LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER

ACTUARIAL SCIENCE

MASTER´S FINAL WORK

DISSERTATION

DEFAULT RISK: ANALYSIS OF A CREDIT RISK MODEL

RICARDO MIGUEL DO BRITO PENHA

OCTOBER - 2016

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER

ACTUARIAL SCIENCE

MASTER´S FINAL WORK

DISSERTATION

DEFAULT RISK: ANALYSIS OF A CREDIT RISK MODEL

RICARDO MIGUEL DO BRITO PENHA

SUPERVISION:

MARIA DE LOURDES CENTENO

OCTOBER - 2016

ACKNOWLEDGEMENTS

Quero primeiro agradecer à instituição bancária que disponibilizou os dados utilizados neste trabalho. Às pessoas que me acompanharam mais de perto, um muito obrigado por toda a preocupação que demonstraram e ajuda nas dúvidas colocadas.

À minha orientadora, Professora Doutora Maria de Lourdes Centeno, uma palavra de gratidão por se ter disponibilizado sempre em ajudar. Pelos valiosos conselhos e críticas construtivas, o trabalho pôde tomar um rumo melhor.

Tenho também muito a agradecer à minha família: pais, irmãs e avós. Aos meus pais em especial, por terem sempre criado condições para que este trabalho, e todo o restante percurso académico, tivesse sucesso.

Aos meus amigos, em particular ao Aires, João, José, Lourenço e Tomás. Muito obrigado por todo o apoio.

Por último um agradecimento muito especial à Luísa, por ter sempre acompanhado de perto os desenvolvimentos do trabalho e por todo o apoio, compreensão e conselhos.

ABSTRACT

A considerable part of the banking business includes the lending of money. Inherently, a bank incurs the risk of not receiving back the money lent. In this work, default risk is studied through the distribution function of the aggregate losses.

After making the link between the characteristics of a portfolio of loans and of a life insurance policies portfolio, Risk Theory results are applied to the portfolio of loans under study. CreditRisk$^+$, usually classified as the actuarial model, is a credit risk model which uses this link. As an input to this model, both the individual probabilities of default for each obligor and the exposure at risk are needed.

The first part of this work focus on the estimation of the probability of default through a logit model, taking into account some financial indicators of the company. Then, in the context of a collective risk model, Panjer's recursive algorithm is applied.

Following the methodology of CreditRisk$^+$, the portfolio is then divided into sectors and default volatility is introduced in each sector, reaching a different aggregate loss distribution function.

At the end, we find that similar results are obtained with less time consuming approximation methods, particularly with NP approximation.

Finally, the average interest rate that the bank should have charged to the loans in the portfolio is found as well as the amount of money that should have been reserved to account for losses.

Keywords: Loan portfolio, Probability of default, Logit, Collective risk model, Aggregate loss, Panjer, CreditRisk+

RESUMO

Uma parte considerável do negócio bancário inclui naturalmente o empréstimo de dinheiro. Inerentemente, o risco de não receber de volta o montante emprestado é assumido pela instituição bancária. Neste trabalho, o risco de incumprimento é estudado através da função de distribuição das perdas agregadas.

Depois de feita a ponte entre as características de uma carteira de empréstimos de um banco e as características de uma carteira de apólices de seguros vida, os resultados da Teoria de Risco podem ser aplicados à carteira em estudo. O CreditRisk^{+}, geralmente classificado como o modelo actuarial, é um modelo de risco de crédito que tem por base esta ponte. Para aplicação deste modelo, é necessária informação relativa às probabilidades de incumprimento de cada devedor e a exposição ao risco, que no nosso caso é igual ao montante em dívida.

Na primeira parte deste trabalho é estimada a probabilidade de incumprimento através de um modelo logit, tendo em conta alguns indicadores financeiros da empresa. Seguidamente, no contexto de um modelo de risco coletivo, é aplicado o método iterativo de Panjer.

Seguindo a metodologia proposta pelo modelo CreditRisk^{+}, a carteira é seguidamente dividida em setores e, em cada setor, é introduzida volatilidade à probabilidade de incumprimento.

No final, conclui-se que conseguem ser obtidos resultados semelhantes utilizando métodos de aproximação menos dispendiosos, nomeadamente com a aproximação NP.

Finalmente, a taxa de juro média que o banco deveria aplicar aos empréstimos em carteira é calculada, assim como a reserva que deveria ter sido constituída.

Palavras-chave: Empréstimos bancários, Probabilidade de incumprimento, Logit, Modelo de risco coletivo, Perda agregada, Panjer, CreditRisk+

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

A considerable part of the business of bank institutions is to lend money. Implicitly, the risk of not receiving back the amount of money lent is incurred. This risk is called default risk and its quantification assumes a fundamental role in the risk management of a bank.

This work aims to study and quantify the default risk. To apply the methodologies presented throughout the essay, a portfolio of loans owned by a national bank institution was provided. However, it should be remarked that the ultimate interest of this work is not to study this particular portfolio, but to show the application of Risk Theory models, usually applied in the insurance context, to the banking framework.

In a first part, the probability of default is briefly studied. We are interested in finding what financial indicators of a company can explain default and how. For this, we will align our approach with what is commonly done in the literature, as far as possible, given some data limitations.

Then, these estimated probabilities will be used as an input to the credit risk model under study in this work, CreditRisk$^+$, also known as the actuarial risk model among the most used ones. We are going to show why this risk model is considered to be the actuarial one. Particularly, we are going to follow CreditRisk$^+$ ideas, but formalizing every step in the Risk Theory framework. This is the second part of this work, which consists of Sections 3 and 4.

In Section 5, we test whether similar results can be obtained with approximation methods which depend only on some moments of the aggregate loss distribution.

At the last section, the average interest rate that the bank should have charged to obligors such that the portfolio is self sustainable is found for a certain probability level as well as the initial reserve that should have been accounted.

2. PROBABILITY OF DEFAULT

When a bank lends money to an obligor there is no guarantee that the obligor will pay back the amount in debt. Each obligor has intrinsically associated a probability of default. It is common sense that this probability of default is driven by some factors. For example, it is more likely to observe a start-up company defaulting than a multinational one as it is more likely that default comes from a company with negative profit in the previous year rather than from one with positive profit. The first part of this work aims to decode what factors may influence a company to default, estimating it through a logit model.

2.1. Generalized Linear Models

Linear regression models aim to quantify how a set of independent variables affect a response variable. In its simplest form, the response variable y is estimated as a linear combination of the explanatory variables $x_1, x_2, ..., x_n$ such that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

where β_i are parameters to be estimated and $\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$ is an error term. Therefore, it is in fact the expected value of the response variable y that is being estimated as a linear combination of the explanatory variables, i.e.

$$\mathrm{E}\left(y|x\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

As a result of this, it is implicit under linear regression models estimation that

$$y|x \sim \mathrm{N}\left(\beta x, \sigma^2\right)$$

where $\beta x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$. In some practical applications, this might not be a proper assumption. This is particularly obvious when modelling a binary response variable. When this is the case, the problem enters in the scope of generalized linear models.

Given a response random variable Y, a generalized linear model is characterized by

(i) A distribution function

The probability density function of the response variable is assumed to be a member of the exponential family, i.e. the set of distributions whose density function can be written as

$$f_Y\left(y,\theta,\varphi\right)=\exp\left[\frac{y\theta-b\left(\theta\right)}{a\left(\varphi\right)}+c\left(y,\varphi\right)\right]$$

where θ, the natural parameter, is a function of the expected value of Y and φ is the dispersion parameter. To this family belong for instance distributions such as Normal, Poisson or Binomial.

(ii) A linear predictor

The linear predictor η is defined as the linear combination of the explanatory variables $x_1, x_2, ..., x_n$, i.e.

$$\eta=\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n .$$

(iii) A link function

The link function g is a monotonic differentiable function which establishes the relationship between the expected value μ of Y and the linear predictor, i.e. $g\left(\mu\right)=\eta$. It is common practice to consider as link function the canonical link function, which is defined as the function h such that $\theta=h\left(\mu\right)$.

In the context of this work, we want to estimate the probability of default of the obligors in our portfolio through a generalized linear model. Being D_i the random variable that models the default of obligor i and p_i its probability of default, we have that $D_i \sim \mathrm{Bi}\left(1,p_i\right)$. It is worth noting that the expected value of D_i is p_i, and therefore, under an appropriate generalized linear model and after estimating the linear predictor η_i for obligor i, our estimate for p_i is the output by the inverse link function of the estimated linear predictor, i.e. $p_i=g^{-1}\left(\eta_i\right)$.

When modelling a Binomial response random variable, the link function must be chosen in such a way that its inverse can only take values between 0 and 1. The canonical link function is

$$g_c\left(\mu\right)=\ln\left(\frac{\mu}{1-\mu}\right)$$

Using this link function, this model is known as the logit model. Other common choices are the probit (g_p) and the complementary log-log (g_l) link functions, which are as follow

$$g_p(\mu) = \Phi^{-1}(\mu)$$

$$g_l(\mu) = \log\left(-\log(1-\mu)\right)$$

where Φ^{-1} is the inverse of the distribution function of a standard normal random variable.

2.2. Literature Review

The literature on the topic of what financial indicators might drive future default is extensive and remote. Edward Altman is amongst the first to investigate this topic. Back in 1968, Altman (1968) studied how a set of financial indicators can predict corporate bankruptcy. He started with 22 ratios under the categories of liquidity, profitability, leverage, solvency and activity, concluding by the significance of 5 of them.

It is common practice to consider financial ratios from different categories. Intuitively, this allows for different aspects of a firm to be captured. Profitability, efficiency and liquidity are amongst ratio categories that are more frequently used to predict default.

Besides firm-specific factors, macroeconomic risk factors are also frequently taken into account to capture systemic risk, as in Carling *et al*. (2007), Bonfim (2009) and Hamerle (2004). For instance, Bonfim (2009) estimates the probability of default for a sample of companies through a probit model using only firm-specific information as explanatory variables in a first approach. Then, by incorporating some macroeconomic variables, an improvement in the model is registered, which suggests some important and reasonable links between credit risk and macroeconomic dynamics.

Along with macroeconomic variables, factor or dummy variables can also be considered. Still related with systemic risk effect, Volk (2014) concludes that taking time dummies as explanatory variables performs slightly better than models with macroeconomic variables. That is, instead of introducing a set of macroeconomic variables, Volk (2014) concludes that the inclusion of a dummy variable accounting for the reference year of the financial information is sufficient. Other factor variables that usually have a good explanatory power include firm’s size and sector of activity, as in Volk (2014).

Transversally to all referred papers, explanatory variables are considered with some lag. Particularly, before estimating a model, Bonfim (2009) starts by analysing the lag effect that must be considered in each variable through its correlation with credit overdue some years later. Intuitively, this is a natural effect to account for, since the default of a company in a given year is the realization of its past activity and performance.

The choice of the framework under which the estimation is going to be performed is also a point to highlight. Huang and Fang (2011) analyse six major credit risk models, including the logit and probit model. According to their results, these two are among the ones with better accuracy ratio, although there is not a significant difference between them. The models in Bonfim (2009) and in Volk (2014) are probit models. However, when comparing logit and probit, Gurny and Gurny (2013) concludes that logit model is more appropriate.

