Development of an Intervened Forecasting Model for Credit Life Cycle Based on a Previous Evaluation

Carolina González-Restrepo
Mathematical Science Department
Mathematical Engineering
EAFIT University
Email:cgonza49@eafit.edu.co

Milton Alfonso Martinez-Negrete
Mathematical Science Department
EAFIT University
Bancolombia
Email: mmarti68@eafit.edu.co

***Abstract*—We performed a deep diagnosis of a credit life cycle forecasting model used in the risk area of Bancolombia. The diagnosis was developed based on different errors between the real and forecasted different concepts of the credit life cycle and a measurement of the impact those error generates in the bank's profit-loss statement. Based on the results obtained in the diagnosis an intervened forecasting model was proposed for the most critical concept; the written offs.**

1. Problem Statement

Before raising the problematic that was addressed during the research practice, it's considered necessary to define the area where it is framed. Bancolombia counts with a tactic for an integral administration of risks, that points towards the identification, measurement, monitoring and mitigation of the inherent risks of the organization; in order to support the decision making processes and the execution of different strategies [1].

It should be noted that the concept of risk is quite large, for this reason it's consider important to precise the most relevant aspects of the types of risks present in Bancolombia. The risks are understand as the probability of incurring in losses to different reasons:

- Operational risk: caused by failures or weaknesses in the processes, people, systems, or external events.
- Market and liquidity risks: decrease in the value of the investments portfolios, funds or other resources mostly explained by the fluctuation of the interest and exchanges rates.
- Credit risk: when a third party fails to fulfill its obligations partial or totally. This translates in deterioration of the credit quality.

In this investigation the attention will be centered in the credit risk, therefore is pertinent to understand credits and their life cycle. Credit is a privilege that a bank or funding organization gives to a natural or legal person to receive money now with the commitment of paying it in the future. The life cycle has three stages: granting, tracing and recovery. The granting stage involves an evaluation of several factors that accounts for the debtor's ability to pay based on the bank policies already established. The tracing stage has the purpose to supervise the fulfilling of the latent responsibilities of the client with the bank; this process generates collection strategies that meet the needs of the clients increasing the payout odds. Finally, the recovery stage is meet when the client fulfills the totality of the obligations. This is what the bank expects but, sometimes the clients are not able to pay the credit, this is when the credit scale in days past due and could be write off.

The credit risk is controlled with a forecasting model that projects: past-due portfolio (30, 60, 90 and more than 90 days past due), written offs, balance allowance and allowance expense, the last one understood as:

$$E_t = A_t - A_{t-1} + W_t, \tag{1}$$

where

E_t: Actual allowance expense
A_t: Actual allowance
A_{t-1}: Past allowance
W_t: Actual written offs

Detailing the variables from above, an allowance (A) can be understand as a collection of money that tries to lighten the impact in case of default by the third party. On the other hand, a write off (W) is a countable operation that gives treatment to the losses of an amount before considered as an asset, in other words, when a credit is considered unrecoverable.

The model considers tendency and seasonality of the events occurred in the past to forecast the future using Markov chains, it's important to precise that the model has quarterly forecasts.

The model works in a probability space (Ω, F, P) where Ω is the set of outcome, F the set of subsets and P

is the probability of A; where A belongs to F. If i is called a state and I is the state space, P_{ij} where i, $j \in I$ is the probability of going from state i to state j [2]. Translating this definition to the specific scenario of the credit life cycle, we start with an initial state that reflect the state of the portfolio past due in the present and based on tendency and a seasonality observed in the past we determine the probability of the credit to move between different days past-due. For example, if we have a credit that has 30 days past due today what are the chances, based on the observed transition and seasonality rates, that in three months it stays there, or gets totally recovered, or moves to 60 days past due, to 90 days past due, or after 90 days past due.

Even though the model tries to represent the credit life cycle that is not always possible, this is why expert intervention is needed to achieve more accurate projections, foregoing generates a problematic for the bank. This problematic would be approach in the present investigation trying to determine the precision of the actual model and developing an intervened model that lessen the error of the actual one incorporating missing concepts such as macroeconomic variables. Because of the magnitude of the project, the model will be restricted to a specific product of the portfolio, free investment.

2. Objectives

General: Develop an intervened a forecasting model for the credit life cycle concept(s).

Specific:

- Measure the quality of the actual model based on statistical techniques.
- Study different models capable of forecasting in accurate ways.
- Study macroeconomic variables and how to include them in the intervened model.
- Study external variables and the relationship they could have in the forecasts.
- Extract all the information needed from the data bases of the organization.
- Implement an improved model based on the studied models and the available information.
- Measure the discrepancies of the reality and the forecasts obtained with the intervened model.

