Time Series Final Project

Group Members: Xinke Sun, Jenny Kong, Adam Reevesman, and Minchen Wang

Introduction

In this report, we will use time series modeling methods to forecast three years of the monthly national bankruptcy rate in Canada (January 2015-December 2017). We will use the last 28 years (January 1987-December 2014) of monthly bankruptcy rates, as well as the population, unemployment rate, and housing price index from January 1987 to December 2017. After exploring the available methods, we use a SARIMAX(5,1,6)x$(1,0,2)_{12}$ with explanatory variable Unemployment Rate (the observation occurring 18 months before) as our optimal model. Using this model, we generate forecasts for Canada's bankruptcy rate for January 2015-December 2017.

Part 1. Problem Description

Time series forecasting is the use of a statistical model to predict future values of a variable based on its history. Time series have applications in government, industry, and commerce. The data provided by this project is suitable for time series modeling and forecasting because it contains observations of the several variables that were measured consistently at equally spaced points in time.

1.1 Data Overview

Two datasets will be used. The first is meant for training a model. It contains observations of four variables: population, unemployment rate, housing price index and bankruptcy rate, measured monthly between January 1987 and December 2014. The second dataset meant for testing the model. It has the monthly records of each of the variables above except the bankruptcy rate. three variables from January 1987 to December 2014, except for the bankruptcy rate which is to be predicted.

By combining the two datasets, we can look at the trend of each variable during the recorded years.

- Population (01/1987 - 12/2017)

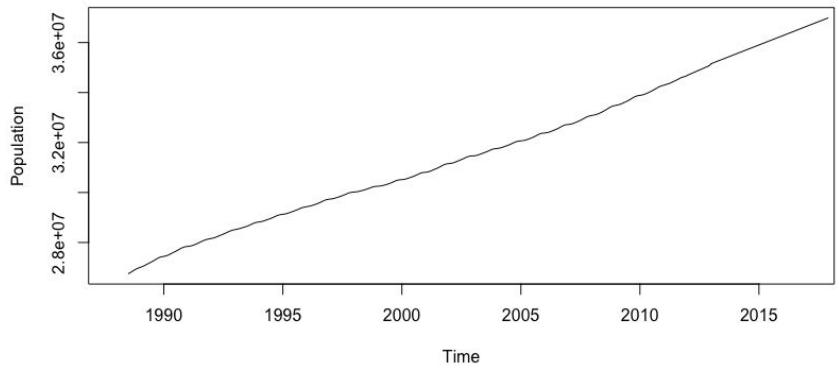

Fig. 1.1. Canada's population, 1987-2017.

- Unemployment Rate (01/1987 - 12/2017)

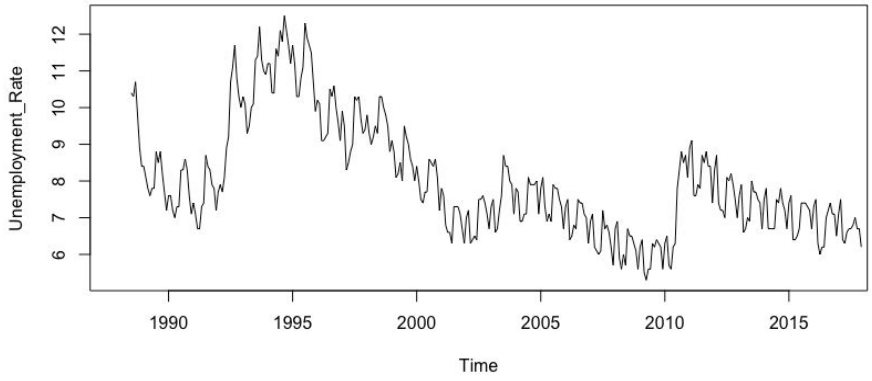

Fig. 1.2. Canada's unemployment rate, 1987-2017.