It should be remarked that all these conclusions, which are presented in the papers considered, are naturally data biased. As it is observed in Altman (1968), the possibility of bias is inherent in any empirical study, since the effectiveness of a set of variables in the sample under study does not imply its effectiveness in any other sample. Nevertheless, we are going to ground our estimation procedure in these conclusions as far as possible, depending obviously on its applicability to our particular database and taking into account the limitations in terms of data provided.

2.3. The Database

In this section, our database is introduced and the proper choice of the linear predictor and its estimation is discussed.

The portfolio under study in this work consists of the portfolio owned by a Portuguese bank institution of loans granted to enterprises. It was provided information about the obligors, namely several ratios based on the companies' balance sheet and profit and loss account throughout some years, as well as the monthly default record and exposure since the beginning of 2014. For confidentiality reasons, the content of this information will not be shown.

Because the format of the information provided was not in a structure that fit our needs, particularly because the information was spread in several files, a new database was constructed to compile the relevant information of each file. In this process, some information was purposely lost, both in terms of variables and of obligors.

The file that contains the economic information has roughly 1.2 million lines of information related to 74 667 different obligors. For each obligor, there might be information in more than one line of the file to account for different reference years and, if the case, different loan contracts. After an insight analysis of this database, we could conclude that there is a considerable number of lines with incomplete information. For estimation purposes, complete data is needed. If only the lines with complete information in all variables were considered, too much information would be lost. To overcome this, we based our analysis in a study conducted by an independent entity on the rating model of the bank. In this study, univariate analysis to both quantitative and qualitative variables led to a conclusion of what variables might be significant in a regression, based on the correlation between them. There are 5 quantitative and 5 qualitative variables to draw attention to and therefore the database is filtered by the obligors which have complete information on all these ten variables. Table I and Table II identify these variables, while Annex A and Annex B shows more detail.

Table I
Quantitative variables description

Variable	Description	Category
ROCEL	Operating Income / Net Economic Capital	Profitability
TVV	(Sales and Services(year n) - Sales and Services(year n-1)) / Sales and Services(year n-1)	Activity
FMNFV	Working Capital / Sales and Services	Operational
AF	Equity / Assets	Financial Structure
JVPS	Interest Expenses / Sales and Services	Banking financing

Category column is based on the classification attributed by the bank

Table II
Qualitative variables description

Variable	Description
info3	Did the exposure of the loan increased in the last 6 months?
info5	Is the company internally identified as a critical case?
info16	Has the company delayed the payment to the bank by more than 30 days?
info18	Has the company delayed any other payment by more that 30 days?
info31	Is the company registering a decrease in the average net income?

Factor variables that take value “Sim” or “Nao” if the answer is positive or negative, respectively

Taking all this considerations into account, the database under study comes down to 11 140 obligors for the year of 2014. This sub-portfolio is going to be considered as if it was the whole portfolio of loans and hence, no conclusion is to be taken for the whole portfolio of loans of the bank. Furthermore, from the 11 140 obligors of our portfolio, 391 were in default. In 2015, the bank continues to be exposed to 10 215 of them. No information was given regarding the reason for the exits.

2.3.1. Estimation Results

In this section, the linear predictor that explains the default variable is discussed and estimated using the software R. Taking into account the limitations of the database that was provided, the ideas and conclusions presented in Section 2.2 are applied as far as possible and considering our sub-portfolio as the entire one.

The chosen link function is the canonical one and so, the probability of default is going to be estimated through a logit model. The idea is, through the estimation of a model for the default in 2014, to apply the model to predict the default in 2015.

In the estimated models presented hereafter, the response variable is naturally the default during 2014. Given the monthly record provided, it is going to be considered that the loan is in default if, in any month of 2014, a delay of 90 days or more in some payment is registered, which is consistent with the definition of default by the bank.

In terms of explanatory variables, default is going to be predicted with information from previous years. Therefore, both quantitative and qualitative variables for reference year 2013 are considered. Furthermore, and to allow for the lag effect of the economic indicators, quantitative variable for reference year 2012 are also considered. The idea is to incorporate all these variables at first into the estimated model and then to check its individual statistically significance.

Along with quantitative and qualitative information, firm's characteristics such as its size and sector of activity are considered too. The variable firm's size, called *Dimensao* in our database, is a factor variable which can take the values "GRE", "PME" and "PE" which stand for large, medium and small firm, respectively. The sector of activity variable, *Setor* in R, takes the values "comercio", "industria" and "servicos", which stand for commerce, manufacturing and services sectors, respectively. Annex C show more detail on these variables.

Before going over the estimation of the model, the macroeconomic context of the years we are considering should be mentioned. The European debt crisis started in 2009, but its effects are still being felt, particularly in Portugal. The year of 2013 was maybe the hardest year for companies in general. It was actually the last year ever since to register a negative real Gross Domestic Product growth rate. Because of this, given that our data is under pressure conditions, all conclusions taken might be limited.

The R output of the model incorporating all these variables, which is presented in Annex D, allows for some interesting conclusions. First, quantitative information for reference year 2012 seems not to be statistically significant as well as firm's size, contradicting the lag effect of more than 1 year in the financial indicators on this particular database. In contrast, all qualitative variables seem to explain default. When it comes to the variable sector of activity, while the estimated parameter $\hat{\beta}_{industria}$ is statistically significant, the estimated parameter $\hat{\beta}_{servicos}$ is not.

In order to reach the best model, the procedure is to eliminate the variable with highest p-value, step by step, ending up only with variables whose estimated parameter is statistically different from zero. In the particular case of the variable *Setor*, instead of disregarding this variable because of the non significance of $\hat{\beta}_{servicos}$, it was substituted by the dummy variable *SetorIndustria*. This variable takes the value "Sim" if the sector of activity is the industry one and "Nao" otherwise. The substitution of *Setor* by *SetorIndustria* permits to conclude about the following hypothesis test

$$H_0 : \hat{\beta}_{servicos} = \hat{\beta}_{comercio}$$

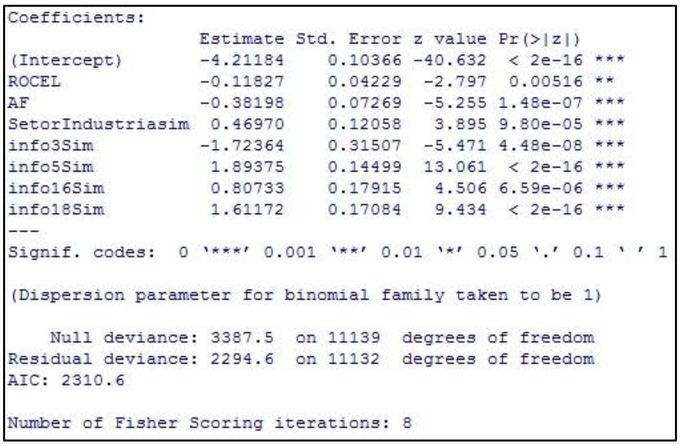
Given that the reduction in the residual deviance from the model that has *SetorIndustria* as explanatory variable to the model that has *Setor* is of 0.1, H_0 is not rejected as the reduction in deviance is less than 3.841, the 95th percentile of $\chi^2(1)$.

Following this procedure the best model, in terms of variables significance, is reached. However, the output of the estimation carried out by the software R for this model returns a warning message that there are obligors where the fitted probability equals 0 or 1. This might be related with the problem of the so-called complete separation. In its simplest form, this problem occurs when running a logit estimation if there is a variable among the whole set of explanatory variables that explains the response variable perfectly. For instance, if in our database we would have some variable which took negative values for firms in default and positive values for firms not in default, then this variable would explain perfectly the event of default. Actually, by the simple knowledge of this variable, default could be predicted. After a careful analysis, no evidence of this situation was found in our database. However, and with no apparent reason, it was discovered that by removing the variable TVV from the estimation, no warning message was returned. Because of this, we restarted the estimation by considering all the variables as before except TVV and, following the procedure explained before, we ended up again with only statistically significant variables. The output by R software of these two models is shown in Annex E, where the model that includes TVV variable is referred to as Model 1a and Model 1b does not. After an analysis to the sign of the estimated parameters of both models, we come to the conclusion that these cannot be our final models.

Given that the inverse of the canonical link function is an increasing function, the highest the linear predictor is, the highest the probability of default. Therefore, in the case of the quantitative variables of both models, we can conclude that their negative sign is reasonable. Theoretically, the higher these ratios are, the better the economic situation of the company, the lower the estimated linear predictor and hence the lower the probability of default of the obligors.

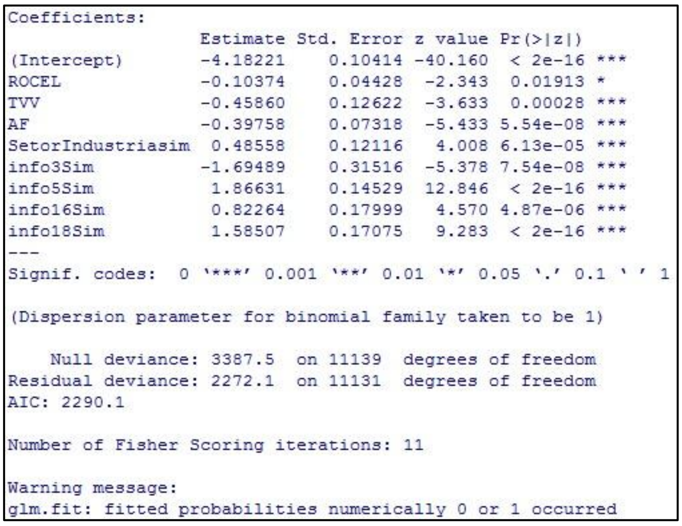
Regarding the qualitative information, we claim that all the sign are reasonable but one. Firstly, when the variable info3 takes the value “Sim”, then the exposure of the loan was increased in the last 6 months. This might mean that the economic situation of the company was reviewed carefully by the bank and so, if the increase in exposure was approved, then this obligor must show good indicators. Hence, the negative sign on this parameter seems reasonable. Secondly, the positive sign on info5, info16 and info18 seems also legitimate, given that if the company has been identified as a critical case or if the company delayed some payment by 30 days or more, then it is more likely to expect a default coming from this obligor. Lastly, concerning info31 variable, its negative sign is at least counterintuitive. This negative sign means that the companies that have been registering a decrease in the average net income are less likely to default. This might be a sign of multicollinearity between explanatory variable. Hence, it is considered appropriate to disregard this variable from the model.