3. Previous Research

Having in mind that knowing in certain degree the behavior of the credit life cycle in the future can help monitoring the credit risk inherent in the organization, a projection model was developed. In virtually every decision they make, executives today consider some kind of forecast; predictions are no longer luxuries, but a necessity [3]. Before establishing the model, there was a previous study of the accuracy, computational cost, and generality of the forecasting existing methods in addition to a review of the state of art in projection models for the portfolio life cycle. Even though the information and bibliography are very limited due to the security and privacy of banks affairs and methodologies, there are certain concepts, as the past due portfolio, that has been more explore. One of the most common tool used in portfolio forecasting are the role rates models; as specified in Figure 7 of [4] the Markov chains can captured the dynamics of the past due portfolio where state transitions are model by historical transitions probabilities. Despite the low evidence of how to model the other concepts in the credit life cycle, a Markov chain forecasting model was developed hoping to obtain accuracy projections for the other concepts as these largely depends on the past due portfolio by range of default. After testing the model and noticing discrepancies and external parameter was included, this parameter is moved in a very empirical way based on expertise. The no automated, inconsistently and not established way of functioning raises alerts that suggests a prompt evaluation.

4. Justification

It is evident that no forecasting model is ideal for representing reality in a perfect way, yet is possible to correct weaknesses in the actual model to have more accurate projections.

The forecasting model is crucial for the bank since it is the sustain for the decisions taken in the management of the portfolio. To be able to forecast the dynamic of the credit life cycle allows not only the management and coverage of the credit risk but to obtain greater profitability. When the past-due portfolio is overestimated, the allowances are too and therefore the allowance expense will be greater; this means that the bank has extra allowance to cover an inferior risk, which impedes the obtaining of probability with that extra money. The contrary case is also unwanted because if the allowance in underestimated the allowance expense will be less and there would be more money able to invest and obtain more profit with, but this scenario is to risky for the bank which is in reality expose to a greater risk that is not being covered with enough allowance.

For the reasons explain above, we can conclude that the forecasting model is the basis for implementing different strategies for the uptake and laying of bank's portfolio. All the action plans on how to invest money, recover past-due portfolio and written offs and how to implement policies for ranting credit are effects of the forecasts ensuing of the actual model. The solution of the problematic has also academic purposes; the solution would be based on a series of concept learned in the university and applied in a financial field that has little bibliographic record because

of the confidentiality of the information the banks uses.

5. Preliminaries

5.1. Concepts

There are some concepts that emerge from the credit life-cycle that should be explained in detail. According to the glossary of terms [5] and the financial superintendence of Colombia.

Performing portfolio: total amount of credits which amortization and interests are up to date according to the contract established during the granting stage of the cycle.

Overdue portfolio: capital, shares or interests of the total amount of credits that have not been paid in a period longer than 30 days from the due date.

Written offs: is a countable operation which consists in providing a treatment of loss to an amount originally recorded as an asset. This measure arises from the establishment will to ascertain the uncollectability of an obligation.

Allowance: raising of money which purpose is to alleviate the impact in case of a default by a third party. Allowances are considered an expense that decreases the value of the portfolio and are proof of the reality; otherwise the financial statements would show accounts receivables or credits, that despite being true show no possibility of recoverabillity. It is important to note that regardless the customer's qualification or score the bank must provision each obligation partially or completely.

5.2. Variations

The formulas for calculating the different variations that were used in this research are the following:

Absolute variation:

$$V.abs_i = value_i - value_{i-1}$$

Monthly variation

$$V.m_i = \frac{value_i}{value_{i-1}} - 1$$

Previous year variation:

$$V.py_i = \frac{value_i}{value_{Previous\ \ December}} - 1$$

Annual variation:

$$V.annual_i = \frac{value_i}{valor_{last\ \ ye}} - 1$$

5.3. Mean absolute deviation(MAD)

Type of error which measures the mean of the absolute deviations of the forecast errors. The way of calculating it is the following [6]:

$$E_{MAD} = \frac{\sum_{i=1}^{n} |x_i - \hat{x}|}{n}$$

where:
x_i: real value
$\hat{x}$: forecasted value

5.4. Root mean square error (RMSE)

As explained in [7] it is an error measure that quantitatively evaluates the accuracy of forecasts. This calculation, compared with MAD, amplifies and strongly penalizes those errors of greater magnitude.

$$E_{RMSE} = \sqrt[2]{\frac{\sum(x_i - \hat{x})^2}{n}}$$

This error measure the quality of the estimator as it reflects its skewness and dispersion.

5.5. Mean perceptual error (MAPE)

As a measure of error independent of any scale is commonly used to evaluate and compare accuracy.

$$E_{MAPE} = \frac{\sum_{i=1}^{n} |\frac{e_i}{x_i}|}{n} * 100$$

where:

$$e_i = y_{real_i} - y_{forecast_i}$$

Thinking in perceptual terms makes MAPE easy to interpret when the typical values each variable takes are unknown. Nevertheless this measure in problematic when [8]:

- Actual values equal to zero
- Actual values almost zero
- Low volume data

5.6. Theil's U

Theil proposed two U statistic, one that measures quality of the forecast (U_2) and another one that measures accuracy of the forecasts (U_1); the last one would be the one used.

