TSA Project Write-Up
11 December 2015
Piyush Bhargava, Chuiyi Liu, Cong Qing, Meg Ellis

INTRODUCTION

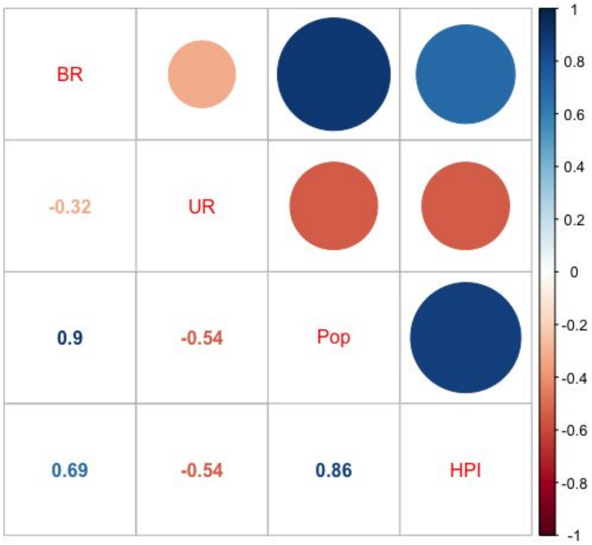
Bankruptcy rate is one of several variables that can serve as an indicator of an economy's well being. High bankruptcy rates are something than any city, state, country, etc. wishes to avoid. In order to do so, it would be useful to know how the past values of bankruptcy rates influence futures rates and what other macroeconomic factors influence bankruptcy. In knowing these macroeconomic factors and how their behaviors affect bankruptcy, it might become possible to anticipate future bankruptcy rates. For this particular problem, we are given three such macroeconomic variables to examine their potential impact on bankruptcy. After examining the variables making a few observations, the question becomes how to model the said data in order to best predict the rate of bankruptcy.

THE DATA AND THE PROBLEM

The dataset appears to be centered around a single location, of which we want to predict future bankruptcy rates. The data was collected once a month, starting in January 1987 up until December 2010, with each observation showing a snapshot of the variables in that given month i.e. 288 data points. We assume that bankruptcy rates takes into account the bankruptcy of corporations as well as individuals. As mentioned earlier, in addition to bankruptcy rate, we are provided three other macroeconomic variables -unemployment rate, house pricing index (HPI), and population that may impact bankruptcy. Each of the variables are numerical, so we do not need to worry about creating any dummy variables to replace categorical data. This will perhaps make for better interpretability of the model that we create. Though it seems that each variable would be useful in predicting the future of bankruptcy rates, it is important we start the process without bias towards which will be the best estimator or which will wield the greatest predictive power. Our goal is to use the final model to predict the monthly bankruptcy rates for the year 2011. The 2011 monthly information for other three macroeconomic variables can also be used.

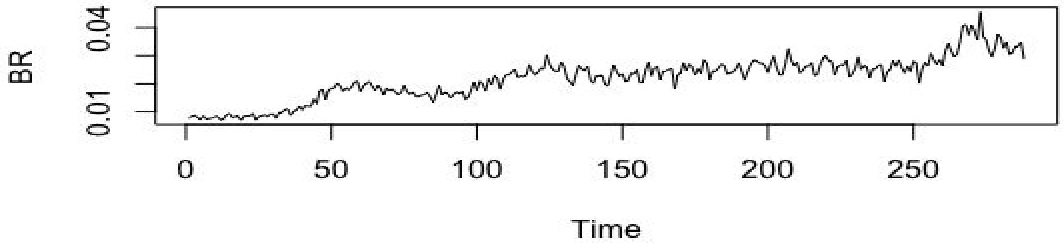
DATA EXPLORATION

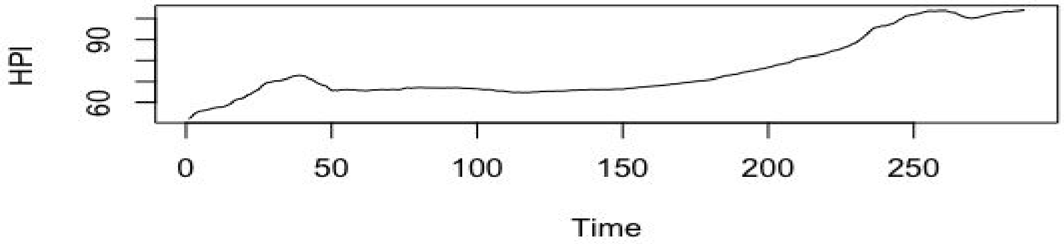
As a preliminary check, we look at the correlation between different variables to get an idea of the relationships that exist in the dataset.