- Housing Price Index (01/1987 - 12/2017)

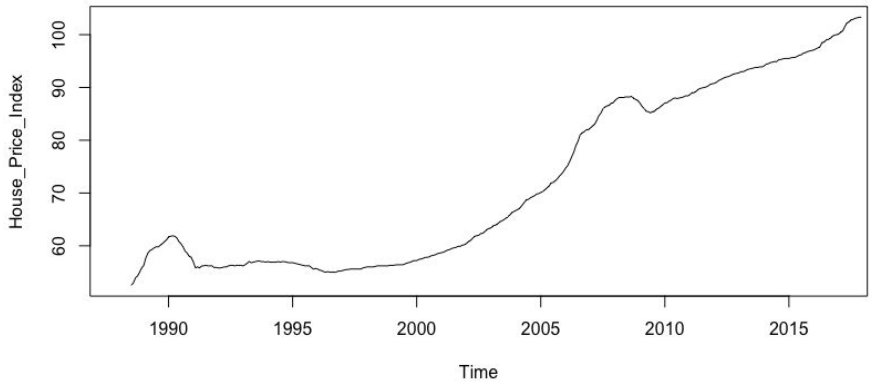

Fig. 1.3. Canada's house price index, 1987-2017.

- Bankruptcy Rate (01/1987 - 12/2014)

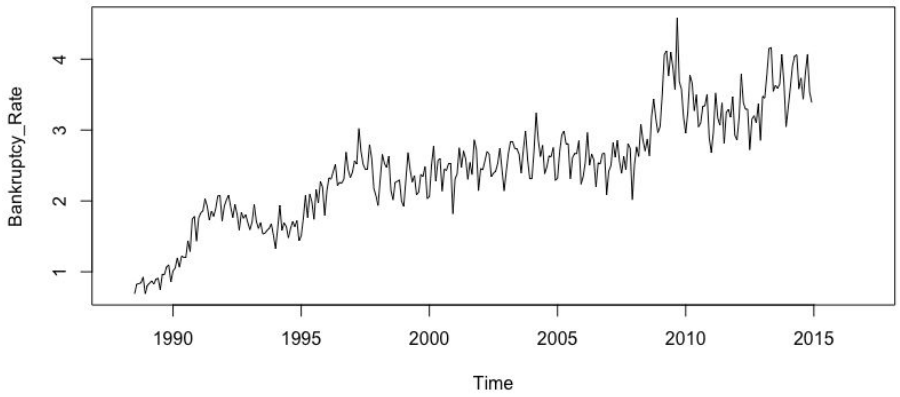

Fig. 1.4. Canada's bankruptcy rate, 1987-2014.

1.2 Project Goal

Based on all the data we have, we aim to choose one optimal time series model to accurately forecast national monthly bankruptcy rates from January 2015 to December 2017 for Canada, which is of interest to national banks, insurance companies, credit-lenders, politicians and so on.

Part 2. Available Methods

In general, there are two kinds of time series models. Univariate models are those that use only the history of the target variable to predict its future values. Multivariate models incorporate measurements of other variables that may add predictive power. There are a variety of methods for fitting univariate and multivariate time series, but one natural approach is to start with univariate models then determine if adding more variables can improve the model's accuracy. Therefore, we will first attempt to forecast future bankruptcy rate using only past values. Then, we will check if models that consider the population, unemployment rate, and housing price index are more useful.

2.1 ARIMA and SARIMA

SARIMA stands for Seasonal Autoregressive Moving Average. The difference between these two models is that ARIMA is often used when there's an obvious up or down trend exists while SARIMA is more properly used when there is a repeating pattern in the data (this is called seasonality). We depend on graphs and plots to give us a first idea as which model should we use. However, there are also formal tests we could use to help us make our decision.

2.2 Holt-Winters

Holt-Winters modeling is another univariate approach and uses what is called "exponential smoothing". These models create up to three equations that are governed by up to three parameters (single, double, or triple exponential smoothing). More specifically, if we observe no trend or seasonality, we use single exponential smoothing (SES). If we believe the data shows trend we use double exponential smoothing (DES), and triple exponential smoothing (TES) should be used when seasonality is present. Using these models does not require us to make many assumptions about the data. Unfortunately, due to the simplicity, they might fail to model complicated data well.

2.3 ARIMAX and SARIMAX

These are multivariate models. The X here in the names represents for "exogenous". An ARIMAX model is simply an ARIMA model topped with exogenous explanatory variables, which means that instead of only predicting the target variable from its own historical record, we include other explanatory variables to help better explain and forecast the target variable. SARIMAX model to a SARIMA model is also the same way that ARIMAX model to an ARIMA model. However, the most important thing to remember is that an exogenous explanatory variable means that it is influencing the predictor variable but not the other way around. The influence in uni-direction, not bi-direction in which case we'll use VAR instead, as we'll discuss later.