Given this, estimation was started again in the way described before. First, all variables were included but info31, and step by step, eliminating the variables with highest p-value, the best model in terms of residual deviance was reached. Again, and disregarding TVV variable because of the warning message already described, another best model was reached. This last model is going to be referred to as Model 1, while the best model including TVV as Model 2.



```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -4.21184    0.10366 -40.632  < 2e-16 ***
ROCEL             -0.11827    0.04229  -2.797  0.00516 ** 
AF                -0.38198    0.07269  -5.255 1.48e-07 ***
SetorIndustriasim  0.46970    0.12058   3.895 9.80e-05 ***
info3Sim          -1.72364    0.31507  -5.471 4.48e-08 ***
info5Sim           1.89375    0.14499  13.061  < 2e-16 ***
info16Sim          0.80733    0.17915   4.506 6.59e-06 ***
info18Sim          1.61172    0.17084   9.434  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3387.5  on 11139  degrees of freedom
Residual deviance: 2294.6  on 11132  degrees of freedom
AIC: 2310.6

Number of Fisher Scoring iterations: 8
```

Figure 1 – R software output for the estimation of Model 1



```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -4.18221    0.10414 -40.160  < 2e-16 ***
ROCEL             -0.10374    0.04428  -2.343  0.01913 *  
TVV               -0.45860    0.12622  -3.633  0.00028 ***
AF                -0.39758    0.07318  -5.433 5.54e-08 ***
SetorIndustriasim  0.48558    0.12116   4.008 6.13e-05 ***
info3Sim          -1.69489    0.31516  -5.378 7.54e-08 ***
info5Sim           1.86631    0.14529  12.846  < 2e-16 ***
info16Sim          0.82264    0.17999   4.570 4.87e-06 ***
info18Sim          1.58507    0.17075   9.283  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3387.5  on 11139  degrees of freedom
Residual deviance: 2272.1  on 11131  degrees of freedom
AIC: 2290.1

Number of Fisher Scoring iterations: 11

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Figure 2 – R software output for the estimation of Model 2

Statistically speaking, we can remark that both models show an acceptable goodness-of-fit, since the residual deviance of each is less than 11 378, the 95th percentile of

$\chi^2(11131)$. When comparing both models, given that they are nested models, Model 2 is preferable in terms of residual deviance, as expected. This is because the increase by one degree of freedom from Model 2 to Model 1 is not worth, since the increase in deviance is greater than 3.841, the 95th percentile of $\chi^2(1)$. With an illustrative purpose, Figure 3 shows the probabilities of default estimated by Model 1.

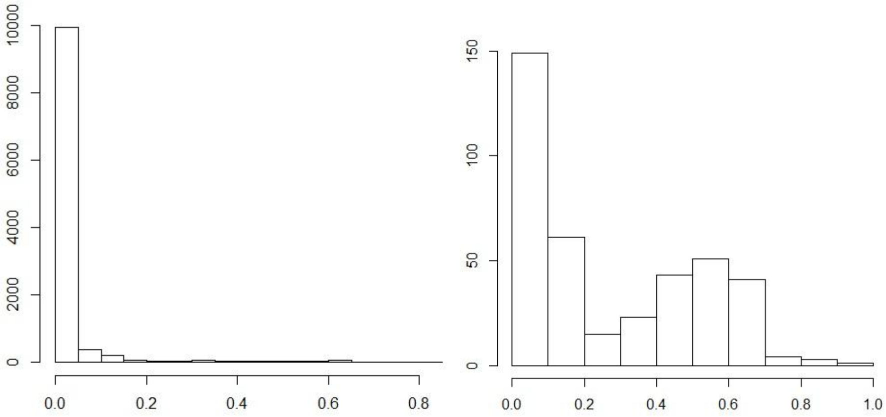


Figure 3 – Fitted probabilities of default for obligors not in default (left) and in default (right) for 2014 according to Model 1

In terms of the fitted probabilities, we can see that the great majority of obligors not in default have a probability of default close to 0. This is not verified for the ones in default. Actually the dispersion of the probability of default is not centred on 0. However, for a considerable number of obligors in default, the estimated probability of default is close to 0, which might show the weaknesses of the model already discussed. Economic conjuncture might also be an important point, since default might have occurred in cases where it was not expected at all.

The prediction power of a model is usually quantified through the Receiving Operator Characteristic (ROC) curve and the area under it. The ROC curve corresponds to the plot of the true positive rate against the false positive rate, for each threshold for considering that default is predicted. These rates are estimated, in our case, as the percentage of obligors for which default was predicted and actually happened and the percentage of obligors for which default was predicted and did not happen, respectively. The closer the area under this curve is to one, the better the model is.

Applying the estimated models to information of 2014, the probability of default for the year of 2015 is estimated. As it was provided data on the default record for the year of

2015, the prediction power of these two models can be calculated. Using the package ROCR of the software R, the ROC curves for both models are shown in Annex F. In terms of predictability, we can conclude that both Model 1 and Model 2 show a good explanatory power, given the area under the curve of 0.9139 and 0.9140, respectively.

Given all similarities between models, we are going to consider only Model 1 in the application of what follows.

3. AGGREGATE LOSS

A proper risk management of an insurance policy portfolio asks for the monitoring of its risks. These risks are quantified in the future, when the company is liable to pay the claim amount. However, it is of interest to predict today the total loss that may arise from the portfolio in the future.

Risk Theory is a branch of actuarial mathematics that aims to describe technical aspects of the insurance business through mathematical models. It might have its roots when a portfolio was first thought as a sum of insurance policies. Considering this, the aggregate loss from the portfolio is the sum of the losses arising from each individual policy.

Let S be the aggregate loss random variable of a portfolio of n independent policies in a given period of time. Therefore, if X_i is the random variable for the loss arising from policy i, then

$$S = \sum_{i=1}^{n} X_i$$

where $\{X_i\}$ are independent random variables, not necessarily identically distributed. This is actually known as the individual risk model. Under this model, we consider that each X_i has a mass point at 0, as it is not expected that all policies result in a claim.

Another way of modelling the random variable S is considering claims as arising from the whole portfolio instead of individual policies. This means that another source of randomness must be taken into account: the claim frequency, i.e. the number of claims that may arise from the portfolio. Therefore, if N is the claim frequency random variable, then the aggregate loss random variable is modelled as

$$S = \sum_{i=0}^{N} X_i$$

with $X_0 \equiv 0$.

For it to be possible to deduct some interesting results, it is usually assumed that the random variables $\{X_i\}$ are independent and identically distributed, and independent of N, where X_i denotes the severity of the i^{th} claim in the portfolio in the period of time under consideration. Under these assumptions, this model is known as the collective risk model.

As a matter of fact, the choice of the risk model depends on the framework of the problem under study. In this work, we want to model aggregate losses from a portfolio of loans of a bank. In fact, besides the fact that Risk Theory was first thought to insurance portfolios applications, it is possible to apply these models to the portfolio in question. By changing the interpretation of the variables in the model, this portfolio of loans is perfectly comparable to a portfolio in the life insurance context.

Let us consider a group life insurance portfolio that pays a fixed amount in the event of death. Interpreting policyholders as obligors, probability of death in a period of time as the probability of default and the amount that the insurance company is liable to pay in the event of death as the amount of money in default, we are in the context of the loans portfolio. In this work, and for prudent reasons, the amount of money in default is going to be considered as the amount lent at the time that the default happens.

In practice, individual risk model is more used in the life insurance context while the collective risk model captures general insurance characteristics the best. This is because individual risk model permits individual specification for the claim severity random variable. Furthermore, assuming claim severity random variables are identically distributed, as we do in the collective risk model, might not be the most proper assumption for life insurance since either the probability of death or the sum assured can be different from policyholder to policyholder. However, methods to approximate an individual risk model to a collective one have been studied. For this, both claim frequency and claim severity distributions are needed.

Considering claim frequency, we remark that a reasonable assumption is that it is Poisson distributed. Let us consider a portfolio of n loans, and group obligors with equal probability of default. Let n_i be the number of obligors with probability of default p_i. In each group, we can say that the number of defaults random variable N_i is binomial distributed, i.e. $N_i \sim \mathrm{Bi}\left(n_i, p_i\right)$. However, for a sufficiently large portfolio and given that p_i is expected to be small, the distribution of N_i can be approximated by a Poisson distribution with parameter $\lambda_i = n_i \times p_i$. It is worth noting that this approximation of a binomial to a Poisson random variable preserves its expected value. Another possible approximation would be by matching the value of the probability at zero, i.e. $\Pr\left(N_i = 0\right)$, which results in a Poisson parameter $\lambda_i = -n_i \ln\left(1 - p_i\right)$. At portfolio level, as the sum of independent Poisson random variables is still Poisson distributed, these approaches result respectively in $\lambda = \sum_i n_i p_i$ and $\lambda = -\sum_i n_i \ln\left(1 - p_i\right)$, where λ is the parameter of the claim frequency random variable N.

For small values of p_i, these two approaches are expected to give similar results. However, in our recent economic situation, this was not the case for many obligors. Therefore, both approaches will be considered further up, being referred to as Approach A and Approach B, respectively.

In terms of claim severity, when an obligor defaults, the amount of default is fixed and equal to the amount of the loan. This means that the claim severity random variable is a multiple of a binomial random variable. Particularly, if L_i is the amount of the loan of obligor i, then the claim severity random variable equals $L_i N_i$. In this context, it is worth considering the following theorem and its corollary.

<u>THEOREM 1</u>: Suppose that S_j has a compound Poisson distribution with Poisson parameter λ_j and severity distribution with distribution function $F_{X_j}\left(x\right)$, for $j = 1, 2, ..., n$. Suppose that $S_1, S_2, ..., S_n$ are independent. Then $S = S_1 + S_2 + ... + S_n$ is

compound Poisson with Poisson parameter $\lambda = \lambda_1 + \lambda_2 + ... + \lambda_n$ and severity distribution function

$$F_X(x) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda} F_{X_j}(x)$$

<u>COROLLARY 1</u>: Let $x_1, x_2, ..., x_n$ be different numbers and suppose that $N_1, N_2, ..., N_n$ are independent random variables, each of them Poisson distributed with parameter λ_i. Then, the random variable $S = x_1 N_1 + x_2 N_2 + ... + x_n N_n$ is compound Poisson with $\lambda = \lambda_1 + \lambda_2 + ... + \lambda_n$ and claim severity probability density function

$$f_X(x) = \begin{cases} \dfrac{\lambda_i}{\lambda} & \text{, if } x = x_i,\ i = 1, 2, ..., n \\ 0 & \text{, otherwise} \end{cases}$$

Given this corollary, we can conclude that the aggregate loss random variable S of our portfolio of loans can be approximated by a compound Poisson random variable. In fact, if we divide our portfolio into groups of obligors with the same characteristics, i.e. same probability of default p_i and same amount of loan L_j, then we have

$$S = \sum_i \sum_j L_j N_{ij}$$

where $N_{ij} \sim \mathrm{Po}(\lambda_{ij})$, with $\lambda_{ij} = n_{ij} p_i$ or $\lambda_{ij} = -n_{ij} \ln(1 - p_i)$, is the claim frequency random variable of the group of the n_{ij} obligors with probability of default p_i and amount of loan L_j. The collective risk model consists in considering the claim frequency random variable N to be Poisson distributed with parameter λ and, when a default occurs, it can take values L_j for $j = 1, 2, ...$ with probability equal to

$$\frac{\sum_i \lambda_{ij}}{\sum_j \sum_i \lambda_{ij}}$$

The validity of this assumption in our particular problem may be questionable. Actually, by estimating the probability of default through a logit model, it is not expected that two obligors have precisely the same probability of default. Therefore, the partition of the portfolio into groups with the same probability of default and same amount of the loan would result in one obligor per group. This fails to verify that the parameter n_{ij} is large enough. We are going to ground this assumption on Credit Suisse Financial Products (1997, p. 35), where it is referred that besides the fact that the probabilities of default are all different, the approximation of the claim frequency random variable to a Poisson random variable is a good approximation. On the other hand, by comparing the probability generating function of the aggregate loss under the individual risk model context with the probability generating function of an approximation by the corresponding compound Poisson distribution, Gerber (1990, p. 97) points out that the smaller the probabilities of default are, the better the compound Poisson approximation is. In our particular case, even with relatively large fitted probabilities of default, given that the variance of the Poisson random variable is higher than the Binomial one, this assumption will actually be applied since it is a prudent one.

