Modeling and forecasting the default rate over 90 days in the presence of explanatory variables

Jennifer Kung
HSBC Bank, Curitiba, Brazil - jenniferkung109@gmail.com

Guilherme Zubatch da Cunha
Biostatistics Graduate Program, Maringa State University, UEM, Maringa, Brazil - zubatch@hotmail.com

Eniuce Menezes de Souza
Statistical Department, Maringa State University, UEM, Maringa, Brazil - emsouza@uem.br

Tiago de Almeida Cequeira Lima
HSBC Bank, Curitiba, Brazil - tcl_almeida@yahoo.com.br

Abstract

Nowadays, large financial institutes have been worried in investing more and more in default rate modeling. Predicting precisely the default rate can result in huge benefits for an institute. The comprehension of how and where explanatory time series influence the default rate, allow a more profitable and better control and management of the client base. In this research, the default rate (90-plus days) that represents the seriously delinquent category was investigated in the period from 2003 to 2014, two terms of the Lula government and the first term of Dilma Rousseff. SARIMA and SARIMAX methodologies were taken into account to built forecasting models, evaluating the improvement of including explanatory variables. The benefit and concerns about the model building in the presence of explanatory variables are discussed. From the analysis, SARIMAX model showed better forecasting results regarding to SARIMA model.

Keywords: SARIMAX; SARIMA; Default rate.

1. Introduction

The credit policy adopted during the two terms of the Lula government, from 2003 to 2010, and also during the present government, was a policy to achieve inflation targets. This policy is the reduction of basic interest rates that eased credit operations, massively encouraging consumption. As a result, this policy had impacts on the economy, strengthening the internal market and contributing to the maintenance to rise the Gross Domestic Product (GDP), but also increasing consumer default rate.

In this research, we intend to study the default rate over 90 days in Brazil, from Lula government period until the government Dilma from 2003 to 2014. Due to the serial correlation of these data, it is necessary to build appropriate models for time series. It is possible to estimate a model for the default rate based only on the past default information. In this kind of series the models SARIMA has been widely used (Box and Jenkins, 1970), however, in the economic scenario, other exogenous variables can influence the default rate. In this sense, the SARIMA models can be extended to the SARIMAX class so that it can incorporate information from other variables in the model (Andrews, 2013). Considering the ability to accurately predict the rate of default can result in huge benefits to an organization, we aim to identify and estimate SARIMA and SARIMAX models to forecast the default rate and compare the accuracy of the predictions of this two class of models.

2. SARIMA and SARIMAX models

For SARIMA model, the usual process of identification, estimation and validation were applied before forecasting. The the sample ACF and PACF functions can help with the selection of the model, e.g., the number and type of parameters to be estimated, i.e., $p$ autoregressive and $q$ moving average. For the estimation of models, the maximum likelihood method was used. Details can be found in Box, Jenkins and Reisel (1994). In SARIMAX model (Andrews, 2013), where exogeneous variables can be considered, the cross correlation
were evaluated to identify which explanatory variables and in which lag they should be included in the model. The required assumptions were evaluated before building the models. Differences, including seasonal, and Box-Cox transformations were applied when the stationarity and/or homoscedasticity needed to be induced. Considering \( W_t \) and \( Z_t \) stationary time series and, \( W_t = \nabla^d \nabla_S^D Y_t \) and \( Z_t = \nabla^d \nabla_S^D X_t \) \((i = 1, \ldots, n)\) are the \( d \)-th successive difference and \( D \)-th seasonal difference of the time series \( Y_t \) and \( X_t \), respectively. Thus SARIMAX model can be written as:

\[
W_t = \gamma_1 Z_{t-\alpha_1} + \gamma_2 Z_{t-\alpha_2} + \cdots + \gamma_n Z_{t-\alpha_n} + \alpha_1 W_{t-1} + \alpha_2 W_{t-2} + \cdots + \alpha_p W_{t-p} + \\
\phi_1 W_{t-S} + \phi_2 W_{t-2S} + \cdots + \phi_p W_{t-PS} + \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2} + \cdots + \beta_q \epsilon_{t-q} + \\
+ \theta_1 \epsilon_{t-S} + \theta_2 \epsilon_{t-2S} + \cdots + \theta_Q \epsilon_{t-QS}
\]