2.4 VAR

VAR stands for Vector Autoregression and is another multivariate model. The biggest difference between a SARIMAX model and VAR model is that the explanatory variables in the VAR framework are considered "endogenous". This means we assume that they influence the target variable and the target variable also influences them. In order to model this type of relationship, a VAR model does not differentiate the target and the explanatory variables. In fact, it creates one equation for each variable that describes how it depends on past observations of the others. Furthermore, if we believe that both exogenous and endogenous variables exist, a VARX model can be considered.

Part 3. Optimal Model

To find the optimal model to forecast, we need to understand the nature of our data (as described in the first section). We attempted each of the methods described above and chose to model the data using SARIMAX(5,1,6)x(1,0,2)$_{12}$ + Unemployment Rate (based on the observation occurring 18 months before).

3.0 Data Preprocessing

In order to compare the performance among different available modeling methods, we split the training dataset even further. We hold out the data between January 2012 and December 2014 (in blue below) to use as a validation set. And we only use the data before 2012 to fit models and will now call this our training set. The models are then tested on the validation set. We compare different models by looking at goodness-of-fit values and how much the forecasted values differ from the actual value, which will be further described in more detail in section 3.2.

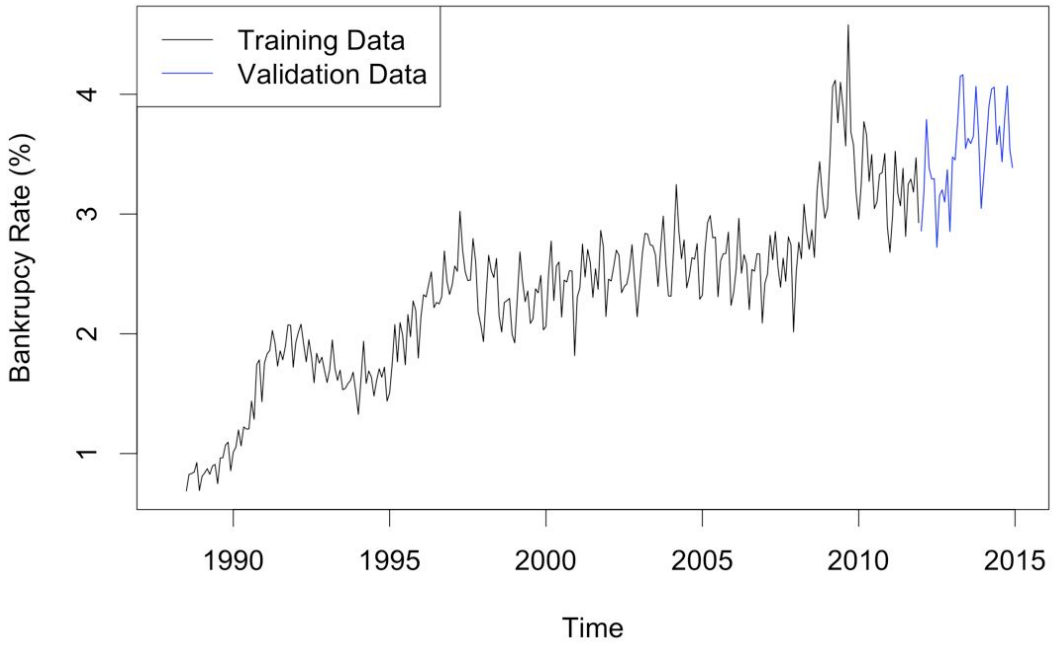

Fig. 3.1. Bankruptcy rate training (1987-2011) and validation (2012-2014) data.

Before start fitting models, we first adjust our data to make it suitable for SARIMA(X) models to fit. Because the raw training data varies differently within different time frames, and that would be troublesome to forecast. Hence, we perform a transformation (specifically, Box-Cox transformation) to the data so that the time series varies by about the same amount regardless of time. The data also show trend and cycles, which might cause non-stationarity and make the

forecast spurious. Hence, we take another transformation (the first ordinary difference) to mitigate the issue. The final result of the transformation is shown as Fig.3.2. After the transformations, we can see the variation of the data looks more constant, and the scope of variables is much smaller than before. This means that it is good for being modeled.