The next step after defining the aggregate loss random variable is to find its distribution function, which depends upon the distribution of the random variables N and $\{X_i\}$. It is possible to find its exact distribution function by convoluting the distribution functions of $\{X_i\}$. However, when considering practical applications on relatively large portfolios, this method is time consuming in terms of calculations. To overcome this, iterative methods involving fewer amounts of computations were developed to approximate the distribution function of the aggregate loss random variable.

3.1. Panjer's recursion formula

Let us consider the set of discrete random variables X that satisfy the following formula, being $p_n = \Pr(X = n)$ and $a, b \in \mathbb{R}$,

$$p_n = \left(a + \frac{b}{n}\right) p_{n-1} \quad , \; n = 1, 2, ...$$

This set of random variables is known as $(a, b, 0)$ family and to it belong distributions such as Poisson, Binomial and Negative Binomial (which includes Geometric), as pointed in Panjer (1981). Actually, these are the only members of this family, as proved by Sundt and Jewell (1981). In the particular case of the Poisson distribution, we have that it belongs to the $(a,b,0)$ family with $a=0$ and $b=\lambda$, since

$$p_n=\frac{e^{-\lambda}\lambda^n}{n!}=\frac{\lambda}{n}\frac{e^{-\lambda}\lambda^{n-1}}{(n-1)!}=\frac{\lambda}{n}p_{n-1}$$

Panjer's recursion method is an iterative method to find aggregate loss probability density function in the context of the collective risk model. For an aggregate loss random variable such that the claim frequency distribution belongs to the $(a,b,0)$ family, let g the aggregate loss density function and f the claim severity density function, taking only values on the non-negative integers. Panjer's recursive formula is

$$\begin{cases} g(x)=\dfrac{1}{1-af(0)}\displaystyle\sum_{j=1}^{x}\left(a+b\frac{j}{x}\right)f(j)g(x-j) \quad , x=1,2,... \\ g(0)=P_N\left(f(0)\right) \end{cases}$$

where P_N is the probability generating function of N.

If we consider the aggregate loss to be compound Poisson distributed, Panjer's recursive method is

$$\begin{cases} g(0)=e^{-\lambda(1-f(0))} \\ g(x)=\dfrac{\lambda}{x}\displaystyle\sum_{j=1}^{x}jf(j)g(x-j) \end{cases}$$

3.2. Discretization of the claim frequency random variable

Considering the application of Panjer's recursion method in a practical environment, one should be aware of the frequent need for a discretization of the claim severity distribution. In fact, beyond this need, it is actually needed to transform claim severity distribution into an arithmetic distribution. An arithmetic distribution is meant to be a discrete distribution function such that all points at which a step happens are multiples of some positive number.

In the case under study in this work, as all loans are integer numbers, the claim severity distribution is discrete. Nonetheless, as the amounts of the loans vary widely, it is convenient to set a reasonable monetary unit. After the determination of this unit, criteria need to be set on how to deal now with non integer amounts. For instance, if a monetary unit of 1500 is defined, a loan of then 7500 is now considered as a loan of 5. However, it might be the case that there is a loan of 14000, which corresponds to a loan amount of 9.33 in the unit set.

There are sundry methods to "arithmitize" the claim severity distribution. The simplest ones are methods of rounding, either up, down or to the nearest. According to these methodologies, and considering the previous example, a loan of 9.33 would be considered as a loan of 10 in the first method and a loan of 9 in the last two methods. In terms of probability density function, the value $f(9.33)$ is now accounted as $f^*(10)$ or $f^*(9)$, respectively. Gerber (1990, p. 94) describes a forth method, which he calls Rounding and which consists of a rounding method to the nearest that keeps the expected value of the distribution, after an adjustment to the individual probability of default.

Another possible method, which is the one to be considered in this work, is called the method that matches the mean of the distribution. Again, and as the name suggests, after the transformation of the claim severity random variable X into an arithmetic random variable X^*, the expected value of the distribution is maintained. Formally, for a monetary unit h, the density function f^* of X^* is defined recursively as

$$f_j^*: \quad f_0^* + f_1^* + ... + f_j^* = \int_j^{j+1} F_X(hy)\,dy$$

In practical terms and because in our particular case F_X is a step function, instead of allocating $f(9.33)$ into $f^*(10)$ or $f^*(9)$ as in the methods of rounding, the method that matches the means implies that $f(9.33)$ is proportionally split contributing to both $f^*(10)$ and $f^*(9)$. Gerber (1990, p.95) calls this method Dispersion and describes it in the context of discrete random variables, where this conclusion is clearly seen. In the context of the example given, we would have that $f(9.33)$ would contribute $\frac{10-9.33}{10-9} f(9.33)$ to $f^*(9)$ and $\frac{9.33-9}{10-9} f(9.33)$ to $f^*(10)$.

The term "contribute" is being used to account for cases where two or more loans in the portfolio have the same upper or lower bound in terms of the chosen monetary unit. For instance, if we would have also a loan of 16000 in our example, which corresponds to 10.67 in the monetary unit set, then besides the contributions already described, $f^*(10)$ would have to have reflected a contribution of $\frac{10.67-10}{11-10}f(10.67)$ from this loan.

In the case of our database, the exposure amount of the loans ranges widely. Due to confidentiality reasons, the amount of the loans will not be shown and therefore the choice of the monetary unit will not be discussed. After expressing all loan amounts in the monetary unit, the highest loan is of 80 000. It was not considered a higher monetary unit, and thus a less thin “arithmatization”, because nearly 56% of the obligors have loans whose amounts are below 50 in the chosen monetary unit.

All in all, after approximating the claim severity distribution accordingly to Corollary 1, it is discretized applying the methodology discussed in this section.

3.3. Results

In this section we are going first to discuss the fitted values for the year of 2014, comparing them to the actual experience. Then, we are going to project the default for the year of 2015.

As already said, for the year of 2014, 391 defaults were registered. According to Approach A, the fitted probabilities sum up 391. This is obviously an expected figure, as the expected number of the Poisson distribution λ is the sum of individual probabilities according to this approach.

For $N \sim \mathrm{Po}(391)$, we have that $\Pr(N=0)=e^{-391}=1.55\times10^{-170}$. However, taking into account the fitted probabilities p_i, the probability of no default in the portfolio actually equals $\prod_i(1-p_i)=1.43\times10^{-210}$. Therefore, it is expected that the Poisson parameter accordingly to Approach B increase, compared to Approach A. Actually, for the year of 2014, we have that the expected number of defaults is 483.19 under Approach B.

Applying Panjer’s algorithm for the year of 2014, the aggregate loss distribution function is estimated. Thus, the percentile of the curve at which the registered loss is can be found. In this year, a total amount of 2 469 693 was lent, in the chosen monetary unit, being the total amount in default equal to 117 700. According to the adopted definition of loss, the actual loss equals 117 700, 4.77% of total loan amount. Given this percentage, the estimated percentile is a curious result.

Table III
Percentile of the loss that actually occurred in 2014

Model	Percentile of 117 700
Po_A	0.9497
Po_B	0.9114

At first these figures seem not to be reasonable. Actually, this emphasizes the questions already pointed out regarding the validity of the model. Besides this, there are three important facts that support why these figures were obtained with this model. First, the majority of the loans in our portfolio are small loans (after expressing its value in the monetary unit): as already pointed out, 56% of the loans are below 50. Second, the estimated probability of defaults for loans greater than 10% of 117 700, which are only 23 loans, are considerably small, having a mean default rate of 1.83%. Therefore, the model for the estimation of the probability of default is limited in predicting default from obligors with the largest loans, which are the ones where principal focus should be. Finally, and concerning the values of the loans that actually defaulted, 385 loans were between 0 and 2000, 4 loans between 2000 and 6000 and 2 loans between 24 000 and 26 000. In fact, the two largest defaulted loans are amongst the largest 12 of the portfolio, which is the reason for the large percentile of the registered loss.

Regarding the year of 2015, the one we are interested in projecting, the following table shows the Poisson parameter, which actually is the expected number of defaults, under Approach A (Po_A) and Approach B (Po_B).

Table IV
Poisson parameter for year of 2015

Model	Poisson parameter
Po_A	384.70
Po_B	586.61

Concerning the aggregate loss distribution, R software was used, particularly its package *actuar*, which includes Panjer's algorithm. Given the considerable large values for λ, Panjer's approximation may be questioned about its validity, as its starting value is a very small value. The function *aggregateDist* of the referred package of R draws attention to this problem, saying that Panjer's algorithm might not start or end if the value of λ is too large. The truth is that no error or warning was returned, maybe because our values for λ do not reach the too large threshold.

In this section and subsequent, the analysis of the estimated aggregate loss distribution will be made considering five percentiles in the tail of the distribution. Besides this, the estimated probability density functions for both models are shown in Annex G. In terms of percentile amounts, results are presented in Table V.

Table V
Tail percentiles of the compound Poisson aggregate loss for the year of 2015

Model	Percentile				
	90	95	97.5	99	99.5
Po_A	96 018	112 194	131 101	149 531	161 686
Po_B	120 154	136 323	153 942	173 136	185 783

4. CREDITRISK$^+$

There are four credit risk models that are recurrently considered as the most relevant ones: CreditMetrics, KMV PortfolioManager, CreditPortfolioView and CreditRisk$^+$.

Briefly describing them, CreditMetrics and KMV Portfolio Manager are usually classified as market value models. In the case of CreditMetrics, risk groups are defined accordingly, for instance, to credit quality classification of the company, being the worst risk group related to default. The probability of default is therefore equal to all obligors in the same risk group. Then, and based on historical record, the probability of moving from one state to another is estimated, entering in the credit migration framework. Using Monte Carlo simulation, portfolio default loss distribution is then generated according to the market value change of the asset portfolio of the company due to credit migration only. Market value change is tracked consistently with Merton's Model, an option pricing model for the valuation of equity based on Black-Scholes, extending it to incorporate credit migration.

Concerning KMV PortfolioManager, the approach is to derive individual probabilities of default, the Expected Default Frequency (EDF), of each obligor rather than historical transition frequency. Following Merton's Model too, the term "distance to default" is defined. An extension to Merton's Model is also done, to account for the refinancing abilities of companies. EDF is defined as a function of the "distance to default", which depends on the firm's financial structure. Based on the estimation of the correlation between default probabilities and default record, credit rating migration matrix can be derived as well as default loss distribution.

CreditPortfolioView is classified as the econometric model, as the probability of default is defined to depend on macroeconomic scenarios. By setting up a multi-factor model to account for systemic risk, probability of default is estimated through a logit model. According to this model, default loss distribution is derived taking into account the relationship between credit migration matrix and macroeconomic indicators.

Among credit risk models, CreditRisk$^+$ is classified as the actuarial model. It is going to be studied and described in detail in the next two subsections. In the first one, we are going to compare the simplest form of this model to the work developed in Section 3. Then, it is going to be briefly shown how to reach CreditRisk$^+$ formula in its generalized form and put it into practice in our database.

By the end of Section 4, it should be clear the reason why CreditRisk$^+$ is considered to be the actuarial model.

4.1. CreditRisk+ with fixed default rate

CreditRisk$^+$ model does not include a methodology for the estimation of the probabilities of default. Nevertheless, this is required as an input to the model.

Assuming that the probabilities of default of each individual obligor are known, Credit Suisse Financial Products (1997), referred henceforward as CSFP (1997), deduces the probability generating function of the claim frequency random variable, concluding by a Poisson random variable. Concerning the Poisson parameter, Approach A is used, assuming probabilities of default are small enough.

Concerning the arithmatization of the default severity random variable, exposure is adjusted by some unit amount. Then, and to preserve the expected loss, a rounding adjustment is made to the expected number of defaults. This is actually the referred Rounding method described by Gerber (1990, p. 94).