U1 is bound between 0 and 1, with values closer to 0 indicating greater forecasting accuracy [9].

$$U_1 = \frac{\sqrt[2]{\sum (x_i - \hat{x})^2}}{\sqrt{\sum x_i^2} + \sqrt{\sum \hat{x_i}^2}} n * 100$$

5.7. Mathematical weighting

Statistics is a tool for calculation and analysis that applies to any reality that has quantitative values; one of its methods is the weighting of variables when studying data in a given area. The weighting then consists in providing specific values to variables depending on the importance will be given in the analysis [10].

5.8. Forecasts

The main function of forecasting is to predict as accurately as possible the future using known data in order to support the decisions making process [11]. The forecasts are highly used in money lending activities where the high volatility of the field demands the achievement of objectives such as:

- Predicting extended time horizons
- Incorporate macroeconomic factors in the model
- Incorporate management scenarios in the model
- Modeling the return on the life cycle of the portfolio

Any forecast of this kind is highly changeable against scenarios is on marketing, sales plans, management policies and the macroeconomic environment, implicitly or explicitly; It is for this reason that should clarify how these assumptions affect the projection.

5.9. Time series

A time series is a sequence of observations measured at a given instant, arranged chronologically, and under a constant sampling period. The time series are normally used to make forecasts, due to the strong assumption, the values variables takes are the result of a tendential, seasonal and random component present in past observations [12].

5.10. Moving averages method

Used as a forecast average of n latest observations of the time series, where n is the number of periods backwards to be considered in the average. Mathematically it can be expressed as:

$$M_t^n = \frac{y_t + y_{t-1} + y_{t-2} + ... + y_{t-(n-1)}}{n}$$

As its name implies, the average is as mobile as new data is calculated, because the average is then modified.

For forecasting through moving averages simply follow the formula [13]:

$$\hat{y_t} = M_{t-1}^n$$

Is worth clarifying that the value of n will bring advantages and disadvantages. A value of $n = 12$ will completely remove the seasonal components of the series, but delayed the evolution of the variable over time, while a value of $n = 3$ will not delay the evolution of the variable but will not completely eliminate seasonality, sometimes reaching even greater effect of the mentioned component.

5.11. Exponential smoothing method

As with moving averages, the forecast is derived from past observations, the difference of this method is that the data are weighted to give greater weight to the nearest observations and a smaller one to the most distant [13]. The greatest weight is called α and is assigned to the immediately preceding observation from this assignment so on weights of $(1-\alpha)$, $(1-\alpha)^2$, $(1-\alpha)^3$ and so on until the last observation that will be consider. This is summarized as follows:

$$P_{t+1} = \alpha Y_t + (1 - \alpha) P_t$$

where
Y_t : value of the series in t
P_{t+1}: forecast in $t + 1$
P_t : forecast in t

An initial value P_0 is necessary to used this method.

5.12. Correlation Analysis

It is a statistical analysis that allows the study of sample data to determine the degree of association or correlation between two or more variables in a population [14]. The degree of correlation is expressed by the correlation coefficient, understood as a value between -1 and 1. There are several types of correlation depending on the direction and magnitude the coefficient has, see Fig. 1.

According to the practice objectives using the multiple correlation method is required; which is nothing more than an extension of the simple correlation, understood as the association between two and only two variables, only in that in the multivariate case there will be multiple simple correlations two to two regardless the influence of third variables.

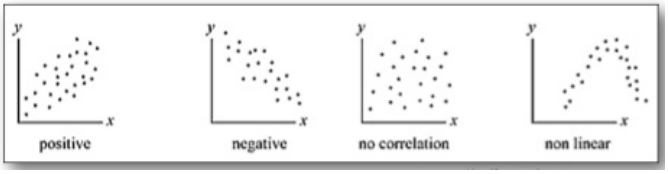

Figure 1: Types of correlation. Taken from [15]

5.13. Linear regression

Regression analysis aims to establish a function: $Y = f(x)$ statistically describing the association between the study variables under the assumption that the forecast variable depends on the values other variables take. There are simple regressions for the case of two variables and multiple regressions for more than two variables; both explore and quantify the relationship between a dependent variable and one or more independent or predictor variables.

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + ...b_n * x_n + u \quad (2)$$

Let the previous be a multiple linear regression model, where x_i with $i = 1, 2, .., n$ explanatory variables and b_i with $i = 0, 2, .., n$ coefficients indicating the increase in weight per unit increase in the explanatory variable.

By the Gauss-Markov theorem which assumes linearity, homoscedasticity, independence and normality in the variables studied, it is possible to establish that using a least squares estimation is possible to find a parameter vector **B**, the above is described in depth in [16].

5.14. Significance test

The significance test is a statistical process that uses data as evidence to approve or reject the hypothesis by comparing sample estimates with predicted values. In this research the following assumptions are used:

H_0 = b_0 = $b_1 \dots b_n$ = 0

H_a = there is at least one $b_n \neq 0$.