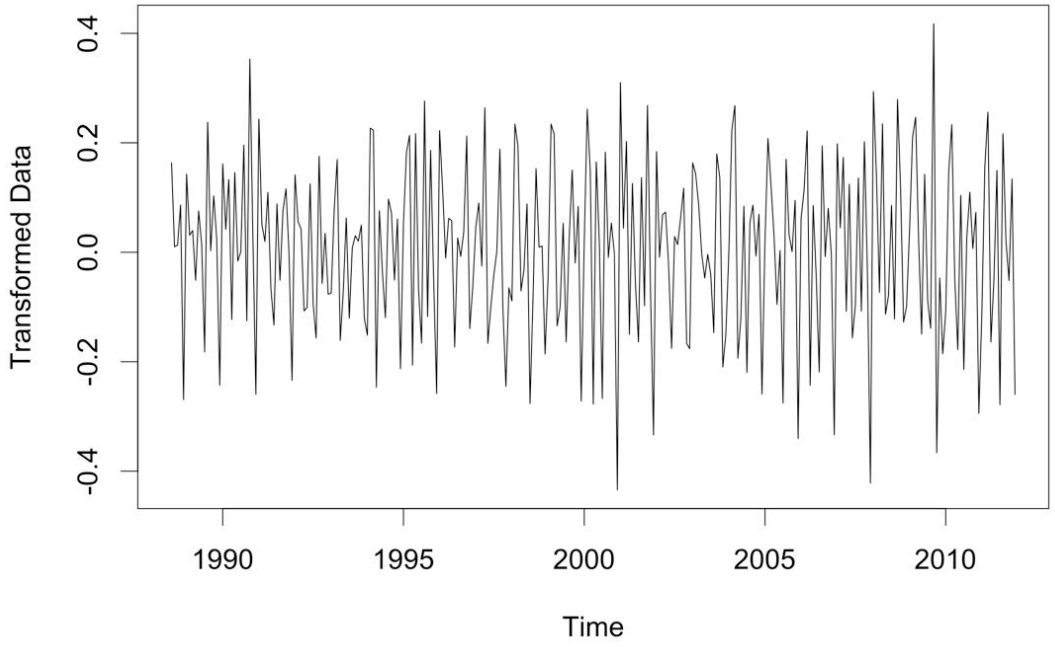

Fig. 3.2. Transformed bankruptcy rate training data.

Since the dataset also contains explanatory variables presented, we might want to include them in our multivariate models to better explain the outcome. At first thought, it is intuitive that Bankruptcy Rate, Population, Unemployment Rate, and Housing Price Index are more or less correlated to one another. Moreover, we found that the correlation between Unemployment Rate and Bankruptcy rate is actually strongest at a lag of 1.5 years. This means, for example, the Unemployment Rate in January 2000 is most correlated with the Bankruptcy Rate in July 2001 than in other years, so it would be better to reconstruct the dataset so that the stronger correlation is captured. Thus, we shift the observations of Unemployment Rate 18 months forward for comparing with all the other variables.

3.1 Evaluation Metrics

RMSE: Root Mean Squared Error (on the validation set). This is the square root of the average squared distance between each predicted value and its corresponding actual value. Taking squared distances is more common than absolute distances because it allows for nice mathematical properties. Since we want our prediction closer to the truth (in this case closer to our validation data), the lower the RMSE, the better.

Log-Likelihood: The "Likelihood function" is a measure of the probability of obtaining the dataset we have if it were generated by a model with given parameters. Finding the parameters that maximize the likelihood function means that we find the model that is most likely to have generated the observed data. The "Log-Likelihood" is the logarithm of the maximum value of the likelihood function and having a large value suggests that the model is a good fit of the data.

Other Metrics: Metrics like AIC and BIC (Akaike/Bayesian Information Criteria) are similar to Log-Likelihood but penalize models that are too complex. The complexity of the models we fit was not too drastic, so we generally found these metrics to agree with the Log-Likelihood. Therefore, we mostly compared models using the two metrics described above.

3.2 Model Selection

The potential models are the ones we described in Part 2: SARIMA, Holt-Winters, SARIMAX, and VAR. We firstly fit the training data and test the validation data in order to find the best model for each method. From the results shown as Tab3.1, we could find SARIMA and SARIMAX which have lower RMSE perform better than Holt-Winters and VAR methods.