where, \( Y_t \) is the dependent variable, \( X_t^1, X_t^2, \cdots, X_t^n \) are the n’s explanatory variables (independent), \( \epsilon_t \) is a white noise process, \( a_1, \ldots, a_n \) indicate the lags that the explanatory time series \( (X_t^1, X_t^2, \cdots, X_t^n) \) influence the dependent variable \( (Y_t) \), default rate in this case.

The interpretation of \( \gamma_i \) is not as the same as in regression models (Hyndman, 2013). The lagged dependent variable in Eq 1 implies that \( \gamma_i \) cannot be interpreted.

An option is to consider the correlated errors modeled by a SARIMA:

\[
Y_t = \gamma_1 X_t^1 + \gamma_2 X_t^2 + \cdots + \gamma_n X_t^n + \eta_t
\]

with,

\[
\eta_t = \alpha_1 \eta_{t-1} + \alpha_2 \eta_{t-2} + \cdots + \alpha_p \eta_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t.
\]

In this case, the regression coefficient has its usual interpretation. There is not much to choose between the models in terms of forecasting ability, but the additional ease of interpretation.

As in the construction of the SARIMA model, for SARIMAX model, it is also necessary that the series is stationary. However, beyond the dependent, explanatory series must also be stationary in order to build the SARIMAX model. Whereby the series is not stationary, it is necessary to try to induce it by some transformation of the original data and/or taking successive differences to remove the trend. After obtaining stationary dependent and explanatory series, the next steps are similar to the steps of the construction of the SARIMA model. The correlation and the lags between the response series and explanatory series were observed from the cross-correlation function (CCF) estimated. For verification and adequacy of the fitted model, initially, to assess the stationary of the series, we used the unit root test, Dickey-Fuller test (Dickey, 1984) and test Phillips-Perron (Perron, 1988). The residuals were analyzed by the residual correlograms and Box-Pierce and Ljung-box. To investigate the presence of autocorrelation, we analyzed the residual histogram, quantile-quantile graphic and corroborated the results with the Jarque-Bera normality test. The best fitted model was chose based on the lowest estimated variance, significance of the estimated parameters, parsimony, absolute error, mean square error and the Akaike information criterion. To make the comparison between the predictions of the models, in addition to graphical analysis we also evaluated the mean absolute error and root mean square error of prediction and U-Theil statistics (Makridakis et al, 1998).

3. Application and Results

All the evaluated time series are available in the site of the Central Bank of Brasil. The studying period was from 2003 to 2014, and the trial period was from January 2014 to July 2014. The monthly default rate series (Figure 1 - De90 ) was modeled in a first moment by a SARIMA model, and then, by a SARIMAX (Equation 1). The investigated explanatory monthly time series, which are plotted in Figure 1, were:

- GPD - the Gross domestic product in millions of Reais;
- UR - Unemployment rate in percentage is metropolitan region;
- CPI - The Consumer Price Indexes program produces monthly data on changes in the prices paid by urban consumers for a representative basket of goods and services;
• IBC-Br - Brazil Central Bank’s Economic Activity Index, is widely considered a leading indicator for gross domestic product growth, although has sometimes been criticized for not capturing tepid economic activity;

• IPI - The Industrial Production Index is an economic indicator published by the Federal Reserve Board of the Brazil that measures the real production output of manufacturing, mining, and utilities.

The default rate (Def90 in figure 1) series presents trend and a cyclical effect. However, it is necessary that the series is stationary for the construction of models of SARIMA class. After successive difference the trend was removed and the series could be considered stationary (unit root test p-value < 0.05). From the autocorrelation and partial autocorrelation functions, significant autocorrelations were indicated in lag 1 and multiples lags of S = 3. Thus, the identified model was:

\[
(1 - \alpha_1 B)(1 - \phi_1 B^3)(1 - B)Y_t = (1 + \theta_1 B^3)(1 + \beta_1 B)\epsilon_t
\]

We can observe Figure 1 that the series presented trends. To build the models of SARIMAX class all the series must be stationary, thus successive difference was applied to series to accomplish this requirement. Then, the cross-correlation functions (CCF) of differenced default series with different explanatory series were evaluated to indicate the SARIMAX models presented in Table 2.