Using a p value, defined as the probability of obtaining an approximate result to the actually obtained assuming the null hypothesis is true, the probability of obtaining a statistically significant linear model is determined; so H_0 should be rejected.

5.15. Neural networks

Inspired by biological systems, particularly the human brain, it is demonstrated that artificial neural networks (ANNs) have a powerful pattern recognition capacity that leads to a great estimation and forecasting technique [17]. In 1986 the development of the back propagation algorithm by Rumelhart, Hinton and Williams [18] represented the birth of the most popular learning algorithm to train a set of multilayer perceptron. The multilayer perceptron is a feed forward artificial neural network model that maps input data to a set of output data, using several layers of nodes (neurons).

6. Methodology

Before proposing an improved forecasting model for credit life-cycle, a deep diagnosis of the actual model is proposed to conduct a more structured intervention plan. The diagnosis will show the performance of the current model, in order word the accuracy of forecasts made, the error between projections and reality, the impact this error have for the business and will also be a starting point to determine where to begin the intervention of the model.

Diagnosis stages:

- Forecasts and real data collection
- Execution of the current model without any intervention
- Convert quarterly data in monthly
- Calculate the four errors mentioned before
- Evaluate the distribution each concept has in the allowance expense
- Calculate contribution of each concept in the expense
- Calculate the weighted error in terms of the expense each concept generates

After determining and understanding the error of the forecasting model, a study was conducted to the variable that presented a higher percentage of error. The variable analysis is to assess methods and forecasting techniques different than Markov chains (implemented in the current model) seeking to reduce the current error.

Development of an intervened model:

- Collection of historical data of the interest variable
- Assume a time series model
  - Smoothing techniques (moving average, exponential)
  - ARIMA (auto regressive integrated moving average) type model
- Assume a linear model
  - Collect possible explanatory variables (bank and macro-economical)
  - Correlation analysis
  - Linear regression of the variables
- Construction of a neural network
  - Training of the neural network
- Re-calculate the errors of the forecasts obtained

7. Results

This section describes some of the results obtained by different methods and techniques implemented for addressing the problem, explained in the previous section. In addition to the difficulties found as the problem was broken down.

Table 1 contains the results of the measurement of the proposed errors, the measurement is performed for the projections obtained by the model intervened by expert criteria and the forecasting model without intervening; both projections are compared against actual observations. The exercise is performed from the second half of 2014; it should be noted that it was not possible to reproduce the exercise to previous months since the model projections is recent.

Errors	**REAL/FORECAST 2014-2015**			
	RMSE	MAD	MAPE (%)	U1
Balance	128.269	85.407	2.8	0.084
Performing Loans	127.693	80.705	3.0	0.094
Disbursement	55.490	37.775	15.5	0.541
$>$ 30 past due	9.907	7.583	5.4	0.156
31-60 past due	5.037	4.582	11.1	0.249
$>$ 60 past due	6.816	5.486	5.8	0.160
A. Expense	3.509	2.692	27.0	0.686
Written offs	3.030	2.584	37.9	0.696

Table 1: Errors of the forecasting model

Errors	**REAL/CLEAN 2014-2015**			
	RMSE	MAD	MAPE (%)	U1
Balance	128.269	85.407	2.8	0.084
Performing Loans	127.693	80.705	3.0	0.094
Disbursement	55.490	37.775	15.5	0.541
$>$ 30 past due	8.319	5.707	4.1	0.132
31-60 past due	5.085	4.500	10.7	0.250
$>$ 60 past due	5.556	4.829	5.2	0.129
A. Expense	3.902	3.294	35.9	0.719
Written offs	3.979	3.594	52.2	0.856

Table 2: Errors of the forecasting clean model

It is important to use common errors, and not own measurements, since by using own errors objectivity is lost in the analysis. For this reason the quality of the model is shown in terms of the root means square error (RMSE), the absolute deviation from the mean (DAM), Theil's U ($U1$) and the mean percentage error (Errors in Table 1 and 2).

From the previous tables is possible to observe how although balance and disbursements errors do not vary (because they are inputs to the model) the performing loans, calculated from the above variables, varies because of the high influence of more than 30 days past due portfolio. Besides, the higher the existing error in forecasting non performing loans the higher the error in the allowance expense, which is expected by the nature of provision expense (see Equation 1).

Intervening the model with expert knowledge benefits the forecast provision expense and penalties but hurts the prognosis of non performing loans in different ranges of default, however this does not imply that intervention is wrong. On the contrary the advantages in terms of error rate (compare Table 1 and 2) are greater when intervening the model, thereby increasing the accuracy of forecasts by up to 10 %; for concepts whose forecasts detract, have a decrease in accuracy of only 1 %, which is considered insignificant. Despite the improvements, allowance expense and written offs still have the highest percentage of error in the model.