The next step is to find the probability generating function of the aggregate loss arising from the portfolio. Without referring to the theoretical background, CSFP (1997) concludes that the probability generating function of the aggregate loss random variable is of the form of a compound Poisson random variable, besides that they do not classify it as a compound Poisson explicitly. Actually, the probability generating function of the claim severity is consistent with Corollary 1.

Finally, an iterative algorithm to find the density function of the aggregate loss is deduced. In their notation, the algorithm is presented as

$$A_n = \sum_{j:\nu_j \le n} \frac{\varepsilon_j}{n} A_{n-\nu_j}$$

where A_n is the probability that an aggregate loss of amount n occurs and υ_j and ε_j are respectively the exposure amount and the expected loss in exposure band j, expressed in the settled monetary unit. The relation between these two quantities is

$$\varepsilon_j = \upsilon_j \times \mu_j$$

where μ_j is the expected number of defaults in exposure band j.

In our notation, μ_j is λ_j and υ_j is simply j. Therefore, the algorithm presented in CSFP (1997) in our notation is

$$g(n) = \sum_{j=1}^{n} \frac{j\lambda_j}{n} g(n-j) = \frac{\lambda}{n} \sum_{j=1}^{n} j \frac{\lambda_j}{\lambda} g(n-j)$$

This is in fact Panjer’s recursive formula since, according to the Corollary 1,

$$f(j) = \frac{\lambda_j}{\lambda}$$

Concluding, as it is now perfectly clear, the simplest form of CreditRisk+ is a direct application of Panjer’s algorithm within the formalization described in Section 3.

4.2. CreditRisk+ with variable default rate

CreditRisk$^+$ model generalizes the simpler model discussed in the previous section. After introducing volatility to the probability of default and sector analysis, a new iterative formula is deduced following the same reasoning.

The concept of sector is user adaptable. A sector might be interpreted as the sector of activity, the size of the company or even the country of domicile of the obligor. The idea is to make a partition in the set of obligors in such a way that the probability of default of the obligors in a specific sector is influenced by the same external uncontrollable factors. As in CSFP (1997), we are going to assume that each sector is driven by only one factor.

The underlying factor of each sector will influence it through the total expected rate of defaults. Therefore, the total number of defaults arising in sector k is going to be a random variable N_k with mean μ_k and standard deviation σ_k.

Formally, instead of having $N_k \sim \mathrm{Po}(\lambda_k)$, where λ_k is the expected number of losses in sector k, which corresponds to the sum of individual probabilities of default of obligors in that sector, we are now going to assume that N_k given $\Theta_k = \theta_k$ follows a Poisson distribution with Poisson parameter $\lambda_k \theta_k$. Therefore, Θ_k is a random variable that accounts for the volatility in the individual probability of default. The key assumption of CreditRisk$^+$ is that Θ_k follows a Gamma distribution. For the parameterization of the Gamma distribution, we are going to follow the one also used by Klugman *et al.* (2004).

To find the parameters of this distribution, we are going to impose that the expected value of Θ_k is 1, so that the expected the number of claims in sector k is λ_k. Hence, being ω_k^2 the variance of $\Theta_k \sim \mathrm{Gamma}(\alpha_k, \beta_k)$, we have that

$$\Theta_k \sim \mathrm{Gamma}\left(\frac{1}{\omega_k^2}, \omega_k^2\right).$$

We can easily deduct that

$$P_{N_k}(z)=E\left[z^{N_k}\right]=E\left[E\left[z^{N_k}|\Theta_k\right]\right]=E\left[e^{\lambda_k\Theta_k(z-1)}\right]=M_{\Theta_k}\left(\lambda_k(z-1)\right)=\left(1-\omega_k^2\lambda_k(z-1)\right)^{-\frac{1}{\omega_k^2}}$$

where M_{Θ_k} is the moment generating function of Θ_k. This last expression is the probability generating function of a Negative Binomial random variable with parameters $r=1/\omega_k^2$ and $\beta=\omega_k^2\lambda_k$, in the Klugman *et al*. (2004) parameterization. We can therefore conclude that N_k follows a Negative Binomial distribution.

After finding the distribution of the claim frequency, we are interested in finding the aggregate loss distribution within each sector. Let us find its probability generating function. For simplicity reasons, the subscript k is going to be dropped in the following proof, but it must be kept in mind that we are within the sector. Hence, in the context of mixed frequency models, it is known that

$$P_S(z)=\int P_{S|\Theta=\theta}(z)\,f_\Theta(\theta)\,\partial\theta$$

where P_S is the probability generating function of S and f_Θ the probability density function of Θ. It is important to remark that $S|\Theta=\theta$ is the aggregate loss random variables for the fixed default rate case. Knowing $\Theta=\theta$, $S|\Theta=\theta$ is a compound Poisson random variable with probability generating function

$$P_{S|\Theta=\theta}(z)=e^{\lambda\theta\left(P_X(z)-1\right)}$$

where $P_X(z)$ is the probability generating function of the claim severity random variable with density function accordingly to Corollary 1, eventually after arithmatization. Hence, we have that

$$\begin{aligned}P_S(z)&=\int e^{\lambda\theta\left(P_X(z)-1\right)}\frac{\theta^{\left(1/w^2\right)-1}e^{-\theta/w^2}}{\left(w^2\right)^{1/w^2}\Gamma\left(1/w^2\right)}\partial\theta\\&=\frac{1}{\left(w^2\right)^{1/w^2}}\int\frac{\theta^{\left(1/w^2\right)-1}e^{-\theta\left[\left(1/w^2\right)-\lambda\left(P_X(z)-1\right)\right]}}{\Gamma\left(1/w^2\right)}\partial\theta\\&=\frac{\left[\left(1/w^2\right)-\lambda\left(P_X(z)-1\right)\right]^{-\left(1/w^2\right)}}{\left(w^2\right)^{1/w^2}}\int\frac{\theta^{\left(1/w^2\right)-1}e^{-\theta\left[\left(1/w^2\right)-\lambda\left(P_X(z)-1\right)\right]}}{\left[\left(1/w^2\right)-\lambda\left(P_X(z)-1\right)\right]^{-\left(1/w^2\right)}\Gamma\left(1/w^2\right)}\partial\theta\end{aligned}$$

Given that the expression inside the integral is the probability density function of a Gamma random variable with parameters $\left(1/w^2\right)$ and $\left[\left(1/w^2\right)-\lambda\left(P_X\left(z\right)-1\right)\right]$, then

$$\begin{aligned} P_S\left(z\right) &= \left[w^2\left(\left(1/w^2\right)-\lambda\left(P_X\left(z\right)-1\right)\right)\right]^{-\left(1/w^2\right)} \\ &= \left[1-w^2\lambda\left(P_X\left(z\right)-1\right)\right]^{-\left(1/w^2\right)} \end{aligned}$$

The last expression allows reaching the conclusion that the aggregate loss random variable within each sector k follows a compound Negative Binomial distribution, with Negative Binomial parameters $r_k=\left(1/w_k^2\right)$ and $\beta_k = w_k{}^2\lambda_k$, and claim severity distribution as in Corollary 1. In other words, given $\Theta_k=\theta_k$, if S_k is a compound Poisson $CP\left(\lambda_k\theta_k, F_X\right)$, then S_k is unconditionally a compound Negative Binomial with the same severity distribution F_X .

Regarding the whole portfolio, the sum of independent compound Negative Binomial random variables might not be compound negative Binomial distributed. In our case, the aggregate loss is not a compound negative Binomial random variable, as β_k are different in each sector. Therefore, to find the aggregate loss distribution, convolution techniques are applied. For instance, if the portfolio is divided into two sectors, then the probability density function of aggregate loss of the whole portfolio would be such that

$$\Pr\left(S=n\right)=\Pr\left(S_1+S_2=n\right)=\sum_{m=0}^{n} f_1\left(m\right) f_2\left(n-m\right)$$

Considering the case in which the portfolio is divided into three sectors, then

$$\begin{aligned} \Pr\left(S=n\right)=\Pr\left(S_1+S_2+S_3=n\right) &= \sum_{m=0}^{n} f_1\left(n-m\right) f_{2+3}\left(m\right) \\ &= \sum_{m=0}^{n} f_1\left(n-m\right)\sum_{s=0}^{m} f_2\left(m-s\right) f_3\left(s\right) \end{aligned}$$

In CSFP (1997) an iterative formula to find the aggregate loss probability density function of the whole portfolio is deduced. In their notation,

$$A_{n+1}=\frac{1}{b_0\left(n+1\right)}\left(\sum_{i=0}^{\min(r,n)} a_i A_{n-i} - \sum_{j=0}^{\min(s-1,n-1)} b_{j+1}\left(n-j\right)A_{n-j}\right)$$

where $A_n = \Pr(S = n)$ and a_i and b_j are the coefficients of the polynomials $A(z)$ and $B(z)$ such that

$$\frac{\partial}{\partial z}\log\left(P_S\left(z\right)\right) = \frac{1}{P_S\left(z\right)}\frac{\partial P_S\left(z\right)}{\partial z} = \frac{A\left(z\right)}{B\left(z\right)} = \frac{a_0 + a_1 z + \ldots + a_r z^r}{b_0 + b_1 z + \ldots + b_s z^s}$$

In the form this formula is presented, it is first needed to find the coefficients a_i and b_j and then apply the recursive formula. This might be computationally demanding, when comparing to the algorithms already available in R software. Because of this, in the practical application R commands are going to be used for the calculation of the distribution function within each sector and then convolve them to find the aggregate loss distribution function of the whole portfolio.

4.2.1. Results

As remarked in the previous section, N_k follows a Negative Binomial distribution whose parameters depend on λ_k and ω_k^2, from which we only know λ_k. To determine ω_k^2, as we lack data to estimate it empirically, we are going to ground our assumption on CSFP (1997), where it is said that, according to historical experience, the standard deviation of the number of defaults observed, year on year in the same sector, is typically of the same order as the average annual number of defaults. Therefore, we are going to assume that, for some constant ρ

$$\sqrt{\text{var}\left(\lambda_k \Theta_k\right)} = \rho \text{E}\left[\lambda_k \Theta_k\right]$$

Given that the expected value and the standard deviation of the number of default are λ_k and $\lambda_k \omega_k$, respectively, solving the equation leads to $\omega_k = \rho$. In the practical application, we considering $\rho = 1.1$ and $\rho = 1.5$. This implies that N_k is such that

$$N_k \sim \text{NB}\left(\frac{1}{\rho^2}\ ,\ \rho^2 \lambda_k\right)$$

When it comes to the number of sectors, two approaches will be addressed. At a more simple level, we are going to consider only one sector. This might be interpreted, for instance, as partition by domicile country, as all obligors in the portfolio are Portuguese entities. Then, we are going to consider three sectors, accounting for the sector of activity: commerce, manufacturing and services. In this case, as we are not following CreditRisk$^+$ formula directly, we are going to apply Panjer's algorithm in each sector, and then apply convolution to find the aggregate loss distribution in each sector.

Let us first analyse the one sector case. Let NB1 and NB2 stand for the compound Negative Binomial models studied within one sector for $\rho = 1.1$ and $\rho = 1.5$, respectively. For each value of ρ, both Approach A and B that determine the values λ_k are applied. Table VI shows the values obtained for the chosen percentiles.

Table VI
Tail percentiles of the compound Negative Binomial aggregate loss considering 1 sector for the year of 2015

Model	Percentile				
	90	95	97.5	99	99.5
NB1_A	169 422	226 514	284 170	360 975	419 396
NB1_B	221 281	295 164	369 783	469 188	544 802
NB2_A	192 473	279 190	370 182	494 868	591 560
NB2_B	252 268	365 475	484 274	647 070	773 318

In the three sectors case, let NB3 and NB4 stand for the compound Negative Binomial models studied within three sector for $\rho = 1.1$ and $\rho = 1.5$, respectively. The obtained percentiles for the same models as above are shown in Table VII.

Table VII
Tail percentiles of the compound Negative Binomial aggregate loss considering 3 sectors for the year of 2015

Model	Percentile				
	90	95	97.5	99	99.5
NB3_A	139 880	174 710	209 228	254 632	288 939
NB3_B	180 381	223 897	266 974	323 654	366 483
NB4_A	157 405	208 000	260 326	331 826	387 380
NB4_B	204 572	268 894	335 155	425 361	495 286

As we can see in both Table VI and Table VII, percentiles increase when considering $\rho = 1.5$ instead of $\rho = 1.1$. As the volatility in the number of claims is now higher, higher amounts of losses are more likely to occur.

Comparing Table VI to Table V, the compound Negative Binomial model considering only one sector is comparable to the compound Poisson model, as it only introduces more volatility to the number of defaults. By increasing the standard deviation by 10%

and 50%, much higher percentiles are obtained.

When comparing the results considering different number of sectors, we can remark that the tail percentiles of the aggregate loss distribution decrease when considering three sectors instead of only one. Again, this is an expected result, because the volatility of the number of defaults within the three sectors framework is lower, as sectors are assumed to be independent and because it was assumed the same value of ρ for the one-sector case and for each sector in the three-sector case. Given this, being N_i the claim frequency random variable in sector i**,** we have that the variance of the default frequency when considering three sectors is

$$\operatorname{var}\left(N_1+N_2+N_3\right)=\sum_{i=1}^{3}\operatorname{var}\left(N_i\right)=\sum_{i=1}^{3}\lambda_i\left(1+\rho^2\lambda_i\right)$$

Comparably, the variance of the number of claims when default volatility is driven by only one sector is higher, given that we are always assuming the same value for ρ.

$$\begin{aligned}\operatorname{var}\left(N\right)&=\lambda\left(1+\rho^2\lambda\right)=\left(\lambda_1+\lambda_2+\lambda_3\right)\left(1+\rho^2\left(\lambda_1+\lambda_2+\lambda_3\right)\right)\\&=\lambda_1\left(1+\rho^2\left(\lambda_1+\lambda_2+\lambda_3\right)\right)+\lambda_2\left(1+\rho^2\left(\lambda_1+\lambda_2+\lambda_3\right)\right)+\lambda_3\left(1+\rho^2\left(\lambda_1+\lambda_2+\lambda_3\right)\right)\\&\geq\lambda_1\left(1+\rho^2\lambda_1\right)+\lambda_2\left(1+\rho^2\lambda_2\right)+\lambda_3\left(1+\rho^2\lambda_3\right)\end{aligned}$$

5. APPROXIMATIONS TO THE AGGREGATE LOSS DISTRIBUTION