The SARIMAX1 and SARIMAX2 models of Table 1 can be expressed as:

\[
W_t = \nabla Y_t = \gamma_1 \nabla UR_{t-1} + \gamma_2 \nabla CPI_{t-5} + \gamma_3 \nabla IPI_{t-3} + \gamma_3 \nabla GDP_{t-3} + \epsilon_t
\]

\[
W_t = \nabla Y_t = \gamma_1 \nabla UR_{t-1} + \gamma_2 \nabla CPI_{t-5} + \gamma_3 \nabla IPI_{t-3} + \gamma_3 \nabla GDP_{t-3} + \phi_1 W_{t-3} + \epsilon_t
\]

The models were estimated as briefly described in section 2. The estimative of the parameters are presented in Table 3. Observing the confidence intervals of both SARIMAX1 and SARIMAX2, there are some non-significant parameters (Table 2). We still can see that adding the autoregressive term in SARIMAX1, the estimate of the effect of covariate UR (\(\gamma_1\)) became not significant in SARIMAX2, possibly the autoregressive term is entering the information represented by the variate UR. Despite the estimated variance and the Akaike
Table 1: SARIMA and SARIMAX competitive models.

<table>
<thead>
<tr>
<th>Modelo</th>
<th>AR and MA orders</th>
<th>Explanatory Series</th>
<th>UR</th>
<th>IBC-Br</th>
<th>IPCA</th>
<th>IPI</th>
<th>GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SARIMA</td>
<td>p d q P D Q S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 1 1 1 0 1 3</td>
<td>lag = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SARIMAX1</td>
<td>− 1 − − − − −</td>
<td>lag = 5 lag = 3 lag = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SARIMAX2</td>
<td>0 1 0 1 0 0 −</td>
<td>lag = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: SARIMA and SARIMAX estimatives confidence intervals and corrected AIC.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Estimative</th>
<th>SE</th>
<th>IC (95%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>SARIMA</td>
<td>α₁</td>
<td>0.7842</td>
<td>0.1152</td>
<td>0.5584</td>
<td>1.009</td>
</tr>
<tr>
<td></td>
<td>β₁</td>
<td>−0.5534</td>
<td>0.1421</td>
<td>−0.8319</td>
<td>−0.2748</td>
</tr>
<tr>
<td></td>
<td>φ₁</td>
<td>0.8555</td>
<td>0.1113</td>
<td>0.6373</td>
<td>1.0736</td>
</tr>
<tr>
<td></td>
<td>θ₁</td>
<td>−0.6467</td>
<td>0.1804</td>
<td>−1.0003</td>
<td>−0.2931</td>
</tr>
<tr>
<td></td>
<td>σ²</td>
<td>0.0215</td>
<td></td>
<td>AICc: −120.10</td>
<td></td>
</tr>
<tr>
<td>SARIMAX1</td>
<td>γ₁</td>
<td>0.0820</td>
<td>0.0315</td>
<td>0.0202</td>
<td>0.1437</td>
</tr>
<tr>
<td></td>
<td>γ₂</td>
<td>0.0838</td>
<td>0.0275</td>
<td>0.0299</td>
<td>0.1377</td>
</tr>
<tr>
<td></td>
<td>γ₃</td>
<td>−0.0009</td>
<td>0.0024</td>
<td>−0.0056</td>
<td>0.0038</td>
</tr>
<tr>
<td></td>
<td>γ₄</td>
<td>−0.0076</td>
<td>0.0039</td>
<td>−0.0152</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>σ²</td>
<td>0.0244</td>
<td></td>
<td>AICc: −100.26</td>
<td></td>
</tr>
<tr>
<td>SARIMAX2</td>
<td>φ₁</td>
<td>0.4654</td>
<td>0.0807</td>
<td>0.3072</td>
<td>0.6236</td>
</tr>
<tr>
<td></td>
<td>γ₁</td>
<td>0.0295</td>
<td>0.0190</td>
<td>−0.0238</td>
<td>0.0828</td>
</tr>
<tr>
<td></td>
<td>γ₂</td>
<td>0.0697</td>
<td>0.0247</td>
<td>0.0201</td>
<td>0.1192</td>
</tr>
<tr>
<td></td>
<td>γ₃</td>
<td>−0.0028</td>
<td>0.0027</td>
<td>−0.0063</td>
<td>0.0007</td>
</tr>
<tr>
<td></td>
<td>γ₄</td>
<td>−0.0050</td>
<td>0.0029</td>
<td>−0.0107</td>
<td>0.0007</td>
</tr>
<tr>
<td></td>
<td>σ²</td>
<td>0.0195</td>
<td></td>
<td>AICc: −125.78</td>
<td></td>
</tr>
</tbody>
</table>