In order to assess the impact of these inaccuracies of the forecasting model, it was decided to calculate a weighted error. The calculus would be focus on the allowance expense those errors generate to the bank. This approach to the problem is an approximation due to the fact that the only precise information known is the total expense but not the expense each concept generates. For this approach, an historical distribution of allowance expense is calculated to analyze in monetary terms how much each concept contributes to the total amount. The results of the distributions are shown in Table 3:

%	**Disburs.**	**P. loans**	**31-60**	**$>$60**	**Cancel**	**W. offs**
2014	65.6	-62.5	24.7	122.6	-50.3	-104.3
2015	70.4	-48.4	20.8	104.5	-47.3	-81.8
total	67.8	-56.2	22.9	114.4	-49	-94.1
Avg.	67.1	-52.4	22.3	109.4	-49.8	-88.9

Table 3: Historical distribution of the allowance expense

There is a spending released each time a credit is canceled for this reason should take into account cancellations C_t:

$$C_t = B_{t-1} + D_t - B_t$$

where:
B_{t-1}: real balance $t-1$
B_t: forecasted balance t
D_t: forecasted disbursements in t

From the results in Table 3 an approximate expending per concept is obtained. Subsequently a percentage of participation in the total expenditure is calculated. When multiplying the percentage error found in Table 1 with the contribution in the allowance expense we obtained a weighted error in terms of costs, which is shown in Table 4:

The results in Table 4 show how errors can be mitigated or increased in terms of the impact they generate in the allowance expense. It is noted that although there are large absolute percentage error the impact they have in the allowance expense and therefore the bank's profit

Concept	MAPE	Expense	contribution	W. Error
P. Loans	3.0%	-16.505	54%	1.6%
Cancel	18.5%	-15.717	51%	9.4%
Disburs.	15.5%	17.698	58%	9.0%
31-60	11.1%	7.571	25%	2.8%
>60	5.8%	37.506	123 %	7.1%
Written.off	37.9%	32.641	107%	40.6%
Total		30.553		

Table 4: Weighted Error

and loss statement is not as significant as it seems. The absolute percentage error can be reduced even by half when weighted by the expense, this is the case of cancellations; or conversely increased further as written offs which directly affect the cost of allowance (see Equation 1)

While analyzing the errors measurements described in Table 1 and the weighted error, it can be seen that the concept has the greater error while being forecasted are the written offs; and it is for this reason that the following results section shows a series of processes performed in order to improve the forecasts of these concept. Following the explained methodology, a time series of written offs is obtained and explored as shown below:

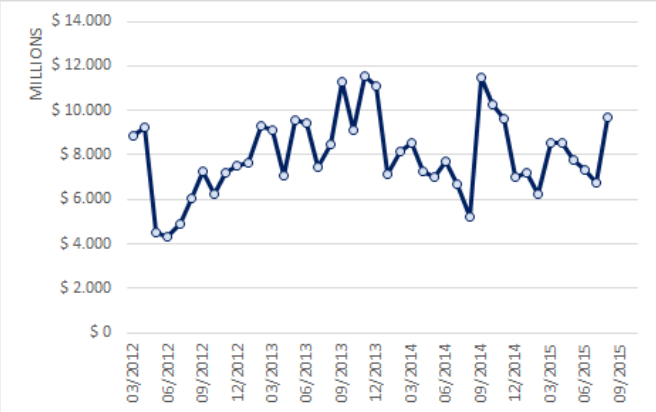

Figure 2: Historical series of written offs

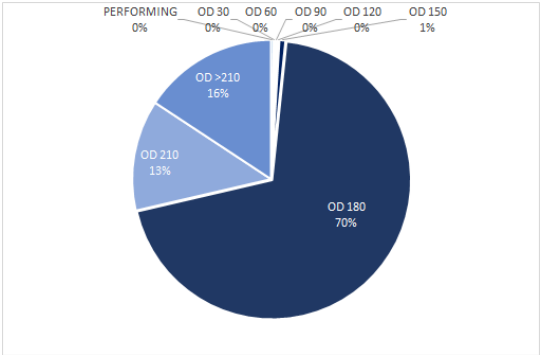

Figure 3: Distribution of written offs depending on default

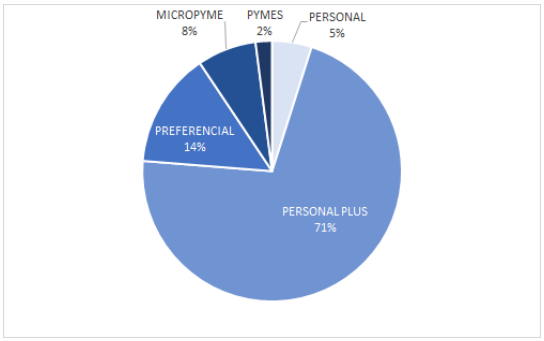

Figure 4: Distribution of written offs depending on the client

Under the assumption that the written offs can be forecasted under a time series model, smoothing techniques as moving averages with different n are used; indicating that future values are forecast taking a different amounts of past observations, the results were:

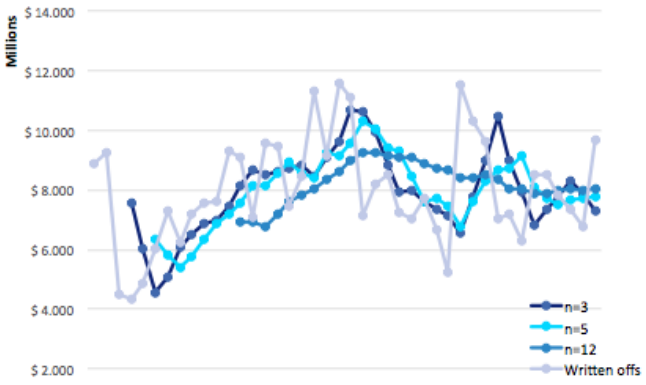

Figure 5: Forecasted written off using moving average

From the figure above it is possible to determine that the moving average that best fits the series of written offs is the moving average of order 3, which means that the value of today is an average of the last 3 periods. The following chart shows in detail the actual and projected series.

Although the forecast for moving averages tries to adopt the behavior of the actual series, does not completely succeed; it is significantly inaccurate sometimes, as illustrated by the red markers. These points show the difference between projected and actual data for the same instant of time. Trying to solve this type of impressions, and looking for a more accurate forecast an exponential smoothing method is proposed, which uses a α to give greater weight to the most recent observations. The parameter α is optimized obtaining a value of $\alpha = 1$ with which the least mean square error is obtained. Graphically:

As with last method a coarse adjustment is obtained but lagged causing significantly high errors as illustrated

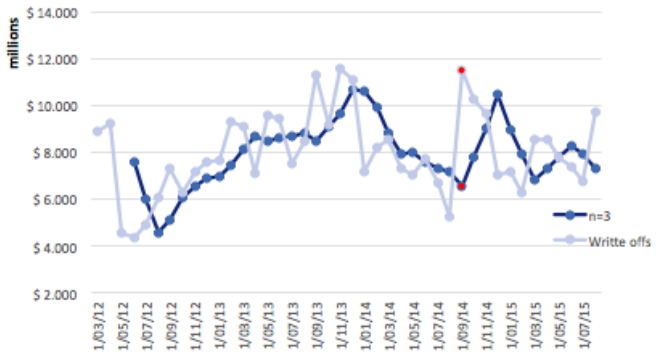

Figure 6: Written offs moving average n=3

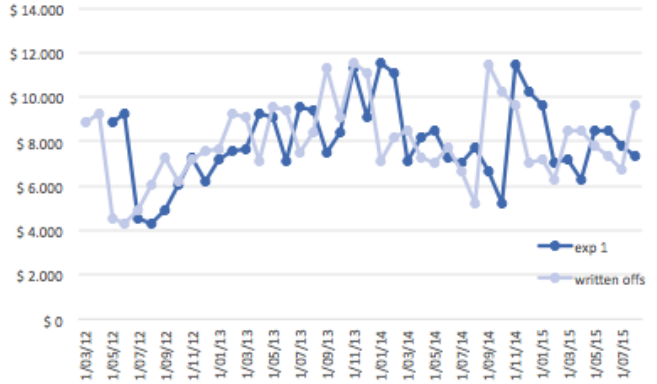

Figure 7: Exponential smoothness alpha=1

in the figure. Before continue implementing other time series forecasting techniques a test is performed to validate that the current series can be approximated to an ARIMA (auto-regressive integrated moving average model) the test is performed in R using the function (aunto.arima) in the statistical package 'forecast', specializing in projections. The exercise result was:

$$result = ARIMA(0, 0, 0) \tag{3}$$

This implies that the series does not have an auto regressive, or integrated, or moving average; therefore it is a random walk. This indicates, that the historical series of write offs can not be fitted using a time series model. The walks are a random process where the position of the variable in some instant depends only on its position at some previous moment and some random variable. For this reason it must be said that write offs need different variables than itself in order to be predicted.

Following this approach a linear regression model is proposed address the written offs estimation dilemma. For this, a collection of internal bank and macroeconomic variables such as DTF, CPI, GDP, among other was performed. In total there were 50 base variables that generated 500 variables which includes their respective annual, previous year and absolute variations and the remnants of themselves.

The biggest problem that may be found when performing linear regression is the collinearity; if when estimating the coefficients described in **(2)**, some independent variable is a linear combination of other the model has no solution. Collinearity between variables can be expressed in terms of the correlation coefficients, when this coefficient takes values near or equal to 1, there is perfect correlation and collinearity between variables.

Once raised the problem, a correlation test is considered necessary to determine which variables should be eliminated because they are already being represented by others within the model.