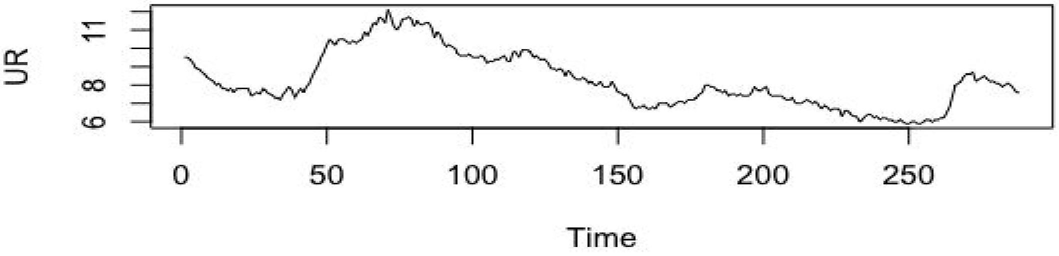


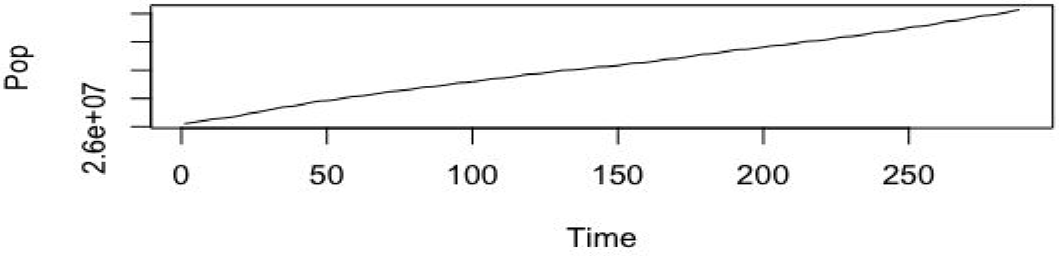
As can be seen from the plot above, there is a fair amount of collinearity among variables. Bankruptcy rate, Population and HPI are highly correlated with each other. Given these results, it is not likely that we will use both, HPI and Population to model Bankruptcy rate.

Following are the plots of the variables against time:









The above plots indicate that Population may not be a good variable to model bankruptcy rate since it has an increasing trend with time and seems to be dependent only on it. Bankruptcy plot looks similar to HPI than other variables.

MODELING TECHNIQUES

Suppose someone wanted to argue that there was no trend to the data, or rather, as how can we be sure that there is. If a time series appears to have no trend but still has irregularities that need to be addressed, smoothing techniques can be used. That is to say that we forecast using smoothing methods rather than creating a model using the data we are given, which is our intention. In order to exhaust our resources, and ensure that we are predicting as best we can, we attempted a few smoothing techniques just in case, although we felt that plots and tests in R revealed the modeling and prediction should be handled differently.

The Exponential Fitting smoothing technique is a sort of weighted moving average approach that uses a constant, alpha, as a smoothing parameter. Alpha penalizes data that is from further back in time, or rather, gives it less weight based on the notion that data closer to the actual observation will be able to better predict it than data that is further away from it. It is important to note though that the Simple Exponential fitting does not estimate trend. Double Exponential Fitting however does, and Triple estimates both trend and seasonality. We attempt only Simple and Double Exponential Fitting (an attempt at Triple returns an error that the time series has less than 2 periods). Using a training and test set approach, we evaluate the forecasting abilities of the Simple and Double Exponential Fitting. Simple actually performs rather well in terms of returning a low error rate, but it predicts the same Bankruptcy rate for each of the points in the future, with varying confidence intervals. This does not really make sense and beyond that, we can see that the data has trend and we want to use a method that accounts for said trend. Double Exponential Fitting can estimate trend, but in our case, it returned a lower accuracy rate than the models we attained for the approach we ultimately end up using.