Tab. 3.1. Best Methods Selection

Method	**Best Model**	**Validation RMSE**
SARIMA	SARIMA(5,1,6)x$(1,0,2)_{12}$	0.2125
Holt-Winters	TES (multiplication)	0.3006421
SARIMAX	SARIMA(5,1,5)x$(1,0,2)_{12}$ + Unemployment Rate	0.2164
VAR	VAR(10) + All explanatory variables	0.3703617

As for SARIMA and SARIMAX methods, we have several competitive models to choose. Therefore, we also fit the data with each potential models based on different train-validation splits and the whole original train-set to see the evaluation metrics and choose our optimal model. In this process, we also filter the SARIMA(X) models which MA part is not invertible, which means it couldn't be exclusively written as past observations and are not stable enough in forecasting. We finally could see from the Tab.3.2 that all these rest models have close outcomes, while SARIMAX(5,1,6)$(1,0,2)_{12}$ with explanatory variable Unemployment Rate performs the best in general and is relatively simple. Up to this point, we can choose it as our optimal model.

Tab. 3.2. Optimal Model Selection

	36 months test		12 months test		full train (before split)	
Model	test loglik	test RMSE	test loglik	test RMSE	full train loglike	full train RMSE
SARIMAX(5,1,5)(1,0,2) + unemploy	299.4	0.2164388	319.1	0.2245777	332.45	0.1511835
SARIMAX(5,1,6)(1,0,2) + unemploy	299.09	0.2167993	322.79	0.1468894	333.43	0.150336
SARIMAX(5,1,6)(1,0,4) + unemploy	299.28	0.2214541	325.47	0.1561526	336.12	0.1484518

3.3 Model Interpretation

The SARIMAX(5,1,6)(2,0,2) + Unemployment Rate is a model that forecasts the future based on previous Bankruptcy Rate observations, and also models with preceding Unemployment Rate as the explanatory variable. The Unemployment Rate used here is the observation occurred 18 months before, which means the unemployment rate which happened 18 months earlier could help us predict the new bankruptcy rate point.

Also, the optimal model estimates a sigma squared value of 0.007096, which means the variance of the data remains close to constant at this value. The log-likelihood of 333.43 is the highest among the 12-months test trial. And the full train-set RMSE is the lowest as 0.150336, which means it performs the best when fitting the train dataset. In all, we chose a model that balances a high goodness of fit score with high predictive accuracy.

3.4 Model Limitation

To check if a time series model is valid, there are certain assumptions it needs to satisfy. The plot below is one tool for assessing how well it meets those assumptions.

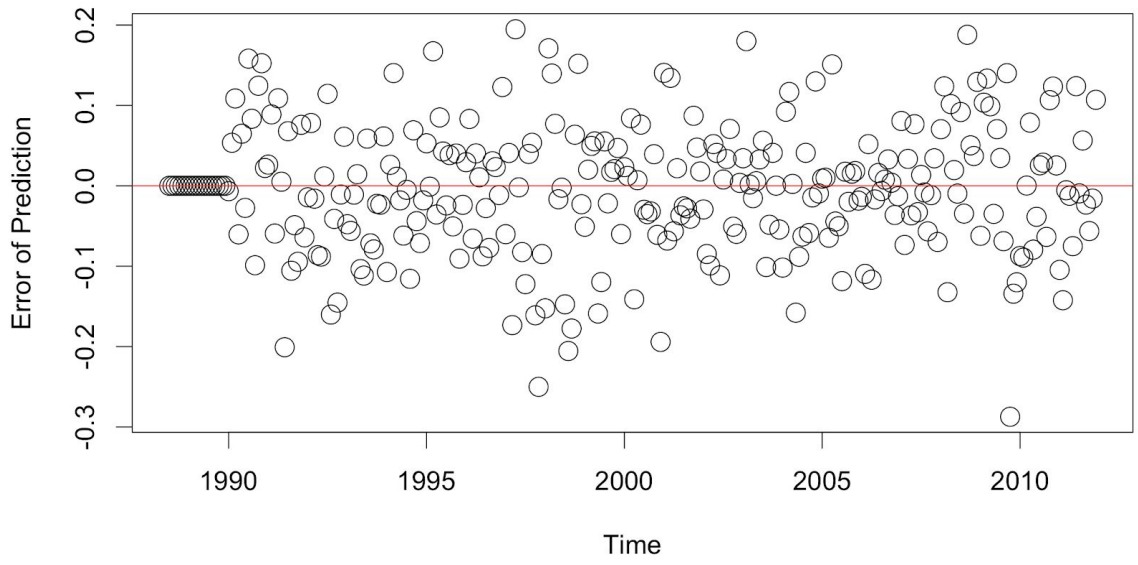

Fig.3.3. Error of predictions of the training data.