In this chapter, other methods of approximating the aggregate loss distribution are going to be presented. They are usually considered as an alternative to Panjer's recursive algorithm because, as any other iterative process, Panjer relies heavily on the first term, namely $\Pr(S=0)$. For a large portfolio, as in our case, this term is very small, which might imply that the algorithm will have some problems. The Normal Power (NP) and the Translated Gamma approximations are usually used when the skewness coefficient γ_S of the aggregate loss distribution is higher than 0.1, giving good approximation for the tail of the distribution.

Despite this, it is important to highlight that these approximations rely only on the knowledge of the first three moments of the aggregate loss distribution. Therefore, they might be preferable to Panjer's, as they are much less time consuming.

5.1. NP approximation

Let Z be the standardized aggregate loss random variable, i.e.

$$Z = \frac{S - \mu_S}{\sigma_S}$$

The NP approximation is based on a formula known as Edgeworth series. Approximating the distribution function of Z by the first two terms of this series,

$$F_Z(z) \approx \Phi(z) - \frac{\gamma_S}{6}\Phi^{(3)}(z)$$

where Φ is the distribution function of a standard normal random variable and $\Phi^{(3)}$ its third derivative. After some mathematics,

$$F_Z\left(z + \frac{\gamma_S}{6}\left(z^2 - 1\right)\right) \approx \Phi(z)$$

Therefore, solving the equation $z + \frac{\gamma_S}{6}\left(z^2 - 1\right) = y$ in z, we have that

$$F_Z(y) \approx \Phi\left(-\frac{3}{\gamma_S} + \sqrt{\frac{9}{\gamma_S^2} + 1 + \frac{6}{\gamma_S} y}\right) \Leftrightarrow F_S(x) \approx \Phi\left(-\frac{3}{\gamma_S} + \sqrt{\frac{9}{\gamma_S^2} + 1 + \frac{6}{\gamma_S}\frac{x - \mu_S}{\sigma_S}}\right)$$

This last formula is known as NP approximation.

5.2. Translated Gamma approximation

The Translated Gamma approximation, as the name suggests, approximates the aggregate loss random variable S by a Gamma(α,θ) random variable Y translated k units, in such a way that both random variables S and $k+Y$ have the same mean, variance and skewness coefficient. Therefore, given μ_S, σ_S^2 and γ_S, the following equations define this approximation

$$\begin{cases} \mu_S = k + \alpha\theta \\ \sigma_S^2 = \alpha\theta^2 \\ \gamma_S = \dfrac{2}{\sqrt{\alpha}} \end{cases}$$

After solving these equations for k, α and θ, we can conclude that S is approximated by the random variable $\mu_S - \dfrac{2\sigma_S}{\gamma_S} + Y$, where $Y \sim \text{Gamma}\left(\dfrac{4}{\gamma_S^2}, \dfrac{\sigma_S\gamma_S}{2}\right)$.

5.3. Results

In this chapter we are going to compare the percentiles for the models presented in both Table V and Table VI, according to the NP and the Translated Gamma approximations. As already pointed out, both these approximations rely on μ_S, σ_S and γ_S. Here, being under a collective risk model, calculations become simpler, as the moments of S depend on the moments of N and of X, according to

$$\begin{cases} \mathrm{E}[S] = \mathrm{E}[N]\mathrm{E}[X] \\ \mathrm{var}[S] = \mathrm{E}[N]\mathrm{var}[X] + \mathrm{var}[N]\mathrm{E}^2[X] \\ \mu_3[S] = \mu_3[N]\mathrm{E}^3[X] + 3\mathrm{var}[N]\mathrm{E}[X]\mathrm{var}[X] + \mathrm{E}[N]\mu_3[X] \end{cases}$$

where $\mu_3[S]$ stands for the third central moment of S. From this, skewness coefficient may be derived as

$$\gamma_S = \frac{\mu_3[S]}{(\mathrm{var}[S])^{3/2}}$$

For each model presented in both Table V and Table VI, expected value, standard deviation and skewness coefficient information is displayed in Annex H. In addition, it is also included the compound Binomial case (Bi model) such that

$$S = \sum_i \sum_j L_j N_{ij}$$

where $\mu_S = \sum_{i,j} n_{ij} p_i L_j$, $\sigma_S = \sum_{i,j} n_{ij} p_i (1 - p_i) L_j^2$ and $\mu_3(S) = \sum_{i,j} n_{ij} p_i (1 - p_i)(1 - 2p_i) L_j^3$.

The p^th percentile of the aggregate loss under the NP approximation is given by

$$x_p = \mu_S + \sigma_S \left[\Phi^{-1}(p) + \frac{\gamma_S}{6}\left(\left(\Phi^{-1}(p) \right)^2 - 1 \right) \right]$$

while under the Translated Gamma approximation, being y_p the p^th^ percentile of Y,

$$x_p = \mu_S - \frac{2\sigma_S}{\gamma_S} + y_p$$

Given this, Table VIII and Table IX show the results obtained.

Table VIII
Tail percentiles of the NP approximation for the aggregate loss for the year of 2015

Model	Percentile				
	90	**95**	**97.5**	**99**	**99.5**
Po_A	101 594	116 560	130 892	149 131	162 521
Po_B	124 960	139 853	153 989	171 848	184 884
NB1_A	185 819	241 930	296 137	365 635	416 939
NB1_B	243 272	317 261	388 855	480 764	548 680
NB2_A	234 811	325 428	414 568	530 525	617 061
NB2_B	308 443	428 267	546 253	699 858	814 556
Bi	99 678	114 324	128 432	146 474	159 768

Table IX
Tail percentiles of the Translated Gamma approximation for the aggregate loss for the year of 2015

Model	Percentile				
	90	**95**	**97.5**	**99**	**99.5**
Po_A	98 538	113 159	127 529	146 263	160 292
Po_B	122 634	137 289	151 485	169 722	183 342
NB1_A	170 973	225 211	179 387	350 940	405 033
NB1_B	222 778	294 129	365 622	460 278	531 964
NB2_A	195 399	279 658	367 311	486 678	578 864
NB2_B	254 940	365 999	481 844	639 914	762 147
Bi	96 122	110 336	124 454	143 017	157 005

There are interesting conclusions to be taken. For this, Annex I shows the percentage variation of each percentile considering these two approximations when compared to the percentile obtained following Panjer's algorithm.

Regarding the NP approximation, some tendencies on the goodness of this approximation are clear. The least the variance of the claim frequency random variables is, the better the NP approximation. In fact, from the Poisson model to the Negative Binomial with $\rho = 1.1$ and then to Negative Binomial with $\rho = 1.5$, NP approximation worsens. Nevertheless, the approximation is quite good when considering the claim frequency as Poisson distributed.

Generically speaking, Translated Gamma gives a better approximation. Interestingly, it overestimates the first two percentiles considered (except for NB1) and underestimates the other ones. This means that the Gamma distribution has a comparatively less heavy tail, still not significant. There is actually no pattern to deduce in what cases the approximation would be even better, as it relies on matching the moments of both distributions.

Besides the fact that the Translated Gamma is generically a better approximation in the percentiles considered, it underestimates the highest percentile considered. On the other hand, as we consider higher percentiles in the NP approximation, the better it is, being actually better than the Translated Gamma one for Po and NB1 models.

As a conclusion, and from the perspective of the risk management of a bank, it is important to highlight that the underestimation of a loss in the future might be critical. Nevertheless, and from the practical point of view, after the computation of the moments of the aggregate loss, these approximation methods return instantaneous results. Depending on the portfolio size, these methods are definitely worth to consider.

6. AVERAGE INTEREST RATE

Interestingly, by finding the pth percentile of the aggregate loss distribution, which can be done by recurring to the aggregate loss distributions estimated on the previous chapters of this work, the average interest rate r can be determined. Let us denote by U the surplus of the bank after one year with respect to this portfolio of loans. Being u the initial reserve that the bank might have to account for future losses, then U equals

$$U = u + Cashflows_{in} - Cashflows_{out}$$

In the context of the problem under study, we have that the cashflows in equal the interest rate earned on the loans that do not default. On the other hand, cashflows out equal the amount of the aggregate loss registered in the one year period considered. Therefore, being V the total amount lent by the bank, we have that

$$U = u + r\left(V - S\right) - S$$

Thus, the probability that the bank has enough money to cover losses within one year, which might be interpreted as a survival probability, is given by

$$\Pr(U \geq 0) = \Pr\left(u + r(V - S) - S \geq 0\right) = \Pr\left(S \leq \frac{u + rV}{1 + r}\right)$$

Defining k such that $\Pr(S \leq k) = p$, the interest rate r can be determined as

$$\Pr\left(S \leq \frac{u + rV}{1 + r}\right) = p \Leftrightarrow \frac{u + rV}{1 + r} = k \Leftrightarrow r = \frac{k - u}{V - k}$$

On the other hand, if the interest rate is settled, the amount of money that should be reserved in the beginning of the year to account for losses is given by

$$u = (1 + r)k - rV = k - r(V - k)$$

6.1. Results

For the application part of this chapter, the percentiles of models Po_A and NB1_A shown in Table V and Table VI, respectively, are used. To determine the interest rate r, the worst case scenario is considered in terms of initial reserve, i.e. $u = 0$.

Table X
Average interest rate for the models obtained from Panjer algorithm

Model	Survival probability		
	90	97.5	99.5
Po_A	4.80%	6.67%	8.36%
NB1_A	8.79%	15.68%	25.01%

As a matter of fact, the higher the percentile amount and the probability level of survival are, the higher the average interest rate to be charged. These interest rates might be thought of as the maximum interest rate to be charged, for each survival probability, given that we are considering that the bank has no reserve to cover defaults. On the other hand, the initial reserve can be determined as a function of the average interest rate charged. In Table XI, we can conclude that it naturally increases as the survival probability increases and the average interest rate decreases.

Table XI
Initial reserve considering the models obtained from Panjer algorithm

	r = 3%			r = 5%		
Model	Survival probability			Survival probability		
	90	97.5	99.5	90	97.5	99.5
Po_A	36 001	72 137	103 640	-4 010	32 828	64 942
NB1_A	111 608	229 798	369 081	73 065	193 550	335 537

7. CONCLUSION

All calculations performed throughout this work depend on the estimated probabilities of default. As identified, the chosen logit model has some limitations, as it was estimated using data from a year where firm were under stressed conditions. Therefore, it should be advised the model to be reviewed in the coming years. As remarked before, the ultimate purpose was not to study this particular portfolio of loans, but use it to illustrate the application of the theories discussed.

In this work we were interested in quantifying default risk. This was done through some percentiles of the aggregate loss distribution function, obtained with a varied set of methodologies. First, the simpler version of CreditRisk$^+$, which is actually Panjer's algorithm for a compound Poisson distribution, was applied. Then, and approximating the methodology to the CreditRisk$^+$ one, Panjer's algorithm was again applied but now for a compound Negative Binomial distribution. As remarked, this transition is accomplished by changing the claim frequency distribution. In terms of aggregate loss distribution, the more volatile the claim frequency random variable is, the more significant the right tail of the aggregate loss distribution is. This was noted by the increasing amount of each percentile.

Questioning if similar results could be obtained with more simple approximation methods, the NP and the Translated Gamma approximation were tested and results were satisfactory, supporting that these methods can be used instead. Generally, Translated Gamma approximation gives better results. However, NP approximation might be an alternative for really high percentiles.

The work developed in the last section was limited. As our estimated probabilities for a given year depend on the financial information of the previous year, it is not possible to project future probabilities of default. This could be interesting, for instance, to apply Ruin Theory reasoning in order to quantify whether this portfolio of loans might be profitable. For this, a Markov chain could have been estimated, where obligors were mapped to a given state which would have a probability of default associated. Intrisically linked to a Markov process, transitions between states would allow for the estimation of the future probability of default. This idea stays as suggestion for future investigation.

8. ANNEX

A. Summary of the quantitative variables to be considered

Variable	min	1st quartile	median	mean	3rd quartile	max
ROCEL	-29	0.015	0.059	0.188	0.143	161
TVV	-1	-0.107	0.027	1.199	0.191	4180
FMNFV	-1622	0.066	0.233	4.981	0.519	29210
AF	-14	0.147	0.284	0.265	0.455	1
JVPS	0	0.003	0.012	0.256	0.029	1127

Table A.I – Summary of the quantitative variables for the year of 2013

Variable	min	1st quartile	median	mean	3rd quartile	max
ROCEL	-680	0.010	0.062	0.087	0.150	580
TVV	-1	-0.094	0.030	0.761	0.183	3104
FMNFV	-542	0.068	0.231	2.833	0.492	11610
AF	-47	0.144	0.294	0.242	0.465	1
JVPS	0	0.004	0.012	1.077	0.029	8784

Table A.II – Summary of the quantitative variables for the year of 2014

B. Summary of the qualitative variables to be considered

Variable	Sim	Nao
info3	2632	8508
info5	700	10440
info16	416	10724
info18	797	10343
info31	9043	2097

Table B.I – Summary of the qualitative information for the year of 2013

Variable	Sim	Nao
info3	2616	7599
info5	668	9547
info16	393	9822
info18	773	9442
info31	8629	1586

Table B.II – Summary of the qualitative information for the year of 2014

C. Summary of the variables *Dimensao* and *Setor*

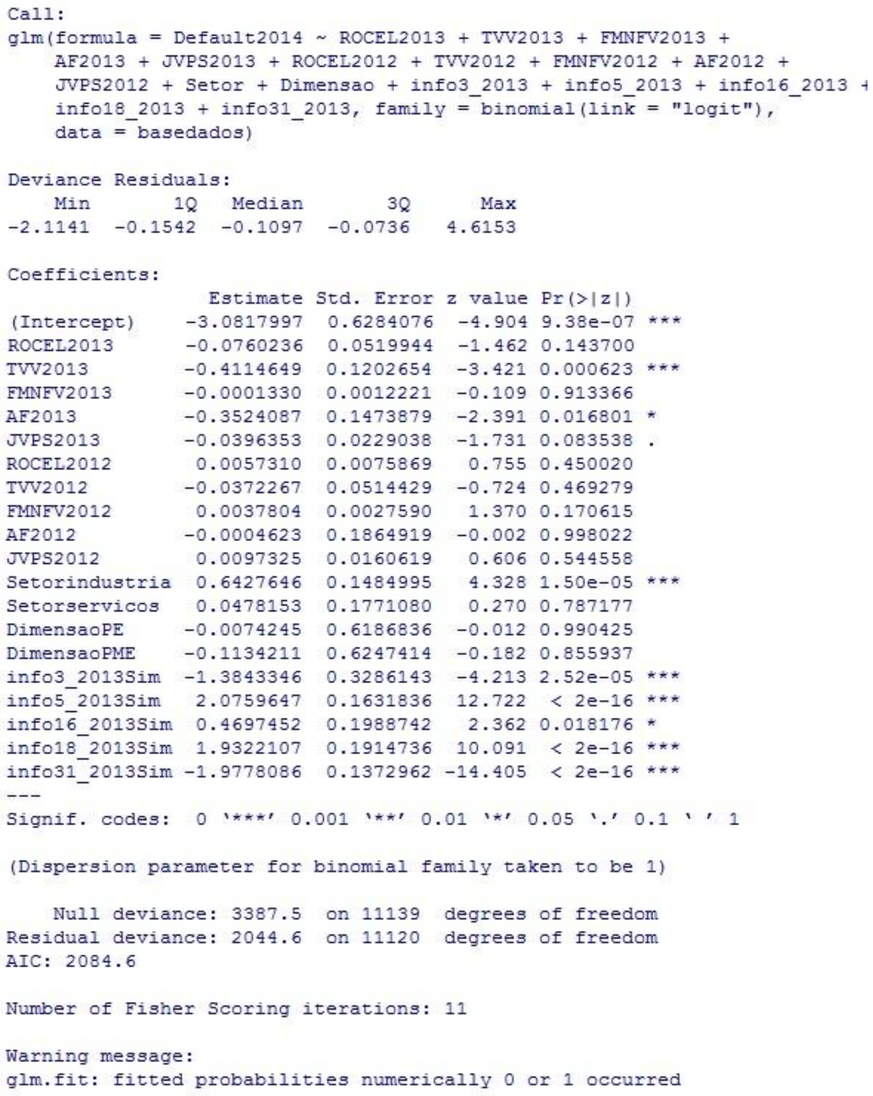
Year	GRE	PME	PE
2014	108	3527	7505
2015	101	3294	6820

Table C.I – Summary of the variable *Dimensao*

Year	comercio	servicos	industria
2014	4109	2650	4381
2015	3763	2385	4067

Table C.II – Summary of the variable *Setor*

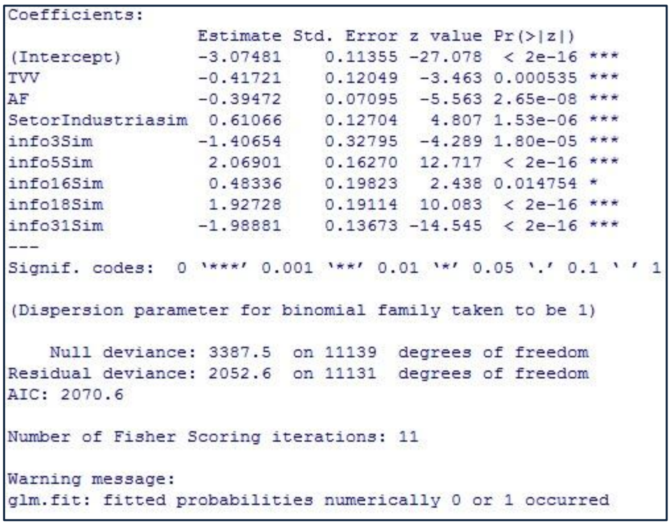
D. Linear predictor estimation taking into account all variables