We also considered using OLS (Ordinary Least Squares) regression as we could have used other macroeconomic variables to model bankruptcy rates. Since, bankruptcy rates is a time series and its value at any point is dependent on past values, OLS assumptions would have been violated if we had built that model.

MODEL SELECTION

There are a variety of methods available to model this data set in order to make predictions on the Bankruptcy variable. There are certain clues that help to decide which of these is best. For instance, in just plotting Bankruptcy Rates vs. Time, it is clear that the time series exhibits some sort variation that is not due to random chance, but rather is attributable to a certain increasing trend. This implies that we do not want to just use smoothing techniques. The dependency in data will have to be decomposed to independent components and certain transformations will have to be made prior to modeling, because in order to build a model that forecasts with any kind of accuracy and precision, we require an assumption that something (a noise component) doesn't vary with time.

While just looking at the plot is telling, there is also a statistical test that tells us as much, the Augmented Dickey-Fuller (ADF) test. "Passing" this test means that the time series doesn't have a trend and is in fact stationary (it has both a mean and variance that are unaffected by shifts in time). But the data in the problem is not stationary and the results of the Dickey-Fuller test imply that we might want to difference the model. Ordinary differencing, or more specifically, differencing consecutive values, is applied to help remove any trend or dependence of series on time and Seasonal differencing is applied to remove any seasonal /periodic trends. It stands to reason that we want a model that includes the differencing process. This model is the Seasonal Auto-Regressive Integrated Moving Average Model also called SARIMA.

SARIMA model has seven individual parameters. Parameter (s) pertains to the length of seasonal periods. We examine Autocorrelation Function (ACF) plot to choose ‘s’. ACF plot shows the correlation between observations at different time points. ACF plot exhibits seasonality and indicates periodicity. Examining the plot in our case shows s = 12 i.e. the period is monthly. This makes sense intuitively also since bankruptcy rates may be different during different months of year. Two other parameters pertain to the process of differencing the data. Data can be differenced to remove both trend (d) and seasonality (D). These two parameters are set to the number of times that ordinary differences and seasonal differences will have to be performed to remove trend and periodic fluctuations i.e. to make the time series stationary. Performing ADF test after every difference indicates whether we have differenced enough. If the test indicates non-stationarity, we should keep differencing. ACF plot can also be used to indicate if the series has become stationary. The plot should exhibit exponential decay once stationary indicating that a trend no longer exists.

In our case, ADF Test implies that differencing should occur so we difference once. In plotting to see if the one difference has improved conditions, we notice that the variance of the differenced time series plot is not constant and can perhaps be addressed by using logarithmic scale (i.e. taking the log of the bankruptcy rates) and then differencing. Both the ACF plot and the Augmented Dickey-Fuller test confirm that no more ordinary differencing needs to take place. Though the graph has spikes at certain intervals but they are not sustained indicating that no differencing for seasonality is required. Other evidence, for example, a built-in test that provides a number of seasonal differences required to stabilize the dataset - ‘nsdiffs’ also shows that no differencing for seasonality is required. As such, in addition to seasonal period (s = 12), we know at least two more parameters of our model - those pertaining to the trend and seasonal differences: d = 1 and D = 0.

In order to find the remaining parameters ‘p’,’q’,’P’ and ’Q’, SARIMA model can be thought of building a “within” the season (year in this case) model with parameters p and q and a “between” the season model with parameters P and Q. “Within” the season model repeats for every year. Since the period is monthly, “between” the year model will be same for every month. Now, we examine ACF and PACF plots. Similar to ACF, PACF also measures the relationship between the data at two different time periods, otherwise called lags but it does so after having accounted for any possible relationships with observations at intermediate lags. These can both be used to examine if the current data point depends on the past values. There is a simple heuristic to determining the values after which the data is no longer correlated. In examining the ACF for q (represents Moving-Average component of “within” the season model) , it appears that there are two significant spikes prior to decay which would imply q = 2. Similarly, based on the same criteria for the PACF, p (represents Auto-regressive component of “within” the season model) also appears to equal 2. i.e. the point under consideration is correlated with 2 immediate lags. Although we applied no seasonal differencing, we can still have parameters P (represents Auto-regressive component of “between” the season model) and Q (represents Moving-Average component of “between” the season model). Observing both the ACF and PACF it would seem that there is seasonality where the period is monthly, or s = 12. We then look for spikes on both plots on multiples of 12. In the ACF there are two – at lags 12 and 24 indicating that Q=2. In the PACF graph there is only one at lag 12 indicating that P=12. Now that we have assigned values to each value in a SARIMA model, we can build the model and ensure that it passes the assumptions, applying diagnostics or trying different values if it does not.