	TRM	TRM(t-1)	TRM(t-2)	TRM(t-3)
TRM	1	*0,9832*	*0,95621*	0,93439
TRM(t-1)	*0,9832*	1	0,9832	0,95621
TRM(t-2)	*0,95621*	0,9832	1	0,9832
TRM(t-3)	0,93439	0,95621	0,9832	1

Figure 8: Correlation matrix

Figure 8 shows the correlation matrix for four variables, in this case the TRM is removed lagged one and two periods as they have a correlation greater than 0.95 (cutoff value), implying that they are not necessary and that the information provided is already contained. The above process is performed for the 500 variables and a total of 196 variables were removed because of the high correlation. The remaining variables will be considered in the linear regression.

With the free-correlation database we proceed with the linear regression, which is performed in SAS (application for data mining) [19]. The written offs are included as the dependent variable and the other variables as explanatory variable. Then forward selection is performed, evaluating the quality of a variable to explain the dependent prior to adding it to the model. The following result were obtained:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	3.727097E19	2.866997E18	2.6E11	<.0001

Figure 9: Linear regression results in SAS

As observed in the last column, the p-value is less than 0.05 indicating a rejection in the null hypothesis, and therefore indicating the linear model is significant. Additional, the program returns the variables that make part of the resulting

model and their significance. Table 5 shows the obtained results:

Forward selection method		
Step	**Variable**	**P-value**
1	V. abs OD>180	< 0.0001
2	V. py TRM_2	< 0.0001
3	V. abs DTF_3	< 0.0001
4	$CV60_2$	0.0342
5	V. abs OD> 180_2	0.0500
6	V. annual expense	0.0106
7	V. abs TRM	0.0112
8	V. abs $REPO_1$	0.0019
9	V. abs TES_3	0.0094
10	V. abs employment	0.0252
11	V. py petroleum	0.0016
12	V. annual employment	0.0043
13	V. annual GDP	0.0091

Table 5: Results of SAS model

In the previous table the variables that can be found are either contemporary, this means form the current period or lagged variables (small subscript) one, two and even three periods. Plotting the forecasted written offs and the actual series we obtained:

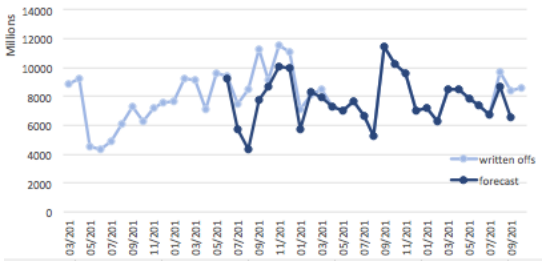

Figure 10: Linear regression in SAS

Figure 10 shows a very appropriate fit between the actual and projected series, but has some limitations. It can be observed that the estimation does not start from the starting point of the written offs series, that is due to the lack of information of these observations from one or more of the significant variables for the model; it is also clear that the projected series not find forecasts because several of the significant variables are contemporary. This means that in our case, to forecast punishment in the next month we should know the absolute variation of nonperforming loans greater than 180 days past due for the same month; this indicates that the minimum value of the lags of the variables in the model indicate the maximum number of periods that may be forecast. For this case as there are contemporary variables, no projections will be generated.

Alternatively to the previous model a free contemporary variables model was proposed, that is, the explanatory variables that enter the linear regression should be lagged at least one period. The regression results where:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	14	4.045517E19	2.889655E18	2.37E11	<.0001

Figure 11: Linear regression results in SAS

Forward selection method		
Step	**Variable**	**P-value**
1	V. py $colcap_2$	0.0015
2	V. py $colcap_3$	0.0085
3	V. abs TES_2	< 0.0001
4	V. annual TES_1	< 0.0001
5	v. annual $employ_2$	< 0.0001
6	V. abs $unemploy_3$	< 0.0001
7	V. abs $LICC_2$	< 0.0001
8	V. abs $LICC_3$	< 0.0001
9	V. py GDP_2	< 0.0001
10	V. abs OD_1	< 0.0001
11	OD 30_3	< 0.0001
12	V. abs OD 30_1	< 0.0001
13	V. abs OD 120_2	< 0.0001
14	$writtenoffs_8$	< 0.0001

Table 6: Result SAS model 2

The respective parameters for each of the variables are shown in the table below, plus the independent term, B_0.

Forward selection method	
Variable	**Parameter**
Intercept	$1.5E10$
V. py $colcap_2$	$-2.9E7$
V. py $colcap_3$	$7.2E6$
V. abs TES_2	$2.04E10$
V. annual TES_1	$3.6E9$
V. annual $employ_2$	$-5.6E10$
V. abs $unemploy_3$	$-1.2E11$
V. abs $LICC_2$	-9079
V. abs $LICC_3$	85.6
V. py $GDPn_2$	$-1.5E10$
V. abs OD_1	0.01
OD 30_3	-0.02
V. abs OD 30_1	-0.09
V. abs OD 120_2	-0.24
$writtenoffs_8$	-0.69

Table 7: Parameter estimation

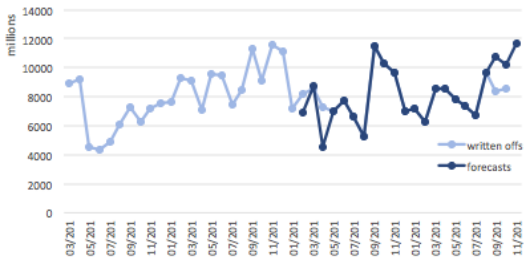

Figure 12: Linear regression SAS model 2

The previous graphical results (Figure 12) show a significant forecast accuracy, and the possibility of

forecasting the next period written offs. The following table contains the error measurements for the concept being estimated for both proporsed linear models.