Our optimal model makes errors that are zero on average and have a fairly constant variance. In addition, the errors do not appear to be related to each other in any meaningful way. All of this suggests that it will make accurate predictions.

Since our model relied on the observation of Unemployment Rate that came 18 months previously, we could not train it on the first 18 months of data. This has the potential to decrease model performance, but we expect that the accuracy gained in using this predictor is worth the information lost.

Part 4. Forecast Result

After choosing SARIMAX(5,1,6)$(1,0,2)_{12}$ with explanatory variable Unemployment Rate (the observation occurring 18 months before) as our optimal model, we fit the model and use it to forecast the monthly bankruptcy rate in the next 36 months.

The forecast results are shown as follows.

4.1 Forecast Data

Subject to the length of the report, we only include the first 6 months' forecasting results in the table.

Tab. 4.1. Tabular depiction of forecasting bankruptcy rate *

Month	Prediction
1/2015	3.4557

2/2015	3.5967
3/2015	4.3024
4/2015	4.1623
5/2015	3.9460
6/2015	3.9855

* part of forecasting results (January 2015 - June 2015)

4.2 Visualization

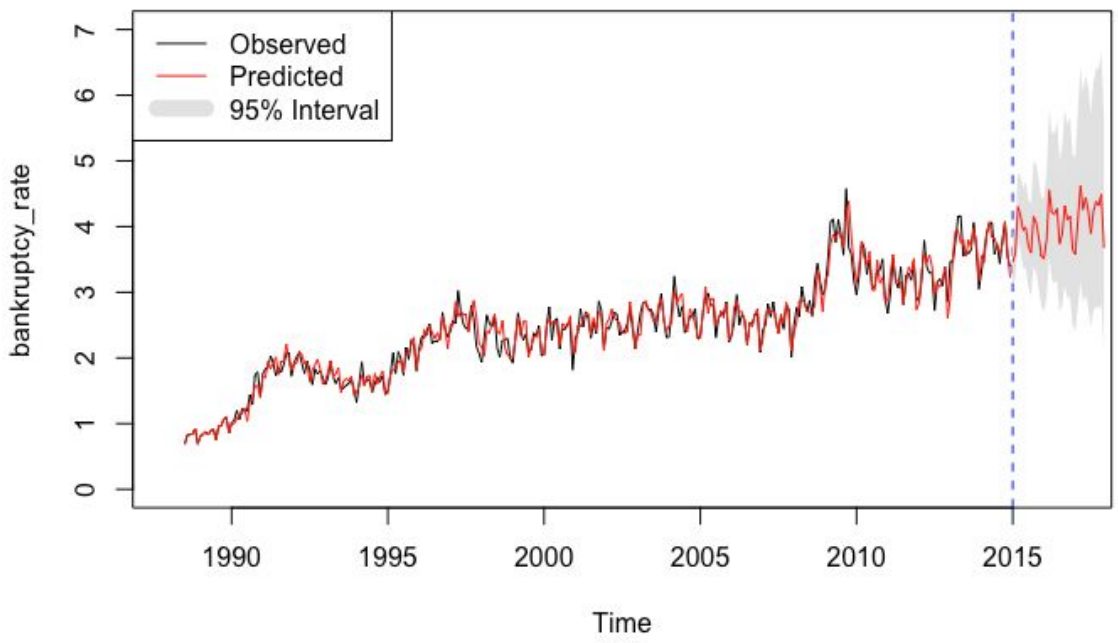

Fig. 4.1. Canada's bankruptcy rate forecasting, 2014-2017.

The forecasting result can be seen more clearly by the visualization. In Fig 4.1, the observed value of the bankruptcy rate is shown in black, and the fitted and predicted value is shown in red. We also add the 95% confidence interval of the predictions as grey zones in the picture, which means the true value of the bankruptcy rate in the predicted years has 95% probability being in this field. From the graph, we can see the model performs roughly well on the training time series, except for some extreme values. The graph of forecasted monthly bankruptcy rates shows a slow upward trend overall with some seasonal changes from 2015 to 2017.