```
Call:
glm(formula = Default2014 ~ ROCEL2013 + TVV2013 + FMNFV2013 +
    AF2013 + JVPS2013 + ROCEL2012 + TVV2012 + FMNFV2012 + AF2012 +
    JVPS2012 + Setor + Dimensao + info3_2013 + info5_2013 + info16_2013 +
    info18_2013 + info31_2013, family = binomial(link = "logit"),
    data = basedados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1141  -0.1542  -0.1097  -0.0736   4.6153

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.0817997  0.6284076  -4.904 9.38e-07 ***
ROCEL2013      -0.0760236  0.0519944  -1.462 0.143700
TVV2013        -0.4114649  0.1202654  -3.421 0.000623 ***
FMNFV2013      -0.0001330  0.0012221  -0.109 0.913366
AF2013         -0.3524087  0.1473879  -2.391 0.016801 *
JVPS2013       -0.0396353  0.0229038  -1.731 0.083538 .
ROCEL2012       0.0057310  0.0075869   0.755 0.450020
TVV2012        -0.0372267  0.0514429  -0.724 0.469279
FMNFV2012       0.0037804  0.0027590   1.370 0.170615
AF2012         -0.0004623  0.1864919  -0.002 0.998022
JVPS2012        0.0097325  0.0160619   0.606 0.544558
Setorindustria  0.6427646  0.1484995   4.328 1.50e-05 ***
Setorservicos   0.0478153  0.1771080   0.270 0.787177
DimensaoPE     -0.0074245  0.6186836  -0.012 0.990425
DimensaoPME    -0.1134211  0.6247414  -0.182 0.855937
info3_2013Sim  -1.3843346  0.3286143  -4.213 2.52e-05 ***
info5_2013Sim   2.0759647  0.1631836  12.722  < 2e-16 ***
info16_2013Sim  0.4697452  0.1988742   2.362 0.018176 *
info18_2013Sim  1.9322107  0.1914736  10.091  < 2e-16 ***
info31_2013Sim -1.9778086  0.1372962 -14.405  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3387.5  on 11139  degrees of freedom
Residual deviance: 2044.6  on 11120  degrees of freedom
AIC: 2084.6