Errors	Real/forecast intervened model			
	RMSE	MAD	MAPE (%)	U1
Written offs (contemponary)	786	634	5.46	0.23
Written offs (lagged model)	904	389	4.87	0.21

Table 8: Errors of the forecasting models proposed

Even though the results obtained with linear regression where significant, we decided to train an artificial neural network with the resultant significant variables obtained from the past linear model. The neural network was trained using 37 patterns and 5000 epochs but it did not ended in the expected time nor learned as expected. This might be explained because of the volatility of the series or output and the little amount of patterns or input information had.

8. Conclusions

The term with greater systematic error are the written offs, which generate a high cost to the bank as seen in the weighted error, which provides valuable information on the magnitude of error in terms of impact to the business.

The written offs series of can not be modeled by a time series model since it lacks auto regressive integrated moving average. This makes written offs a random walk.
The models obtained with SAS are statistically significant as the p values evidenced. The p-values obtained reject the null hypothesis in the ANOVA test, which support the fact of modeling and forecasting written offs with a linear model.

The results obtained by the unrestricted model in terms of lag are accurate but require forecasting other concepts needed as input to forecasts written offs. The results obtained by the lagged variables model are an alternative because they retain the accuracy and do not require contemporary variables facilitating the calculation of the inputs and therefore the written offs.

The errors decreased significantly with the intervened forecasting model, evidenciating the fact written offs could be better estimated using methods different from Markov chains.

The linear model proposed is a feasible alternative due to its simple replicability for other bank products.

It is recommended to calculate a model with greater lags to achieve more than one outcome. For the neural network it is recommended to generate more pattern. the generation could be by using a bootstrapping method or by running another linear regression model with a database containing variables with more observations.

References

[1] Valores.Bancolombia, "Gestion de riesgo." http://www.valoresbancolombia.com/cs/SatellitePagecid2,. accessed 04-02-2016.

[2] J. Norris, *Markov Chains*, ch. 1. Cambridge University Press, 1997.

[3] J. ChambersSatinder, K. MullickDonald, and D. Smith, "How to choose the right forecasting technique," *Harvard Bussiness Review*, pp. 1–4, 1971.

[4] J. Breeden, "Portfolio forecasting tools: What you need to know.," *The RMA journal*, pp. 6–10, 2003.

[5] B. C. de Bolivia, "Términos frecuentes."

[6] J. Garzon, "Desviación media absoluta." http://es.scribd.com/doc/50145519/Desviacion-media-absoluta.

[7] G. de Operación, "Cálculo de la raíz del error cuadrático medio."

[8] G. de Operación, "Error porcentual absoluto medio."

[9] A. García Santillán, "Coeficiente u de theil." http://www.forecastingprinciples.com/data/definitions/theil's%20u.html. accessed 04-03-2016.

[10] Investopedia, "Weighted average."

[11] P. Reyes Aguilar, "Métodos de pronósticos," *Administración de operaciones*, Agosto 2009.

[12] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*. Springer Science & Business Media, 2013.

[13] L. Allen, "Métodos de pronósticos," *Técnicas de Suavización*, 1998.

[14] U. de Cordoba, "Correlación multiple y correlación canónica," Departamento de Producción Animal.

[15] M. O. Suárez Ibujes, "Coeficiente de correlación de karl pearson."

[16] J. M. Rojo, "Resgresión lineal múltiple," *Laboratorio de Estadística*, 2007.

[17] S. Haykin and N. Network, "A comprehensive foundation," *Neural Networks*, vol. 2, no. 2004, 2004.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

[19] S. E. Guide, "Sas products & solutions." http://support.sas.com/software/products/guide/. accessed 06-03-2016.

[20] S. F. de Colombia, "Reglas relativas a la gestión del riesgo crediticio," *Régimen general de Evaluación, Calificación y Provisionamiento de cartera de Crédito. Circular 11 de 2002.*, 2002.

[21] Consejo.Superior, "Reglamento de propiedad intelectual de la universidad eafit." http://www.eafit.edu.co/institucional/\\reglamentos/Documents/Reglamento_Propiedad_Intelectual.pdf. accessed 08-02-2016.

[22] D. T.W., "Sas versus r part two." http://thomaswdinsmore.com/2014/12/15/sas-versus-r-part-two/. accessed 05-02-2016.