We try the model with two different methods – least squares and maximum likelihood. The former is more robust to non-normality, while the later must have normally distributed residuals. However, the results returned from both methods are very similar; that is to say, the coefficients assigned to each parameter by the model are similar for both methods. This observation implies that we can in fact assume normality and use the ML method. All of the coefficients for this model are significant except for the second MA coefficient. This in addition to a desire to be thorough leads us to attempt other models with different values for p, P, Q, and q. Using log-likelihood, AIC, and sigma-squared values as criteria; we are looking for low AIC values, high log-likelihood values and small sigma-squared values. Model 2, which sets p = 2, d = 1, q = 1, P = 1, D = 0, and Q = 2, is the model we ultimately choose. Though this model seems ideal, we must first ensure that it passes all the necessary assumptions required by the Maximum Likelihood method. These assumptions each pertain to properties of the residuals i.e. the values we get from subtracting the model predicted bankruptcy rates from the actual bankruptcy rates in the dataset.

The first assumption is that the residuals have zero mean. This can be examined by plotting the standardized residuals vs. time, and based on the plot that assumption appears to hold, as the values seem to be evenly plotted around the x = 0 line. A regular hypothesis test also confirms this. The residuals vs. time plot addresses the assumption of residuals’ constant variance or homoscedasticity. The formal hypothesis tests - Bartlett’s and Levene’s tests for homoscedasticity also confirm our assumption of constant variance with 95% confidence. The third assumption is that of the residuals’ zero correlation with lags. For this assumption, we use the Ljung-Box Test, the Runs Test and the ACF plot of the residuals. The assumption holds since the Runs Test returns a p-value above .05 meaning that we are 95% confident that the null hypothesis of uncorrelated residuals holds. It is worth mentioning however that at lags 9 and 10, the Ljung Box Test does not pass, but since the other lags are alright and the Runs tests results satisfies the assumption then we say that the model does indeed satisfy the Zero correlation assumption. Finally, we have the assumption of normal distribution for the residuals. Using both the Shapiro-Wilk Test and the QQ Plot of the residuals, we confirm that this assumption holds at 95% confidence level.

Now that we know that the SARIMA model we picked is a viable choice, we can consider the covariates: the other macroeconomic variables in the data set. As previously discussed, because we want to avoid collinearity, it is a distinct possibility that we will not use all of the variables. It was established in a previous section that ‘Population’ and ‘HPI’ are both highly correlated with bankruptcy. They are also pretty correlated with each other. So which do we use? Since ‘Population’ only seems to have an upward trend we chose ‘HPI’. The other variable to consider is ‘Unemployment’, which we will initially include it in the covariate model and see if these two work well together. We refit the SARIMA model we have already chosen with the addition of these two covariates.

This model shows that ‘Unemployment’ is not a significant variable since its coefficient is not statistically different from zero at 95% confidence level. As such, we eliminate it from the model, and refit again this time only using ‘HPI’ as a covariate. This particular model returns a significant coefficient for ‘HPI’ as well as a better sigma-squared, AIC, and log-likelihood than the SARIMA model we created without covariates. Still, we must ensure that the assumptions hold. We perform residual diagnostics again using the same process outlined before with the solely-SARIMA model, and, just as before, each assumption does in fact hold. Since both the purely SARIMA and the SARIMA with covariates model pass these assumptions, we can trust the validity of any predictions made by this model.

Prior to moving on to the actual forecasting, the ultimate goal for this model, it is important to point out that we attempted ARMA-GARCH method for this data, thinking there might be volatility, but ultimately, these attempts did not outperform the model outlined above in terms of the root-mean-square error (RMSE) on ‘test’ (10 percent hold out sample) data. Since we are using this model for prediction, we place higher importance on performance of model on a ‘test’ data than other model fit measures such as AIC and log-likelihood.

