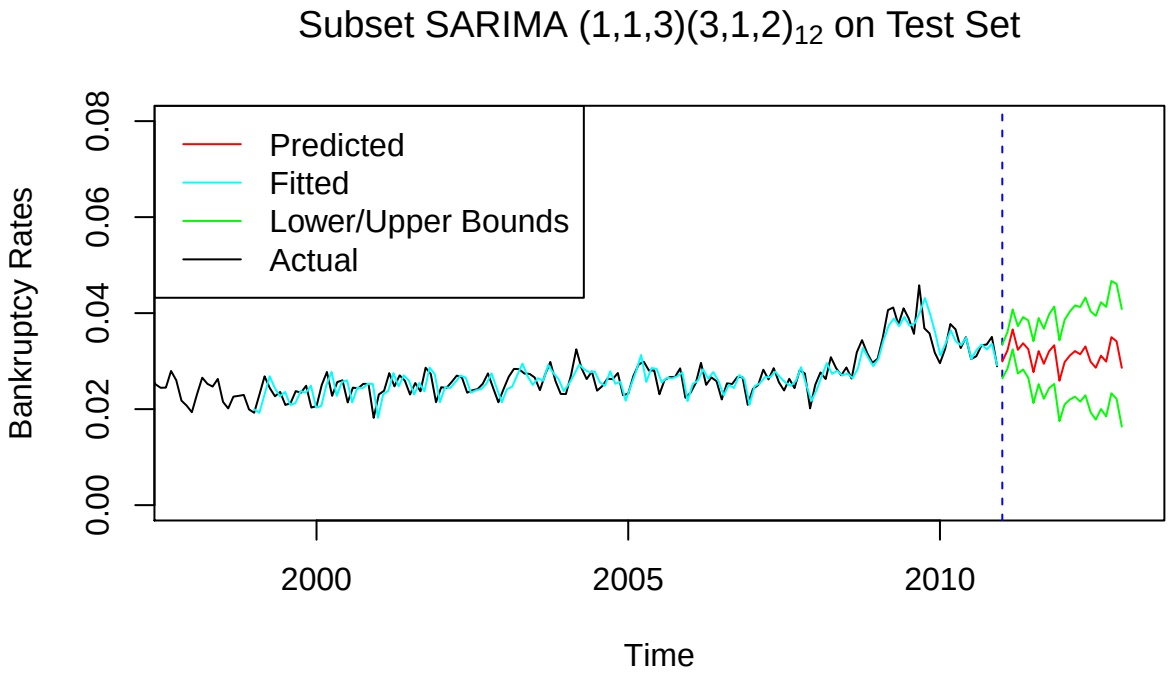
Conclusion

Comparison of RMSE 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 $(\alpha = 0.01, \beta = 1, \gamma = 0.04)$	0.0044
SARIMA $(0,1,3)(2,1,3)_{12}$	0.00372

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