indicate that the model SARIMA is better, it is still necessary to analyze the residuals and the prediction of these models to decide which model to choose. The residuals from SARIMAX1 model, as well as, their ACF and PACF are presented in Figure 2.

We can see in Figure 2 that the model that did not consider the errors modeled by a SARIMA still has significant autocorrelation in the residuals (mainly in lag 3). Thus the presence of only explanatory variables is not enough (SARIMAX1) and so SARIMAX2 (Table 1), with errors modeled by a SARIMA and the explanatory variables, was chosen as the best model. The residuals and their respective ACF and PACF can be seen in Figure 3. Comparing the adjustments and forecasts of SARIMA and SARIMAX2 models (Table 2), the SARIMA model fitting errors are similar to SARIMAX errors, however, when comparing the predictions, SARIMAX2 model errors are smaller than SARIMA errors, as it can be seen in Figure 3.

In U-Theil statistics, note that for both models, the forecast fell short of the naive model. But comparing the two models, we have that the U-Theil is smaller for SARIMAX2 than for SARIMA model. Furthermore, the forecast error was reduced about 46 % with SARIMAX2 model in relation to SARIMA model.

Through the Figure 4, the values observed in the trial period are within the confidence intervals for both models, but the model with the explanatory variables SARIMAX accompanies more accurately the time
5. Conclusions

Taking into consideration all the aspects analyzed in the development of this work, one realizes that the final models \( \text{SARIMA}(1, 1, 1) \times (1, 0, 1) \) and \( \text{eSARMAX}(0, 1, 0) \times (1, 0, 0)(UR_{t-3}, CPI_{t-5}, IPI_{t-3}, GDP_{t-3}) \) in red and blue, respectively, provide a good fit for the series behaviors (Table 3).
Table 3: Adjustment and Prediction Errors for SARIMA and SARIMAX2.

<table>
<thead>
<tr>
<th>Errors</th>
<th>Modelo SARIMA</th>
<th>Modelo SARIMAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.1460</td>
<td>0.1746</td>
</tr>
<tr>
<td>MSE</td>
<td>0.1104</td>
<td>0.1375</td>
</tr>
<tr>
<td>U-Theil</td>
<td>1.66</td>
<td>1.23</td>
</tr>
</tbody>
</table>

presented similar results in the fitting period. However, the SARIMAX model showed more efficiency in forecasting aspect, what is the main interest.

Even though SARIMAX2 model has shown good forecast, some estimates were not significant, it shows that the model can still be improved, an alternative is to investigate other variates or also apply transformations in the series in order to leave all serious with the same scale. In addition, further work is to create economic scenarios for explanatory series model in order to perform stress testing on the model. By creating scenarios, it can be seen how the 90-plus default rate behaves on these economic scenarios, which may be positive or negative scenarios.

References
HYDMAN, R. J., & ATHANASOPOULOS, G. Forcasting: Principles and Practice, Texts.