Number of Fisher Scoring iterations: 11

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

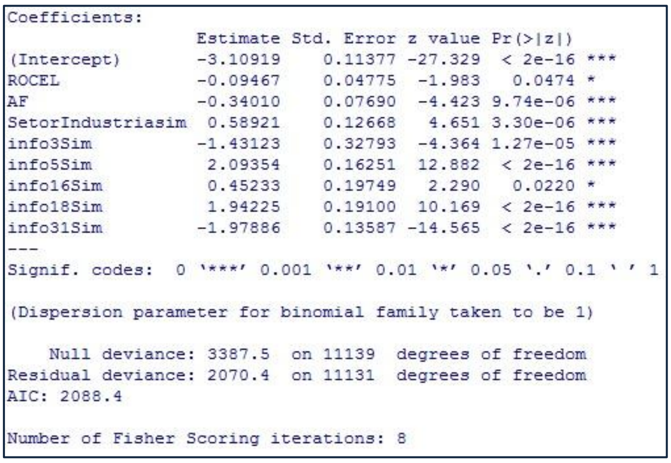
Figure D.1 – R software output for the estimation of the linear predictor of a logistic regression taking into account all variables

E. Linear predictor estimation of Model 1a and Model 1b



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.07481    0.11355 -27.078  < 2e-16 ***
TVV                -0.41721    0.12049  -3.463 0.000535 ***
AF                 -0.39472    0.07095  -5.563 2.65e-08 ***
SetorIndustriasim   0.61066    0.12704   4.807 1.53e-06 ***
info3Sim           -1.40654    0.32795  -4.289 1.80e-05 ***
info5Sim            2.06901    0.16270  12.717  < 2e-16 ***
info16Sim           0.48336    0.19823   2.438 0.014754 *
info18Sim           1.92728    0.19114  10.083  < 2e-16 ***
info31Sim          -1.98881    0.13673 -14.545  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3387.5  on 11139  degrees of freedom
Residual deviance: 2052.6  on 11131  degrees of freedom
AIC: 2070.6

Number of Fisher Scoring iterations: 11

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Figure E.1 – R software output for the estimation of Model 1a



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.10919    0.11377 -27.329  < 2e-16 ***
ROCEL              -0.09467    0.04775  -1.983   0.0474 *
AF                 -0.34010    0.07690  -4.423 9.74e-06 ***
SetorIndustriasim   0.58921    0.12668   4.651 3.30e-06 ***
info3Sim           -1.43123    0.32793  -4.364 1.27e-05 ***
info5Sim            2.09354    0.16251  12.882  < 2e-16 ***
info16Sim           0.45233    0.19749   2.290   0.0220 *
info18Sim           1.94225    0.19100  10.169  < 2e-16 ***
info31Sim          -1.97886    0.13587 -14.565  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3387.5  on 11139  degrees of freedom
Residual deviance: 2070.4  on 11131  degrees of freedom
AIC: 2088.4

Number of Fisher Scoring iterations: 8
```

Figure E.2 – R software output for the estimation of Model 1b

F. ROC curves of Model 1 and Model 2

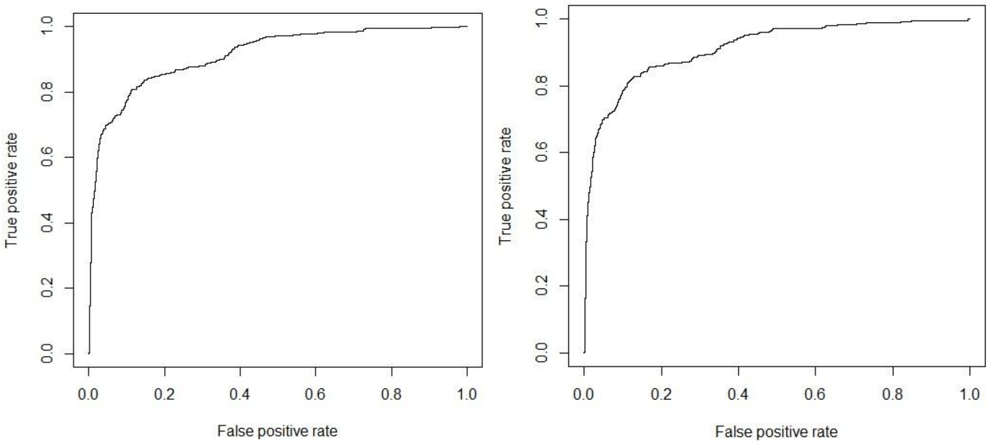


Figure F.1 – ROC curve for Model 1 (left) and for Model 2 (right)

G. Aggregate Loss probability density functions for models Po_A and Po_B for the year of 2015

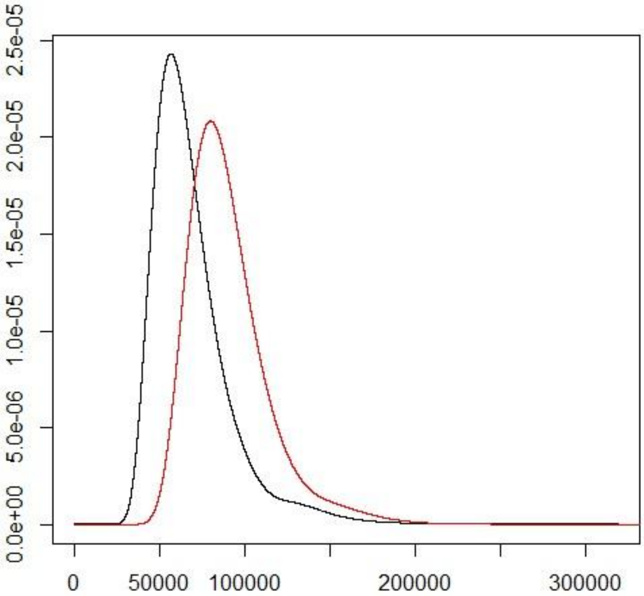


Figure G.1 – Compound Poisson aggregate loss probability density function for the year 2015 for Model Po_A (black) and for Model Po_B (red)

H. Expected value, standard deviation and skewness coefficients for the estimated Aggregate Loss

	Aggregate Loss		
Model	**Expected value**	**Standard deviation**	**Skewness coefficient**
Po_A	68 421	22 720	1.67
Po_B	90 141	24 313	1.41
NB1_A	68 421	78 617	1.98
NB1_B	90 141	102 093	2.04
NB2_A	68 421	105 116	2.81
NB2_B	90 141	137 380	2.87
Bi	68 421	21 099	1.87

Table H.I – Expected value, standard deviation and skewness coefficients for the estimated Aggregate Loss under some models

I. Percentage increase of NP and Translated Gamma approximation percentiles when compared to Panjer

Model	NP vs. Panjer					Translated Gamma vs. Panjer				
	90	95	97.5	99	99.5	90	95	97.5	99	99.5
Po_A	5.8%	3.9%	-0.2%	-0.3%	0.5%	2.6%	0.9%	-2.7%	-2.2%	-0.9%
Po_B	4.0%	2.6%	0.0%	-0.7%	-0.5%	2.1%	0.7%	-1.6%	-2.0%	-1.3%
NB1_A	9.7%	6.8%	4.2%	1.3%	-0.6%	0.9%	-0.6%	-1.7%	-2.8%	-3.4%
NB1_B	9.9%	7.5%	5.2%	2.5%	0.7%	0.7%	-0.4%	-1.1%	-1.9%	-2.4%
NB2_A	22.0%	16.6%	12.0%	7.2%	4.3%	1.5%	0.2%	-0.8%	-1.7%	-2.1%
NB2_B	22.3%	17.2%	12.8%	8.2%	5.3%	1.1%	0.1%	-0.5%	-1.1%	-1.4%

Table I.I – Percentage increase of the percentiles according to NP and Translated Gamma approximation percentiles when compared to percentiles obtained with Panjer’s algorithm.

9. BIBLIOGRAPHY

[1] Altman, E. I. (1968). Finantial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance , XXIII* (4), 589-609.

[2] Arthur Charpentier. (2015). *The R Series - Computational Actuarial Science with R.* Taylor & Francis Group.

[3] Carling, K., Jacobson, T., Lindé, J., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking & Finance , 31*, 845-868.

[4] Centeno, M. L. (2003). *Teoria do Risco na Actividade Seguradora.* (C. Editora, Ed.)

[5] Credit Suisse Financial Products. (1997). *CreditRisk+: A Credit Risk Management Framework.* London: Credit Suisse Financial Products.

[6] Gerber, H. U. (1990). *Life insurance Mathematics.* Zurich: Springer - Verlag Berlin Heidelberg and Swiss Association of Actuaries.

[7] Gerhold, S., Schmock, U., & Warnung, R. (2013). *A generalization of Panjer's recursion and numerically stable risk aggregation.*

[8] Gordy, M. B. (1998). A Comparative Anatomy of Credit Risk Models. Board of Governors of the Federal Reserve System.

[9] Gurny, P., & Gurny, M. (2013). Comparison of credit scoring models on probability of default estimation for US banks. *Prague Economic Papers , 2*, 163-181.

[10] Heller, G. Z. (2013). *Generalized Linear Models for Insurance Data.* Cambridge University Press.

[11] Huang, H., & Fang, K. (2011). Variable Selection for Credit Risk Model Using Data Mining Technique. *Journal of Computers , 6* (9), 1868-1874.

[12] Kern, M., & Rudolph, B. (2001). Comparative Analysis of Alternative Credit Risk Models- an Application on German Middle Market Loan Portfolios. *03*.

[13] Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2004). Loss Models - from Data to Decisions. Hoboken, New Jersey, United States of America: JohnWiley&SonsInc.

[14] Panjer, H. H. (1981). *Recursive Evaluation.* Ontario, Canada: Astin Bulletin 12.

[15] Real GDP growth rate in Portugal. Acceded on 14 October 2016, from PORDATA: http://www.pordata.pt/en/Portugal/Real+GDP+growth+rate-2298

[16] Ramsay, C. M. A note on the normal power approximation. *Astin Bulletin , 21* (1), 147-150.

[17] Sundt, B., Jewell, W.S., 1981. Further results on recursive evaluation of compound distributions. *Astin Bulletin 12.* 27-39.