FORECASTING

Since we have two models that work well, we choose our model by comparing its ability to predict bankruptcy rates on data we already have. The data is divided into a training and test set, where we fit the model using the training data and use this newly fitted model to predict the bankruptcy rates in the test set. We use a rolling window approach, so that we are using the most “up-to-date” data (even though the most “recent” is a prediction) to make a more informed prediction. We predict one at a time, add that prediction to the existing dataset, refit the model, and predict using this new model that had incorporated the new prediction. Comparing the error between different models, we finally choose the SARIMA model with HPI as a covariate. Here we take the logarithm of HPI as we did with bankruptcy rates.

After deciding our final model, we use it for prediction. The plot of the previous time series and the predicted with its confidence interval is in Figure.1. The tabular form of the forecasting result and its confidence interval is in Table.1.

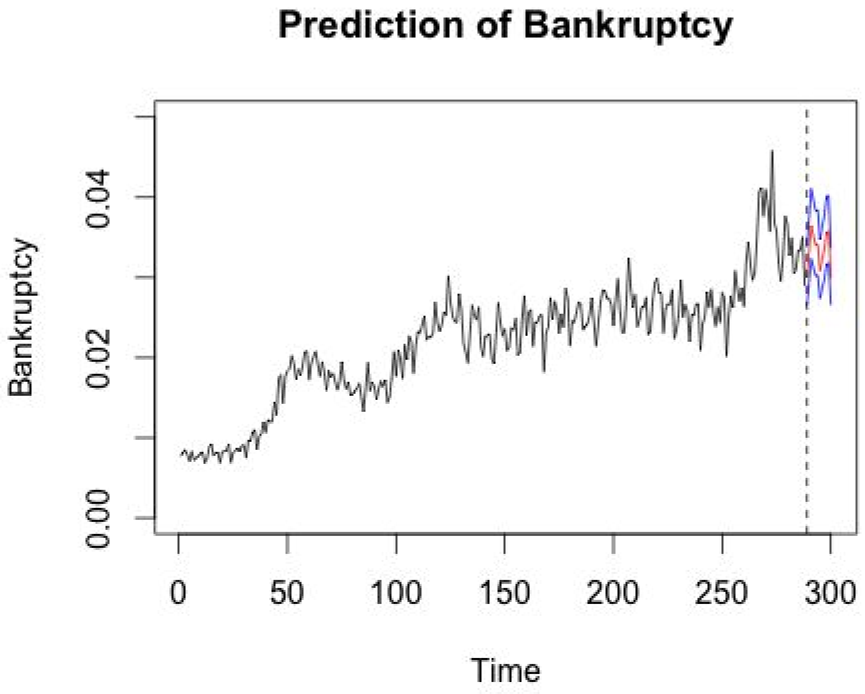


Figure. 1 Prediction of Bankruptcy

	mean	lower	upper
1	0.02995982	0.02656665	0.03378637
2	0.03319114	0.02943894	0.03742158
3	0.03643629	0.03232468	0.04107087
4	0.03518602	0.03122348	0.03965143
5	0.03397475	0.03015488	0.03827851
6	0.03409939	0.03027207	0.03841061
7	0.03082918	0.02737474	0.03471954
8	0.03224749	0.02864027	0.03630904
9	0.03346011	0.02972342	0.03766656
10	0.03556058	0.03159570	0.04002301
11	0.03573947	0.03176185	0.04021522
12	0.02987100	0.02655187	0.03360504

Table.1 Prediction of Bankruptcy

CONCLUSION

Our approach predicts bankruptcy rates using the past values of bankruptcy rates and the values of HPI. Bankruptcy rate has a positive relation with HPI in the model which means that as HPI increases, Bankruptcy rate will increase. If HPI increases, real estate may become expensive and unaffordable for people. This may result in higher bankruptcy rates. Our model also incorporated seasonal variations during the year in bankruptcy rates. Typically, month of ‘March’ had highest bankruptcy rates during any year. Since the financial year ends in month of ‘March’, higher number of companies may be filing for bankruptcy during that time. So, accounting for seasonal variation in our model looks reasonable. We have also tried to keep our model simple and ensure that the dependency of bankruptcy rates on its past values doesn’t go too far back in time. The model also has certain limitations. We haven’t considered ‘HPI’ as time-series but as a normal covariate. To build a more powerful model, we should have made both ‘Bankruptcy rate’ and ‘HPI’ stationary and then built the model. Alternately, we could have considered Co-integration approach using which we could have modeled the combination of the two time-series. Also, a better model could have been built if data for more covariates was available.